Chapter 9
Relevant and Non-Redundant Amino Acid Sequence Selection for Protein Functional Site Identification

Chandra Das
West Bengal University of Technology, India

Pradipta Maji
Indian Statistical Institute, India

ABSTRACT

In order to apply a powerful pattern recognition algorithm to predict functional sites in proteins, amino acids cannot be used directly as inputs since they are non-numerical variables. Therefore, they need encoding prior to input. In this regard, the bio-basis function maps a non-numerical sequence space to a numerical feature space. One of the important issues for the bio-basis function is how to select a minimum set of bio-basis strings with maximum information. In this paper, an efficient method to select bio-basis strings for the bio-basis function is described integrating the concepts of the Fisher ratio and "degree of resemblance". The integration enables the method to select a minimum set of most informative bio-basis strings. The "degree of resemblance" enables efficient selection of a set of distinct bio-basis strings. In effect, it reduces the redundant features in numerical feature space. Quantitative indices are proposed for evaluating the quality of selected bio-basis strings. The effectiveness of the proposed bio-basis string selection method, along with a comparison with existing methods, is demonstrated on different data sets.

DOI: 10.4018/978-1-4666-0264-9.ch009

INTRODUCTION

Cognitive informatics is the cross fertilization between computer science, artificial intelligence, systems science, cognitive science, neuropsychology, life science, and so forth (Wang, 2009a; Wang, 2009b; Wang et al., 2006b). It investigates the internal information processing mechanisms and processes of natural intelligence, and forges links between a number of natural science and life science disciplines with informatics and computing science (Wang et al., 2006a; Wang et al., 2006b).

The fundamental methodology of cognitive informatics uses informatics and computing techniques to investigate cognitive science problems such as memory, learning, and reasoning in one direction, although it is bidirectional and comparative in nature (Wang, 2009a; Wang, 2009b; Wang et al., 2006a; Wang et al., 2009; Wang et al., 2006b). In this paper, a new learning algorithm is presented to select a set of relevant and non-redundant amino acid sequences for identification of protein functional sites.

Recent advancement and wide use of high-throughput technology for biological research are producing enormous size of biological data. The successful analysis of biological data has become critical. Although laboratory experiment is the most effective method to analyze the biological data, it is very financially expensive and labor intensive. Pattern recognition techniques and machine learning methods provide useful tools for analyzing the biological data (Arrigo et al., 1991; Ferran et al., 1991; Cai et al., 1998; Baldi et al., 1995).

The prediction of functional sites in proteins is an important issue in protein function studies and hence, drug design. As a result, most researchers use protein sequences for the analysis or the prediction of protein functions in various ways (Baldi et al., 1998; Yang, 2004). Thus, one of the major tasks in bioinformatics is the classification and prediction of protein sequences. There are two types of analysis of protein sequences. The first is to analyze whole sequences aiming to annotate novel proteins or classify proteins. In this method the protein function is annotated through aligning a novel protein sequence with a known protein sequence. If the similarity between a novel sequence and a known sequence is very high, the novel protein is believed to have the same or similar function as the known protein. The second is to recognize functional sites within a sequence. The latter normally deals with subsequences (Yang, 2004).

The problem of functional sites prediction deals with the subsequences; each subsequence is obtained through moving a fixed length sliding window residue by residue. The residues within a scan form a subsequence. If there is a functional site within a subsequence, the subsequence is labeled as functional; otherwise it is labeled as non-functional. Therefore, protein subsequence analysis problem is to classify a subsequence whether it is functional or non-functional (Yang, 2004). The major objective in classification analysis is to train a classification model based on labeled data. The trained model is then used for classifying novel data. Classification analysis requires two descriptions of an object: one is the set of features that are used as inputs to train the model and the other is referred to as the class label. Classification analysis aims to find a mapping function from the features to the class label.

Many powerful pattern recognition algorithms like back-propagation neural networks (Cai et al., 1998; Qian et al., 1988; Narayanan et al., 2002), Kohonen's self-organising map (Arrigo et al., 1991), feed-forward and recurrent neural networks (Baldi et al., 1995; Baldi et al., 1998), bio-basis function neural networks (Thomson et al., 2003; Berry et al., 2004; Yang et al., 2005a; Yang, 2005b; Yang et al., 2005b; Yang et al., 2004), and support vector machines (Yang, 2004; Cai, 2002; Minakuchi, 2002), have been used to predict different functional sites in proteins such as protease cleavage sites of HIV (human immunodeficiency virus) and Hepatitis C Virus,