

# Extracting Features from Protein Primary Structure with Feature Selection to Enhance Protein Structural and Functional Prediction

Mary Qu Yang<sup>1</sup>, Jack Y. Yang<sup>2</sup>, YiZhi Zhang<sup>2</sup>

**Keywords:** feature extraction, feature selection, protein structure, protein function, hybrid predictor

## 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 fall 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 complexes. The hypothesis is that the primary structure of a protein codes for all higher level structures and associated functions [1]. Given an amino acid sequence, we extract more than 500 features from sequence information only. To overcome the “curse of dimensionality”, we incorporate a feature selection step to reduce the dimensionality of the feature space and obtain the most important features. We input those feature into our predictor [2] for predicting 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 or a protein sequence) is derived from the first order statistics, which are just the probabilities that each of 20 amino acids occurs in a window [2] of interest (of length  $L$ ). We then construct a second set of 400 additional features derived from the second order statistics, which are just the probabilities that two consecutive given amino acids [3] both occur in the same window. Since amino acids have different biochemical and physical properties that influence their relative replaceability 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, and calculate the first and second order statistics of the 9-gram encoding; the first order statistics of this encoding generate 9 additional features, while the second order statistics generate 81 additional features. Other features that we use include complexity (as measured by Shannon’s Entropy [4]) and the relative hydrophobicity of each amino acid (also called hydropathy [5]). To obtain values for the complexity and hydrophobicity associated with a given window, we calculate each of these features for each position in the window and then average them over the entire window [2]. Since hydropathy is an important determinant of protein folding [1], it could provide information that is useful for learning protein structure and function [1].

---

<sup>1</sup> 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

<sup>2</sup> Indiana University School of Medicine, Center for Computational Biology and Bioinformatics, 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, sulfur containing
3	N, Q	Amides, polar
4	D, E	Acids, polar, charged
5	S, T	Alcohols
6	P, A, G	Small
7	I, V, L	Aliphatic
8	F, Y, W	Aromatic
9	H, K, R	Bases, charged, positive

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

At this point, we have generated a total of 512 features; to reduce this number for more efficient learning, we incorporate a feature selection [3] step. Let  $D(X)$  be a measure of the separability of two classes based on feature  $X$ , and  $D(Y)$  be a measure of the separability of two classes based on feature  $Y$ . If  $D(X) > D(Y)$ , then feature  $X$  should be selected [2], since the interclass distance for feature  $X$  is larger than that for feature  $Y$ ; otherwise, feature  $Y$  should be selected.

### 3 Results and Discussion

The features calculated above (after feature selection) are fed into a hybrid unsupervised-supervised predictor [2] that we previously developed that is based on our sequential bifurcation tree algorithm [2] and also uses ensemble methods [2] such as Boosting [2], Consensus Networking [2], Bootstrap Aggregation (Bagging [2]). We applied our predictor to predict protein secondary structure and to predict protein function. Our results show that augmenting features derived from protein sequences with features derived from biophysical properties of amino acids such as hydrophathy and complexity, in conjunction with a 9-gram encoding scheme, proved beneficial for learning protein structural and functional classes.

### 4 Acknowledgement

This research is supported by a post doctoral fellowship (JYY) and a multi-disciplinary computer engineering and biological physics dual-degree doctoral fellowship (MQY). We thank Dr. A. Keith Dunker, Director of IU Medical School Bioinformatics Center for his guidance and financial support.

### 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 hydrophathical 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 Y., Yang, Mary Q., Zhang, Y., and Dunker, A.K. 2004. A hybrid unsupervised-supervised approach to predict protein disorder. *First Indiana Bioinformatics Conference*, Indiana University Medical School, IUPUI.