John von Neumann Institute for Computing

Dihedral Angle Patterns in Coil Regions of Protein Structures

O. Zimmermann, U. H. E. Hansmann

published in

From Computational Biophysics to Systems Biology (CBSB07),
Proceedings of the NIC Workshop 2007,
Ulrich H. E. Hansmann, Jan Meinke, Sandipan Mohanty,
Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **36**, ISBN 978-3-9810843-2-0, pp. 301-303, 2007.

http://www.fz-juelich.de/nic-series/volume36

Dihedral Angle Patterns in Coil Regions of Protein Structures

Olav Zimmermann[1] **and Ulrich H. E. Hansmann**[1,2]

[1] John von Neumann Institute for Computing,
Research Centre Jülich, 52425 Jülich, Germany
E-mail: {*olav.zimmermann, u.hansmann*}*@fz-juelich.de*

[2] Department of Physics,
Michigan Technological University, Houghton, MI 49931-1295, USA
E-mail: hansmann@mtu.edu

We report first results on a Decision Tree classifier for subclassifying β-turn types from sequence. It is based on previous work where we developed a software DHPRED that predicts dihedral angle regions for each residue in a polypeptide chain. For 5 out of 8 β-turn (sub)classes we obtain good prediction results with Matthew's correlation coefficients between 0.3 and 0.6, thus providing additional geometric constraints in those critical regions which determine the fold topology of a protein. Three β-turn classes however can not be classified with sufficient accuracy. We discuss possible sources of misclassifications and outline further research directions.

1 Introduction

Many structure prediction strategies involve the use of dihedral angle constraints from secondary structure prediction. From the usual 3-class predictions useful dihedral bounds can however only be derived for α-helices and β-sheets. We recently developed a program (DHPRED) to predict dihedral angle ranges for all residues, including those in coil regions[1]. The most abundant structural motifs within the coil regions of proteins are β-turns which are defined as tetrapeptide motifs with a maximum C-α distance of 7 $\mathring{A}$. They have been originally described by Venkatachalam[2] and Lewis[3] and classified by the dihedral angle regions of their central two residues. β-turns are of prime importance for the tertiary structure of proteins as most of them characterize positions where the peptide chain reverses its direction. Several algorithms have been described to predict the location of β-turns from sequence, however to our knowledge there is no prediction program available which can distinguish between the different types of β-turns. In this study we use the dihedral angle regions predicted by DHPRED to develop a classification algorithm for individual β-turn classes.

2 Methods

As reference for the training of the classification algorithm we use a nonredundant set of protein structures from the Protein Data Bank (PDB) with pairwise sequence identities $\leq 25\%$ and x-ray resolutions ≤ 2.0 $\mathring{A}$. The assignment of the different turn types described by Lewis is performed by the PROMOTIF software[4]. The Decision Tree algorithm[5] as implemented in[6] is employed for multi-class classification. For this initial study we use

only minimal information. For the central two residues we encode (a) the amino acid type as either Glycine, Proline or all other, (b) the 3-class secondary structure prediction from PSIPRED[7] and (c) the dihedral angle regions predicted by DHPRED. We use the definitions in Table 1 for prediction outcome types:

Prediction	Observation	
	+1	-1
+1	TP (True Positive)	FP (False Positive)
-1	FN (False Negative)	TN (True Negative)

Table 1. Definition of prediction outcome types.

and employ as performance measures the accuracy $Acc = (TP + TN)/(TP + TN + FP + FN)$, specificity $Spec = TN/(TN + FP)$, sensitivity $Sens = TP/(TP + FN)$ and Matthew's correlation coefficient MCC $= \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$.

3 Results and Discussion

For the ca. 1,000 proteins of the nonredundant PDB-set, PROMOTIF assigns about 9,000 β-turns. We devide the turns randomly into two sets of same size and use one set for training of the Decision Tree classifier and the other for testing. Table 2 shows the confusion matrix of the test set and the distribution of the different β-turn classes.

		predicted								
	β-turn class	I	I'	II	II'	IV	VIa	VIb	VIII	sum
observed	I	1099	65	17	5	449	2	1	34	1672
	II	72	305	26	0	77	0	0	3	483
	I'	21	55	89	1	26	0	0	1	193
	II'	26	1	8	12	74	1	0	1	123
	IV	744	104	35	16	697	4	16	50	1666
	VIa	4	0	0	0	5	10	15	0	34
	VIb	0	0	0	0	5	2	34	0	41
	VIII	107	3	2	2	195	0	0	43	352
	sum	2073	533	177	36	1528	19	66	132	4564

Table 2. Confusion matrix for β-turn classifier.

Table 3 shows the performance of the Decision Tree classifier using different measures. While the Matthew's correlation coefficients for the classes show reasonable (I, VIa) or good (I', II, VIb) classification performance, the classes II', IV and VIII can not be identified correctly. The confusion matrix shows that the pseudo-class IV, which is just defined as all β-turns which do not belong to any other class, is responsible for most of the misclassifications.

β-turn class	Acc	Spec	Sens	MCC
I	65.7%	0.66	0.66	0.31
I'	63.1%	0.94	0.63	0.55
II	46.1%	0.98	0.46	0.46
II'	9.8%	0.99	0.10	0.17
IV	41.8%	0.71	0.42	0.13
VIa	29.4%	1.00	0.29	0.39
VIb	82.9%	0.99	0.83	0.65
VIII	12.2%	0.98	0.12	0.16

Table 3. Performance measures for β-turn classifier: accuracy (Acc.), specificity (Spec.), sensitivity (Sens.), and Mathew's correlation coefficient (MCC).

Our initial results show the general feasibility to predict β-turn types from the amino acid sequence. This study does not address the separation of β-turns from other secondary structure elements but such algorithms are already available[8]. In the future we will direct our research to assemble a comprehensive library for the prediction of local structure features. For the β-turn classification we will e.g. use a derivate of the DHPRED software which is targeted to different dihedral regions as used in Lewis' definition of β-turns and employ Support Vector Machine algorithms for the classification.

Acknowledgments

This work is supported in part by an NIH research grant (GM62838).

References

1. O. Zimmermann and U. H. E. Hansmann, *Support Vector Machines for prediction of dihedral angle regions*, Bioinformatics **22**, 3009–3015, 2006.
2. C. M. Venkatachalam, *Stereochemical criteria for polypeptides and proteins V. Conformation of a system of three linked peptide units.*, Biopolymers **6**, 1425–1436, 1968.
3. P. N. Lewis, F. A. Monany, and H. A. Scheraga, *Chain reversals in proteins*, Biochim. Biophys. Acta **303**, 211–229, 1973.
4. E. G. Hutchinson and J. M. Thornton, *PROMOTIF - A program to identify structural motifs in proteins*, Protein Science **5**, 212–220, 1996.
5. S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, (Prentice Hall, Englewood Cliffs, NJ, 2003).
6. P. Norvig, Python implementation of the Decision Tree algorithm decribed in[5], `http://aima.cs.berkeley.edu/python/`.
7. D. T. Jones, *Protein secondary structure prediction based on position-specific scoring matrices*, J Mol Biol **292**, 195–202, 1999.
8. Q. Zhang, S. Yoon and W. J. Welsh, *Improved method for predicting β-turn using support vector machine*, Bioinformatics **21**, 2370–2374, 2005.