BIOINFORMATICS

Vol. 19 no. 18 2003, pages 2390–2396
DOI: 10.1093/bioinformatics/btg331

Classification of protein quaternary structure with support vector machine

Shao-Wu Zhang, Quan Pan, Hong-Cai Zhang, Yun-Long Zhang and Hai-Yu Wang*

Department of Automatic Control, Northwestern Polytechnical University, Xi'an 710072, People's Republic of China

Received on December 17, 2002; revised on March 22, 2003; accepted on June 16, 2003

ABSTRACT

Motivation: Since the gap between sharply increasing known sequences and slow accumulation of known structures is becoming large, an automatic classification process based on the primary sequences and known three-dimensional structure becomes indispensable. The classification of protein quaternary structure based on the primary sequences can provide some useful information for the biologists. So a fully automatic and reliable classification system is needed. This work tries to look for the effective methods of extracting attribute and the algorithm for classifying the quaternary structure from the primary sequences.

Results: Both of the support vector machine (SVM) and the covariant discriminant algorithms have been first introduced to predict quaternary structure properties from the protein primary sequences. The amino acid composition and the autocorrelation functions based on the amino acid index profile of the primary sequence have been taken into account in the algorithms. We have analyzed 472 amino acid indices and selected the four amino acid indices as the examples, which have the best performance. Thus the five attribute parameter data sets (COMP, FASG, NISK, WOLS and KYTJ) were established from the protein primary sequences. The COMP attribute data set is composed of amino acid composition, and the FASG, NISK, WOLS and KYTJ attribute data sets are composed of the amino acid composition and the autocorrelation functions of the corresponding amino acid residue index. The overall accuracies of SVM are 78.5, 87.5, 83.2, 81.7 and 81.9%, respectively, for COMP, FASG, NISK, WOLS and KYTJ data sets in jackknife test, which are 19.6, 7.8, 15.5, 13.1 and 15.8%, respectively, higher than that of the covariant discriminant algorithm in the same test. The results show that SVM may be applied to discriminate between the primary sequences of homodimers and non-homodimers and the two protein sequence descriptors can reflect the quaternary structure information. Compared with previous Robert Garian's investigation, the performance of SVM is almost equal to that of the Decision tree models, and the methods of extracting feature vector from the primary sequences are superior to Robert's binning function method.

Availability: Programs are available on request from the authors.

Contact: quanpan@nwpu.edu.cn; shaowuzhang@hotmail.com

INTRODUCTION

It is generally accepted that the amino acid sequence of most, not all, proteins contains all the information needed to fold the protein into its correct three-dimensional structure (Afinsen *et al*., 1961; Afinsen, 1973). At the next level of protein organization, tertiary structures associate into quaternary structures forming multimeric proteins. The association of tertiary structure subunits depends upon the existence of complementary 'patches' on their surfaces. The patches are buried in the interfaces formed by the subunits, thus, play a role in both tertiary and quaternary structure. This suggests that primary sequences contain quaternary structure information (Robert Garian, 2001). Since protein sequence information grows significantly faster than information on three-dimensional (3D) structures of proteins, the need for predicting the structure of a given protein sequence naturally arises. Thus, predicting the spatial structure based a given protein primary sequence information could play a significant role, in conjunction with experimental methods.

The concept of quaternary structure was first put forward by Bernal in 1958 (Klotz *et al*., 1975). Quaternary structure is the interaction of non-covalently bound monomeric protein subunits to form oligomers. Such complexes are involved in various biological processes (Terry and Richard, 1998), including metabolism, signal transduction and chromosome replicating etc. The oligomeric proteins have more advantages than the monomers in the scope of functional evolution of biomacromolecules (Price, 1994). Thus, the study of quaternary structure is very interesting in biology.

Robert Garian investigated the prediction of quaternary structure from primary structure using decision-tree models and the feature extraction function (the simple binning

*To whom correspondence should be addressed.

function), and found that protein sequences contain quaternary structure information (Robert Garian, 2001). However, up to now, we have not seen other feature extracting methods and algorithms to predict the protein quaternary structure. In this paper, we try to apply support vector machine (SVM) (Vapnik, 1995, 1998), covariant discriminant algorithm (Duba and Hart, 1973; Chou and Elord, 1999) and two of protein sequence descriptors to approach this problem.

SVM is a new type of learning machine based on statistical learning theory. Due to its powerful discrimination, it was applied with success to medicine, bioinformatics, computational biology, and structure–activity relationships, such as translation initiation sites (Zien *et al*., 2000), membrane protein types (Cai *et al*., 2002, http://www.biochempress.com), protein–protein interactions (Bock and Gough, 2001), protein subcellular localization (Hua and Sun, 2001), protein fold (Ding and Dubchak, 2001), etc.

MATERIALS AND METHODS

Database

We use Robert Garian's Database R (Robert Garian, 2001). The Database R consisted of 1639 homo-oligomeric protein sequences, 914 of which are homodimers and 725 non-homodimers. It was obtained from Release 34 of the SWISS-PROT database (Bairoch and Apweiler, 1996) and limited to the prokaryotic, cytosolic subset of homo-oligomers in order to eliminate membrane proteins and other specialized proteins.

Support vector machine

The basic idea of applying SVM to pattern classification can be stated briefly as follows. First, map the input vectors into one feature space (possible with a higher dimension). Then, within this feature space, construct a hyperplane which can separate two classes. The mapping function will involve only the relatively low-dimensional vectors in the input space and dot products in the feature space. These dot products are represented by kernel functions. Thus the 'curse of dimensionality' can be avoided. SVM training always seeks a globally optimized solution and avoids over-fitting, so it is of the ability to deal with a large number of features.

The decision function implemented by SVM can be written as:

$$f(\vec{x}) = \operatorname{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i k(\vec{x}, \vec{x}_i) + b\right).$$

Two typical kernel functions are listed below:

Polynomial function

$$k(\vec{x}_i, \vec{x}_j) = (\vec{x}_i \bullet \vec{x}_j + 1)^d.$$

Radial basis function (RBF)

$$k(\vec{x}_i, \vec{x}_j) = \exp(-\gamma \|\vec{x}_i - \vec{x}_j\|^2).$$

The software used to implement SVM was SVM$^{\text{light}}$ by Joachims (1999) which can be freely downloaded from http://ais.gmd.de/~thorsten/svm_light for academic use.

Covariant discriminant algorithm

Suppose the jth protein in the ρ-class can be denoted as:

$$x_j^\rho = \left[x_{j1}^\rho, x_{j2}^\rho, \ldots, x_{jn}^\rho\right]^{\mathrm{T}},$$
$$j = 1, 2, \ldots, N_\rho; \quad \rho = 1, 2, \ldots, l,$$

where N_ρ is the numbers of ρ-class protein, l is the numbers of protein classes.

Denoted by X^ρ the average vector for the proteins in the ρ-class, we have

$$X^\rho = [x_1^\rho, x_2^\rho, \ldots, x_n^\rho]^{\mathrm{T}},$$

where

$$x_i^\rho = \frac{1}{N_\rho} \sum_{j=1}^{N\rho} x_{ji}^\rho \quad i = 1, 2, \ldots, n \quad \rho = 1, 2, \ldots, l.$$

Denoted by C^ρ the covariance matrix for the protein in the ρ-class, we find

$$C^\rho = \begin{bmatrix} C_{11}^\rho & C_{12}^\rho & \cdots & C_{1n}^\rho \\ C_{21}^\rho & C_{22}^\rho & \cdots & C_{2n}^\rho \\ \vdots & \vdots & \vdots & \vdots \\ C_{N_\rho 1}^\rho & C_{N_\rho 2}^\rho & \cdots & C_{N_\rho n}^\rho \end{bmatrix},$$

where

$$C_{ji}^\rho = \frac{1}{N_\rho - 1} \sum_{s=1}^{N_\rho} \left(x_{sj}^\rho - x_j^\rho\right)\left(x_{si}^\rho - x_i^\rho\right).$$

Suppose that the attribute vector associated with the query protein is denoted by X. The similarity or dissimilarity between the average vector X^ρ and X is characterized by the covariant discriminant function $F(X, X^\rho)$ and

$$F(X, X^\rho) = D^2(X, X^\rho) + \ln|C^\rho|,$$

where

$$D^2(X, X^\rho) = (X - X^\rho)^{\mathrm{T}} (C^\rho)^{-1} (X - X^\rho).$$

The criterion to perform the classification is based on the principle of the least covariant discriminant function.

$$F(X, X^\tau) = \operatorname{Min}\{F(X, X^1), F(X, X^2), \ldots, F(X, X^l)\}.$$

If $\tau = \rho, (\rho = 1, 2, \ldots, l)$, then the query protein is classified as a member of ρ class.

Extraction of the sequence descriptor

According to the studies of Nakashima *et al*. (1986); Klein (1986) and Chou and Maggiora (1998); Chou and Elord (1999) etc, the 20D(dimension) attribute vector $\vec{x}_a$ is used to represent a protein primary sequence.

$$\vec{x}_a = [f_1, f_2, \ldots, f_{20}]^{\mathrm{T}}$$

where $f_i (i = 1, 2, \ldots, 20)$ are the occurrence frequencies of the 20 amino acids in the protein concerned, arranged alphabetically according to their signal letter codes, and T means a transpose operator.

Since the information within the primary sequence is greatly reduced by considering the amino acid composition alone, the sequence orders of amino acids in the query protein have been taken into account. Thus the auto-correlation functions based on the physicochemical properties of amino acid along the primary sequence of the query protein have been considered here. In other words, in addition to the 20D components of the amino acid frequencies, other mD components should be added in to form a $(20+m)$D vector. Thus the attribute vector will be defined as:

$$\vec{x}_b = [f_1, f_2, \ldots, f_{20}, r_1, r_2, \ldots, r_m]^{\mathrm{T}}$$

where $r_j (j = 1, 2, \ldots, m)$ are the auto-correlation functions, and m is an integer to be determined by the optimum classification. In order to calculate the auto-correlation functions, we replace each residue in the primary sequence by its amino acid index (Shuichi Kawashima *et al*., 1999). Here an amino acid index is a set of 20 numerical values representing any of the different physicochemical properties of the 20 amino acids. Consequently, the replacement results in a numerical sequence: $h_1, h_2, \ldots, h_L$.

The auto-correlation functions r_j are defined as (Cornette *et al*., 1987; Zhang *et al*., 1998):

$$r_j = \frac{1}{L-j} \sum_{i=1}^{L-j} h_i h_{i+j}, \quad j = 1, 2, \ldots, m, \qquad (1)$$

where h_i is the amino acid index for the ith residue, and L is the length of protein sequence.

According to the above description, we extract five attribute parameter sets from protein primary sequences, which are clearly shown in Table 1.

Table 1. Five parameter data sets extracted from protein primary sequences

Symbol	Parameter data set
COMP	This set is composed of amino acid compositions
FASG[a]	This set is composed of amino acid compositions and the auto-correlation functions of amino acid residue index of Fasman
NISK[b]	This set is composed of amino acid compositions and the auto-correlation functions of amino acid residue index of Nishikawa–Ooi
WOLS[c]	This set is composed of amino acid compositions and the auto-correlation functions of amino acid residue index of Wold *et al*.
KYTJ[d]	This set is composed of amino acid compositions and the auto-correlation functions of amino acid residue index of Kyte–Dolittle

These index values can be found in the web http://www.genome.ad.jp/dbget/aaindex.html
[a]FASG760104 pK-N (Fasman, 1976).
[b]NISK860101 14 A contact number (Nishikawa–Ooi, 1986).
[c]WOLS870101 Principal property value zl (Wold *et al*., 1987).
[d]KYTJ820101 Hydropathy index (Kyte–Doolittle, 1982).

Classification of system assessment

The classification quality can be examined using the jackknife test and 10-fold cross-validation (10CV) test (Weiss and Kulikowski, 1990), which are objective and rigorous testing procedures. In the jackknife test, each protein is singled out in turn as a test protein and the remaining proteins are used as training protein set. In 10CV test, the protein data set will be randomly partitioned into 10 blocks of proteins of approximately equal size, and one of them is singled out in turn as test proteins and the other 9 blocks are used as training protein set. This process is repeated for 10 iterations, each time setting aside a different test block. Thus, each of the blocks will serve both as part of the training set and as a test block. The overall classification accuracy (Q), the true positive rate (TPR), the false positive rate (FPR) and Matthew's Correlation Coefficient (MCC) (Fasman, 1976) for assessment of the classification system are respectively defined as:

$$Q = (\mathrm{TP} + \mathrm{TN})/N$$

$$\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$$

$$\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

Here, N is the total number of sequences, TP and TN are the numbers of correctly classified sequences of positive and negative samples, respectively, FP and FN are the numbers of incorrectly classified sequences of negative and positive samples, respectively.

RESULTS

The results of the SVM and the covariant discriminant algorithms in the jackknife test for Database R are shown in Table 2. The overall accuracy of SVM for COMP, FASG, NISK, WOLS and KYTJ data sets are 78.5, 87.5, 83.2, 81.7 and 81.9%, respectively, which are 19.6, 7.8, 15.5, 13.1 and 15.8% higher than that of covariant discriminant algorithm. Using the same covariant discriminant algorithm, the overall accuracy for FASG, NISK, WOLS and KYTJ

Table 2. The results of the SVM and the covariant discriminant algorithms in the jackknife test

	SVM ($C = 1000$)					Covariant Discriminant				
	COMP $\gamma = 0.044$	FASG $m = 14$ $\gamma = 0.026$	NISK $m = 26$ $\gamma = 0.024$	WOLS $m = 24$ $\gamma = 0.022$	KYTJ $m = 21$ $\gamma = 0.032$	COMP	FASG $m = 14$	NISK $m = 26$	WOLS $m = 24$	KYTJ $m = 21$
Sensitivity	0.845	0.899	0.888	0.860	0.884	0.357	0.686	0.542	0.554	0.498
Specificity	0.709	0.844	0.760	0.763	0.737	0.882	0.938	0.848	0.854	0.866
Positive predictive rate	0.785	0.879	0.824	0.820	0.809	0.793	0.933	0.818	0.827	0.824
Negative predictive rate	0.784	0.869	0.844	0.812	0.834	0.521	0.703	0.595	0.603	0.578
TPR	0.845	0.899	0.888	0.860	0.884	0.357	0.686	0.542	0.554	0.498
FPR	0.291	0.156	0.240	0.237	0.263	0.117	0.062	0.152	0.146	0.134
Overall accuracy	0.785	0.875	0.832	0.817	0.819	0.589	0.797	0.677	0.686	0.661
MCC	0.561	0.746	0.658	0.628	0.632	0.274	0.630	0.401	0.418	0.383

Table 3. Performance comparisons of SVM and decision tree methods in 10CV test

	Decision tree SDDV	SVM ($C = 1000$)					
		SDDV $\gamma = 0.2$	COMP $\gamma = 0.044$	FASG $m = 14$ $\gamma = 0.026$	NISK $m = 26$ $\gamma = 0.024$	WOLS $m = 24$ $\gamma = 0.022$	KYTJ $m = 21$ $\gamma = 0.032$
FPR	0.402	0.407	0.302	0.160	0.250	0.247	0.273
TPR	0.781	0.788	0.847	0.893	0.888	0.858	0.879
Overall accuracy	0.699	0.702	0.781	0.870	0.827	0.812	0.812
MCC	0.386	0.390	0.554	0.735	0.648	0.617	0.618

data sets are 20.8, 8.8, 9.7 and 7.2%, respectively, higher than that of COMP data set. These results indicate that the classification accuracy can be significantly improved using the same classification information (such as amino acid composition) with a more powerful algorithm and the auto-correlation functions of amino acid index profile of the primary sequence appear to capture the information of protein sequence orders.

Using the same Database R, we have compared the performance of SVM and the decision tree models, and also compared our methods of extracting feature vector from the protein primary sequences with Robert's method that the feature vectors are extracted by a binning function from the sequence profiles created with an amino acid index (the feature attribute data set is denoted by the symbol SDDV). Table 3 shows that the performance of SVM is almost equal to the Decision tree models, and the methods of extracting feature vector based on the amino acid composition and the auto-correlation function of the amino acid residue index from the primary sequences are superior to the binning function. The standard deviations of the accuracy measures for SDDV, COMP, FASG, NISK, WOLS and KYTJ in 10CV test are also shown in Table 4.

Table 4. The standard deviation of the accuracy measures using SVM in 10CV test

	SDDV	COMP	FASG	NISK	WOLS	KYTJ
Standard deviation	0.0019	0.0024	0.0041	0.0035	0.0033	0.0046

The process of randomly partitioning each of the data sets was repeated five times.

These results show that both the methods of extracting protein sequence attributes are feasible and effective, and the SVM may be applied to the classification of the protein homodimers and non-homodimers.

The optimal number of amino acid auto-correlation functions

The classification results can be affected by m, the number of the auto-correlation functions used in the FASG, NISK, WOLS and KYTJ parameter sets. We take FASG set as an example to investigate this problem. The results are clearly shown in Figures 1 and 2.

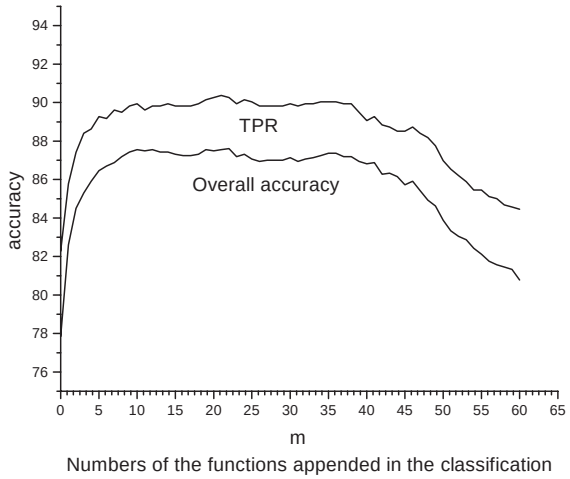

Fig. 1. The relationship between the number of the auto-correlation functions used in the classification (x-axis) and the overall accuracy and the TPR (y-axis) in the 10CV test.

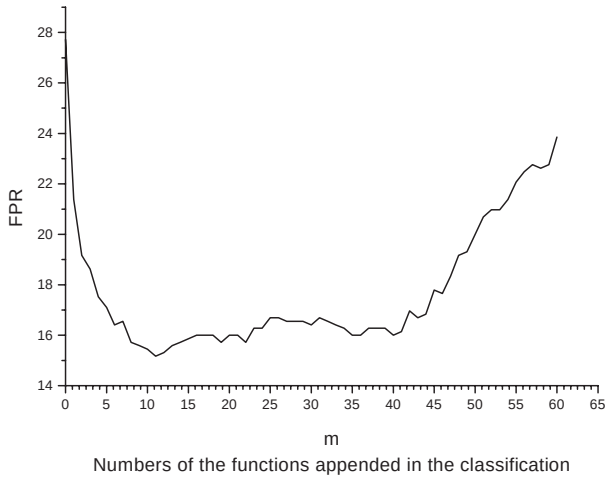

Fig. 2. The relationship between the number of the auto-correlation functions used in the classification (x-axis) and the FPR (y-axis) in the 10CV test.

From Figures 1 and 2, it is seen that when m is changed in the scope of 8–40, the classification results are almost unchanged. Here we select $m = 14$. The best overall accuracy, TPR and FPR will be 87.0, 89.3 and 16.0%, respectively, using the SVM method in 10CV test.

DISCUSSION AND CONCLUSION

The performance of the classification system influenced by the size and the homology of the database

To investigate the influence of the size and the homology of the database, we established two subsets (Database A and Database B). The Database A is randomly selected from the Database R, which consists of 711 homo-oligomeric protein sequences, 417 of which are homodimers and 294 non-homodimers. The Database B is derived from the Database R based on the sequence alignment program BLAST (Altschul *et al*., 1997) with pairwise sequence identity less than 50%, 417 of which are homodimers and 294 non-homodimers. The sequence identity of the Database R and Database A is higher than that of the Database B and the size of the Database A and Database B is equal. The results of the two subsets are shown in Table 5. The results of the Database A are the mean of five random selections. It is seen that when the size and the homology of the database decreases, the performance of the classification system also lowers. This may result in a memorization in the classification. If the training data and the testing data are highly identical or homologous, then the classification accuracy could be misleadingly high. The best solution to such memorization problem appears to increase the size of database and decrease the sequence identity.

Table 5. Influence by the homology and size of the database based on the amino acid composition attribute data set using SVM ($C = 1000$, $\gamma = 0.044$) in the jackknife test

	Database R	Database A	Database B
FPR	0.291	0.403	0.517
TPR	0.845	0.841	0.782
Overall accuracy	0.785	0.740	0.658
MCC	0.561	0.455	0.278

SVM parameters selection

SVM still has a few adjustable parameters to be determined. SVM training includes the selection of the proper kernel function parameters and the regularization parameter C. Both of polynomial kernel and RBF kernel are to be investigated. Simulations show that when $C \geq 10$, it has almost no effect on the classification performance of two types of kernel functions, so we select the default $C = 1000$ of SVM^{light} program. For polynomial kernel, the algorithm may be divergent or the training speed is very slow, therefore we did not select it for investigation. The parameter γ of RBF kernel has a different effect on classification performance for a different data set.

The selection of the amino acid index and the m values

We have analyzed 472 sets of indices in Aaindex ver.5.0. The database may be accessed through the DBGET/LinkDB system at GenomeNet (http://www.genome.ad.jp/dbget). These indices were found to cluster into a small number of groups: α and turn propensity, β propensity, composition, hydrophobicity, physicochemical properties, and other properties (Tomi and Kanehisa, 1996). Each of the 472 sets of indices is

tested separately. The overall accuracy, TPR and FPR in 10CV test are used to evaluate the classifying ability of sequence descriptors based on each amino acid index. Among 472 sets of data, about 40% may differently improve the classification results, and the indices listed here are four examples of 472 indices, whose results are the best. Although the feature vectors composed of amino acid composition and auto-correlation functions of amino acid residue index can reflect the quaternary structure information at a certain extent, this method of representing protein sequence has a certain limitation. With the different amino acid indices and *m* values of auto-correlation functions, there are many integrating forms of amino acid composition and auto-correlation functions. Thus, the best classification result can be obtained for a given data set by carefully selecting amino acid index and *m* value.

These simulation results show that both the methods of extracting attributes based on the amino acid composition and the auto-correlation functions are feasible and effective; the SVM can be applied to protein quaternary classification, and its performance is better than that of the covariant discriminant algorithm and almost equal to the decision tree method. The primary sequences of homo-oligomeric proteins contain quaternary structure information. The feature vectors composed of amino acid composition and the auto-correlation functions of amino acid residue index appear to capture essential information about the composition and hydrophobicity of residues in the surface patches that buried in the interfaces of associated subunits. It is anticipated that the current classification method would be a useful tool for the large-scale analysis of genome data, and may provide some useful information for biologists who are at the investigation of the biomacromolecules. We will try to find other feature extracting methods for classifying homodimers and non-homodimers, and apply them to predict homo-multimers and hetero-multimers in the future work.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Robert Garian (School of Computational Sciences, George Mason University, USA) for providing the database. This work was supported by a Doctor Innovation Grant of Northwestern Polytechnical University (China).

REFERENCES

Altschul,S.F., Thomas,L.M., Alejandro,A.S., Zhang,J.H., Zhang,Z., Webb,M. and David,J.L. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Anfinsen,C.B., Haber,E., Sela,M. and White,F.H. Jr. (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl Acad. Sci. USA*, **47**, 1309–1314.

Anfinsen,C.B. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.

Bairoch,A. and Apweiler,R. (1996) The SWISS-PROT protein data bank and its new supplement TrEMBL. *Nucleic Acids Res.*, **24**, 21–25.

Bock,J.R. and Gough,D.A. (2001) Predicting protein–protein interactions from primary structure, *Bioinformatics*, **17**, 455–460.

Cai,Y.D., Liu,X.J., Xu,X. and Chou,K.C. (2002) Support vector machines for predicting membrane protein types by incorporating quasi-sequence-order effect. *Internet Electron. J. Mol. Des.*, **1**, 219–226,

Cornette,J.L., Cease,K.B., Margali,H., Spouge,J.L., Berzofsky,J.A. and Delisi,C. (1987) Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.*, **195**, 659–685.

Chou,K.C. and Maggiora,G.M. (1998) Domain structural prediction. *Protein Eng.*, **11**, 523–538.

Chou,K.C. and Elord,D.W. (1999) Protein subcellular location prediction. *Protein Eng.*, **12**, 107–118.

Ding,C.H.Q. and Dubchak,I. (2001) Multi-class protein fold recognition using support vector machines and neural networks, *Bioinformatics*, **17**, 349–358.

Duba,R.O. and Hart,P.E. (1973) *Pattern Classification and Scene Analysis*, Chap. 2. Wiley, New York.

Fasman,G.D. (1976) *Handbook of Biochemistry and Molecular Biology*, 3rd edn., Proteins—Volume 1. CRC Press, Cleveland.

Hua,S.J. and Sun,Z.R. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.

Joachims,T. (1999) Making large-scale SVM learning practical. In Scholkopf, B., Burges,C. and Smola, A. (eds), *Advances in Kernal Methods—Support Vector Learning*. MIT Press, Cambridge, MA, pp. 42–56.

Kawashima,S., Ogata,H. and Kanehisa,M. (1999) Aaindex: amino acid index database. *Nucleic Acids Res.*, **27**, 368–369.

Klotz,I.M., Darnall,D.W. and Langerman,N.R. (1975) *The Protein*, 3rd edn. Academic Press, New York, Vol. 1, pp. 293–411.

Klein,P. (1986) Prediction of protein structural class by discriminant analysis. *Biochim. Biophys. Acta.*, **876**, 205–275.

Kyte,J. and Doolittle,R. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Nakashima,H., Nishikawa,K. and Ooi,T. (1986) The folding type of a protein is relevant to the amino acid composition. *J. Biochem.*, **99**, 152–162.

Nishikawa,K. and Ooi,T. (1986) The folding type of a protein is relevant to the amino acid composition. *J. Biochem.*, **99**, 152–162.

Price,N.C. (1994) Assembly of multi-subunit structure. In Pain,R.H. (ed.), *Mechanisms of Protein Folding*. Oxford University Press, New York, pp. 160–193.

Robert,G. (2001) Prediction of quaternary structure from primary structure. *Bioinformatics*, **17**, 551–556.

Terry,B.F. and Richard,M.C. (1998) Determination of protein–protein interactions by matrix-assisted laser desorption/ionization mass spectrometry. *J. Mass Spectrom.*, **33**, 697–704.

Tomi,K. and Kanehisa,M. (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.*, **9**, 27–36.

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

Vapnik,V. (1998) *Statistical Learning Theory*. Wiley, New York.

Weiss,S.M. and Kulikowski,C.A. (1990) *Computer Systems That Learn*. Morgan Kaufmann, San Mateo, CA.

Wold,S., Eriksson,L., Hellberg,S., Jonsson,J., Sjostrom,M., Skagerberg,B. and Wikstrom,C. (1987) Principal property values for six non-natural amino acids and their application to a structure-activity relationship for oxytocin peptide analogues. *Can. J. Chem.* **65**, 1814–1820.

Zhang,C.T., Lin,Z.S., Zhang,Z.D. and Yan,M. (1998) Prediction of the helix/strand content of globular proteins based on their primary sequences [J]. *Protein Eng.*, **11**, 971–979.

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