Computers and Chemistry 26 (2002) 293–296

Computers & Chemistry

www.elsevier.com/locate/compchem

Short Communication

Prediction of protein structural classes by support vector machines

Yu-Dong Cai [a,*], Xiao-Jun Liu [b], Xue-biao Xu [c], Kuo-Chen Chou [d]

[a] *Shanghai Research Centre of Biotechnology, Chinese Academy of Sciences, Shanghai 200233, China*
[b] *Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK*
[c] *Department of Computing Science, University of Wales, College of Cardiff, Queens Buildings, Newport Road, P.O. Box 916, Cardiff CF2 3XF, UK*
[d] *Computer-Aided Drug Discovery, Upjohn Laboratories, Kalamazoo, MI 49001-4940, USA*

Received 31 May 2001; received in revised form 7 June 2001; accepted 6 July 2001

Abstract

In this paper, we apply a new machine learning method which is called support vector machine to approach the prediction of protein structural class. The support vector machine method is performed based on the database derived from SCOP which is based upon domains of known structure and the evolutionary relationships and the principles that govern their 3D structure. As a result, high rates of both self-consistency and jackknife test are obtained. This indicates that the structural class of a protein inconsiderably correlated with its amino acid composition, and the support vector machine can be referred as a powerful computational tool for predicting the structural classes of proteins. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Support vector machine; Protein structural class; Jackknife test; Self-consistency

1. Introduction

In general, protein structure can be classed into all-α, all-β, α/β, and α + β (Levitt and Chothia, 1976) and ζ protein (Chou and Zhang, 1993) according to protein chain folding topologies. The so-called ζ proteins are highly irregular that contain very little or no α-helices and β-sheets at all. Prediction of protein structural class is very important to many aspects of molecular biology. Previous studies have shown evidence that some correlation between the protein structural class and amino acid composition does exist, and the protein structural class can be predicted according to amino acid composition alone to some extent (Chou, 1980, 1989; Nakashima et al., 1986; Klein and Delisi, 1986; Metfessel et al., 1993; Dubchak et al., 1993; Chou and Zhang, 1994; Mao et al., 1994; Chou, 1995; Chandonia and Karplus, 1995; Bahar et al., 1997; Chou et al., 1998; Zhou, 1998; Cai and Zhou, 2000; Cai et al., 2000). This implies that protein structural class is significantly dictated by the interactions among the components of amino acid composition, although it is well known that the three-dimensional structure of protein is determined by the amino acid interactions over the entire sequence chain.

In this paper, we try to apply Vapnik's support vector machine (Vapnik, 1995) to approach this prob-

[*] Corresponding author. Present address: Biomolecular Sciences Department, UMIST, P.O. Box 88, Manchester M60 1QD, UK. Tel.: +44-161-200-4191; fax: +44-161-236-0409.

E-mail address: y.cai@umist.ac.uk (Y.-D. Cai).

0097-8485/02/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S0097-8485(01)00113-9

lem. In this work, the support vector machine was tested based on a new paradigmatic dataset (Chou, 1999) derived from the SCOP database (Murzin et al., 1995). As a result, it reached high rates of self-consistency and the jackknife test. This shows that the structural class of a protein is considerably correlated with its amino acid composition, and the support vector machines can become a useful tool for predicting the structural classes of proteins.

2. Support vector machine

Support vector machine (SVM) is a kind of learning machine based on statistical learning theory. 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), either linearly or non-linearly, which is relevant with the selection of the kernel function. Then, within the feature space from the first step, seek an optimized linear division, i.e. construct a hyperplane which separates two classes (this can be extended to multi-class). SVM training always seeks a global optimized solution and avoids over-fitting, so it has the ability to deal with a large number of features. A complete description to the theory of SVMs for pattern recognition is in Vapnik's book (Vapnik, 1998).

SVMs have been used in a range of problems including drug design (Burbidge et al., 2000), image recognition and text classification (Joachims, 1998).

In this paper, we apply Vapnik's support vector machine (Vapnik, 1995) for predicting the structural classes of proteins. We download the SVMlight, which is an implementation (in C language) of SVM for the problem of pattern recognition. The optimization algorithm used in SVMlight can be found in Joachims (1999a,b). The code has been used in text classification, and image recognition (Joachims, 1998).

Suppose we are given a set of samples, i.e. a series of input vectors

$$X_i \in R^d \quad (i = 1, ..., N)$$

with corresponding labels $y_i \in \{+1, -1\}$ $(i = 1, ..., N)$.

Here -1 and $+1$ are used to represent, respectively, the two classes. The goal here is to construct one binary classifier or derive one decision function from the available samples, which has small probability of misclassifying a future sample. Both the basic linear separable case and the most useful linear non-separable case for most real life problems are considered here.

2.1. The linear separable case

In this case, there exists a separating hyperplane whose function is $\vec{W} \cdot \vec{x} + b = 0$, which implies

$$y_i(\vec{W} \cdot \vec{x}_i + b) \geq 1, \quad i = 1, ..., N$$

By minimizing $(1/2)\|\vec{W}\|^2$ subject to this constraint, the SVM approach tries to find a unique separating hyperplane. Here $\|\vec{W}\|^2$ is the Euclidean norm of $\vec{w}$, which maximizes the distance between the hyperplane (optimal separating hyperplane or OSH in Cortes and Vapnik (1995)) and the nearest data points of each class. The classifier is called the largest margin classifier.

By introducing Lagrange multipliers α_i, the SVM training procedure amounts to solving a convex QP problem. The solution is a unique globally optimized result, which can be shown to have the following expansion:

$$\vec{W} = \sum_{i=1}^{N} y_i \alpha_i \cdot \vec{x}_i$$

Only if the corresponding $\alpha_i > 0$, these $\vec{x}_i$ are called support vectors.

When an SVM is trained, the decision function can be written as

$$f(\vec{x}) = \mathrm{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i \cdot \vec{x} \cdot \vec{x}_i + b \right)$$

where sgn( ) in the above formula is the given sign function.

2.2. The linear non-separable case

2.2.1. 'Soft margin' technique

In order to allow for training errors, Cortes and Vapnik (1995) introduced slack variables

$$\xi_i > 0, \quad i = 1, \ldots, N$$

The relaxed separation constraint is given as

$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 - \xi_i, \quad (i = 1,\ldots,N)$$

and the OSH can be found by minimizing

$$\frac{1}{2}\|\vec{w}\|^2 + C \sum_{i=1}^{N} \xi_i$$

where C is a regularization parameter used to decide a trade-off between the training error and the margin.

2.2.2. 'Kernel substitution' technique

SVM performs a non-linear mapping of the input vector $\vec{x}$from the input space R^d into a higher dimensional Hilbert space, where the mapping is determined by the kernel function. Then like in Section 2.1, it finds the OSH in the space H corresponding to a non-linear boundary in the input space. Two typical kernel functions are listed below

$$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i \cdot \vec{x}_j + 1)^d$$

$$K(\vec{x}_i, \vec{x}_j) = \exp(-r\|\vec{x}_i - \vec{x}_j\|^2)$$

and the form of the decision function is

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

For a given dataset, only the kernel function and the regularity parameter C must be selected to specify one SVM.

3. The training and prediction of protein structural class

The dataset studied here was taken from Chou (1999). It consists of 204 protein chains, of which 52 are all-α proteins, 61 all-β proteins, 45 α/β proteins and 46 α + β proteins. The dataset was derived from SCOP according to the following conditions. (1) Any protein in the dataset must, as a whole, clearly and unambiguously belong to one of the four structural classes. (2) Each subset in the datset must contain a statistically significant number of proteins that belong to a same structural class. The process of constructing such a paradigmatic working dataset has been clearly elaborated in Chou (1999), and there is no need to repeat it here. The sequence matches performed between all members in each subset have indicated that the average sequence identity percentages for the all-α, all-β, α/β, and α + β are 21, 30, 15, and 14%, respectively, indicating that the majority of pairs in each of the subsets concerned have the low relative sequence identity.

According to its amino acid composition, a protein can be represented by a point or a vector in a 20-D space. However, of the 20 amino acid composition components, only 19 are independent due to the normalization condition (Chou, 1995). Accordingly, strictly speaking, if based on amino acid composition, a protein should be represented by a point or a vector in a 19-D, rather than a 20-D space as defined in a conventional manner. Furthermore, according to Chou's invariance theorem, the final predicted result will remain the same regardless of which one of the 20 components is left out for forming the 19-D space. It is extremely important to realize this, particularly when the calculations involve a covariance matrix such as in the case of Chou (1995), Chandonia and Karplus (1995), Bahar et al. (1997) and Chou et al. (1998). The amino acid composition is taken as the input of the SVM.

The SVM method applies to two-class problems. In this paper, for the four-class problems, we use a simple and effective method: 'one-against-others' method (Brown et al., 2000; Ding and Dubchak, 2001) to transfer it into two-class problems.

The computations were carried out on a Silicon Graphics IRIS Indigo workstation (Elan 4000).

In this research, for the SVM, the width of the Gaussian RBFs is selected as that which minimized an estimate of the VC-dimension. The parameter C that controls the error-margin trade-off is set at 150. After being trained, the hyperplane output by the SVM was obtained. This indicates that the trained model, i.e. hyperplane output which is including the important information, has the function to identify protein structural classes.

We first test the self-consistency of the method, and later test the method by cross-validation (jackknife test). As a result, the rates of both self-consistency and cross-validation were quite high.

4. Results and discussion

4.1. Success rate of self-consistency of SVMs

In this research, the examination for the self-consistency of the SVMs method was tested for a dataset from Chou (1999) that contains 204 proteins: 52 all-α, 61 all-β, 45 α/β, 46 α + β. All the rates of correct prediction for the four structural classes reach 100%, indicating that after being trained, the SVM model has grasped the complicated relationship between the amino acid composition and protein structure. The result is also the same as the result obtained by Chou (1999) using the 2nd-order component-coupled algorithm.

4.2. Success rate of jackknife test of SVMs

As is well known, the single-test-set analysis, subsampling and jackknife analysis are the three tests often used for cross-validation examination. Among these three, the jackknife test is deemed as the most effective and objective one (Chou and Zhang, 1995). The jackknife test is also called leave-one-out test, in which each protein in the dataset is in turn singled out as a tested protein and all the rule-parameters are calculated without using this protein. In this paper, we use the jackknife test to the SVM method. As a result, the rates of correct prediction for the four structural classes of 204 domains were 152/204 = 74.5% (all-α: 39/52 = 75%; all-β: 55/61 = 90%; α/β domains: 29/45 = 64%; α + β domains: 29/46 = 64%). Such a rate is very close to 77%, the overall success rate obtained by Chou (1999) using the 2nd-order component-coupled algorithm to perform the jeckknife test for the same dataset.

4.3. Comparison to neural network method

In this research, we also applied the neural network method (Kohonen, 1988) to this problem. The comparison of its results to the SVM method is given in Table 1 (self-consistency test and jackknife test). We can see that the rates of both self-consistency test and jackknife test of SVM are higher than those of neural network.

Table 1
Results of the self-consistency test and jackknife test

Algorithm	Rate of correct prediction for each class (%)				Overall rate of correct prediction (%)
	All-α	All-β	α/β	α+β	
Self-consistency test					
Neural network	98.6	93.4	96.3	84.6	93.5
SVM	100	100	100	100	100
Jackknife test					
Neural network	86.0	96.0	88.2	86.0	89.2
SVM	88.8	95.2	96.3	91.5	93.2

5. Conclusions

The above results, together with those obtained by the other prediction algorithms (Chou, 1980, 1989; Nakashima et al., 1986; Klein and Delisi, 1986; Metfessel et al., 1993; Dubchak et al., 1993; Chou and Zhang, 1994; Mao et al., 1994; Chou, 1995; Chandonia and Karplus, 1995; Bahar et al., 1997; Chou et al., 1998; Zhou, 1998; Cai and Zhou, 2000; Cai et al., 2000; Chou, 1999), indicate that the structural class of a protein is considerably correlated with its amino acid composition. It is anticipated that the SVM method and the elegant covariant discriminant algorithm (Chou, 1995, 1999; Chou et al., 1998; Zhou, 1998), if complemented with them, will become a very useful tool for predicting the structural classes of proteins.

References

Bahar, I., Atilgan, A.R., Jernigan, R.L., Erman, B., 1997. Protein: Struct. Funct. Genet. 29, 172–185.

Burbidge, R., Trotter, M., Holden, S., Buxton, B., 2000. Proceedings of the AISB'00 Symposium on Artificial Intelligence in Bioinformatics, pp. 1–4.

Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C., Ares, M. Jr., Haussler, D., 2000. Proc. Natl. Acad. Sci. USA 97, 262–267.

Cai, Y.D., Zhou, G.P., 2000. Biochimie 82 (8), 783–785.

Cai, Y.D., Li, Y.X., Chou, K.C., 2000. Biochim. Biophys. Acta 1476 (1), 1–2.

Chandonia, J.M., Karplus, M., 1995. Proteins Sci. 4, 275–285.

Chou, J.J., Zhang, C.T., 1993. J. Theor. Biol. 161, 251–262.

Chou, K.C., 1995. Proteins: Struct. Funct. Genet. 21, 319–344.

Chou, K.C., 1999. Biochem. Biophys. Res. Commun. 264, 216–224.

Chou, K.C., Zhang, C.T., 1994. J. Biol. Chem. 269, 22014–22020.

Chou, K.C., Zhang, C.T., 1995. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.

Chou, K.C., Liu, W.M., Maggiora, G.M., Zhang, C.T., 1998. Proteins: Struct. Funct. Genet. 31, 97–103.

Chou, P.Y., 1980. Second Chemical Congress of the North American Continent, Las Vegas, NV, Abstracts of Papers, Part I.

Chou, P.Y., 1989. In: Fasman, G.D. (Ed.), Prediction of Protein Structure and the Principles of Protein Conformation. Plenum Press, New York, pp. 549–586.

Cortes, C., Vapnik, V., 1995. Machine Learning 20, 273–293.

Ding, C.H.Q., Dubchak, I., 2001. Bioinformatics 4 (17), 349–358.

Dubchak, I., Holbrook, S.R., Kim, S.H., 1993. Proteins: Struct. Funct. Genet. 16, 79–91.

Joachims, T., 1998. Proceedings of the European Conference on Machine Learning, Springer, Berlin.

Joachims, T., 1999a. Making large-scale SVM learning practical. In: Schölkopf, B., Burges, C., Smola, A. (Eds.), Advances in Kernel Methods-Support Vector Learning. MIT Press, Cambridge, MA.

Joachims, T., 1999b. Proceedings of the International Conference on Machine Learning.

Klein, P., Delisi, C., 1986. Biopolymers 25, 1659–1672.

Kohonen, T., 1988. Neural Networks 1, 3–16.

Levitt, M., Chothia, C., 1976. Nature 261, 552–557.

Mao, B., Chou, K.C., Zhang, C.T., 1994. Protein Eng. 7, 319–330.

Metfessel, B.A., Saurugger, P.N., Connelly, D.P., Rich, S.T., 1993. Proteins Sci. 2, 1171–1182.

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. J. Mol. Biol. 247, 536–540.

Nakashima, H., Nishikawa, K., Ooi, T., 1986. J. Biochem. 99, 152–162.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, Berlin.

Vapnik, V., 1998. Statistical Learning Theory. Wiley-Interscience, New York.

Zhou, G.P., 1998. J. Protein Chem. 8, 729–738.