

Enhancing Joint Protein Structural Prediction by Ensemble Methods

Mary Qu Yang¹, Jack Y. Yang²

Keywords: protein structure, Ensemble Methods, Consensus Networking, boosting, joint predictor.

1 Introduction

Proteins exhibit several types of structure. In particular, regions of certain proteins, which have no definite tertiary structure, are known as Intrinsic Unstructured or disordered regions of Proteins (IUP). The secondary structure is defined by local, repetitive spatial arrangements of amino acids, which falls into three basic categories: helix, strand, and coil. IUP are associated with a wide range of protein functions [1] and are correlated with secondary structure. Identification of IUP regions can aid both structure determination and sequence alignment, and is also useful for drug design and discovery [2].

2 Ensemble Methods

We developed a new algorithm to predict IUP and protein secondary structure by utilizing Ensemble Methods, which are a diverse class of methods that seek to combine the decisions of several classifiers in order to improve performance. Our hybrid predictor uses the following:

1. Consensus Networking – In this approach, the test instance is fed into several classifiers and a majority vote of the classification decisions of these classifiers is taken. We combine the decisions from our predictors [3] with other classifiers such as Support Vector Machine Tree [4], Supervised-Unsupervised Tree [5] and Parallel Self-Organizing Neural Networks [4].
2. Bootstrap Aggregation (“Bagging”) – In this approach, the original data set is sampled (with replacement) to form M “bags” of data, each equal in size to the original dataset; a classifier is constructed based on each of the M bags. Then, given an instance to be classified, we feed it into each of the M classifiers and take the majority vote of these classifiers to form the final classification decision. There is a strong theoretical basis for “Bagging” in that it can be shown that under certain conditions “Bagging” will reduce the variance component of the error [3].
3. Boosting – In this approach, a series of classifiers are constructed based on the training data. To aid in learning the training data, a distribution over the training data is supplied to the classifier construction procedure; this distribution becomes more concentrated on the instances that are the most difficult to learn. Boosting transforms a weak learner into a strong one and has been shown to improve the performance of a wide range of classifiers. We have experimented with a particular boosting algorithm that uses confidence information returned by the classifier [6]; this approach has been combined with other techniques that we developed to improve the accuracy of our combined classifier.

¹ Purdue University, School of Electrical and Computer Engineering, Computer Engineering and Physics Department, Biological Physics, West Lafayette, Indiana, 47907 USA. E-mail: purduexy@purdue.edu

² Indiana University School of Medicine, Center for Computational Biology and Bioinformatics and Department of Biochemistry and Molecular Biology, Indiana University Purdue University Indianapolis, 714 North Senate Avenue EF 250, Indianapolis, Indiana 46202 USA. E-mail: jayyang@iupui.edu

Joint Secondary Structure and IUP Prediction

In this section, we discuss how to achieve more accurate secondary structure prediction by utilizing IUP information. Combining secondary structure information with primary structure and IUP information may also lead to more accurate tertiary protein structure prediction.

To better predict secondary structure, we generate homology information augmented with IUP/order and hydrophobicity information. The homology information is generated by using the PSI-Blast program iteratively to perform multiple sequence alignment over the protein database, guided by a score matrix. The initial score matrix used in these multiple alignment calculations is BLOSUM62. The final score matrix is of dimension M by 20, where M is the length of the protein sequence, and 20 is the number of possible amino acid residues. Since we want to augment the homology information with IUP/order information and hydrophobicity information, we use an M by 22 matrix, where the two additional columns contain structured (ordered)/IUP (disordered) and hydrophobicity information.

In this approach, two stages are used in predicting secondary structure. In the first stage, the structure is predicted from the sequence information, while in the second stage, the structure information is modified to yield a physically plausible structure. As there appears to be a correlation between secondary structure and the presence of IUP regions, it is advantageous to predict these jointly. Our joint IUP secondary structure predictor use the Ensemble Methods as described above.

3 Results and Discussion

Missing coordinators in PDB of a protein from X-ray Crystallography experimental data indicate those regions are IUP. Also IUP regions are obtained from NMR and other experimental data in PDB. Our IUP predictors had reached an overall performance of 80% based on n -fold cross validations [3].

We found that bagging improved the performance of our predictors slightly. A possible explanation for this is that our algorithm is known to be fairly stable under perturbations of the training data, so it is not ideally suited for bagging, which requires a diverse population of classifiers [3]. We also designed a filter on the output of the predictor; the overall performance was then increased by 2-3%. Our predictors contain architectures of Ensemble Methods to boost the performances of protein structural prediction from the amino acid sequences.

4 Acknowledgement

This research is partially supported by a Bilsland Interdisciplinary Doctoral Dissertation Fellowship for Computer Engineering and Biological Physics Dual-Degrees (MY) and a Post Doctoral Faculty Fellowship in Medical Science and Computer Engineering (JY). We thank A. Keith Dunker.

References.

- [6] Codrington, Craig.W., 2001 Boosting with Confidence Information, *Int'l Conf of Machine Learning*.
- [1] Dunker, A.K., Brown, C.J., and Obradovic, Z. 2002. Identification and functions of usefully disordered proteins. *Advances in Protein Chemistry* 62:25-49.
- [2] Dunker, A. Keith et al. 2001 "The protein trinity - linking function and disorder," *Nature Biotech* 19, 805-806.
- [4] Ersoy, O. K et al. 2002 "Support Vector Machine Decision Trees with Rare Event Detection" *IJSES* 225-242.
- [3] Yang, Jack, Yang, Mary, Zhang, Y, Dunker, A.K. 2003. A hybrid unsupervised-supervised approach to predict protein Disorder, *The first Indiana Bioinformatics Conference*, Indiana Univ. Purdue Univ. Indianapolis
- [5] Yang, Jack Y., Yang, Mary Qu and Ersoy, O. K. 2003. "Exploring Protein Functional Relationships Using Genomic Information and Data Mining Techniques," *Lecture Notes in Computer Science*, Springer-Verlag. ISSN 0302-9743 V. 2714