BIOINFORMATICS

Vol. 20 no. 12 2004, pages 1948–1954
doi:10.1093/bioinformatics/bth186

Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry

Yan Fu[1,2,*], Qiang Yang[3], Ruixiang Sun[2], Dequan Li[1], Rong Zeng[4], Charles X. Ling[5] and Wen Gao[1,2]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China, [2]Graduate School of Chinese Academy of Sciences, Beijing 100039, China, [3]Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, [4]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 and [5]Department of Computer Science, The University of Western Ontario, London, Ontario, Canada N6A 5B7

Received on September 4, 2003; revised on January 6, 2004; accepted on February 10, 2004
Advance Access publication March 25 2004

ABSTRACT

Motivation: The correlation among fragment ions in a tandem mass spectrum is crucial in reducing stochastic mismatches for peptide identification by database searching. Until now, an efficient scoring algorithm that considers the correlative information in a tunable and comprehensive manner has been lacking.

Results: This paper provides a promising approach to utilizing the correlative information for improving the peptide identification accuracy. The kernel trick, rooted in the statistical learning theory, is exploited to address this issue with low computational effort. The common scoring method, the tandem mass spectral dot product (SDP), is extended to the kernel SDP (KSDP). Experiments on a dataset reported previously demonstrate the effectiveness of the KSDP. The implementation on consecutive fragments shows a decrease of 10% in the error rate compared with the SDP. Our software tool, pFind, using a simple scoring function based on the KSDP, outperforms two SDP-based software tools, SEQUEST and Sonar MS/MS, in terms of identification accuracy.

Contact: yfu@ict.ac.cn

Supplementary Information: http://www.jdl.ac.cn/user/yfu/pfind/index.html

INTRODUCTION

In recent years, mass spectrometry (MS) has been recognized as one of the most successful techniques in proteomics research (Aebersold and Mann, 2003). The method of protein identification via tandem mass spectrometry (MS/MS) plays a fundamental and indispensable role in current proteomic laboratories. The relevant research has drawn much attention from both biological and computational fields.

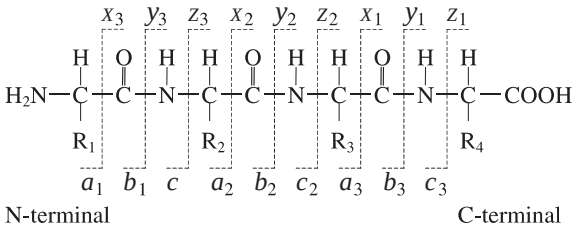

Fig. 1. Fragment ions resulting from peptide bonds cleavage by collision-induced dissociation.

A peptide is a string of amino acid residues joined together by peptide bonds. In the mass spectrometer, peptides derived from digested proteins are ionized. Peptide precursors of a specific mass–charge ratio (m/z) are selected and further fragmented by collision-induced dissociation (CID). Product ions are detected. The measured m/z and intensity of the product ions form finally the peaks in the tandem mass spectrum (MS/MS spectrum).

By CID, three kinds of backbone cleavages on peptide bonds can produce six series of fragment ions, denoted by N-terminal a, b and c type fragments and C-terminal x, y and z type fragments, as illustrated in Figure 1. For low-energy CID, usually just one backbone cleavage occurs in a peptide and a, b and y are dominant fragment types. Fragments can be singly charged or multiply charged and may possibly lose a neutral water or ammonia molecule. Besides these fragments, other types of product ions also present themselves as peaks

*To whom correspondence should be addressed.

Table 1. Notations for fragment description

Notation	Meaning
i	Subscript denoting the number of residues in the fragment (examples in Fig. 1)
0	Superscript indicating a loss of a water molecule
*	Superscript indicating a loss of an ammonia molecule
++	Superscript indicating a double charge state (single charge as default)

in the MS/MS spectrum as well as noise and contaminants. Table 1 gives the notations used in this paper for describing fragments.

Three computational approaches have been employed to identify peptide sequences via MS/MS: database searching, *de novo* and sequence tag query. The database searching approach (Eng *et al*., 1994; Fenyö *et al*., 1998; Clauser *et al*., 1999; Perkins *et al*., 1999; Pevzner *et al*., 2000; Bafna and Edwards, 2001; Field *et al*., 2002; Zhang *et al*., 2002) searches the protein database for the target peptide sequence by simulating protein digestion and peptide fragmentation and comparing the theoretical MS/MS spectra with the experimental spectrum. The *de novo* approach (Taylor and Johnson, 1997; Danick *et al*., 1999; Chen *et al*., 2001; Taylor and Johnson, 2001; Ma *et al*., 2003) tries to recover the full peptide sequence directly from the observed MS/MS spectrum without any dependence on databases. The sequence tag query approach (e.g. Mann and Wilm, 1994; Perkins *et al*., 1999) searches the database for the full peptide sequence with partial sequence information, which is recovered from the observed MS/MS spectrum either automatically or manually. The database searching approach has been applied widely to high-throughput protein identification. This paper focuses on peptide-scoring algorithms in the database searching approach.

Various strategies have been proposed for scoring peptides in the existing database search tools. The shared peak count in early work counts the number of matched peaks between the theoretical and the experimental spectra. SEQUEST (Eng *et al*., 1994) uses the notion of cross-correlation between spectra. Sonar MS/MS (Fenyö and Beavis, 2003) adopts the vector representation (Wan *et al*., 2002) of spectra and calculates the dot product of two spectral vectors as the score. Mascot (Perkins *et al*., 1999), SCOPE (Bafna and Edwards, 2001) and ProbID (Zhang *et al*., 2002) deal with the peptide identification problem in the probabilistic framework with different assumptions. As a rule, predicted fragments are compared with the observed peaks, and all the matches contribute equally to the final score. However, there is no guarantee for the correctness of these matches. Stochastic mismatches are accepted falsely and may lead finally to false assignment of peptides to spectra. In fact, the fragmentation pattern of a peptide and the relative intensity of the peaks in the MS/MS spectrum are not fully predictable. Noise and unexpected product ion types may result in many uninterpretable peaks. The databases of proteins sequenced or predicted from genomes expand rapidly. All these facts increase the possibility of stochastic mismatches and make the peptide-scoring algorithm less satisfactory.

In this paper, we make use of the correlative information among fragments to reduce stochastic mismatches. One of our most important observations about the MS/MS spectrum is that the fragments resulting from peptide fragmentation by CID rarely occur independently; most often they tend to occur correlatively with each other. However, few scoring algorithms make good use of this kind of correlative information to reduce stochastic mismatches. SCOPE assumes independence between fragments in order to make its complex probability computable. SEQUEST increases its S_p score for each matched consecutive fragment. ProbID calculates the ratio of matched consecutive and complementary fragment pairs as a component of its probability. However, an efficient algorithm that considers correlations among more fragments in a tunable manner is still lacking. There are two computational difficulties in such an algorithm. One is how to count exhaustly all possible combinations of correlated ions while excluding others. The other is how to deal with the exponential combinations in a computationally efficient way. With the observation that current scoring algorithms are based mostly on the spectral dot product (SDP), we propose to extend it with kernels, whose concept is deeply rooted in the machine learning discipline, to introduce the correlative information. By applying the locally improved kernels, the two difficulties are overcome gracefully. Experiments demonstrate the effectiveness of our new approach.

ALGORITHM

Spectral vector

Thereotical and experimental spectra can be expressed as N-dimensional vectors, where N is the number of different m/z values used. Let $\mathbf{c} = [c_1, c_2, \ldots, c_N]$ denote the experimental spectrum and vector $\mathbf{t} = [t_1, t_2, \ldots, t_N]$ the theoretical one. c_i and t_i take binary values $\{0, 1\}$ or the intensity of the fragment with the i-th m/z value in the spectra (assumed intensities for predicted fragments). Thus, both spectra are elements in the N-dimensional vector space, which we call the initial spectral space.

Spectral dot product

The tandem mass SDP between the experimental and theoretical spectra is defined as

$$\mathrm{SDP} = \langle \mathbf{c}, \mathbf{t} \rangle = \sum_{i=1}^{N} c_i t_i. \quad (1)$$

The SDP-based cosine value of the angle between spectral vectors was used as a similarity measure for MS/MS spectra (Wan *et al*., 2002; Tabb *et al*., 2003). In current peptide-scoring algorithms, the SDP is often involved directly or indirectly and plays an important role. The vector representation and the dot product were adopted explicitly in the Sonar MS/MS for scoring. In SEQUEST, the cross-correlation of two spectra is actually the SDP, and the score Xcorr is the SDP minus the mean of a series of τ-displaced SDPs intended to reduce the stochastically high SDP. The shared peak count is the special case of the SDP where c_i and t_i are binary values. In addition, some *de novo* algorithms, e.g. PEAKS (Ma *et al*., 2003), also use scoring functions similar to the SDP as the objectives to be maximized.

The SDP is conceptually simple, computationally efficient and effective in many cases. However, an inherent drawback of the SDP is that it ignores all correlative information among the dimensions of the spectral vector, corresponding to different fragments. Any match between the fragments in experimental and in theoretical spectra is arbitrarily accepted as the true match. However, due to the reasons mentioned in the above section, many matches may well be stochastic and result finally in false positives.

Kernel SDP (KSDP)

Some fragments tend to occur together in the MS/MS spectrum. Positively correlated fragments potentially include consecutive fragments, complementary fragments, isotopic fragments, and fragments differing in units of charges or the neutral loss of a water or ammonia molecule. Intuitively, when positively correlated fragments are matched together, they are more likely to be true matches and should be assigned a higher score as a whole than as individuals.

One way of improving the SDP by including the correlative information among fragments is to map the spectral vector space non-linearly into a high-dimensional space in which all combinations of correlated fragments have their corresponding dimensions. We call this space the combinational correlative space. The dot product in this space counts all the matched combinations of correlated fragments other than matched individual fragments. However, one problem is how to map the spectral space efficiently to the correlative space so that the dimensions in the correlative space only correspond to combinations of truly correlated fragments. Another issue is that the dimensionality of the correlative space might be too high to compute efficiently in it; i.e. there may be too many combinations of correlated fragments to count one by one.

An idea inspired by the kernel trick (Boser *et al*., 1992; Vapnik, 1995) is to compute directly the dot product in the correlative space with a proper kernel without an explicit mapping from the spectral space to the correlative space. To this end, all predicted fragments are arranged in a manner we call the correlative matrix, as shown in Figure 2. Thus, all correlated fragments cluster together and can be included naturally

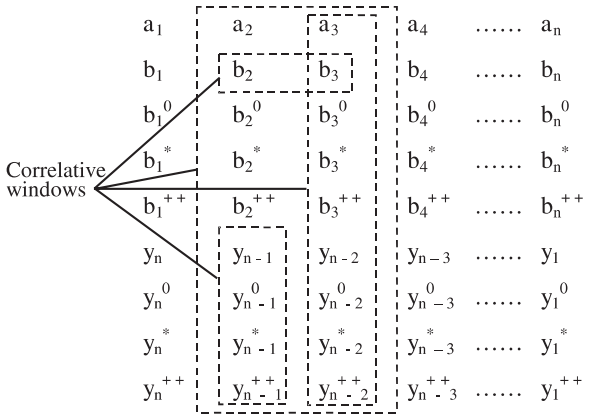

Fig. 2. Correlative matrix and correlative windows.

into the local correlative windows, e.g. the dashed frames in Figure 2. This makes it possible to exploit the locally improved polynomial kernel (Schölkopf *et al*., 1998; Zien *et al*., 2000) to capture this kind of local correlation among fragments.

All predicted fragments are assumed to possess unique m/z values so that all non-zero dimensions in the theoretical spectral vector, $\mathbf{t}$, can be extracted and rearranged into the matrix $\mathbf{T} = (t_{pq})_{m\times n}$ in a manner according to their fragment types and fragmentation positions as shown in Figure 2, where m is the number of fragment types and $n+1$ is the residue number of the peptide precursor. For example, $t_{2,3}$ corresponds to the fragment b_3 in Figure 2. The corresponding dimensions in experimental spectral vector $\mathbf{c}$ are also extracted and rearranged into the matrix $\mathbf{C} = (c_{pq})_{m\times n}$. This can be accomplished as follows. If the predicted fragments are present in the experimental spectrum within a given matching tolerance, then the corresponding position in the matrix $\mathbf{C}$ is set to 1 or the observed intensity; otherwise, it is set to zero. Under the assumption above, we have

$$\mathrm{SDP} = \langle \mathbf{c}, \mathbf{t} \rangle = \sum_{p=1}^{m}\sum_{q=1}^{n} c_{pq}t_{pq}.$$

To introduce correlative information, we define the function $K_{\mathrm{pep}}(\mathbf{c}, \mathbf{t})$ given the peptide pep as

$$\mathrm{win}_j(\mathbf{c}, \mathbf{t}) = \left[\sum_{(p,q)\in U_j} w_{pq}(c_{pq}t_{pq})^{1/d}\right]^d,$$

$$K_{\mathrm{pep}}(\mathbf{c}, \mathbf{t}) = \sum_j \mathrm{win}_j(\mathbf{c}, \mathbf{t}), \qquad (2)$$

where U_j is the set of all subscripts of elements in the j-th correlative window in the matrices $\mathbf{C}$ and $\mathbf{T}$, w_{pq} are weights and d is called the window power. We call the function defined in Equation (2) the KSDP. It is a kernel function (Schölkopf *et al*., 1998) that maps implicitly the spectral space to the

correlative space. For example, when d is equal to two and w_{pq} equal to one, we have

$$\text{win}_j(\mathbf{c},\mathbf{t}) = \langle \phi_j(\mathbf{c}), \phi_j(\mathbf{t}) \rangle,$$
$$\phi_j(\mathbf{s}) = \left[\sqrt{s_{uv} s_{pq}} : (u,v),(p,q) \in U_j \right],$$

where $\mathbf{s}$ represents the spectral vector $\mathbf{c}$ or $\mathbf{t}$, and $\phi_j(\mathbf{s})$ is the vector with the elements $\sqrt{s_{uv}s_{pq}}$, $(u,v),(p,q) \in U_j$. Hence, when c_{pq} and t_{pq} take binary values, $\text{win}_j(\mathbf{c},\mathbf{t})$ amounts to the number of all matched fragment pairs in the j-th correlative window. For d greater than two, the KSDP given in Equation (2) incorporates correlative information of more than two fragments. When d is 1, the KSDP is reduced to the SDP given in Equation (1).

In principle, the window power, d, is required to be a positive integer for the function given in Equation (2) to be a valid kernel. In such a case, d defines the maximum number of correlated fragments. However, since we are not dealing with a kernel-based learning algorithm, this restriction can be removed without side effects. Here, d controls the degree of correlation of fragments for scoring candidate peptides. Therefore, d could be used conveniently as a continuous variable. In addition, such usage allows the two parameters l and d to be tuned effectively.

The correlative window should be defined according to expert knowledge about which fragments are correlated and how or should be learned from labeled data. For instance, when we believe that consecutive fragments are positively correlated, the KSDP can be implemented as

$$K_{\text{pep}}(\mathbf{c},\mathbf{t}) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\sum_{k=j-l_1}^{j+l_2}(w_{|k-j|}(c_{ik}t_{ik})^{1/d})\right]^d, \quad (3)$$

where positive integers l_1 and l_2 equal $\lfloor (l-1)/2 \rfloor$ and $\lceil (l-1)/2 \rceil$, respectively. Integer l is the size of the correlative window. c_{ik} and t_{ik} are set to zero for $k \le 0$ and $k > n$. The weight $w_{|k-j|}$ reflects the assumed correlating strength between the fragments in the position (i,j) and its neighbor with $|k-j|$ residues near it. The KSDP given in Equation (3) can be computed in $O(lmn)$ time in general and in $O(mn)$ time if $w_{|k-j|}$ is equal to one. The experimental analysis of this KSDP will be presented below.

We summarize some advantageous properties of the KSDP. First, the KSDP incorporates correlative information among fragments. It is locally improved, which implies that combinations of uncorrelated fragments are excluded. Second, the KSDP is fragment-type scalable. It is applicable to any fragment series that can be arranged into a correlative matrix. Third and most importantly, the KSDP can be computed efficiently with a low time complexity, often $O(mn)$, the same as that of the SDP.

Relation to traditional approach

SEQUEST and ProbID consider the continuity of matched fragments by counting explicitly the matched consecutive fragment pairs. For comparison, we define the score function PSCORE, which is a simplified form of the score function S_p used by SEQUEST:

$$\text{PSCORE} = n_{\text{f}}(1 + n_{\text{p}}\alpha), \quad (4)$$

where n_{f} is the number of matched fragments, n_{p} is the number of matched consecutive fragment pairs and α is a small positive real number. Such a score function is expected to perform equivalently to the special case of the KSDP given in Equation (4) where the window size, l, is fixed to two. This is demonstrated by experiments.

IMPLEMENTATION

During the above derivation of the KSDP, we assumed that all the predicted fragments of the candidate peptide had a unique m/z value. When several predicted fragments shared a common m/z value, we treated them just as if each of them had a unique m/z. The consequence of such a treatment is that the dimension of this m/z is emphasized to some extent in the final score. Suppose that the candidate peptide did produce the experimental spectrum and several fragments of it share the common m/z; it is rather unlikely that none of these fragments had been detected and measured. On the contrary, if this common m/z is observed in the experimental spectrum, then we consider it as good evidence for the candidate peptide. Therefore, the emphasis on this common m/z is reasonable.

When $w_{|k-j|}$ in Equation (3) is set to one, the KSDP given in Equation (3) can be computed efficiently in time of $O(mn)$ as follows.

$K\_ct = 0;$
for$(i = 1; i \le m; i{+}{+})$
{
  $\text{win}_{i1} = 0;$
  for$(j = 1; j \le 1 + l_2; j{+}{+})$
    $\text{win}_{i1} = \text{win}_{i1} + (c_{ij}t_{ij})^{1/d};$
  $K\_ct = K\_ct + \text{win}_{i1}^d;$
  for$(j = 2; j \le n; j{+}{+})$
  {
    $\text{win}_{ij} = \text{win}_{i,j-1} + (c_{i,j+l_2}t_{i,j+l_2})^{1/d}$
      $- (c_{i,j-l_1-1}t_{i,j-l_1-1})^{1/d};$
    $K\_ct = K\_ct + \text{win}_{ij}^d;$
  }
}

The KSDP algorithm was implemented in our database search tool, pFind, written in C/C++. A Windows interface is provided for entering search parameters and MS/MS data. pFind reports the peptide and protein identification results in HTML/XML files.

RESULTS

Data

The MS/MS spectra used for experiments come from a previously reported dataset of ion trap spectra (Keller *et al*., 2002). A total of 18 purified proteins with different physicochemical properties were mixed together and divided into two mixtures, A and B, which were then digested with trypsin. A total of 14 LC/MS/MS runs and 8 LC/MS/MS runs were performed on the mixtures A and B, respectively. The resulting MS/MS spectra were analyzed using SEQUEST (Eng *et al*., 1994). All spectra were searched against a database including human protein sequences plus the 18 control mixture proteins (denoted by 'human plus mixture database'). From the SEQUEST search results, 2757 spectra were confirmed manually as having been correctly identified.

Out of the 2757 spectra, those with their peptide terminus consistent with the substrate specificity of trypsin were selected for our experiments. The 731 spectra derived from the mixture B (denoted by dataset B) were used to tune the parameters in Equation (3), while the 1323 spectra derived from mixture A (denoted by dataset A) were used to compare pFind with the existing database search tools.

Noise reduction

If unprocessed, the numerous noise peaks in the raw spectra may have led to a heavy computational cost and decreased identification accuracy. Since our purpose is to show the effectiveness of the kernel trick to correlate fragments and tune the parameters in the KSDP, simple data preprocessing was performed. Only the 200 most intense peaks were reserved in each spectrum.

Database and search parameters

Parameter tuning was performed against the human plus mixture database. Comparison with other software tools was performed against the SWISS-PROT database. The trypsin enzyme and up to two missed cleavage sites were specified for digesting theoretically the sequences in their databases. The matching tolerance for the precursor and the fragment, respectively, is set at 3 and 1 Da. The predicted fragment types include b, b^{++}, b^0, y, y^{++} and y^0.

Tuning parameters of KSDP

Although we have constructed the formula for the KSDP given in Equation (3), the choice for the values of the parameters is another issue. In this part of the experiment, two important parameters, the correlative window size, l, and the window power, d, were tuned on the dataset B against the human plus mixture database. The KSDP given in Equation (3) was used directly as the scoring function, with c_{ik} and t_{ik} being binary values and $w_{|k-j|}$ equal to one.

Since d is allowed to be continuous, it must be quantized somehow for the experiments. In addition, it is expected that $d$is more finely tuned for smaller l and more coarsely tuned for larger l. To this end, an auxiliary parameter β was introduced to correlate l and d:

$$d = 1 + \beta(l - 1).$$

Experiments are performed for $l \in \{1, 2, 3, 4, 5, 6\}$ and $\beta \in \{0.2, 0.4, 0.5, 0.6, 0.7, 0.8, 1.0, 1.2, 1.4\}$. A spectrum is regarded as falsely identified if its correct peptide sequence does not rank first in the search result. Figure 3 illustrates the error rate against the changing window size and the window power.

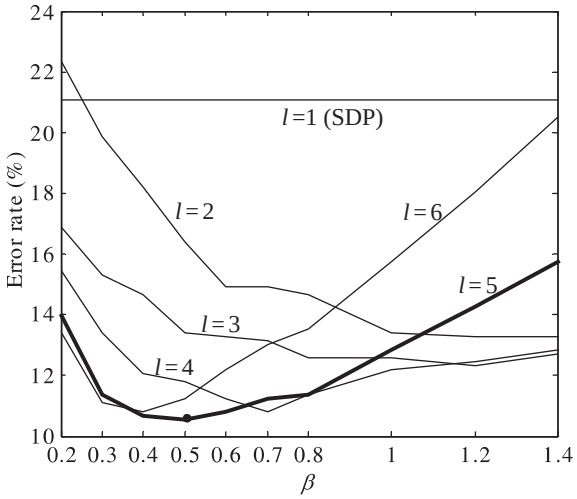

Fig. 3. Curves of the error rate versus the window power for the KSDP given in Equation (3) with various window sizes l. [Note that the window power $d = 1+\beta(l-1)$. The point on the curve indicates the lowest point.]

When $l = 1$, the KSDP given in Equation (3) reduces to the SDP given in Equation (1). When l becomes greater than one, the kernel function takes effect and the error rate drops rapidly. It is seen clearly in Figure 3 that nearly all the error rates for $l > 1$ are remarkably lower than that for $l = 1$. The lowest error rate is obtained when $l = 5$ and $\beta = 0.5$ ($d = 3$). Compared with the SDP, the KSDP decreases the error rate by 10% at best in this experiment. The role of the correlative window in reducing stochastic mismatches is significant.

Figure 4 presents the curve of the error rate on dataset B for the scoring function PSCORE given in Equation (4). The curve of the error rate for the KSDP with the window size 2 is also plotted for comparison. Although the variable parameters for the two curves are different, interesting phenomena can be observed from this forcible combination. The two curves share the lowest value of about 13% and overlap each other largely. This suggests that PSCORE is almost equivalent in performance to the special case of the KSDP, where the window size is fixed to two. It can be explained by observing that both the scoring functions only consider the continuity of two matched fragments. When the continuity of more matched

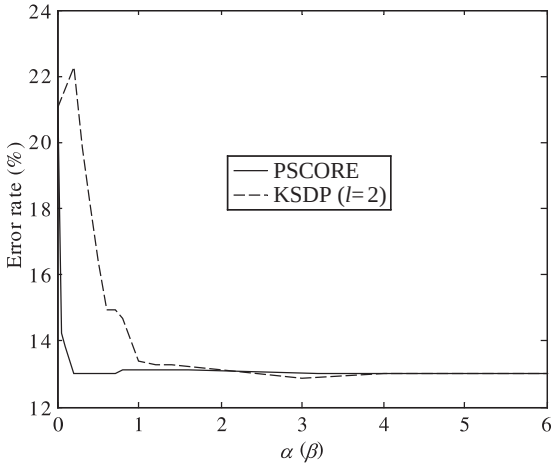

Fig. 4. Curve of the error rate versus the parameter α for the PSCORE given in Equation (4) and curve of the error rate versus the parameter β for the KSDP given in Equation (3) with the window size, l, fixed at two. (Note that window power, d, here equals $1+\beta$.)

fragments is considered, it is possible to obtain a lower error rate as illustrated in Figure 3.

Comparing pFind with existing software tools

SEQUEST and Sonar MS/MS are two popular software tools that use SDP-based scoring algorithms. In this part of the experiment, pFind is compared with them in terms of identification accuracy.

The KSDP tuned above ignores some important information such as the observed intensities of fragments and the length of peptides. To include such information, the following score, based on the KSDP, is used as a practical scoring method in pFind.

$$\left(\sum_i \sqrt{I_i}\right) \cdot K/L, \tag{5}$$

where I_i is the observed intensity of the i-th matched fragment, K is the KSDP defined in Equation (3), with c_{ik} and t_{ik} being binary values and $w_{|k-j|}$ equal to one, and L is the length of the candidate peptide being scored. We found empirically that reducing appropriately the observed intensities could improve the identification accuracy. Therefore, the square roots of the intensities are summed instead of their original values. The denominator, L, plays the normalizing role in avoiding the inherent bias of the KSDP toward large peptides.

The spectra in dataset A described above are searched against the SWISS-PROT database. Table 2 shows the search results of the three software tools pFind, SEQUEST (version 2.7) and Sonar MS/MS (version 2002.07.01.04). The same search parameters are set for the three search engines. The scoring method given in Equation (5) is used in pFind with the experimentally optimized correlative window size 5 and the window power 3.

Table 2. Comparison of three software tools on the dataset A against the SWISS-PROT database

Dataset	Total	Correctly identified		
		pFind	SEQUEST	Sonar MS/MS
A-1, 2	230	218	214	184
A	1323	1262	1257	—

Since the software Sonar MS/MS available on the Internet cannot batch spectra files, only the 230 spectra derived from the first two LC/MS/MS runs on mixture A (denoted by A-1, 2) are searched with Sonar MS/MS.

As shown in Table 2, pFind performs remarkably better than Sonar MS/MS. But the advantage over the industry standard software, SEQUEST, is slim. This may be due partly to the rough data preprocessing and simple form of the scoring function in the current version of pFind. It is also important to realize that the test dataset may tend to favor SEQUEST as it was with the analysis by SEQUEST that the correct peptide sequences of these spectra were recovered.

CONCLUSIONS

This paper provides a novel approach to utilizing the correlative information among fragment ions in a tandem mass spectrum to improve the peptide identification accuracy by database searching. The common scoring method, the tandem mass SDP, is extended to the KSDP. By virtue of the kernel trick, the KSDP avoids enumeration of the exponential combinations of correlated fragments and thereby has a low computational complexity. The experiments on a dataset reported previously demonstrate the effectiveness of the KSDP in correlating consecutive fragments. The error rate decreases by 10% at best compared with the SDP because the SDP ignores all correlative information. Experiments also suggest that the traditional approach to considering correlative information resembles a special case of the KSDP that is much less optimal. Our software tool, pFind, using a simple scoring function based on the KSDP, outperforms two SDP-based software tools, SEQUEST and Sonar MS/MS, in terms of identification accuracy.

Correlative windows for more types of correlation will be investigated as our future work. The flexible definition of correlative windows, as a key for the KSDP to succeed, can potentially be learned from training data. The elaborate spectrum preprocessing and sophisticated scoring algorithm based on the KSDP will be added to the new version of the pFind system. Finally, we emphasize that any existing scoring algorithm based on the SDP can be extended to incorporate

the correlative information among fragment ions simply by replacing the SDP with the KSDP.

ACKNOWLEDGEMENTS

This work was supported by the National Key Basic Research & Development Program (973) of China under Grant No. 2002CB713807. We thank Simin He, Yiqiang Chen, Hu Zhou and Xiaobiao Wang for valuable discussions. We acknowledge Dr Andrew Keller for providing the MS/MS dataset, as well as the referees for helpful comments.

REFERENCES

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

Bafna,V. and Edwards,N. (2001) SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, **17**(Suppl. 1), S13–S21.

Boser,B.E., Guyon,I.M. and Vapnik,V.N. (1992) A training algorithm for optimal margin classifiers. In Haussler,D. (ed.), *5th Annual ACM Workshop on COLT*. ACM Press, Pittsburgh, PA, pp. 144–152.

Chen,T., Kao,M.Y., Tepel,M., Rush,J. and Church,J. (2001) A dynamic programming approach to *de novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.*, **8**, 325–337.

Clauser,K.R., Baker,P. and Burlingame,A.L. (1999) Role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.*, **71**, 2871–2882.

Danick,V., Addona,T.A., Clauser,K.R., Vath,J.E. and Pevzner,P.A (1999) *De novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.*, **6**, 327–342.

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

Fenyö,D., Qin,J. and Chait,B.T. (1998) Protein identification using mass spectromic information. *Electrophoresis*, **19**, 998–1005.

Fenyö,D. and Beavis,R.C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.*, **75**, 768–774.

Field,H.I., Fenyö,D. and Beavis,R.C. (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics*, **2**, 36–47.

Keller,A., Purvine,S., Nesvizhskii,A.I., Stolyar,S., Goodlett,D.R. and Kolker,E. (2002) Experimental protein mixture for validating tandem mass spectral analysis. *Omics*, **6**, 207–212.

Ma,B., Zhang,K.Z., Hendrie,C., Liang,C.Z., Li,M., Doherty-Kirby,A. and Lajoie,G. (2003) PEAKS: powerful software for peptide de novo sequencing by MS/MS. *Rapid Commun. Mass Spectrom.*, **17**, 2337–2342.

Mann,M. and Wilm,M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem.*, **66**, 4390–4399.

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.

Pevzner,P.A., Dancik,V. and Tang,C.L. (2000) Mutation-tolerant protein identification by mass-spectrometry. *J. Comput. Biol.*, **7**, 777–787.

Schölkopf,B., Simard,P., Smola,A. and Vapnik,V. (1998) Prior knowledge in support vector kernels. In Jordan,M., Kearns,M. and Solla,S. (eds), *Advances in Neural Information. Processing Systems 10*. MIT Press, Cambridge, MA, pp. 640–646.

Tabb,D.L., MacCoss,M.J., Wu,C.C., Anderson,S.D. and Yates,J.R. (2003) Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal. Chem.*, **75**, 2470–2477.

Taylor,J.A. and Johnson,R.S. (1997) Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, **11**, 1067–1075.

Taylor,J.A. and Johnson,R.S. (2001) Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. *Anal. Chem.*, **73**, 2594–2604.

Vapnik,V.N. (1995) *The Nature of Statistical Learning Theory*. Springer, New York.

Wan,K.X., Vidavsky,I. and Gross,M.L. (2002) Comparing similar spectra: from similarity index to spectral contrast angle. *J. Am. Soc. Mass Spectrom.*, **13**, 85–88.

Zhang,N., Aebersold,R. and Schwikowski,B. (2002) ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*, **2**, 1406–1412.

Zien,A., Raetsch,G., Mika,S., Schölkopf,B., Lengauer,T. and Mueller,K. (2000) Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, **16**, 799–807.