Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

PEPTIDES

Peptides 24 (2003) 629–630

www.elsevier.com/locate/peptides

Short communication

Support Vector Machine for predicting α-turn types

Yu-Dong Cai [a,*], Kai-Yan Feng [b], Yi-Xue Li [c], Kuo-Chen Chou [d]

[a] *Shanghai Research Centre of Biotechnology, Chinese Academy of Sciences, Shanghai 200233, China*
[b] *Imaging Science and Biomedical Engineering, Medical School, Stopford Building, Oxford Road, University of Manchester, Manchester M13 9PT, UK*
[c] *Bioinformation Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200030, China*
[d] *Gordon Life Science Institute, Kalamazoo, MI 49009, USA*

Received 18 October 2002; accepted 5 February 2003

Abstract

Tight turns play an important role in globular proteins from both the structural and functional points of view. Of tight turns, β-turns and γ-turns have been extensively studied, but α-turns were little investigated. Recently, a systematic search for α-turns classified α-turns into nine different types according to their backbone trajectory features. In this paper, Support Vector Machines (SVMs), a new machine learning method, is proposed for predicting the α-turn types in proteins. The high rates of correct prediction imply that that the formation of different α-turn types is evidently correlated with the sequence of a pentapeptide, and hence can be approximately predicted based on the sequence information of the pentapeptide alone, although the incorporation of its interaction with the other part of a protein, the so-called "long distance interaction", will further improve the prediction quality.

Keywords: Long distance interaction; Support Vector Machine; Tight turns

1. Introduction

Like α-helices and β-sheets in proteins, tight turns also play a very important role from both the structural and functional points of view (see, e.g. a recent review by Chou [8], where a rigorous definition for each of different types of tight turns has been systematically given). Among the tight turns, less work has been done on α-turns in comparison with β-turns and γ-turns due to lower occurrence in protein. Based on the work in predicting β-turns [7,17] and β-turns types [9], Chou [6] proposed a new sequence-coupled model for prediction of the α-turns in proteins. According to Chou's study, the prediction quality is significantly improved in comparison with the prediction results reported previously. Subsequently, Cai and Chou [1] used neural network to predict α-turns. In this paper, we applied Vapnik's Support Vector Machine (SVM) [16] to approach this problem, and good results are obtained.

2. Materials and methods

Support Vector Machine is a machine learning paradigm based on the statistical learning theory. The theory and algorithms SVM can be found in Refs. [12,15,16].

Up to now, SVM has been applied to many biology areas, such as splicing sites prediction in eukaryotic RNA [13], gene expression data analysis [14], prediction of protein subcellular location [4,10], structural classes [2] and the prediction of the specificity of GalNAc-transferase [3].

Following the same procedures and rationale as given by Chou [6], the α-turn types were clustered into 10 categories, i.e. type I-α-RS, type I-α-LS, type II-α-RS, type II-α-LS, type I-α-RU, type I-α-LU, type II-α-RU, type II-α-LU, type I-α-C and non-α-turn.

Since the α-turn structure is a pentapeptide that contains five consecutive residues (i), $(i + 1)$, $(i + 2)$, $(i + 3)$ and $(i + 4)$, its sequence can be represented as: $R(i)$, $R(i + 1)$, $R(i+2)$, $R(i+3)$, $R(i+4)$, where $R(i)$ represents the amino acid residue at the protein sequence position (i), $R(i+1)$ the amino acid residue at the protein sequence position $(i + 1)$, and so forth.

In this research, 20 bases of pentapeptides are coded as 20-D vectors composed of only 0 and 1 ($A = 100000\ldots$

[*] Corresponding author. Present address: Department of Biomolecular Sciences, UMIST, P.O. Box 88, Manchester, M60 1QD, UK.
Tel.: +44-161-200-4191; fax: +44-161-236-0409.
E-mail address: ycai@umist.ac.uk (Y.-D. Cai).

0196-9781/03/$ – see front matter © 2003 Elsevier Science Inc. All rights reserved.
doi:10.1016/S0196-9781(03)00100-1

000, $C = 010000\ldots000, \ldots, Y = 000000\ldots001$), which are taken as the input of the SVMs.

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

The training dataset was taken from Chou [6] that contains 25,416 pentapeptides, of which 238 are type I-α-RS α-turns, 5 type I-α-LS α-turns, 39 type II-α-RS α-turns, 8 type II-α-LS α-turns, 28 type I-α-RU α-turns, 14 type I-α-LU α-turns, 9 type II-α-RU α-turns, 8 type II-α-LU α-turns, 7 type I-α-C α-turns and 25,060 non-α-turns. In this research, for the SVMs, the width of the Gaussian RBFs is selected as that which minimised 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 SVMs was obtained. This indicates that the trained model, i.e. hyperplane output which is including the important information, has the function to identify the α-turns.

In this research, we predict α-turns and their types from the entire primary sequence of Rhe (114 residues, 110 pentapeptides). As a result, the predicting rate is quite promising.

3. Results

First, the method was tested by the re-substitution operation. The so-called re-substitution test is an examination for the self-consistency of a prediction method [5,11,18]. When the re-substitution test is performed for the current study, the α-turn type of each pentapeptide in the dataset is in turn predicted using the rule parameters derived from the same dataset, the so-called training dataset. The rates of correct prediction, thus, obtained reaches 100% (I-α-RS), 100% (I-α-LS), 94.9% (II-α-RS), 87.5% (II-α-LS), 100% (I-α-RU), 100% (I-α-LU), 100% (II-α-RU), 100% (II-α-LU), 100%(I-α-C) and 100% (non-α-turn). This indicates that after being trained, the hyperplanes output of the SVMs has grasped the complicated relationship between the pentapeptides and α-turns, and it can be used to predict the unknown pentapeptides.

As a demonstration to show how to use the current method to predict the α-turn types in proteins, we took Rhe as an example. Rhe contains 110 residues, and its primary sequence is given below [6,8]:

ESVLTQPPSASGTPGQRVTISCTGSATDIGSNSVIWY-
QQVPGKAPKLLIYYNDLLPSGV
SDRFSASKSGTSASLAISGLESEDEADYYCAAWNDS-
LDEPGFGGGTKLTVLGQPK

Along with the sequence, $114 - 5 + 1 = 110$ pentapeptides were automatically extracted in succession, and predicted by the SVMs. As a result, 110 pentapeptides are totally correctly predicted, which is slightly higher than 98.2% which is the result obtained by the neural network method [1].

4. Discussion

The above results obtained by the SVMs indicates that the formation of different α-turn types or non-α-turns is considerably correlated with the sequence of a pentapeptide, fully consistent with the earlier reports by the different approaches [1,6,8].

References

[1] Cai YD, Chou KC. Artificial neural network for predicting alpha-turn types. Anal Biochem 1999;268:407–9.

[2] Cai YD, Liu XJ, Xu XB, Chou KC. Prediction of protein structural classes by support vector machines. Comput Chem 2002;26:293–6.

[3] Cai YD, Liu XJ, Xu XB, Chou KC. Support vector machines for predicting the specificity of GalNAc-transferase. Peptides 2002;23:205–8.

[4] Cai YD, Liu XJ, Xu XB, Chou KC. Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. J Cell Biochem 2002;84:343–8.

[5] Chou KC. A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space. Proteins: Struct Funct Genet 1995;21:319–44.

[6] Chou KC. Prediction and classification of alpha-turn types. Biopolymers 1997;42:837–53.

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

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

[13] Sun YF, Fan XD, Li YD. Identifying splicing sites in eukaryotic RNA: support vector machine approach. Comput Biol Med 2003;33:17–29.

[14] Valentini G. Gene expression data analysis of human lymphoma using support vector machines and output coding ensembles. Artif Intell Med 2002;26:281–304.

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

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

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

[18] Zhou GP. An intriguing controversy over protein structural class prediction. J Protein Chem 1998;17:729–38.