Feature-matching Pattern-based Support Vector Machines for Robust Peptide Mass Fingerprinting*†§

Youyuan Li‡, Pei Hao§, Siliang Zhang‡¶, and Yixue Li§

Peptide mass fingerprinting, regardless of becoming complementary to tandem mass spectrometry for protein identification, is still the subject of in-depth study because of its higher sample throughput, higher level of specificity for single peptides and lower level of sensitivity to unexpected post-translational modifications compared with tandem mass spectrometry. In this study, we propose, implement and evaluate a uniform approach using support vector machines to incorporate individual concepts and conclusions for accurate PMF.

We focus on the inherent attributes and critical issues of the theoretical spectrum (peptides), the experimental spectrum (peaks) and spectrum (masses) alignment. Eighty-one feature-matching patterns derived from cleavage type, uniqueness and variable masses of theoretical peptides together with the intensity rank of experimental peaks were proposed to characterize the matching profile of the peptide mass fingerprinting procedure. We developed a new strategy including the participation of matched peak intensity redistribution to handle shared peak intensities and 440 parameters were generated to digitalize each feature-matching pattern. A high performance for an evaluation data set of 137 items was finally achieved by the optimal multi-criteria support vector machines approach, with 491 final features out of a feature vector of 35,640 normalized features through cross training and validating a publicly available “gold standard” peptide mass fingerprinting data set of 1733 items. Compared with the Mascot, MS-Fit, ProFound and Aldente algorithms commonly used for MS-based protein identification, the feature-matching patterns algorithm has a greater ability to clearly separate correct identifications and random matches with the highest values for sensitivity (82%), precision (97%) and F1-measure (89%) of protein identification.

Several conclusions reached via this research make general contributions to MS-based protein identification. Firstly, inherent attributes showed comparable or even greater robustness than other explicit. As an inherent attribute of an experimental spectrum, peak intensity should receive considerable attention during protein identification. Secondly, alignment between intense experimental peaks and properly digested, unique or non-modified theoretical peptides is very likely to occur in positive peptide mass fingerprinting. Finally, normalization by several types of harmonic factors, including missed cleavages and mass modification, can make important contributions to the performance of the procedure. Molecular & Cellular Proteomics 10: 10.1074/mcp.M110.005785, 1–17, 2011.

In MS-based proteomics, MS\textsuperscript{1} or MS\textsuperscript{2}, or even MS\textsuperscript{n}, data for peptides produced by proteolysis are obtained and used for peptide mass fingerprinting (PMF),\textsuperscript{1} peptide fragment fingerprinting (PFF), and de novo sequencing for qualitative analysis or quantification of proteins. State-of-the-art proteomics has adopted the use of tandem MS (MS/MS) because of its growing usefulness in protein identification (1). PMF is often used in conjunction with the PFF method because of the comparative lack of accuracy. However, the practical choice between PMF and PFF is the subject of debate. PMF was the first commonly used method for protein identification and is still in widespread use, sometimes as a filtering process before MS/MS, which is expensive and time-consuming (1, 2).

As summarized by Damodaran et al. (3), there are two remarkable advantages of PMF that are often ignored in comparison to MS/MS analysis of a single peptide. First, the mass matching and scoring process of PMF is less sensitive than that of MS/MS to unexpected post-translational modifications. Second, using multiple peptides in PMF allows more extensive coverage of the protein, whereas MS/MS can be confused by common peptides shared by a number of different proteins. Thus, in some cases, MS/MS can be less spe-

\textsuperscript{1} The abbreviations used are: MS/MS, Tandem mass spectrometry; PMF, Peptide Mass Fingerprinting; PFF, Peptide Fragment Fingerprinting; SVMs, Support Vector Machines; TS, Theoretical Spectrum; ES, Experimental Spectrum; TP, Theoretical Peptide; EP, Experimental Peak; PUD, Peptide Uniqueness Database; MPR, Matched Peak Intensity Redistribution; PCT, Peptide Cleavage Type; PUN, Peptide Uniquenesses; PMA, Peptide Mass Altering; PII, Peak Intensity Rank; FMP, Feature-matching Pattern.

Want to cite this article? Please look on the last page for the proper citation format.
Feature-matching Pattern-based SVM for Robust PMF

PMF searching tools: (1) the number of peptides matched; (2) algorithms. There are four important parameters used in most PMF methods to form an efficient scoring system. Shadforth et al. have provided excellent reviews of the evolution of PMF as a method for protein identification.

Several parameters have been formalized and used in PMF in order to form an efficient scoring system. Shadforth et al. (6) have reviewed 13 well-known and commonly used PMF algorithms. There are four important parameters used in most PMF searching tools: (1) the number of peptides matched; (2) the threshold for mass tolerance; (3) the percentage of sequence coverage; and (4) the number of missed cleavage sites. Additional parameters such as pI, molecular mass, post-modification, and chemical composition (7), methods such as a proteotypic peptide library (8) and integrated information (9) have been taken into account. The introduction of new parameters, such as spectrum similarity (10), mass spectrum alignment (11), peak bagging (12), negative ionization (13), probability-based scoring function (14), peak intensity prediction (15, 16), mass accuracy (17, 18), mass tolerance (19), and a validation system (20) has improved the accuracy of PMF identification.

Scoring algorithms vary in sensitivity, selectivity and accuracy, and each has advantages, disadvantages, and applications to each commonly used tool. In order to evaluate these algorithms, Chamrad et al. have reported the results of a study of performance in terms of specificity and sensitivity of the three most popular tools (21); i.e. Mascot (22), MS-Fit (23) and ProFound (24). They found the performance of Mascot and ProFound to be similar, correctly identifying around 53% of proteins from the reference set at a significance level of p < 0.05 but, surprisingly, MS-Fit correctly identified only 32% using the same input parameters.

In response to that disappointing evaluation of performance, new rigorous statistical methods have been used to integrate the scores returned by each algorithm. The strength of some packages can be exploited through such a consensus method, and weaknesses in others are mitigated, so that marginal identification can be corroborated or rejected by complementary packages (25, 26). Two generally similar analysis techniques termed multisearch and multipass have been developed for a tandem mass approach (27) and these consensus methods are now being adopted for large-scale projects (28) but are still not popular in smaller laboratories. The only significant shortcoming of these methods is the expense and difficulty in handling multiple search algorithms and scoring systems.

The objective of this study was to incorporate earlier conclusions and new strategies into one process. We proposed, implemented, and evaluated a machine learning approach to integrate a series of parameters for accurate and robust PMF. We focused on the inherent attributes and critical issues of theoretical spectrum (peptides), experimental spectrum (peaks), and spectrum (masses) alignment. Instead of assigning a statistical model or a scoring scheme for how well the identification was achieved, the machine learning approach, called support vector machines (SVMs), was applied to intelligently integrate a series of inherent features for protein identification without the dilemma of choosing a tradeoff between false positives and false negatives.

The study design was as follows: a multi-dimensional parameter space derived from theoretical peptides, experimental peaks, and mass alignment was initially constructed according to earlier conclusions and new strategies. A total of 81 feature-matching patterns for characterizing the PMF procedure were applied to extract features for the SVM approach. Second, a set of 35,640 normalized features vector was subjected to a multicriteria SVM to cross train and validate publicly available "gold standard" PMF sets of 1733 items. Finally, the optimal SVM model with 491 select features achieved a high-performance evaluation for standard PMF sets of 137 items.

EXPERIMENTAL PROCEDURES

Proteomics Data Set Collection—A high-quality data set, especially a training set, is essential to the machine learning approach. Highly confident (gold standard) PMF data should consist of a mass spectrum of the proteolytic digest of a single protein that has been positively identified by orthogonal methods, such as an antibody test, an ELISA assay, MS/MS-based identification, etc (19). We report our results using validated data from three different sources.

The Aurum data set presented by Falkner et al. (29) is a high-quality data set of 246 known human proteins expressed in Escherichia coli, purified by the use of N terminus T7 and His tags, further purified by SDS-PAGE and tryptic digestion and individually spotted 4 times for matrix-assisted laser desorption ionization/time of flight (MALDI-TOF)/TOF analysis. Data were acquired to represent at least the top 29 most intense MS peaks, and validated by the published decoy database.

The Avermectin data set presented by Yin (30) stemmed from a proteomics study of Streptomyces avermitilis protein complexes. With the optimized protein extraction protocol (31), the protein expression patterns of S. avermitilis during the onset of avermectin production in complex medium were analyzed. The obviously regulated proteins have been analyzed in detail by 2-dimensional electrophoresis TOF MS/MS analysis.

The YPRC data set was downloaded from ProteomeCommons.org, which was uploaded by the Yonsei Proteome Research Center. The identification results made available by Mascot and Aldente (32) can be downloaded from the ProteomeCommons.org Tranche network using given hash (see supplemental file, SI1-YPRCSourceHash).

Standard PMF Sets Construction—Raw spectra files (.t2d format) from the 3 proteomics datasets were subjected to software AuotDataexplorer@DataWisdom(v1.0), the automatic implementation of Data Explorer(v4.3), to extract m/z and automatically generate peak list files with peak intensity (.pmi format). To construct positive PMF sets, despite the pre-validation of each data set, all the proteomics data were subjected to software BatchMascot@DataWisdom(v1.0) to perform automated pipeline Mascot searching with default parameters. A PMF was assigned as positive only if the Mascot result was significant (expectation value E < 0.05) and the first ranked protein
was identical with the true protein confirmed by the original data set provider using other methods.

During a PMF search procedure, a set of experimental peaks is matched to a massive database of theoretical peptides produced by protein in silico digestion. Because only one protein is expected to be distinct out of hundreds and thousands of candidate proteins, the negative PMF sets should be larger than the positive sets in order to provide sufficient and various information to train the machine learning approach. For each of the positive PMF data, insignificant proteins (expectation value $E > 0.05$) from the top 2–10 ranked proteins in the Mascot result list were paired and assigned as negative sets. This procedure constructed a positive data set of 266 pair-wise PMF sets and a negative data set of 1604 pair-wise PMF sets from the three proteomics datasets (for detailed statistics, see supplemental file SI2-ProteomicDatasetStatistics). A ratio of 1:8 was applied to the positive plus negative PMF sets selected from the Aurum and Avermectin data sets to randomly extract the training set ($D_1$) and a validating set ($D_2$). All the PMF sets selected from the YPRC data set were assigned to the evaluating set ($D_3$). The statistics of each PMF set are given in Table I.

### Multi-dimensional Parameter Space Construction—A typical PMF searching procedure starts with comparing the experimental peak $m/z$ to the theoretical peptide mass. The three individual but complex objects of the experimental spectrum (peaks), the theoretical spectrum (peptides) and the spectrum (masses) alignment are worth intensive investigation during the PMF procedure (see Fig. 1).

Each object was characterized by explicit, implied and statistical parameters, which were classified into three categories, spectrum attributes, alignment statistics and SVM parameters, according to their formative stages as shown in Fig. 2.

The experimental spectrum, the theoretical spectrum and the spectrum alignment have parameters of 14, 17, and 11, respectively (see supplemental file SI3-Parameters Table). Those 42 parameters contributed mainly to the multi-dimensional parameter space. In this section, we discuss the newly applied parameters of each object.

### Newly Applied Parameters of the Experimental Spectrum (ES)—An experimental spectrum or peak list of length $n$ is a list $ES = \{esp_1, \ldots, esp_n\}$ of experimental peak $esp_i \in N \times l$

Each experimental peak has a $m/z$ $esPMZ_i \in N \subseteq R +$ and an original intensity (henceforth, refers specifically to peak relative intensity)

| PMF Set Name | Positive | Negative | Source/Species |
|--------------|----------|----------|---------------|
| Training     | 146      | 1184     | Aurum/Eukaryota + Avermectin/Bacteria |
| Validating   | 43       | 360      |               |
| Evaluating   | 77       | 60       | YPRC/Eukaryota |

### Table I

Statistics of each standard PMF sets. Training and validating sets are selected entirely from the YPRC data set

During a PMF search procedure, a set of experimental peaks is matched to a massive database of theoretical peptides produced by protein in silico digestion. Because only one protein is expected to be distinct out of hundreds and thousands of candidate proteins, the negative PMF sets should be larger than the positive sets in order to provide sufficient and various information to train the machine learning approach. For each of the positive PMF data, insignificant proteins (expectation value $E > 0.05$) from the top 2–10 ranked proteins in the Mascot result list were paired and assigned as negative sets.

This procedure constructed a positive data set of 266 pair-wise PMF sets and a negative data set of 1604 pair-wise PMF sets from the three proteomics datasets (for detailed statistics, see supplemental file SI2-ProteomicDatasetStatistics). A ratio of 1:8 was applied to the positive plus negative PMF sets selected from the Aurum and Avermectin data sets to randomly extract the training set ($D_1$) and a validating set ($D_2$). All the PMF sets selected from the YPRC data set were assigned to the evaluating set ($D_3$). The statistics of each PMF set are given in Table I.

### Multi-dimensional Parameter Space Construction—A typical PMF searching procedure starts with comparing the experimental peak $m/z$ to the theoretical peptide mass. The three individual but complex objects of the experimental spectrum (peaks), the theoretical spectrum (peptides) and the spectrum (masses) alignment are worth intensive investigation during the PMF procedure (see Fig. 1).

Each object was characterized by explicit, implied and statistical parameters, which were classified into three categories, spectrum attributes, alignment statistics and SVM parameters, according to their formative stages as shown in Fig. 2.

The experimental spectrum, the theoretical spectrum and the spectrum alignment have parameters of 14, 17, and 11, respectively (see supplemental file SI3-Parameters Table). Those 42 parameters contributed mainly to the multi-dimensional parameter space. In this section, we discuss the newly applied parameters of each object.

### Newly Applied Parameters of the Experimental Spectrum (ES)—An experimental spectrum or peak list of length $n$ is a list $ES = \{esp_1, \ldots, esp_n\}$ of experimental peak $esp_i \in N \times l$

Each experimental peak has a $m/z$ $esPMZ_i \in N \subseteq R +$ and an original intensity (henceforth, refers specifically to peak relative intensity)

### Table I

Statistics of each standard PMF sets. Training and validating sets are selected from both the Aurum and the Avermectin data sets. The ratio of positive to negative data is approximately 1:8. Evaluation sets are selected entirely from the YPRC data set

| PMF Set Name | Positive | Negative | Source/Species |
|--------------|----------|----------|---------------|
| Training     | 146      | 1184     | Aurum/Eukaryota + Avermectin/Bacteria |
| Validating   | 43       | 360      |               |
| Evaluating   | 77       | 60       | YPRC/Eukaryota |

During a PMF search procedure, a set of experimental peaks is matched to a massive database of theoretical peptides produced by protein in silico digestion. Because only one protein is expected to be distinct out of hundreds and thousands of candidate proteins, the negative PMF sets should be larger than the positive sets in order to provide sufficient and various information to train the machine learning approach. For each of the positive PMF data, insignificant proteins (expectation value $E > 0.05$) from the top 2–10 ranked proteins in the Mascot result list were paired and assigned as negative sets.

This procedure constructed a positive data set of 266 pair-wise PMF sets and a negative data set of 1604 pair-wise PMF sets from the three proteomics datasets (for detailed statistics, see supplemental file SI2-ProteomicDatasetStatistics). A ratio of 1:8 was applied to the positive plus negative PMF sets selected from the Aurum and Avermectin data sets to randomly extract the training set ($D_1$) and a validating set ($D_2$). All the PMF sets selected from the YPRC data set were assigned to the evaluating set ($D_3$). The statistics of each PMF set are given in Table I.

### Multi-dimensional Parameter Space Construction—A typical PMF searching procedure starts with comparing the experimental peak $m/z$ to the theoretical peptide mass. The three individual but complex objects of the experimental spectrum (peaks), the theoretical spectrum (peptides) and the spectrum (masses) alignment are worth intensive investigation during the PMF procedure (see Fig. 1).

Each object was characterized by explicit, implied and statistical parameters, which were classified into three categories, spectrum attributes, alignment statistics and SVM parameters, according to their formative stages as shown in Fig. 2.

The experimental spectrum, the theoretical spectrum and the spectrum alignment have parameters of 14, 17, and 11, respectively (see supplemental file SI3-Parameters Table). Those 42 parameters contributed mainly to the multi-dimensional parameter space. In this section, we discuss the newly applied parameters of each object.

### Newly Applied Parameters of the Experimental Spectrum (ES)—An experimental spectrum or peak list of length $n$ is a list $ES = \{esp_1, \ldots, esp_n\}$ of experimental peak $esp_i \in N \times l$

Each experimental peak has a $m/z$ $esPMZ_i \in N \subseteq R +$ and an original intensity (henceforth, refers specifically to peak relative intensity)

| PMF Set Name | Positive | Negative | Source/Species |
|--------------|----------|----------|---------------|
| Training     | 146      | 1184     | Aurum/Eukaryota + Avermectin/Bacteria |
| Validating   | 43       | 360      |               |
| Evaluating   | 77       | 60       | YPRC/Eukaryota |

### Table I

Statistics of each standard PMF sets. Training and validating sets are selected from both the Aurum and the Avermectin data sets. The ratio of positive to negative data is approximately 1:8. Evaluation sets are selected entirely from the YPRC data set

### Feature-matching Pattern-based SVM for Robust PMF

#### Fig. 1. An illustration of the PMF spectrum alignment. There are three objects in PMF-based protein identification. The experimental peak $m/z$ is matched to the theoretical peptide mass with given mass tolerance. In this illustration, experimental peaks are sorted by intensity rank and theoretical peptide masses are sorted by cleavage order. There are one-to-one, one-to-many and many-to-many matching in the spectrum alignment because of the mass tolerance.

$$esPI_i \in I \subseteq R +$$

where $R +$ means the set of nonnegative real numbers.

**Peak Intensity Proportion (esPIP)**—The extraction of $m/z$ from spectra is a signal processing task that cannot be done perfectly. Data from this domain are always very noisy and contain errors introduced by preprocessing steps in the wet laboratory as well as in the signal processing. Values of peak original intensity can be distorted because of the inconsistent scale of different spectra. Therefore, esPIP was proposed to normalize the flexible peak original intensity:

$$esPIP = \frac{esPI_i}{\sum_{j=1}^{n} esPI_j} \times 100\%$$

Here, the individual peak original intensity is divided by the sum of all intensities, so that the sum of the peak intensity proportion in each peak list equals 1:

$$\sum_i esPI_i = 1.$$  

**Peak Intensity Rank (esPIR)**—In the inherent attribute esPIR was introduced to fully eliminate the absolute rank of a given peak original intensity or intensity proportion. We obtained a new parameter, the experimental peak intensity rank

$$esPIR_i < esPIR_j \quad (i < j, 1 \leq i, j \leq |ES|)$$

when a typical experiment spectrum, normally displayed by $m/z$ order, was sorted by descending order of peak intensity. The esPIR of highest intense peak is 1.

**Parameter Series of Top Matching**—Parameter series of top matching describe the matching profile of the top intense peaks in the experimental spectrum. The range of top intensity rank (TopIR) is initialized with 10, and it will be optimized in a later section.

The matched peak intensity set ($esPIs_{\text{matched}}$) and the matched intensity ranks set ($esPIRs_{\text{matched}}$) of $t$ matched experimental peaks are defined as:
esPIRs matched

- \{esPI: all the intensity ranks of matched experimental peaks\} of length t

esPIsi matched

- \{esPI: all the intensities of matched experimental peaks\} of length t

esPIPsi matched

- \{esPIP: all the intensity proportions of matched experimental peaks\} of length t.

The top matching value (esTMV), top matching peak intensities esTMI and top matching peak intensities proportion (esTMIP) are calculated as:

\[
esTMV = \sum_{i=1}^{t} 2^{TopIR_i} \times \text{esPIRs}_i \text{matched}
\]

\[
esTMI = \sum_{i=1}^{t} (2^{TopIR_i} \times \text{esPIsi}_i \text{matched}) \times \text{esPIsi}_i \text{matched}
\]

\[
esTMIP = \sum_{i=1}^{t} (2^{TopIR_i} \times \text{esPIPsi}_i \text{matched}) \times \text{esPIPsi}_i \text{matched}
\]

The harmonic factors total matched peak intensities (esMI) and total matched peak intensities proportion (esMIP) are calculated as:

\[
esMI = \sum_{i=1}^{t} \text{esPIsi}_i \text{matched}
\]

\[
esMIP = \sum_{i=1}^{t} \text{esPIPsi}_i \text{matched}
\]

Then, four harmonized parameters esTMI/MI, esTMV/TMV, esTMIP/MIP, and esTMIP/TMV are calculated as:

\[
esTMI/MI = \frac{esTMI}{esMI}
\]

\[
esTMV/TMV = \frac{esTMV}{esTMV}
\]

\[
esTMIP/MIP = \frac{esTMIP}{esMIP}
\]

\[
esTMIP/TMV = \frac{esTMIP}{esTMV}
\]

Newly Applied Parameters of the Theoretical Spectrum (TS)—A theoretical spectrum or peptide list of length m is a list of theoretical peptides tsPj \( \in M \). Each peptide has a mass:

\[
tsPM_j \in M \subseteq \mathbb{R} +
\]

Peptide Cleavage Order (tsPCO)—In TS = \{tsP_1, ..., tsP_m\}, the subscript numbers from 1 to m indicate the original order of the theoretical cleavage sites from N terminus to C terminus in the protein sequence. In other words, those theoretical peptides are located in the protein sequence according to its tsPCO from N terminus to C terminus.

Peptide Cleavage Type (tsPCT)—tsPCT indicates the type of cleavage that generated the peptide, which can be proper cleavage or missed cleavage. We have divided the missed cleavage type into theoretical missed cleavage and random missed cleavage. Siepen (33) presented a simple predictive method based on information theory that was able to identify experimentally defined missed cleavages with up to 90% accuracy from amino acid sequence alone. We used the web tool available at http://ispider.smith.man.ac.uk/Missed-Cleave to predict theoretical missed cleavage sites from random missed cleavages.

Peptide Mass Altering (tsPMA)—Because of differences of protease specificity, chemical composition of the protein and digestion...
environment, the in silico digestion should take differences of mass into consideration for theoretical peptides. Here, tsPMA is commonly classified into three categories of modification: none, fixed, and variable.

**Peptide Uniqueness (tsPUN)** —Another implied property for each theoretical digested peptide is the repeat number of its sequence in the whole protein sequence database. A peptide is classified as unique if it exists in only one protein. A local secondary database PUD (peptide uniqueness database) derived from SwissProt (version 57.12) has been constructed to generate tsPUN for each peptide.

**Parameter Series of Consecutive Matching**—Presumably, when matched theoretical peptides are distributed contiguously rather than randomly within a protein sequence, the protein is more likely to be a true protein rather than a false one. Neighbor-matching concepts, such as KSDP (34) and NMOWSE (14), have been used successfully in MS/MS and PMF, respectively. In this study, we expanded the parameter series of consecutive matching to describe the matching profile of a theoretical spectrum.

The consecutively matched peptides number set (tsCMN\(_\text{sc\_matched}\)) and consecutively matched intensities set (tsMIs\(_\text{sc\_matched}\)) of \(c\) segments of consecutively matched theoretical peptides are defined as:

\[
\text{tsCMN}\_\text{sc\_matched} = \{\text{the number of cleavage order in each segment} - 1: \text{segments consisting of consecutive cleavage orders of matched theoretical peptides}\}\]

\[
\text{tsMIs}\_\text{sc\_matched} = \{\text{sum of experimental peak intensity matched to consecutive cleavage orders in each segment: segments consisting of consecutive cleavage orders of matched theoretical peptides}\}\]

The discretely matched peptides number (tsCMN\(_\text{d\_matched}\)), discretely matched intensities (tsMIs\(_\text{d\_matched}\)) and discretely matched intensities proportions (tsMIsP\(_\text{d\_matched}\)) of all discretely matched theoretical peptides are defined as:

\[
\text{tsCMN}\_\text{d\_matched} = \text{sum of the number of discrete single cleavage order of matched theoretical peptides}
\]

\[
\text{tsMIs}\_\text{d\_matched} = \text{sum of the experimental peak intensity matched to the discrete single cleavage order of matched theoretical peptides}
\]

\[
\text{tsMIsP}\_\text{d\_matched} = \text{sum of the experimental peak intensity proportion matched to the discrete single cleavage order of matched theoretical peptides}
\]

The consecutive matching value (tsCMV), consecutive matching peak intensities (tsCMI) and consecutive matching peak intensities proportions (tsCMIP) are calculated as:

\[
\text{tsCMV} = \sum_{j=1}^{c} 2^{\text{tsCMN}\_\text{d\_matched}}
\]

\[
\text{tsCMI} = \sum_{j=1}^{c} (2^{\text{tsCMN}\_\text{d\_matched}} \times \text{tsMIs}\_\text{d\_matched})
\]

\[
\text{tsCMIP} = \sum_{j=1}^{c} (2^{\text{tsCMN}\_\text{d\_matched}} \times \text{tsMIsP}\_\text{d\_matched})
\]

To take average peak intensity into consideration, the consecutive matching average peak intensities (tsCMAI) and consecutive matching average peak intensities proportions (tsCMAPI) are varied as:

\[
\text{tsCMAI} = \sum_{j=1}^{c} \left(2^{\text{tsCMN}\_\text{d\_matched}} \times \frac{\text{tsMIs}\_\text{d\_matched}}{\text{tsCMN}\_\text{d\_matched} + 1}\right)
\]

\[
\text{tsCMAPI} = \sum_{j=1}^{c} \left(2^{\text{tsCMN}\_\text{d\_matched}} \times \frac{\text{tsMIsP}\_\text{d\_matched}}{\text{tsCMN}\_\text{d\_matched} + 1}\right)
\]

**Newly Applied Parameters of Spectrum Alignment**

**Matched Peak Intensity Redistribution (MPIR)** —A spectrum alignment is commonly done by matching masses with a predefined mass tolerance between two spectra of

\[
\text{TS} = \{\text{tsp}_{1}, \ldots, \text{tsp}_{m}\}
\]

\[
\text{ES} = \{\text{esp}_{1}, \ldots, \text{esp}_{n}\}
\]

A single and simple matching is a bijective process that can be described as:

\[
\text{tsp}(\text{tsPMZ}, \text{tsPCO}_i, \text{tsPCT}_j, \text{tsPMA}, \text{tsPUN}) \rightarrow \text{saMME}_{ij} \rightarrow \text{esp}(\text{esPM}_i, \text{esPI}_j, \text{esPIP}, \text{esPIR})
\]

where saMME is matched mass error in spectrum alignment:

\[
\text{saMME}_{ij} = |\text{tsPMZ} - \text{esPM}_i|
\]

However, the bijective procedure would result in a many-to-many rather than a bijection matching because of the mass tolerance. Many-to-many mapping introduces a new difficulty in recalculating matched peak intensity for each matching. It is known that peak intensity is additive. The -trimmed mean has been used to deal with one-to-many (15). For the many-to-one case, two or more theoretical peptides that are matched to the same experimental peak should share the intensity of the latter. In our study, a new strategy that we have named matched peak intensity redistribution (MPIR) was used to recalculate peak intensity or peak intensity proportion for each matched theoretical peptide.

Because of different types of cleavage and peptide modification, a given peptide will generate several masses during proteolysis. As shown in Fig. 3, there are nine probable approaches to generate differ-
Feature-matching Pattern-based SVM for Robust PMF

Fig. 3. Probable approaches to generate theoretical peptide masses for a given peptide. During proteolysis, a peptide can be generated through any of three types of cleavage: proper, theoretical missed, and random missed. The peptide can suffer no modification or one or more fixed or variable modifications. The theoretical peptide masses have different distributions of occurrence. Instead of evaluating the prior probability of each occurrence, we assign a distribution weight to each approach. Each of 9 theoretical peptide masses for a given peptide has a distribution weight which is the product of the weights of its cleavage type and peptide modification.

During proteolysis, a peptide can be generated through any of three types of cleavage: proper, theoretical missed, and random missed. The peptide can suffer no modification or one or more fixed or variable modifications. The theoretical peptide masses have different distributions of occurrence. Instead of evaluating the prior probability of each occurrence, we assign a distribution weight to each approach. Each of 9 theoretical peptide masses for a given peptide has a distribution weight which is the product of the weights of its cleavage type and peptide modification.

The optimized weights are inspected in a later section.

In MPIR, a matched peptides cluster (MPC) contains several theoretical peptides with very close masses. A many-to-one matching can be redescribed as:

Feature-matching Pattern (FMP) and SVM Feature Extraction—In this study, we assigned the three attributes of PCT, PUN, and PMA to a theoretical peptide and one attribute, PIR, to the experimental peak as shown in Fig. 4A. Each attribute has two values, which are digitalized as 1 and 0. When considering all the values of one attribute, an asterisk (*) is assigned to it and then each attribute has three states: 0, 1, and *

A PMF procedure consists of a set of mass alignments between theoretical peptides and experimental peaks. For a given matching, each attribute of the related theoretical peptide and the experimental peak has one certain status and those four items of status generate one FMP for a given mass alignment. The number of all possible
In an attempt to describe the matching profile of a PMF procedure in as much detail as possible, we calculated 440 SVM features for each FMP as described below.

First, 29 statistical SVM parameters from parameter space are calculated for each FMP. To eliminate the negative effect of the number of experimental peaks, theoretical peptides and mass alignment, parameters 15–23, 25–34, and 36–42 are normalized by \( \text{esMatchedNum} \) (\( \text{esMN} \), the number of matched experimental peaks), \( \text{tsMatchedNum} \) (\( \text{tsMN} \), the number of matched theoretical peptides), and \( \text{saAlignNum} \) (\( \text{saAN} \), the number of mass alignment) respectively, which increases the number of SVM parameters from 29 to 55. Further, these 55 parameters are normalized by eight harmonic factors to eliminate the effect of protein size (see supplemental file SI5-HF).

Hence, for a spectrum alignment, there are 35,640 features calculated from 81 (the number of feature-matching pattern) multiplied by 55 (the number of parameters for one feature-matching pattern) and by eight (harmonic factors).

SVM Implementation with Multi-criteria Evaluation—SVM is a supervised learning method that is widely used to solve binary classification problems; we used the LIBSVM 2.91(35) software in this work. A radial basis function was selected as the kernel function. Two parameters, the regularization parameter \( C \) and the kernel width parameter \( g \), were optimized using a grid search approach.

In this study, positive instances were collected from double-checked PMF data sets, and there were eight times more negative instances than positive instances. However, for handling unbalanced datasets, accuracy might not be a good criterion for evaluating a model. Maximizing accuracy would encourage predictors to favor the majority group; for example, a predictor will achieve an accuracy of 88.89% in a 1:8 positive/negative data set by simply predicting all samples as negative. We applied implementation of binary class cross-validation with different criteria (36) to enable LIBSVM to conduct cross-validation and prediction with respect to different criteria. The six measurements used in our SVM approach are given in Table II.

F-score-based Binary Partial Feature Selection—In addition to setting the proper parameters, feature subset selection can improve the accuracy of SVM classification. The F-score is a simple technique that measures the discrimination of two sets of real numbers. Given training vectors \( x_k \), \( k = 1, 2, \ldots, m \), if the number of positive and negative instances are \( n_+ \) and \( n_- \), respectively, then the F-score of the \( i \)th feature is defined as (37):

---

**Fig. 4. Construction of 81 FMPs in SVM feature space.** There are three statuses for each attribute of a theoretical peptide and experimental peak. **A**, Status * means consideration of both 1 and 0. For single mass alignment, the number of all possible attribute combination is 81 (3^4). All these attribute combinations provide 81 FMPs for spectrum alignment in the SVM feature space. **B**, A total of 55 parameters multiplied by 8 harmonic factors are calculated for each FMP. The 35,640 (81 × 55 × 8) features are assembled to describe the PMF procedure in detail.
Feature-matching Pattern-based SVM for Robust PMF

**TABLE II**

| Measurement                  | Abbreviation | Equation |
|------------------------------|--------------|----------|
| Precision                    | Prec.        | TP/(TP + FP) |
| Sensitivity (Recall)         | Sens.        | TP/(TP + FN) |
| Accuracy                     | Acc.         | (TP + TN)/(TP + TN + FP + FN) |
| F-score                      | Fs.          | 2 × Precision × Recall/(Precision + Recall) |
| Balanced Accuracy            | BAC          | (Sensitivity + Specificity)/2 |
| Area under the ROC Curve     | AUC          |          |

**TABLE III**

| Parameter                        | Algorithm                |
|----------------------------------|--------------------------|
| Database                         | SwissProt                |
| Taxonomy                         | Homo sapiens(human)     |
| Maximum missed cleavage sites    |                          |
| Cleavage enzyme                  | Trypsin                  |
| Fixed modifications              |                          |
| variable modification            | Carbamidomethyl(C)       |
| Mass Type, Charge state          | Oxidation(M)             |
| Mass tolerance                   | Monoisotopic, MH⁺        |
|                                  | 100ppm                   |

$$F(i) = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (x_{ik}^{(+)} - \bar{x}_i)^2 + \frac{1}{n_+ - 1} \sum_{k=1}^{n_+} (x_{k}^{(+)} - \bar{x}_i)^2$$

Where $\bar{x}_i$, $\bar{x}_i^{(+)}$, and $\bar{x}_i^{(-)}$ are the averages of the $i$th feature of the whole, positive and negative data sets, respectively; $\bar{x}_i^{(+)}$ is the $i$th feature of the $k$th positive instance, and $\bar{x}_i^{(-)}$ is the $i$th feature of the $k$th negative instance. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the discrimination within each set. The larger the $F$-score, the more likely this feature is to be more discriminative.

Like a binary search, which searches a sorted array by repeatedly halving the search interval, the $F$-score-based binary partial feature selection begins with fivefold cross-validation of the whole sorted feature array, which is then divided into two parts according to the mean $F$-score. If fivefold cross-validation of the upper partial features is not less than that of the initial whole features, the new partial features are restricted to the upper half; otherwise, the lower half is included. The new partial features are repeatedly cross-validated until the minimum but sufficient feature set is found.

The $F$-score-based binary partial feature selection procedure consists of five steps:

1. Calculate the $F$-score for each feature. Here, we use modified script $fselect.py$ (37) to calculate $F$-score only.

2. Sort $F$-score by descending order; use fivefold cross-validation to train the training set $D_u$ with whole features and obtain a predictor. Use the predictor to predict the validating set $D_v$. Assign the evaluation value $AUC$ with whole features as initial $EV_+$ and the evaluation value $AUC$ of one feature with the highest $F$-score as initial $EV_-$. Calculate the mean $F$-score of the whole features.

3. Select features whose $F$-scores are higher than the mean to be the new partial features. Use fivefold cross-validation to train the training set $D_u$ with the new partial features and obtain a new predictor. Use the new predictor to predict validating set $D_v$. Assign the evaluation value $AUC$ as the new $EV_+$ if it is more sufficient than the previous $EV_+$; otherwise, assign it as the new $EV_-$. Repeat step 3 based on the new mean $F$-score.

4. Calculate the mean $F$-score of the new partial features assigned $EV_+$ and $EV_-$. Repeat step 3 based on the new mean $F$-score.

5. Repeat step 4 until the minimum but sufficient feature set is found.

**Comparison Workflow of PMF-based Algorithms**

Implementation of PMF-based Algorithms—The feature-matching pattern-based algorithm described here was compared with the most commonly used PMF-based algorithms, Mascot (http://www.matrixscience.com), MS-Fit (http://prospector.ucsf.edu), ProFound (http://prowl.rockefeller.edu/prowl-cgi/profound.exe) and Aaldente (http://expasy.org/tools/aldente). A series of automatic implementation of the four algorithms (BatchMascot@DataWisdom(v1.0), BatchMS-Fit@DataWisdom(v1.0), BatchProFound@DataWisdom(v1.0) and BatchAldente@DataWisdom(v1.0)) were used for automated pipeline database searching and to extract and store the information necessary for further analysis (38). In order to establish an objective comparison, a set of standard parameter and a standard comparing set ($D_C$) were developed for all PMF-based approaches. The comparing set ($D_C$) was selected from the Aurum data set, which contains 225 proteins with identity confirmed by analysis of at least two peptides via tandem mass spectrometry (see supplemental file, S16-ComparingSetResult). All the PMF sets from $D_C$ were assigned as positive.

The standard search parameters are summarized in Table III. The variable oxidation of methionine and the complete carbamidomethylation of cysteine were investigated because of the sample preparation procedure (19).

Evaluation Criteria—The PMF search results were assessed automatically using rigorous criteria. The evaluation was based on the first identifications ranked by the PMF programs; all other ranks were ignored. The putative protein hits were classified on the basis of their score and whether they were correct. The hit was assigned as true positive if the score for a correct protein hit was above the threshold or as false positive if the score was below the threshold. Each algorithm has its own significance criterion to evaluate the quality of its search result (24). The thresholds for each algorithm are given in the supporting information (see supplemental file, S16-ComparingSetResult).

We use standard performance metrics in information retrieval, including precision, recall and F1-measure, to evaluate the identification.
Feature-matching Pattern-based SVM for Robust PMF

RESULTS

Feature Selection—The accuracy, speed and interpretability of a machine learning approach can be improved by reducing the number of features. The main task for feature selection is to reduce the size of input feature space and reveal only the discriminant features for a given task.

First, a feature matrix of 1743 items from $D_T$ and $D_V$ plus 35,640 features was calculated. The F-scores of each feature were calculated and sorted by descending order (see supplemental file SI7-Fscore_35640Features). Second, we obtained a minimum but sufficient partial features of 491 after cross-validation 14 times in six rounds according to the procedure of F-score-based binary partial feature selection (see supplemental file SI8-BinaryPartialFeatureSelection).

The 11 out of 55 SVM parameters that contributed to 491 select SVM features are given in Table IV. Traditional parameter series of counting, $tsMatchedNum$ and $esMatchedNum$ play the same important roles in other protein identification methods. $tsMatchedNum$ and $esMatchedNum$ are used to calculate sequence coverage and mass value matched. However, there is an important distinction in the connotative meaning of such parameters. Unlike the new parameter $saAlignNum$ used to describe spectrum alignment, $tsMatchedNum$ and $esMatchedNum$ do not score the whole matched peaks or peptides (refer to FMP HF0_****) but describe mainly other special feature matching patterns (see supplemental file SI9-MatchedNumDistribution).

Table V ranks 21 out of 81 FMPs that contributed to the 491 select SVM features. It is interesting that FMP ****1 has displaced feature matching pattern **** (rank 2) and ***0 (rank 17) as the top rank, which indicates that experimental peaks with top rank intensity contribute more to the performance than any other peak.

Fig. 5 demonstrates the contribution of four attributes and their own three statuses. At first glance, statuses 1 and * of all attributes were the main contributors to the performance. There are four important points: first, proper digestion of theoretical peptides contributes more than missed digestion. Second, unique peptides other than repetitive peptides have a marked effect on performance. Third, changes of mass including fixed and variable modifications play a supplementary role for the original mass. Lastly, top-rank intensity experimental peaks have a dominant effect on performance and

| Table IV | Average F-score of 11 select SVM parameters that contributed to the SVM approach |
|----------|---------------------------------------------------------------------------------|
| Rank     | Average F-score            | Parameter ID       | Count | Object                                      |
| 1        | 2.947693                   | $esMatchedNum$     | 66    | Experimental Spectrum                       |
| 2        | 2.937659                   | $esMatchedIntensities$ | 65    | Experimental Spectrum                       |
| 3        | 2.915842                   | $esTopMatchingValue$ | 64    | Experimental Spectrum                       |
| 4        | 2.906978                   | $saAlignNum$       | 65    | SpectrumAlignment                           |
| 5        | 2.844247                   | $tsMatchedNum$     | 60    | Theoretical Spectrum                        |
| 6        | 2.793061                   | $esTopMatchingIntensities$ | 44    | Experimental Spectrum                       |
| 7        | 2.714127                   | $esTopMatchingIntensities/TopMatchingValue$ | 74    | Experimental Spectrum                       |
| 8        | 2.563682                   | $esMatchedIntensities$ | 41    | Experimental Spectrum                       |
| 9        | 2.553685                   | $tsTopMatchingIntensitiesProportion$ | 8     | Experimental Spectrum                       |
| 10       | 2.271217                   | $tsConsecutiveMatchingValue_All$ | 3     | Theoretical Spectrum                        |
| 11       | 2.206425                   | $tsConsecutiveMatchingValue_All/tsMatchedNum$ | 1     | Theoretical Spectrum                        |

| Table V  | Average F-score of 21 select feature matching patterns that contributed to the SVM approach |
|----------|---------------------------------------------------------------------------------|
| Rank     | Average F-score            | FMP Count | Rank | Average F-score            | FMP Count | Rank | Average F-score            | FMP Count |
| 1        | 3.260269                   | "****1"  | 49    | 2.739475                   | "11"     | 32    | 2.592058                   | 111*      | 10    |
| 2        | 3.098448                   | "****"   | 34    | 2.714126                   | "11"     | 25    | 2.577115                   | 1111      | 19    |
| 3        | 2.965369                   | "1**"    | 59    | 2.699814                   | "11"     | 36    | 2.546275                   | "**0"     | 5     |
| 4        | 2.919238                   | "1**"    | 40    | 2.686240                   | "11"     | 18    | 2.497017                   | 110*      | 4     |
| 5        | 2.812000                   | "111"    | 34    | 2.657608                   | "11"     | 19    | 2.341434                   | 10*       | 4     |
| 6        | 2.809335                   | "111"    | 37    | 2.630326                   | "1**"    | 23    | 2.334302                   | 10*       | 4     |
| 7        | 2.770610                   | "1**"    | 26    | 2.629654                   | "11**"   | 20    | 2.232498                   | 110*      | 1     |

**TP:** the number of true positives (correct hits with scores above the threshold)  
**FP:** the number of false positives (incorrect hits with scores above the threshold)  

Precision $= \frac{TP}{TP + FP}$, the proportion of identified ground-truth proteins to all identified proteins.

Sensitivity $= \frac{TP}{P}$, the proportion of identified ground-truth proteins to all ground-truth proteins.

$F1$-measure $= 2 \times \text{precision} \times \text{sensitivity} / (\text{precision} + \text{sensitivity})$, the harmonic mean of precision and sensitivity.
experimental peaks with medium and low intensity have a comparable importance at the same time.

Average $F$-scores of each harmonic factor are given in Table VI. Normalization by different harmonic factors has higher average $F$-scores than the original value, which proves that it is necessary to eliminate the bias derived from large proteins in the experimental spectrum.

Optimizations in the later studies were based on the 491 selected features.

**Optimization of MPIR Parameters**—We used two steps (orthogonal test design plus response surface methodology) to optimize four parameters: peptide mass distribution weights for missed cleavage and mass altering in MPIR.

First, an orthogonal test design $L_9(3)^4$ was used to find the appropriate range for each parameter. Nine sets of MPIR (weight of proper cleavage and none mass modification are pre-assigned 1) were applied to $D_T$, which gave nine values of CV AUC for each run. Four parameter ranges of MPIR were determined from their effects on the CV AUC of $D_T$ (for a table of orthogonal test design, results and analysis, see supplemental file SI10–491-MPIR-OTD).

Second, response surface methodology was used to find the optimum parameters. Low and high actual values of each parameter were set according to the parameter ranges determined from orthogonal test design (see supplemental file SI11–491-MPIR-RSM). A total of 30 runs of MPIR were
applied to $D_t$, $D_v$, and $D_e$ to generate SVM sets (weight of proper cleavage and non-modification of mass are pre-assigned 1). We analyzed the results from 30 runs of SVM performance on $D_e$ and derived a polynomial formula to fit the experimental data. The $F$-value of the model was significant, which made the model practicable. Three curves (Fig. 6) illustrate the interaction of the parameters theoretical missed cleavage, random missed cleavage, fixed modification and variable modification on the balanced accuracy (BAC) of $D_e$.

The predictive optimum values for $A$ (theoretical missed cleavage), $B$ (random missed cleavage), $C$ (fixed modification), and $D$ (variable modification) are 0.75, 0.85, 1.09, and 0.33, respectively. Contours for each interactive parameter are shown in Fig. 7.

The optimum values of each MPIR parameter were used in $D_t$ and $D_v$ to evaluate the optimum performance. The optimum BAC of $D_e$ is 0.883117, which is among the highest ranks of performance and those optimum MPIR parameters were used in the next part of the study.

**Optimization of the Number of Experimental Peaks**—The extraction of PMF from experimental spectra is a signal processing task that is affected by mass calibration, S/N, threshold of m/z and intensity etc. Except peak mass and intensity, the number of peaks extracted from a given experimental spectrum can vary between laboratories and even between different operators in the same laboratory. However, neither insufficient nor redundant peaks have a positive influence on protein identification. The former cannot provide sufficient information and the latter might mislead the algorithm by offering redundant information. For these reasons, it is necessary to optimize the peak number used for protein identification. We used a progressive approach strategy to determine the optimum peak number.

First, the top $2^m \times 10^0 \leq n \leq 6$ peaks of each PMF set in $D_t$, $D_v$, and $D_e$ were selected and used for the SVM approach. Training models from $D_t$ were applied to evaluate $D_e$ (see supplemental file SI12–491-EPNum). A range $2 \leq n \leq 4$ of potential high performance was determined by analyzing 7 groups of performance evaluation of $D_e$ (see Fig. 8A).

Second, progressive peak numbers were further investigated by refining $n$ between 2 and 4. By approaching the highest AUC, BAC, precision and accuracy of $D_e$, the optimum experimental peak number was determined as 75, which was kept constant for the rest of the study.

**Optimization of the Top Rank Range of Experimental Peak Intensity**—On the basis of the inference that intense peaks should contribute more to protein identification, a parameter series of top matching was aimed mainly at highlighting the effect of intense peaks on performance. It is useful to determine the top rank of experimental peak intensity. Here, we used a progressive approach to determine the optimum top rank.

First, top rank ranges of $2^m \times 5^0 \leq m \leq 4$ with an additional 0 were used to train $D_v$ and to predict $D_v$. Training models were then used to evaluate $D_e$ (see supplemental file SI13–491-TIR). A range $2 \leq m \leq 3$ of potential high performance was found by analyzing 6 groups of the evaluation parameter $D_e$ (see Fig. 8B).

Second, progressive top rank ranges were investigated further by refining $m$ between 2 and 3. By approaching an AUC of 1, precision of 100% and the greatest accuracy, recall, $F$-score and BAC of $D_e$, the top rank range of experimental peak intensity was set at 30 as the default optimum value.

**Comparative Evaluation of PMF-based Protein Identification Algorithms**—Standard MALDI fingerprint data sets (255 items) were analyzed with Mascot, MS-Fit, ProFound, Aldenet and FMP using a set of identical search parameters (see supplemental file, SI6-ComparingSetResult) and Table VII gives the performance comparison.

Mascot and MS-Fit have achieved performances similar to that reported by Chamrad et al. (24). Among the algorithms that use the MOWSE score, Mascot outperformed MS-Fit and Profound, with Mascot identifying 62% of proteins whereas Profound and MS-Fit identified 35 and 40% of the proteins, respectively. Aldente identified 48% of the proteins. The FMP algorithm described here outperformed all of the earlier methods by achieving the highest sensitivity (82%), precision (97%), and $F1$-measure (89%) values of protein identification.

Plots of scoring distributions are shown in Fig. 9. Both FMP and ProFound have greater ability to clearly separate correct identifications and random matches but this separation was less obvious in the case of MS-Fit.

The numbers of correctly identified peptide hits ranked by the five PMF search algorithms are shown in a five-way Venn
Feature-matching Pattern-based SVM for Robust PMF

Fig. 6. Response surface of the interaction of the parameters theoretical cleavage, random cleavage, fixed modification, and variable modification on the BAC of DB. The parameters random cleavage, fixed modification, and variable modification are set at 0.99, 0.80, and 0.1, respectively, to plot response surfaces for the interaction of theoretical cleavage on the BAC of DB. Each plot has convergent ranges and a highest point with a predictive value.

Diagnosis in Fig. 10. Out of 201 correct hits from the five algorithms (union), 53 proteins were identified by all five algorithms (intersection), 43 proteins were identified by a single algorithm and 17% of these protein hits were identified independently by FMP.

DISCUSSION

PMF, regardless of becoming complementary to PFF for protein identification, remains the commonly used method in proteomics research, accompanied by in-depth studies in every aspect. It is necessary to incorporate earlier conclusions and ideas into a uniform approach to improve the performance of PMF-based protein identification.

In this study, three types of early concept that were adopted as candidate parameters gave different contributions. First, although parameters derived from matched mass error have insufficient F-score to be selected for the final SVM parameters, they still gave us a new insight into the PMF procedure. Second, prediction of missed cleavage has been applied in MPIR, which helped us to learn that it is necessary to deal with theoretical and random missed cleavages separately. Last, consideration of matched peptides hit on a protein represented competitive information for protein identification. It will be worthwhile to further inspect parameter series of consecutive matching.

Peak intensity, which has been ignored by traditional methods because it is influenced by various ionization processes and the unequal trypsin specificity for different amino acids, is now becoming a hot issue. The combination of earlier findings with successful prediction about the relationship between peak intensity and protein sequence suggests that intense experimental peaks are more likely to be generated from true
Fig. 7. Contour diagrams of the predictive maximum BAC of D_E generated by each optimum parameter. The contour diagrams are drawn to show the optimum parameters and the predictive maximum BAC of D_E. The predictive optimum values for A (theoretical missed cleavage), B (random missed cleavage), C (fixed modification), and D (variable modification) are 0.75, 0.85, 1.09, and 0.33, respectively.
peptides rather than noisy peptides. Hence, we gave prominence to peak intensity in this approach by using it rather than predicting it. First, peak intensity proportion was used to make it more comparable between different PMF data. Second, peak intensity and peak intensity proportion participated in the generation of all series of parameters. Third, parameter series of top matching are newly proposed to highlight the effect of intense experimental peaks on performance. Finally, parameters derived from peak intensity and peak intensity proportion contributed five out of 11 to the selected SVM parameters. Peak intensity is an inherent attribute of the experimental spectrum and should receive more attention during protein identification.

The purpose of MPIR is to handle multiple shared peak intensities during spectrum alignment. Differentiated weights make it rational to allocate shared peak intensities to peptides of various origins. The inherent mechanism of MPIR allows both peptide missed cleavage and mass modification to be taken into consideration concurrently. The tuning of MPIR showed that missed cleavages should no longer be an obstacle to protein identification. When normalized by several types of harmonic factors, missed cleavages and mass modification can offer competitive information on performance.

SVM is sensitive to irrelevant or correlated inputs, and the algorithm itself uses stratified sampling to reduce the size of the training data as needed. There is no upper limit on the number of features; the only constraints are those imposed by hardware. In an attempt to describe the PMF procedure in as much detail as possible, we divided it into three objects and assembled four attributes of the experimental and theoretical spectrum to generate 81 FMPs. A total of 11 select parameters calculated for each FMP at the same time could offer ultrafine and sufficient information to distinguish positive sets from large numbers of negative data. According to our conclusion, FMPs with status 1 play the same roles in performance as those with status *. In other words, alignments between intense experimental peaks and theoretical peptides digested from proper cleavage or unique in the database or with the original mass are very likely to have occurred in positive PMFs. Meanwhile, parameters derived from implied attributes of the experimental spectrum and the theoretical spectrum contributed most to the 11 select parameters, among which five out of seven parameters were derived from peak intensity rank and peak intensity proportion of the experimental spectrum, and two parameters were derived from the peptide cleavage order of the experimental spectrum. The outstanding performance of the implied attributes shows their comparable or even greater robustness than other explicit attributes.

The good generalization performance, absence of local minima and sparse representation of solution make SVM a powerful tool in classification systems. In this study, SVM was
used to select sufficient and efficient parameters from 35,640 features generated from a series of earlier findings and new opinions. Using 491 select features of the PMF procedure, we obtained high performance with the validating data set by our feature-matching pattern-based SVM approach (see supplemental file SI14-PerformanceSummary). The FMP algorithm outperforms earlier methods and has a greater ability to discriminate correctly identified proteins from randomly matched proteins, achieving the highest sensitivity, precision and F1-measure values. The high level of performance is mainly the result of the more robust parameters and elaborate features in our procedure.

As demonstrated by the tuning of the number of experimental spectra, it is necessary to extract proper peaks for single PMF data before submission to this SVM approach. The feature-matching pattern-based SVM approach described here remains to be further evaluated by more PMF data. More precise data will surely contribute more to its performance by consistent re-training. A web service for pro-

**FIG. 9.** Distributions of the scores for the top candidates of the PMF algorithms (A) Mascot, (B) MS-Fit, (C) ProFound, (D) Aldente, and (E) FMP. Distributions of the scores for the top candidates of the PMF algorithms (A) Mascot, (B) MS-Fit, (C) ProFound, (D) Aldente and (E) FMP are shown as histograms. The continuous curve represents the group of correctly identified proteins and the broken curve represents the group of failed matches. FMP shows the clearest separation between correct and failed identifications.

**FIG. 10.** Five-way Venn diagram of the correct protein numbers identified by five PMF search algorithms. The number of correctly identified proteins by one or more algorithms is indicated; e.g. 53 protein hits are correctly identified by all five algorithms (intersection), and 43 protein hits are correctly identified by one or more algorithms (union).
tein identification based on the current SVM approach is available at http://bioinformatics.datawisdom.net/fmp.

Acknowledgments—We want to thank Prof. Yonghong Wang (State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai 200237, China) for providing the original mass spectrometry data of Avermectin Data set and Prof. Dianhui Wang (School of Engineering and Mathematical Science, La Trobe University, Melbourne, Victoria 3086, Australia) for helpful discussions.

This work was supported by the National Basic Research Program of China (2007CB714303) and the Open Project Program of the State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology. This article contains supplemental Parameter Tables.

To whom correspondence should be addressed: State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai 200237, P. R. China. Tel.: +86 21 64253658; E-mail: sliangz@ecust.edu.cn.

REFERENCES
1. Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punta, T., Peregrin-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canavil, V., Laele, A., Mena, F., Wong, P., Starostine, A., Canete, M. E., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanh, S., Lam, M. H. Y., Bullard, G., Athaf-Ul, A. M., Kanaya, S., Shilatifard, A., O’Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440, 637–643.
2. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dumpefield, B., Edelmann, A., Heurtier, M. A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A. M., Schneider, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440, 631–636.
3. Damodoran, S., Wood, T. D., Nagarajan, P., and Rabin, R. A. (2007) Evaluating peptide mass fingerprinting-based protein identification. Genomics Proteomics Bioinformatics 5, 152–157.
4. Henzel, W. J., Watanabe, C., and Stults, J. T. (2003) Protein identification: the origins of peptide mass fingerprinting. J. Am. Soc. Mass Spectrom. 14, 931–942.
5. Palagi, P. M., Hernandez, P., Walther, D., and Appel, R. D. (2006) Proteome informatics: bioinformatics tools for processing experimental data. Proteomics 6, 5435–5444.
6. Shadforth, I., Crowther, D., and Bessant, C. (2005) Evaluation of sequence databases using mass spectrometry data. Proteomics 5, 4082–4095.
7. Parker, K. C. (2002) Scoring methods in MALDI peptide mass fingerprinting: ChemScore, and the ChemApex program. J. Am. Soc. Mass Spectrom. 13, 22–39.
8. Craig, R., Cortens, J. P., and Beavis, R. C. (2005) Use of automated peptidic library for protein identification. Rapid Commun. Mass Spectrom. 19, 1844–1850.
9. Gras, R., Muller, M., Gasteiger, E., Gay, S., Binz, P. A., Bienvenut, W., Hoogland, C., Sanchez, J. C., Barocho, A., Hochstrasser, D. F., and Appel, R. D. (1999) Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. Electrophoresis 20, 3535–3550.
10. Monigatti, F., and Berndt, P. (2005) Algorithm for accurate similarity measurements of peptide mass fingerprints and its application. J. Am. Soc. Mass Spectrom. 16, 13–21.
11. Kaltenbach, H. M., and Wilke, A., Böcker, S. (2007) SAMPI: protein identification with mass spectra alignments. BMC Bioinformatics 8, 102–112.
12. He, Z., Yang, C., and Yu, W. (2008) Peak bagging for peptide mass fingerprinting. Bioinformatics 24, 1293–1299.
13. Sanaki, T., Suzuki, M., Lee, S. H., Goto, T., and Oe, T. (2010) A simple and efficient approach to improve protein identification by the peptide mass fingerprinting method: concomitant use of negative ionization. Anal. Methods 2, 1144–1151.
14. Song, Z., Chen, L., Ganapathy, A., Wan, X. F., Brechenmacher, L., Tao, N., Emerich, D., Stacey, G., and Xu, D. (2007) Development and assessment of scoring functions for protein identification using PMF data. Electrophoresis 28, 864–870.
15. Timm, W., Scherbert, A., Böcker, S., Kohlbacher, O., and Nattkemper, T. W. (2008) Peak intensity prediction in MALDI-TOF mass spectrometry: A machine learning study to support quantitative proteomics. BMC Bioinformatics 9, 443–460.
16. Yang, D., Ramkissoon, K., Hamlett, E., and Giddings, M. C. (2008) High-Accuracy Peptide Mass Fingerprinting Using Peak Intensity Data with Machine Learning. J. Proteome Res. 7, 62–69.
17. Haas, W., Faherty, B. K., Gerber, S. A., Elias, J. E., Beausoleil, S. A., Bakalakis, C. E., Li, X., Villen, J., and Gyi, S. P. (2006) Optimization and use of peptide mass measurement accuracy in shotgun proteomics. Mol. Cell. Proteomics 5, 1326–1337.
18. Dodds, D. E., Clowers, B. H., Hagerman, P. J., and Lebrilla, C. B. (2008) Systematic characterization of high mass accuracy influence on false discovery and probability scoring in peptide mass fingerprinting. Anal. Biochem. 372, 156–166.
19. Jain, R., and Wagner, M. (2010) Kolmogorov-Smirnov Scores and Intrinsic Mass Tolerances for Peptide Mass Fingerprinting. J. Proteome Res. 9, 737–742.
20. Song, Z., Chen, L., and Xu, D. (2009) Confidence assessment for protein identification using by peptide-mass fingerprinting data. Proteomics 9, 3090–3099.
21. Chamrad, D. C., Körtting, G., Stühler, K., Meyer, H. E., Klose, J., and Blüggen, M. (2004) Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. Proteomics 4, 619–628.
22. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20, 3551–3567.
23. Clauer, K. R., Baker, P., and Burlingame, A. L. (1999) Role of accurate mass measurement (±0.1 ppm) in protein identification strategies employing MS or MS/MS and database searching. Anal. Chem. 71, 2471–2478.
24. Zhang, W., and Chait, B. T. (2000) ProFound: An Expert System for Protein Identification Using Mass Spectrometric Peptide Mapping Information. Anal. Chem. 72, 2482–2489.
25. Sreevatsa, A. N., Badrunnisa, S., Shaukat, A. M., and Vinitha, R. P. (2010) Computational diagnostics based on proteomic data—review on approaches and algorithms. Int. J. Bioinformatics Res. 2, 56–66.
26. McHugh, L., and Arthur, J. W. (2008) Computational Methods for Protein Identification from Mass Spectrometry Data. PLoS Comput Biol 4, e12.
27. Tharakan, R., Edwards, N., and Graham, D. R. M. (2010) Data maximization by multipass analysis of protein mass spectra. Proteomics 10, 1160–1171.
28. Chamrad, D. C., Koerting, G., Gobom, J., Thiele, H., Klose, J., Meyer, H. E., and Blueggel, M. (2003) Interpretation of mass spectrometry data for high-throughput proteomics. Anal. Bio. Anal. Chem. 376, 1014–1022.
29. Falkner, A. J., Zachman, M., Veine, D. M., Walker, A., Strahler, J. R., and Andrews, P. C. (2007) Validated MALDI-TOF/TOF mass spectra for protein standards. J. Am. Soc. Mass Spectrom. 18, 850–855.
30. Yin, P. (2008) The regulation of avermectin fermentation process and its metabolic analysis by differential proteome. Doctoral Dissertation, East China University of Science and Technology.
31. Yin, P., Wang, Y. H., Zhang, S. L., Chu, J., Zhang, Y. P., Wang, M. L., and Zhou, J. (2008) Isolation of soluble proteins from an industrial strain-Streptomyces avermitilis in complex culture medium for two-dimensional gel electrophoresis. J. Microbiol. Methods 73, 105–110.
32. Touloup, M., Hernandez, C., Coro, I., Hoogland, C., Binz, P. A., and Appel, R. D. (2003) Aldente and BioGraph: An improved peptide mass fingerprinting protein identification environment. Proceedings of the Swiss
33. Siepen, J. A., Keevil, E. J., Knight, D., and Hubbard, S. J. (2007) Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics. J. Proteome Res. 6, 399–408
34. Fu, Y., Yang, Q., Sun, R., Li, D., Zeng, R., Ling, C. X., and Gao, W. (2004) Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. Bioinformatics 20, 1948–1954
35. Chang, C. C., and Lin, C. J. (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
36. Yu, H. F., and Ho, C. H. http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/eval/index.html
37. Chen, Y. W., and Lin, C. J. (2006) Combining SVMs with various feature selection strategies. Studies in Fuzziness and Soft Computing 207, 315–324. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/fselect/fselect.py
38. Li, Y. Y., Zhuang, Y. P., and Zhang, S. L. (2011) Automation Strategies for Bioinformatics Software and Their Implementation. Comput. Eng. 37, 81–83, 86