BIOINFORMATICS

Vol. 20 no. 5 2004, pages 735–741
DOI: 10.1093/bioinformatics/btg477

Bio-support vector machines for computational proteomics

Zheng Rong Yang[1,*] *and Kuo-Chen Chou*[2,3]

[1]*Department of Computer Science, Exeter University, Exeter EX4 4PT, UK,* [2]*Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA and* [3]*Tianjin Institute of Bioinformatics & Drug Discovery (TIBDD), Tianjin, China*

Received on June 25, 2003; revised on August 7, 2003; accepted on August 19, 2003
Advance Access publication January 29, 2004

ABSTRACT

Motivation: One of the most important issues in computational proteomics is to produce a prediction model for the classification or annotation of biological function of novel protein sequences. In order to improve the prediction accuracy, much attention has been paid to the improvement of the performance of the algorithms used, few is for solving the fundamental issue, namely, amino acid encoding as most existing pattern recognition algorithms are unable to recognize amino acids in protein sequences. Importantly, the most commonly used amino acid encoding method has the flaw that leads to large computational cost and recognition bias.

Results: By replacing kernel functions of support vector machines (SVMs) with amino acid similarity measurement matrices, we have modified SVMs, a new type of pattern recognition algorithm for analysing protein sequences, particularly for proteolytic cleavage site prediction. We refer to the modified SVMs as bio-support vector machine. When applied to the prediction of HIV protease cleavage sites, the new method has shown a remarkable advantage in reducing the model complexity and enhancing the model robustness.

Contact: Z.R.Yang@exeter.ac.uk

INTRODUCTION

Pattern recognition algorithms including artificial neural networks (ANNs) have been widely used in analysing biological sequences. The neural learning algorithms used in analysing protein sequence data are back-propagation neural networks (BPNNs), self-organizing maps (SOM), and recurrent neural networks (RNNs). The application of a BPNN to Human Immunodeficiency Virus (HIV) protease cleavage sites prediction yielded a prediction accuracy of 90–92% (Cai and Chou, 1998; Yang, 2001; Narayanan *et al*., 2002). The use of BPNNs resulted in about 65–70% of accuracy in secondary structure prediction (Reczko, 1993; Baldi *et al*., 2000 Pollastri *et al*., 2002), while the use of the RNNs has improved the prediction accuracy of secondary structure up to 75% (Baldi *et al*., 2000). SOM has also been used to identify motifs and families in the context of unsupervised learning (Arrigo *et al*., 1991). However, the problem with using ANNs to analyse biological data is that most ANNs cannot recognize non-numerical features such as the biochemical codes of amino acids. Investigating a proper encoding process prior to modelling the amino acids is then critical.

The most common encoding of amino acids is the distributed method, in which 20 binary bits are used to represent each amino acid (Qian and Sejnowski, 1988). For instance, 'Alanine' is expressed by 0000000000 00000000001, 'Cysteine' 0000000000 0000000010 and 'Aspartate' 0000000000 0000000100. However, there are the following three problems with this method. (1) The input space will be unnecessarily over-expanded, leaving a large part of the space unused. (2) As a consequence, the ratio of the sequence number against the space dimension would be significantly decreased, causing difficulty in neural learning. (3) The use of Euclidean space has no theoretic foundation in biology or chemistry and hence might reduce the accuracy of a model. According to the numerical assignment in the distributed method, the distance between any two different amino acids is $\sqrt{2}$. This would conflict with the reality in biology. Actually, different distances for different amino acid pairs have been defined by various mutation matrices, and validated (Dayhoff *et al*., 1978; Johnson and Overington, 1993). There are 15 commonly used mutation matrices, of which Dayhoff's is generally superior (Johnson and Overington, 1993). However, they cannot be used for encoding an amino acid to a unique numerical value directly. The development of bio-basis function neural network (BBFNN) using amino acid similarity matrix has improved the time complexity and robustness of pattern recognition algorithms in analysing biological sequences. For instance, the use of BBFNNs in a variety of proteolytic cleavage activity prediction and others has shown the above advantages. The application of BBFNNs has covered trypsin protease cleavage activity prediction (Thomson and Yang, 2002; Thomson *et al*., 2003), HIV protease cleavage activity (Thomson *et al*., 2003), hepatitis C virus protease cleavage activity (Yang *et al*., 2003), factor

*To whom correspondence should be addressed.

Xa protease cleavage activity (Yang *et al*., 2003) and the O-linkage site prediction in glycoproteins (Chou, 1995; Yang and Chou, 2004).

This paper is devoted to further investigate the use of amino acid similarity measurement matrices in a more general pattern recognition algorithm, namely support vector machines (SVMs). SVMs is recently of increasing interest due to its promising empirical performance compared with other learning techniques (Vapnik, 1995, 1998; Scholkopf *et al*., 1997). Instead of using empirical risk minimization (ERM), which is commonly used in statistical learning, SVM is founded on structural risk minimization (SRM). ERM only minimizes the error occurred to training data whilst SRM minimizes an upper bound on the generalization error. This enables SVM to generalize well. The basic principle of SVM is to map the input space to a high-dimensional feature space using kernel techniques. A linear discriminant analysis is then formulated in the feature space to maximize the margin between two classes so as to maximize the generalization ability. Moreover, a discriminant analysis process is conducted based on a set of support vectors which are selected automatically from training data.

SVMs have been already used to deal with many biological problems, such as the analysis of microarray gene data (Brown *et al*., 2000), glycoprotein linakge site prediction (Cai *et al*., 2002a,b), predicting rRNA-, RNA-, and DNA-binding proteins (Cai and Lin, 2003), predicting protein subcellular location (Cai *et al*., 2000, 2003c; Chou and Cai, 2002), the prediction of protein domain structural class (Cai *et al*., 2002a,b, 2003a), the prediction of protein signal sequences and their cleavage sites (Cai *et al*., 2003b), DNA expression profiling (Rahman and Miles, 2001), and secondary structure prediction (Hua and Sun, 2001). Except for the use of chemical descriptors (Cai and Lin, 2003), all the others encoded amino acids using the distributed encoding method (Qian and Sejnowski, 1988).

The algorithm presented in this paper is referred to as bio-support vector machines (bSVM). The basic principle of bSVM is the replacement of kernel functions of SVMs with amino acid similarity matrices. We have applied bSVM to a specific problem, the prediction of HIV protease cleavage sites in proteins. The overall success rate is about 91%.

SYSTEM AND METHOD

Suppose we have N input patterns, $\mathbf{x}_i \in \Re^d$ be the i-th input pattern, where d is the number of the input variables, and y_i be the corresponding label of $\mathbf{x}_i$. Each label is either 1 or -1. A SVM classifier based on the support vectors found through learning is defined as

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b\right), \tag{1}$$

where, $K(\mathbf{x}, \mathbf{x}_i)$ is called a kernel function. In this study, $\mathbf{x}$ and $\mathbf{x}_i$ are protein sequences or oligopeptides, which contain non-numerical attributes, namely amino acids. We refer to $\mathbf{x}$ as a novel sequence while $\mathbf{x}_i$ as a support sequence (corresponding to a support vector). A major revision is then made to determine the kernel function using amino acid similarity matrices (Dayhoff *et al*., 1978; Johnson and Overington, 1993; Henikoff and Henikoff, 1993) for recognizing non-numerical attributes. The kernel function of the i-th sequence is defined as

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\left(\alpha \frac{s(\mathbf{x}, \mathbf{x}_i) - b_i}{b_i}\right), \tag{2}$$

where $s(\mathbf{x}, \mathbf{x}_i)$ is a pair-wise similarity measurement between $\mathbf{x}$ and $\mathbf{x}_i$, b_i is the maximum similarity measurement associated with the i-th support sequence and α is a constant. It can be seen that $[s(\mathbf{x}, \mathbf{x}_i) - b_i]/b_i$ is in general negative. In this study, we have revised SVM$^{\text{light}}$ (Joachims, 1999) for bio-SVM.

Most real applications involve how to minimize the probability of misclassification using the Bayes rule, by which the selection of the optimal threshold for making decisions depends on the prior knowledge $P(k)$ (Duda and Hart, 2002)

$$P(k|\mathbf{x}) = \frac{P(\mathbf{x}|k)P(k)}{P(\mathbf{x})}. \tag{3}$$

$P(\mathbf{x}|k)$ is independent from a trained model and dominates a decision process after a classifier has been built. In other words, different prior knowledge may lead to different decision outcomes. If the prior probability of class A is larger than that of class B, i.e. $P(A) > P(B)$, the sensitivity and specificity rates will be changed. A model with a small change is preferred as it shows robustness. The receiver operating characteristic (ROC) curve can be used to assess this robustness (Metz, 1978). In a ROC curve, the true positive fraction (TPf) is used as the vertical axis and the false positive fraction (FPf) the horizontal one. For a fixed FPf, a model with a higher TPf will be preferred. Therefore, the larger the area under the ROC curve, the better the performance a classifier has. TPf is $TPf = Tp/(Tp + Fn)$ and FPf $FPf = Fp/(Tn + Fp)$. Suppose the sensitivity is defined as (Olsson and Laurio, 2002)[†]

$$\text{Sensitivity} = \frac{Tp}{Tp + Fp}. \tag{4}$$

The relationship between TPf, FPf and sensitivity is then

$$\text{Sensitivity} = \frac{TPf}{TPf + FPf(Nn/Np)} \xrightarrow{Nn=Np} \frac{1}{1 + FPf/TPf}, \tag{5}$$

[†] Note that there are different definitions of sensitivity. The definition of Olsson and Laurio is shown in Equation (4). While the other definition of sensitivity is Sen $= Tp/(Tp + Fn)$ (Zhang *et al*., 2002). Our explanation of sensitivity is the likelihood that a prediction is true positive if the model's output is positive. We therefore adopt the definition of Olsson and Laurio in this study.

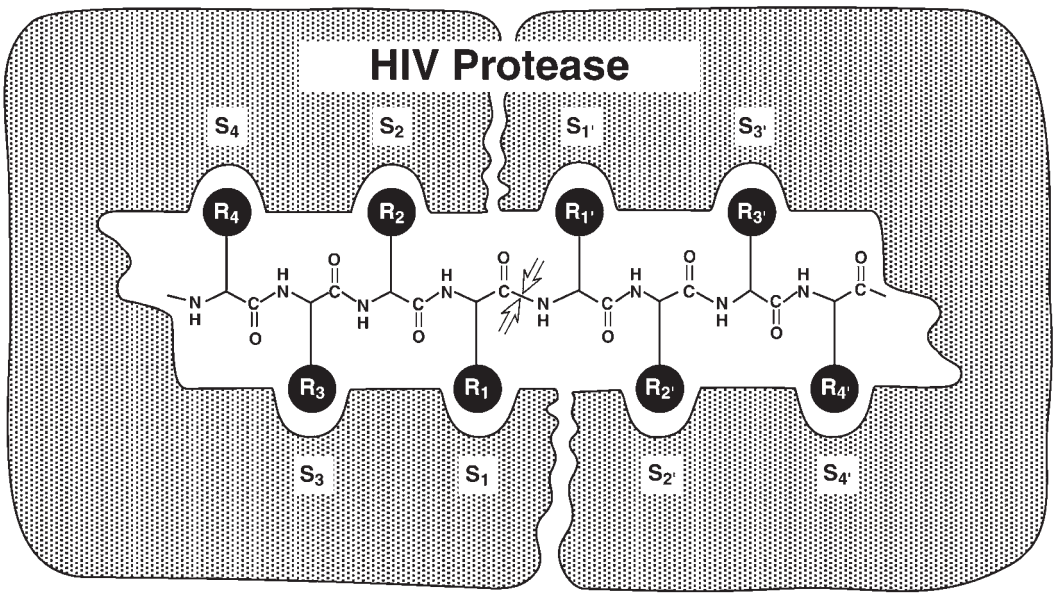

Fig. 1. Schematic representation of substrate bound to HIV protease based on an analysis of protease-inhibitor crystal structures. The active site of enzyme is composed of eight extended 'subsites', S_4, S_3, S_2, S_1, $S_{1'}$, $S_{2'}$, $S_{3'}$, $S_{4'}$, and their counterparts in a substrate extend to an octapeptide region, sequentially symbolized by R_4, R_3, R_2, R_1, $R_{1'}$, $R_{2'}$, $R_{3'}$, $R_{4'}$, respectively. The scissile bond is located between the subsites R_1 and $R_{1'}$. Reproduced with permission from Figure 3 of Chou,K.C. (*Anal. Biochem.*, 1996, **233**, 1–14).

where Nn and Np represent the numbers of negative and positive patterns, respectively. It can be seen that maximizing the area under a ROC curve implies maximizing the sensitivity value. This results from the fact that either maximizing TPf for a given FPf or minimizing FPf for a given TPf can maximize the area under a ROC curve.

IMPLEMENTATION

This method was encoded in java on a PC containing a 500 MHz Pentium and Linux operating system.

DISCUSSION

In this session, we discuss the application of the method described above for the prediction of HIV protease cleavage sites in proteins. Since the initial clinical reports in 1981, AIDS (acquired immunodeficiency syndrome) has become a synonym of terror to human beings. Threatened by such a severe disease, scientists in all areas are facing a significant challenge, i.e. how to provide useful knowledge and technology that will lead to effective method for designing drugs against AIDS. A key step in fighting against AIDS is how to effectively suppress HIV, the primary culprit of AIDS (Barre-Sinoussi *et al*., 1983; Gallo *et al*., 1984; Chou, 1996). Because a specific enzyme called HIV protease is indispensable for processing the viral gag and gag/pol polyproteins which takes place during the final maturation step of the viral life cycle (Kohl *et al*., 1988; Hellen *et al*., 1989; Navia *et al*., 1989; Wlodawer *et al*., 1989), blocking of HIV protease action by inhibitors (Ashorn *et al*., 1990; McQuade *et al*., 1990; Meek *et al*., 1990; Roberts *et al*., 1990; Chou, 1993c) or by mutagenesis (Kohl *et al*., 1988) results in production of immature, non-infectious viral particles so as to stop the replication of HIV. Discovering inhibitors of the HIV protease with antiviral activity has therefore been a critical issue over the last decade (Henderson *et al*., 1988; Hellen *et al*., 1989; Putney, 1992). In order to ensure effective inhibitors design against HIV protease, knowledge about the specificity or a successful prediction of what kind of peptides can be cleaved by HIV protease and what kind cannot is particularly useful, and this is the most important step.

HIV protease belongs to the family of the aspartyl proteases, which has been well-characterized as proteolytic enzymes. In HIV protease, the catalytic mechanism is composed of carboxyl groups from two aspartyl residues situated in both NH_2- and COOH-terminal halves of the enzyme molecule (Toh *et al*., 1985; Pearl and Taylor, 1987). Their property of strongly substrate-selective and cleavage-specific shows that they can cleave large, virus-specific polypeptides called polyproteins between a specific pair of amino acid (Hellen *et al*., 1989). It has been found that the cleavable sites in a given protein extend to an octapeptide region (Miller *et al*., 1989). The amino acid residues within this octapeptide region are denoted by eight subsites R_4, R_3, R_2, R_1, $R_{1'}$, $R_{2'}$, $R_{3'}$, $R_{4'}$ in order. The counterparts in the HIV protease are denoted by S_4, S_3, S_2, S_1, $S_{1'}$, $S_{2'}$, $S_{3'}$, $S_{4'}$ (Chou, 1993c). A diagram of HIV protease structure is shown in Figure 1. The cleave site is between R_1 and $R_{1'}$. The susceptible sites in some proteins

may contain one subsite less or one subsite more, corresponding to the case of a heptapetide or nonapeptide, respectively. However, they occur rarely due to the result of a balance between the following two factors. First, according to the 'rack mechanism' (Chou *et al*., 1981; Chou, 1988; Martel, 1992), the active site of HIV protease can be compared to a 'rack' during the peptide-cleaving process, meaning that the more residues that are bound to the rack of enzyme, the more stained the peptide and hence the more efficient the cleavage process. On the other hand, however, according to the dimension of the active site of an HIV protease, it can hardly accommodate more than eight residues. That is why most protease-susceptible sites in proteins are sequences of octapeptides.

There have been many methods for predicting HIV protease cleavage site in proteins, for instance, the *h* function (Poorman *et al*., 1991), correlation angle method (Chou, 1993a,b), vector sequence-coupling model (Chou, 1993c), descriminant function method (Chou *et al*., 1996), feed-forward neural networks with a back-propagation algorithm (Cai and Chou, 1998; Narayanan *et al*., 2002), binary probabilistic model (Yang, 2001), decision tree methods (Narayanan *et al*., 2002), bio-basis function neural networks (Thomson *et al*., 2003) and genetic programming method (Yang *et al*., 2003).

In this study, 362 HIV octapeptides were collected from (Poorman *et al*., 1991; Chou, 1996; Chou *et al*., 1996; Cai and Chou, 1998), of which 114 were positive (with cleavage sites) and 248 negative (without cleavage sites). Here, 300 were randomly selected for training bSVM models and the rest were used for testing. This process was repeated 10 times. The final prediction on the unseen testing (62) sequences was based on jackknife estimation. As is well known, the independent data set test, sub-sampling test and jackknife test are often used for cross-validation to examine the prediction quality. Among them the jackknife test is deemed as the most effective and objective one; see, e.g. Chou and Zhang (1995) for a comprehensive discussion about this, and Mardia *et al*. (1979) for the mathematical principle. Jackknife test is particularly useful for checking the cluster-tolerant capacity, and hence was often used for the case when the training data sets were far from complete yet [see, e.g. (Zhou, 1998; Chou, 1999; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003; Zhou and Troy, 2003)]. During jackknifing, each sample in the training data set is in turn singled out as a tested sample and all the rule-parameters are calculated based on the remaining samples.

Shown in Figure 2 were the prediction results when the '*C*' values were 1, 10, and 100 on the independent testing sequences, which were not involved in any process of modelling. The mean total accuracies were 91 ± 3.5, 91.5 ± 3.1, and $91.9 \pm 3.0\%$. When the '*C*' value was increased from 100, the performance was not changed, see Table 1. While a comparison among bSVM, decision tree method (C5 program), BPM (binary probabilistic model), and a feed-forward neural network with a back-propagation algorithm on the same data set was given in Table 2. It can be seen that bSVMs not only outperformed the others in prediction accuracy but also was superior to the others in model robustness. In decision tree method, there was no report of CPU time since the licence expired when we wanted the CPU time. There was also no report of the number of parameters since C5 did not involve any parameters in modelling.

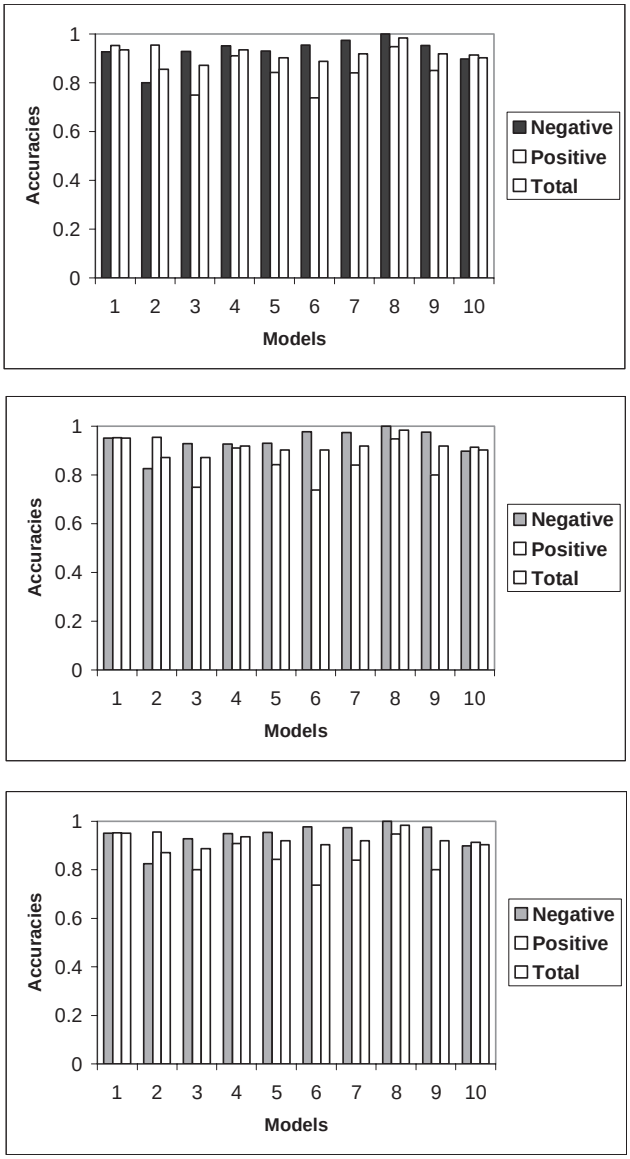

Fig. 2. The performance of bSVM for using different '*C*' values: (**a**) 1; (**b**) 10; (**c**) 100. There are three groups of measurements of the performance, true negative fraction (*TNf*), true positive fraction (*TPf*) and total accuracy. *TNf* is used to measure the prediction accuracy for non-cleaved HIV octapeptides. *TPf* is used to measure the prediction accuracy for cleaved HIV octapeptides.

Shown in Figure 3 were the ROC curves of the models when the '*C*' value was 100. It can be seen that the models

Table 1. Performance measured by jackknife on 10 models

	TNf (%)	*TPf* (%)	Sensitivity (%)	Total (%)
$C = 10^0$	91.1 ± 5.1	87.0 ± 7.6	87.6 ± 6.5	91.1 ± 3.2
$C = 10^1$	93.8 ± 4.8	86.5 ± 7.9	88.7 ± 7.0	91.5 ± 3.2
$C = 10^2$	94.3 ± 4.8	87.0 ± 7.2	89.7 ± 6.8	91.9 ± 3.1
$C = 10^3$	94.3 ± 4.8	87.0 ± 7.2	89.7 ± 6.8	91.9 ± 3.1
$C = 10^4$	94.3 ± 4.8	87.0 ± 7.2	89.7 ± 6.8	91.9 ± 3.1
$C = 10^5$	94.3 ± 4.8	87.0 ± 7.2	89.7 ± 6.8	91.9 ± 3.1

Table 2. Comparison among decision tree method (C5), BPNN, BPM and bSVMs

	C5	BPNN	BPM	bSVMs
Mean accuracy	85.5%	90.0%	88.6%	91.2%
Standard deviation	3.3%	8.2%	6.3%	3.0%
Parameters	—	12 880	1600	185
CPU (10 models)	—	5 h	78 min	19 min

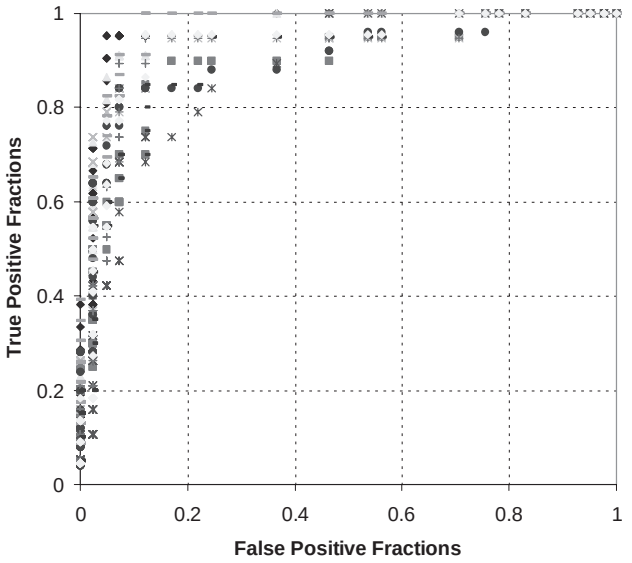

Fig. 3. ROC curves of bSVM when the '*C*' value was 100. The horizontal axis indicates the false positive fraction and the vertical axis means the true positive fraction. Each point in the graph means a model with a specifically selected threshold for discrimination between two classes, non-cleaved octapeptides and cleaved HIV octapeptides. This selection of the threshold is determined by a prior knowledge, which weights a preferred class. If the prediction accuracy of cleaved HIV octapeptides is more important, the cost function related with it will be larger than that of non-cleaved HIV octapeptides.

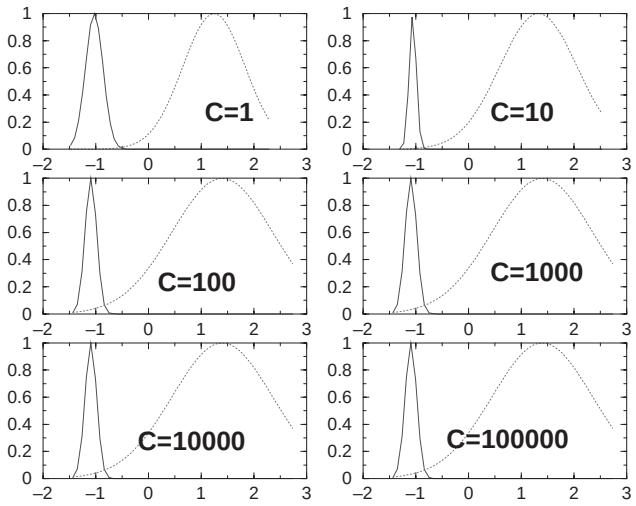

Fig. 4. Estimated probability density functions of the model outputs. The discrimination power is commonly determined by the separability of the probability density functions of two classes. If they are far separated, the discrimination power will be large. In this graph, we can see that two probability density functions are far away.

were robust since the ROC curves were away from the diagonal line. The estimated probability density functions of the two classes (non-cleaved HIV octapeptides and cleaved HIV octapeptides) were shown in Figure 4. It can be seen that the two classes were well separated and the discrimination power of the models for these two classes is large.

This paper has presented a new pattern recognition method, called the 'bio-support vector machine', for analysing protein sequences. The major principle is to replace kernel functions of SVMs with amino acid similarity measurement matrices. The method has been successfully applied to the prediction of HIV protease cleavage sites in proteins. The mean accuracy is from 91 ± 3.5 to $91.9 \pm 3.0\%$. The prediction accuracy on the unseen independent data set is higher than that from BPNN. The standard deviation of the mean prediction accuracy of bSVM is smaller than that of BPNN. In this study, we only presented the result based on the Dayhoff matrix (Dayhoff *et al*., 1978; Johnson and Overington, 1993). Further work is to use more matrices such as Blosum matrices (Henikoff and Henikoff, 1993).

ACKNOWLEDGEMENTS

The authors would like to thank Dr T. Joachims for permitting us to use and revise the SVM^{light} package.

REFERENCES

Arrigo,P. Giuliano,F., Scalia,F., Rapallo,A. and Damiani,G. (1991) Identification of a new motif on nucleic acid sequence data using Kohonen's self-organising map. *Comput. Appl. Biosci.*, **7**, 353–357.

Ashorn,P., McQuade,T.J., Thaisrivongs,S., Tomasselli,A.G., Tarpley,W.G. and Moss,B. (1990) An inhibitor of the protease blocks maturation of human and simian immunodeficiency viruses and spread of infection. *Proc. Natl Acad. Sci. USA*, **87**, 7472–7476.

Baldi,P., Pollastri,G., Andersen,C.A. and Brunak,S. (2000) Matching protein beta-sheet partners by feedforward and recurrent neural networks. *Proceedings of International Conference on Intelligent Systems for Molecular Biology, ISMB*, Vol. 8, pp. 25–36.

Barre-Sinoussi,F., Chermann,J.C., Rey,F., Nugeyre,M.T., Chamaret,S., Gruest,J., Dauguet.C., Axler-Blin,C., Vezinet-Brun,F., Rouzioux,C., Rozenbaum,W. and Montagnier,L. (1983) Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, **220**, 868–871.

Brown,M.P.S., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W. Furey,T.S., Jr. Ares,M. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *PNAS*, **97**, 262–267.

Cai,Y.D. and Chou,K.C. (1998) Artificial neural network model for predicting HIV protease cleavage sites in protein. *Adv. Engng Software*, **29**, 119–128.

Cai,Y.D., Liu,X.J., Xu,X.B. and Chou,K.C. (2000) Support vector machines for prediction of subcellular location. *Mol. Cell Biol. Res. Commun.*, **4**, 230–233.

Cai,Y.D., Liu,X.J., Xu,X.B. and Chou,K.C. (2002a) Support vector machines for predicting the specificity of GalNAc-transferase, *Peptides*, **23**, 205–208.

Cai,Y.D., Liu,X.J., Xu,X.B. and Chou,K.C. (2003a) Support vector machines for prediction of protein domain structural class. *J. Theor. Biol.*, **221**, 115–120.

Cai,Y.D., Liu,X.J., Xu,X.B. and Chou,K.C. (2002b) Prediction of protein structural classes by support vector machines. *Comput. Chem.*, **26**, 293–296.

Cay,Y.D. and Lin,S.L. (2003) Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochim. Biophys. Acta*, **1648**, 127–133.

Cay,Y.D., Lin,S.L. and Chou,K.C. (2003b) Support vector machines for prediction of protein signal sequences and their cleavage sites. *Peptides*, **24**, 159–161.

Cai,Y.D., Zhou,G.P. and Chou,K.C. (2003c) Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys. J.*, **84**, 3257–3263.

Chou,J.J. (1993a) A formulation for correlating properties of peptides and its application to predicting human immunodeficiency virus protease-cleavable sites in proteins. *Biopolymers*, **33**, 1405–1414.

Chou,J.J. (1993b) Predicting cleavability of peptide sequences by HIV protease via correlation-angle approach. *J. Protein Chem.*, **12**, 291–302.

Chou,K.C., Chen,N.Y. and Forsen,S. (1981) *Chem. Scr.*, **18**, 126–132.

Chou,K.C. (1988) Review: low-frequency collective motion in biomacromolecules and its biological functions. *Biophys. Chem.*, **30**, 3–48.

Chou,K.C. (1993c) A vectorised sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J. Biol. Chem.*, **268**, 16938–16948.

Chou,K.C. (1995) A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase. *Protein Sci.*, **4**, 1365–1383.

Chou,K.C. (1996) Prediction of human immunodeficiency virus protease cleavage sites in proteins. *Anal. Biochem.*, **233**, 1–14.

Chou,K.C., Tomasselli,A.L., Reardon,I.M. and Heinrikson,R.L. (1996) Predicting HIV protease cleavage sites in proteins by a discriminant function method. *Proteins: Structure, Function, and Genetics*, **24**, 51–72.

Chou,K.C. (1999) A key driving force in determination of protein structural classes. *Biochem. Biophys. Res. Commun.*, **264**, 216–224.

Chou,K.C. and Zhang,C.T. (1995) Review: prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.*, **30**, 275–349.

Chou,K.C. and Cai,Y.D. (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.*, **277**, 45765–45769.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. Matrices for detecting distant relationships. In Dayhoff,M.O. (ed.) *Atlas of Protein Sequence and Structure*, Vol. 5, National Biomedical Research Foundation Washington, DC, pp. 345–358.

Duda,R.O. and Hart,P.E. (2002) *Pattern Classification and Scene Analysis*. Wiley, New York.

Gallo,R.C., Salahuddin,S.Z., Popovic,M., Shearer,G.M., Kaplan,M., Haynes,B.F., Palker,T.J., Redfield,R., Oleske,J., Safai,B. *et al*. (1984) Frequent detection and isolation of cytopathic retroviruses (HTLV III) from patients with AIDS and at risk for AIDS. *Science*, **224**, 500–503.

Hellen,C.U.T., Krausslich,H.G. and Wimmer,E. (1989) Proteolytic processing of polyproteins in the replication of RNA viruses. *Biochemistry*, **28**, 9881–9890.

Henderson,L.E., Benveniste,R.E., Sowder,R., Copeland,T.D., Schultz,A.M. and Oroszlan,S. (1988) Molecular characterization of gag proteins from simian immunodeficiency virus (SIVMne). *J. Virol.*, **62**, 2587–2595.

Henikoff,S. and Henikoff,J.G. (1993) Performance evaluation of amino acid substitution matrices. *Proteins*, **17**, 49–61.

Hua,S. and Sun,Z. (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J. Mol. Biol.*, **302**, 397–407.

Joachims,T. (1999) Making large-scale SVM learning practical. In Scholkopf,B. and Burges,C. (eds) *Advances in Kernel Methods—Support Vector Learning*. MIT Press, Cambridge, MA.

Johnson,M.S. and Overington,J.P. (1993) A structural basis for sequence comparisons—an evaluation of scoring methodologies. *J. Mol. Biol.*, **233**, 716–738.

Kohl,N.E., Emini,E.A. and Sigal,I.S. (1988) Active human immunodeficiency virus protease is required for viral infectivity. *Proc. Natl Acad. Sci. USA*, **85**, 4686–4690.

Mardia,K.V., Kent,J.T. and Bibby,J.M. (1979) *Multivariate Analysis*. Academic Press, London, pp. 322 and 381.

Martel,P. (1992) *Prog. Biophys. Mol. Biol.*, **57**, 129–179.

McQuade,T.J., Tomasselli,A.G. and Liu,L. (1990) A synthetic HIV-1 protease inhibitor with antiviral activity arrests HIV-like particle maturation. *Science*, **247**, 454–456.

Meek,T.D., Lambert,D.M. and Dreyer,G.B. (1990) Inhibition of HIV-1 protease in infected *t*-lymphocytes by synthetic peptide analogues. *Nature (London)*, **343**, 90–92.

Metz,C.E. (1978) Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, **8**, 283–298.

Miller,M., Schneider,J., Sathayanarayana,B.K., Toth,M.V., Marshall,G.R., Clawson,L., Selk,L., Kent,S.B.H. and Wlodawer,A. (1989) Structure of complex of synthetic HIV-1 protease with substrate-based inhibitor at 2.3 Å resolution. *Science*, **246**, 1149–1152.

Narayanan,A., Wu,X.K. and Yang,Z.R. (2002) Mining viral protease data to extract cleavage knowledge. *Bioinformatics*, **18**, 5–13.

Navia,M.A., Fitzgerald,P.M.D., McKeever,B.M., Leu,C.T., Heimbach,J.C., Herber,W.K., Sigal,I,S., Drake,P.L. and Springer,J.P. (1989) Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. *Nature*, **337**, 615–620.

Olsson,B. and Laurio,K. (2002) Towards a comprehensive collection of diagnostic patterns for protein sequence classification. *Inform. Sci.*, **143**, 1–11.

Pearl,L.H. and Taylor,W.R. (1987) A structural model for the retroviral proteases. *Nature*, **329**, 351–354.

Pollastri,G., Przybylski,D., Rost,B. and Baldi,P. (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, **47**, 228–235.

Poorman,R.A., Tomasselli,A.G., Heinrikson,R.L. and Kezdy,F.J. (1991) A cumulative specificity model for protease from human immunodeficiency virus types 1 and 2, inferred from statistical analysis of an extended substrate data base. *J. Biol. Chem.*, **22**, 14554–14561.

Putney,S. (1992) How antibodies block HIV infection: paths to an AIDS vaccine. *Trends Biochem. Sci.*, **17**, 91–196.

Qian,N. and Sejnowski,T.J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, **202**, 865–884.

Rahman,S. and Miles,F. (2001) Identification of novel ethanol-sensitive genes by expression profiling. *Pharmacology & Therapeutics*, **92**, 123–134.

Reczko,M. (1993) Protein secondary structure prediction with partially recurrent neural networks. *SAR QSAR Environ. Res.*, **1**, 153–159.

Roberts,N.A., Martin,J.A., Kinchington,D., Broadhurst,A.V., Craig,J.C., Duncan,I.B., Galpin,S.A., Handa,B.K., Kay,J., Krohn,A. *et al*. (1990) Rational design of peptide-based HIV proteinase inhibitors. *Science*, **248**, 358–361.

Scholkopf,B., Sung,K.K., Burges,C.J.C., Girosi,F., Niyogi,P., Poggio,T. and Vapnik,V. (1997) Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Trans. Signal Process.*, **45**, 2758–2765.

Thomson,R. and Yang,Z.R. (2002) A novel bio-basis function neural network. *Proceedings of 9th International of Neural Information Processing*, Singapore.

Thomson,R., Hodgman,T.C., Yang,Z.R. and Doyle,A.K. (2003) Characterising proteolytic cleavage site activity using bio-basis function neural networks. *Bioinformatics*, **19**, 1741–1747.

Toh,H., Ono,M., Saigo,K. and Miyata,T. (1985) Retroviral protease-like sequence in the yeast transposon Ty1. *Nature*, **315**, 691.

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

Vapnik,V. (1998) The support vector method of function estimation. In Suykens,J.A.K. and Vandewalle,J. (eds) *Nonlinear Modeling: Advanced Black-Box Techniques*. Kluwer, Boston, MA, pp. 55–85.

Wlodawer,A., Miller,M., Jaskolski,M., Sathyarayana,B.K., Baldwin,E., Weber,I.T., Selk,L., Clawson,L., Schneider,J. and Kent,S.B.H. (1989) Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. *Science*, **245**, 616–621.

Yang,Z.R. (2001) A binary probabilistic model and genetic algorithm for HIV protease cleavage sites prediction and search. *The 8th International Conference on Neural Information Processing*. Shanghai, China, pp. 847–852.

Yang,Z.R., Thomson,R., Hodgman,T.C., Dry,J., Doyle,K., Narayanan,A. and Wu,X.K. (2003) Genetic programming method for proteolytic cleavage site prediction. *BioSystems*, (in press).

Yang,Z.R. and Chou,K.C. (2004) Predicting the linkage sites in glycoproteins using bio-basis function neural network. *Bioinformatics* **20**, in press.

Zhang,C.T., Wang,J. and Zhang,R. (2002) Using a Euclid distance discriminant method to find protein coding genes in the yeast genome. *Comput. Chem.*, **26**, 195–206.

Zhou,G.P. (1998) An intriguing controversy over protein structural class prediction. *J. Protein Chem.*, **17**, 729–738.

Zhou,G.P. and Assa-Munt,N. (2001) Some insights into protein structural class prediction. *Proteins: Structure, Function, and Genetics*, **44**, 57–59.

Zhou,G.P. and Doctor,K. (2003) Subcellular location prediction of apoptosis proteins. *Proteins: Structure, Function, and Genetics*, **50**, 44–48.

Zhou,G.P. and Troy,F.A. (2003) Characterization by NMR and molecular modeling of the binding of polyisoprenols and polyisoprenyl recognition sequence peptides: 3D structure of the complexes reveals site of specific interactions. *Glycobiology*, **13**, 51–71.