Chapter XV

Predicting Protein Secondary Structure Using Artificial Neural Networks and Information Theory

Saad Osman Abdalla Subair, Al-Ghurair University, United Arab Emirates

Safaai Deris, University of Technology Malaysia, Malaysia

Abstract

Protein secondary-structure prediction is a fundamental step in determining the 3D structure of a protein. In this chapter, a new method for predicting protein secondary structure from amino-acid sequences has been proposed and implemented. Cuff and Barton 513 protein data set is used in training and testing the prediction methods under the same hardware, platforms, and environments. The newly developed method utilizes the knowledge of the GOR-V information theory and the power of the neural networks to classify a novel protein sequence in one of its three secondary-structures classes (i.e., helices, strands, and coils). The newly developed method (NN-GORV-I) is further improved by applying a filtering mechanism to the searched database and hence named NN-GORV-II. The developed prediction methods are rigorously analyzed and tested together with the other five well-known prediction methods in this domain to allow easy comparison and clear conclusions.
Introduction

Proteins are a series of **amino acids** known as polymers linked together into contiguous chains. In a living cell, the DNA of an organism encodes or **transcribes** its proteins into a **sequence** of nucleotides that are copied to the mRNA which are then **translated** into protein (Branden & Tooze, 1991). Protein has three main structures: (1) **primary structure**, which is essentially the linear **amino-acid sequence**; (2) **secondary structures** which are α helices, β sheets, and coils that are formed when the sequences of primary structures tend to arrange themselves into regular conformations (Kendrew, 1960; Pauling & Corey, 1951); and (3) **tertiary or 3D structure**, where secondary structure elements are packed against each other in a stable configuration. However, coils or loops usually serve as connection points between alpha-helices and beta-sheets; they do not have uniform patterns like alpha-helices and beta-sheets, and they could be any other part of the protein structure rather than helices or strands. In the molecular-biology laboratory, protein secondary structure is determined experimentally by two lengthy methods: X-ray crystallography method and nuclear magnetic resonance (NMR) spectroscopy method.

Advances in molecular biology in the last few decades, and the availability of equipment in this field, lead to the rapid sequencing of considerable genomes of several species. Several bacterial genomes, as well as those of some simple eukaryotic organisms, have been completely sequenced until now. These large genome-sequencing projects; including human genome projects, generate a huge number of protein sequences in their primary structures that are difficult for conventional molecular-biology laboratory techniques like X-ray crystallography and NMR to determine their corresponding 3D structures (Heilig et al., 2003).

One of the main approaches of predicting protein structures from sequence alone is based on data sets of known protein structures and sequences. This approach attempts to find common features in these data sets, which can be generalized to provide structural models of other proteins.

Prediction Methods

Since Anfinsen (1973) concluded that the amino-acid sequence is the only source of information to survive the denaturing process, and hence the structured information must be somehow specified by the primary protein sequence, researchers have been trying to predict secondary structure from protein sequence. Anfinsen’s hypothesis suggests that an ideal theoretical model of predicting protein secondary structure form its sequence should exist anyhow.

Many researchers used the approach of predicting protein structures from sequence alone, which is based on the data sets of known protein structures and sequences,