# AN SVM SCORER FOR MORE SENSITIVE AND RELIABLE PEPTIDE IDENTIFICATION VIA TANDEM MASS SPECTROMETRY*

HAIPENG WANG[1, 2 †], YAN FU[1, 2], RUIXIANG SUN[1], SIMIN HE[1], RONG ZENG[3], and WEN GAO[1, 2]

[1]*Digital Technology Lab, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China*

[2]*Graduate University of Chinese Academy of Sciences, Beijing 100039, China*

[3]*Research Center for Proteome Analysis, Key Lab of Proteomics, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*

Tandem mass spectrometry (MS/MS) has become increasingly important and indispensable in high-throughput proteomics for identifying complex protein mixtures. Database searching is the standard method to accomplish this purpose. A key sub-routine, peptide identification, is used to generate a list of candidate peptides from a protein database according to an experimental MS/MS spectrum, and then validate these candidate peptides for protein identification. Although currently there are many algorithms for peptide identification, most of them either lack an effective validation module or only validate the first-ranked peptide, thus leading to a low identification reliability or sensitivity. This paper proposes a new algorithm, named pepReap, to overcome the above drawbacks. It consists of a two-layered scoring scheme based on machine learning. The first layer is a rough scoring function which uses some simple and heuristic factors to measure the degree of the matches between an experimental MS/MS spectrum and the candidate peptides; thus a ranked list of candidate peptides is generated at a relatively low computational cost. The second layer is a fine scoring function which re-ranks the candidate peptides generated in the first layer and determines which one among them is the true positive. The fine scoring function was designed based on support vector machines (SVMs) using more comprehensive factors, such as the correlations between ions, the mass matching errors of fragment and peptide ions, *etc*. Consequently, the SVM classifier serves as not only a scorer but also a validation module. Experimental comparison with the popular SEQUEST algorithm coupled with threshold validation criteria on a reported dataset demonstrates that the pepReap algorithm achieves higher performance in terms of identification sensitivity with comparable precision.

## 1. Introduction

The essential mission of proteomics is to identify and quantify all the levels of proteins found in a cell or tissue under various physiological conditions [1]. Tandem mass spectrometry (MS/MS), which can measure the mass-to-charge ratios ($m/z$) of ionized molecules, has become increasingly important and indis-

* This work is supported by the National Key Basic R&D Program (Grant No. 2002CB713807) and the National Key Technologies R&D Program (Grant No. 2004BA711A21) of China.

† To whom correspondence should be addressed. E-mail: hpwang@jdl.ac.cn.

pensable for identifying complex protein mixtures in high-throughput proteomics experiments [2–4].

In a typical "bottom-up" experiment, protein mixtures are directly digested with a site-specific protease, usually the trypsin, into complex peptide mixtures which are subsequently separated by liquid chromatography (LC). The separated peptides eluted from LC are then ionized with one or more units of charges (to form precursor ions), selected according to their $m/z$ values, and analyzed through fragmentation by MS/MS. In this process, hundreds of thousands of MS/MS spectra are produced which are then computationally interpreted to generate their candidate peptide sequences. Finally the candidate peptides are validated and the correct ones are grouped to identify the proteins from which the peptides derived. The peptide identification including peptide scoring and subsequent validation is a critical step in the process of protein identification [5,6].

Aiming at the drawbacks of existing algorithms for peptide identification, we developed a more robust algorithm, pepReap, which consists of a two-layered scoring scheme based on support vector machines (SVMs) using some elaborated features characterizing the matches between peptides and MS/MS spectra, to obtain a positive or negative score for each candidate peptide, explicitly distinguishing correct matches from incorrect matches dispensing with setting significant thresholds.

## 2. Background and Related Work

### 2.1. *Peptide MS/MS Spectrum*

The basic molecular building blocks of proteins are amino acids which are differentiated from each other by the side chain $R$ (shown in Figure 1(a)). A protein or peptide is a chain that consists of amino acid residues linked together by peptide bonds formed by condensation reactions (shown in Figure 1(b)). For MS/MS, peptides are the products of enzymatic digestion of proteins.

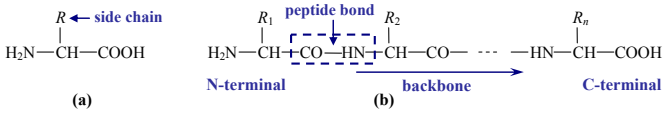


Figure 1. The general structures for (a) amino acids and (b) peptides.

In a tandem mass spectrometer, precursor ions within a given range around a specific $m/z$ value are selected and subjected to fragmentation by collision-induced dissociation (CID) resulting in various types of fragment ions. Fragment ions are measured to obtain a number of spectral peaks each comprising an $m/z$

and an intensity value. Peaks plus the *m*/*z* value and charge state of the precursor ion constitutes a peptide MS/MS spectrum which normally corresponds to a unique peptide. Besides the peaks from the peptide to be identified, there are also many noisy peaks brought by chemical contaminants or electronic fluctuations.

Common fragmentation patterns of parent ions and nomenclature for fragment ions are shown in Figure 2. According to the fragmentation position relative to the peptide bond along the backbone and the terminal (N or C-terminal) where the charge(s) is (are) retained, fragment ions are classified as *a*-, *b*-, *c*-, *x*-, *y*-, or *z*-ions. The pairs of (*a*, *x*), (*b*, *y*) and (*c*, *z*) are complementary ion types. Fragment ions can be singly or multiply charged and possibly lose a neutral water ($H_2O$) or ammonia ($NH_3$).

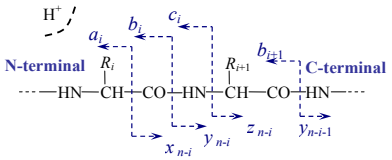


Figure 2. Fragmentation patterns of parent ions and nomenclature for various types of fragment ions.

The notations used for describing fragments are listed in Table 1. Based on the notations, $T_{\text{cnsc}} = \{t_i, t_{i+1}, \cdots, t_{i+k}\}$ denotes a set of consecutive ions, $T_{\text{cmpl}} = \{t_i, \overline{t}_{n-i}\}$ a pair of complementary ions, and $T_{\text{homo}} = \{t_i, t_i^{++}, t_i^0, t_i^*\}$ a set of homologous ions, with *t* denoting any ion type and $0 < k < (n\text{-}i)$. Ions of *a*, *b*, *y* and their homologues are dominant fragments observed in MS/MS spectra.

Table 1. Notations for fragments.

| Notation | Meaning |
|---|---|
| $i$ ($0 < i < n$) | Subscript denoting the cleavage site, i.e., the number of residues contained in the fragment; $n$ is the number of residues in the precursor, i.e., the peptide length |
| ‾ | Overline denoting the complement of an ion type, e.g., $\overline{b}$ is *y* |
| 0 | Superscript indicating a neutral loss of water |
| * | Superscript indicating a neutral loss of ammonia |
| ++ | Superscript denoting a double-charge state (single-charge state by default) |

### 2.2. *Approaches to Peptide Identification from MS/MS Spectra*

There are mainly three approaches to peptide identification from MS/MS spectra: *de novo* sequencing [7–9], sequence tagging [10,11] and database searching [13–23].

*De novo* sequencing tries to infer the complete peptide sequence directly from the *m*/*z* differences between peaks in an MS/MS spectrum without any help of databases. It is capable of identifying the peptides which are not present in

databases or suffer from unanticipated post-translational modifications (PTMs) or mutations; however, it is generally less tolerant of low-quality mass spectra than the database searching approach.

Sequence tagging yields one or more partial sequences (called sequence tag) by manual interpretation or *de novo* sequencing. Candidate peptides containing this sequence tag can be found by homologous sequence searching. The sequence tags can serve as an effective filter for candidate peptide generation in database searching especially when PTM identification is involved [12].

Database searching is the most widely used method in high-throughput proteomics experiments due to its sensitivity, rapidness, and tolerance for low-quality spectra. It compares an experimental MS/MS spectrum with theoretical ones predicted from the peptide sequences resulting from *in silico* digestion of proteins in databases, whereby the experimental spectrum is correlated by a scoring function with a ranked list of candidate peptides. The match of the highest score is normally regarded as the peptide corresponding to the spectrum. However, random matches often occur between theoretical fragments and noisy peaks, or between false isobaric theoretical fragments and signal peaks. In addition, unanticipated fragment ions or modifications can lead to missing matches. Therefore the best matching peptide may not be correct and inversely the correct peptide may not be the top-ranked one. Consequently the peptide identifications need to be carefully validated.

Scoring functions, validation method, and fragmentation model [24,25] are central and fundamental to all the above three approaches.

### 2.3. *Peptide Scoring and Validation Methods in Database Searching*

To measure the similarity between experimental and theoretical MS/MS spectra, two strategies are adopted based on the descriptive framework or probabilistic framework [13].

In the probabilistic strategy, a probability is calculated for the event that the match between a peptide and an experimental spectrum is completely random [18], or that a peptide actually generated the experimental spectrum [19–21]. The combination of the above two means leads to the likelihood-ratio score [22,23].

In the descriptive strategy, an experimental and a theoretical MS/MS spectrum are represented as vectors $\mathbf{S} = (s_1, s_2, \cdots, s_n)$ and $\mathbf{T} = (t_1, t_2, \cdots, t_n)$, respectively, where $n$ denotes the number of predicted fragments, $s_i$ and $t_i$ are binary values or the observed and predicted intensity values of the $i^{\text{th}}$ fragment, respectively. The spectral dot product (SDP) between **S** and **T** serves as their similarity measure [15]. In SDP, correlative information among fragments is totally ig-

nored. Many scoring algorithms are based on the SDP [14–16]. A representative of the descriptive strategy is the popular SEQUEST algorithm [14], in which consecutive ion pairs are considered in a preliminary scoring function and then the cross-correlation analysis is performed between the experimental and theoretical spectra. The KSDP algorithm adopted in the pFind software extends the SDP by using the kernel technique to comprehensively incorporate correlative information among ions [16,17].

In routine experiments, the threshold method is widely used to validate peptide identifications of SEQUEST. However, there are no uniform rules to set the cutoff values [3,4,26,27]. Some sophisticated validation algorithms based on probability (or pseudo-probability) [28–31] and machine learning [30,32,33] have been developed to improve the reliability of peptide identifications. Most algorithms take advantage of some outputs of SEQUEST, such as *XCorr*, $\Delta Cn$, *Sp*, *RSp*, and *Ions* [14]. Keller *et al.* discriminated between positive and negtive identifications according to a Gaussian and a gamma distribution [28]. MacCoss *et al.* proposed a scoring scheme which normalizes XCorr values to be independent of peptide length and then derived a confidence of an identification [29]. RScore combines the XCorr and matched intensity value to get a measurement of randomicity [31]. Anderson *et al.* and Baczek *et al.* performed classification tasks via machine learning approaches using outputs of SEQUEST and some additional factors (for example, the peak count, the ratios of matched peaks and matched intensities; isoelectric value, hydrophobicity, molecular weight, and charge state of peptides) [32,33]. These algorithms are remedies for the validation of SEQUEST results to some extent. However, they all focus on deciding whether the first-ranked peptide is correct while ignoring all other lower-ranked peptides that often include the correct one. Moreover, they do not fully exploit the features characterizing the quality of matches between MS/MS spectra and peptides.

In this paper, the scoring and validating processes are combined together by directly using an SVM classifier for scoring the peptide-spectrum matches based on a variety of matching features.

## 3. Methods

### 3.1. *The Framework of pepReap*

The pepReap algorithm comprises a rough scorer and a fine SVM scorer shown in Figure 3.

In the first step, a ranked list of candidate peptides is generated by the rough scoring function:

$$RS = (\sum_{i=1}^{N_{match}} I_i) \times N_{match} \Big/ L_{pep}\ , \tag{1}$$

where $N_{match}$ is the number of the predicted fragment ions matching the peaks in the experimental spectrum, $I_i$ is the intensity value of the $i^{th}$ matched peak and $L_{pep}$ is the peptide length. Similar formulas have been adopted in SEQUEST [14] and pFind [16].

In the second step, an SVM-based scoring function gives a signed decision value according to the features constructed from the matching matrix (see section 3.2). The parameters of the SVM scorer are tuned on a training dataset by cross validation.

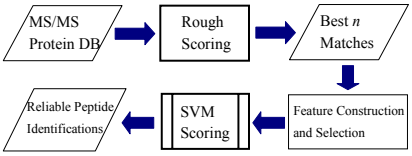


Figure 3. The framework of the pepReap algorithm.

### 3.2. *Matching Matrix*

To make data processing convenient, a matching matrix between a peptide and a spectrum is constructed, as shown in Figure 4. The matching matrix is an $m \times n$ array, where $m$ denotes the number of different ion types under consideration, $n$ is the length of a peptide, the column indexes $(1, 2, \cdots, n)$ represent the cleavage sites of a peptide, the row indexes $(a, b, \cdots, y)$ denote various ion types and the element $p_{ti}$ ( $t \in \{a, b, \cdots, y\}$ ) holds the information of the corresponding matched peak, or keeps null.

| | 1 | 2 | ⋯ | $n$ |
|---|---|---|---|---|
| $a$ | $p_{a1}$ | $p_{a2}$ | ⋯ | $p_{an}$ |
| $b$ | $p_{b1}$ | $p_{b2}$ | ⋯ | $p_{bn}$ |
| ⋮ | ⋮ | ⋮ | ⋱ | ⋮ |
| $y$ | $p_{y1}$ | $p_{y2}$ | ⋯ | $p_{yn}$ |

Figure 4. Matching matrix between a peptide and a spectrum.

### 3.3. *Support Vector Machines*

Support vector machines are developed by Vapnik and his coworkers based on the statistical learning theory [34]. The principle of structural risk minimization establishes the basis of the good generalization performance of SVMs. For a binary classification problem, the input to the SVM training algorithm is a set of $n$ samples denoted as

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)\}, \tag{2}$$

where $\mathbf{x}_i \in \mathbb{R}^d$ is the $i^{\text{th}}$ sample and $y_i \in \{-1, 1\}$ is its class label. The objective of SVMs is to find an optimal separating hyperplane that maximizes the margin between two classes in a high dimensional feature space into which the input vectors are mapped by a kernel function, as shown in Figure 5. The kernel function implicitly calculates a dot product in the feature space with all necessary computations performed in the input space. One advantage of it is that it can get linearly non-separable samples in the input space to be linearly separable in the feature space.

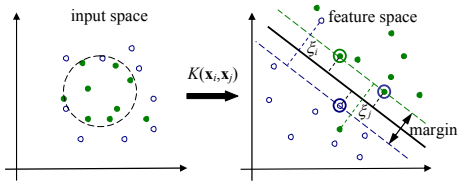


Figure 5. A linear separating hyperplane (the solid line in the right coordinates) in the feature space corresponding to a non-linear boundary (the dashed line in the left coordinates) in the input space. The data points in circles are support vectors (SVs).

The decision function of the SVM classifier is

$$f(\mathbf{x}) = \sum_{i \in \text{SVs}} y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b, \tag{3}$$

where the coefficients $\alpha_i$ are solved in the interval [0, $C$] by a convex quadratic programming. $C$ is a tradeoff between maximizing the margin and minimizing the empirical risks and can be specified for positive and negative samples respectively in the case of unbalanced datasets. The radial basis function (RBF) kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2)$ is popular for practical use due to its approximate behaviors to other kernels under certain conditions and the less number of parameters to be tuned (only $C$ and $\gamma$) [35].

### 3.4. *Performance Measurement*

For a binary classification problem, let *tp*, *fp*, *tn* and *fn* denote the number of true positives, false positives, true negatives and false negatives respectively. The Matthews correlation coefficient (MCC) [36] incorporates all four prediction indexes into a single statistic to measure the performance of classifiers:

$$MCC = \frac{(tp \times tn) - (fp \times fn)}{\sqrt{(tp + fp) \times (fp + tn) \times (tn + fn) \times (fn + tp)}}. \tag{4}$$

The MCC is a number in the interval [-1, 1], with 1 indicating completely correct classification, -1 indicating completely incorrect classification and 0 indicat-

ing no correlations between predictions and the true class labels. The MCC is superior to the accuracy which is defined as the proportion of correctly classified samples, especially when datasets are unbalanced, because the accuracy is dominated by the majority class and thus can be misleading. Therefore the MCC is employed in the cross-validation training process of SVMs.

Sensitivity ($SEN = tp / (tp + fn)$) and precision ($PRE = tp / (tp + fp)$) are used as the performance measures for comparing pepReap with SEQUEST.

### 3.5. *Features Characterizing the Quality of Matches*

In the step of SVM scoring, we constructed some features from the matching matrix to characterize the quality of matches between peptides and spectra. The features fall into seven categories: outputs of rough scoring, total matched intensities for various ion types, correlations bwtween matched ions, residue composition and properties of candidate peptides, statistics of cleavage sites, ratios and mass errors of matched peaks, and some other descriptive features such as missed proteolytic cleavage site and charge state.

The measures for consecutive, complementary, and homologous ions are

$$f_{\text{cnsc}} = \sum_{i \in \mathbf{T}_{\text{cnsc}}} \left( \prod_{t \in T^i_{\text{cnsc}}} I_t \right), \tag{5}$$

$$f_{\text{cmpl}} = \sum_{i \in \mathbf{T}_{\text{cmpl}}} \left( \sum_{t \in T^i_{\text{cmpl}}} I_t \times I_{\bar{t}} \right), \tag{6}$$

and

$$f_{\text{homo}} = \sum_{i \in \mathbf{T}_{\text{homo}}} \left( \prod_{t \in T^i_{\text{homo}}} I_t \right), \tag{7}$$

respectively, for all $I_t > 1$, where $I_t$ is the intensity of a matched peak in the matching matrix, and $\mathbf{T}_j$ is a set comprising all $T_j^i$ with $j \in \{\text{cnsc,cmpl,homo}\}$.

The average matching error is calculated by

$$f_{\text{ame}} = \sqrt{\frac{1}{M} \sum_{i=1}^{M} (emz_i - tmz_i)^2}, \tag{8}$$

where $M$ is the number of matching fragments and $emz_i$ and $tmz_i$ are the observed and the theoretical $m/z$ values of a predicted fragment respectively.

Fragmentation patterns of peptides are influenced not only by collision energy of the tandem mass spectrometer but by physical and chemical properties (gas-phase basicity, hydrophobicity, *etc.*) of amino acid composition, as is explained by the mobile proton model [37]. Therefore we believe that such features as the number of certain residues, the hydrophobicity of peptides, and the

statistics of cleavage sites jointly provide better clues to describing the quality of matches. These statistics are normalized by the peptide length.

## 4. Experiments

### 4.1. *Datasets*

The ion trap MS/MS spectra reported in Ref. 38 were used for our experiments. These spectra were divided into two datasets, A and B, according to the different concentration of two mixtures of 18 purified proteins with known sequences which were digested by trypsin. All the spectra were searched using SEQUEST against a database combining the human proteins with the 18 proteins; then the peptide identification results were validated manually. Consequently, there are totally 2054 spectra identified correctly with their peptide terminus consistent with the substrate specificity of trypsin. In our experiments, 731 validated identifications in dataset B were used for training the SVM scorer in pepReap and 1323 validated identifications in dataset A were used for comparing pepReap with SEQUEST.

### 4.2. *Noise Reduction and Intensity Normalization*

To weaken the influence of noises on peptide identification and eliminate the diversities of total ion currents of different spectra, noise reduction and intensity normalization are performed. All the peaks lower than 2% intensity of the highest peak are removed and the intensities of the remaining peaks are normalized:

$$I_{Ni} = 100 \times \sqrt{I_{Oi} \Big/ \sum_{j} I_{Oj}} \tag{9}$$

where $I_{Ni}$ is the normalized intensity and $I_{Oi}$ is the original intensity.

### 4.3. *Protein Database and Search Parameters*

The protein database searched is a union of the SWISS-PROT protein database and the 18 known proteins. The search parameters used in pepReap and SEQUEST is: maximum number of missed cleavage sites: 2; tolerance of fragment ions: 1.0 Da; tolerance of precursor: 3.0 Da; ion types: $b$, $b^{++}$, $b^0$, $y$, $y^{++}$, $y^0$; and enzyme: trypsin.

### 4.4. *Feature Selection*

A total of 56 features are extracted from each match between a spectrum and its corresponding candidate peptides. These features are scaled into the interval [0,

1]. The best 20 features are selected by information gain ratio [39] on dataset B through cross validation, as listed in Table 2.

Table 2. The top 20 features selected by information gain ratio.

| FEATURES | GAIN RATIO | FEATURES | GAIN RATIO |
|---|---|---|---|
| *delta_rough_score* | 0.409 ± 0.096 | *missed_cleavage_site* | 0.045 ± 0.001 |
| *rough_score* | 0.409 ± 0.096 | *total_intensity_b* | 0.044 ± 0.003 |
| *match_ratio_intensity* | 0.274 ± 0.035 | *intensity_b* | 0.037 ± 0.001 |
| *rank_rough_score* | 0.275 ± 0.001 | *average_match_error* | 0.032 ± 0.001 |
| *intensity_y* | 0.156 ± 0.004 | *homologous_y* | 0.031 ± 0.004 |
| *match_ratio_peak* | 0.160 ± 0.026 | *peptide_match_error* | 0.027 ± 0.001 |
| *complementary_by* | 0.146 ± 0.013 | *peptide_mass* | 0.027 ± 0.001 |
| *total_intensity_y* | 0.092 ± 0.003 | *number_HKR_pep* | 0.016 ± 0.000 |
| *total_consecutive_y* | 0.088 ± 0.012 | *hydrophobicity_pep* | 0.012 ± 0.001 |
| *total_consecutive_b* | 0.056 ± 0.001 | *clvgsite_median_b* | 0.012 ± 0.001 |

### 4.5. *Results*

A ranked list of 500 candidate peptides was first generated by rough scoring, whereby we observed that all correct peptide identifications ranked in the top ten except for four spectra in dataset A and two spectra in dataset B. The SVM scorer was trained and tested on the top ten rough-scoring results of dataset B and dataset A, respectively. LIBSVM, an implementation of SVMs, with the RBF kernel, was employed [40]. A peptide is regarded as the correct answer if its SVM prediction value is the highest and above a given threshold; otherwise, peptides are considered incorrect answers. Performance comparison between the pepReap algorithm using the SVM scorer and the SEQUEST using threshold validation criteria are shown in Table 3.

Table 3. Performance comparison of pepReap and SEQUEST.

| pepReap | | | | | | SEQUEST(threshold[2]) | |
|---|---|---|---|---|---|---|---|
| Training Set (Dataset B) | | | | Test Set (Dataset A) | | Test Set (Dataset A) | |
| *weight*[1] | best *C* | best $\gamma$ | *MCC* | *SEN* | *PRE* | *SEN* | *PRE* |
| 1:1 | 4.0000 | 0.03125 | 0.9212±0.0158 | 0.9106 | 0.9128 | 0.8715 [a] | 0.9107 [a] |
| 10:1 | 0.0625 | 0.12500 | 0.9269±0.0177 | 0.9116 | 0.9204 | 0.5397 [b] | 0.9420 [b] |
| 50:1 | 1.0000 | 0.12500 | 0.9300±0.0105 | 0.9175 | 0.9257 | 0.5548 [c] | 0.9410 [c] |
| | | | | | | 0.6757 [d] | 0.9391 [d] |

Note: [1]The *weight* is used to set the parameter *C* of class 1 and -1 to *weight* × *C*. [2]The commonly used threshold criteria for evaluating SEQUEST identification results are, (a) *XCorr* ≥ 1.5, 2.0, 2.0 [3], (b) Δ*Cn* ≥ 0.1 and *XCorr* ≥ 1.9, 2.2, 3.75 [26], (c) Δ*Cn* ≥ 0.1 and *XCorr* ≥ 1.8, 2.2, 3.7 [27], and (d) Δ*Cn* ≥ 0.08 and *XCorr* ≥ 2.0, 1.5, 3.3 [4], for +1, +2, +3 charged fully tryptic peptides, respectively.

From Table 3, it can be seen that the high precision of SEQUEST is obtained at the cost of a very low sensitivity. In contrast, pepReap achieves much higher sensitivity than SEQUEST with some insignificant loss of precision. Both measures tend to increase when higher weight ratios for positives and negatives are applied to the SVM scorer.

## 5. Conclusions and Future Work

We have presented a novel and promising peptide identification algorithm, named pepReap, based on support vector machines. The characteristics distinguishing the pepReap from other algorithms lie in the flexible use of an SVM classifier both as the scoring function and the validation module and comprehensive features we used for measuring the match between a spectrum and a peptide. Preliminary experimental results on a dataset demonstrate that the pepReap algorithm can achieve much higher identification sensitivity without significant loss in identification precision compared with the popular SEQUEST algorithm that uses simple threshold validation criteria. A prerequisite of the pepReap algorithm is a set of mass spectra with known peptide sequences. Such training dataset for a given instrument can be obtained by first applying an independent identification algorithm to the spectra to be identified and then picking out high-confidence identifications. Our future work includes exploiting more informative features based on improved fragmentation models, testing the pepReap algorithm on more datasets and comparing it with sophisticated validation algorithms coupled to the SEQUEST (e.g. algorithms in Ref. [28], [32]).


## Acknowledgments

We would like to thank Dr. Andrew Keller from the Institute for Systems Biology for providing the dataset of MS/MS spectra. We would also like to thank Dr. Chih-Jen Lin from the National Taiwan University for helpful discussions on support vector machines.


## References


1. A. Pandey and M. Mann, *Nature* **405**, 837 (2000).
2. R. Aebersold and M. Mann, *Nature* **422**, 198 (2003).
3. A. J. Link *et al.*, *Nat. Biotechnol.* **17**, 676 (1999).
4. J. Peng *et al.*, *J. Proteome Res.* **2**, 43 (2003).
5. A. I. Nesvizhskii and R. Aebersold, *Drug Discov. Today* **9**, 173 (2004).
6. H. Steen and M. Mann, *Nat. Rev. Mol. Cell Biol.* **5**, 699 (2004).
7. B. Lu and T. Chen, *Drug Discov. Today*: *Biosilico* **2**, 85 (2004).
8. Z. Zhang, *Anal. Chem.* **76**, 6374 (2004).
9. A. Frank and P. Pevzner, *Anal. Chem.* **77**, 964 (2005).
10. M. Mann and M. Wilm, *Anal. Chem.* **66**, 4390 (1994).
11. D. L. Tabb, A. Saraf, and J. R. Yates III, *Anal. Chem.* **75**, 6415 (2003).
12. A. Frank, *et al*., *J. Proteome Res.* **4**, 1287 (2005).
13. R. G. Sadygov, D. Cociorva and J. R. Yates III, *Nat. Methods* **1**, 195 (2004).

14. J. K. Eng, A. L. McCormack and J. R. Yate III, *J. Am. Soc. Mass Spectrom.* **5**, 976 (1994).
15. H. I. Field, D. Fenyö and R. C. Beavis, *Proteomics* **2**, 36 (2002).
16. Y. Fu *et al.*, *Bioinformatics* **20**, 1948 (2004).
17. D. Li *et al.*, *Bioinformatics* **21**, 3049 (2005).
18. D. N. Perkins *et al.*, *Electrophoresis* **20**, 3551 (1999).
19. V. Bafna and N. Edwards, *Bioinformatics* **17**, S13 (2001).
20. N. Zhang, R. Aebersold and B. Schwikowski, *Proteomics* **2**, 1406 (2002).
21. M. Havilio, Y. Haddad and Z. Smilansky, *Anal. Chem.* **75**, 435 (2003).
22. J. Colinge *et al.*, *Proteomics* **3**, 1454 (2003).
23. J. E. Elias *et al.*, *Nat. Biotechnol.* **22**, 214 (2004).
24. F. Schütz *et al.*, *Biochem. Soc. Trans.* **31**, 1479 (2003).
25. Z. Zhang, *Anal. Chem.* **76**, 3908 (2004).
26. M. P. Washburn, D. Wolters and J. R. Yates III, *Nat. Biotechnol.* **19**, 242 (2001).
27. X.-S. Jiang *et al.*, *Mol. Cell. Proteomics* **3**, 441 (2004).
28. A. Keller *et al.*, *Anal. Chem.* **74**, 5383 (2002).
29. M. J. MacCoss, C. C. Wu and J. R. Yates III, *Anal. Chem.* **74**, 5593 (2002).
30. J. Razumovskaya *et al.*, *Proteomics* **4**, 961 (2004).
31. F. Li *et al.*, *Rapid Commun. Mass Spectrom.* **18**, 1655 (2004).
32. D. C. Anderson *et al.*, *J. Proteome Res.* **2**, 137 (2003).
33. T. Baczek *et al.*, *Anal. Chem.* **76**, 1726 (2004).
34. V. N. Vapnik, *Statistical Learning Theory*, New York: John Wiley and Sons (1998).
35. S. S. Keerthi and C.-J. Lin, *Neural Comput.* **15**, 1667 (2003).
36. B. W. Matthews *et al.*, *Biochim. Biophys. Acta* **405**, 442 (1975).
37. A. R. Dongré *et al.*, *J. Am. Chem. Soc.* **118**, 8365 (1996).
38. A. Keller *et al.*, *OMICS* **6**, 207 (2002).
39. R. Quinlan, *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufmann Publishers Inc. (1993).
40. C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (2001).