PREFACE

Protein is an essential molecule for defining the structure and function of body cells. However, protein sequences have enough information for determining their structure and function. There is a large gap between the number of protein sequences and the number of protein structures. Protein secondary structure prediction from its amino acid sequence is an important problem in molecular biology. Protein structural information is critical for understanding their function, but many protein sequences do not have their structural information. The large gap between protein sequence and structure motivates computational methods and, in particular, deep learning methods predicting structural information from the protein primary sequences.

In this thesis, we focused on predicting protein secondary structure from the primary sequences. Protein primary sequences represented using subsequence-based representations of amino acid sequences. The dense embedding vectors of primary sequences are trained over a large dataset of primary sequences. These dense embedding vectors combined with a bidirectional long short term memory network to capture the contextual information of amino acid residue for predicting secondary structure.

Firstly the thesis work focuses on primary sequence representation for deep learning methods. The primary sequences are segmented into twenty amino acid characters.

The dense embedding vector of amino acid residues combined with deep learning techniques for protein secondary structure prediction. Further, the primary sequences are a bag of character n-gram words. The different sizes of n-gram words computed and the dense vector of character n-gram words used to capture the local contextual information for secondary structure prediction. The variable-length character n-gram words dense embedding vector used for secondary structure prediction to overcome the drawback of fixed size character n-gram word. The recurrent neural network used for secondary structure prediction, the bidirectional long short-term memory network, better capture the long contextual information.

Further protein secondary structure prediction mapped as a sequence to sequence modeling. The amino acid residues are labeled with secondary structure elements. The state vectors for encoder and decoder are computed using long short term memory network.