LETTER

Extraction and Optimization of Fuzzy Protein Sequences Classification Rules Using GRBF Neural Networks

Dianhui Wang*, Nung Kion Lee[+] , and Tharam S. Dillon*

*Department of Computer Science and Computer Engineering
La Trobe University, Melbourne, VIC 3083, Australia
E-mail: csdhwang@ieee.org

[+]Faculty of Cognitive Sciences and Human Development
University Malaysia Sarawak, Kota Samarahan, Sarawak, Malaysia
E-mail: nklee@fcs.unimas.my

(Submitted on May 14, 2003; Accepted on October 24, 2003)

Abstract—Traditionally, two protein sequences are classified into the same class if their feature patterns have high homology. These feature patterns were originally extracted by sequence alignment algorithms, which measure similarity between an unseen protein sequence and identified protein sequences. Neural network approaches, while reasonably accurate at classification, give no information about the relationship between the unseen case and the classified items that is useful to biologist. In contrast, in this paper we use a generalized radial basis function (GRBF) neural network architecture that generates fuzzy classification rules that could be used for further knowledge discovery. Our proposed techniques were evaluated using protein sequences with ten classes of super-families downloaded from a public domain database, and the results compared favorably with other standard machine learning techniques.

Keywords—Neural classification systems, data mining, rules extraction and optimization, generalized radial basis function networks, protein sequence

1. Introduction

A protein super-family consists of protein sequence members that are evolutionally related and therefore functionally and structurally relevant with each other [1]. One of the benefits from this category grouping is that some molecular analysis can be carried out within a particular super-family instead of individual protein sequence. It has also become apparent that the function of most genes is still unknown and classification into functionally related groups will provide valuable information on the protein function. Traditionally, two protein sequences are classified into the same class if they have high homology in terms of feature patterns extracted through sequence alignment algorithms. These algorithms, for instance, iPro-Class [4], SAM[5], MEME[6], compare an unseen protein sequence with all the identified protein sequences and provide a score based on similarity of sequences. As the size of the protein sequence databases is large, it is a very time consuming job to perform exhaustive comparison of existing protein sequences. Therefore, it is useful and helpful to build an intelligent classification system for effectively searching protein sequences in some large protein databases. Motivated by this, recently neural networks have been successfully applied in this domain and the results obtained demonstrate some merits of the methodology [1,2]. Neural networks have been chosen as technical tools for the protein sequence classification task due to the following two reasons: (i) the extracted features of protein sequences are distributed in a high dimensional space with complex characteristics which is difficult to satisfactorily model using some statistical or parameterized approaches; and (ii) neural networks are able to use the raw continuous values as system inputs. Basically, there are two types of neural models applicable for protein sequences classification task, i.e., unsupervised self-organizing mapping (SOM) networks [7] and supervised feed-forward neural networks (FNNs) [8,12]. The use of the SOM networks is to discover relationships within a set of protein sequences by clustering them into different groups. In contrast, the FNN based classification systems emphasizes on matching patterns through supervised learning. Once off-line training of the neural