Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Biology and Chemistry 29 (2005) 95–100
Computational Biology and Chemistry
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 (φ, ψ, ω) = (−83°, +158°, 0°) (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 (φ, ψ, ω) = (−78°, +149°, 180°) (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).

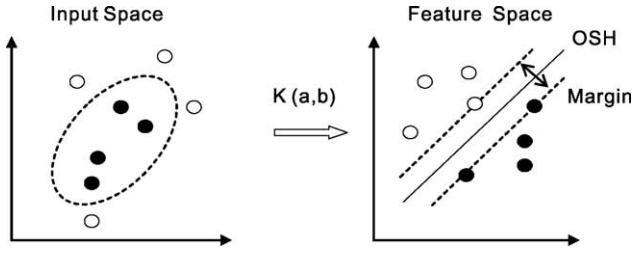

Fig. 1. Two classes denoted by circles and disks, respectively, are linear non-separable in the input space. SVM constructs the optimal separating hyperplane (OSH) (continuous line) which maximizes the margin between two classes by mapping the input space into a high dimensional space, the feature space. 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,

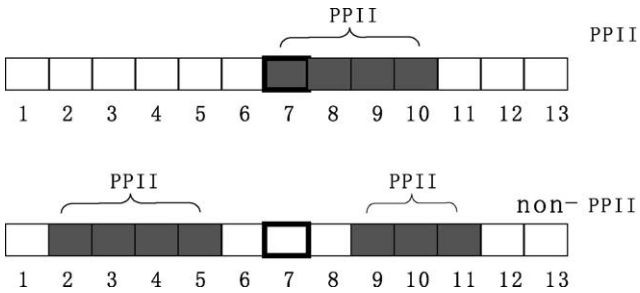

Fig. 2. The grey positions indicate PPII structures. This windowing technique accepts a local sequence of the exact window of 13-residue length in the PPII class if the local sequence's middlemost position, i.e. the seventh position, was one position in the PPII structure.

Table 1
The numbers of the local sequences in the training and testing data set with sequence identity less than 25%

Training set	
PPII class	7152
Non-PPII class	7152
Total	14304
Testing set	
PPII class	3576
Non-PPII class	3576
Total	7152

respectively. Because the non-PPII class local sequences were much more numerous than the PPII class local sequences, the non-PPII class local sequences, whose number was equal to that of the PPII class local sequences, were randomly chosen from all. The final numbers of the local sequences in the training and testing data sets are summarized in Table 1.

2.4. Implementation of SVM

We downloaded the SVM^{light}, (ftp://ftp-ai.cs.uni-dortmund.de/pub/Users/thorsten/svm_light/current/svm_light_windows.zip), which was an implementation of Vapnik's support vector machine for the problem of pattern recognition, for the problem of regression, and for the problem of learning a ranking function (Joachims, 1999). To set a kernel on constructing one SVM binary classifier for PPII/non-PPII class local sequences, we selected the polynomial kernel function to train the SVM. The polynomial kernel function was defined as $K(\vec{a}, \vec{b}) = (s\vec{a} \times \vec{b} + c)^d$ with the parameters s, c, the default value in SVM^{light}, $d = 2$–8.

2.5. Prediction system assessment

To measure the performance of the SVM classifier, we defined four numbers, first:

TP: the number of local sequences observed PPII class, predicted PPII class (true positive);
TN: the number of local sequences observed non-PPII class, predicted non-PPII class (true negative);
FP: the number of local sequences observed non-PPII class, predicted PPII class (false positive);
FN: the number of local sequences observed PPII class, predicted non-PPII class (false negative).

We can measure the performance by using sensitivity, specificity, total accuracy, and Matthew's correlation coefficient (MCC), which can provide a better summary of performance in this case (Matthews, 1975; Baldi et al., 2000).

$$\text{Specificity} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Total accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

The jackknife/leave-one-out procedure is an objective and rigorous testing procedure, but it is also very time-consuming, so we set parameter $x = 1$ in SVM^{light} to efficiently compute jackknife/leave-one-out estimates of the error rate, the sensitivity, and the specificity.

This estimator is based on the general leave-one-out method, but requires an-order-of-magnitude-less computation time due to particular properties of the SVM. In particular, it does not require actually performing re-sampling and retraining, but can be applied directly after training the learner on the training set (Joachims, 2000).

2.6. Computing Z-score

To investigate the correlation between the PPII structure formation and the different residue at different position around the structure, Z-score was defined as:

$$Z_i(a) = \frac{n_i(a) - Np_i(a)}{\sigma_i(a)}$$

$$\sigma_i(a) = \sqrt{Np_i(a)[1 - p_i(a)]}$$

where $n_i(a)$ stood for the number of times residue a is located at position i; N stood for the total number of PPII class local sequences; $p_i(a)$ stood for the probability residue a that is located at position i in all local sequences including PPII and non-PPII class. $\sigma_i(a)$ stood for the standard deviation. If $Z_i(a) > 0$, the residue a at position i is in favor of PPII structure formation, if $Z_i(a) < 0$, unfavorable. If $Z_i(a) = 0$, the correlation is not statistically significant.

3. Results

3.1. Prediction accuracy of SVM classifier

Success rates of correct prediction of the SVM classifiers with different parameter d for the independent testing dataset are depicted in Table 2. The jackknife estimates of the total

Table 2
Results of prediction for the independent testing dataset with sequence identity <25% and different SVM^{light} parameter d

d	2	3	4	5	6	7	8
Specificity (%)	69.84	71.11	73.19	74.19	74.54	74.53	74.51
Sensitivity (%)	66.36	65.74	64.07	61.97	59.54	58.11	58.98
Total accuracy (%)	68.85	69.52	70.30	70.20	69.60	69.13	68.41

Table 3
Estimates of jackknife testing with sequence identity <25% and different SVM^{light} parameter *d*

d	2	3	4	5	6	7	8
Specificity (%)	70.44	71.66	73.18	74.18	74.75	74.51	73.18
Sensitivity (%)	66.89	65.25	63.59	60.95	59.31	58.53	56.35
Total accuracy (%)	69.41	69.72	70.14	69.87	69.64	69.25	67.85

Table 4
MCC for the independent testing dataset with sequence identity <25% and different SVM^{light} parameter *d*

d	MCC
2	0.377
3	0.391
4	0.409
5	0.410
6	0.400
7	0.392
8	0.380

accuracy, the sensitivity, and the specificity are depicted in Table 3. MCC for the independent testing dataset different SVM^{light} parameter *d* is depicted in Table 4.

3.2. Z-score

The distribution of the *Z*-scores for each amino acid as a function of its position in the PPII segments in the downloaded PDB list is reported in Fig. 3. Most of the values of *Z*-scores were very close to each other, so most curves in the figure overlapped each other. All the data could be found in supplementary material.

3.3. Comparison with prediction results of a profile hidden Markov model

We also set up a profile hidden Markov model of PPII structure by the HMMER 2.3.1 package (Eddy, 1998). The aligned PPII local sequences were used to build a model for global alignment using the 'hmmbuild' program in the HMMER package. It was critical to tune the architecture prior parameter since the default setting failed to give a model with correct PPII structures. Those details would be described in another paper. This method was also based on local sequences. The prediction results for the dataset with sequence identity less than 25% are summarized in Table 5. From those results, it was obvious that the SVM method performance was better than the performance of profile hidden Markov model.

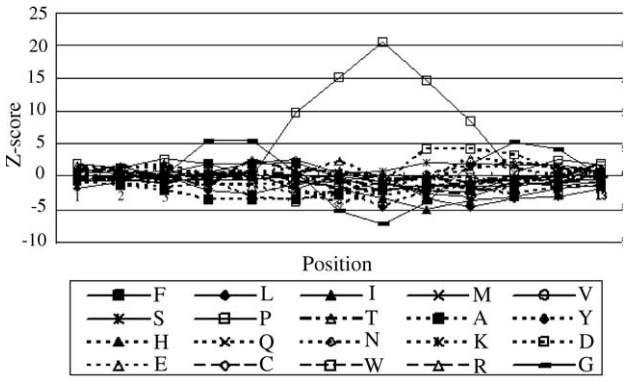

Fig. 3. *Z*-scores computed of different residues at different positions, respectively, with the sequences identity less than 25%.

4. Discussion

PPII conformation is a type of important but rare secondary structure of proteins. Although increasing amount of research work on PPII has been reported, no accurate method for predicting PPII segments was published. In this study, we present a novel prediction method by SVM. Total accuracy for the independent testing set and estimate of jackknife testing both reached approximately 70%. MCC could reach about 0.4. The parameters of kernel function obviously influenced the final performance of the SVM method. The best results were obtained with SVM^{light} parameter $d = 4$. The results obtained through this study indicated that the SVM method would become a powerful tool for predicting PPII conformation. If additional information was added to SVM or more appropriate kernel function was adopted, combined with other methods, the performance could be better.

It is difficult to compare our method with the neural networks method by Siermala et al. because we cannot obtain some important parameters of their neural networks, such as the weights and biases. Their results appeared slightly better than ours, but their datasets were much smaller than ours. The average sequence identity of their sequences was 30–40%.

Table 5
The prediction results for the dataset with sequence identity <25% by using the profile hidden Markov model of PPII structure

	Architecture prior parameter											
	0.90				0.95				0.99			
	Sp[a] (%)	Se[a] (%)	Total accuracy (%)	MCC	Sp (%)	Se (%)	Total accuracy (%)	MCC	Sp (%)	Se (%)	Total accuracy (%)	MCC
Cut-off score												
−7	73.9	29.1	59.4	0.236	73.7	28.8	59.3	0.233	77.6	26.6	59.5	0.251
−8	64.8	55.3	62.6	0.255	66.1	52.6	62.8	0.261	68.5	50.7	63.7	0.284
−9	54.8	88.4	57.8	0.197	56.3	82.7	59.3	0.209	58.2	77.9	60.9	0.231

[a] Sp stands for specificity; Se stands for sensitivity.

Table 6
Z-scores of residues, which were beyond the low/high threshold (−5/+5)

Favor of PPII		Not favor of PPII
G_4(5.378)	P_7(15.157)	G_7(−5.367)
G_5(5.400)	P_8(20.423)	G_8(−7.252)
G_{11}(5.196)	P_9(14.738)	G_9(−5.060)
P_6(9.665)	P_{10}(8.436)	

By convention, $X_i(Z)$ stood for the residue X at position i. The value in the brackets was the computed *Z*-score.

Even the high threshold of 65% was applied to obtain maximum amount of data in their research (Siermala et al., 2001). And the sequence identity between each pair of the sequences in our datasets was less than 25%.

To describe the correlation between residues and PPII structure, *Z*-scores were computed in this study. Similar work was done by Siermala et al. (2001). They found that amino acids G, H, L, N, P, S, V, and Y were prevalent in the PPII structures, whereas G was under-represented by scrutinizing frequencies of different amino acids in their selected data. By spectrum of neural network, they again noticed that amino acids G, D, N, Y, and W were under-represented in PPII structures (Siermala et al., 2003).

By convention, $X_i(Z)$ stood for residue X at position i. The value in the bracket was the computed *Z*-score. From Fig. 3 of *Z*-score distribution, if the low/high threshold of *Z*-score was simply extended −5/+5, PPII-forming propensity of these residues could be found clearly. The results were depicted in Table 6. *Z*-score reflected the influence of the residue abundance on PPII formation. From the computed results, P and G were the two most important residues to PPII structure. If the 4th position was G and the 6th–10th positions were Ps, the PPII-forming propensity became high. If the seventh and eighth positions were Gs, then the local sequence would generally not be in a PPII conformation. So P was still very important to PPII structure though not indispensable. This was consistent with previous work of Rucker et al. (2003). P has the highest measured PPII-forming propensity. G propensity was correlated with its location. PPII-forming propensity of G was relatively high when located the immediately preceding position of PPII and low when located the middle position. This was not entirely consistent with the result obtained by Rucker et al. (2003). But we are not able to scan the datasets simply against a PROSITE-like motif such as XXXGXPPPPP(P, D)(P, D) or XXXGXPPPPP(P, D)(P, D, E) since *Z*-score simply reflects PPII-forming propensity of the residues at different positions and does not indicates the existence of some motif which consists of those residues with the highest *Z*-score.

We also computed *Z*-scores of different residues at different positions, respectively, with the predicted PPII and non-PPII structure sequences by SVM. It was found that the distribution of the *Z*-scores (Fig. 4) was similar to that of *Z*-scores computed with all the sequences in the list (Fig. 3). This suggested that the SVM method had grasped the character of the PPII class local sequences.

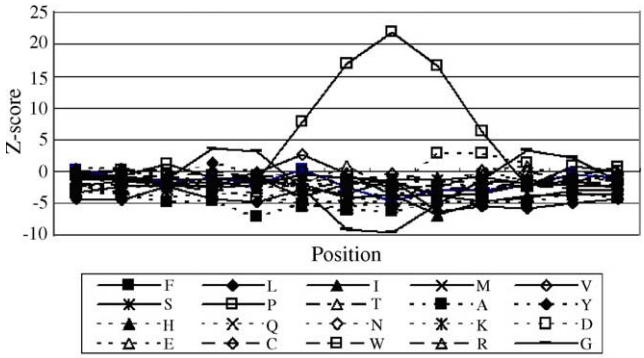

Fig. 4. *Z*-scores computed of different residues at different positions, respectively, with the predicted PPII and non-PPII structure sequences by SVM.

5. Summary

We presented here a SVM prediction method of poly-L-proline type II conformation based on local sequence. The parameter of SVM kernel function was important to the performance of the method. The propensities of residues located at different positions were also analyzed. By computing *Z*-score, we found that P and G were the two most important residues to PPII structure conformation. Our data are useful in future studies of theoretical prediction of protein structures.

Acknowledgments

We thank the anonymous referees for their useful comments to this paper. We are grateful to Dr. Min Fang for some language corrections.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.compbiolchem.2005.02.002.

References

Adzhubei, A.A., Sternberg, M.J.E., 1993. Left-handed polyproline II helices commonly occur in globular proteins. J. Mol. Biol. 229, 472–493.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16, 412–424.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242.

Bochicchio, B., Tamburro, A.M., 2002. Polyproline II structure in proteins: identification by chiroptical spectroscopies, stability, and functions. Chirality 14, 782–792.

Cai, Y.D., Lin, S.L., Chou, K.C., 2003. Support vector machines for prediction of protein signal sequences and their cleavage sites. Peptides 24, 159–161.

Cai, Y.D., Liu, X.J., Xu, X.B., Chou, K.C., 2000. Support vector machines for prediction of protein subcellular location. Mol. Cell Biol. Res. Co. 4, 230–233.

Cai, Y.D., Liu, X.J., Xu, X.B., Chou, K.C., 2002a. Prediction of protein structural classes by support vector machines. Comput. Chem. 26, 293–296.

Cai, Y.D., Liu, X.J., Xu, X.B., Chou, K.C., 2002b. Support vector machines for predicting the specificity of GalNActransferase. Peptides 23, 205–208.

Ding, C.H.Q., Dubchak, I., 2001. Multiclass protein fold recognization using support vector machines and neural networks. Bioinformatics 17, 349–358.

Eddy, S.R., 1998. Profile hidden Markov models. Bioinformatics 14, 755–763.

Ferris, P.J., Woessner, J.P., Waffenschmidt, S., Kilz, S., Drees, J., Goodenough, U.W., 2001. Glycosylated polyproline II rods with kinks as a structural motif in plant hydroxyproline-rich glycoproteins. Biochemistry 40, 2978–2987.

Hua, S.J., Sun, Z.R., 2001a. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J. Mol. Biol. 308, 397–407.

Hua, S.J., Sun, Z.R., 2001b. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17, 721–728.

Jardetzky, T.S., Brown, J.H., Gorga, J.C., Stern, L.J., Urban, R.G., Strominger, J.L., Wiley, D.C, 1996. Crystallographic analysis of endogenous peptides associated with HLA-DR1 suggests a common, polyproline II-like conformation for bound peptides. Proc. Natl. Acad. Sci. U.S.A. 93, 734–738.

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

Joachims, T., 2000. Estimating the generalization performance of a SVM efficiently. In: Proceedings of the International Conference on Machine Learning, Morgan Kaufman.

Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637.

Kay, B.K., Williamson, M.P., Sudol, M., 2000. The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. FASEB J. 14, 231–241.

Kelly, M.A., Chellgren, B.W., Rucker, A.L., Troutman, J.M., Fried, M.G., Miller, A.F., Creamer, T.P., 2001. Host-guest study of left-handed polyproline II helix formation. Biochemistry 40, 14376–14383.

Kim, H., Park, H., 2004. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. Proteins 54, 557–562.

Krimm, S., Tiffany, M.L., 1974. The circular dichroism spectrum and structure of unordered polypeptides and proteins. Israel J. Chem. 12, 189–200.

Mahoney, N.M., Janmey, P.A., Almo, S.C., 1997. Structure of the profilin-poly-L-proline complex involved in morphogenesis and cytoskeletal regulation. Nat. Struct. Biol. 4, 953–960.

Matthews, B.W., 1975. Comparison of predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405, 442–451.

Pauling, L., Corey, R.B., 1951. The structure of fibrous proteins of the collagen-gelatin group. Proc. Natl. Acad. Sci. U.S.A. 37, 272–281.

Rucker, A.L., Pager, C.T., Campbell, M.N., Qualls, J.E., Creamer, T.P., 2003. Host-guest scale of left-handed polyproline II helix formation. Proteins 52, 68–75.

Siermala, M., Juhola, M., Vihinen, M., 2000. Neural network prediction of polyproline type II secondary structures. In: Hasman, A., Blobel, B., Dudeck, J., Engelbrecht, R., Gell, G., Prokosch, H.-U. (Eds.), Medical Infobahn for Europe, Proc MIE2000 and GMDS2000, Studies in Health Technology and Informatics. IOS Press, Amsterdam, pp. 475–479.

Siermala, M., Juhola, M., Vihinen, M., 2001. On preprocessing of protein sequences for neural network prediction of polyproline type II secondary structures. Comput. Biol. Med. 31, 385–398.

Siermala, M., Vihinen, M., Juhola, M., 2003. On postprocessing of neural network prediction of polyproline type II secondary structures: network spectrum, response analysis, and scattering. Neural Comput. Appl. 11, 238–243.

Sreerama, N., Woody, R.W., 1994. Poly(pro)II helices in globular proteins: identification and circular dichroism analysis. Biochemistry 33, 10022–10025.

Stapley, B.J., Creamer, T.P., 1999. A survey of left-handed polyproline II helices. Protein Sci. 8, 587–595.

Sun, Y.F., Fan, X.D., Li, Y.D., 2003. Identifying splicing sites in eukaryotic RNA: support vector machine approach. Comput. Biol. Med. 33, 17–29.

Tiffany, M.L., Krimm, S., 1968. New chain conformations of poly(glutamic acid) and polylysine. Biopolymers 6, 1379–1382.

Traub, W., Shmueli, U., 1963. Structure of poly-L-proline (I). Nature 198, 1165–1166.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, Berlin.

Vapnik, V., 1998. Statistical Learning Theory. Wiley-Interscience, New York.

Wang, G., Dunbrack Jr., R.L., 2003. PISCES: a protein sequence culling server. Bioinformatics 19, 1589–1591.

Wang, M.L., Li, W.J., Xu, W.B., 2004. Support vector machines for prediction of peptidyl prolyl cis/trans isomerization. J. Pept. Res. 63, 23–28.

Williamson, M.P., 1994. The structure and function of proline-rich regions in proteins. Biochem. J. 297, 249–260.

Wilson, G., Hecht, L., Barron, L.D., 1996. Residual structure in unfolded proteins revealed by raman optical activity. Biochemistry 35, 12518–12525.

Woody, R.W., 1992. Circular dichroism and conformation of unordered polypeptides. Adv. Biophys. Chem. 2, 37–79.

Zavaljevski, N., Stevens, F.J., Reifman, J., 2002. Support vector machine with selective kernel scaling for protein classification and identification of key amino acid positions. Bioinformatics 18, 689–696.