Chapter 58
Protein Secondary Structure Prediction Approaches:
A Review With Focus on Deep Learning Methods

Fawaz H. H. Mahyoub
https://orcid.org/0000-0003-4094-7023
School of Computer Sciences, Universiti Sains Malaysia, Malaysia

Rosni Abdullah
School of Computer Sciences, Universiti Sains Malaysia, Malaysia

ABSTRACT

The prediction of protein secondary structure from a protein sequence provides useful information for predicting the three-dimensional structure and function of the protein. In recent decades, protein secondary structure prediction systems have been improved benefiting from the advances in computational techniques as well as the growth and increased availability of solved protein structures in protein data banks. Existing methods for predicting the secondary structure of proteins can be roughly subdivided into statistical, nearest-neighbor, machine learning, meta-predictors, and deep learning approaches. This chapter provides an overview of these computational approaches to predict the secondary structure of proteins, focusing on deep learning techniques, with highlights on key aspects in each approach.

INTRODUCTION

Proteins constitute most of the dry mass of the cell. They are not just the building blocks of the cells; they also perform most of the functions of the cells. Proteins act as antibodies, antifreeze molecules, elastic fibres, signal integrators, enzymatic catalysis, toxins, hormones, transmembranal, and so on (Bruce et al., 2015). Majority of these functions are dependent on the 3D structure of proteins (Bruce et al., 2015). Accurate prediction of this 3D structure from the protein sequence is a very intricate task in computational biology (Yang et al., 2018). Since the rapid expansion in the fields of genomics and proteomics;

DOI: 10.4018/979-8-3693-3026-5.ch058

particularly, the DNA and protein sequencing technologies, there has been an enormous accumulation of protein sequence data. However, predicting the 3D structures of protein from its sequence data remains a key challenge facing bioinformatics (Jiang, Jin, Lee, & Yao, 2017). A proffered approach to resolving this prediction difficulty involves breaking down the problem into smaller structural problems, in the hope that their solutions will eventually result in a solution of predicting the 3D structure of proteins.

A number of these structural problems can be symbolized as 1D vectors along the protein sequence. Hence, they can be categorized as 1D structural features. Commonly used 1D structural features of protein are secondary structures (local conformations of the backbone of protein) (Voet & Voet, 2011), backbone torsion angles (rotation angles in the protein's backbone) (Voet & Voet, 2011), residue depth (the distance of an amino acid residue in the protein sequence from the adjacent solvent molecule) (Chakravarty & Varadarajan, 1999), residue accessible surface area (solvent accessibility) (Pedersen et al., 1991), residue contact number (the count of spatially-close amino acid residues within a cut-off space) (Pollastri, Baldi, Fariselli, & Casadio, 2002), and half-sphere exposure (orientation-dependent contact numbers) (Hamelryck, 2005).

This chapter focuses mainly on predicting the secondary structure of proteins. This is imperative as precisely predicting the secondary structures of proteins is crucial for several 3D protein structure related predictions (Hanson, Yang, Paliwal, & Zhou, 2017; Heffernan, Yang, Paliwal, & Zhou, 2017). The prediction of secondary structures has an extensive background, starting by the early work on the secondary structures of the backbone of proteins (Pauling & Corey, 1951a, 1951b; Pauling, Corey, & Branson, 1951), but it is only with the application of modern deep learning techniques and the growth of resolved protein structures that we seem to be approaching the theoretical limit of 3-state prediction accuracy (88-90%) (Rost, 2001; Yang et al., 2018).

The purpose of this chapter is to review the various predicting approaches of the secondary structure of proteins with emphasis on deep learning methods. The paper begins with a brief overview of proteins and their structure levels, followed by sections presenting the approaches utilized for the prediction of the secondary structure of proteins.

BACKGROUND

Proteins

Proteins are the centre of many biological activities, including transporting molecules (Klingenberg, 1981), responding to stimuli (Yoshida, Sanematsu, Shigemura, Yasumatsu, & Ninomiya, 2005), and catalysing metabolic reactions necessary for life possible (Margolis, 2008). Furthermore, proteins act as chemical go-betweens to sustain internal communication, as regulators to turn genes on and off, and as storages to store nutrients and energy-rich molecules for later use (Zvelebil & Baum, 2007). Most of the proteins are globular. Globular proteins are simpler to crystallize because of their chemical characteristics. Conversely, non-globular proteins such as fibrous and membrane proteins are frequently defined by numerous repeated amino acid sequences with less distinctive chemical characteristics (Yoo, Zhou, & Zomaya, 2008).