Chapter 1

PROTEIN CLASSIFICATION USING NEURAL NETWORKS: A REVIEW

Nung kion Lee, Dianhui Wang, and Kock wah Tan

Abstract

This paper presents an overview on the application of neural networks (NN) in bioinformatics, specifically in the classification of protein family/superfamily. Protein classification is important for both biological data analysis and knowledge discovery. NN has been one of the most widely used methods for protein classification. In this paper, detailed discussion of protein classification processes using NN is presented with emphasis on the various protein sequences feature extraction and neural classifier design. Some other related issues and future challenges are discussed at the end of the chapter.

INTRODUCTION

A protein superfamily comprises set of protein sequences that are evolutionary and therefore functionally and structurally related. One of the benefits from this grouping is that some molecular analysis can be carried out within a particular superfamily instead of individual protein sequence. This can facilitate investigation of the functions of genes of an unknown sequence and could help provide valuable information. There are numerous ways to establish superfamily/family. For example, in Protein Information Resource (PIR) protein annotated databases [1], each superfamily is a collection of families. Sequences in these databases are grouped into the same superfamilies if they share at least 50% in overall identity. These identities are the end-to-end sequence similarity, including common domain architecture, and do not differ too much in overall length [2].

Usually, two protein sequences are assigned to the same class if they have high homology in the sequence level. Evidence of homology from these data shows that the genes may share a common evolutionary past (i.e., common ancestor). This is based on the first fact of biology sequence analysis that "if two peptides stretches exhibit sufficient similarity at the sequence level, then they are likely to be biologically related" [3]. Two of the classic well known algorithms to establish this homology measures are Smith-Waterman and Needleman and Wunsch.

Due to increasing number of molecular sequences, comparison between a query protein and proteins in database is an expensive operation. With improvement in the speed of sequence alignment algorithms (e.g., BLAST, and PatternHunter) and advances made in computer power, these tools are still practical for small to medium sized biological sequence databases. Currently and in the future, where the number of genomes is expected to be in the millions and with the availability of more complete genomes, these methods would become less practical.

Artificial intelligence technique such as neural network has been one of the most frequently used machine learning techniques in bioinformatics. The input-output mapping capability of NN can predict the degree to which a query sequence belongs to a superfamily/family; subsequent further analysis with reduced scope can be carried out by using sequence alignment tools. The neural network is also known for its tolerance to noise data due to mistake in the process of acquiring molecular data or incompleteness in the sequence data. The neural network can be used for proteins classification based on the information content of the protein sequences.

A protein sequence comprises a sequence of amino acid combination derived from twenty known amino acids. The set of abbreviation for the amino acids is represented by å ={A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}. An example of protein sequence is