John von Neumann Institute for Computing

Prediction of Parallel and Antiparallel Beta Sheets Based on Sequence Profiles Using Support Vector Machines

Longhui Wang, Olav Zimmermann, Ulrich H.E. Hansmann

published in

NIC Workshop 2006,
From Computational Biophysics to Systems Biology,
Jan Meinke, Olav Zimmermann,
Sandipan Mohanty, Ulrich H.E. Hansmann (Editors)
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **34**, ISBN-10: 3-9810843-0-6,
ISBN-13: 978-3-9810843-0-6, pp. 201-204 , 2006.

http://www.fz-juelich.de/nic-series/volume34

Prediction of Parallel and Antiparallel Beta Sheets Based on Sequence Profiles Using Support Vector Machines

Longhui Wang[1], 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: {l.wang, olav.zimmermann, u.hansmann}@fz-juelich.de

[2] Dept. of Physics, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, USA
E-mail: hansmann@mtu.edu

In this paper, we use SVM to construct classifiers to distinguish parallel and antiparallel beta sheets. Sequences are encoded as psiblast profiles. With seven-cross-validation carried on a 2686 non-homologous protein dataset, we obtain MCC (Matthew Correlation Coefficient) of 0.539836. The result shows that this two categories are separable by sequence profile.

1 Introduction

Beta-sheets are important secondary structures in the three-dimensional structures of peptides and proteins[1]. Analysis of their properties may therefore help us to understand the general mechanism of folding in proteins. In this paper, we assume that we know which part of the amino acid sequence is beta-sheet, and then use a SVM based classifier to distinguish parallel and antiparallel beta sheets based on sequence profiles generated by PSI-BLAST[2].

2 Dataset

The original dataset includes 2686 non-homologous proteins with pairwise sequence identity less than 25%. The secondary structure is assigned from the experimentally determined tertiary structure by DSSP[3]. There are eight secondary structure classes: H (α-helix), G (3_{10}-helix), I (π-helix), E(β-strand), B(isolated β-bridge), T(turn), S(bend)and -(rest). There are 412,748 residues in the whole dataset. In our experiments, we only use E(β-strand), B(isolated β-bridge) are excluded. There are 89,500 β-strand residues in the dataset, 22.28% of them are parallel, 78.29% of them are antiparallel, and include mixed samples. Residues from mixed samples are assigned as parallel.

3 Method

3.1 Coding Scheme

We use PSI-BLAST profiles (PSSM) to describe each amino acid. The PSSM has $20 \times N$ elements with N the length of protein sequence. Each element of PSSM is an integer

between -7 and 7, and is scaled to [0,1] range by the following function:

$$f(x) = \begin{cases} 0.0 & \text{if } x \geq 5; \\ 0.5 + 0.1x & \text{if } -5 < x < 5\ ; \\ 1.0 & \text{if } x \leq -5. \end{cases}$$

Here x is the value from the raw profile matrix[4].

For each protein, we perform 3-iteration blast on non-redundant(nr) database to get PSSM. The window length is 15 amino acid residues, and is fixed during the whole process. This value is commonly used from protein secondary structure prediction based on Support Vector Machine and Neural Network.

3.2 Support Vector Machine

Support Vector Machine (SVM) is a machine learning method proposed by Vapnik and his coworkers[4]. Some samples are inseparable in low dimensional space, SVM use a kernel function to map them into high dimensional space, and seek a hyper-plane to divide them. It performs very well on a lot of pattern recognition problems. In our paper, we use C-SVM. In this kind of SVM, we need to select a kernel function and the regularization parameter C for the classifier. We use radial basis function (RBF):

$$K(X_i, X_j) = exp(-\gamma \|X_i - X_j\|^2)$$

Different C and γ correspond to different classifiers, so we have to select best value for them our problem. And we use libsvm toolbox to construct classifiers[5], which can be download from http://www.csie.ntu.edu.tw/ cjlin/libsvm/

4 Experiment and Results

To select the best C and γ for SVM, we randomly group the dataset (2686 proteins) into 56 subsets, and take seven of them to do seven-cross-validation. Each of the seven subsets has similar parallel and antiparallel composition as the original one. Detailed information about these seven subsets are omitted. All the results are evaluated by percentage and Matthew Correlation Coefficient (MCC). Initial test shows C=512 and $\gamma = 0.003175$ as the best parameters. With this values, we do seven-cross-validation on 2686 proteins. Results are shown in Table 1.

From the results in Table 1, we can see that the total prediction accuracy on the whole dataset for parallel and antiparallel beta sheets is 85.35%, with MCC of 0.539836, which shows that sequence profiles are efficient to distinguish parallel and antiparallel beta sheets using SVM classifiers.

5 Discussion

The presented results illustrate that parallel and antiparallel beta sheets residues can be separated only using sequence information. Changing window length and adding penalty

Test set	Number of res	Antiparallel β-sheet%	Total predict Accuracy	MCC
Test1	11831	0.781844	0.860367	0.557849
Test2	13046	0.751265	0.84622	0.558996
Test3	11089	0.785824	0.847777	0.500422
Test4	13920	0.788578	0.863721	0.549801
Test5	12704	0.784635	0.854534	0.531594
Test6	14215	0.785227	0.858248	0.548517
Test7	12695	0.763056	0.840567	0.522223
Total	89500	0.7772	0.853552	0.539836

Table 1. Results on whole dataset

parameters in SVM may further improve the results. We should also do the analysis about which kind will the mixed samples be predicted to. An upcoming paper will discuss the results in more details[6].

Acknowledgment

Ulrich H.E. Hansmann is supported by a research grant (GM62838) of the National Institute of Health (USA).

References

1. S. S Zimmermann and H. A. Scheraga, *Local interactions in bends of proteins. Fast parallel algorithms for short-range molecular dynamics*, Proc. Natl. Acad. Sci. USA **9**, 4126–4129 (1977).
2. F. A. Stephan, L. T. Madden, A. A. Schaffer, J. Zhang, W. Miller and D. J. Lipman, *Capped BLAST and PSI-BLAST: a new generation of protein database search programs*, Nucleic Acids Res. **25**, 3389–3402 (1997).
3. W. Kabsh and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features*, Biopolymers. **22**, 2577-2637 (1983).
4. H. Kim and H. Park, *Protein secondary structure prediction based on an improved support vector machines approach*, Protein Engineering. **16(8)**, 553–560 (2003).
5. C. C. Chang and C. J. Lin, *LIBSVM: a library for support vector machines.* (2001).
6. L. Wang, O. Zimmermann and U.H.E. Hansmann, *Prediction of parallel and antiparallel beta sheet on sequence profiles using support vector machines.* In Preparation (2006).