Template-based Prediction of Protein 8-state Secondary Structures

Ashraf Yaseen
Department of Computer Science
Old Dominion University, Norfolk, VA
ayaseen@cs.odu.edu

Yaohang Li
Department of Computer Science
Old Dominion University, Norfolk, VA
yaohang@cs.odu.edu

Abstract — **Accurately predicting protein secondary structures is important to many protein structure modeling applications. In this paper, we investigate a template-based approach to enhance 8-state secondary structure prediction accuracy. The rationale is to construct structural templates from known protein structures with certain sequence similarity. The information contained in templates is then incorporated as features with sequence, evolutionary, and heuristic information to train neural networks. Our computational results show that templates containing structural information are effective features to enhance 8-state secondary structure prediction. A 7-fold cross-validated Q8 score of 78.85% is obtained.**

Keywords- Protein Secondary Structures, Neural Networks, Homology Templates

I. INTRODUCTION

An important intermediate step in modeling the three-dimensional structure of a protein is to accurately predict the secondary structures [1]. Most often, the secondary structures are classified into three general states, i.e., helices (H), strands (E), and coils (C). Correspondingly, performance of secondary structure prediction is typically measured by the Q3 accuracy. Many machine learning methods, including neural networks, hidden Markov chain, support vector machines, have been developed to predict secondary structures. Correspondingly, there are many secondary structure prediction servers available, including PSI-Pred [2], PHD [3], Porter [4], JPred [5], SSPRO [6], NETSURF [7], and many others. The modern secondary structure prediction servers can generate prediction results of ~ 80% Q3 accuracy.

Compared to the general three secondary structure states, the DSSP program [8] has more detailed classifications by assigning secondary structures to eight states, including 3-10 helix (G), α-helix (H), π-helix (I), β-stand (E), bridge (B), turn (T), bend (S), and others (C). The 8-state secondary structures convey more precise structural information than 3-state, which is particularly important for a variety of applications.

Most current secondary structure perdition methods, including those for 8-state predictions [6, 9], do not rely on similarity to known protein structures. However, many protein sequences have some degree of similarity among themselves. Actually, over half of all known protein sequences have some detectable similarity (> 25%) to one or more sequences of known structures [10]. Consequently, taking advantage of structural similarity of proteins with sequence similarity may lead to significant improvement of protein structure prediction. In fact, the latest version of porter has used homology-based templates for 3-state secondary structure prediction [10]. Porter has been reported to achieve prediction accuracy improvement when known structures with >30% sequence similarity are available and even reach theoretical upper bound when such sequence similarity is higher than 50%.

In this paper, we investigate a template-based method for 8-state secondary structure prediction. We extract structural information from known structures of chains with certain sequence similarity to build structural templates. Then, the structural template is incorporated as features together with sequence, evolutionary, and heuristic information for neural network training and validation. We test our prediction method on several popularly used benchmarks.

II. MATERIALS AND METHODS

A. The Protein Data Sets

We use the protein dataset Cull5547 from PISCES server [11] for neural network training. Cull5547 contains 5547 protein chains with at most 25% sequence identity and 2.0A resolution cutoff. Public benchmarks, including CB513 [12], Manesh215 [13], Carugo338 [14], and CASP9 targets [15] are used to benchmark our method.

B. Template Construction

For a given protein sequence target, PSI-BLAST [16] is used to search against the NR (Non-Redundant) database with E-value=0.001 and at most 3 iterations to generate the PSSM (Position Specific Scoring Matrix) data. Then, the PSSM is used to search against the Protein Data Bank (PDB) [17] for alignments with E-value=10.0. If known structures are available in PDB, their 8-state assignments are determined by the DSSP program and then a structural template is built for the correspondent residue positions.

C. Encoding and Nural Network Model

We use a window size of 15 residues for input encodings. Each residue is represented with 20 values from the PSSM data, 1 extra input to indicate if the residue window overlaps C- or N-terminal, 1 value for degree of similarity, and 8 values for structural information from template or context-based secondary structure scores. For a residue with available structural information, the corresponding secondary structure state is set to 1 while the other states are set to 0. At the same time, the degree of similarity is set for the sequence similarity. On the other hand, if the structural information for a resi-

due is not available, the degree of similarity is set to 0 and the context-based scores are incorporated instead. The context-based scores are heuristic statistics to specify the favorability of a residue adopting a certain secondary structure in its amino acid context. Detailed description of generating context-based scores can be found in [18].

We incorporate two phases of standard feed-forward neural network training. We use 7-fold cross validation on the training of protein sets. Q8 and SOV8 (Segment overlap [19]) scores are used to measure the prediction qualities.

III. Results and Discussions

Upon the selection of the best alignment with similarity < 95% for all chains in Cull5547, an overall of 78.85% Q8 accuracy and 80.10% SOV8 accuracy are achieved in 7-fold cross validation. Table 1 compares the accuracy of using predictions with and without templates on the benchmarks. Clearly, when homology structural information is available, the prediction accuracy is significantly improved.

Table 1. Comparison between 8-state predictions with and without template on CB513, CASP9, Manesh215, and Carugo338

		CB513	CASP9	Manesh215	Carugo338
Q_8	No Template	67.22	71.54	69.71	68.44
	With Template	79.39	76.36	81.10	80.39
SOV_8	No Template	67.66	73.47	70.79	69.50
	With Template	80.64	78.15	82.99	81.95

IV. Conclusions

We describe a template-based approach to enhance 8-state secondary structure prediction accuracy in this paper. Our computational results show that the templates can help improve the prediction accuracy. Overall, 78.85% Q_8 and 80.10% SOV_8 accuracies are achieved in 7-fold cross validation. The effectiveness of using templates has been demonstrated on the benchmarks. A webserver (C8-Scorpion) implementing 8-state secondary structure prediction is available at http://hpcr.cs.odu.edu/c8scorpion. The integration of template-based prediction into the C8-Scorpion webserver is currently under development.

Achknowledgements

This work is partially supported by NSF grant 1066471 and ODU 2013 Multidisciplinary Seed grant.

References

[1] B. Rost, "Review: Protein secondary structure prediction continues to rise," *Journal of Structural Biology,* vol. 134, pp. 204-218, May-Jun 2001.

[2] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *Journal of Molecular Biology,* vol. 292, pp. 195-202, Sep 17 1999.

[3] B. Rost and C. Sander, "Combining evolutionary information and neural networks to predict protein secondary structure," *Proteins-Structure Function and Bioinformatics,* vol. 19, pp. 55-72, May 1994.

[4] G. Pollastri and A. McLysaght, "Porter: a new, accurate server for protein secondary structure prediction," *Bioinformatics,* vol. 21, pp. 1719-1720, Apr 15 2005.

[5] C. Cole, J. D. Barber, and G. J. Barton, "The Jpred 3 secondary structure prediction server," *Nucleic Acids Research,* vol. 36, pp. W197-W201, Jul 2008.

[6] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles," *Proteins-Structure Function and Genetics,* vol. 47, pp. 228-235, May 1 2002.

[7] B. Petersen, T. N. Petersen, P. Andersen, M. Nielsen, and C. Lundegaard, "A generic method for assignment of reliability scores applied to solvent accessibility predictions," *Bmc Structural Biology,* vol. 9, Jul 31 2009.

[8] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers,* vol. 22, pp. 2577-637, Dec 1983.

[9] Z. Wang, F. Zhao, J. Peng, and J. Xu, "Protein 8-class secondary structure prediction using conditional neural fields," *Proteomics,* vol. 11, pp. 3786-92, Oct 2011.

[10] G. Pollastri, A. J. M. Martin, C. Mooney, and A. Vullo, "Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information," *Bmc Bioinformatics,* vol. 8, Jun 14 2007.

[11] J. Roland L. Dunbrack. A protein sequence culling server. Available: http://dunbrack.fccc.edu/Guoli/pisces_download.php

[12] J. A. Cuff and G. J. Barton, "Application of multiple sequence alignment profiles to improve protein secondary structure prediction," *Proteins-Structure Function and Genetics,* vol. 40, pp. 502-511, Aug 15 2000.

[13] S. Ahmad, M. M. Gromiha, and A. Sarai, "Real value prediction of solvent accessibility from amino acid sequence," *Proteins-Structure Function and Genetics,* vol. 50, pp. 629-635, Mar 1 2003.

[14] O. Carugo, "Predicting residue solvent accessibility from protein sequence by considering the sequence environment," *Protein Engineering,* vol. 13, pp. 607-609, Sep 2000.

[15] L. N. Kinch, S. Shi, H. Cheng, Q. Cong, J. M. Pei, V. Mariani, T. Schwede, and N. V. Grishin, "CASP9 target classification," *Proteins-Structure Function and Bioinformatics,* vol. 79, pp. 21-36, 2011.

[16] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research,* vol. 25, pp. 3389-402, Sep 1 1997.

[17] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research,* vol. 28, pp. 235-42, Jan 1 2000.

[18] Y. Li, H. Liu, I. Rata, and E. Jakobsson, "Building a Knowledge-Based Statistical Potential by Capturing High-Order Inter-residue Interactions and its Applications in Protein Secondary Structure Assessment," *J Chem Inf Model,* vol. 53, pp. 500-8, Feb 25 2013.

[19] A. Zemla, C. Venclovas, K. Fidelis, and B. Rost, "A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment," *Proteins-Structure Function and Genetics,* vol. 34, pp. 220-3, Feb 1 1999.