A Kernel-Based Case Retrieval Algorithm with Application to Bioinformatics

Yan Fu[1,2], Qiang Yang[3], Charles X. Ling[4], Haipeng Wang[1], Dequan Li[1], Ruixiang Sun[2], Hu Zhou[5], Rong Zeng[5], Yiqiang Chen[1], Simin He[1], and Wen Gao[1,2]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
{yfu,hpwang,dqli,yqchen,smhe,wgao}@ict.ac.cn
[2] Graduate School of Chinese Academy of Sciences, Beijing 100039, China
{rxsun,wgao}@gscas.ac.cn
[3] Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
qyang@cs.ust.hk
[4] Department of Computer Science, The University of Western Ontario, London, Ontario N6A 5B7,Canada
cling@csd.uwo.ca
[5] 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
{hzhou,zr}@sibs.ac.cn

Abstract. Case retrieval in case-based reasoning relies heavily on the design of a good similarity function. This paper provides an approach to utilizing the correlative information among features to compute the similarity of cases for case retrieval. This is achieved by extending the dot product-based linear similarity measures to their nonlinear versions with kernel functions. An application to the peptide retrieval problem in bioinformatics shows the effectiveness of the approach. In this problem, the objective is to retrieve the corresponding peptide to the input tandem mass spectrum from a large database of known peptides. By a kernel function implicitly mapping the tandem mass spectrum to a high dimensional space, the correlative information among fragment ions in a tandem mass spectrum can be modeled to dramatically reduce the stochastic mismatches. The experiment on the real spectra dataset shows a significant reduction of 10% in the error rate as compared to a common linear similarity function.

1 Introduction

Case-based reasoning (CBR) relies on the use of a similarity function to rank previous cases for solving new problems [13, 14]. Over the years, CBR has enjoyed tremendous success as a technique for solving problems related to knowledge reuse, with many important industrial applications [22]. The central component of a CBR system is a similarity function, based on which cases are retrieved and ranked for adaptation and further solution [13, 14]. Because of its importance, various methods have been proposed to compute the similarity between cases, including that of [2, 3, 14, 18, 19].

C. Zhang, H.W. Guesgen, W.K. Yeap (Eds.): PRICAI 2004, LNAI 3157, pp. 544–553, 2004.

Although various methods for learning the weights have been designed, no specific similarity function has been designed to take advantage of the correlative information between features using a nonlinear similarity function. An exception is the collaborative filtering framework, in which a linear weighting function is used to represent the correlative information.

This paper presents a general approach to engineering nonlinear similarity functions for scoring cases, and highlights an application [11] of the new method to the peptide retrieval problem in bioinformatics. A central characteristic of this problem is that the correlated features should play a more important role in scoring the cases than other, non-correlated features. In order to emphasize the correlations, we apply the kernel functions to those correlated features. This implicitly translates the cases from the original space to a feature space with new dimensions for combinations of correlated features. Thanks to the kernel trick, nonlinear similarity functions can be constructed in the original space with slight overhead. We show that the resulting similarity function dramatically improves the retrieval accuracy.

Mass spectrometry is currently one of the most important techniques for proteomics research [1]. Protein identification via tandem mass spectrometry (MS/MS) is the central task in MS/MS based proteomics. For example, for the diagnosis and therapy of diseases, investigation on the differently expressed proteomes in normal and abnormal cells is very important. High precision and high-throughput protein identification via MS/MS needs not only elaborate biophysical instruments but also powerful computer algorithms. The basic computational problem is to retrieve the peptide sequence from which the observed MS/MS spectrum was derived through a search for the most similar theoretical MS/MS spectrum in a large database of known peptides. In this paper, we show that the peptide retrieval problem can be expressed as a case-based reasoning problem, in which the peptide sequences correspond to the cases while MS/MS spectra correspond to the features of cases. By using a kernel function to improve a common linear similarity measure for comparing MS/MS spectra, we show that much better retrieval accuracy can be obtained.

Below, we first introduce how to design kernel-based nonlinear similarity functions for the case retrieval. Then we apply the proposed approach to the peptide retrieval problem.

2 Applying the Kernel Trick to Similarity Measurement

For the measurement of the similarity between cases, they are usually presented as feature vectors. One of the simplest similarity measures between two vectors is their dot product, i.e. $\langle \mathbf{x}, \mathbf{y} \rangle$, where $\mathbf{x}$, $\mathbf{y}$ are n-dimensional feature vectors. For the binary features, the dot product counts the number of features that two cases possess in common. The cosine of the angle between vectors and the Euclidean distance can both be expressed in terms of the dot products, i.e.

$$\cos(\mathbf{x},\mathbf{y}) = \langle \mathbf{x},\mathbf{y} \rangle \Big/ \sqrt{\langle \mathbf{x},\mathbf{x} \rangle \cdot \langle \mathbf{y},\mathbf{y} \rangle}, \text{ and}$$

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\|\mathbf{x}-\mathbf{y}\|^2} = \sqrt{\langle \mathbf{x},\mathbf{x} \rangle + \langle \mathbf{y},\mathbf{y} \rangle - 2\langle \mathbf{x},\mathbf{y} \rangle}.$$

In reality, similar cases may not be close to each other geometrically in the original vector space, which we call the input space. In such cases, we may wish to map the original space to a new, usually higher dimensional space with an aim for the similar cases to get closer in the new space. The transformed space is called the feature space. For instance, when the elements of the input vector highly correlate with each other, we may want the feature space to include as new dimensions all the *d*-order products of the dimensions in the input space. However, the dimensionality of the feature space might be too high to compute efficiently and explicitly.

The kernel trick, popularly used in the machine learning [20], overcomes this difficulty gracefully. A kernel is a function k such that for all $\mathbf{x}, \mathbf{y} \in A$ (usually $A = R^n$),

$$k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle,$$

where ϕ is a mapping from the input space A to a feature space B. Usually, ϕ is a nonlinear mapping and the feature space B is of very high, or even infinite, dimensions. Therefore, any computation that is exclusively based on the dot product in the feature space can be performed with the kernel $k(\mathbf{x}, \mathbf{y})$ from the input space, thus avoiding the explicit mapping ϕ from the input space to the feature space.

For example, the polynomial kernel, $\langle \mathbf{x}, \mathbf{y} \rangle^d$, implicitly maps the n-dimensional input space to a much higher dimensional feature space with all d-order products of the dimensions of the input space as new dimensions. For instance, when $d = 2$ and $\mathbf{x}, \mathbf{y} \in R^2$, we have

$$\langle \mathbf{x}, \mathbf{y} \rangle^d = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle,$$

$$\phi(\mathbf{x}) = (x_1^2, x_2^2, x_1 x_2, x_2 x_1).$$

Kernels have been widely used recently to extend linear learning algorithms to nonlinear versions, e.g. [5, 17]. However, kernels are not limited to learning algorithms. Being the dot products in the feature space, kernels can be used to construct nonlinear version of any linear algorithm as long as only the dot product is involved. In the case of the case retrieval problem, we obtain the following kernel-based similarity and distance measures for cases $\mathbf{x}$ and $\mathbf{y}$ (where we use a $\cos'(\mathbf{x},\mathbf{y})$ and $d'(\mathbf{x},\mathbf{y})$ for the new cosine and distance functions):

$$k(\mathbf{x}, \mathbf{y}),$$

$$\cos'(\mathbf{x}, \mathbf{y}) = k(\mathbf{x}, \mathbf{y}) \Big/ \sqrt{k(\mathbf{x}, \mathbf{x}) \cdot k(\mathbf{y}, \mathbf{y})}, \text{ and}$$

$$d'(\mathbf{x}, \mathbf{y}) = \sqrt{k(\mathbf{x}, \mathbf{x}) - 2k(\mathbf{x}, \mathbf{y}) + k(\mathbf{y}, \mathbf{y})}.$$

The success of these similarity and distance measures depends on the proper definition of the kernel $k(\mathbf{x}, \mathbf{y})$, which should incorporate the available apriori knowledge in the specific domain. In the following, we show how a kernel can incorporate the domain knowledge and is directly used as the similarity measure in a bioinformatic application. We first introduce the peptide retrieval problem in detail.

3 The Peptide Retrieval Problem

Via MS/MS, protein identification problem is divided into peptides identification sub-problem. In the mass spectrometer, peptides derived from digested proteins are ionized. Peptide precursor ions of a specific mass-charge ratio (*m*/*z*) are fragmented by the collision-induced dissociation (CID). Product ions are detected. The measured *m*/*z* and intensity of the product ions form the peaks in the MS/MS spectrum. A peptide is a string of amino acid residues joined together by peptide bonds. For the low-energy CID, the *b*-*y* type backbone cleavage is most frequent and usually occurs only once in each peptide, resulting in *b* and *y* series of fragment ions, as shown in Fig. 1. The *b* fragments are the N-terminal sub-strings of amino acid residues dissociated from the cleaved peptide precursors, while the *y* fragments are the C-terminal sub-strings. The fragments can be singly charged or multiply charged and may possibly lose a neutral water or ammonia molecule. Besides these fragments, the noise and product ions derived from unexpected peptide cleavages also present themselves as peaks in the MS/MS spectrum.

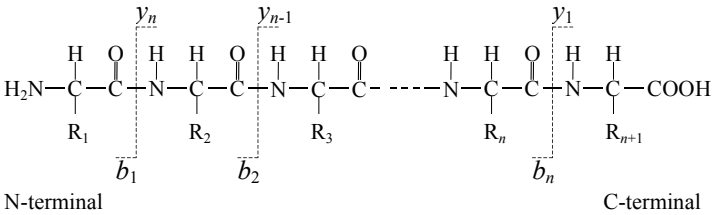

Fig. 1. *b* and *y* fragment ions resulting from peptide bonds cleavage by collision-induced dissociation

To identify the peptide sequence of the observed MS/MS spectrum, the database searching approach has been widely used. The peptide retrieval problem can be expressed as follows: given the experimental MS/MS spectrum *S*, the peptides database *D*, and background conditions *C*, find the peptide *pep** in *D* from which *S* derived. During a retrieval, peptide sequences in the database are fragmented theoretically to construct the theoretical MS/MS spectra. The experimental and theoretical MS/MS spectra are compared to find the target peptide. Expressed in terms of case retrieval, peptide sequences correspond to the cases while MS/MS spectra correspond to the features of cases. This paper focuses on the use of dot product similarity to compare MS/MS spectra for scoring the peptides in the peptide retrieval problem.

Various strategies have been proposed for scoring peptides in existing peptide retrieval software tools [4, 6, 7, 8, 10, 15, 23]. In existing peptide-scoring algorithms, the *Spectral Dot Product* (*SDP*) is often involved directly or indirectly and plays an important role. In *SDP*, the thereotical and experimental MS/MS spectra are represented as two *N*-dimensional spectral vectors, denoted by $\mathbf{c} = [c_1, c_2, \ldots, c_N]$ and $\mathbf{t} = [t_1, t_2, \ldots, t_N]$, respectively, where *N* is the number of different *m*/*z* values used, c_i and t_i are binary values {0, 1} or the intensities of the peaks at the *i*-th *m*/*z* value in MS/MS spectra. The *SDP* is defined as

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

The *SDP*-based cosine function of the angle between spectral vectors was used as an MS/MS spectrum similarity measure [21]. Sonar MS/MS [9, 10] explicitly adopted the spectral vector representation and the *SDP* for scoring peptides. The notion of cross-correlation in the SEQUEST [7] is in nature equivalent to the *SDP*. The shared peak count in early work is the special case of the *SDP* with c_i and t_i being binary values. An inherent drawback of the *SDP* is that it does not especially leverage correlative information among the dimensions of spectral vectors corresponding to different fragments. This increases the possibility of stochastic mismatches.

4 The Kernel-Based Correlative Similarity Function

Our most important observation about the MS/MS spectrum is that the fragments resulting from peptide bonds cleavage by CID rarely occur independently; most often they tend to occur correlatively with each other. Intuitively, when positively correlated fragments are matched together, the matches are more reliable and should be emphasized somehow for scoring the candidate peptide.

Example 1. Two peptide sequences TSDANINWNNLK and FQDLVDAVRAEK, denoted by pep_{corr} and $\text{pep}_{\text{incorr}}$ respectively, have the same length and nearly the same peptide mass. Suppose that an observed MS/MS spectrum was derived from the peptide precursors with the sequence pep_{corr}. To identify the peptide sequence, a retrieval is performed against the database containing the two peptide sequences pep_{corr} and $\text{pep}_{\text{incorr}}$. For simplicity, the *y* series of fragments is used to construct theoretical MS/MS spectra. Compared with the observed spectrum, the correct peptide pep_{corr} has six matched fragments including y_3, y_4, y_5, y_6, y_7, and y_8 as shown in Fig. 2, while the peptide $\text{pep}_{\text{incorr}}$ has seven matched fragments including y_2, y_4, y_5, y_6, y_9, y_{10}, and y_{11}. Although there are more matched fragments for the peptide $\text{pep}_{\text{incorr}}$, the matches for the pep_{corr} are more consecutive and therefore should be considered as strong indicators of correct answer.

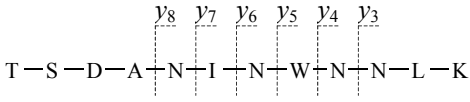

Fig. 2. Consecutive fragments produced from the fragmentation of the peptide precursors with sequence TSDANINWNNLK

To consider the correlation among fragments, we may exhaustively examine whether each possible combination of correlated fragments is matched as a whole. However, there may be too many such combinations to count one by one. Alternatively, since we are only interested in the overall similarity between two MS/MS spectra rather than the detailed matching results of the individual fragment combina-

tions, we can design the similarity function with kernels as a final sum of all the matching results. To this end, all predicted fragments in the theoretical MS/MS spectrum are arranged in a manner we call the correlative matrix, as shown in Fig. 3, thus making correlated fragments cluster into the local correlative windows. This kind of local correlation can be emphasized with the locally improved polynomial kernel [16].

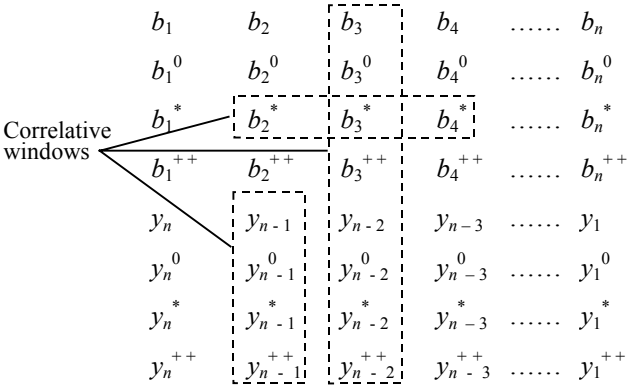

Fig. 3. Correlative matrix and correlative windows. The subscript number indicates the fragmentation position as illustrated in Fig. 1, superscripts 0 and * indicate a neutral loss of H_2O and NH_3, respectively

We assume that all predicted fragments have their corresponding unique *m*/*z* values. The only influence of this assumption is that the shared *m*/*z* values are emphasized to some extent. We regard such emphasis as reasonable, since the shared *m*/*z* values should be of more importance than other, unique *m*/*z* values. Under this assumption, all non-zero dimensions in the theoretical spectral vector **t** can be arranged into a matrix $\mathbf{T} = (t_{pq})_{m\times n}$ in accordance with their fragment types and fragmentation positions, 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_3^0 in Fig. 3. For an experimental spectral vector **c**, the dimensions at the *m*/*z* value corresponding to t_{pq} are also arranged into a matrix $\mathbf{C} = (c_{pq})_{m\times n}$. Under the above assumption, we have

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

The correlative window may be defined according to the biologists' expert knowledge about how fragments are correlated. With the observation that the continuity of matched fragments is the most important correlation, we define the *Kernel Spectral Dot Product* (*KSDP*) [11] for consecutive fragments with locally improved kernel as follows,

$$K(\mathbf{c}, \mathbf{t}) = \sum_{i=1}^{m} \sum_{j=1}^{n} win_{ij}(\mathbf{c}, \mathbf{t}), \tag{2}$$

$$win_{ij}(\mathbf{c},\mathbf{t}) = \left[\sum_{k=j-l_1}^{j+l_2} \left(w_{|k-j|} \left(c_{ik} t_{ik} \right)^{1/d} \right) \right]^d ,$$

where positive integers l_1 and l_2 are $\lfloor (l-1)/2 \rfloor$ and $\lceil (l-1)/2 \rceil$ respectively, the window power d is a positive integer and defines the maximum number of consecutive fragments to be considered. 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 to it. The *KSDP* given in Eq. (2) can be computed in O(lmn) time in general and in O(mn) time if $w_{|k-j|}$ equals one. The experimental analysis of this *KSDP* is presented in the next section.

In Example 1, $m = 1$ and $n = 11$. When c_{ik} and t_{ik} are binary values, we have

$$\mathbf{T}_{\text{corr}} = \mathbf{T}_{\text{incorr}} = \mathbf{T} = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1),$$

$$\mathbf{C}_{\text{corr}} = (0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0),$$

$$\mathbf{C}_{\text{incorr}} = (0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1).$$

Therefore,

$$SDP_{\text{corr}} = \langle \mathbf{T}_{\text{corr}}, \mathbf{C}_{\text{corr}} \rangle = 6,$$

$$SDP_{\text{incorr}} = \langle \mathbf{T}_{\text{incorr}}, \mathbf{C}_{\text{incorr}} \rangle = 7,$$

$$KSDP_{corr}\Big|_{l=5,d=3,w=1} = \sum_{j=1}^{11} \left[\sum_{k=j-2}^{j+2} \left(c_{corr1,k} t_{1,k} \right) \right]^3 = 450 ,$$

$$KSDP_{incorr}\Big|_{l=5,d=3,w=1} = \sum_{j=1}^{11} \left[\sum_{k=j-2}^{j+2} \left(c_{incorr1,k} t_{1,k} \right) \right]^3 = 289 .$$

Thus, The correct peptide obtains a lower *SDP* score but a higher *KSDP* score in virtue of the kernel function to correlate consecutive fragments.

5 Experimental Results

The MS/MS spectra used for experiments come from a dataset of ion trap spectra reported in [12]. 18 purified proteins were mixed and digested with trypsin. 22 LC/MS/MS runs were performed on this digested mixture. The generated MS/MS spectra were searched using the SEQUEST software against a database including human protein sequences and the 18 control mixture proteins (denoted by "human plus mixture database"). Search results were manually examined, and 2757 of them were confirmed as true positives.

From the 2757 spectra with their peptide sequences correctly recovered, 2054 spectra with their peptide terminus consistent with the substrate specificity of trypsin are

selected for our experiments to make the experiments more manageable. To reduce the noise in the original spectra, only the 200 most intense peaks are retained in each spectrum. In retrieval, trypsin and up to two missed cleavage sites are specified for theoretically digesting the sequences in the database. 3 daltons and 1dalton are set as the matching tolerances for the precursor and the fragment respectively. b, b^{++}, b^0, y, y^{++}, and y^0 are specified as the predicted fragment types.

To tune the two parameters, window size l and window power d, experiments are performed for $l \in \{1, 2, 3, 4, 5, 6, 7, 8\}$ and $d \in \{2, 3, 4\}$ against the human plus mixture database. The *KSDP* given in Eq. (2) is directly used as the similarity function with c_{ik} and t_{ik} being binary values and $w_{|k-j|}$ equal to one. The error rates vs. the parameters are illustrated in Fig. 4, in which erroneous identification indicates the fact that the correct peptide does not rank first in the search result.

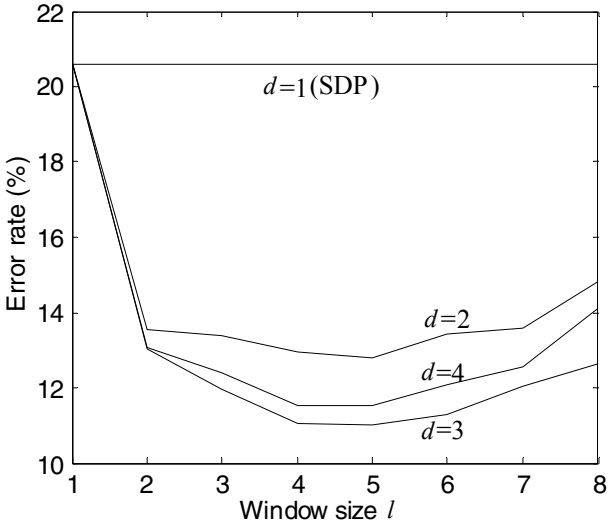

Fig. 4. Error rates vs. the window size l and window power d in *KSDP* given in Eq. (2)

Compared with the *SDP*, the *KSDP* decreases the error rate by 10% at best in this experiment. The lowest error rate is obtained when $d = 3$ and $l = 4$ or 5. It can also be observed that for all tested values of l, the lowest error rate is obtained when $d = 3$; and, for all tested values of d, the lowest error rate is obtained when $l = 4$ or 5. Therefore, we have a good reason to regard window size 4 or 5 and window power 3 as the approximate optimal parameters.

When $l = 1$ or $d = 1$, the *KSDP* given in Eq. (2) reduces to the *SDP* given in Eq. (1). When l and d become larger than one, the kernel function takes effect and the error rate drops rapidly. It is clearly shown in Fig. 4 that nearly all the error rates for $l > 1$ are remarkably lower than that for $l = 1$. The same claim stands for the parameter d. The role of the kernel to reduce stochastic mismatches is significant.

6 Conclusions and Future Work

This paper provides an approach to utilizing the correlative information among features to compute the similarity of cases for case retrieval. This is achieved by extending the dot product-based linear similarity measures to their nonlinear versions with kernel functions. An application to the peptide retrieval problem in bioinformatics shows the effectiveness of the approach. The common linear similarity measure for tandem mass spectra, *Spectral Dot Product* (*SDP*), is extended to the *Kernel SDP* (*KSDP*) to take advantage of the correlative information among fragment ions. The experiments on a previously reported dataset demonstrate the effectiveness of the *KSDP* to reduce stochastic mismatches. In the future, we wish to apply the proposed method to other case retrieval problems.

Acknowledgements

This work was supported by the National Key Basic Research & Development Program (973) of China under Grant No. 2002CB713807. Qiang Yang was also supported by a Hong Kong RGC grant. We acknowledge Dr. Andrew Keller for providing the MS/MS dataset.

References

1. Aebersold, R., Mann, M.: Mass Spectrometry-Based Proteomics. Nature 422(2003) 198–207
2. Agnar, A., Plaza, E.: Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications 7 (1994) 39–59
3. Aha, D.W., Kibler, D., Albert, M.K.: Instance-Based Learning Algorithms. Machine Learning 6 (1991) 37–66
4. Bafna, V., Edwards, N.: SCOPE: a Probabilistic Model for Scoring Tandem Mass Spectra against a Peptide Database. Bioinformatics 17 Suppl. 1 (2001) S13–S21
5. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A Training Algorithm for Optimal Margin Classifiers. In: Haussler, D. (ed.): Proceedings of the Fifth Annual Workshop on Computational Learning Theory. ACM Press, Pittsburgh, PA (1992) 144–152
6. Clauser, K.R., Baker, P., Burlingame, A.L.: Role of Accurate Mass Measurement (± 10 ppm) in Protein Identification Strategies Employing MS or MS/MS and Database Searching. Anal. Chem. 71 (1999) 2871–2882
7. Eng, J.K., McCormack, A.L., Yates, J.R.: An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. J. Am. Soc. Mass Spectrom. 5 (1994) 976–989
8. Fenyö, D., Qin, J., Chait, B.T.: Protein Identification Using Mass Spectromic Information. Electrophoresis 19 (1998) 998–1005
9. Fenyö, D., Beavis, R.C.: A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications Using General Scoring Schemes. Anal. Chem. 75 (2003) 768–774
10. Field, H.I., Fenyö, D., Beavis, R.C.: RADARS, a Bioinformatics Solution that Automates Proteome Mass Spectral Analysis, Optimises Protein Identification, and Archives Data in a Relational Database. Proteomics 2 (2002) 36–47

11. Fu, Y., Yang, Q., Sun, R., Li, D., Zeng, R., Ling, C.X., Gao, W. Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. Bioinformatics (2004) 10.1093/bioinformatics/bth186
12. Keller, A., Purvine, S., Nesvizhskii, A.I., Stolyar, S., Goodlett, D.R., Kolker, E.: Experimental Protein Mixture for Validating Tandem Mass Spectral Analysis. Omics 6 (2002) 207–212
13. Kolodner, J.L.: Case-Based Reasoning. Morgan Kaufmann Publisher, California (1993)
14. Leake, D.B., Kinley, A., Wilson, D.: Case-Based Similarity Assessment: Estimating Adaptability from Experience. In: Proceedings of the 14th National Conference on Artificial Intelligence. AAAI Press, Menlo Park, California (1997)
15. Perkins, D.N., Pappin, D.J., Creasy, D.M., Cottrell, J.S.: Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. Electrophoresis 20 (1999) 3551–3567
16. Schölkopf, B., Simard, P., Smola, A.J., Vapnik, V.: Prior Knowledge in Support Vector Kernels. In: Jordan, M., Kearns, M., Solla, S. (eds.): Advances in Neural Information. Processing Systems 10. MIT Press, Cambridge, MA (1998) 640–646
17. Schölkopf, B., Smola, A.J., Müller, K.R.: Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation 10 (1998) 1299–1319
18. Smyth, B., Keane, M.T.: Remembering to Forget: A Competence Preserving Deletion Policy for Case-Based Reasoning Systems. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Morgan-Kaufmann (1995) 377–382
19. Smyth, B., Keane, M.T.: Adaptation-Guided Retrieval: Questioning the Similarity Assumption in Reasoning. Artificial Intelligence 102 (1998) 249–293
20. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1995)
21. Wan, K.X., Vidavsky, I., Gross, M.L.: Comparing Similar Spectra: from Similarity Index to Spectral Contrast Angle. J. Am. Soc. Mass Spectrom. 13 (2002) 85–88
22. Watson, I.: Applying Case-Based Reasoning: Techniques for Enterprise Systems. Morgan Kaufmann Publisher, Inc., California, USA (1997)
23. Zhang, N., Aebersold, R., Schwikowski, B.: ProbID: A Probabilistic Algorithm to Identify Peptides through Sequence Database Searching Using Tandem Mass Spectral Data. Proteomics 2 (2002) 1406–1412