

Extracting Features from Protein Primary Structure to Enhance Structural and Functional Prediction

Mary Qu Yang¹, Jack Y. Yang², YiZhi Zhang²

Keywords: protein structure, protein function, feature extraction, feature selection, hybrid unsupervised-supervised predictor, ensemble method.

1 Introduction.

Proteins are composed of one or more chains of amino acids, and exhibit several levels of structure. The primary structure is defined by the sequence of amino acids comprising each chain, while the secondary structure is defined by local, repetitive spatial arrangements, which falls into three basic categories: helix, strand, and coil. The tertiary structure is defined by how the chain folds into a three-dimensional configuration, while the quaternary structure is concerned with how different chains combine into multisubunit or oligomeric, protein (protein complexes). The hypothesis is that the primary structure of a protein codes for all higher level structures and its function [1]. Given an amino acid sequence, we extract more than 500 features from sequence information only. Due to “curse of dimensionality”, we implement a feature selection algorithm to reduce the dimension of feature space and obtain the most important features. We applied our predictor by using those features to predict protein structural and functional classes.

2 Feature Extraction and Feature Selection

There are 20 different amino acids that occur in proteins, which can be denoted as {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}. We analyze the amino acid composition of a given sequence. The first set of 20 features of a given instance (e.g. an amino acid residue, a protein sequence) is derived from the first order statistics regarding “probability” of each of 20 amino acids in window [2] of length L of interest (for an instance that is protein sequence, the length of the window is the length of whole sequence). We then construct the second set of 400 more features by the second order statistics [3] regarding the pattern of one amino acid followed by another amino acid [3] in the same window of length L . Since amino acids have different biochemical and physical properties that influence their relative replace-ability in evolution, we next reclassify all 20 amino acids into a 9-gram encoding based on their biophysical and biochemical properties, as illustrated in table 1. The first order statistics on the 9-gram is performed in a similar way for 20 amino acids. This generates 9 more features. We then perform the second order statistics on the 9-gram and thus 81 new features are further added. We also use complexity in our features. Complexity is measured by Shannon’s Entropy [4]. We calculate the entropy of each position in the window of length L and average the entropy over the window. In addition, we add the relative hydrophobicity of each amino acid, called hydrophathy [5]. Since hydrophathy is important determinant of protein folding [1], calculation of hydrophathy could provide useful information for learning protein function. The feature for hydrophathy, H , is the average of hydrophathy of the same window of length L .

¹ Purdue University, College of Engineering, School of Electrical and Computer Engineering, Division of Computer Engineering, 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, Indianapolis, Indiana 46202 USA. E-mail: jayyang@iupui.edu

Group	Residues	Description
1	C	Cysteine, remains strongly during evolution
2	M	Hydrophobic
3	N, Q	Amides, polar
4	D, E	Acids, positive, polar
5	S, T	Alcohols
6	P, A, G	Aliphatic, small
7	I, V, L	Aliphatic
8	F, Y, W	Aromatic
9	H, K, R	Bases, charged

Table 1: 9-gram encoding scheme for amino acids based on biophysical proprieties.

At this point, we generated a total of 512 features. Feature selection [3] is necessary. Assume two features: feature X, x is the value of feature X; and feature Y, y is the value of feature Y. Distance $D(X)$ is the measurement how far the 2-class is separate by using feature X and distance $D(Y)$ is the measurement how far the two class is separate by using feature Y. If $D(X) > D(Y)$, then feature X should be selected, since the interclass distance for feature X is larger than feature Y. Otherwise, feature Y should be selected.

3 Results and Discussion

We input those features (after feature selection) to into our hybrid unsupervised-supervised predictor [2] based on the sequential bifurcation algorithm and ensemble method that we previously developed. We apply our predictor to predict protein secondary structure and to predict protein function. Our results show that augmenting features derived 9-gram encoding scheme and from biophysical properties of amino acids such as hydrophobicity, complexity etc. proved beneficial for learning protein structural and functional classes.

References.

- [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.
- [5] Kyte, J. and Doolittle R.F. 2001 A simple method of displaying the hydropathical character of a protein *J. Mol Biol.*, 157, 105-132.
- [4] Shannon C.E, Weaver W. 1949. The mathematical theory of communication. *University of Illinois Press.*
- [3] Wu, C.H, Whitson G, McLarty J., Ermongkonchain A., and Chang T. C. (1992), "Protein Classification Artificial Neural System," *Protein Science* 1, No. 5, 667-677.
- [2] 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