Chapter 1: Introduction

This chapter presents an introduction to the problems discussed in this thesis, the motivation behind the present work, and the thesis's objectives. The chapter concludes with a list of contributions to this thesis focus on recent advances in deep learning and representation learning, and subsequence-based processing, which facilitates the primary sequences for secondary structure prediction.

1.1 Background

Proteins are macromolecules that are crucial elements for the structure and function of cells with a wide array of responsibilities including structural support, intra- and inter-cellular transport, catalytic activity, defense against bacteria and viruses, muscle contraction, signaling and regulation. Proteins accomplish their diverse functions in interactions with their environments, which can be other macromolecules such as proteins, DNA, or RNA, chemical compounds, or factors such as the pH or temperature[1]. Proteins are polymers of small molecules called amino acids, of which there are 20 different types, represented by the characters {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}. Protein sequences can be as large as chains of 10Ks amino acids; meaning that the space of possible protein sequences is very large. Proteins fold to form a particular three-dimensional structure. It has been proven that the protein’s linear sequence can determine their tertiary structures. The functions of proteins are highly tied to their 3D structures. Hence, a protein sequence should theoretically hold enough information determining its function. However, finding a mapping from the protein primary sequence to the structure is one of the open challenges in molecular biology[2]. The large gap between the number of known protein sequences and the number of known protein 3D structures motivates computational methods and in particular machine learning methods predicting structural information from the protein primary sequences. Protein structure can be described at three main levels: (i) primary structure referring to a linear sequence of amino acids, (ii) secondary structure referring to the structure of the local segments of the protein sequence categorized into 8 or 3 of secondary structures, and (ii) tertiary structure referring to the 3D structure of protein macromolecules. In this thesis, we focused on predicting protein secondary structure from the primary sequences. While it is challenging to predict a protein's structure directly from its sequence alone, accurate structure prediction can now be made using restraints derived from correlated mutations located from Multiple Sequence Alignments (MSAs) of homologous sequences, if a large number of homologous sequences are known. Similarly, the accuracy of predicting protein secondary structure, an important sub problem of protein structure prediction, increased from approximately 60% by early single-sequence-based techniques [1] to beyond 70% with the introduction of evolutionary information from MSA,[3] and to 82%–84% with the latest deep long-range learning techniques also with evolutionary information as the key input [4–7] However, the majority of proteins (>90%) have few, if any, known homologous sequences. In these cases, evolutionary information is limited or non-existent, and poor prediction accuracy is expected. It is quite possible that inaccurate evolutionary information might reduce the accuracy of prediction. Indeed, a recent single-sequence-based prediction of solvent Accessible Surface Area (ASA) is more accurate than evolution-profile trained methods for those proteins with few homologous sequences[8]. Thus, it is possible that one can simply improve prediction accuracy by the alternative use of single sequence and evolution-based methods, depending on the size of the homologous sequence cluster for a given protein. Moreover, such single-sequence-based prediction is computationally efficient because greater than 99% of the computational time is spent on generating evolutionary sequence profiles. Increasingly inexpensive sequencing techniques have led to an exponential increase in the number of known sequences. As a result, the computational time requirement for finding sequence profiles is continuing to increase. For example, PSSM generation, by PSI-BLAST, can take in the region of 30 min for a short protein (around 100 resides) up to multiple hours for a longer sequence (around 1000 residues). More importantly, a single-sequence-based prediction directly addresses the original sub problem of protein structure prediction: how far can we push the accuracy of predicting protein secondary structure from its sequence alone? Although secondary structure prediction is dominated by methods based on evolutionary information, progress has been making in single-sequence-based prediction.

1.2 Motivation

Proteins are macromolecules that are crucial elements for the structure and function of cells with a wide array of responsibilities, including structural support, intra-, and inter-cellular transport, catalytic activity, defense against bacteria and viruses, muscle contraction, signaling, and regulation. Proteins fold to form a particular three-dimensional structure. The functions of proteins are dependent on their three-dimensional structures. The Structure-based drug design relies on knowledge of the target protein's three-dimensional structure obtained through computational methods. It has been proven that the protein’s linear sequence can determine their tertiary structures (3D structure). Hence, a protein sequence should theoretically hold enough information determining its function. However, finding a mapping from the protein primary sequence to the structure is one of the open challenges in molecular biology. The large gap between the number of known protein sequences (UniProt database contains 116 million protein sequences to date) and the number of known protein 3D structures (Protein Databank contains only 142K entries for protein 3D structures to date) motivates computational methods and in particular machine learning methods predicting structural information from the protein primary sequences. It can help us understand the effects of mutations on structure and function. Protein primary structure is a linear combination of amino acids. The tertiary structure shows a three dimensional arrangement. The secondary structure connects both primary sequences and tertiary structures. Primary sequences are the basis of secondary structure prediction. The accurate prediction of secondary structure is essential for precise three-dimensional structure prediction, and beneficial for finding relationships within protein primary sequence and their function.

1.3 Problem Statement

Protein secondary structure prediction using deep learning and subsequence based representation of protein primary sequences. The Protein secondary structure prediction improved by extracting the character sequence based information of primary sequences such as character tokenization and a character n-gram words. The protein secondary structure prediction task can be viewed as a sequence labeling machine learning task type, i.e. assigning a categorical label $y_t \in Y$ to each element of a sequence of input elements, $x_t \in X$, where t indicates the position in the sequence. There exist eight possible secondary structure categories (Q8 labeling) for each amino acid at position t in the sequence: the 3-10 helix (G), α helix (H), π helix (I), turn (T), β sheet (E), β bridge (B), bend (S), and loop (L). A simpler labeling scheme is Q3, where the categories are divided into three main classes: helix, strand, and loop/coil.

- Given a protein sequence (primary structure)

GHWIATRGQLIREAYEDYRHFSSECPFIP

- Predict its secondary structure content

(C=Coils H=Alpha Helix E=Beta Strands)

INPUT: GHWIATRGQLIREAYEDYRHFSSECPFIP

OUTPUT: CEEEEECHHHHHHHHHHHCCCHHCCCCCC

1.4 Research Objectives

The key objectives of this research work are as follows:

1. Study of the computational methods for protein secondary structure prediction.
2. Literature review of the existing state-of-the-art methods for solving the problem of protein secondary structure prediction and identifying the research gaps and challenges.
3. Tokenization of protein primary sequences with the help of character based preprocessing techniques to extract the local contextual information.
4. Proposing the deep learning based architectures, its implementation, analysis and comparative study with other standard methods in literature for protein

secondary structure prediction. In this context, four methods have been proposed as reported in the thesis.

5. Deep learning network architectures are proposed to effectively process both local and global interactions between amino acids in making accurate secondary structure prediction.

1.5 Research Contributions

1. To study and compare the performances of the various conventional as well as state-of-the-art methods for protein secondary structure prediction.
2. Design, development, and implementation of a framework to learn a generic representation for protein sequences based on amino acid characters dense vector. In the proposed character-level representation of protein sequences, the models are compact and able to better handle chameleon sequences.
3. Design, development, and implementation of Protein Secondary Structure Prediction architecture by utilizing Character n-gram Embedding and Bi-LSTM. The primary sequences are represented with character n-gram words. We utilize the n-gram words embeddings in the context of disambiguation. These character n-gram embedding combined with a recurrent neural network to predict secondary structure.
4. Design, development, and implementation of Variable Length Character N-Gram Embedding of Protein Sequences for Secondary Structure Prediction. A variable length character n-gram based representation of primary sequences for protein secondary structure prediction is proposed. The variable length character n-gram words are selected based on their frequency in sequences.

5. Design, development, and implementation of Sequence-to-Sequence modeling for Secondary Structure Prediction: The character-based sequence-to-sequence models are used for mapping protein primary sequence to secondary structure. Their input is a sequence of amino acid characters, and they are directly optimized for secondary structure prediction.

   All above mentioned proposed methods have been implemented using Python software on a standard Intel Core i7 PC with 8 GB RAM and tested for standard datasets. Their performances were evaluated using various performance measures and also the performance of each of the proposed methods is compared against state-of-the-art methods available in literature. The obtained results and their performance analyses justify the applicability of the proposed approaches.

1.6 Outline of the Thesis

- Chapter 1 presents a brief introduction of the problems addressed in this thesis followed by the objectives of the thesis. Finally, the chapter concludes with a brief account on contributions of this thesis in the field of protein secondary structure prediction.
- Chapter 2 presents extensive survey of conventional and deep machine learning based approaches for protein secondary structure prediction. Further, Research Gaps have been identified.
- Chapter 3 presents Protein Secondary Structure Prediction using Sequence Embedding and Bidirectional Long Short Term Memory.

- Chapter 4 presents a Protein Secondary Structure Prediction using Character n-gram Embedding of primary sequences and Bidirectional Long Short Term Memory.
- Chapter 5 of the thesis presents Variable Length Character N-Gram Embedding of Protein Sequences for Secondary Structure Prediction.
- Chapter 6 presents a Sequence-to-Sequence Modeling using LSTM for Protein Secondary Structure Prediction.
- Chapter 7 concludes the thesis and summarizes main findings of the work done. This chapter also proposes some possible future perspectives of the thesis.