Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computational Biology and Chemistry

Computational Biology and Chemistry 29 (2005) 95–100

www.elsevier.com/locate/compbiolchem

Prediction by support vector machines and analysis by *Z*-score of poly-L-proline type II conformation based on local sequence

Ming-Lei Wang[a,b,*], Hui Yao[c], Wen-Bo Xu[c]

[a] *Laboratory of Bioinformatics, The Key Laboratory of Industrial Biotechnology, Ministry of Education, Southern Yangtze University, Wuxi 214036, China*
[b] *School of Biotechnology, Southern Yangtze University, Wuxi 214036, China*
[c] *School of Information Technology, Southern Yangtze University, Wuxi 214036, China*

Received 15 September 2004; received in revised form 8 January 2005; accepted 18 February 2005

Abstract

In recent years, the poly-L-proline type II (PPII) conformation has gained more and more importance. This structure plays vital roles in many biological processes. But few studies have been made to predict PPII secondary structures computationally. The support vector machine (SVM) represents a new approach to supervised pattern classification and has been successfully applied to a wide range of pattern recognition problems. In this paper, we present a SVM prediction method of PPII conformation based on local sequence. The overall accuracy for both the independent testing set and estimate of jackknife testing reached approximately 70%. Matthew's correlation coefficient (MCC) could reach 0.4. By comparing the results of training and testing datasets with different sequence identities, we suggest that the performance of this method correlates with the sequence identity of dataset. The parameter of SVM kernel function was an important factor to the performance of this method. The propensities of residues located at different positions were also analyzed. By computing *Z*-scores, we found that P and G were the two most important residues to PPII structure conformation.

Keywords: Poly-L-proline type II; Support vector machine; Local sequence; *Z*-score; Protein structure

1. Introduction

The poly-L-proline is assumed to adopt basically two different helical conformations, i.e. type I and type II polyproline. Type I poly-L-proline is a right-handed helix with an axial translation of 1.90 Å composed of 3.3 prolyl residues per turn, linked by *cis*-amide bonds and adopting backbone dihedral angles of $(\varphi, \psi, \omega) = (-83^\circ, +158^\circ, 0^\circ)$ (Traub and Shmueli, 1963). In theory, type I poly-L-proline is possible, but was never detected in nature. Type II poly-L-proline is a left-handed helix with an axial translation of 3.20 Å composed of three prolyl residues per turn, joined by transpeptide bonds with backbone dihedral angles of $(\varphi, \psi, \omega) = (-78^\circ, +149^\circ, 180^\circ)$ (Bochicchio and Tamburro, 2002).

The poly-L-proline type II (PPII) conformation used to be considered a relatively rare and apparently uninteresting secondary structure. In recent years, however, it has become known as surprisingly common and of the utmost importance. This structure plays vital roles in processes such as signal transduction, transcription, cell motility, and the immune response. PPII helices are major features of collagens (Pauling and Corey, 1951) and plant cell wall proteins (Ferris et al., 2001). Proline-rich ligands of the cytoskeletal protein profiling (Mahoney et al., 1997), as well as those of the SH3, WW, and EVH1 protein interaction domains, are bound in this conformation (Kay et al., 2000). The peptide ligands of class II MHC molecules are also bound in the PPII conformation (Jardetzky et al., 1996). The PPII helix is believed to be the dominant conformation for many proline-rich regions of sequence (PRRs) (Williamson, 1994). Sequences not rich in proline, such as poly(lysine), poly(glutamate), and poly(aspartate) peptides, can also adopt this conformation (Woody, 1992). Around 2% of all residues in known protein structures are found in PPII helices at least four residues long (Adzhubei and Sternberg, 1993; Stapley and Creamer,

[*] Corresponding author. Tel.: +86 510 5880679; fax: +86 510 5869645.
E-mail address: wml_yh@yahoo.com.cn (M.-L. Wang).

1476-9271/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2005.02.002

1999). As many as 10% of all residues are found in the PPII conformation, although not necessarily as part of PPII helices (Sreerama and Woody, 1994). PPII helices have also been hypothesized to be a major component of a protein at its denatured states, giving them a role in a most fundamental process (Wilson et al., 1996; Tiffany and Krimm, 1968; Krimm and Tiffany, 1974; Kelly et al., 2001).

Information of such important conformation cannot be derived directly from amino acid sequences. Numerous studies on PPII conformation were reported, most of which were laboratory works. Few attempts have been made to predict PPII secondary structures computationally. Siermala et al. (2000, 2001, 2003) developed a method on the basis of feed-forward multilayer neural networks with the back propagation learning algorithm to predict PPII and investigated the preprocessing and postprocessing of neural networks prediction.

In this paper, we tried to apply the support vector machine (SVM) to reveal the hidden correlation between PPII and local sequence. The SVM method, initially proposed by Vapnik (1995), is a very effective method for general-purpose pattern recognition. It is a learning system that uses a hypothetical space of linear functions in a high dimensional feature space trained with a learning algorithm based on an optimization theory implementing a learning bias derived from statistical learning. Intuitively, the SVM method learns the boundary between samples belonging to two classes by mapping the input samples into a high dimensional space, and seeking a separating hyper-plane in this space (see Fig. 1). This hyper-plane, termed optimal separating hyper-plane (OSH), is chosen in a way to maximize its distance from the closest training samples. As a supervised machine learning technology, the SVM approach is attractive because it is based on an extremely well-developed statistical learning theory (SLT) and has superior performance in practical applications (Vapnik, 1995, 1998). It has been widely used in biological fields, especially in prediction of protein structure (Cai et al., 2000, 2002a,b, 2003; Ding and Dubchak, 2001; Hua and Sun, 2001a,b; Zavaljevski et al., 2002; Sun et al., 2003; Kim and Park, 2004; Wang et al., 2004).

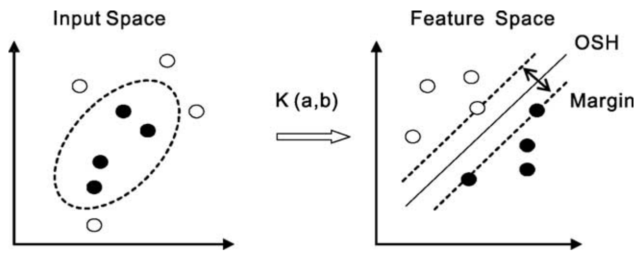

Fig. 1. Two classes denoted by circles and disks, respectively, are linear non-separable in the input space. SVM constructs the optimal separating hyperplane (OSH) (continuous line) which maximizes the margin between two classes by mapping the input space into a high dimensional space, the feature space. The mapping is determined by a kernel function. Support vectors are the circle and disks crossed by the broken lines.

2. Materials and methods

2.1. PDB List

The Protein Data Bank (PDB) (Berman et al., 2000) code list was used in this work, which was provided by a protein sequence culling server called PISCES (http://www.fccc.edu/research/labs/dunbrack/pisces) (Wang and Dunbrack, 2003). All structures in the list had a resolution better than 2.5 Å. Sequence identity between each pair of the sequences in the list was less than 25%. The *R*-factor was less than 0.25. The list was generated on 2 January 2004. The number of chains in each list was 2567.

2.2. Localization of PPII structures

The DSSP method (Kabsch and Sander, 1983) was employed to compute the secondary structures of the PDB files consistently. In this paper, we employed the method of Adzhubei and Sternberg (1993) and Siermala et al. (2001) to localize the PPII structures. After various experiments, the local sequence of 13-residue length is appropriate (Siermala et al., 2001). In order to choose local sequences for SVM, we used the windowing technique 1 described by Siermala et al. (2001). The local sequence was considered in the PPII class when the middlemost position, i.e. the seventh position, of the window was one position in the PPII structure (Fig. 2). Finally, from the PDB list with sequence identity less than 25%, we gained 10,728 local sequences, which were considered in the PPII class, and 561,006 local sequences, which were considered in the non-PPII class, respectively. (The list and local sequences are available by E-mail.)

2.3. Training and testing data sets

In this research, 20 residues were coded as 20-D vectors composed of only 0 and 1 ($A=100000\cdots000$, $C=010000\cdots000$, $\cdots$, $Y=000000\cdots001$). So each 13-residue local sequence was denoted by a vector of 260 bits. 1 and -1 denoted the PPII class and non-PPII class,

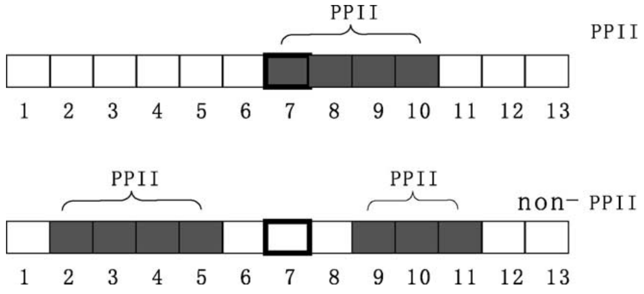

Fig. 2. The grey positions indicate PPII structures. This windowing technique accepts a local sequence of the exact window of 13-residue length in the PPII class if the local sequence's middlemost position, i.e. the seventh position, was one position in the PPII structure.