Mining Biological Data Using Self-Organizing Map

Zheng Rong Yang*,† and Kuo-Chen Chou‡,§

Department of Computer Science, Exeter University, Exeter EX4 4PT, U.K., Gordon Life Science Institute, San Diego, California 92130, and Tianjin Institute of Bioinformatics and Drug Discovery (TIBDD), Tianjin, China

Received July 8, 2003

This paper presents a novel method of mining biological data using a self-organizing map (SOM). After partitioning a set of protein sequences using SOM, conventional homology alignment is applied to each cluster to determine the conserved local motif (biological pattern) for the cluster. These local motifs are then regarded as rules for prediction and classification. In the application to the prediction of HIV protease cleavage sites in proteins, we found that the rules derived from this method are much more robust than those derived from the decision tree method.

INTRODUCTION

Data mining has been of increasing interest to biology scientists. The main topic in using data mining techniques for analyzing biological data is to extract hidden knowledge for the purpose of interpolation or extrapolation, i.e., for annotating biological functions of novel genes, proteins, and compounds.

One of the tasks in mining biological data is to find a mapping function from input data (commonly, nucleic or protein sequences) to their structure/function class. Most supervised pattern recognition algorithms including artificial neural networks (ANNs) have therefore been used or modified for this purpose. The supervised pattern recognition algorithms used in analyzing protein sequence data are decision tree method, back-propagation neural networks (BPNNs), recurrent neural networks (RNNs), support vector machines (SVMs), and biobasis function neural networks (BBFNN).[1]

The application of BPNNs to human immunodeficiency virus (HIV) protease cleavage sites prediction yielded a prediction accuracy of approximately 90%.[2,3] BPNNs and RNNs have been used for secondary structure prediction.[4] SVMs are recently of increasing interest due to a promising empirical performance compared with other pattern recognition algorithms.[5−7] SVMs have been used to deal with many biological problems, such as the analysis of microarray gene data,[8] glycoprotein linakge site prediction,[9,10] predicting rRNA-, RNA-, and DNA-binding proteins,[11] protein subcellular location prediction,[12−14] the prediction of protein domain structural class,[15] the prediction of protein signal sequences and their cleavage sites,[16] DNA expression profiling,[17] and secondary structure prediction.[18] The decision tree method has been used for mining HIV and hepatitis C virus data.[3] In the application, C5 is used to extract some decision rules, which are later used for decision making. The development of BBFNN using an amino acid similarity matrix has improved the time complexity and robustness of pattern recognition algorithms in analyzing biological sequences. BBFNN has been applied to a variety of proteolytic cleavage activity predictions and others. For instance, trypsin protease cleavage activity prediction,[1] HIV protease cleavage activity,[1] hepatitis C virus protease cleavage activity,[19] factor Xa protease cleavage activity,[19] and the prediction of *O*-linkage sites in glycoproteins.[20]

Except for the use of the supervised pattern recognition algorithms, unsupervised pattern recognition algorithms have also been paid attention. The most popular unsupervised pattern recognition algorithms used in bioinformatics are various clustering algorithms, such as hierarchical clustering algorithms, *K*-means algorithms, and model-based clustering algorithms. The use of these algorithms mainly aims to partition biological data, for instance, hierarchical cluster analysis of gene expression data,[21−26] model-based cluster analysis of gene data,[27] and SVD method for gene cluster analysis.[28]

Except for these clustering algorithms, self-organizing map (SOM) are a powerful alternative for mining biological data. The basic principle of SOM is to map multidimensional vectors to a two-dimensional grid space called an output map, in which similar vectors would be mapped onto the same or close grid in the output map. Such mapping is based on the inherent topological structure hidden in the vectors. This means that SOM can work without using the target information and hence is an unsupervised learning algorithm. SOM has been used to identify motifs and families in the context of unsupervised learning.[29−32] For instance, SOM was used to cluster 444 protein sequences.[33] These 444 protein sequences were classified into 13 families by using statistical and hybrid methods prior to that simulation in SwissProt database (release 19.0, 8/91). The study showed that the hidden biological information contained in sequence protein databases could be well-organized using SOM. A further work based on 1758 human protein sequences stored in SwissProt database (release 19.0) showed this capability of SOM.[31] In the study, 1758 human protein sequences were clustered in a dendrogram tree to demonstrate the hierarchical

* Corresponding author phone: 01392−264045; fax: 01392−264047. e-mail: Z.R.Yang@exeter.ac.uk.

† Exeter University.

‡ Gordon Life Science Institute.

§ TIBDD.

10.1021/ci034138n CCC: $25.00
Published on Web 11/07/2003

relationship among the protein families. An interesting report came out using SOM to identify a motif on nucleic acid sequence data,[29] in which the SOM net was modified to a one-dimensional output map. A single nucleic acid sequence was input into the net, and the motif was finally determined. In a sequence of HSIRPR cDNA release (EMBL data bank) with a 5180 basis, the identified motif was *GCGGG*. SOM has also been used for partitioning gene data.[34] However, these studies did not show how the methods could be generalized to unseen data. It should be noted that if knowledge extracted using an unsupervised learning algorithm can be verified on unseen data, the credibility of the extracted knowledge could be further enhanced.

This study is aimed to mine protein oligopeptides using SOM and verify the extracted knowledge for the purpose of prediction and classification on unseen data. SOM is first used to partition protein oligopeptides. Local motifs are then found for these clustered protein oligopeptides using conventional homology alignment techniques. The extracted motifs are regarded as hidden knowledge which governs the distribution of protein oligopeptides and the formulation of the amino acids in protein oligopeptides for different biological functions. To verify the extracted knowledge, two supervised pattern recognition processes are involved. First, a number of in-house testing oligopeptides are used for simulating out-of-house prediction and classification. It should be noted that these in-house oligopeptides are never used for knowledge extraction. Second, a rule must be more tolerant to noise, which can be caused by possible mutation of some residue in oligopeptides. A rule, which is less sensitive to possible mutation of some residue in oligopeptides, will be considered as a robust rule. In this study, we have applied our method to the prediction of HIV protease cleavage sites in proteins. The comparison between the SOM and the C5 rules shows that the SOM rules are more robust than C5 rules.

METHOD

Decision Tree Method. The basic principle of decision trees is to *divide and conquer*. The decision tree method has been used for mining HIV and hepatitis C virus data.[3] In the application, C5 is used to extract some decision rules, which are later used for decision making. However, most decision tree algorithms, such as ID3, C4.5, and C5, are based on frequency estimation without taking into account the interaction among the amino acids in protein oligopeptides. This may not be able to determine the true probability density function of protein oligopeptides, and hence the prediction accuracy may be reduced. In fact, it has been found that the interaction among amino acids in HIV octapeptide exists.[35,36]

Self-Organizing Map. SOM[37] is a sheetlike network structure, in which a squared array of output units is commonly adopted. The array of the output units (grids) is called an output or feature map. After completing training of a SOM net, the input vectors used for training will be mapped onto different output units. If two input vectors are mapped onto the same output unit, it means that these two input vectors have similar characteristics. In this study, it is more likely that their protein oligopeptides share the same biological properties. A simple way to formulate clusters is to regard an output unit with at least two input vectors mapped onto it as a cluster. More complicated methods can determine clusters through clustering the weight vectors in SOM if the net size is large.

The algorithm is as follows

step 1	to encode each oligopeptide using the distributed encoding method[38] (this means that each amino acid will be coded using a 20-bit binary vector)
step 2	to organize a SOM net with desired output units
step 3	to divide data into two parts, one for training and the other for in-house testing
step 4	to train SOM using the training data set of input vectors encoded from oligopeptides
step 5	to select output units, each of which has at least two input vectors corresponding to "positive" oligopeptides mapped on, for searching for local motifs (note that positive means cleaved oligopeptides in the latter application to the prediction of HIV protease sites in proteins)
step 6	to search for a local motif for each cluster using a homology alignment package available in the Internet
step 7	to use the extracted local motifs as decision rules for prediction and classification
step 8	to analyze the sensitivity or robustness of the rules

APPLICATION TO THE PREDICTION OF HIV PROTEASE CLEAVAGE SITES IN PROTEINS

Since the clinical report in 1981, AIDS (acquired immunodeficiency syndrome) has been identified as a synonym of terror to human beings. In addressing the threat, scientists with different disciplines are facing a significant challenge about designing effective drugs against AIDS. It is understood that effectively suppressing HIV, the primary culprit of AIDS, is the key to fighting against AIDS.[36,39,40] It is also understood that HIV protease as a specific enzyme is necessary for processing the viral gag and gag/pol polyproteins which occur in the maturation stage of the viral life cycle.[41−44] Blocking HIV protease action using inhibitors[35,45−48] or mutagenesis[42] can lead to the production of immature, noninfectious viral particles. Discovering those inhibitors has therefore been an area of focus over the past decade.[41,49,50] To design effective inhibitors of HIV protease, knowledge about the specificity or a successful prediction of HIV protease cleavage capability is critical.[36]

HIV protease is a member 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 N- and C-terminal halves of the enzyme molecule.[51,52] They can cleave large, virus-specific polypeptides called polyproteins between a specific pair of amino acids.[41] It is known that the cleavable sites in a given protein extends to an octapeptide region.[35−36,53,55,56] The amino acid residues are denoted by eight subsites R_4, R_3, R_2, R_1, $R_{1'}$, $R_{2'}$, $R_{3'}$, and $R_{4'}$, and their counterparts in the HIV protease are S_4, S_3, S_2, S_1, $S_{1'}$, $S_{2'}$, $S_{3'}$, and $S_{4'}$. Shown in Figure 1 is such a diagram.[36] The cleave site is between R_1 and $R_{1'}$. The susceptible sites in some proteins may contain more than one subsite. However, they occur rarely due to the result of a balance between the following two factors. First, according to the "rack mechanism",[54] the active site of HIV protease has a "rack" during the peptide-cleaving process. This means that the more residues that are bound to the rack of the enzyme, the more stained the peptide and hence the more efficient the cleavage process.[35,36] Besides, according to the dimension of the active site of an HIV protease, it can hardly accommodate more than eight residues.

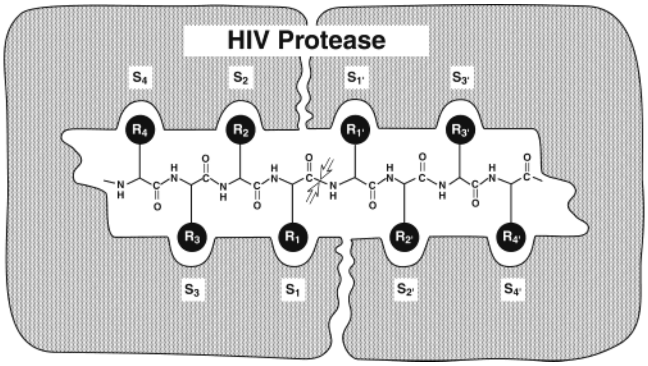


Figure 1. Schematic representation of substrate bound to HIV protease based on an analysis of protease-inhibitor crystal structures. The active site of the enzyme is composed of eight extended "subsites", S_4, S_3, S_2, S_1, $S_{1'}$, $S_{2'}$, $S_{3'}$, and $S_{4'}$, and their counterparts in a substrate extend to an octapeptide region, sequentially symbolized by R_4, R_3, R_2, R_1, $R_{1'}$, $R_{2'}$, $R_{3'}$, and $R_{4'}$, respectively. The scissile bond is located between the subsites R_1 and $R_{1'}$. Reproduced with permission from ref 36. Copyright 1996 Academic Press.

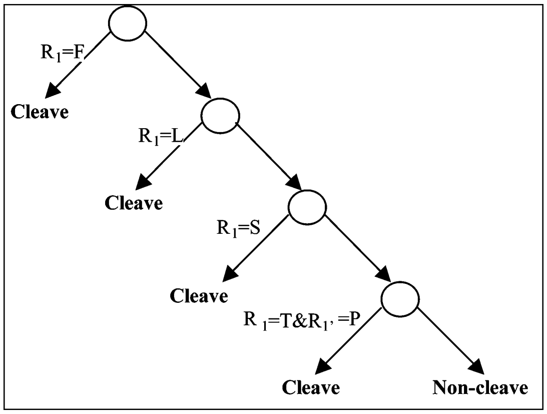
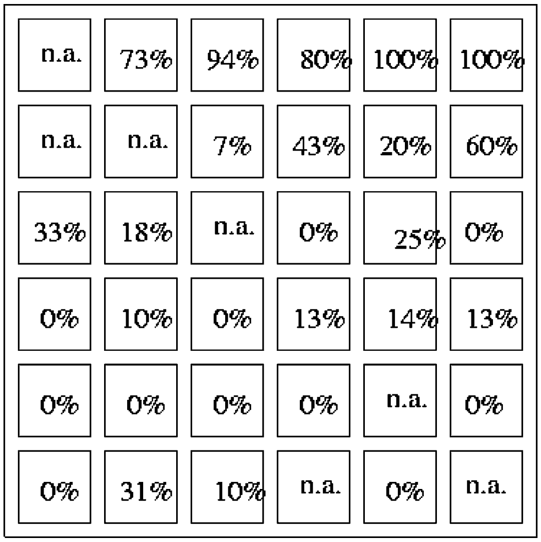


Figure 2. Decision tree obtained by C5 method.



n.a.	73%	94%	80%	100%	100%
n.a.	n.a.	7%	43%	20%	60%
33%	18%	n.a.	0%	25%	0%
0%	10%	0%	13%	14%	13%
0%	0%	0%	0%	n.a.	0%
0%	31%	10%	n.a.	0%	n.a.

Figure 3. Frequency map after training.

Table 1. Local Conserved Motifs Found

motifs ($R_4R_3R_2R_1R_{1'}R_{1'}R_{2'}R_{3'}R_{4'}$)	cleavage prob (%)	homology %	coverage %
SQNYPXVQ	73	78	21
SXNFPXXX	94	64	15
AXTFYXXX	80	23	7
ARXLFXAL	100	59	14
XXXLFEXX	100	12	13

The data of 362 HIV octapeptides was used in a series of the earlier studies.[2,35,36,55−58] Among 362 octapeptides, 114 are "positive" (with cleavage sites) and the rest "negative" (without cleavage sites).

Shown in Figure 2 is the result of the C5 model, where phenylalanine, leucine, serine, tyrosine, and proline show critical importance to the cleavage capability when they are located at R_1 and $R_{1'}$. Interestingly, the importance of serine and leucine close to the cleavage point is shared with the Rous sarcoma virus (RSC), which also has an octapeptide recognition substrate. However, cleavage needs both serine and leucine immediately on both sides of the cleavage site.[3]

In using SOM, each HIV octapeptide was encoded into a 160-bit binary vector referred to as an input vector using the distributed encoding method. These vectors were fed to SOM with 6 × 6 output units. If some HIV octapeptides are evolved from the same local motif, their corresponding input vectors should be close to each other in input space. They therefore should be mapped onto the same output unit after training. Correspondingly the HIV octapeptides can be partitioned. The basic assumption is that the hidden motif, which we do not know in advance, governs the data distribution. It is then believed that the cluster structure should be the best representation of the motif structure; i.e., each cluster should have a motif to represent all the members (HIV octapeptides) in the cluster. Homology alignment method was therefore applied to search for these hidden local motifs. We used the GeneBee (www.genebee.msu.su) for homology alignment for each cluster. We were in fact particularly interested in searching for motifs for cleaved HIV octapeptides since this knowledge would be very useful for drug design. Homology alignment was therefore applied to the clusters which mainly contained cleaved HIV octapeptides.

Before determining where to apply homology alignment, we first defined a concept called cleavage probability, which was actually the frequency that the cleaved HIV octapeptides were mapped to a corresponding output unit. A frequency of 100% therefore means that all the HIV octapeptides mapped onto the unit are positive. A frequency of 0% means that all the HIV octapeptides mapped onto the unit are negative. If the frequency of an output unit is large, the confidence that the corresponding cluster contains a useful motif for cleaved HIV octapeptides should be high. We selected five clusters whose cleavage probabilities were larger than 70%. For each of these five clusters, homology alignment was therefore applied. Shown in Figure 3 is a frequency map, in which the cleavage frequency for each output unit is marked. Note that "n.a." means that no input vector was mapped onto the corresponding output unit.

After applying homology alignment for each cluster, we determined a local motif, which had the highly conserved amino acids. Shown in Table 1 are the local motifs of the five clusters, where "X" means any amino acids. In Table 1, cleavage probability ("cleavage prob") is defined above, "homology %" is given by the GeneBee (www.genebee.msu.su) measuring the probability of multiple homology alignment, "coverage %" means the coverage percentage of the cleaved HIV octapapetides clustered into the cluster. For instance, the first motif (SQNYPXVQ) covers 21% of

Table 2. SOM Rules

rule no.	rules
1	R_1 = tyrosine and $R_{1'}$ = proline
2	R_1 ≃ phenylalnine and $R_{1'}$ = proline
3	R_1 = phenylalnine and $R_{1'}$ = tyrosine
4	R_1 = leucine and $R_{1'}$ = phenylalnine

cleaved HIV octapeptides. From the table, the following biological patterns are extracted:

(a) Tyrosine (Y) and proline (P) are more conserved at R_1 and $R_{1'}$ for cleaved HIV octapeptides. It is observed early that a Tyr-Pro bond is at the cleavage site among the matrix protein and the capsid protein.[59] As indicated in ref 60, replacement of the rapidly cleaved methionine−methionine bond with tyrosine−proline or replacement of the tyrosine−proline bond with methionine−methionine results in sites that cannot be efficiently cleaved.

(b) Phenylalanine (F) and proline (P) are more conserved at R_1 and $R_{1'}$ for cleaved HIV octapeptides. The elementary reaction properties of the Phe-Pro structure have been studied.[61] It is also shown that the scissile bond (Phe-Pro) within the gag-pol polyprotein is a competitive inhibitor of HIV-1 protease.[62] The importance of the Phe-Pro bond in inhibitor design has also been the focus of researchers.[63−65]

(c) Phenylalanine (F) and tyrosine (Y) are more conserved at R_1 and $R_{1'}$ for cleaved HIV octapeptides. The specificity of the Phe-Tyr bond has been intensively studied because of its importance for inhibitor design.[66−69]

(d) Leucine (L) and phenylalanine (F) are more conserved at R_1 and $R_{1'}$ for cleaved HIV octapeptides. It had also been observed early that leucine is critical to cleavage sites.[59] The other studies also addresses the importance of the Leu-Phe bond in inhibitor design.[70−72]

In conclusion, phenylalanine, tyroline, proline, and leucine are more critical to HIV protease cleavage activity. The result is consistent with previous studies, either biological experiments or machine learning.[3] It can be seen that a pattern with a small amount of conserved amino acids is more specific for pattern recognition. This means that it is less affected by noise. For instance, the pattern “XXXLFEXX” has three conserved amino acids and gains 100% of the cleavage probability, while the pattern “SQNYPXVQ” has seven conserved amino acids and gains only 73% of the cleavage probability. It is also interesting to note that the relationship between cleavage probability and homology percent is reversed. In total, 70% of cleaved HIV octapeptides are covered by these five motifs. This means that the use of these five motifs can ensure at least 70% accuracy in identifying cleaved HIV octapeptides.

Since the local motif represented an invariant pattern of amino acids, the local motifs are regarded as the rules for decision making. The information provided by the amino acids at the subsites besides the cleavage sites (R_1 and $R_{1'}$) is actually critically important; we are interested in using the motifs formulated at these two subsites, and hence the 4th and 5th motifs are merged as one SOM rule. Shown in Table 2 are four SOM rules.

For short, they are denoted as Y-P, F-P, F-Y, and L-F. In correspondence, we also consider four major C5 rules, F-X, L-X, S-X, and Y-P, where X means no conserved amino acid at the specific subsite.[2] Note that these rules are derived based on 80% of randomly selected HIV octapeptides. The remaining 20% of octapeptides are then used for cross-validation.[14,73,74] The classification accuracies of the SOM rules and the C5 rules on the in-house testing are 84 and 80%, respectively. It is not surprising that some C5 rules have a high identification rate for the cleaved HIV octapeptides since they only use one-side information besides the cleavage site, whose consequence is that the misclassification rate (false positive rate) of the noncleaved HIV octapeptides is high. The false positive rate of C5 rules is 8.4 times larger than that of SOM rules.

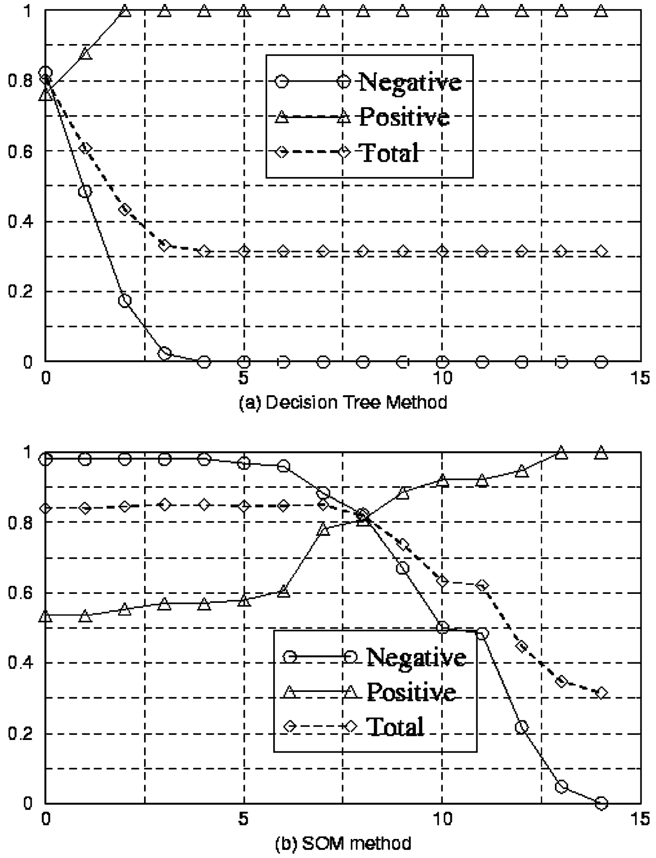


Figure 4. Comparison of sensitivity of the rules to evolutionary mutation.

The other important issue associated with the decision rules is whether they are sensitive to evolutionary mutation particularly at the subsites R_1 and $R_{1'}$. A rule is regarded as a robust rule if it is not sensitive to possible evolutionary mutation (regarded as possible noise). To assess their sensitivities, artificially evolutionary mutation was conducted. The amino acids at the subsites were mutated artificially using the Dayhoff similarity matrix.[75] We relaxed the amino acid similarity one by one to select the most probable and random amino acid to replace the original amino acid besides the cleavage sites of the octapeptides. We then verified the identification rates of the rules. Shown in Figure 4 is a demonstration that the SOM rules are much less sensitive (hence more robust) than the C5 rules when the evolutionary mutation takes place at the subsites beside the cleavage sites. Note that the horizontal and vertical axes in Figure 4 represent the relaxed similarity and the prediction accuracies, respectively. For C5 rules, when the evolutionary mutation starts, the correct classification rate for negative (noncleaved HIV octapeptides) decreases very quickly. When the similarity is relaxed to 4, the correct classification rate of noncleaved

HIV octapeptides is zero, and the total prediction accuracy is then 30%. On the other hand, the total accuracy of SOM rules on the in-house testing data maintains until the similarity is relaxed to 8. The total accuracy drops to 30% until the similarity is relaxed to 14.

SUMMARY

This paper has presented a method for mining biological data using a self-organizing map (SOM). The novelty of this method is the use of the pattern recognition method for verifying the mined knowledge for prediction and classification on unseen data. The method has been successfully applied to the prediction of HIV protease cleavage sites in proteins. The study shows that tyrosine, proline, phenylalnine, and leucine are more important since they appear at the cleavage sites of HIV octapeptides, R$_1$ and R$_{1'}$ much more often. The result is highly consistent with biochemical experiments. The SOM rules are more accurate than the C5 rules when they are used for decision making. Importantly, the SOM rules are much more robust (less sensitive to noise) than the C5 rules.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their valuable comments.

REFERENCES AND NOTES

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

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

(3) Narayanan, A.; Wu, X. K.; Yang, Z. R. Mining viral protease data to extract cleavage knowledge. *Bioinformatics* **2002**, *18*, s1−s5.

(4) Baldi, P.; Pollastri, G.; Andersen, C. A.; Brunak, S. Matching protein beta-sheet partners by feedforward and recurrent neural networks. *Proc. Int. Conf. Intell. Syst. Mol. Biol., ISMB* **2000**, *8*, 25−36.

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

(6) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, 1995.

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

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

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

(10) Zhou, G. P.; Troy, F. A. 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* **2003**, *13*, 51−71.

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

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

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

(14) Chou, K. C.; Zhang, C. T. Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275−349.

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

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

(17) Rahman, S.; Miles, M. F. Identification of novel ethanol-sensitive genes be expression profiling. *Pharmacol. Ther.* **2001**, *92*, 123−134.

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

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

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

(21) Chu, S.; DeRisi, J. L.; Eisen, J.; Mulholland, D.; Botstein, D.; Brown, P. O.; Herskowitz, I. The transcriptional program of sporulation in budding yeast. *Science (Washington, D.C.)* **1998**, *282*, 699−705.

(22) DeRisi, J. L.; Iyer, V. R.; Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science (Washington, D.C.)* **1997**, *278*, 680−686.

(23) Eisen, M.; Spellman, P.; Brown, P.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. *PNAS* **1998**, *95*, 14863−14868.

(24) Iyer, V. R.; Horak, C. E.; Scale, C. S.; Botstein, D.; Snyder, M.; Brown, P. O. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature (London)* **2001**, *409*, 533−538.

(25) Lukashin, A. V.; Fuchs, R. Analysis of temporal gene expression profiles: Clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics* **2001**, *17*, 405−414.

(26) Spellman, P. T.; Sherlock, G.; Zhang, M. Q.; Iyer, V. R.; Anders, K.; Eisen, M. B.; Brown, P. O.; Botstein, D.; Futcher, B. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol.* **1998**, *9*, 3273−3297.

(27) Yeung, K. Y.; Ruzzo, W. L. Principal component analysis for clustering gene expression data. *Bioinformatics* **2001**, *17*, 763−774.

(28) Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P. O.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R. B. Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001**, *17*, 520−525.

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

(30) Bengio, Y.; Pouliot, Y. Efficient recognition of immunoglobulin domains from amino acid sequences using a neural network. *CABIOS* **1990**, *6*, 319−324.

(31) Ferran, E. A.; Ferrara, P. Topological maps of protein sequences. *Biol. Cybernetics* **1991**, *65*, 451−458.

(32) Wang, H. C.; Dopazo, J.; Carazo, J. M. Self-organising tree growing network for classifying amino acids. *Bioinformatics* **1998**, *14*, 376−377.

(33) Ferran, E. A.; Pflugfelder, B. A hybrid method to cluster protein sequences based on statistics and artificial neural networks. *CABIOS* **1993**, *9*, 671−680.

(34) Tamayo, P.; Slonim, D.; Mesirov, J.; Zhu, Q.; Kitareewan, S.; Dmitrovsky, E.; Lander, E. S.; Golub, T. R. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *PNAS* **1999**, *96*, 2907−2912.

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

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

(37) Kohonen, T. *Self organization and associative Memory*, 3rd ed.; Springer: Berlin, 1989.

(38) Qian, N.; Sejnowski, T. Predicting the secondary structure of globular proteins using neural network models. *Proc. Int Jt. Conf. Neural Networks* **1988**, 865−884.

(39) 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.; Montagnier, L. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science (Washington, D.C.)* **1983**, *220*, 868−871.

(40) 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.; White, G.; Foster, P.; Markham, P. D. Frequent detection and isolation of cytopathic retroviruses (HTLV III) from patients with AIDS and at risk for AIDS. *Science (Washington, D.C.)* **1984**, *224*, 500−503.

(41) Hellen, C. U. T.; Krausslich, H. G.; Wimmer, E. *Biochemistry* **1989**, *28*, 9881−9890.

(42) Kohl, N. E.; Emini, E. A.; Sigal, I. S. Active human immunodeficiency virus protease is required for viral infectivity. *PNSA* **1988**, *85*, 4686−4690.

(43) 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.; Springer, J. P. Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. *Nature (London)* **1989**, *337*, 615−620.

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

(45) Ashorn, P.; McQuade, T. J.; Thaisrivongs, S.; Tomasselli, A. G.; Tarpley, W. G.; Moss, B. An inhibitor of the protease blocks maturation of human and simian immunodeficiency viruses and spread of infection. *PNAS* **1999**, *87*, 7472−7476.

(46) McQuade, T. J.; Tomasselli, A. G.; Liu, L. A synthetic hiv-1 protease inhibitor with antiviral activity arrests hiv-like particle maturation. *Science (Washington, D.C.)* **1990**, *247*, 454−456.

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

(48) 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.; Lambert, R. W.; Merrett, J. H.; Mills, J. S.; Parkes, K. E. B.; Redshaw, S.; Ritchie, A. J.; Taylor, D. L.; Thomas, G. J.; Machin, P. J. Rational design of peptide-based hiv proteinase inhibitors. *Science (Washington, D.C.)* **1990**, *248*, 358−361.

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

(50) Putney, S. How antibodies block HIV infection: Paths to an AIDS vaccine. *Trends Biochem. Sci.* **1992**, *17*, 191−196.

(51) Pearl, L. H.; Taylor, W. R. A structural model for the retroviral proteases. *Nature (London)* **1987**, *329*, 351−354.

(52) Toh, H.; Ono, M.; Saigo, K.; Miyata, T. *Nature (London)* **1985**, *315*, 691.

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

(54) Chou, K. C. Review: Low-frequency collective motion in biomacromolecules and its biological functions. *Biophys. Chem.* **1988**, *30*, 3−48. See also: Martel, P. *Prog. Biophys. Mol. Biol*. **1992**, *57*, 129−179. Chou, K. C.; Chen, N. Y.; Forsen, S. The biological functions of low-frequency phonons: 2. Cooperative effects. *Chem. Scr.* **1981**, *18*, 126−132.

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

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

(57) Poorman, R. A.; Tomasselli, A. G.; Heinrikson, R. L.; Kezdy, F. J. 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.* **1991**, *22*, 14554−14561.

(58) Thompson, T. B.; Chou, K. C.; Zheng, C. Neural network prediction of the HIV-1 protease cleavage sitesy. *J. Theor. Biol.* **1995**, *177*, 369−379.

(59) Hong, L. Active-site mobility in human immunodeficiency virus, type 1, protease as demonstrated by crystal structure of A28s mutant. *Protein Sci.* **1998**, *7*, 300−305.

(60) Cheng, T. R. J.; Yin, Y. E.; Erickson-Viitanen, S. Mutagenesis of protease cleavage sites in the human immunodeficiency virus type 1 gag polyprotein. *J. Virol.* **1991**, *65*, 922−930.

(61) Okimoto, N. Protein hydrolysis mechanism of HIV-1 protease investigation by the ab initio MO calculations. *RIKEN Rev.* **2000**, *29*, 100−102.

(62) Pivazyan, A. D.; Matteson, D. S.; Fabry-Asztalos, L.; Singh, R. P.; Lin, P. F.; Blair, W.; Guo, K.; Robinson, B.; Prusoff, W. H. Inhibition of HIV-1 protease by a boron-modified polypeptide. *Biochem. Pharmacol.* **2000**, *60*, 927−936.

(63) Fournout, S.; Roquet, F.; Salhi, S. L.; Seyer, R.; Valverde, V.; Masson, J. M.; Jouin, P.; Pau, B.; Nicolas, M.; Hanin, V. Development and standardization of an immuno-quantified solid-phase assay for HIV-1 aspartyl protease activity and its application to the evaluation of inhibitor. *Anal. Chem.* **1997**, *69*, 1746−1752.

(64) Reich, S. H.; Melnick, M.; Pino, M. J.; Fuhry, M. A.; Trippe, A. J.; Appelt, K.; Davies, J. F.; Wu, B. W.; Musick, L. Structure-based design and synthesis of substituted 2-butanols as nonpeptidic inhibitors of HIV protease: secondary amide series. *J. Med. Chem.* **1996**, *39*, 2781−2794.

(65) Tran, T. T.; Patinoa, N.; Condoma, R.; Frogiera, T.; Guedja, R. Fluorinated peptides incorporating a 4-fluoroproline residue as potential inhibitors of HIV protease. *J. Fluorine Chem.* **1997**, *82*, 125−130.

(66) Glenn, M. P.; Pattenden, L. K.; Reid, R. C.; Tyssen, D. P.; Tyndall, J. D. A.; Birch, C. J.; Fairlie, D. P. Beta-strand mimicking macrocyclic amino acids, templates for protease inhibitors with antiviral activity. *J. Med. Chem.* **2002**, *45*, 371−381.

(67) Kassel, D. B.; Green, M. D.; Wehbie, R. S.; Swanstrom, R.; Berman, J. HIV-1 protease specificity derived from a complex mixture of synthetic substrates. *Anal. Biochem.* **1995**, *34*, 259−266.

(68) Marastoni, M.; Bortolotti, F.; Salvadori, S.; Tomatis, R. Structure−activity relationships of HIV-1 protease inhibitors containing gem-diaminoserine core unit. *Arzneim.-Forsch.* **1998**, *48*, 709−712.

(69) Tossi, A.; Antcheva, N.; Romeo, D.; Miertus, S. Development of pseudopeptide inhibitors of HIV-1 aspartic protease, analysis and tuning of the subsite specificity. *Pept. Res.* **1995**, *8*, 328−334.

(70) Carrillo, A.; Stewart, K. D.; Sham, H. L.; Norbeck, D. W.; Kohlbrenner, W. E.; Leonard, J. M.; Kempf, D. J.; Molla, A. In vitro selection and characterization of human immunodeficiency virus type 1 variants with increased resistance to ABT-378, a novel protease inhibitor. *J. Virol.* **1998**, *72*, 7532−7541.

(71) Dreyer, G. B.; Lambert, D. M.; Meek, T. D.; Carr, T. J.; Tomaszek, T. A.; Fernandez, A. V.; Bartus, H.; Cacciavillani, E. Hydroxyethylene isostere inhibitors of human immunodeficiency virus-1 protease: structure−activity analysis using enzyme kinetics, X-ray crystallography, and infected T-cell assays. *Biochemistry* **1992**, *31*, 6646−6659.

(72) Polgar, L.; Szeltner, Z.; Boros, I. Substrate-dependent mechanisms in the catalysis of human immunodeficiency virus protease. *Biochemistry* **1994**, *33*, 9351−9357.

(73) Zhou, G. P.; Assa-Munt, N. Some insights into protein structural class prediction. *PROTEINS*: *Struct.*, *Funct.*, *Genet.* **2001**, *44*, 57−59.

(74) Zhou, G. P.; Doctor, K. Subcellular location prediction of apoptosis proteins. *PROTEINS*: *Struct.*, *Funct.*, *Genet.* **2003**, *50*, 44−48.

(75) Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. A model of evolutionary change in proteins. Matrices for detecting distant relationships. In *Atlas of protein sequence and structure*; Dayhoff, M. O., Ed.; National Biomedical Research Foundation: Washington, DC, 1978; Vol. 5, pp 345−358,

CI034138N