PEPTIDES

ELSEVIER

Peptides 23 (2002) 205–208

Support vector machines for predicting the specificity of GalNAc-transferase

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, PO Box 916, Cardiff CF2 3XF, UK*
[d]*Computer-Aided Drug Discovery, Upjohn Laboratories, Kalamazoo, MI 49001-4940, USA*

Received 26 March 2001; accepted 30 July 2001

Abstract

Support Vector Machines (SVMs) which is one kind of learning machines, was applied to predict the specificity of GalNAc-transferase. The examination for the self-consistency and the jackknife test of the SVMs method were tested for the training dataset (305 oligopeptides), the correct rate of self-consistency and jackknife test reaches 100% and 84.9%, respectively. Furthermore, the prediction of the independent testing dataset (30 oligopeptides) was tested, the rate reaches 76.67%. © 2002 Elsevier Science Inc. All rights reserved.

Keywords: Support Vector Machines; GalNAc-transferase; Self-consistency; Jackknife test

1. Introduction

A GalNAc-transferase has an extended active site, which is composed of nine amino acid residues denoted by R4, R3 R2, R1 R0, R1′, R2′, R3′, R4′ [3,5,7]. The central amino acid residue R0 is either Ser or Thr, where the reducing monosaccharide is being anchored (see, e.g. Fig. 1 of ref. 3). To find out what kind of peptides can function as a competitive inhibitor against the enzyme, understanding of the specificity of the enzyme is essential. However, the number of the possible nonapeptides with either Ser or Thr at the central position is very large. It is time-consuming and painful to test so many peptides solely based on experiments. To help reach such a goal, Cai and Chou [2] used the neural network method for predicting the specificity of GalNAc-transferase. In this paper, we apply Vapnik's Support Vector Machine [13] for this problem and good results are obtained.

2. Materials and methods

2.1. Support Vector Machine

Support Vector Machine (SVM) is one kind of learning machines 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, whichever 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 the 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 [12].

SVMs have been used in a range of problems including drug design [1], image recognition and text classification [8].

In this paper, we apply Vapnik's Support Vector Machine [13] for predicting the specificity of GalNAc-transferase. 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

[*] Corresponding author. 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).

PII: S0196-9781(01)00597-6

can be found in [9,10]. The code has been used in text classification, image recognition [8].

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

$$\vec{X}_i \in R^d \, (i = 1,\ldots,N) \, .$$

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

Here -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:

2.1.1. The linear separable case

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

$$Y_i(\vec{W} \bullet \vec{X}_i + b) \geq 1, \quad i = 1,\ldots,N$$

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 hyper plane, i.e., the Optimal Separating Hyperplane (OSH) defined by Cortes and Vapnik [6], and the nearest data points of each class. For the latter reason, the above hyperplane is called the largest margin classifier.

By introducing Lagrange multipliers α_i, the SVM training procedure amounts to solving the following convex QP problem:

$$Max\colon \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \cdot Y_i \cdot Y_j \cdot \vec{X}_i \bullet \vec{X}_j$$

subject to the following two conditions:

$$\alpha_i \geq 0$$

$$\sum_{i=1}^{N} \alpha_i Y_i = 0, i = 1,\ldots,N$$

The solution includes b and a unique globally optimized result which can be shown having the following expansion:

$$\vec{W} = \sum_{i=1}^{N} Y_i \alpha_i \cdot \vec{X}_i$$

Only if the corresponding $\alpha_i > 0$, are these $\vec{X}_i$ called Support Vectors.

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

$$f(\vec{X}) = \text{sgn}\left(\sum_{i=1}^{N} Y_i \cdot \alpha_i \cdot \vec{X} \cdot \vec{X}_i + b \right)$$

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

2.1.2. The linear non-separable case

Two important techniques needed for this case are given respectively as below.

(1) The "soft margin" technique.

In order to allow for training errors, Cortes and Vapnik [6] introduced slack variables:

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

The relaxed separation constraint is given as:

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

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 (2.1.1). Here C is a regularization parameter used to decide a trade-off between the training error and the margin.

(2) The "kernel substitution" technique

SVM performs a nonlinear 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 case (2.1.1), 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:

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

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

Here 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 RBF (Radial Basic Function) kernel.

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

$$Max\colon \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \cdot Y_i Y_j \cdot K(\vec{X}_i \cdot \vec{X}_j)$$

subject to:

$$0 \leq a_i \leq C$$

$$\sum_{i=1}^{N} \alpha_i \cdot Y_i = 0, i = 1,\ldots,N$$

And the form of the decision function is

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

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

2.2. The training and prediction of the specificity of GalNAc-transferase

A GalNAc-transferase has an extended active site [3,11]. For studying the specificity of the enzyme, oligopeptides can be classified into two categories: the positive set and negative set. The positive set consists of those which can be glycosylated by the enzyme; while the negative set consists of those which cannot be glycosylated.

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

Given a nonapeptide, its assignment to the positive set or the negative set can be formulated by a 9-D (dimension) vector. In this research, 20 bases of oligopeptides are coded as 20-D vectors composed of only 0 and 1 (A = 100000. . . 000, C = 010000. . . 000, ... Y = 000000. . . 001).

The training data set was taken from Chou [3]. It contains 305 samples, of which 195 are "positive" samples and 110 "negative" samples. For the SVMs, the width of the Gaussian RBFs [12] was selected as that which minimized an estimate of the VC-dimension [12]. The parameter C that controls the error-margin trade-off was set at 100. After being trained, the hyperplane output by the SVMs was obtained. This indicates that the trained model, i.e., hyperplane output which is including the important information, gives the function to identify the glycosylation.

The independent testing dataset was also taken from Chou [3]. It contains 30 samples, of which 26 are "positive" samples and 4 "negative" samples.

3. Results

In this research, the examination for the self-consistency of the SVMs method was tested. The rates of correct prediction for the two classes reaches 100%.

And the results by the jackknife test indicates that the rate of correct prediction for the two classes of 305 oligopeptides is 258/305 = 84.9%.

Furthermore, in order to test the performance of the established model, 30 testing samples are recognized. As a result, the correct predicting rate reaches 23/30 = 76.67%.

4. Discussion

4.1. The success rate of self-consistency and prediction

In this study, the rate of self-consistency reaches 100%. This indicates that after being trained, the hyperplane output of the SVMs has grasped the complicated relationship between the oligopeptides and glycosylation, and it can be used to predict the unknown oligopeptides. And the rate of prediction reaches 76.67%, which further indicates that the trained SVM model is really successful.

4.2. The success rate of jackknife test

In this study, the rate of jackknife test reach 84.9%. As is well known, the single-test-set analysis, sub-sampling 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. See Chou and Zhang [4] for a comprehensive discussion about this. The jackknife test is also called leave-one-out test, in which each oligopeptide in the dataset is in turn singled out as a tested oligopeptide and all the rule-parameters are calculated without using this oligopeptide. In other words, the glycosylation of each oligopeptide is predicted by the rules derived using all other oligopeptides except the one which is being predicted. During the process of jackknife analysis both the training dataset and testing dataset are actually open, and a oligopeptide will in turn move from each to other. Therefore the high rate of jackknife test indicates the good performance of SVMs.

The results obtained through this study indicate that the SVM method, the sequence-coupled vector-projection method [3], and the neural network method [2], if complemented with each other, will become a powerful tool for predicting the specificity of GalNAc-transferase. Such a powerful tool might be very useful for rational drug design because understanding the specificity of the GalNAc-transferase is vitally important for finding effective inhibitors against the enzyme.

References

[1] Burbidge R, Matthew T, Sean H, Bernard B. Drug design by machine learning: Support Vector Machine for pharmaceutical data analysis. Proceedings of the AISB'00 Symposium on Artificial Intelligence in Bioinformatics. 2000. p. 1–4.

[2] Cai YD, Chou KC. Artificial neural network model for predicting the specificity of GalNAc-transferase. Anal Biochem 1996;243(2):284–5.

[3] Chou KC. A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase. Protein Science 1995;4:1365–84.

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

[5] Chou KC, Zhang CT, Kezdy FJ, Poorman RA. A vector projection method for predicting the specificity of GalNAc-transferase. Proteins: Structure, Function, and Genetics 1995;21:118–26.

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

[7] Elhammer AP, Poorman RA, Brown E, Maggiora LL, Hoogerheide JG, Kezdy FJ. The specificity of UDP-GalNAc:polypeptide N-acetylgalactosaminyl transferase as inferred from a database of in vivo substrates and from the in vitro glycosylation of proteins and peptides. J Biol Chem 1993;268:10029–38.

[8] Joachims T. Text categorization with Support Vector Machines: learning with many relevant features. Proceedings of the European Conference on Machine Learning. Springer, 1998.

[9] Joachims T. 11 in: Making large-scale SVM learning practical. Advances in kernel methods—support vector learning. Schölkopf B, Burges C, Smola A, editors. MIT Press, 1999.

[10] Joachims T. Transductive inference for text classification using support vector machines. International Conference on Machine Learning (ICML). 1999.

[11] Schechter I, Berger A. On the size of the active site in proteases. I. Papain Biochem Biophys Res Commun 1967;27:157–62.

[12] Vapnik V. Statistical learning theory. New York: Wiley-Interscience, 1998.

[13] Vapnik V. The nature of statistical learning theory. Springer, 1995.