J. theor. Biol. (2003) **221,** 115–120
doi:10.1006/jtbi.2003.3179, available online at http://www.idealibrary.com on IDEAL®

Support Vector Machines for Prediction of Protein Domain Structural Class

YU-DONG CAI*†, XIAO-JUN LIU‡, XUE-BIAO XU§ AND KUO-CHEN CHOU||

†*Shanghai Research Centre of Biotechnology, Chinese Academy of Sciences, Shanghai* 200233, *China*
‡*Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH*9 3*JT, U.K.* §*Department of Computing Science, University of Wales, College of Cardiff, Queens Buildings, Newport Road, P.O. Box* 916, *Cardiff CF*2 3*XF, U.K. and* ||*Upjohn Laboratories, Pfizer, Kalamazoo, MI* 49001-4940, *U.S.A.*

(*Received on* 14 *June* 2002, *Accepted in revised form on* 26 *September* 2002)

The support vector machines (SVMs) method was introduced for predicting the structural class of protein domains. The results obtained through the self-consistency test, jack-knife test, and independent dataset test have indicated that the current method and the elegant component-coupled algorithm developed by Chou and co-workers, if effectively complemented with each other, may become a powerful tool for predicting the structural class of protein domains.

1. Introduction

Protein domains can be classified into one of the following seven classes: all-α, all-β, α/β, $\alpha+\beta$, multi-domain, small protein and peptide (Chou & Maggiora, 1998). The structural class of a protein domain is correlated with its amino acid composition. However, given the amino acid composition of a protein domain, how may one predict its structural class? Various efforts have been made in addressing this problem (Bahar *et al.*, 1997; Cai *et al.*, 2000; Chou, 1995; Chou *et al.*, 1998; Chou & Maggiora, 1998; Chou & Zhang, 1994, 1995; Zhou, 1998).

We applied Vapnik's support vector machine (SVM) (Vapnik, 1995) to this problem. In this work, SVM was performed according to the database derived from SCOP (Murzin *et al.*, 1995), which was established based on domains of known structure and the evolutionary relationships, and the principles that govern their 3-D structure. As a result, high success rates in both the self-consistency test and jack-knife test were obtained. This indicates that the structural class of protein domain is considerably correlated with its amino acid composition, and the SVM may become a powerful tool for predicting the structural classes of protein domains.

2. Support Vector Machine

SVMs are kinds of learning machines based on statistical learning theory. The basic idea of applying SVMs 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 nonlinearly, 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.

*Corresponding author. Current address. Biomolecular Sciences Department, UMIST, P.O. Box 88, Manchester M60 1QD, U.K.
E-mail address: y.cai@umist.ac.uk (Y.-D. Cai).

0022-5193/03/$35.00

construct a hyperplane which separates two classes (this can be extended to multi-classes). The SVM training process 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 given in the book by 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 applied Vapnik's SVM (Vapnik, 1998) for predicting the structural classes of protein domains. We downloaded the SVMlight from the website at http://svmlight.-joachims.org. The SVMlight program 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, image recognition (Joachims, 1998).

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

$$x_i \in R^d \ (i = 1, \ldots, N)$$

with corresponding labels $y_i \in \{+1, -1\}$ $(i = 1, \ldots, N)$, where -1 and $+1$ are used to stand, respectively, for 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

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

By minimizing $\frac{1}{2}\|\vec{W}\|^2$ 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 & Vapnik, 1995)] and the nearest data points of each class. The classifier is called the largest margin classifier.

By introducing Lagrange multipliers α_i, using the Karush–Kuhn–Tucker (KKT) conditions and the Wolfe dual theorem of optimization theory, the SVM training procedure amounts to solving the following convex QP problem:

$$\textit{Maximize} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j \cdot y_i y_j \cdot \vec{x}_i \cdot \vec{x}_j$$

$$\textit{subject to} \quad \alpha_i \geqslant 0,$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0, \quad i = 1, \ldots, N.$$

The solution is a unique globally optimized result and 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

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

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

2.2. THE LINEAR NON-SEPARABLE CASE

Two important techniques needed for this case are given below.

(i) *The "soft margin" technique*: In order to allow for training errors, Cortes & Vapnik, (1995) introduced slack variables:

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

And relaxed separation constraint is given as

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

And the OSH can be found by minimizing

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

instead of $\frac{1}{2}||\vec{W}||^2$ for the above two constraints in Section 2.1, where C is a regularization parameter used to decide a trade-off between the training error and the margin.

(ii) The "kernel substitution" technique: SVM performs a nonlinear mapping of the input vector $\bar{x}$ from the input space R^d into a higher dimensional Hilbert space, where the mapping is determined by the kernel function. Then like in case (i), it finds the OSH in the space H corresponding to a nonlinear boundary in the input space. Two typical kernel functions are listed below:

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

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

where the first one is called the *polynomial kernel function of degree d* which will eventually revert to the linear function when $d = 1$, the latter one is called the radial basic function (RBF) kernel.

Finally, for the selected kernel function, the learning task amounts to solving the following QP problem:

$$\mathit{Maximize} \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j \cdot y_i y_j \cdot K(\vec{x}_i \cdot \vec{x}_j)$$

$$\mathit{subject\ to} \quad 0 \leqslant a_i \leqslant C,$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0, \quad i = 1, \ldots, N.$$

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 Structural Class of Protein domain

Following the procedures and rationale given by Chou & Maggiora (1998), the protein domains to be considered fall into one of the following seven classes: (1) all-α, (2) all-β, (3) α/β, (4) $\alpha+\beta$, (5) multi-domain, (6) small protein, (7) peptide. For brevity in formulation, we shall use μ, σ and ρ to represent multi-domain, small-domain and peptide classes, respectively, as done by Chou & Maggiora (1998).

According to its amino acid composition, a protein domain can be represented by a point or a vector in a 20-D space (Chou, 1995). The amino acid composition is taken as the input of the SVM.

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 100. 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 structural classes of protein domain.

We first tested the self-consistency of the method, then tested it by cross-validation (jack-knife approach), and finally as a demonstration, we tested the prediction quality by an independent dataset. All the results thus obtained have shown that the success rates are quite high.

4. Results and Discussion

4.1. SUCCESS RATE OF SELF-CONSISTENCY TEST

In this research, the examination for the self-consistency of the SVMs methods was tested for the following four datasets from Chou & Maggiora (1998): (1) 138 domains (36 all-α domains, 29 all-β domains, 32 α/β domains, 41 $\alpha+\beta$ domains); (2) 253 domains (63 all-α domains, 58 all-β domains, 61 α/β domains, 71 $\alpha+\beta$ domains); (3) 359 domains (82 all-α

domains, 85 all-β domains, 99 α/β domains, 93$\alpha+\beta$ domains); (4) 1601 domains (273 all-α domains, 461 all-β domains, 332α/β domains, 297$\alpha+\beta$ domains, 31 μ domains, 168 σ domains, 39 ρ domains). As a result, the overall success rates reach 100, 100, and 93% for the datasets of 138, 253, and 359 domains, respectively, indicating that after being trained, the SVMs have grasped the complicated relationship between the amino acid composition and protein domain structure. The corresponding rates as reported in Chou & Maggiora (1998) are 98, 95, and 94%, respectively. For the dataset of 1601 domains, the overall success rate by the SVM method was 87%. However, no self-consistency test result was reported by Chou & Maggiora (1998) for the same dataset.

4.2. SUCCESS RATE OF JACK-KNIFE TEST

We test cross-validation (jack-knife test) to the SVMs method. The jack-knife test is also called leave-one-out test, in which each domain in the dataset is in turn singled out as a tested domain and all the rule parameters are calculated without using this domain. In other words, the structural class of each domain is predicted by the rule derived using all other domains except the one that is being predicted. During the process of jack-knife analysis, both the training dataset and testing dataset are actually open, and a protein will in turn move from each to the other. The overall rates of correct prediction thus obtained for the three structural classes of 138, 253 and 359 domains are 57, 83, and 95%, respectively (Table 1), in contrast to 64, 79, and 84% as reported by Chou & Maggiora (1998). For the dataset of 1601 domains classified into seven classes, the overall rate of correct prediction by the current method was 84% (Table 2), and no jack-knife test result was reported by Chou & Maggiora (1998) for the same dataset.

4.3. SUCCESS RATE OF INDEPENDENT DATASET TEST

Furthermore, the prediction quality was also examined by the independent dataset test. This procedure generally consists of the following two steps: (1) construct a training dataset; (2) construct an independent testing set for which the prediction is performed using the SVMs model trained by the training dataset. Here both the training and testing datasets were taken from Chou & Maggiora (1998). The training dataset for the four structural classes (all-α; all-β; α/β domains: $\alpha+\beta$ domains) from Chou & Maggiora (1998) consists of 225 protein domains (61 all-α, 45 all-β, 56 α/β domains, 63 $\alpha+\beta$ domains), and the corresponding testing dataset contains 510 domains, of which 109 are all-α, 130

TABLE 1

Predicted results for the 138, 253, *and* 359 *domains by jack-knife test**

Dataset	Rate of correct prediction (%)				
domains	All-α domains	All-β domain	α/β domain	$\alpha+\beta$ domain	Overall
138	19/36 = 53	22/29 = 78	14/32 = 44	24/41 = 59	79/138 = 57
253	53/63 = 84	46/58 = 79	50/61 = 82	62/71 = 87	211/253 = 83
359	76/82 = 93	82/85 = 98	94/99 = 95	90/93 = 97	342/359 = 95

*All the domain data in this table were taken from Chou & Maggiora (1998).

TABLE 2

Predicted results for the 1601 *domains by the jack-knife test**

Rate of correct prediction (%)							
All-α	All-β	α/β	$\alpha+\beta$	μ	σ	ρ	Overall
243/273 = 89	401/461 = 87	299/332 = 90	258/297 = 87	12/31 = 39	110/168 = 65	23/39 = 59	1346/1601 = 84

*All the domain data in this table were taken from Chou & Maggiora (1998).

all-β, 135$\alpha+\beta$ and 136 α/β domains. The overall rate of correct prediction reaches $484/510 = 94.9\%$.

The training dataset for the seven structural classes (all-α, all-β, α/β, $\alpha+\beta$, μ, σ, ρ) from Chou & Maggiora (1998) consists of 1601 protein domains (273 all-α, 461 all-β domains, 332α/β domains, 297 $\alpha+\beta$ domains, 31 μ domains, 168 σ domains, 39 ρ domains), and the corresponding testing dataset contains 2438, of which 393 are all-α domains, 704 all-β domains, 509 α/β domains, 608 $\alpha+\beta$ domains, 46 μ domains, 158 σ domains and 20 ρ domains (Bahar *et al.*, 1997). The overall success rate predicted for the 2438 independent domains by the current method using the 1601 domains as a training dataset reached $2304/2438 = 94.5\%$.

Finally, it is instructive to mention that all the datasets used in this paper were taken from Chou & Maggiora (1998). It was clearly described in that paper that all the protein sequences in the datasets were taken from SCOP according to different functional families. Therefore, no redundant sequences exist whatsoever in the datasets. Actually, the sequence matches performed between all members in each class of the datasets indicated that most pairs have very low sequence identity. For example, for the database of 1601 protein domains, the average sequence identities in the all-α, all-β, α/β, $\alpha+\beta$, μ, σ, and ρ classes are only 0.081, 0.087, 0.071, 0.078, 0.082, 0.127, and 0.109, respectively.

5. Conclusion

A comparison of the above results with those reported by Chou & Maggiora (1998) indicates that for small datasets the component-coupled algorithm yielded better results, but for large datasets the SVMs method yielded better results. Nevertheless, both the component-coupled algorithm (Chou *et al.*, 1998) and the SVMs method are much more powerful than the simple geometry prediction algorithms as demonstrated by Chou *et al.*, (1998) and Chou & Maggiora (1998). Accordingly, it is anticipated that the current SVMs methods and the elegant component-coupled algorithm developed by Chou and his co-workers (Chou, 1995; Chou *et al.*, 1998; Chou & Maggiora, 1998; Chou & Zhang, 1994, 1995), if effectively complemented with each other, will become a powerful tool for predicting the structural classes of protein domains. It should be pointed out, however, that using amino acid composition as an input to predict the structural class of protein domains is merely an approximate approach. This is because no sequence order effects are included in the amino acid composition. Unfortunately, as analysed by Chou (1999), if using the domain sequence as an input, we are to face a formidable barrier in formulating a feasible prediction algorithm and constructing a workable training dataset. As a compromise, the amino acid composition has been widely used to deal with these kinds of problems. To improve such an approximate treatment, one promising approach is to use the pseudo-amino acid composition (Chou, 2001), which contains some quasi-sequence order effects and was recently developed by Chou to predict protein subcellular location. And this will certainly further improve the prediction quality of protein structural class as well.

REFERENCES

Bahar, I., Atilgan, A. R., Jernigan, R. L. & Erman, B. (1997). Understanding the recognition of protein structural classes by amino acid composition. *Proteins* **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.

Cai, Y. D., Li, Y. X. & Chou, K. C. (2000). Using neural networks for prediction of domain structural classes. *Biochim. Biophys. Acta.* **1476,** 1–2.

Chou, K. C. (1995). A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins: Struct. Funct. Genet.* **21,** 319–344.

Chou, K. C. (1999). Using pair-coupled amino acid composition to predict protein secondary structure content. *J. Protein Chem.* **18,** 473–480.

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino-acid-composition. *Proteins: Struct. Funct. Genet.* **43,** 246–255 (erratum: *Proteins: Struct. Funct. Genet.* 2001, **44,** 60).

Chou, K. C., Liu, W., Maggiora, G. M. & Zhang, C. T. (1998). Prediction and classification of domain structural classes. *Proteins: Struct. Funct. Genet.* **31,** 97–103.

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

Chou, K. C. & Zhang, C. T. (1994). Predicting protein folding types by distance functions that make allowances for amino acid interactions. *J. Biol. Chem.* **269,** 22014–22020.

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

Cortes, C. & Vapnik, V. (1995). Support vector networks. *Mach. Learning* **20,** 273–293.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In: *Proceedings of the European Conference on Machine Learning*, Springer, Berlin.

Joachims, T. (1999a). Making large-scale SVM learning practical. In: *Advances in Kernel Methods—Support Vector Learning* (Schölkopf, B., Burges, C.J.C. & Smola, A. J., eds), pp. 169–184. Cambridge, MA: MIT Press.

Joachims, T. (1999b). *International Conference on Machine Learning* (*ICML*).

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of protein database for the investigation of sequence and structures. *J. Mol. Biol.* **247,** 536–540.

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

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Berlin: Springer Verlag.

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