B.Sc. in Computer Science and Engineering Thesis

Protein Secondary Structure Prediction using Machine Leaning Methods

Submitted by

Md. Jakaria
1305024

Mostofa Rafid Uddin
1305039

Md. Mohaiminul Islam
1305077

Supervised by

Md. Shamsuzzoha Bayzid

Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology

Dhaka, Bangladesh

October 2018

CANDIDATES' DECLARATION

This is to certify that the work presented in this thesis, titled, "Protein Secondary Structure Prediction using Machine Leaning Methods", is the outcome of the investigation and research carried out by us under the supervision of Md. Shamsuzzoha Bayzid.

It is also declared that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Md. Jakaria
1305024

Mostofa Rafid Uddin
1305039

Md. Mohaiminul Islam
1305077

CERTIFICATION

This thesis titled, **"Protein Secondary Structure Prediction using Machine Leaning Methods"**, submitted by the group as mentioned below has been accepted as satisfactory in partial fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in October 2018.

Group Members:

Md. Jakaria

Mostofa Rafid Uddin

Md. Mohaiminul Islam

Supervisor:

Md. Shamsuzzoha Bayzid
Asst. Professor
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology

ACKNOWLEDGEMENT

Setting off by thanking our mentor, adviser Dr. Md. Shamsuzzoha Bayzid would be the correct gesture. It was due to him we were introduced to this novel field of research. We are really grateful for the opportunities, direction and the freedom that was provided while we were undergoing our research. His invaluable supervisions and insights inspired us a lot in developing the work we have done.

We are thankful to our friends from CSE, BUET who constantly provided us a suitable environment for carrying out our research work.

Finally, we want to thank our families for always being supportive and being there when we needed them the most.

Dhaka
October 2018

Md. Jakaria

Mostofa Rafid Uddin

Md. Mohaiminul Islam

Contents

List of Figures

List of Tables

ABSTRACT

Protein is the material basis of all living organisms and a widely discussed in bioinformatics. Knowing ther structure precisely is important for knowing their functions and that knowledge is crucial for disease reasoning and drug design. Protein secondary structure is a bridge between protein primary sequence and tertiary structure. But as finding out this structure information is very expensive via wet lab experiments, we have to rely on computational prediction based methods. This an extensively explored field in bioinformatics and numerous approaches have been experimented but deep learing based techniques provided best result till now. According to state of the art, scientists are using various hybrid deep learning models that utilizes single sequence information, evolutionary information via sequence profile and various physio-chemical properties of protein residues. However, though 3-state secondary structure prediction has advanced much, but state of the art 8-state prediction is yet to be improved for better precision and insight into protein structures. In our work, we have also used a hybrid model using CB6133 as training data and taking PSSM matrix and solvent accessibility of amino acid residues as input featues. Our experimented model gives 65% Q8 accuracy in benchmark CB513 dataset which is yet less than state of the art. However, we trying to incorporate and design special attention mechanism in PSSP. Attention mechanism is a recently discovered model which yielded great result in various fields of deep learning like NMT, Image Captioning etc. We believe it may give us better results along with a good insight in PSSP. This is still an ongoing research.

Chapter 1

Introduction

Proteins are large biomolecules, consisting of one or more long chains of amino acid residues that are produced in cells of all organisms. They are certainly the most important thing in living cells and the material basis of all living organisms as the play a crucial role in several life processes. They perform a vast array of functions including catalyzing metabolic reactions, DNA replication, responding to stimuli, providing structure to cells and organisms, and transporting molecules from one location to another. Usually they form complex networks together to perform a specific task and any pitfall in the process may cause to deadly diseases. So, knowing about the functions of protein has always been a major concern for the biologists and medical scientists. It is evident that the functions of proteins is largely dependent on its structures. Till now, due to the rapid advancement of genome and protein sequencing technologies a lot of protein sequences have been discovered. But the corresponding structure is known for only 0.2% of them. The known functionality is even more less. So, predicting unknown structures and functions of these proteins from their known sequences or unravelling the relation between amino acid sequences in protein and its structure has always been a grand challenge for the scientists and researchers in the field of computational biology. Generally, protein structures hierarchies are classified into four unique levels: the primary (peptide chain) , secondary, tertiary and quaternary. Protein tertiary structure and quaternary structure determines the 3-D structure of a protein and further determines its functional characteristics. The secondary structure is a connection between the primary and tertiary structure, and it is the early folding stage of protein molecule as the foundation of protein 3-D structure. As a result, the research of protein secondary structures is regarded as the first and the most important step in 3-D structure studies that will help us to understand the underlying relationship between the function and primary structure or peptide sequences of proteins Except to learn the 3-D spatial structure of protein, it can also be used in many protein science fields , such as the prediction of native tertiary structure , prediction of transition-state position , real value prediction of solvent accessibility , prediction of protein-protein interactions , prediction of protein structural classes , prediction of protein domains , prediction of -turns in proteins and so on. The importance and the far reaching implications of being able to predict the structure of a protein from its amino acid sequence is manifested by the ongoing biennial competition on Critical Assessment of Protein Structure Prediction (CASP) that started in 1994. CASP is designed to assess the performance of current structure prediction methods and over the years the number of groups that have been participating in it continues to increase. The prediction results for protein secondary structure faced a great improvement after the emerge and development of deep learning based methods in 90s. Though the protein secondary structure prediction work started in 1976 by Chothla and Levitt, but the third generation methods that used multiple sequence alignment (MSA) profile as in the input of machine learning began after 1992. Since then, a lot of machine learning algorithms and methods have been developed and the protein database used for feature extraction enriched a lot, too. In this book, we will discuss about all of them briefly in the Background and Related Works section. In this work, he have particularly focused on the problem of capturing both short range and long range interactions between amino acid sequences while folding to secondary structure using most updated machine learning methods while keeping the model as simple as possible. Hence we follow occams razor that simple model is always the better one. As protein secondary structure prediction is a very common problem in the field of computational biology, a lot of research groups are continually working on it and new works have been being released at about each months changing the state of the art of accuracy measures. In this book, we discuss the state of the art and highest accuracy obtained till June 2018. In our work, we have tried to incorporate the very recent attention mechanism in our deep learning based model and studied the effect of existing attention models in the context of protein secondary structure prediction.

1.1 Motivation

The rapid development in DNA sequencing and protein sequencing technologies resulted in enormous accumulation of protein sequence data. But technologies for getting accurate protein structure information of high precision via wet lab experiments have not developed in pace. Generally, protein structure is perfectly obtained using X-ray crystallography and multidimensional magnetic resonance in laboratory. These method have the disadvantages of extremely difficult, cost prohibitive, time consuming, limited molecular weight.Consequently, the experimental methods apparently do not cope with the challenge of the rapidly growing protein sequences data .So, scientists rely a lot on computational methods and bioinformatics tools as they are simple, low cost and fast speed which overcomes the disadvantage of experimental methods. However, simplicity comes with some cost. The accuracy obtained by the computational methods is still much below than the experimental ones which is cent percent accurate. So, the effort of developing the necessary computational methods is a live issue in the field of bioinformatics and this trend is not going to end in near future as more and more protein sequences are discovered now-a-days whose structure is unknown. In 1.1, we can observe where number of known protein sequences till 2014 is nearly 45 million and it is increasing at a rapid rate, but the available known structure till that time is far below, near 0.1 million only and not increasing accordingly. So, for the rest of the proteins (44.9 million) , we need to rely on prediction based computational methods for having a concept of their structures.

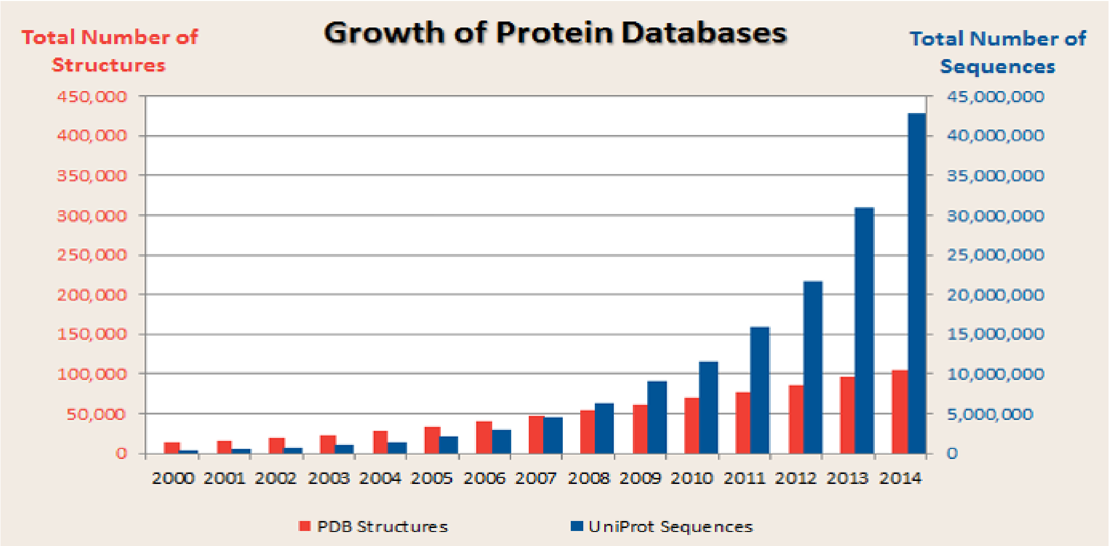


Figure 1.1: Comparison of growth of protein sequences in PDB and growth of known protein structures in Uniprot

1.2 Overview of Protein Secondary Structure Prediction

Protein Secondary Structure Prediction is basically predicting protein secondary structures from the amino acid residue sequence of proteins. There are 20 different types of amino acids. alanine - ala - A arginine - arg - R asparagine - asn - N aspartic acid - asp - D cysteine - cys - C glutamine - gln - Q glutamic acid - glu - E glycine - gly - G histidine - his - H isoleucine - ile - I leucine - leu - L lysine - lys - K methionine - met - M phenylalanine - phe - F proline - pro - P serine - ser - S threonine - thr - T tryptophan - trp - W tyrosine - tyr - Y valine - val - V

A protein consists of a polypeptide chain of these amino acid residues. The chains are usually of different lengths. However, while folding each amino acid residue converts to a certain structure primarily classified into three states Helix (H), Strand (E) and Coil (C). However 3 state is extended to 8-state now. Those are : H (α -helix), G(310-helix), I (π-helix), E (β-strand), B (isolated β-bridge), T (turn), S (bend), and C (Others).. While folding each amino acid turns into one of these shapes. So protein secondary structure transformation is basically a mapping or translation from amino acid residue sequence to a secondary structure state sequence of a protein and the length of those two sequences are always same. So, it can be reduced to a problem of transforming a sequence of {A,R,N,D,C,Q,E,G,H,I,L,K,M,F,P,S,T,W,Y,V}* to a equal length sequence of {H,E,C}* if 3 state prediction is used or {H,G,I,E,B,T,S,C}* if 8 state prediction is used. However, it may seem much similar like machine translation but is significantly different than popular machine translation problems like natural language processing, bilingual translation, text to speech, speech to text etc. problems as there is no bag of words in this case and each residue maps to different states in different protein sequences or in different positions in same protein sequence. However, it is due to the fact that while folding each amino acid residue bonds with its nearby residues and also sometimes residues of large distances in the sequence.

1.3 Problem

While protein folds from polypeptide chain and turns into secondary structure, the conversion of a residue to amino acid depends on the conversion of nearby residues to amino acid as while folding, usually nearby residues folds with each other. This phenomenon is called short range interaction. Again some residues bond with distant residues in poplypeptide chain while folding. This effect is called long range interaction. While a lot of machine learning and deep learning based methods have been developed and applied on protein secondary structure prediction problem, none could capture both of short range and long range interactions among residues of a polypeptide chain sufficiently. In our work, we have addressed this problem extensively and tried to come up with some simple better approach than existing ones.

1.4 Solution Overview

The problem of capturing both short and long range interactions of amino acid residues while forming secondary structure is not new and a significant amount of research works have been done in this field. Convolutional Neural Networks (CNN) works very well for location dependent features and so can capture short range interactions. But they notably fail to reflect the long range interactions. However, after the invention of long short term memory (LSTM) and Gated Reccurent Unit(GRU) networks, the problem of capturing long range interactions have been reduced a lot. However, though LSTM or GRU can capture long range interactions way better than CNNs, in case of short range interactions they are much less promising than CNNs. A simple hybrid approach is to use a combination of them. In our work, we build a cascaded model of multi-layer CNN-GRU network. However, we are also trying to incorporate attention mechanism with it inspired by the very recent breakthrough in performance in Neural Machine Translation(NMT) ,Image captioning etc tasks via attention mechanism. However, the naive attention layer doesnt seem to work good apparently in our case. We are exploring more of various existing attention mechanisms and trying to modify them to achieve better performance than the state of the art methods.

1.5 Outline

This dissertation book is composed of a total of six chapters.

Chapter 1 is the current chapter and introduces the topic of the thesis. The motivation of our works is the main focus of this chapter. An overview of protein primary and structures, problem definition, proposed model and our contribution are described in this chapter.

In chapter 2, we discuss about protein structures, history of protein secondary structure prediction, datasets and performance assessment in PSSM.

In chapter 3, we present many machine learning methods applied in the field of protein structure prediction.

chapter 4 describes the state of the art for 8-state protein secondary structure prediction.

Chapter 5 describes the proposed machine learning architecture, dataset and input features used by this model, and the performance measure of this model.

Finally, in chapter 6, we conclude our research work and explore the future possibilities of our approach.

Chapter 2

Background

2.1 Proteins

Proteins are polypeptide chains of amino acids folding into larger dimensions and operating most essential functions in living cells of an organism. They are the most basic building block of all organisms. Their formation, structure etc are essential for understanding our work. Hence we have provided a brief description of that.

2.1.1 Protein Formation

Proteins are actually governed by the information encoded in the genes of organisms. The formation of protein from genetic code is at the same time a complex and interesting process. The process includes two major steps: transcription and translation. Together, they are known as gene expression.

Genome sequence consists of 4 nucleotide base pairs : Adenine (A), Thymine (T) , Cytosine (C), Guanine (G). The whole genome sequence resides inside the nucleus. Messenger RNA or mRNA binds with the protein coding parts or genes of the chromosome. Each mRNA binds with three nucleotides at once forming a nucleotide triplet codon inside it. This mRNA then comes out of nucleus to cytoplasm. In cytoplasm it is captured by the ribosome which is also known as protein factory. Inside ribosome each mRNA bonds with a transfer RNA or tRNA according to the codon of mRNA. Each tRNA arriving at the ribosome carries a specific amino acid at one end and has a specific nucleotide triplet, an anticodon, at the other that bonds with the codon of mRNA. Codon by codon, tRNAs deposit amino acids in the prescribed order and the ribosome joins them into a polypeptide chain.Thus a protein is formed. The polypeptide chain of protein contains 20 different amino acid residues. Though a single amino acid is formed after a nucleotide triplet or 3 nucleotide base pairs (ATCG s) and there are 4^3 or 64 different combinations of triplet, but the number of possible amino acid is 20 as a number of combinations may result into a single amino acid.

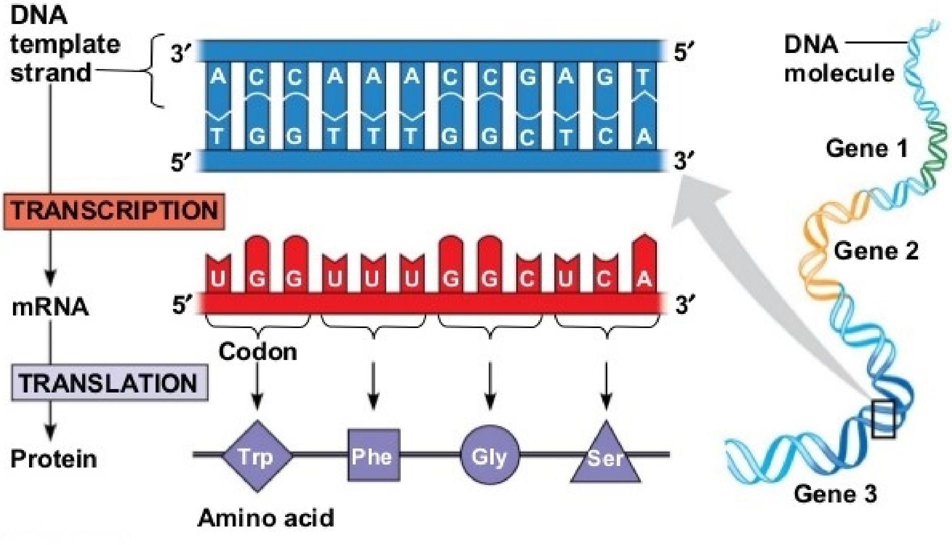


Figure 2.1: Formation of protein (amino acid residue sequences) form genome

2.1.2 Structure of protein

Protein structure is divided into 4 distinct hierarchical classes : Primary structure, secondary structure, tertiary structure and quaternary structure. The primary structure is simply the amino acid residue sequence : a linear one. The secondary structure is generated by the effect of hydrogen bonds, short and long range interaction of primary structure residues in one-dimensional space. Tertiary structure is formed by folding of secondary structures in three dimensional space. Such structure is enough to explain the function of mono polypeptide chain proteins. But understanding the functions of proteins with several polypeptide chains requires the concept of Quarternary strcuture. They are protein complexes formed by multiple three dimensional protein chains.

2.1.3 3 state and 8 state secondary structure of protein

Protein secondary structure gains special attention for being the basis of spatial structure of protein. Hence, each of the amino acid residues in primary structure presents different states by the impact of internal hydrogen bond in polypeptide chain. Initially only three possible secondary structures for each residues were assumed to exist : Helix (H), Strand (E), Coil (C). H is a helical configuration which are strengthened by the hydrogen bonds between every fourth

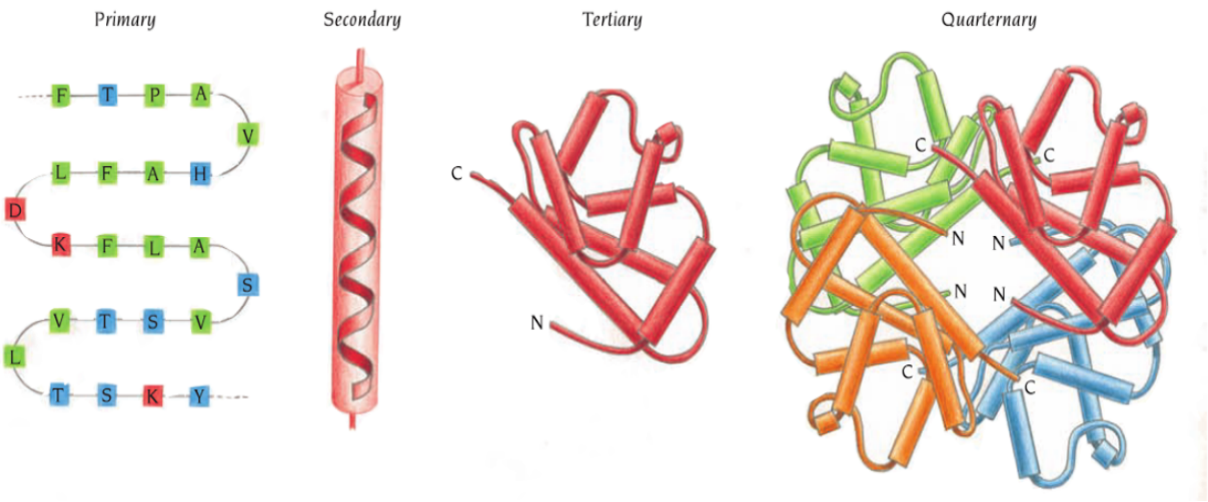


Figure 2.2: Protein Structure Hierarchy

amino acid; E is a strand segments structure generated by hydrogen bonds among interacting amino acids; C is a default class for those amino acids that do not belong to H or E classes [6].

However, later 3 states were extended to 8 states that classifies residues into 8 different secondary structure states : H (α -helix), G(3_{10}-helix), I (π-helix), E (β-strand), B (isolated β-bridge), T (turn), S (bend), and C (Others). 8 state prediction provides much detail information than 3- state prediction but they are much more challenging off course. In fact, the first distinct work solely based on 8 class secondary structure prediction was made 7 years ago in 2011 by Wang et al [7]. That was the first time 8-class prediction reached above 60% accuracy before that it was near 50% only in SSpro8. After that a number of work has been done which is briefly described in chapter 4 . However, there are ways to transform 8-state secondary structure into 3 states. But many pointed this mapping as vague as they were done using different classification criterions. However 5 methods of such 8-state to 3-state reduction are used which are depicted in Table 2.1 . Among them, method 3 is most popular.

Class name	3-state	8-state				
		method 1	method 2	method 3	method 4	method 5
Helix	H	H,G	H	H,G,I	H,G	H,G,I
Strand	E	E,B	E	E,B	E	E
Coild	C	S,T,I,C	G,S,T,B,I,C	S,T,C	S,T,B,I,C	S,T,B,C

Table 2.1: 8-state to 3-state reduction methods

2.2 History of Protein Secondary Structure Prediction

The first Protein Secondary Structure Prediction was presented more than 50 years ago. It was by Chothla and Levitt in 1976. After that, three generations of prediction technologies have been observed The first generation used statistical probability methods that utilized individual residue information to predict 3 states of protein. The most popular among such methods is Chou-Fasman's method. However those method were hardly expected to satisfy the requirement of protein 3-D structure analysis and prediction with an overall accuracy below 60%. The second generation lasted from 1980 to 1992. That also used statistical methods but in more sophisticated way and utilized neighbouring residue information and physio-chemical properties of amino acid residues. The representative of such method can be GORIII. However, the accuracy was still below 65%. The third generation prediction began in 1992 after the merge of some advanced machine learning methods. Notable among them are PHD and PSIPRED. They utilized information for multiple sequence alignment profile, homogeneous information and long-range correlation. These features are still the most used ones and they have been deemed to be the most useful ones in PSSP. However form 1992 to 2006, the overall accuracy of PSSP boosted to 76%-80%. From last decade, improved machine learning methods, hybrid models and incorporation of protein natural and physio-chemical properties as new features have boosted the accuracy over 80%. However, all the accuracy measures stated above are for 3-state secondary structure prediction. Till 2011, accuracy for 8-state prediction was only near 50%. So we may refer the time before that as first generation of 8-state prediction. A boost occured in 8-state prediction by the work of Wang et al. who developed a model targeting particularly 8-state prediction and achieved an accuracy of 64.9% on benchmark CB513 Dataset using Conditional Neural Fields (CNF). The CB513 is briefly discussed in 2.3.However, this started a new era in the way of 8-state secondary structure prediction. Later in 2014, Zhou and Troyanskaya from Princeton developed an Generative Stochastic Network(GSN) based model that achieved a 66.4% Q-8 accuracy on the same dataset which was a little improvement. Later complex and latest machine learning methods have been used for 8-state structure prediction. Among them,those achieved more than 70% accuracy till June 2018 are CNNH_PSS : A convolutional highway network (70.3% Q-8 accuracy on CB513) and MUFOLD-SS :A Deep Inception-Inside-Inception (Deep3I) network (70.63% Q-8 accuracy on CB513). Both of them are very recent works and have used very advanced and recent machine learning methods. Before them BLSTM , DeepCNF, DCRNN etc where used all having Q-8 accuracy in 67% to 70% range.

2.3 Datasets

Dataset choice for training and testing is one of the most vital part of any machine learning algorithm. In the context of PSSP, the fact is more intense. Some PSSP prediction methods are very much dataset dependent which is assumed to be a problem. R Researchers provide many sub-datasets, hybrid datasets for PSSP training and testing. So, there is actually a lot of datasets built and available for PSSP. Hence we limit our discussion for the most popular and benchmark ones.

2.3.1 PISCES CullPDB

PISCES is a public sever for culling sets of proteins from PDB (Protein Data Bank). Culling is a process of segregation or selection and hence it was done using sequence identity and structural quality criteria by Wang and Dunbrack in 2003. sequence identities are obtained from PSI-BLAST alignments with position-specific substitution matrices derived from the non-redundant(nr) protein sequence database. This was used by Zhou and Troyanskaya in 2014 and that contained 6128 proteins. The filtered data set of this CullPDB had a sequence identity of less than 25% with the CB513 test data, and it contained 5534 protein sequence after filtering. This dataset is publicly available at https://www.princeton.edu/~jzthree/datasets/ICML2014/ and used by about all the papers in PSSP field afterwards.

2.3.2 CB6133

This dataset is mostly used now-a-days in 8-state PSSP and was made by Zhou and Troyanskaya in 2014. This dataset contains evolutionary information, solvent accessibility and N- ,C-terminals of proteins. Hence they retrieved a subset of solved protein structures with better than 2.5A resolution and less than 30% identity, (Same set was used in (Wang et al, 2011). They also removed protein chains with less than 50 or more than 700 amino acids or discontinuous chains. They inferred 8-states secondary structure labels and solvent accessibility score from the 3D PDB structure by the DSSP program (Kabsch & Sander, 1983). The dataset is available at https://www.princeton.edu/~jzthree/datasets/ICML2014/

2.3.3 CB513 and CB396

CB513 dataset is developed by Cuff and Barton and comprises of 513 protein sequences and comprises 84,107 residues. It is the most used benchmark dataset and state of the art is mainly judged by accuracy on this unique dataset.It is non-homologous and sequence similarity of all 513 proteins are less that 25% so that there can be very little homology in training set. Little homology in training set is important for models so that model doesn't become much data-dependent.

CB396 consists of 396 proteins from CB513 having sequence identity less than 34% and having an average sequence length of 157 residues.

2.3.4 RS126

RS126 is also one of the most frequently used benchmark dataset for 3-state PSSP. It has 126 protein sequences and comprises 26,846 residues. It was developed by Rost and Sandar The average sequence length is 185 residues and average sequence identity is less than 31%

2.3.5 CASP

CASP stands for Critical Assessment of protein Structure Prediction. This is an biennial competition for protein strcuture prediction and a community wide effort to advance the state of the art in modelling protein structure from its amino acid sequences. It started in 1994 by CASP1. Since then CASP is held periodically in two years interval. CASP 13 will held in December 2018. CASP 14 will held in 2020. In each CASP competition, a new dataset is provided as benchmark dataset for testing and in many recent papers, models are evaluated based on CASP10, CASP11 and CASP12 datasets.

2.4 Feature Extraction

Feature extraction is a key issue of PSSP as the performance of models crucially depends on the choice of feature vectors used to classify the residue structures. Appropriate features are necessary for improvement of prediction accuracy of PSSP.

2.4.1 Single Sequence

Single sequence based methods were the only considered ones during the first generation of PSSP. However it can not utilize the evolutionary informations for other proteins. But today's evolutionary information based high accuracy models haven't thrown off this as input feature yet. Because most of the proteins identified in genome sequencing projects have no referable sequence similarity to any known protein.

2.4.2 Evolutionary information

After 1992, protein evolutionary informations have been incorporated in PSSP and this was possible due to the fact that a lot of sequences had stated being added to the protein data bank. Evolutionary information can be extracted using multiple sequence alignment profile of homologous proteins. There are many methods for generating multiple sequence alignment profile such as PSI-BLAST, PSI-Search,HMMER3, AMPS and CLUSTALW. All of these methods generate position specific profile and PSI-BLAST is the most popular among them. Some frequently used position specific profile formats are BLOSUM62 matrix and PSSM matrix. Between them, PSSM matrix are mostly used. Here is a brief description of them.

Position Specific Scoring Matrix (PSSM)

Position specific scoring matrices are generally obtained by PSI-BLAST algorithm. PSI-BLAST algorithm is applied both on genomic and proteomic data. In proteomics, it takes an amino acid residue sequence as input and searches a large database for similar sequences. The database can be specified by the user. There are a lot of popular databases like nr-database, Swiss_Prot (0.08GB), UniRef50 (4.3GB), UniRef90 (12GB) and UniRef100 (25GB). The later four are available at https://www.uniprot.org/downloads. PSI-BLAST can be run in the server available at https://blast.ncbi.nlm.nih.gov/Blast.cgi? For PSSP, protein blast is used. The algorithm parameters (e.g. expected threshold,max target sequence, number of iterations) can also be specified there. After that the PSI-BLAST program runs and assigns a score value to each of the amino acid residues of input sequence according to the number of sequence alignment found in the specified database with specified parameters. If the length of input sequence is l, it generates a matrix of $l \times 20$ dimension where each amino acid residue of input has a score value against all different types of residue. This is called PSSM matrix and used as input feature by most of the PSSM algorithm.

BLOSUM62 Matrix

BLOSUM62 matrix is also a scoring matrix and it takes into account the chemical properties of amino acids. It can effectively capture the difference between distantly related proteins. Hence, amino acid pairs with similar chemical property is assigned positive value and pairs with very different physciochemical properties are assigned negative values. The values are 'log-odds' score. It was proposed by HeniKoff et al. However, it is still less used for its more complexity and runtime.

2.4.3 Physiochemical Properties

With the advancement of molecular biology, several physciochemical properties of amino acid residues have been unveiled and those are effectively used as modern PSSP methods. The 20 amino acids can be grouped according to 4 most crucial physcio chemical properties. The properties and regarding types are :-

- **Hydrogen bond** : Hydrophobic and Hydrophilic

- **Polarity** : Polar and Non-Polar
- **Size** : Small and Large
- **Charge** : Charged and Uncharged

Above propeties and classifications can provide powerful insights in PSSP and used as input features in PSSP models.

2.5 Prediction Accuracy Assessment

Numerous assessment indexes were introduced and adopted by many scholars and researchers to in order to evaluate the prediction quality of PSSP methods. To find the optimal parameters for PSSP algorithms, these prediction quality assessment methods mostly use the measure of prediction accuracy. Researchers use these assessment methods to evaluate and intuitive represent the effectiveness of the PSSP approaches. Q score and segment overlap (SOV) are the two most frequently used integrative assessment methods. Between these two assessment methods, SOV was found more appropriate measure of prediction accuracy by critical assessment of methods of protein structure prediction (CASP) [8] [9]. Many other assessment methods like Matthews correlation coefficient, average absolute error, mean absolute error etc. are also used to assess the performance of PSSP techniques.

2.5.1 Q Score

Q score is the simplest and most popular measure methods for PSSP. There are two types of scoring functios (2.1) which are mostly adopted: the former one is three-state-per-residue accuracy (Q3) and the other is eight-state-per-residue accuracy (Q8). Q score calculates the percent of residues for each secondary structure which are correctly predicted [10] [9] [11].

$$Q_m = 100\frac{1}{N_{res}}\sum_{i=1}^{m} M_{ii} \tag{2.1}$$

Here $m = 3$ and $m = 8$ is referred as Q_3 and Q_8 accuracy, respectively. N_{res} is the total number of residues, and M_{ii} is correctly predicted number of residues in state i .

The per-state accuracy is the percentage of correctly predicted residues in a particular state, as (2).

$$Q_i = 100\frac{M_{ii}}{obs^i} \tag{2.2}$$

where obs^i is the number of residues observed in state i.

2.5.2 Segment Overlap

Segment Overlap (SOV) is second most frequently used assessment measure that is based on the average overlap between the observed and the predicted segments instead of the average per-residue accuracy shown in (2.3). This method takes into account the segments of continuous structure types instead of simple calculation of the number of correct residues.

$$SOV = 100\frac{1}{\sum_i N(i)}\sum_i \sum_{S(i)} \frac{\min ov(s_1, s_2) + \delta(s_1, s_2)}{\max ov(s_1, s_2)} \times len(s_1) \tag{2.3}$$

where $N(i)$ is the number of residues in state i , s_1 and s_2 are the observed and predicted structure segments, $\min ov(s_1, s_2)$ is the length of actual overlap of s_1 and s_2 , $\max ov(s_1, s_2)$ is the length of the total extent for which either of the segments s_1 and s_2 has a residue in state i . $len(s_1)$ is the number of residues in the segment of s_1 , and $\delta(s_1, s_2)$ is defined as (2.4) :

$$\delta(s_1, s_2) = \min((\max ov(s_1, s_2) - \min ov(s_1, s_2)); \min ov(s_1, s_2); int(len(s_1)/2); int(len(s_2)/2)) \tag{2.4}$$

From these two equations, it is evinced that SOV tolerates some small number of errors at the ends of segments, but it seriously penalizes the mistakes in the middle region of a secondary structure segment [10] [7].

2.5.3 Matthews Correlation Coefficient

Matthews correlation coefficient (MCC), popular method used to measure the quality of binary classifications, a more robust measure of correlation coefficient which takes into account both over- predictions and under-predictions. It is generally regarded as a balanced measure and can be used in prediction algorithms with different size classes. It returns a value between 1 and +1, the +1 represents a perfect prediction, 0 represents an average random prediction and 1 represents an inverse prediction. The formulation of MCC is given as (2.5)

$$MCC = \frac{TP \times TN - FP \times FN}{[(TN + FN)(TN + FP)(TP + FN)(TP + FP)]^{\frac{1}{2}}} \tag{2.5}$$

where TP (true positives) is the number of residues correctly predicted, TN (true negatives) is the number of residues that are not predicted, FP (false positives) is the number of residues incorrectly predicted, and FN (false negatives) is the number of residues observed in the secondary structure and incorrectly predicted [12].

2.5.4 Average Absolute Error

Average absolute error (δ^{Θ}) is the average of the absolute deviations between measured value and the mean value of data set.(δ^{Θ}) is calculated for each secondary structural element from each protein and is formulated as follows [13]:

$$\delta^{\Theta} = \frac{1}{N_{res}} \sum_{k=1}^{N} |\Theta_k - y_k^{\Theta}| \tag{2.6}$$

where Θ_k is the predicted content of the secondary structural element for the k-th protein, and y_k^{Θ} is the content actually observed.

The standard deviation of the average absolute error σ^{Θ} is formulated as follows:

$$\sigma^{\Theta} = sqrt\left(\frac{1}{N-1} \sum_{k=1}^{N} (\delta^{\theta} - |\Theta_k - y_k^{\Theta}|)^2\right) \tag{2.7}$$

And the overall average error $\langle\delta\rangle$ is given by

$$\langle\delta\rangle = \frac{1}{8} \sum_{\Theta} \delta^{\Theta} \tag{2.8}$$

2.5.5 Mean Absolute Error

The mean absolute error (MAE), which takes in account the periodicity of dihedral angles, is the average of the absolute distance between the observed and predicted value. MEA is formulated as follows:

$$MEA = \frac{1}{N_{res}} \sum_{i=1}^{N_{res}} \min(|p_i - x_i|, |360^{\circ} - (p_i - x_i)|) \tag{2.9}$$

where, N_{res} is the total number of residues, x_i and p_i represent the observed and the predicted value, respectively at state i.

In addition to above mentioned quality assessment methods, there are some other less frequently used methods such as cross validation tests, self-consistency, fuzzy Overlap (FOV), fuzzy correlation coefficient (Forr), standard error of prediction (SEP), k-state correlation coefficients and so on [10] [13].

Chapter 3

Machine Learning in PSSP

3.1 The Beginning

In 1988, Qian and Sejnowski proposed a Neural Network based method for PSSP which is considered to be one of the earliest machine learning work in the field of PSSP [14]. Machine learning based approaches quickly gained popularity in this field due to their achievements. And now machine learning is the most widely used method in PSSP. Hence, a number of machine learning methods have been used extensively by the researchers like Deep Neural Network (DNN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Support Vector Machine (SVM) etc. We are providing a brief discussion on each of them.

3.2 Deep Learning

Deep Learning is a very powerful artificial intelligence tool in bioinformatics. The first deep learning based PSSP method was DNSS, proposed by Spencer et al. in 2014. They used PSSM as input feature and developed a deep belief network (DBN) that had an Q3 accuracy of 80.7% and SOV of 74.2% on a subset mixture of CASP9 and CASP10 datasets [9]. Heffernan et al. also proposed at a Deep Learning (DL) based method at the same year using more additional features including solvent accessibe surface area, backbone angles and dihedral based on Cα atoms [15]. They achieved a Q-3 accuracy of 81.8% in CASP11 dataset using these physio-chemical properties. Wang et al proposed deep convolutional neural fields (DeepCNF) model that achieved an state of the art accuracy in both Q-3 and Q-8 prediction [7]. They used PSSMs as input features and achieved Q-3 accuracy of 82.3%, SOV of 84.8% and Q-8 accuracy of 68.3% in the benchmark dataset CB513. Bushia et al. form Google Brain developed a PSSP model based on Deep Convolutional Neural Networks (CNN) named NCNN in 2017 that achieved a Q-8 accuracy of 70.3% on benchmark CB513 dataset [3]. For 8-state secondary structure prediction Deep Learning has become a necessity as the non-linearity and variability increases much more in 8-state prediction. However, the cons of deep learning based method is its black-box characteristics and man can not understand its operation mechanism completely. So, improving the performance of DL based methods rely on changing and tuning hyperparameters.

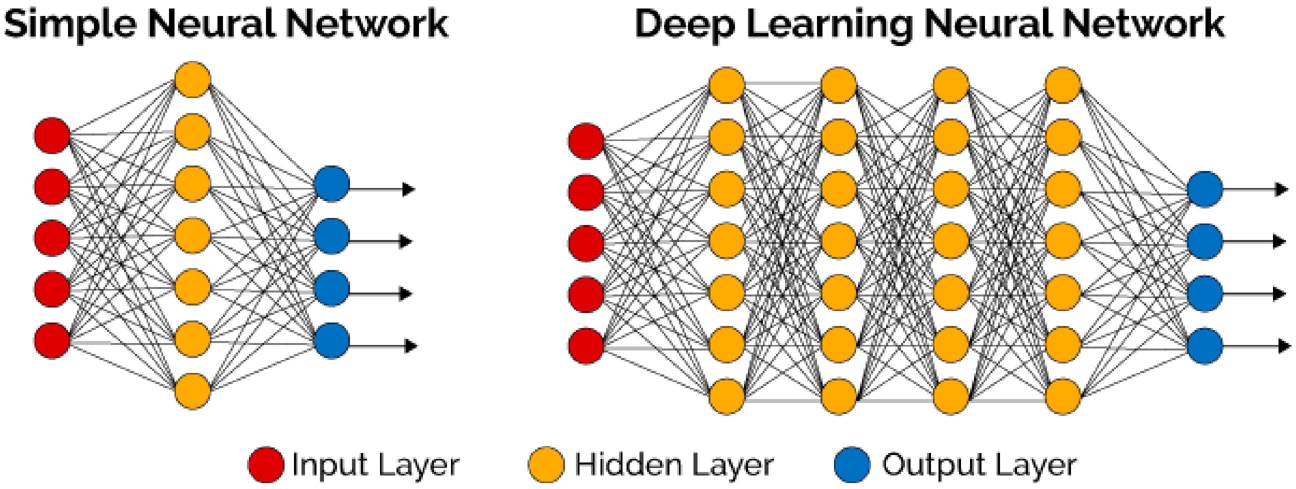


Figure 3.1: Deep Learning Neural Networks

3.3 Recurrent Neural Network (RNN)

Recurrent Neural Network or RNNs are powerful tools in the case of sequential information that was introduced in 1980s. Traditional neural network assumes all inputs (and outputs) to be independent of each other. But in many real life scenarios the situation is different. Often the data has persistance where traditional neural networks fail. However, this issue is addressed by RNN. With self loops, it combines the previous step and hidden representation into current step, allowing information to persist. In PSSP, the amino acid residue and corresponding structure labels are related to their previous or next residues sometimes at a large distance. The effect is termed as long-range correlations that feedforward neural networks fail to capture. But RNNs come handy in this case.

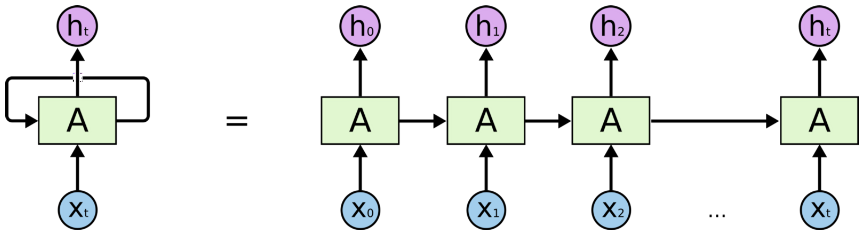
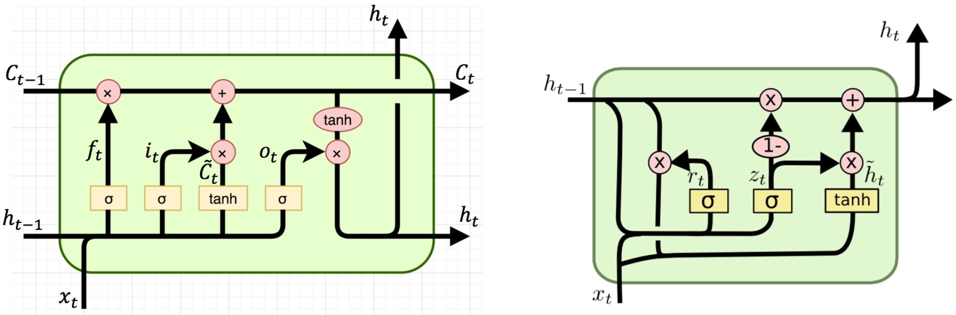


Figure 3.2: Recurrent Neural Networks

However, unidirectional RNN or vanilla RNN can not incorporate future input information form current state. In 1999, Pierre et al.introduced Bidirectional Recurrent Neural Network(BRNN) into PSSP, using PSSM as input feature. They used RS126 set for training. However, BRNN seemed to improve accuracy better than RNN and later many researchers tried to achieve better accuracy using it. However, RNN suffers from vanishing or exploding gradient problems that sometimes make the task of capturing long-term dependencies of proteins tough for RNNs. For this disadvantage of RNN, In 2010, Babaei et al. proposed a modular prediction system for PSSP that integrated a multilayer bidirectional RNN (MBRNN) to capture the strong short range interactions and a multilayer reciprocal recurrent neural network (MRRNN) that was to capture long range interactions [16]. Their method achieved Q-3 score of 79.36% on PSIPRED dataset. However, the speciality of the work was it was the first to use combination of models to capture short and long range interactions separately. Before that in 2006, J. Chen et al proposed segmented memory recurrent neural network inspired by human memory type and it would memorize segments from entire sequence [17]. Their method was more effective in capturing the long-term dependencies in proteins and acieved a Q-3 accuracy of 73.1% and SOV = 63.0% in CB396 dataset. (They trained using RS126 dataset). However, the variants of RNNs that can solve the vanishing gradient problem of RNN are Long Short Term Memory Networks (LSTM) and Gated Recurrent Unit (GRU) 3.3. Though LSTM was first proposed back in 1997 by Sepp Hochreiter and Jrgen Schmidhuber, but it gained popularity and ease of use by the researchers much later. In 2017, Heffernan et al proposed LSTM-BRNN network for PSSP and achieved Q3 score of 83.9% on TS115 dataset [18]. The lastest state of the art methods for 8-state PSSP are also using GRUs and LSTMs extensively for capturing long range interactions. More Details on them are provided in chapter 4



(a) Long Short-Term Memory

(b) Gated Recurrent Unit

Figure 3.3: Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) : variants of RNN that solves vanishing gradient problem via gated mechanism

3.4 Support Vector Machine

Support Vector Machine (SVM) is a statistical learning theory based machine learning algorithm and widely used in pattern recognition field including bioinformatics. It creates a optimal unique separating hyperplane that separates data into two classes with maximum margin in linearly separable problem. It also works for non linearly separable classes by using a kernel function which maps the input data to a high-dimensional feature space from its low dimensional space. Kernel function is the most powerful tool of SVM that makes it popular parallel to deep learning. One advantage for SVM over deep learning based methods are SVM is not black box like deep learning and can provide insight into actual phenomenon. However, it is less used as it is harder to implement in specific way.

The works in PSSP field based on SVM is mainly based on freuency patters. Birzele et al. used SVM based on frequent pattern phenomenon of consecutive amino acids in protein in 2006 [19]. They used this frequent amino acid patters, structure to structure layer and level wise search and achieved a Q3 accuracy of 75.34% - 77% on EVA150 dataset. In 2007, Chen et al. took into account pair-coupled amino acid composition or pair occurence frequency to train a SVM regressing system and achieved an average absolute error value of 0.018 only in CB513 dataset for 8-state prediction [13]. Karypis et al. used cascaded SVMs, exponential kernel functions and achieved a Q3 accuracy of 77.83% and SOV of 75.05% on benchmark CB513 dataset [20]. Later in 2009, Kountouris et al. considered backbone dihedral angles as input features in SVM along with PSSM and achieved 80% Q3 accuracy on CB513 [11]. In 2011, Chatterjee et al. used physiochemical properties and PSSM as inputs of his two-stage cascaded SVM classifier and obtained Q3 score of 75% on CASP9 [21].

However, SVM is very suitable in PSSP due to its pattern recognition capacity and powerful kernel functions that can map non-linear characteristics of PSSP task to a linear one in high-dimensional space. However, SVM is actually designed for binary classification and doing multiclass classification with it effectively is still an ongoing research

3.5 Hybrid Methods

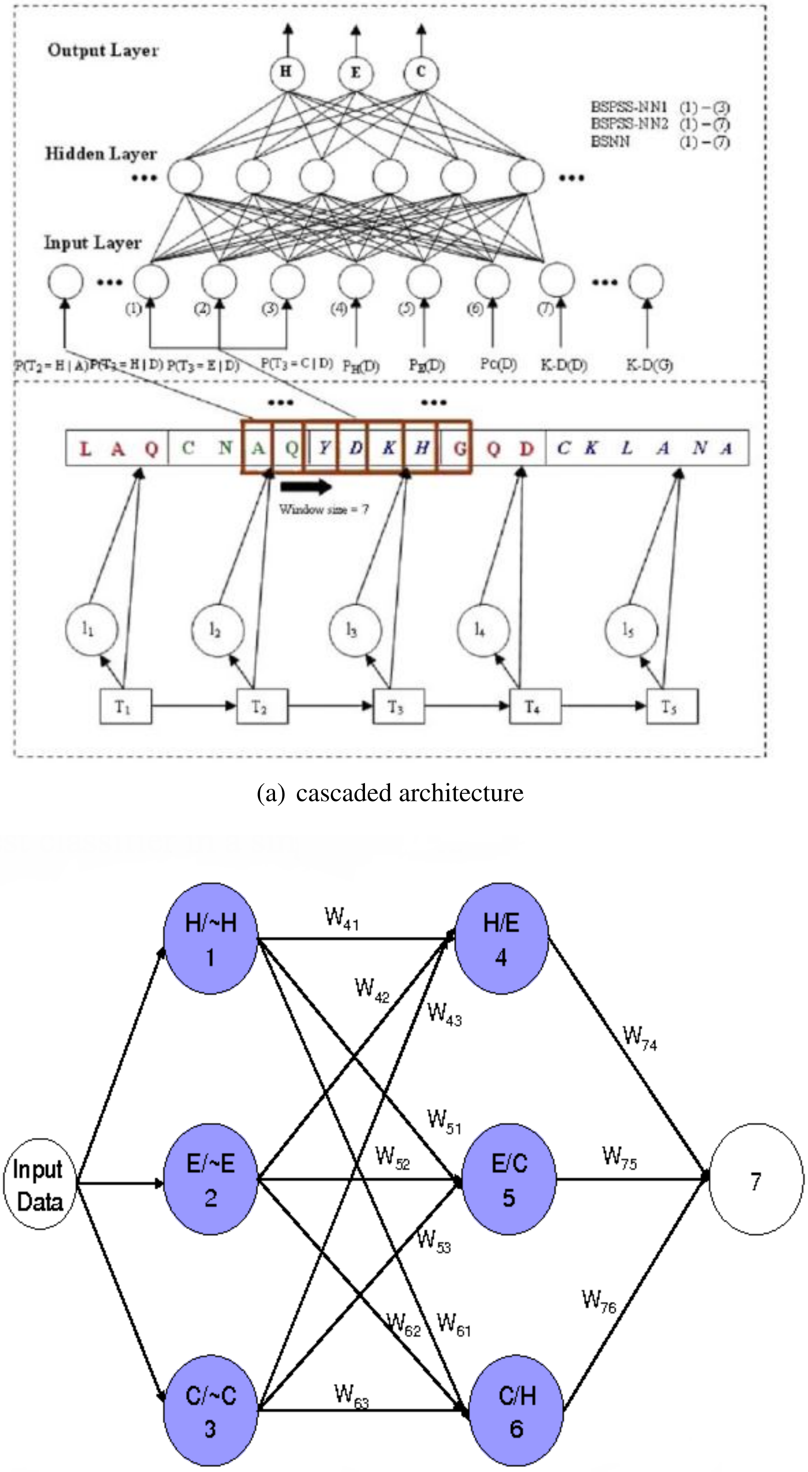
New advanced technology has geared the improvement of machine learning tools and many hybrid models has been developed in different prediction methods. These tools and methods accelerated the enhancement in prediction accuracy of protein secondary structure. In previous sections, we have seen that any single model have its limitations in predicting protein secondary structure. So scholars conducted exhaustive study on employment multiple machine learning model on same data set and combine their outputs to get better performance. This type of learning model is called Ensemble Learning [22]. The models employed to train multiple learning machines are treated as “committee” of learners. Each member of this committee learn individually, and their individual prediction are combined which gives better overall accuracy then the individual accuracy because the shortcomings of one model is focused by other models and remedied thereby [23]. This type of hybrid models have created a new trends in PSSP in recent years for its better accuracy. There is another type of prediction models which have become popular among scholars for it high performance, referred as Multi-step processed based hybrid methods. These prediction processes are composed of two learning methods, one is for feature extraction and the other is for prediction. The prior learning processes are specialized on effective extraction of features from protein data by applying suitable methods. The second-mentioned process is applied on the output of the former process which adopts proper learning methods to yield prediction of secondary structure.

3.5.1 The Combination of Multiple Methods

Every model has its inherent characteristic. If it was possible to build a "perfect" machine learning model that will yield best possible answer, there would be no need for hybrid models. Each of the models has some limitations that cause the errors. For example, it is difficult to capture the evolutionary information of proteins by probabilistic graphical model while neural network is good at capturing this information. Again, neural network yield poor performance in capturing the distant interactions among residues. The combination of these two models remedy the shortcomings of individual models and produce better accuracy than any of the two individual models [24]. The field of hybrid prediction models can be divided into two categories: one is cascade architectures and the other is cross architecture [25]. In the first architecture, multiple learning methods are combined in a way that output of one stage is the input for another stage [24]. In second architecture, multiple learning methods are combined in crossed fashion [26]. The graphical depiction of these two class of hybrid models are shown in Figure 3.4(a) and 3.4(b).

1. Neural networks-probabilistic graphical models

In 2009, Bidargaddi et al. proposed a hybrid model for PSSP which consist of three stages. In first stage they captured mom-local interaction by using the Bayesian segmentation based on Segmental semi Markov model (SSMM) through a joint sequence-structure probability distribution. In second stage they used neural networks to capture local correlations. This network is fed by a local window of residues with segmentation scores and physical-chemical properties. At the last stage, they used posterior probability scores based neural network ensembles. The structural states of each residue sample were used for training and this method performed better for single sequences without homogeneous information [24]. Malekpour et al. proposed a similar type of hybrid model with some slight changes to the model proposed by Bidargaddi et al. They changed the order of first two stages from the previously discussed model because it was difficulty to consider evolutionary information by probability graph model. So in first stage they three neural networks and SSMM as second stage. The three neural networks were fed by multiple sequence alignment profile to produce four types of outputs and SSMM take these output as input to produce final prediction results [25]. Zhang et al. proposed a model for PSSP, where they used three neural networks as first stage and a Bayesian model as second



(a) cascaded architecture

(b) cross structure

Figure 3.4: The structure of multiple methods

stage. The output from first stage is fed to second stage to obtain the final prediction results. This model is known as MDOAO+Bayesian model for PSSP [27].

2. Neural networks - SVM

In 2009, Anjum et al. proposed a hybrid PSSP method which included two different tertiary classifiers; the first classifier was neural network and the other was a granular decision tree. They employed SVM as the neurons in neural network architecture and input for SVM was formed by utilizing sliding multi-window scheme. The neural network architecture was optimized by genetic algorithms. The other of this method was a granular decision tree based on granular computing, decision tree and SVM [26]. Two methods: NSVMps and NSVM, based on Neuro-SVM were proposed by Pradip et al. in 2013. NSVMps utilized position-specific probability-based features where NSVM utilized position-independent probability-based feature. The two methods employed single-sequence as its turning data and did not used any sequence alignment profile information [28].

3. Neural networks - Others

Neural network is potential element in hybrid prediction models. It is combined with other methods besides previously discussed methods, such as fuzzy k-nearest neighbor algorithm, knowledge base, cascaded nonlinear components analysis and conditional random fields.

- Neural network - fuzzy k-nearest neighbor algorithm

  The fuzzy k-nearest neighbors algorithm (fuzzy k-NN) is fuzzy sets based method used for protein secondary structure classification where input consists of local sequence similarity of the segments of known proteins. The success of this method depends on existence of well-defined sequence profile in the query protein. It is possible to build a sequence database like PSSM with many similar sequences that boosts the performance of sequence profile based methods. Bondugula et al. used this reasoning and presented a MUPRED system. This method used two types of features: the first was membership values of each residue in the 3-state structure classes generated by fuzzy k-nearest neighbor algorithm, and the second was normalized PSSM. A feed-forward neural network was trained by standard back-propagation learning method with these two types of feature as input. This method used the information from both the sequence and structure databases and yielded better performance [29].

- Neural network - conditional random fields

  It is challenging to receive much attentions in 8-state prediction than 3-state prediction especially when there are few homologous sequences. To extract the inter-dependency among adjacent secondary structures, Wang et al. proposed conditional neural fields (CNFs) based method. CNFs are the combination of conditional random fields (CRFs) and neural networks (NNs). This method used probabilistic model similar to HMM for 8-states PSSP. Neural network is added to this model to extract the nonlinear features between protein sequence and its corresponding secondary structure. They combined the advantages of the two models to extract the complex relationship between sequence features and secondary structures, and the inter-dependency among adjacent residues. Because of CNFs stage, it was possible to output a probability distribution over all the possible secondary structures types and non-evolutionary information of proteins [1].

- Neural network - knowledge base

  Patel et al. proposed a two stage hybrid model for PSSP named KB-PROSSP-NN which used knowledge based method (KB-PROSSP) as first stage and neural network as second stage. The first stage of this model was again consist of two phases: the first one was KB-PROSSP which was used to extract the statistics of association between the 5-residue words and corresponding secondary structures by using hierarchical lateral-validation technique and the second one was a pre-trained feed-forward neural network (BPNN) which was used to correct the discrepancies of the knowledge base. The second stage of the model i.e., neural network was trained by the output of of first stage to predict the actual protein secondary structure [30].

- Neural network - cascaded nonlinear components analysis

  Botelho et al. proposed a hybrid model with reduced complexity named as (C-NLPCA). They reduced the high complexity of the prediction model by using effective pre-processing of input data set which also contributed enough for the improvement of accuracy of the model. The pre-processing of the data set was carried by Principal components analysis (PCA), an algorithm that can effectively provide to statistical analysis of the data. This method utilized the nonlinear potential of neural networks and reduced dimensionality of protein data to acquire the useful information for classification phase. Three neural networks with different topologies were trained by the output of PCA using resilient propagation method and the output from these networks were combined for the attainment of better classification results [31].

4. Other hybrid models

In addition to above mentioned hybrid models, there are some other mixed-breed methods, such as knowledge discovery in databases model, SVM-decision tree and k-mers - CRF etc. Short descriptions of these models are provided below.

- Knowledge discovery in databases model

  The general single-method models and hybrid models which are the combinations of simple prediction methods may not obtain satisfactory prediction results for PSSP and other non-trivial problems.To remedy this shortcoming, the data mining technologies receive a good recognition in recent years and have a good application prospect in PSSP. In 2002, Yang et al. proposed several compound pyramid models (CPM) for PSSP based on data mining and machine learning approaches. This compound model was made up of several layers that took multi-hierarchical configuration looking like pyramid shape. Each layer in the hierarchy works in close coordination with neighbouring layers and adopts a gradually refining focused on independent functions [32]. The core technology in CPM is the knowledge discovery in databases (KDD*) process. The two most essential mechanisms of KDDs are heuristic coordinator which simulates the intention creation in cognitive psychology, so that the shortage of knowledge could be detected by the system itself and a maintaining coordinator. Yang et al. proposed one of such CPM in 2009 which is shown in Figure 3.5

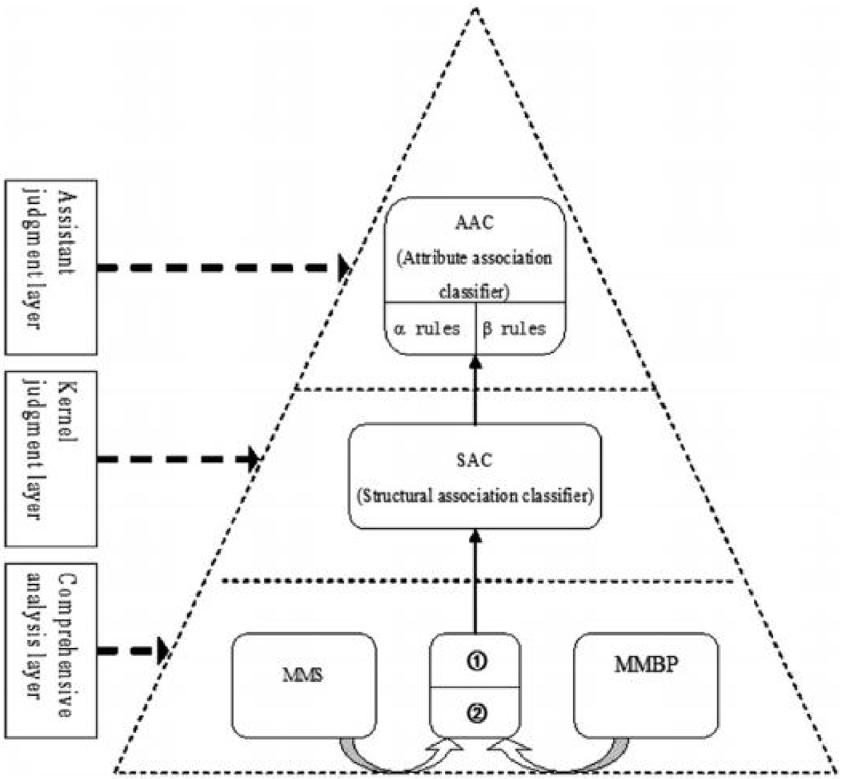

Assistant judgment layer
Kernel judgment layer
Comprehensive analysis layer
AAC
(Attribute association classifier)
α rules
β rules
SAC
(Structural association classifier)
MMS
①
②
MMBP

Figure 3.5: The compound pyramid model (CPM)

This was a four layers CPM where KDD* association analysis protein secondary structure prediction (KAAPRO) was proposed based on two mining algorithms: Maradbcm algorithm which was induced by KDD* model and D-CBA which was derived from a classical CBA algorithm based on complex measure. KAAPRO worked as the kernel in CMP which was followed by two SVM layers of CMP [33].

- SVM - decision tree

We have previously seen that SVM have higher potential in PSSP. Unfortunately this method was not used any comprehensible model so that prediction accuracy of SVM did not reflected properly. Decision tree is a good complement to SVM and as a result can be used combinedly to produce a comprehensible model. In 2002, He et al. presented a hybrid model named SVM_DT that combined the advantages of two divergent models, the strong generalization ability of SVM and the strong comprehensibility of rule induction of decision tree. In this method, SVM was used as an encoder that preprocess data for decision tree by combining orthogonal matrix and BLOSUM62 matrix. The output from this stage is used to train decision tree learning system. This hybrid model perform much better than single methods based SVM model [34].

- K-mers - conditional random field (CRF)

  A general framework based on k-mers and CRF was proposed by Madera et al. which could take into account variable range interactions among amino acid sequence. They considered all possible range interactions, i.e., short range, medium range and longer range. From real sequence of amino acid, this method measured the occurred frequency of each individual amino acid (1-mer), each possible pair (2-mer) or every combination of up to k amino acids (k-mer). This k-mer model was combined with markov chain model to generate the realistic secondary structure profiles. The longer range interactions in the amino acid sequence were focused in k-mer order model. Katzman et al. used PREDICT-2ND neural networks to generate two local structure profiles for each protein sequences. The former one was to described the amino acid at each position of the target sequence, and the latter one to from the alignment [35]. These local structure were used for PSSP [36].

Chapter 4

State Of The Art For 8-State PSSP

4.1 8-State PSSP Models

Prediction of the three states from protein sequences (i.e., the Q3 prediction problem) has been intensively investigated for decades using many machine learning methods. Recently, the focus of secondary structure prediction has been shifted from Q3 prediction to the prediction of 8-state secondary structures (i.e., the Q3 prediction problem), due to the fact that a chain of 8-state secondary structures contains more precise structural information for a variety of applications. 8-state PSSP is much more complicated than 3-state PSSP. Scholars have introduced many deep learning methods to tackle this problems. Some significant works widely explored by researchers are

- SC-GSN network
- Bidirectional long short-term memory (BLSTM) method
- Deep conditional neural field
- DCRNN
- Next-step conditioned deep convolutional neural network(CNN)
- Deep inception-inside-inception (Deep3I) network etc.

Abstract description of these models are presentd in the following sections.

4.1.1 8-Class PSSP Using CNFs

Unavailability of large dataset made it much more challenging to predict 8-class secondary structure (SS) compared to the 3-class prediction. In 2011, Zhiyong et al. presented a new probabilistic method for 8-class SS prediction using conditional neural fields (CNFs). This CNF method served two purposes: one is to model the complex relationship between sequence features and SS, the other is to exploit the inter-dependency among SS types of adjacent residues. Both the sequence profiles and non-evolutionary information are given as input to this prediction model. They used position-dependent feature, PSSM and position-independent features such as, physico-chemical property, correlated contact potential of amino acids, and primary sequence. CNFs is depicted in Figure 4.1

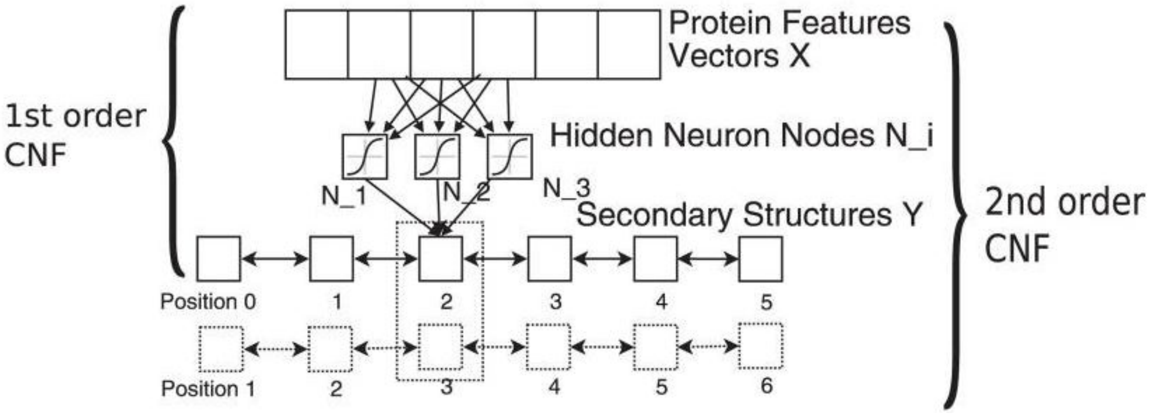


Figure 4.1: The CNF model for 8-class protein SS prediction. [1]

The authors used CullPDB data set for training and two public benchmarks CB513 and RS126 to test the performance of their model. They reported Q8 accuracy of 64.9 and 64.7%, for these two benchmarks respectively.They also endorsed that this method can be used to predict other structure properties (e.g. solvent accessibility) of a protein or the SS of RNA [1].

4.1.2 Deep Supervised and Convolutional GSN

Generative stochastic network (GSN) is a framework based on learning the transition operator of a Markov chain to sample from a conditional distribution, which is used to globally train deep generative models. Zhou et al. presented a GSN based method to predict local secondary structure of proteins with deep hierarchical representations. Protein sequence is a full-sized, high-dimensional dataset. To allows efficient learning across multiple layers of hierarchical representations they introduce a convolutional architecture to scale the model to this dataset. This architecture uniquely focuses on predicting structured low-level labels informed with both low and high-level representations which corresponds to labeling the secondary structure state of each amino-acid residue.

GSN model is fed by evolutionary information i.e., PSSM along with original protein sequence for amino-acid residues encoded by $n \times b$ binary matrices as input features. They used 6128 proteins from PISCES CullPDB dataset, and divided it randomly into training (5600), validation

(256), and testing (272) sets. This model achieved 66.4% Q8 accuracy on the CB513 benchmark dataset [37].

4.1.3 PSSM Using DeepCNF

Deep Convolutional Neural Fields (DeepCNF) is a Deep Learning extension of Conditional Neural Fields (CNF). It combines the advantages of both CNF and deep convolutional neural networks. DeepCNF can capture both the complex sequence-structure relationship and interdependency between adjacent secondary structure labels. In 2016, Wang et al. presented a PSSP model based on DeepCNF which outperform most of the popular predictors. A graphical depiction is shown in Figure 4.2.

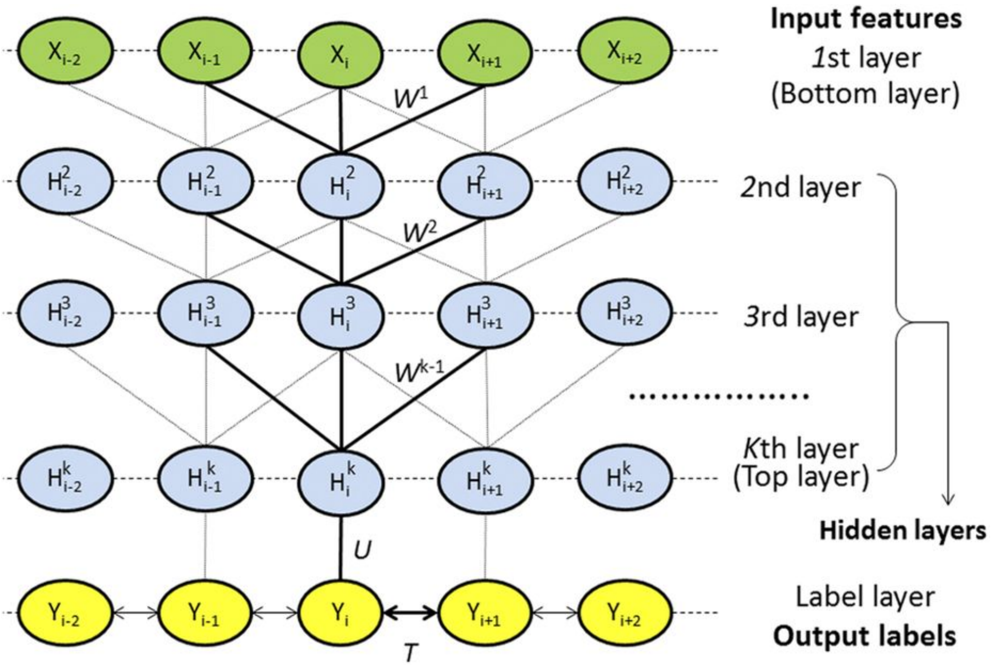


Figure 4.2: The architecture of DeepCNF, where i is the residue index and X_i the associated input features, H^k represents the k-th hidden layer, and Y is the output label [2].

They ran PSI-BLAST to generate PSSM and then transformed each PSSM entry by sigmoid function and used it as input feature. Besides they used a binary vector of 21 elements to indicate the amino acid type at each residue position. They used five publicly available datasets: CullPDB53, CB513, CASP1054 and CASP1155 datasets. This model obtained 84% Q3 accuracy, 85% SOV score, and 72% Q8 accuracy, respectively, on the CASP and CAMEO test proteins [2].

4.1.4 PSSM Using DCRNN

Zhen et al. proposed deep convolutional and recurrent neural network (DCRNN), an end-to-end deep network architecture that leverages convolutional neural networks with different kernel sizes to extract multi-scale local contextual features for prediction of protein secondary structures. They set up a bidirectional neural network consisting of gated recurrent unit to capture global contextual features in addition to the long-range dependencies existing in amino acid sequences. DCRNN consists of four parts:

- One feature embedding layer
- Multiscale convolutional neural network (CNN) layers
- Three stacked bidirectional gated recurren unit (BGRU) layers
- Two fully connected hidden layers.

The feature embedding layer was used to transform sparse sequence feature vectors into denser feature vectors in a new feature space. Both the sequence features and profile features were used as input were fed into multiscale CNN layers with different kernel sizes to extract multiscale local contextual features. The three stacked BGRU layers were used to capture global contexts and the two fully connected hidden layers were used to output 8-category secondary structure and 4-category solvent accessibility classification. They used PSSM as input feature and four publicly available datasets, CB6133,CB513, CASP10, and CASP11 to evaluate the performance of DCRNN. They claimed that DCRNN gained 69.7% Q8 accuracy on the public benchmark CB513, 76.9% Q8 accuracy on CASP10 and 73.1% Q8 accuracy on CASP11 [38].

4.1.5 NCCNN Improve PSSM

Understanding of Convolutional Neural Network (CNN) is the most popular deep learning method for analyzing visual imagery and various speech recognition systems. Busia et al. presented a chained convolutional architecture with next-step conditioning CNN (NCCNN) for improving performance on PSSP problems. They integrated general purpose improvements in the field of deep learning e.g., batch normalization (BN), dropout and weight-norm constraint, residual connections, multi-scale convolutional filters.

The authors used two publicly available benchmark datasets: CullPDB and CB513. They used amino acids sequence of each protein and PSSM which they normalized via mean-centering and scaling by the standard deviation. The performance of a single model is reported as 70.3% Q8 accuracy and the performance of an ensemble of these models is 71.4% Q8 accuracy, both on CB513 dataset [3].

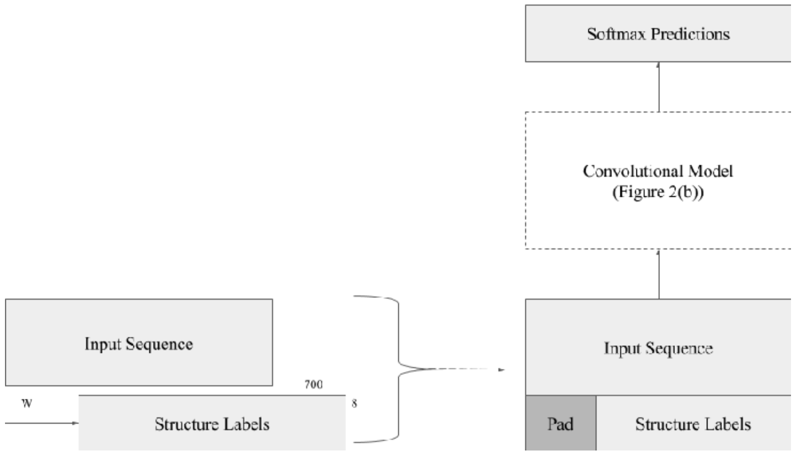


Figure 4.3: Introduction of next-step conditioning on previous labels by including a shifted copy of the input sequences SS labels as input to the convolutional model [3].

4.1.6 CNNH_PSS

In 2018, Jiyun et al. introduced highway in multi-scale CNN and presented a novel deep learning based model, referred to as CNNH_PSS. In this model highway is a path from current layer to the output of the next one to deliver information between two neighbor convolutional layers for the purpose of capturing both local and contexts. In multi-scale CNN, lower layers extract local context while higher layers extract long-range interdependencies. The highways between neighbor layers able to extract both local contexts and long-range interdependencies. The frame of CNNH_PSS is shown in Figure: 4.4.

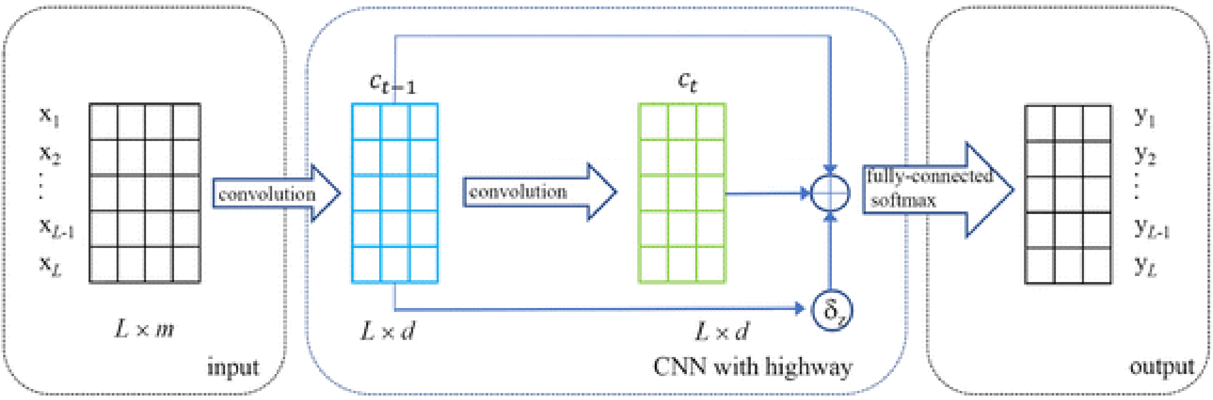


Figure 4.4: A deep inception network consisting of three inception modules, followed by one convolution and two fully-connected dense layers. [4].

This method used m-dimensional feature vector for each residue of protein sequences along with position-specific scoring matrix, PSSM as input feature. Performane of CNNH_PSS is evaluated on CB6133 and CB513 datasets. Multi-scale CNN produce 0.729% on CB6133 and 0.693% on CB513 dataset. This model outperforms the multi-scale CNN without highway by at least 0.010% Q8 accuracy [4].

4.1.7 MUFOLD-SS

One of the latest state of the art fot 8-state PSSP was carried by Chao et al. in 2018. They proposed a new deep neural network architecture, named the Deep inception-inside-inception (Deep3I) network (Fig 4.5) for PSSP. They claimed that the new method outperforms the best existing methods and other deep neural networks significantly. Afterward, they implemented a software tool named MUFOLD-SS. The input for this software is the feature matrix corresponding to the primary amino acid sequence of a protein. This tool effective processes the local and global interactions between amino acids and predict the protein secondary structure fast and accurately.

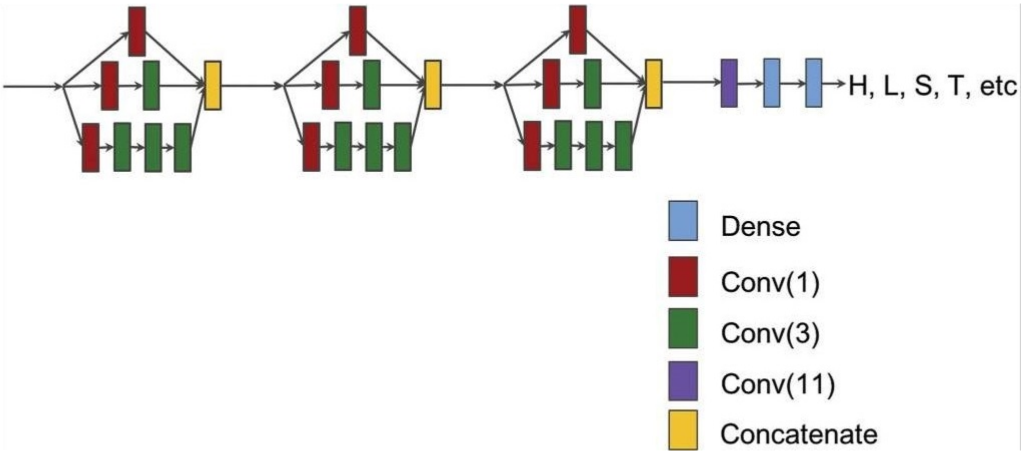


Figure 4.5: A deep inception network consisting of three inception modules, followed by one convolution and two fully-connected dense layers. [5].

This model uses feature matrix composed of physio-chemical properties of amino acids, PSI-BLAST profile, and HHBlits profile. Five publically available datasets: CullPDB, JPRED, CASP, CB513 benchmark and PDB (Protein Data Bank) were used to measure the effectiveness of this method. MUFOLD-SS achieved 76.47%, 74.51%, and 72.1% Q8 accuracy on CASP10, CASP11, CASP12 datasets respectively [5].

4.2 Performance Evaluation Of PSSP Models

A comparison among the performance of different 8-state PSSP models on different benchmark dataset is shown in Table 4.1.

Here we also incorporate a performance comparison among different learning methods on CB513 and CAPS datasets in Table 4.2.

Method	CB513	CASP10	CASP11	CASP12
GSN	66.4	-	-	-
BLSTM	67.4	-	-	-
DeepCNF	68.3	71.8	71.7	0.694
DCRNN	69.7	-	-	-
NCCNN	70.3	-	-	-
MUFOLD-SS	70.5	74.2	71.6	69.5
CRRNN	71.4 ± 0.2	73.8 ± 0.5	71.6 ± 0.7	68.7 ± 0.8

Table 4.1: A comparison of the Q8 accuracy(%) on CB513, CASP10, CASP11 and CASP12 among state-of-the-art methods.

Method	CASP10	CASP11	CASP12	CB513
PSIPRED	81.2	80.7	80.5	79.2
JPRED	81.6	80.4	78.8	81.7
DeepCNF	84.4	84.7	83.2	82.3
DCRNN	-	-	-	84
NCCNN	-	-	-	-
CRRNN	86.1 ± 0.6	84.2 ± 0.5	82.6 ± 1.2	85.3 ± 0.4
eCRRNN	87.8	85.9	83.7	87.3

Table 4.2: Q3 accuracy(%) comparison on CB513 and CASP datasets.

Chapter 5

Our Experiments

We first experminted with a hybrid model consisting of boths Convolutional Networks (CNN) and Recurrent Neural Networks (RNN). We did this as CNNs can better capture the local spatial features (short-range interaction) and RNNs can better capture the global features (long-range interaction). However, RNNs often fail to capture global features due to its vanishing and exploding gradient problem. Long Short Term Memory (LSTM) and Gated Reccurent Unit (GRU) are two variants of RNN that solves these problems with gate mechanism. Both LSTM and GRU are proved to provide similar performance whereas GRU works a bit faster than LSTMs. So, we used GRUs : a improvement over RNN in our experiment. As input data is one-dimensional in our case, we used 1DCNNs and as global dependency can be in both direction, we used BGRUs (Bidirectional GRUs).

5.1 Dataset

We used parts of CB6133 for training and validation. For testing we used both parts CB6133 and CB513 dataset. The brief introduction of the datasets are given in Section 2.3 of Chapter 2. Here we will provide more on the dataset. CB6133 dataset was developed by Zhou and Troyanskaya in 2014 that contained 6128 non-homologous protein sequences (after filtering). This is produced with PISCES Cull PDB server (Wang & Dunbrack, 2003) which is commonly used for evaluating structure prediction algorithms. They retrieved a subset of solved protein structures with better than 2.5A resolution and less than 30% identity and the same set was used in the previous work (Wang et al., 2011). They also removed protein chains with less than 50 or more than 700 amino acids or discontinuous chains. They inferred 8-states secondary structure labels and solvent accessibility score from the 3D PDB structure by the DSSP program (Kabsch & Sander, 1983). They discretized solvent accessibility scores to absolute solvent accessibility and relative solvent accessibility following (Qi et al., 2012). To generate PSSM, they ran PSI-

BLAST against UniRef90 database with threshold 0.001 and 3 iterations. We used pfilt program from PSI-PRED package to pre-filter the database for removing low information content and coiled-coil like regions (Jones, 1999). To use PSSM as the input for the neural network models, they transformed the PSSM scores to 0-1 range by the sigmoid function. The resulting training data including both feature and labels has 57 channels (22 for PSSM, 22 for sequence, 2 for terminals, 8 for secondary structure labels, 2 for solvent accessibility labels), and the overall channel size is 700. The 700 amino-acids length cutoff was chosen to provide a good balance between efficiency and coverage as the majority of protein chains are shorter than 700AA. Proteins shorter than 700AA were padded with all-zero features. Both the dataset was downloaded form https://www.princeton.edu/~jzthree/datasets/ICML2014/ which is publicly available and most the works afterwards used this public dataset due to their' ease of access and robustness.

5.1.1 Data Processing

The data was available in numpy format which made our experments much easier. However, it was in (N protein × k features format. We reshaped it into N proteins ×700 amino acid ×57 features. The 57 features and their value types are described below :
[0,22): amino acid residues, with the order of 'A', 'C', 'E', 'D', 'G', 'F', 'I', 'H', 'K', 'M', 'L', 'N', 'Q', 'P', 'S', 'R', 'T', 'W', 'V', 'Y', 'X','NoSeq' . This is a one-hot encoded vectorer. Only one value is 1 at a row in this vector indicating the amino acid residue. Others are set to zero.
[22,31): Secondary structure labels, with the sequence of 'L', 'B', 'E', 'G', 'I', 'H', 'S', 'T','NoSeq' This is also a one-hot encoded vector. Only one value is set to one indicating the secondary structure of the corresponding amino acid residue. Others are set to one. This is the output of the model.
[31,33): N- and C- terminals. This is also one-hot encoded. However, we have not used this as input or output in our prediction algorithm. [33,35): relative and absolute solvent accessibility, used only for training. (absolute accessibility is thresholded at 15; relative accessibility is normalized by the largest accessibility value in a protein and thresholded at 0.15; original solvent accessibility is computed by DSSP)
[35,57): sequence profile against each type of amino acid residues. Hence the order of amino acid residues against which scoring hit value is stored is ACDEFGHIKLMNPQRSTVWXY and it is different from the order for amino acid residues. The values are output of a sigmoid function are in 0-1 value range.
The 'NoSeq' feature of both amino acid residues and secondary structure labels just mark end of the protein sequence.

5.2 Input Features

The input features we used in our experiments are sequence information via single sequence (captured by the one-hot vector [0,22) for each amino acid residue) , evolutionary information (captured by the sequence profile vector [35,57) against each amino acid residue.) and solvent accessibility (captured by [33,35) vector) of each amino acid. So, our input is a combination of sequence information, evolutionary information and one physio-chemical property.

5.3 Model

We built a hybrid model for in our experiment as hybrid models are now trendy as they prove to give higher accuracy. Hence, for capturing various lengths of short range interactions, we used 3-mers, 7-mers and 11-mers each retrieved by a 1D convolutional network. Then by concatenating them, we used all of them as input feature for the next Bidirectional Gated Recurrent Unit (BGRU). We used 2 stacked BGRUs to effectively capture the long-range interactions among residues.

5.4 Integration of Attention Mechanism

Attention mechanism is a method of giving focus or attention to some particular location or area while decoding information from one domain to another one. It is developed inspired by the visual attention system of humans. Our visual system focuses on a particular object of location corresponding to our visual input. This process is largley studied in neuroscience and recently steps are made to incorporate this into neural network models. This mecchanism is beleived to convey large impact on many machine learning based methods like speech recognition, translation, reasoning, image captioning and visual identification of objects.

5.4.1 What is Attention Mechanism

Attention mechanism works by selectively processing data by focusing on the parts or segments of data that is the most important. By 'most important' we refer that part of the input that is contributing most for a particular output.
An attention model is a method that takes n arguments $y_1, ..., y_n$ (in the precedent examples, the $y_i$would be theh_i), and a context c. It return a vector z which is supposed to be the summary of the y_i, focusing on information linked to the context c. More formally, it returns a weighted arithmetic mean of they_i, and the weights are chosen according the relevance of each y_i given

the context c.

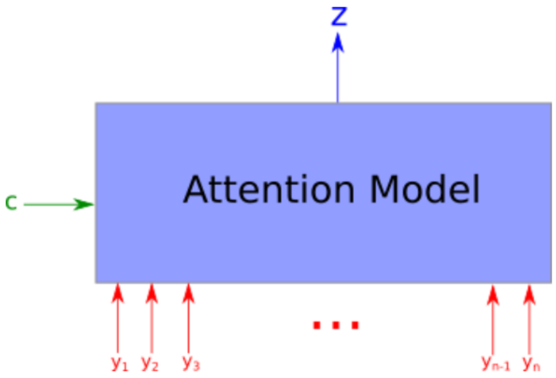


Figure 5.1: An attention unit

Vaswani et al. in 2017, in their famous work in machine translation with attention, described attention function as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors [39]. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. In their work, they provided two different variants of attention mechanism. One is scale dot product matrix attention that performs single attention function on each of the keys and another is multiheaded attention that linearly project the queries, keys and values with different projection to different dimension.

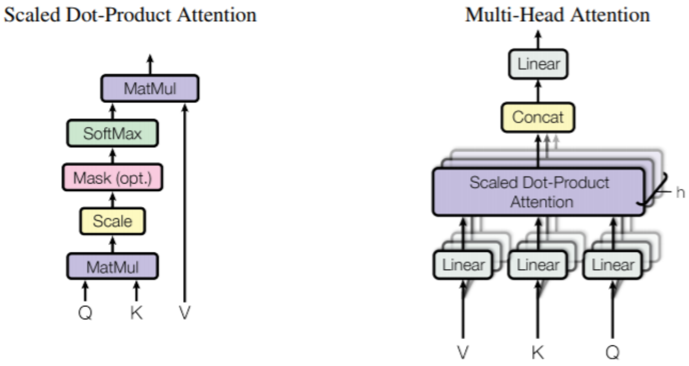


Figure 5.2: Different types of attention mechanism proposed by Vaswani et al.

5.4.2 Various attention models

Attention models can be mainly divided into two classes.

- Soft Attention Model
- Hard Attention Model

Soft Attention Model

This version is easier to understand. The attention model is softly-chooses the variable the most correlated with the context. It is a fully differentiable deterministic mechanism that can be plugged into an existing system, and the gradients are propagated through the attention mechanism at the same time they are propagated through the rest of the network.

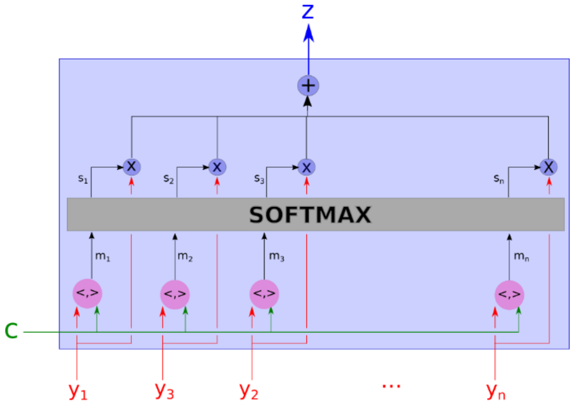


Figure 5.3: Soft Attention Model

Hard Attention Model

Hard attention is a stochastic process: instead of using all the hidden states as an input for the decoding, the system samples a hidden state y_i with the probabilities s_i. In order to propagate a gradient through this process, we estimate the gradient by Monte Carlo sampling.

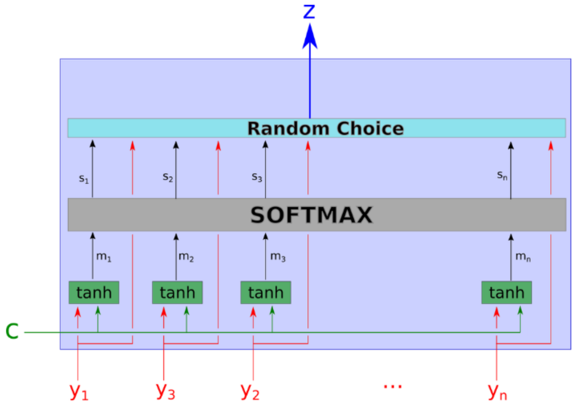


Figure 5.4: Hard Attention Model

5.4.3 Possibilities in our context

The field of attention mechanism is new. However, till now the models using attention method usually works with RNN layer. As RNN is used for capturing global features mostly, attention make their tasks easy by their mechanism. The machine learning methods that begets the use of RNN e.g. Nueral Machine Translation, Image Captioning etc. have proved to give better performance with attention mechanism. As RNN is used for capturing global features or long range interactions, attention mechanism may come handy in this regard.

5.5 Results

Our experimented model gives 65% Q8 accuracy in benchmark CB513 dataset. However, we trying to incorporate and design special attention mechanism in PSSP and we believe it may give us better results along with a good insight in PSSP. For limitation of time and less available resources , this work has not finished yet and it is still an ongoing research.

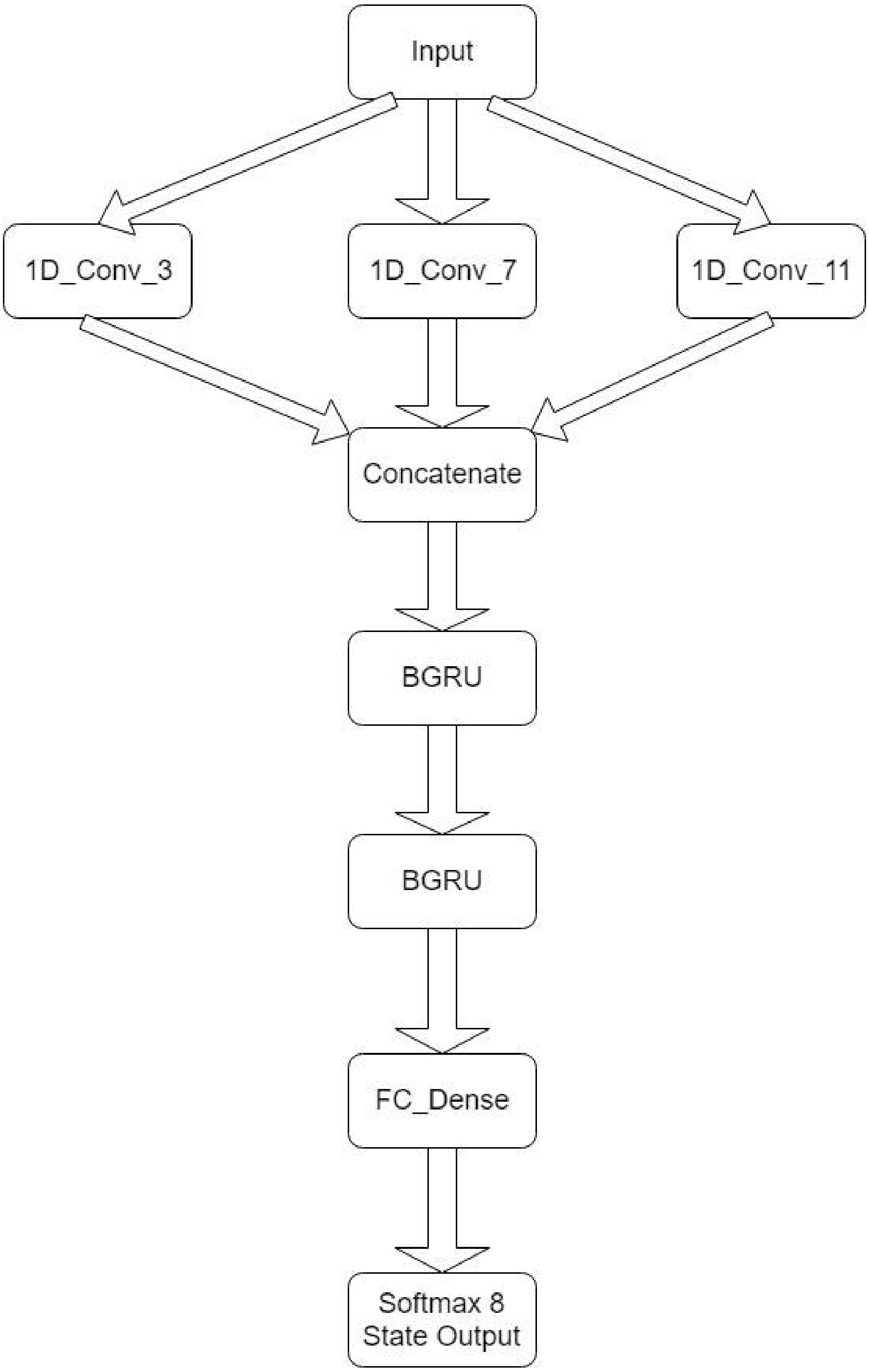


Figure 5.5: Model Architecture of our experiment

Chapter 6

Conclusion

Protein secondary structure prediction is an extensively explored research topic in bioinformatics due to its significance as a fundamental task to learn about protein 3-D structures and functions. Knowledge on protein functions are crucial for disease detection and drug discovery. Also for a large number of discovered protein sequences we need to rely on prediction based methods to know their structures as the number of known structures is far beneath the number of known protein sequences. Though numerous methods have been proposed for PSSP already, still it can not meet the actual need. Also, for precise knowledge, 8 state prediction is necessary where a little has been done in cost of a lot of complex technologies. So, advancing state of the art has become more challenging. However, sometimes simpler model can give us better insights and results. We are proceeding keeping that in mind.

This book provides a overview of the state of the art to learn the latest works in this field. Discussion on popular machine learning tools have also been provided in brief. Recent trends on hybrid models and deep networks have been focused in the literature review discussion.

However, due to shortage of resources and time, we could not come up with a better model than the state of the art. But this is an ongoing work and now we are trying to incorporate the very recently explored attention mechanism in our model that has been proven to be a very effective one in many real life machine learning problems like NMT, Image Captioning etc. to capture global features. Since PSSP has some similarity with those works in some aspects, we assume that we may achieve better result and also better insight via incorporating and designing special attention mechanism for PSSP in future. This book ends with a recommendation of incorporating attention mechanism in the state of the art works on PSSP and ultimately tries to promote the overall development of PSSP.

References

[1] Z. Wang, F. Zhao, J. Peng, and J. Xu, “Protein 8-class secondary structure prediction using conditional neural fields,” in *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 109–114, Dec 2010.

[2] S. Wang, J. Peng, J. Ma, and J. Xu, “Protein secondary structure prediction using deep convolutional neural fields,” in *Scientific reports*, 2016.

[3] A. Busia and N. Jaitly, “Next-step conditioned deep convolutional neural networks improve protein secondary structure prediction,” *arXiv preprint arXiv:1702.03865*, 2017.

[4] J. Zhou, H. Wang, Z. Zhao, R. Xu, and Q. Lu, “Cnnh_pss: protein 8-class secondary structure prediction by convolutional neural network with highway,” *BMC bioinformatics*, vol. 19, no. 4, p. 60, 2018.

[5] C. Fang, Y. Shang, and D. Xu, “Mufold-ss: New deep inception-inside-inception networks for protein secondary structure prediction,” *Proteins: Structure, Function, and Bioinformatics*, vol. 86, no. 5, pp. 592–598, 2018.

[6] S. A. Malekpour, S. Naghizadeh, H. Pezeshk, M. Sadeghi, and C. Eslahchi, “A segmental semi markov model for protein secondary structure prediction,” *Mathematical biosciences*, vol. 221, no. 2, pp. 130–135, 2009.

[7] S. Wang, J. Peng, J. Ma, and J. Xu, “Protein secondary structure prediction using deep convolutional neural fields,” *Scientific reports*, vol. 6, p. 18962, 2016.

[8] J. Moult, K. Fidelis, A. Zemla, and T. Hubbard, “Critical assessment of methods of protein structure prediction (casp)-round v,” *Proteins: Structure, Function, and Bioinformatics*, vol. 53, no. S6, pp. 334–339, 2003.

[9] M. Spencer, J. Eickholt, and J. Cheng, “A deep learning network approach to ab initio protein secondary structure prediction,” *IEEE/ACM transactions on computational biology and bioinformatics (TCBB)*, vol. 12, no. 1, pp. 103–112, 2015.

[10] J. Lee, “Measures for the assessment of fuzzy predictions of protein secondary structure,” *Proteins: Structure, Function, and Bioinformatics*, vol. 65, no. 2, pp. 453–462, 2006.

[11] P. Kountouris and J. D. Hirst, "Prediction of backbone dihedral angles and protein secondary structure using support vector machines," *BMC bioinformatics*, vol. 10, no. 1, p. 437, 2009.

[12] Z. Aydin, Y. Altunbasak, and M. Borodovsky, "Protein secondary structure prediction for a single-sequence using hidden semi-markov models," *BMC bioinformatics*, vol. 7, no. 1, p. 178, 2006.

[13] C. Chen, Y. Tian, X. Zou, P. Cai, and J. Mo, "Prediction of protein secondary structure content using support vector machine," *Talanta*, vol. 71, no. 5, pp. 2069–2073, 2007.

[14] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, "An overview of topic modeling and its current applications in bioinformatics," *SpringerPlus*, vol. 5, no. 1, p. 1608, 2016.

[15] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, and Y. Zhou, "Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning," *Scientific reports*, vol. 5, p. 11476, 2015.

[16] S. Babaei, A. Geranmayeh, and S. A. Seyyedsalehi, "Protein secondary structure prediction using modular reciprocal bidirectional recurrent neural networks," *Computer methods and programs in biomedicine*, vol. 100, no. 3, pp. 237–247, 2010.

[17] J. Chen and N. S. Chaudhari, "Bidirectional segmented-memory recurrent neural network for protein secondary structure prediction," *Soft Computing*, vol. 10, no. 4, pp. 315–324, 2006.

[18] R. Heffernan, Y. Yang, K. Paliwal, and Y. Zhou, "Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility," *Bioinformatics*, vol. 33, no. 18, pp. 2842–2849, 2017.

[19] F. Birzele and S. Kramer, "A new representation for protein secondary structure prediction based on frequent patterns," *Bioinformatics*, vol. 22, no. 21, pp. 2628–2634, 2006.

[20] G. Karypis, "Yasspp: better kernels and coding schemes lead to improvements in protein secondary structure prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 64, no. 3, pp. 575–586, 2006.

[21] P. Chatterjee, S. Basu, M. Kundu, M. Nasipuri, and D. Plewczynski, "Psp_mcsvm: brainstorming consensus prediction of protein secondary structures using two-stage multiclass support vector machines," *Journal of molecular modeling*, vol. 17, no. 9, p. 2191, 2011.

[22] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, (Berlin, Heidelberg), pp. 1–15, Springer Berlin Heidelberg, 2000.

[23] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning*. Springer Publishing Company, Incorporated, 1st ed., 2011.

[24] N. P. Bidargaddi, M. Chetty, and J. Kamruzzaman, "Combining segmental semi-markov models with neural networks for protein secondary structure prediction," *Neurocomput.*, vol. 72, pp. 3943–3950, Oct. 2009.

[25] Q. Jiang, X. Jin, S.-J. Lee, and S. Yao, "Protein secondary structure prediction: A survey of the state of the art," *Journal of Molecular Graphics and Modelling*, vol. 76, pp. 379 – 402, 2017.

[26] A. Reyaz-Ahmed, Y. Zhang, and R. W. Harrison, "Granular decision tree and evolutionary neural svm for protein secondary structure prediction," *Int. J. Comput. Intell. Syst.*, vol. 2, pp. 343–352, 2009.

[27] B. Zhang, Z. Chen, and Y. L. Murphey, "Protein secondary structure prediction using machine learning," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 1, pp. 532–537 vol. 1, July 2005.

[28] P. GHANTY, N. R. PAL, and R. K. MUDI, "Prediction of protein secondary structure using probability based features and a hybrid system," *Journal of Bioinformatics and Computational Biology*, vol. 11, no. 05, p. 1350012, 2013. PMID: 24131056.

[29] R. Bondugula and D. Xu, "Mupred: a tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 66, no. 3, pp. 664–670, 2007.

[30] M. S. Patel and H. S. Mazumdar, "Knowledge base and neural network approach for protein secondary structure prediction," *Journal of theoretical biology*, vol. 361, p. 182189, November 2014.

[31] S. Botelho, G. Simas, and P. Silveira, "Prediction of protein secondary structure using nonlinear method," in *Neural Information Processing* (I. King, J. Wang, L.-W. Chan, and D. Wang, eds.), (Berlin, Heidelberg), pp. 40–47, Springer Berlin Heidelberg, 2006.

[32] B. Yang, H. Sun, and F. Xiong, "Ming quantitative association rules with standard sql queries and its evaluation," *J Comput Res Dev*, vol. 39, no. 3, pp. 307–312, 2002.

[33] B. Yang, W. Qu, Y. Zhai, and H. Sui, "An approach of protein secondary structure prediction based on homology analysis method in compound pyramid model," in *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on*, vol. 1, pp. 450–454, IEEE, 2010.

[34] J. He, H.-J. Hu, R. Harrison, P. C. Tai, and Y. Pan, "Rule generation for protein secondary structure prediction with support vector machines and decision tree," *IEEE Transactions on NanoBioscience*, vol. 5, pp. 46–53, March 2006.

[35] S. Katzman, C. Barrett, G. Thiltgen, R. Karchin, and K. Karplus, "Predict-2nd: a tool for generalized protein local structure prediction," *Bioinformatics*, vol. 24, no. 21, pp. 2453–2459, 2008.

[36] M. Madera, R. Calmus, G. Thiltgen, K. Karplus, and J. Gough, "Improving protein secondary structure prediction using a simple k-mer model," *Bioinformatics*, vol. 26, no. 5, pp. 596–602, 2010.

[37] J. Zhou and O. G. Troyanskaya, "Deep supervised and convolutional generative stochastic network for protein secondary structure prediction," *arXiv preprint arXiv:1403.1347*, 2014.

[38] Z. Li and Y. Yu, "Protein secondary structure prediction using cascaded convolutional and recurrent neural networks," *arXiv preprint arXiv:1604.07176*, 2016.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Generated using Undegraduate Thesis LaTeX Template, Version 1.3. Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh.

This thesis was generated on Saturday 19th October, 2019 at 1:56pm.