# Fuzzy *K*-Nearest Neighbor Classifier to Predict Protein Solvent Accessibility

Jyh-Yeong Chang, Jia-Jie Shyu, and Yi-Xiang Shi

Department of Electrical and Control Engineering
National Chiao Tung University
1001 Ta Hsueh Road, Hsinchu, Taiwan 300, R.O.C
jychang@mail.nctu.edu.tw

**Abstract.** The prediction of protein solvent accessibility is an intermediate step for predicting the tertiary structure of proteins. Knowledge of solvent accessibility has proved useful for identifying protein function, sequence motifs, and domains**.** Using a position-specific scoring matrix (PSSM) generated from PSI-BLAST in this paper, we develop the modified fuzzy *k*-nearest neighbor method to predict the protein relative solvent accessibility. By modifying the membership functions of the fuzzy *k*-nearest neighbor method by Sim *et al.* [1], has recently been applied to protein solvent accessibility prediction with excellent results. Our modified fuzzy *k*-nearest neighbor method is applied on the three-state, E, I, and B, and two-state, E, and B, relative solvent accessibility predictions, and its prediction accuracy compares favorly with those by the fuzzy *k*-NN and other approaches.

## 1 Introduction

The solvent accessibility of amino acid residues plays an important role in tertiary structure prediction, especially in the absence of significant sequence similarity of a query protein to those with known structures. The prediction of solvent accessibility is less accurate than secondary structure prediction in spite of improvements in recent researches. Predicting the three-dimensional (3D) structure of a protein from its sequence is an important issue because the gap between the enormous number of protein sequences and the number of experimentally determined structures has increased [2], [3]. However, the prediction of the complete 3D structure of a protein is still a big challenge, especially in the case where there is no significant sequence similarity of a query protein to those with known structures. The prediction of solvent accessibility and secondary structure has been studied as an intermediate step for predicting the tertiary structure of proteins, and the development of knowledge-based approaches has helped to solve these problems [4]–[8].

Secondary structures and solvent accessibilities of amino acid residues give a useful insight into the structure and function of a protein [8]–[11]. In particular, the knowledge of solvent accessibility has assisted alignments in regions of remote sequence identity for threading [2], [12]. However, in contrast to the secondary structure, there is no widely accepted criterion for classifying the experimentally determined solvent accessibility into a finite number of discrete states such as *buried*,

M. Ishikawa et al. (Eds.): ICONIP 2007, Part II, LNCS 4985, pp. 837 – 845, 2008.

*intermediate* and *exposed* states. Also, the prediction accuracies of solvent accessibilities are lower than those for secondary structure prediction, since the solvent accessibility is less conserved than secondary structure [2], although there has been some progress recently.

The prediction of solvent accessibility, as well as that of the secondary structure, is a typical pattern classification problem. The first step for solving such a problem is the feature extraction, where the important features of the data are extracted and expressed as a set of numbers, called feature vectors. The performance of the pattern classifier depends crucially on the judicious choice of the feature vectors. In the case of the solvent accessibility prediction, using evolutionary information such as multiple sequence alignment and position-specific scoring matrix generally has given good prediction results [13], [14]. Once an appropriate feature vector has been chosen, a classification algorithm is used to partition the feature space into disjoint regions with decision boundaries. The decision boundaries are determined using feature vectors of a reference sample with known classes, which are also called the reference dataset or training set. The class of a query data is then assigned depending on the region it belongs to.

Various classification algorithms have been developed. Bayesian statistics is a parametric method where the functional form of the probability density is assumed for each class, and its parameters are estimated from the reference data. In nonparametric methods, no specific functional form for the probability density is assumed. There are various nonparametric methods such as, for example, neural networks, support vector machines and nearest neighbor methods. In the neural network methods, the decision boundaries are set up before the prediction using a training set. Support vector machines are similar to neural networks in that the decision boundaries are determined before the prediction, but in contrast to neural network methods where the overall error function between the predicted and observed class for the training set is minimized, the margin in the boundary is maximized.

In the $k$-nearest neighbor methods, the decision boundaries are determined implicitly during the prediction, where the prediction is performed by assigning the query data the class most matched among the $k$-nearest reference data. The standard $k$-nearest neighbor rule is to place equal weights on the $k$-nearest reference data for determining the class of the query, but a more general rule is to use weights proportional to a certain power of distance. Also, by assigning the fuzzy membership to the query data instead of a definite class, one can estimate the confidence level of the prediction. The method employing these more general rules is called the fuzzy $k$-nearest neighbor methods [15].

The $k$-nearest neighbor method has been frequently used for the classification of biological and medical data, and despite its simplicity, the performances are competitive compared to many other methods. However, the $k$-nearest neighbor method has few been applied for predicting solvent accessibility, although it has been used to predict protein secondary structure. In this paper, we apply the modified fuzzy $k$-nearest neighbor method to the prediction of solvent accessibility where PSI-BLAST [16] profiles are used as the feature vectors. We obtain relatively high accuracy on various benchmark tests.