Available online at www.sciencedirect.com

SCIENCE DIRECT®

Peptides 24 (2003) 665–669

PEPTIDES

www.elsevier.com/locate/peptides

Prediction of β-turns with learning machines

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

[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] Bioinformation Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200030, China
[d] Department of Computing Science, College of Cardiff, University of Wales, Queens Buildings,
Newport Road, P.O. Box 916, Cardiff CF2 3XF, UK
[e] Gordon Life Science Institute, Kalamazoo, MI 49009, USA

Received 13 January 2003; accepted 8 April 2003

Abstract

The support vector machine approach was introduced to predict the β-turns in proteins. The overall self-consistency rate by the re-substitution test for the training or learning dataset reached 100%. Both the training dataset and independent testing dataset were taken from Chou [J. Pept. Res. 49 (1997) 120]. The success prediction rates by the jackknife test for the β-turn subset of 455 tetrapeptides and non-β-turn subset of 3807 tetrapeptides in the training dataset were 58.1 and 98.4%, respectively. The success rates with the independent dataset test for the β-turn subset of 110 tetrapeptides and non-β-turn subset of 30,231 tetrapeptides were 69.1 and 97.3%, respectively. The results obtained from this study support the conclusion that the residue-coupled effect along a tetrapeptide is important for the formation of a β-turn.

Keywords: Jackknife test; Dataset; Tetrapeptides

1. Introduction

β-Turns are a type of tight turn [8] that consist of four consecutive residues with the distance between $C^{\alpha}(i)$ and $C^{\alpha}(i+3)$ less than 7 Å. However, the tetrapeptides under consideration in this paper do not have α-helical conformation [8,27]. Being the most common type of non-repetitive structure in proteins [22], β-turns have some special importance in structure and function. They are involved in forming highly compacted structure for a protein [24] as well its binding site for ligands [28]. Therefore, prediction of β-turns in protein is very important in both the structural and functional sense. Much effort has been made for predicting β-turns [3,4,7,9,14,15,19,26,31,33]. In this paper, we apply a new learning machine to predict β-turns.

2. Materials and methods

2.1. Support vector machine

Support vector machines (SVMs) are types of learning machines based on statistical learning theory. The most remarkable characteristics of SVMs are the absence of local minima, the sparseness of the solution, and the use of the kernel-induced feature spaces. The basic idea of applying SVMs to pattern classification can be outlined as follows. First, map the input vectors into a feature space (possible with a higher dimension), either linearly or non-linearly, which is relevant to the selection of the kernel function. Then, within the feature space, seek an optimized linear division; i.e. construct a hyper-plane which can separate two classes (this can be extended to multi-classes) with the least error and maximal margin. The SVMs training process always seeks a global optimised 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 [29].

[*] 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).

0196-9781/$ – see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S0196-9781(03)00133-5

SVMs have been used to deal with protein fold recognition [18], protein–protein interaction prediction [1] and protein secondary structure prediction [20].

In this paper, the Vapnik's support vector machine [30] was introduced to predict beta-turns in proteins. Specifically, the SVMlight, which is an implementation (in C language) of SVM for the problems of pattern recognition, was used for computations. The optimization algorithm used in SVMlight is described elsewhere [21]. The relevant mathematical principles can be briefly formulated as follows.

Given a set of N samples, i.e. a series of input vectors:

$$\boldsymbol{X}_k \in \Re^{\boldsymbol{\tau}}, \quad (k = 1, \ldots, N), \tag{1}$$

where $\boldsymbol{X}_k$ can be regarded as the kth sample or vector defined in the relevant space according to the object-studied [12], and $\Re^{\boldsymbol{\tau}}$ is a Euclidean space with $\boldsymbol{\tau}$ dimensions. Since the multi-class identification problem can always be converted into a two-class identification problem, without loss of the generality the formulation below is given for the two-class case only. Suppose the output derived from the learning machine is expressed by $y_k \in \{+1, -1\}(k = 1, \ldots, N)$, where the indexes -1 and $+1$ are used to stand for the two classes concerned, respectively. The goal here is to construct one binary classifier or derive one decision function from the available samples that has a small probability of misclassifying a future sample. Here both the basic linear separable case and the most useful linear non-separable case for most real life problems are taken into consideration.

2.2. The linear separable case

In this case, there exists a separating hyper-plane whose function is $\boldsymbol{W} \cdot \boldsymbol{X} + b = 0$, which implies

$$y_k(\boldsymbol{W} \cdot \boldsymbol{X}_k + b) \geq 1, \quad (k = 1, \ldots, N). \tag{2}$$

By minimizing $1/2\,\|\boldsymbol{W}\|^2$ subject to the above constraint, the SVM approach will find a unique separating hyper-plane. Here $\|\boldsymbol{W}\|^2$ is the Euclidean norm of $\boldsymbol{W}$, which maximizes the distance between the hyper-plane, or the optimal separating hyper-plane (OSH) [16], and the nearest data points of each class. The classifier thus obtained is called the maximal margin classifier. By introducing Lagrange multipliers α_i, and using the Karush–Kuhn–Tucker (KKT) conditions [17,23] as well as the Wolfe dual theorem of optimization theory [32], the SVM training procedure amounts to solving the following convex quadratic programming (QP) problem:

$$\text{Max}: \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N}\alpha_i\alpha_j y_i y_j \boldsymbol{X}_i \cdot \boldsymbol{X}_j, \tag{3}$$

subject to the following two conditions:

$$\alpha_i \geq 0, \quad (i = 1, 2, \ldots, N), \tag{4}$$

$$\sum_{i=1}^{N}\alpha_i y_i = 0. \tag{5}$$

The solution is a unique globally optimized result, which can be expressed with the following expansion:

$$\boldsymbol{W} = \sum_{i=1}^{N} y_i\alpha_i\boldsymbol{X}_i. \tag{6}$$

Only if the corresponding $\alpha_i > 0$, are these $\boldsymbol{X}_i$ called the support vectors. Now suppose $\boldsymbol{X}$ is a query sample defined in the same space as $\boldsymbol{X}_i$ [6,12]. After the SVM has been trained, the decision function for identifying which class the query protein belongs to can be formulated as

$$f(\boldsymbol{X}) = \text{sgn}\left(\sum_{i=1}^{N} y_i\alpha_i\boldsymbol{X} \cdot \boldsymbol{X}_i + b\right), \tag{7}$$

where sgn( ) in the above equation is a sign function, which equals to $+1$ or -1 when its argument is ≥ 0 or <0, respectively.

2.3. The linear non-separable case

For this case two important techniques are needed that are given below, respectively.

2.3.1. The "soft margin" technique

In order to allow for training errors, Cortes and Vapnik [16] introduced the slack variables:

$$\boldsymbol{\xi}_i > 0, \quad (i = 1, \ldots, N), \tag{8}$$

and the relaxed separation constraint given by

$$y_i(\boldsymbol{W} \cdot \boldsymbol{X}_i + b) \geq 1 - \boldsymbol{\xi}_i, \quad (i = 1, \ldots, N). \tag{9}$$

The optimal separating hyper-plane can be found by minimizing

$$\frac{1}{2}\,\|\boldsymbol{W}\|^2 + c\sum_{i=1}^{N}\boldsymbol{\xi}_i, \tag{10}$$

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

2.3.2. The "kernel substitution" technique

The SVM performs a non-linear mapping of the input vectors from the Euclidean space $\Re^d$ into a higher dimensional Hilbert space H, where the mapping is determined by the kernel function. Then, like in the linear separable case, it finds the optimal separating hyper-plane in the Hilbert space H that would correspond to a non-linear boundary in the original Euclidean space. Two typical kernel functions are listed below:

$$K(\boldsymbol{X}_i, \boldsymbol{X}_j) = (\boldsymbol{X}_i \cdot \boldsymbol{X}_j + 1)^{\boldsymbol{\tau}}, \tag{11}$$

$$K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \exp\left(-r\,\|\boldsymbol{X}_i - \boldsymbol{X}_j\|^2\right), \tag{12}$$

where the first one is called the *polynomial kernel function of degree* $\boldsymbol{\tau}$ which will eventually revert to the linear function when $\boldsymbol{\tau} = 1$ the second one is called the radial basic function (RBF) kernel. Finally, for the selected kernel function, the learning task amounts to solving the following quadratic programming problem:

$$\text{Max}: \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N}\alpha_i\alpha_j y_i y_j K(\boldsymbol{X}_i \cdot \boldsymbol{X}_j), \tag{13}$$

subject to:

$$0 \le a_i \le c, \quad (i = 1, 2, \ldots, N), \tag{14}$$

$$\sum_{i=1}^{N}\alpha_i y_i = 0. \tag{15}$$

Accordingly, the form of the decision function is given by

$$f(\boldsymbol{X}) = \text{sgn}\left(\sum_{i=1}^{N} y_i\alpha_i K(\boldsymbol{X}, \boldsymbol{X}_i) + b\right). \tag{16}$$

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

3. The training and prediction of β-turns

The β-turn structure is formed by four consecutive residues, indexed as i, $i + 1$, $i + 2$, and $i + 3$ [7]. Following Chou [8] a tetrapeptide can be generally expressed by R_i R_{i+1} R_{i+2} R_{i+3}, where R_i represents the amino acid at position i, R_{i+1} represents the amino acid at position $i + 1$, and so forth. For the current study, tetrapeptides can be classified into two sets: S^+ consisting of only those tetrapeptides forming β-turns in proteins, and S^- those forming non-β-turns.

The 20 amino acids are coded with 20 digits as given by: A = 1000000000000000000, C = 01000000000000000 00, ... ,Y = 00000000000000000001). Thus, a tetrapeptide can be expressed as a vector or a point in a 4 × 20D = 80D space. The computations were carried out on a Silicon Graphics IRIS Indigo Work Station (Elan 4000).

The training dataset was taken from Chou [7] that contains 4262 tetrapeptide samples of which 455 are of β-turn and 3807 of non-β-turns.

For the SVMs, the width of the Gaussian RBFs was 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 hyper-plane output by the SVMs was obtained. This indicates that the trained model, i.e. the hyper-plane output harboring the relevant important information, has the desired function to identify the β-turns.

4. Results

The demonstration was conducted by the three most typical approaches in statistical prediction [13]; i.e. the re-substitution test, jackknife test, and independent dataset test, as reported below.

4.1. Re-substitution test

The so-called re-substitution test is an examination for the self-consistency of a prediction method. When the re-substitution test is performed for the current study, each tetrapeptide in the dataset concerned is in turn identified using the rule parameters derived from the same data set, the so-called training dataset. The success rate thus obtained for predicting the 455 β-turn tetrapeptides and 3807 non-β-turn tetrapeptides was 100%, indicating that after being trained, the SVMs model has grasped the complicated relationship between the sequence character of the tetrapeptides and their β-turn attribute.

4.2. Jackknife test

During jackknifing, each tetrapeptide in the dataset is in turn singled out as a tested tetrapeptide and all the rule-parameters are calculated based on the remaining tetrapeptides. In other words, the β-turn attribute of each tetrapeptide is identified by the rule parameters derived using all the other tetrapeptides except the one which is being identified. During the process of jackknifing both the training data set and testing data set are actually open, and a tetrapeptide will in turn move from one to the other. The result of jackknife test thus obtained for the 455 β-turn tetrapeptides was 246/455 = 54.1%, and that for the 3807 non-β-turn tetrapeptides was 3746/3807 = 98.4%. The overall success rate is 3992/4262 = 93.6%.

4.3. Independent dataset test

Moreover, as a demonstration of practical application, predictions were also conducted for an independent dataset based on the rule-parameters derived from 455 + 3807 = 4262 tetrapeptides in the training dataset. The independent testing dataset was also taken from Chou [7] that contains 110 β-turn tetrapeptides and 30,229 non-β-turn tetrapeptides. None of these tetrapeptides occurs in the training dataset. The predicted result thus obtained for the 110 β-turn tetrapeptides was 76/110 = 69.2, and that for the 30,231 non-β-turn tetrapeptides was 29,423/30,231 = 97.3%. The overall success rate is 29,423/30,417 = 96.9%.

5. Discussion

As can be seen from the above results, the success rates obtained by the re-substitution test are higher than those by the jackknife test and independent dataset test. Because during the process of the re-substitution test, the rule parameters derived from the training data set include the information of the query tetrapeptide later plugged back in the test. This will certainly underestimate the error and enhance the success rate because the same peptides are used to derive the rule parameters and to test themselves. Accordingly, the success rate obtained by the re-substitution test merely represents the self-consistency of a prediction method, and hence the rate thus derived must bear an optimistic estimation [2,6,10,11,34,35]. Nevertheless, the re-substitution test is absolutely necessary because it reflects the self-consistency of an identification method, especially for its algorithm part. A prediction algorithm certainly cannot be deemed as a good one if its self-consistency is poor. In other words, the re-substitution test is necessary but not sufficient for evaluating an identification method. As a complement, a cross-validation test for an independent testing data set is needed because it can reflect the effectiveness of an identification method in practical application. This is important especially for checking the validity of a training database: whether it contains sufficient information to reflect all the important features concerned so as to yield a high success rate in application.

As is well known, the independent data set test, sub-sampling test and jackknife test are the three methods often used for cross-validation in statistical prediction. Among these three, however, the jackknife test is deemed the most effective and objective one [13,25]. As can be seen from the results, the success rates obtained by the jackknife test are lower than those by the re-substitution test, particularly for the β-turn subset. This is because the current dataset for β-turns is much smaller than the non-β-turn dataset and hence the cluster-tolerant capacity [5] of the former is much lower than that of the latter. As a consequence, the information loss during the jackknife process will have a more negative impact to the prediction of β-turns than non-β-turns. It is anticipated that with the improvement of β-turn subset by adding more newly-found β-turn tetrapeptides into it, the prediction quality for the β-turn subset will be enhanced.

The results obtained from the present study further support the conclusion of previous investigators (see, e.g. Chou [7]) that the formation of β-turns in proteins is considerably correlated with the sequence code of tetrapeptides although the long-distance interaction along an entire protein chain should also be taken into account in order to further improve the prediction quality [8]. The current approach may play a complementary role to the sequence-coded algorithms developed by the previous investigators [7,9,33].

References

[1] Bock JR, Gough DA. Predicting protein–protein interactions from primary structure. Bioinformatics 2001;17:455–60.

[2] Cai YD. Is it a paradox or misinterpretation. Proteins: Structure, Function, and Genetics 2001;43:336–8.

[3] Cai YD, Li YX, Chou KC. Classification and prediction of beta-turn types by neural networks. Adv Eng Software 1999;30:347–52.

[4] Cai YD, Yu H, Chou KC. Prediction of beta-turns. J Protein Chem 1998;17:363–76.

[5] Chou KC. A key driving force in determination of protein structural classes. Biochem Biophys Res Commun 1999;264:216–24.

[6] Chou KC. A novel approach to predicting protein structural classes in a (20-1)-D-amino acid composition space. Proteins: Structure, Function, and Genetics 1995;21:319–44.

[7] Chou KC. Prediction of beta-turns in proteins. J Pept Res 1997;49:120–44.

[8] Chou KC. Review: prediction of tight turns and their types in proteins. Anal Biochem 2000;286:1–16.

[9] Chou KC, Blinn JR. Classification and prediction of β-turn types. J Protein Chem 1997;16:575–95.

[10] Chou KC, Cai YD. Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 2002;277:45765–9.

[11] Chou KC, Elrod DW. Protein subcellular location prediction. Protein Eng 1999;12:107–18.

[12] Chou KC, Zhang CT. Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 1994;269:22014–20.

[13] Chou KC, Zhang CT. Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 1995;30:275–349.

[14] Chou PY, Fasman GD. Conformational parameters for amino acids in helical, beta-sheet and random coil regions calculated from proteins. Biochemistry 1974;13:221–3.

[15] Cohen FE, Abarbanel RM, Kuntz ID, Fletterick RJ. The prediction in proteins using a pattern-matching approach. Biochemistry 1983;25:266–75.

[16] Cortes C, Vapnik V. Support vector networks. Machine Learn 1995;20:273–93.

[17] Cristianini N, Shawe-Taylor J. Support vector machines. Cambridge: Cambridge University Press; 2000.

[18] Ding CH, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 2001;17:349–58.

[19] Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 1978;120:97–120.

[20] Hua SJ, Sun ZR. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol 2001;308:397–407.

[21] Joachims T. Making large-scale SVM learning practical. In: Schölkopf B, Burges CJC, Smola AJ, editors. Advances in kernel methods—support vector learning. Cambridge: MIT Press; 1999. p. 169–84.

[22] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–637.

[23] Karush W. Minima of functions of several variables with inequalities as side constraints. MSc Thesis. Chicago: University of Chicago; 1939.

[24] Lewis PN, Momany FA, Scheraga HA. Chain reversals in proteins. Biochem Biophys Acta 1973;303:211–29.

[25] Mardia KV, Kent JT, Bibby JM. Multivariate Analysis. London: Academic Press; 1979. p. 322 and 381.

[26] McGregor MJ, Flores TP, Sternberg MJE. Prediction of b-turns in proteins using neural networks. Protein Eng 1989;2:521–6.

[27] Richardson JS. The anatomy and taxonomy of protein structure. Adv Protein Chem 1981;34:167–339.

[28] Rose GD, Gierasch LM, Smith JA. Turns in peptides and proteins. Adv Protein Chem 1985;37:1–109.

[29] Vapnik V. Statistical learning theory. New York: Wiley-Interscience; 1998.

[30] Vapnik VN. The nature of statistical learning theory. Berlin: Springer-Verlag; 1995.

[31] Wilmot CM, Thornton JM. Analysis and prediction of the different types of b-turn in proteins. J Mol Biol 1988;203:221–32.

[32] Wolfe P. A duality theorem for nonlinear programming. Q Appl Math 1961;19:239–44.

[33] Zhang CT, Chou KC. Prediction of beta-turns in proteins by 1–4 and 2–3 correlation model. Biopolymers 1997;41:673–702.

[34] Zhou GP, Assa-Munt N. Some insights into protein structural class prediction. Proteins: Structure, Function, and Genetics 2001;44: 57–9.

[35] Zhou GP, Doctor K. Subcellular location prediction of apoptosis proteins. Proteins: Structure, Function, and Genetics 2003;50: 44–8.