SAAS: Short Amino Acid Sequence - A Promising Protein Secondary Structure Prediction Method of Single Sequence

Zhou Yuan Wu[1,2*], Ray P. S. Han[3]

[1]*School of Mathematics and Computational Science*
Xiangtan University
Xiangtan, China
E-mail: greenzyp@126.com

[2]*Civil Construction Engineering Department*
Guangxi University of Technology
Liuzhou, China

[3]*Department of Materials Science and Engineering*
Peking University
Beijing, China
E-mail: ray-han@pku.edu.cn

[*]*Corresponding author*

Received: March 21, 2013 **Accepted: June 14, 2013**

Published: July 29, 2013

Abstract: *In statistical methods of predicting protein secondary structure, many researchers focus on single amino acid frequencies in α-helices, β-sheets, and so on, or the impact near amino acids on an amino acid forming a secondary structure. But the paper considers a short sequence of amino acids (3, 4, 5 or 6 amino acids) as integer, and statistics short sequence's probability forming secondary structure. Also, many researchers select low homologous sequences as statistical database. But this paper select whole PDB database. In this paper we propose a strategy to predict protein secondary structure using simple statistical method. Numerical computation shows that, short amino acids sequence as integer to statistics, which can easy see trend of short sequence forming secondary structure, and it will work well to select large statistical database (whole PDB database) without considering homologous, and Q3 accuracy is ~74% using this paper proposed simple statistical method, but accuracy of others statistical methods is less than 70%.*

Keywords: *Short amino acid sequence, Protein Secondary Structure Prediction, Statistical method.*

Introduction

Protein secondary structure prediction is important one for protein folding [1, 2]. Its development goes from Chou-Fasman [3], GOR [4] statistical methods, to now popular neural network and multiple sequences alignment [5-12]. Traditional statistical methods are simple and easy to implement, but prediction accuracy Q3 is lower than 65%. And research on statistical methods is seldom now. The best prediction accuracy Q3 64.4% was claimed by GOR IV [13]. Improved Chou-Fasman method claimed that prediction accuracy Q3 was 56% [14]. Survey shows that, for pairwise sequence identities of > 35%, these secondary structure mappings are typically more than 90% accurate, and nearly 3/4 of newly deposited PDB structures have sequence identities greater than 25% to a pre-existing structure [15]. So in GOR V, multiple sequence alignment is used [16, 17]. We should consider using multiple sequence alignment if query sequences have homologous sequence of known

structure. This paper does not consider multiple sequence alignment and only uses simple statistics method, but the prediction accuracy is remarkably improved compared to other statistical methods without sequence alignment.

Proposed statistical methods

We compute probability of short sequence of amino acids as an integer forming secondary structure as follow:

$$p_i = \frac{n_i}{n},\ i = \text{H, E, C state,} \tag{1}$$

where n is the number of a short sequence of amino acids in statistical database, n_i is the number of a short sequence of amino acids forming helix (H), sheet (E), coil (C), or mixed in statistical database. It is very difficult to determine mixed secondary structure. So we do not compute its probability.

Next n and n_i are explained. For example, in statistical database, PDB ID of a protein sequence is 1A1U, its amino acids sequence is as follow:

DGEYFTLQIRGRERFEKIREYNEALELKDAQAGKE

If we consider RFE of the short sequence of 3 amino acids, then the 1A1U chain has one short sequence. We add up all the short sequences in statistical database, the total number is n for the short sequence. And we consider the short sequences of amino acids forming helix, where the secondary structure code of the short sequences is HHH. We add up all the short sequences forming helix in statistical database, the total number is n_i for the short sequence.

The prediction rules for proposed method are concluded as follow:

Firstly, probabilities for short sequence of 3, 4, 5, and 6 amino acids forming secondary structure are computed. The statistical sequences and secondary structure data comes from PDBFINDER database [18]. According to the amount of computation, cutoff of forming secondary structure for 3, 4, 5, and 6 amino acids short sequences is 0.4, 0.5, 0.6, and 0.6, respectively. Next, we explain how to select the cutoff. For example, for the purpose of selecting the cutoff of forming secondary structure for 5 amino acids short sequences, we computed secondary structure prediction accuracy rate using the 126 protein set as test set [19] with the cutoff changing from 0.3 to 0.9, the results is in the Fig. 1. From the figure, prediction accuracy rate is not sensitive to the cutoff. In the paper, we select 0.6 as the cutoff.

Then, the initial secondary structure is set to coil. We assign the short sequence a secondary structure if the probability of short sequence of 5 and 6 amino acids forming secondary structure is more than cutoff 0.6. If the number of a short sequence in the database is less than 10, the cutoff is set 1.0.

Results and discussion

Helix and sheet cannot exist in separate amino acid, so short sequence of amino acids should be considered as an integer, the integer may form helix, sheet, coil or mixed secondary structure.

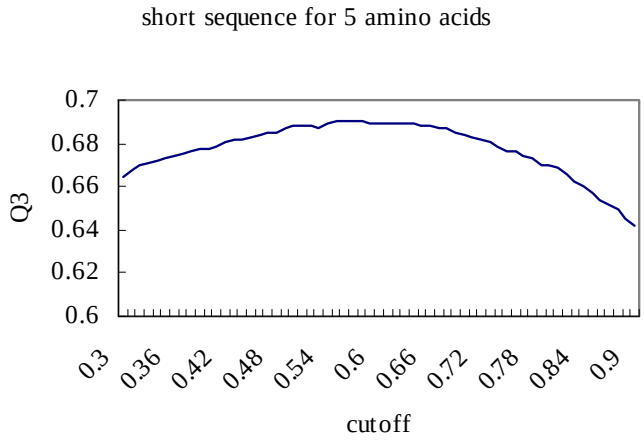

Fig. 1 Changes in the probability of the formation
of secondary structure for five amino acids short sequence

Why does this paper select 3, 4, 5, 6 amino acids' sequence to compute? According to the computation, 3, 4, 5, 6 amino acids' sequence have obvious improvement for prediction accuracy, and 7 and 8 amino acids' sequence hardly have improvement for prediction accuracy. So we only consider 3, 4, 5, 6 amino acids' sequence. Table 1 is comparison of secondary structure prediction for all combination of short sequences of 3, 4, 5, and 6 amino acids. This show the selection of proposed method is rational and the shorter sequence, the blurrier the orientation forming a secondary structure. And combination of whole short sequences of 3, 4, 5 and 6 amino acids is good selection.

Table 1. Q3 accuracy comparison for all combination
of short sequences of 3, 4, 5, and 6 amino acids

Short sequences	126 protein set	396 protein set	2180 protein set
3	60.9	59.9	59.7
4	70.5	66.4	63.8
5	85.3	79.3	77.5
6	87.2	83.2	83.8
3, 4	70.6	66.9	64.7
3, 5	85.4	79.8	78.3
3, 6	87.3	84.0	85.3
4, 5	85.6	79.9	78.2
4, 6	88.5	84.6	85.3
5, 6	89.2	85.2	86.3
3, 4, 5	85.5	79.8	78.3
3, 4, 6	88.1	84.6	85.4
3, 5, 6	89.1	85.3	86.6
4, 5, 6	89.4	85.5	86.5
3, 4, 5, 6	89.3	85.4	86.5

The 126 protein set comes from Rost and Sander [18]. The 396 protein set comes from reference Cuff and Barton [6]. The 2180 protein set comes from http://swift.cmbi.kun.nl/whatif/select/.

Table 2 presents Q3 and S_{ov} accuracy comparison with different methods based on test set, 126 protein set [19]. The 126 protein set was used by many papers as test set. We select new 153 protein set (not included in statistical database and released from 2007-11-30 to 2007-12-14 in PDB), and the set's sequence identities is less than 25%. This shows that the proposed method is remarkably better than GOR IV [13], GOR V [16], and J_{pred} [12] using 126 protein test set. The three methods' results were computed using their web service in April 10, 2008.

Table 2. Q3 and S_{ov} accuracy comparison with several different methods based on 126 protein set

Method	Proposed	GOR IV	GOR V	J_{pred}
Q3%	89.3	66.8	72.5	80.8
S_{ov}	84.3	61.8	68.3	77.4

S_{ov} is a assessment method for protein secondary structure prediction [13].

The results of proposed method comparison with GOR IV [13], GOR V [16], and J_{pred} [12] (see Table 3) using 153 new protein test set (statistical database is all proteins released before 2007-11-15 in PDB).

Table 3. Q3 and S_{ov} accuracy comparison with several different methods based on 153 new protein set

Method	Proposed	GOR IV	GOR V	J_{pred}
Q3%	73.7	60.8	67.6	80.7
S_{ov}	68.2	57.5	64.5	78.8

Our results of Q3 = 73.7% and S_{OV} = 68.2% compare favorably with the averaged values of Q3 = 69.7% and S_{OV} = 66.9% from the 3 techniques. Also, as listed in Table 4, the accuracy of Q3 improves with the volume of data; the greater the volume, the higher the accuracy of the protein secondary structure prediction. The Q3 accuracy of protein sequence 2VCIA (PDB ID, 208 amino acids) secondary structure prediction is 86.1% using the proposed method. Picture about the prediction results using Rasmol software is shown in Fig. 2. In Fig. 3 a comparison with observed and predicted protein secondary structure using 2D representation [21] about protein sequence 2VCIA is presented.

Table 4. Q3 accuracy change
with statistical database development

Latest released date of PDB proteins as statistical database	Q3(%)
2007-01-01	71.1
2006-01-01	69.0
2005-01-01	67.5
2004-01-01	65.6
2003-01-01	64.2
2002-01-01	63.4
2001-01-01	61.8
2000-01-01	60.6

Fig. 2 Prediction results on protein sequence 2VCIA
(using Rasmol software)
(the red is not predicted correctly)

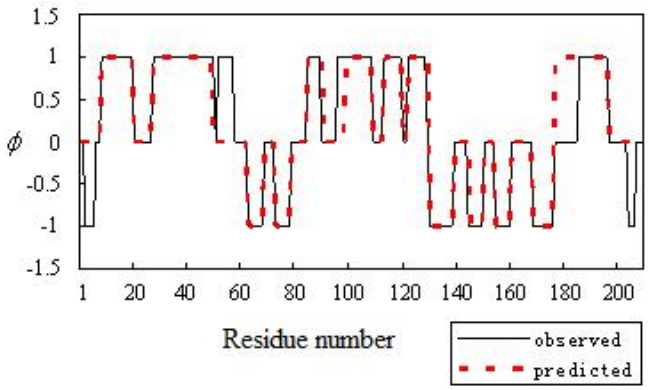

Fig. 3 Prediction results on protein sequence 2VCIA
(using 2D representation [20]).
Note: 1 denotes helix, -1 denotes sheet, 0 denotes coil.

Conclusion

In classic statistical methods of predicting protein secondary structure many researchers focus on single amino acid frequencies in α-helices, β-sheets, and so on, or the impact near amino acids on an amino acid forming a secondary structure. But we consider a short sequence of amino acids (3, 4, 5 or 6 amino acids) as integer, and statistics short sequence's probability forming secondary structure. Also, many researchers select low homologous sequences as statistical database. But this paper selects the entire PDB database. We propose a strategy to predict protein secondary structure using simple statistical method. Numerical computation shows that, short amino acids sequence as integer to statistics, which can easy see trend of short sequence forming secondary structure, and it will work well to select large statistical database (whole PDB database) without considering homologous, and Q3 accuracy is ~74% using this paper proposed simple statistical method, but accuracy of others statistical methods is less than 70%. The proposed method is promising and simple.

Acknowledgements

The investigations in this paper are supported by the Hunan Provincial Natural Science Foundation of China (Grant No. 10JJ7001); the Aid program of Science and Technology Innovative Research Team in Higher Educational Institutions; Science and Technology Planning Project of Hunan province of China (Grant No. 2011FJ2011) and Postdoctoral Research Fund of Xiangtan University of China. This support is gratefully acknowledged.

References

1. Babaei S., A. Geranmayeh, S. A. Seyyedsalehi (2010). Protein Secondary Structure Prediction using Modular Reciprocal Bidirectional Recurrent Neural Networks, Computer Methods and Programs in Biomedicine, 100(3), 237-247.
2. Bidargaddi N. P., M. Chetty, J. Kamruzzaman (2009). Combining Segmental Semi-Markov Models with Neural Networks for Protein Secondary Structure Prediction, Neurocomputing, 72(16-18), 3943-3950.
3. Chen H., F. Gu, Z. Huang (2006). Improved Chou-Fasman Method for Protein Secondary Structure Prediction, BMC Bioinformatics, 7(4), S14.
4. Chou P. Y., G. D. Fasman (1974). Prediction of Protein Conformation Biochemistry, Biochemistry-USA, 13, 222-245.
5. Cole C., J. D. Barber, G. J. Barton (2008). The Jpred 3 Secondary Structure Prediction Server, Nucleic Acids Research, 36(2), W197.
6. Cuff J. A., G. J. Barton (2000). Application of Enhanced Multiple Sequence Alignment Profiles to Improve Protein Secondary Structure Prediction, Proteins: Structure, Function and Genetics, 40, 502-511.
7. Li D., T. Li, P. Cong, W. Xiong, J. Sun (2012). A Novel Structural Position-specific Scoring Matrix for the Prediction of Protein Secondary Structures, Bioinformatics, 28(1), 32-39.
8. Garnier J., J. Gibrat, B. Robson (1996). GOR Method for Predicting Protein Secondary Structure from Amino Acid Sequence, Method Enzymol, 266, 540-553.
9. Garnier J., D. J. Osguthorpe, B. Robson (1978). Analysis of the Accuracy and Implications of Simple Method for Predicting the Secondary Structure of Globular Proteins, Journal of Molecular Biology, 120, 97-120.
10. Green J., M. Korenberg, M. Aboul-Magd (2009). PCI-SS: MISO Dynamic Nonlinear Protein Secondary Structure Prediction, BMC Bioinformatics, 10(1), 222.

11. Hooft R. W. W., C. Sander, G. Vriend (1996). The PDBFINDER Database: A Summary of PDB, DSSP and HSSP Information with Added Value, Computer Applications in the Biosciences, 12, 525-529.
12. Kloczkowski A., K. L. Ting, R. L. Jernigan, J. Garnier (2002). Combining the GOR V Algorithm with Evolutionary Information for Protein Secondary Structure Prediction From Amino Acid Sequence, PROTEINS: Structure, Function, and Genetics, 49, 154-166.
13. Lin H.-N., T.-Y. Sung, S.-Y. Ho, W.-L. Hsu (2010). Improving Protein Secondary Structure Prediction based on Short Subsequences with Local Structure Similarity, BMC Genomics, 11(4), S4.
14. Liu L., T. Wang (2006). 2D Representation of Protein Secondary Structure Sequences and Its Applications, Journal of Computational Chemistry, 27, 1119-1124.
15. Montgomerie S., S. Sundararaj, W. J. Gallin, D. S. Wishart (2006). Improving the Accuracy of Protein Secondary Structure Prediction using Structural Alignment, BMC Bioinformatics, 7, 301.
16. Pollastri G., A. J. Martin, C. Mooney, A. Vullo (2007). Accurate Prediction of Protein Secondary Structure and Solvent Accessibility by Consensus Combiners of Sequence and Structure Information, BMC Bioinformatics, 8(1), 201.
17. Qu W., H. Sui, B. Yang, W. Qian (2011). Improving Protein Secondary Structure Prediction using a Multi-modal BP Method, Computers in Biology and Medicine, 41(10), 946-959.
18. Rost B., C. Sander (1993). Prediction of Protein Secondary Structure at Better Than 70% Accuracy, Journal of Molecular Biology, 20, 584-599.
19. Sen T. Z., R. L. Jernigan, J. Garnier, A. Kloczkowski (2005). GOR V Server for Protein Secondary Structure Prediction, Bioinformatics, 21, 2787-2788.
20. Sundararajan S., P. Gniewek, R. L. Jernigan, A. Kolinski, A. Kloczkowski (2010). Protein Secondary Structure Prediction using Knowledge-based Potentials and an Ensemble of Classifiers, Biophysical Journal, 98, 52.
21. Zemla A., C. Venclovas, K. Fidelis, B. Rost (1999). A Modified Definition of S_{ov}, a Segment-based Measure for Protein Secondary Structure Prediction Assessment, PROTEINS: Structure, Function, and Genetics, 34, 220-223.

Zhou Yuan Wu, Ph.D.
E-mail: greenzyp@126.com

Zhou Yuan Wu is a doctor, and is an active member of Research Group at School of Mathematics and Computational Science, Xiangtan University, P. R. China. He is currently the post doctor of School of Mathematics and Computational Science, Xiangtan University. And he is working in the Civil Construction Engineering Department, Guangxi University of Technology, P. R. China. In research group, he is working on ongoing research projects. He has command over many Bioinformatics data analysis and structure prediction tools. His interests are Protein Structural Bioinformatics, Protein Folding, and Data Mining.

Prof. Ray P. S. Han
E-mail: ray-han@pku.edu.cn

Han Pingchou is a Changjiang Professor of Advanced Materials and Nanotechnology and an Assistant Dean of Engineering at Peking University, Beijing, China. His research interests encompass mechanobiology and microfluidics, biomaterials modeling and mechanics and nanodevices. He has graduated some 45 doctoral and masters students and post-doctoral fellows, and has authored with them 4 research monographs and 170 technical publications. Professor Han serves on the editorial boards of several journals and was a past chair of the ASME Technical Committee on Vibration and Sound. He has received over $6 M from various sources that include NSF, Canada NSERC, NSFC, China 973 Program, Guangdong Research Committee and Shanghai Research Foundation. His teaching experience spans a wide spectrum of courses in mechanical and material engineering. At PKU, he is responsible for the capstone design program and the College of Engineering Global Educational Exchange Initiative (GLOBEX).