Deep learning neural network approaches for one-dimensional structure prediction from protein sequences

Shamima Banu Syed Mohamed Rashid

2018

Shamima Banu Syed Mohamed Rashid. (2018). Deep learning neural network approaches for one-dimensional structure prediction from protein sequences. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/73444>

<https://doi.org/10.32657/10356/73444>

Deep Learning Neural Network Approaches for One-dimensional Structure Prediction from Protein Sequences

A Thesis Submitted to the
School of Computer Science and Engineering
Nanyang Technological University, Singapore

by

**Shamima Banu binte Syed Mohamed Rashid
(G1103160C)**

for the fulfilment of the Degree of Doctor of Philosophy

Supervisor : Assoc. Prof. Suresh Sundaram

2018

Acknowledgements

I express sincere thanks to my supervisor Associate Professor Suresh Sundaram. He is my guru who gave immense guidance throughout my research works and supported me in my studies. Without his critical feedback, I would not have grown. I am also indebted to Associate Professor Kwoh Chee Keong for his encouragement and advice when I was working as a research assistant. He motivated me to apply for the PhD, at a time when I was uncertain and considering several options.

During my time in NTU, I gained many valuable friends such as Su Tran To Chinh, Divya Jain, Kartick Subramanian, Guo Jing and Liu Qian. They were a source of support for me and I had many engaging, intellectual discussions with them. I also gained valuable feedback that enabled me to improve my work. I thank Chinh and Kartick in particular, for project discussions and their critical comments to improve a manuscript that I had written.

I would also like to thank Dr. Savitha Ramaswamy and Abeegithan Jeyasothy for their great help in the use of neural networks applied to the problems presented in this work. In particular, I thank Mr. Abeegithan for modifications to the neural networks used, designing the sequence fitting block and for carrying out experiments. I also thank Dr. Saras Saraswathi for her patient help and collaboration over several datasets that have been of use in this thesis.

I am also indebted to the School of Computer Science and Engineering and Nanyang Technological University, for supporting me with a research scholarship. I also take this opportunity to thank Mr. Tan Sing Yau of the Biomedical Informatics Laboratory (BIL) and Mr. Kesavan Asaithambi of the Computational Intelligence Laboratory (CIL). Their technical support was of great assistance. I thank Mr. Sing Yau in particular for assisting in matters such as card access and problems with my network account, amongst many other things.

Finally, I thank my family members and friends for their kind words of encouragement and motivation throughout my studies.

Contents

1	Introduction	1
1.1	Description of Protein Structure	1
1.1.1	Primary Structure	2
1.1.2	Secondary Structure	3
1.1.3	Tertiary Structure	4
1.1.4	Membrane Protein Topology	5
1.1.5	Protein Data Bank (PDB)	7
1.2	Motivation for Protein Structure Prediction from Primary Sequence . .	8
1.2.1	A Schematic Representation of Secondary Structure and Mem- brane Topology Prediction	10
1.2.2	Challenges in protein SS prediction	12
1.2.3	Challenges in Membrane Protein Topology prediction	14
1.3	Objectives	17
1.3.1	Deep Belief Networks to extract higher level features	18
1.3.2	Feature Extraction using the CABS algorithm	18
1.3.3	Secondary Structure Prediction using a Complex-valued Neural Network	19
1.3.4	Heuristics based approach to detect a small training set	19
1.3.5	Biological Analysis of Predicted Structures	19
1.4	Contributions	20
1.4.1	Predicting Protein Structures with Deep Ensemble Networks . .	21
1.4.2	Predicting Protein Secondary Structures with a Fully Complex- valued Relaxation Network	22
1.4.3	Biological analysis of predicted structures	23
1.5	Organization	24
2	Literature Review	27
2.1	Topology Predictors for α -helical Membrane Proteins	28
2.1.1	Resources for Membrane Proteins	28
2.1.2	Early data driven approaches using single sequences	29
2.1.3	Multiple Sequence Alignment (MSA) and PSI-BLAST	30
2.1.4	Early data driven approaches using multiple sequences	31
2.1.5	Modern data driven approaches combining signal peptide (SP) with TM prediction	33
2.1.6	Other approaches applied to membrane protein predictions . . .	34
2.1.7	Non-topology Prediction methods	35
2.2	Modelling and Predicting Protein Secondary Structures	38
2.2.1	Secondary Structure prediction Based on Information Theory .	38
2.2.2	Secondary Structure Prediction based on Neural Networks . . .	40

2.2.3	Secondary Structure Prediction based on Support Vector Machines	43
2.2.4	Secondary Structure Predictors based on Hidden Markov Models	43
2.3	Deep Learning Neural Networks and their applications in Structure Prediction	47
2.4	Challenges in Secondary Structure Prediction and Membrane Topology Predictions	48
2.4.1	Theoretical upper Limit for Accuracy & Noise from Variation	48
2.4.2	Effects of PSSM based Representation	49
2.4.3	Modelling Structural Properties in Sliding Windows	49
2.5	Summary	51
3	Membrane Protein Topology Prediction using Deep Belief Networks	53
3.1	Introduction	53
3.2	Datasets	55
3.3	Feature Encoding: Generating PSSM for Sequences	57
3.4	Architecture of HiDEN	59
3.5	Sequence Fitting Block	63
3.6	Performance Measures	65
3.7	Comparison of HiDEN Performance	66
3.8	HiDEN Performance Discussion	69
3.8.1	Sequence similarities and Cross-Validation	72
3.9	Effectiveness of TM Helix Prediction	74
3.10	Summary	78
4	Protein Secondary Structure Prediction using Deep Belief Networks (DBN)	80
4.1	Introduction	80
4.2	Dataset	81
4.3	PSSM based Feature Encoding and Performance Measures	81
4.4	Architecture and Learning Algorithm of Deep Belief Network	82
4.5	Performance of Deep Belief Network	85
4.5.1	Motivation for using CABS energy potential based Feature Encoding with DBN	87
4.6	Summary	91
5	Protein Secondary Structure Prediction using a Compact Model	93
5.1	Introduction	93
5.1.1	Compact Model	94
5.2	Datasets	98
5.2.1	CB513	98
5.2.2	G Switch Proteins (GSW25)	99
5.2.3	Secondary Structure Assignments	99
5.3	Methods	100
5.3.1	Feature Generation using the CABS algorithm	100
5.3.2	Fully Complex-valued Relaxation Network Classifier	102
5.3.3	Heuristics based Approach to derive the Compact Model	107
5.4	Performance Measures	112
5.5	Performance Study of FCRN	114
5.6	Performance Study of Compact Model Using FCRN	116
5.7	Structure-based insights from Analysis of Predictions	123

5.8	Summary	131
6	Conclusions and Future Works	133
6.1	Conclusions	133
6.1.1	Summary of Findings in Membrane protein Topology Prediction	134
6.1.2	Summary of Findings in Protein Secondary Structure Prediction	136
6.2	Future Works	139
6.2.1	Improvements to Sequence representation and Feature Encoding	140
6.2.2	Ensemble DBN to Predict Secondary Structures	143
6.2.3	Automated Procedure to Detect Compact Model	143
6.2.4	Relation of Sequence to Structure Space	144
A	Twenty Amino Acids	146
B	Protein Information	148
C	Author's Publications	151

List of Figures

1.1	Ball and Stick Diagram of Peptide Bond and Torsion Angles	2
1.2	Illustration of Three Secondary Structures	4
1.3	Hierarchical Organization of Protein Structure	5
1.4	Alpha Helical Transmembrane Protein	6
1.5	Beta Barrel Transmembrane Proteins	7
1.6	Schematic of SS and MP prediction	11
2.1	Diagram of a Multiple Sequence Alignment	31
2.2	Diagram of a PSSM	32
3.1	Flow Diagram of Membrane Residue Prediction	57
3.2	Schematic of HiDEN	60
3.3	Architecture of the Deep Ensemble Network (DEN)	61
3.4	Topology of α -helical TM Protein	63
3.5	Sequence Fitting Block	64
4.1	Architecture of Deep Belief Network	83
5.1	CABS Lattice	101
5.2	CABS algorithm based Feature Encoding	103
5.3	Architecture of FCRN	105
5.4	Plot of Q_3 vs. number of training sequences	108
5.5	Secondary Structure Ratio in CB513 Sequences	109
5.6	Secondary Structure Composition of training sets	110
5.7	Q_3 accuracies per SCOP class in CB513	125
5.8	Case Study of Two Inhibitors	127
A.1	The 20 standard amino acids	147

List of Tables

2.1	Summary of main Transmembrane (TM) Topology Predictors in the Literature.	37
2.2	Summary of Main SS Predictors in the Literature	46
3.1	TOPCONS Datasets	55
3.2	HiDEN Settings	62
3.3	HiDEN Comparison with other Classifiers	69
3.4	Confusion matrix for Protein Group classification	70
3.5	HiDEN Topology Prediction compared with Reported Methods	71
3.6	GRAVY values of Transmembrane Helices	77
4.1	DBN Performance on PSSM derived features	87
4.2	Comparison of Deep Network based SS Predictors	88
4.3	Comparison of DBN Performance on Different Features	89
5.1	Secondary Structure Datasets	99
5.2	Summary of Protein Training Sets	110
5.3	Performance of FCRN Predictor in Comparison with other classifiers .	114
5.4	FCRN overall accuracy comparison with other protein secondary structure prediction methods	115
5.5	Results on CB513 (51,634 residues)	117
5.6	Results for G Switch Proteins (1400 residues)	119
5.7	Methods Comparison on G Switch Proteins	121
5.8	List of SCOP Structural Classes	124
5.9	Observed and Predicted SS in two Inhibitors by SSP ₅₅	126
5.10	Detected hydrogen bonds of Sheet and Coil residues	130
B.1	GSW25 sequences	149
B.2	SSP ₅₅ compact model proteins and their properties	150

Abbreviations

CABS C-Alpha-C-Beta-Side Group.

DBN Deep Belief Network.

DEN Deep Ensemble Network.

FCRN Fully Complex-valued Relaxation Network.

GRAVY Grand Averaged Hydropathy.

HiDEN Hierarchical Deep Ensemble Network.

HMM Hidden Markov Model.

MP Membrane Protein.

NN Neural Network.

PBL-McRBF Projection-based Meta-cognitive learning Radial Basis Function Network.

RBM Restricted Boltzmann Machine.

SS Secondary Structure.

SSP Secondary Structure Prediction.

SSP₅₅ Compact Model, a training set of 55 proteins.

SSP_{CV} Cross-validated Model, a training set of 308 proteins.

SVM Support Vector Machine.

TM Transmembrane.

TMH Transmembrane Helix.

ABSTRACT

Proteins are macromolecules that carry out important processes in the cells of living organisms, such as signalling, transport, cell replication and catalysis performed by enzymes. Loss of protein activity through problems caused by mutations or misfolding results in malfunction and is the cause of many diseases. In particular membrane proteins, of which the α -helical transmembrane (TM) proteins form the majority of folds, are important drug targets involved in critical roles such as cell signalling and metabolism, amongst other functions. A protein comprises of a chain of amino acid residues that fold into specific 3-*D* conformations to achieve their specified function. However, because it is difficult and resource intensive to obtain the structures of proteins experimentally, it is desirable to predict protein structures directly from sequence. Machine learning methods in the past few decades, have been increasingly used as a complement to experimental approaches to predict structures from protein sequences. The accurate prediction of protein secondary structures and in the case of membrane proteins, the correct detection of the residues' environment preference as well as the topological orientation with respect to the membrane are intermediary steps to aid the full 3-*D* structure determination.

Current protein Secondary Structure (SS) prediction methods such as neural network based approaches report cross-validated accuracies of about 81% for single predictors and up to 85% in the case of consensus based models that combine individual predictors. Nevertheless, the theoretical upper limit for SS prediction is estimated to be at 88-90% and remains a challenge to be reached. As even small errors at the secondary structure level can compound into large deviations in the final, tertiary structures, there is a strong motivation to develop better techniques to handle the SS prediction. Similarly for the prediction of membrane protein (MP) topology, while earlier proposed works regularly reported topology scores of 80% or higher, with increasing size and complexity of topological formations seen in current membrane protein datasets, the recent scores reported by individual predictors are typically around 60% to 70%, although consensus based methods have reported up to 80%.

A common feature in both SS and MP topology prediction problems is the representation of protein sequences as Position Specific Scoring Matrices (PSSM). While they possess the advantage of evolutionary information, it is difficult to accurately represent a sequence with PSSM due to challenges of the sliding window scheme and the overlapping signals found in the sequence space. Additionally in the case of SS prediction, methods are trained on large numbers of proteins for the cross-validation based models which may result in a loss of generalization. Here, it is of interest to obtain a more efficient training set that may generalize well to unseen datasets. Ways of extracting higher level features from the PSSM by using Deep Belief Networks are explored. Additionally, an energy based feature representation procedure, in which a protein sequence is encoded with probabilities derived from energy potentials using the previously developed C α -C β -Side Group protein model (CABS) algorithm is employed. Finally, the use of the Complex-plane in the SS prediction is also explored. Complex-valued neurons had shown to be computationally powerful in classification tasks and have been used here to predict protein secondary structures as well as to

develop a heuristics based procedure to develop a training model distinct from conventionally employed cross-validation approaches. The proposed works and main findings in this thesis are

- (i) the prediction of membrane protein topology and detection of signal peptide and globular sequences, using a Hierarchical Deep Ensemble Network (HiDEN) from PSSM based encoding
- (ii) Deep Belief Networks (DBN) applied to predict protein secondary structures from PSSM based encoding and CABS algorithm based feature encoding
- (iii) the prediction of protein secondary structures using a Fully Complex-valued Relaxation Network (FCRN) using cross-validated training models together with CABS algorithm based feature encoding
- (iv) the development of a heuristics based procedure to determine a small training set of proteins, known as the Compact Model, whose accuracies are similar to cross-validated models, yet remains capable of generalizing to new datasets
- (v) structural analysis of predicted SS to detect the role of hydrogen bonds in residue mis-classification rates. In particular, the effect of water-mediated vs. peptide-backbone hydrogen bonding is investigated, with the finding that residues with distinct hydrogen bonding patterns prove a challenge for many existing predictors, even if such residues are included in the training set.

The findings of the thesis can be extended in several ways. Firstly, for the use of PSSM based encoding, the proposed models may be improved by incorporating other sequence and structure-based properties of interest, such as the amino acid composition, accessible surface areas of residues and so forth. Secondly, in the case of MP structure prediction, the current three or four state residue classification system could be extended to cover more features of interest such as kinks and re-entrant helices, that remain un-explored. Lastly, the heuristics based procedure to obtain the compact model can be extended in a systematic way by applying automated sample selection strategies, such that the best learning model given any training set is automatically selected.

Chapter 1

Introduction

Proteins are a major group of biological macromolecules in cells of living organisms and are produced by translation of the genetic code contained in DNA. They are involved in cellular processes such as enzyme reactions, signalling pathways, transport of substances, replication and many other important functions in the cell. The structures of proteins are shaped by the functional roles they are required to play in cells. For example, membrane proteins comprise hydrophobic helices and signal peptides to anchor them to the membrane. Globular proteins contain predominantly hydrophilic residues on their exposed surface in order to form favourable interactions with a polar environment. Enzymes have precise spatial arrangements of residues in their reactive centres to allow catalysis. Hence, the structure of a protein can provide an insight to its biological function. However, the three dimensional structure of a protein is not easily obtained via experimental methods and many researchers have attempted to predict protein structures through computational means for decades.

1.1 Description of Protein Structure

A protein sequence comprises of a linear sequence of amino acids linked together by peptide bonds. There are twenty naturally occurring amino acids, with side groups (R) organized by their chemical polarity, given in Appendix A. Each position in a protein sequence may be occupied by any one of the twenty amino acids. Figure 1.1 illustrates the linear polypeptide chain, showing Carbon (C), Nitrogen (N) and

Oxygen (O) atoms and the side groups (R). However *in vivo*, the protein molecule folds into a specific conformation and rarely exists as an extended polypeptide chain. The conformation of a protein refers to the spatial arrangement of atoms within its structure. Theoretically, there are hundreds of possible conformations for a given protein, but only one or few states exist since they are energetically favourable. A low energy, stable and folded protein state is also known as its native structure. The protein conformation is further organized in a structural hierarchy, which is described below to frame the biological context of the structure prediction problem.

Figure 1.1: Ball and Stick Diagram of Peptide Bond and Torsion Angles. A protein chain is shown with carbon atoms in grey, carbonyl oxygen atoms (O) in red, amide Nitrogen atoms (N) in blue and amide hydrogen (H) in white. The alpha carbons are marked with (C_{α}). The side groups (R) comprise more atoms but here, are represented with a single large purple sphere. There are up to 20 different commonly occurring side groups in proteins (given in Appendix A). Planar bonds are indicated by atomic bonds that lie within rectangles. These bonds comprise the protein backbone that gives rise to secondary structures according to the torsion angles. The torsion angles ϕ and ψ indicate the degree of rotation between the planes of N- C_{α} and C_{α} -C atoms respectively. The ϕ, ψ values determine the type of secondary structures that the protein adopts. Numbers indicate idealized bond lengths in Angstroms ($\text{\AA}$). The figure is from [1] and was used in this thesis by permission of the publisher, W.H. Freeman and Company.

1.1.1 Primary Structure

The primary sequence of a protein refers to the sequence of amino acid residues comprising the protein chain. A single protein sequence may contain from about a hundred to thousands of amino acids. For a sequence of length M , each position in the sequence may be occupied by any one of the 20 types of amino acids described earlier. Each of the 20 types of amino acids has a full name, as well as standardized three and

one letter abbreviations that are commonly used. The full names and abbreviations of the twenty amino acids are available in Appendix A. Besides the 20 amino acids, others with modified side groups exist, but are not of interest in this thesis.

In order to exist as a polypeptide chain, the amino acids must be linked together covalently by amide bonds (known as peptide bonds), forming the peptide backbone, as illustrated in Figure 1.1. The sequence is typically ordered from the amino terminal (known as the N-terminal) to the Carboxyl terminal (also known as the C-terminal). Amino acids that are integrated with the peptide chain are termed amino acid residues. In this thesis, the terms *amino acid*, *amino acid residue* and *residue* are used interchangeably; the terms *protein backbone*, *peptide backbone*, as well as *polypeptide chain* are also used interchangeably with each other.

1.1.2 Secondary Structure

The secondary structure (SS) consists of several highly regular and well-ordered local structures that arise from the linear polypeptide chain folding into its conformation. These are the α -helix and β -sheet structures formed by regular hydrogen bonded geometry of the polypeptide backbone. Besides the regular secondary structure, several flexible structures such as coils or loops that do not participate in extended hydrogen bonding, are also considered as secondary structures. Other types of helices with differing hydrogen bonded geometries (e.g. π -helix, 3_{10} -helix) also exist in nature. For the purposes of this thesis, all secondary structures will be reduced into three classes for prediction, namely Helix (H), Sheet (E) and Coil (C) following the standard three-state grouping of secondary structure in the literature.

The α -helices form hydrogen bonds with residues that are nearby in the sequence, while β -sheet structures form hydrogen bonds with more distant neighbours in the sequence, as shown in Figure 1.2. As will be described later in this thesis, β -sheet structures pose a challenge to be predicted correctly by current methods and it has been proposed that one of the reasons is the inability of the sliding window schemes to detect interactions between distant neighbours in the sequence [2]. Results from

Figure 1.2: Illustration of Three Secondary Structures. The figure shows the structure of a protein in which Helix, Sheet and Coil structures are displayed in red, yellow and magenta respectively, while the rest of the protein is green. The amino acid sequence above the protein structure represents its primary sequence while the numbers indicate their residue position. The colors of the sequence correspond to those of the structure. For example, the portion of the sequence in red corresponds to residues that adopt the Helix structure, also in red. It can be observed that residues which are displayed yellow in the sequence are not contiguous, unlike the Helix or Coil structures. Sequentially non-contiguous residues can be spatially close and interact in order to form the Sheet structure. These are termed long-ranged interactions and are difficult to capture using current prediction methods. The figure was prepared using the PyMOL structure visualization software and the PDB structure file '1NDH'.

existing works show that the Sheet class (comprising mainly of β -sheets) is the most challenging class to predict [3–6].

1.1.3 Tertiary Structure

The tertiary structure of a protein refers to its full 3-dimensional conformation specified by atomic coordinates. It is obtained by experimental methods such as X-ray crystallography or Nuclear Magnetic Resonance (NMR). Atomic records are prepared according to format specifications and deposited into the Protein Data Bank (PDB) [7]. At the tertiary level, the full polypeptide conformation folded into a native structure is observed. Further, the organization of chains with respect to one another within this conformation is referred to as its quaternary structure. Figure 1.3 shows the overall organization of the primary, secondary, tertiary and quaternary structures in a single molecule.

Figure 1.3: Hierarchical Organization of Protein Structure. The figure shows the organization of a haemoglobin protein molecule into four levels of structural hierarchy. Haemoglobin is a globular protein that transports oxygen to cells. The primary structure comprises of a sequence of ordered amino acid residues, represented with three letter abbreviations. For example, Lys stands for Lysine. The secondary structure contains an α -helix, shown in blue. The tertiary structure refers to the full three-dimensional organization of the polypeptide chain specified by its atomic coordinates (translucent blue). The quaternary level represents the organization of protein chains relative to each other, when they assemble as subunits. The figure is from [1] and was used in this thesis by permission of the publisher, W.H. Freeman and Company.

1.1.4 Membrane Protein Topology

The protein organization described above is typically applied to globular proteins, which are present in polar (water-based) environments like the blood plasma or the cytosol (cell inside). Membrane proteins, however are attached to the cell membranes. The cell membrane consists of a phospholipid bilayer, containing polar phosphate heads on the membrane exterior and hydrophobic lipid molecules in the membrane interior. Membrane proteins are specialized with specific structures and are of great interest since they comprise approximately 50% of drug targets [8]. Their structure is discussed in detail below.

Membrane proteins may either fully span the lipid bilayer (integral proteins) or may be attached to its surface (peripheral proteins). The integral membrane proteins are also known as transmembrane proteins and may be divided into two main categories: α -helical transmembrane (TM) proteins and β -barrel transmembrane proteins (TMB).

Figure 1.4 shows the Bacteriorhodopsin Transmembrane Helix (TMH) protein containing 7 TM helices while Figure 1.5 shows several TMB proteins anchored within a lipid bilayer.

Figure 1.4: Alpha Helical Transmembrane Protein. The figure shows Bacteriorhodopsin, an α -helical transmembrane protein with 7 transmembrane helices (TMH) each shown in different colours. The phospholipid bilayer is indicated with the phosphate head groups shown as light blue circles while lipids are shown as light yellow tails attached to the phosphate heads. The Outside and Inside regions are labelled. In this protein the amino terminus (or N-terminal) is situated 'Outside' the cell while the Carboxyl terminus (or C-terminus) is located 'Inside' the cell. To obtain the correct topology the location of both termini, the number as well as location of TMH within the protein sequence must be predicted correctly. The figure is from [1] and was used in this thesis by permission of the publisher, W.H. Freeman and Company.

The membrane proteins perform specialized functions, such as acting as molecular pores to transport small molecules, or performing light sensitive reactions. To accomplish their specialized functions, the MP are oriented in specific directions with respect to the cell membrane. Figure 1.4 shows the carboxyl (C) terminal of the Bacteriorhodopsin protein within the cell (Inside) while the Amino (N) terminal is located externally to the cell membrane (Outside). In the case of α -helical MP, predicting the topology is defined as the correct prediction of the orientation of the protein structure with respect to the cell membrane. At the single residue level, the residue environment preferences with respect to the membrane are typically regarded as Inside (I), Outside (O) and Membrane (M). Some works consider other types of residues such as Loop and Globular [9]. From single residue preferences, the overall topology, such as the correct orientation of protein N and C termini as well as the number and location of TMH must be predicted accurately to be deemed as the correct topology.

Figure 1.5: Beta Barrel Transmembrane Proteins. The figure shows the structures of three β -barrel Transmembrane (TMB) proteins, from the *E. Coli* outer membrane in purple. FepA with 22 β -strands is a protein involved in iron transport. OmpLA with $(12 \times 2)\beta$ -strands is an enzyme to process lipids. Maltoporin with $(16 \times 3)\beta$ -strands transports maltose for *E. Coli*. The phospholipid bilayer is shown with phosphate head groups in light blue and lipid tails in light yellow. The β -barrel structures are anchored within the lipid bilayer and serve as molecular pores to transport molecules into and out of the cell. The figure is from [1] and was used in this thesis by permission of the publisher, W.H. Freeman and Company.

While in recent years the number of experimentally determined membrane proteins has increased, they still lag far behind that of globular proteins. As of December 2016, the number of TMH proteins is approximately 2600 while that of TMB proteins is about 400 in the PDBTM database [10], or more than 6 times. Since there are more data available for the α -TM helical proteins, they are the focus of predictions in this thesis.

1.1.5 Protein Data Bank (PDB)

The Protein Data Bank (PDB) contains full protein structures in the form of atomic coordinates that have been resolved by experimental methods [7]. Each entry in the database is specified by a 4 letter identification code known as the PDB ID. Besides proteins, the PDB contains other biological macromolecules such as DNA, RNA and complexes of interacting proteins. As of March 2017, the PDB contains approximately 127,000 structures, of which proteins comprise slightly more than 118,000 structures (figures are rounded to the nearest thousand).

The PDB is the standard repository for protein structures and consists of the primary source of all datasets used in the thesis. Membrane protein structures are present in secondary databases such as the PDB TM [10] database that is derived by annotating MP in the PDB.

1.2 Motivation for Protein Structure Prediction from Primary Sequence

There are two main reasons to predict the protein structures from their primary sequence. Firstly, there are comparatively many more sequences than structures, known as the "sequence to structure" gap, which motivates the proposed works in this thesis. Secondly, experimental evidence has shown that the protein sequence alone suffices to form the fully folded and functional native conformation of the protein.

The Sequence to Structure Gap:

A protein's primary sequence may be obtained relatively easily by experimental techniques such as mass spectrometry. In comparison, the full atomic coordinates are much harder to obtain. Experimental techniques to resolve the tertiary structure include X-ray crystallography and Nuclear Magnetic Resonance (NMR), with challenges such as protein purification and expression in large amounts for the former and needing advanced techniques to resolve NMR restraints for the latter being significant bottlenecks.

In recent years, the number of deposited structures in the PDB has seen rapid growth, with large protein complexes as well as many membrane proteins being resolved. This growth motivates some claims that the sequence to structure gap is narrowing [11]. Yet, a check of the sequence and structure databases shows that a significant gap remains between experimentally determined structures and sequences. For example, *RefSeq* database provided by the National Center for Biotechnology Information (NCBI) contains 13 million non-redundant protein sequences [12] while in comparison the PDB contains 118,000 protein structures. Accordingly, 99.09% of sequences lack structures and must rely on templates, needing the assistance of computational models to build them and to assign correct templates based on sequence similarity matches.

Sequence Sufficiency for Structure formation:

In early years it has been shown that the protein sequence alone contains sufficient information to produce the correctly folded, fully functional structure under favourable

conditions. This was demonstrated in a series of denaturation experiments conducted by Anfinsen [13]. Anfinsen demonstrated with a protein ribonuclease that the tertiary structure of the fully folded conformation could be assembled *in vitro* from an unfolded sequence.

The prediction of protein structures directly from sequence is termed *ab initio* prediction and is a challenging problem. Owing to the large degrees of freedom afforded by the ϕ, ψ torsion angles of the peptide backbone and rotational freedom of residue side groups, there is a very large conformational space available to any given sequence. The mechanisms by which the lowest energy structures are quickly obtained by a sequence are still unresolved problems in protein folding. Consequently the estimation of the functional relationship between a sequence and its structure is difficult to elucidate by atomic, physics based models. As it is difficult to directly determine the sequence to structure relationship using physics based models, existing works have posed the problems as multi category classification problems where the underlying relation is estimated by a classifier such as a Neural Network (NN) or a Support Vector Machine (SVM).

Applications of Predicted Secondary Structure and Membrane Protein Topology:

Before the full tertiary structure of a protein can be predicted, it is helpful to pose the problem as a one-dimensional prediction problem (of secondary structure and membrane topology prediction), where each residue is predicted to belong in one of several pre-defined categories. By posing the problem as a lower-resolution, categorical classification problem, machine learning techniques have been used to estimate the functional relationship between an input sequence and its secondary structure (or topology). The predicted secondary structures and membrane protein topologies are a powerful complement to experimental approaches; for example, predicted secondary structures are correlated with other applications like the prediction of residue depth [14] and prediction of protein-ligand interaction sites [15,16]. Reliably predicted secondary structure can improve the quality of multiple sequence alignments [17]. Further, membrane protein and signal peptide detection algorithms may be applied to

new genomes to characterize them [18], or to place correct transmembrane helices for starting models in molecular dynamics simulations of protein folding [19].

1.2.1 A Schematic Representation of Secondary Structure and Membrane Topology Prediction

The works in this thesis focus on two main problem areas, the prediction of secondary structures and Membrane Protein (MP) topologies from sequence. The two problem areas are related, requiring the prediction of residue structures or preferences of environments. The classifier (such as a neural network) should discriminate between

- Secondary Structure Prediction
 - Detect if residue structures belong in
 - **H**elix (**H**), **S**heet (**E**) or **C**oil (**C**)
- Membrane Protein Topology Prediction
 - Detect residue environment preferences with respect to the membrane
 - within **M**embrane (**M**), **O**utside membrane (**O**), **I**nside membrane (**I**)
 - and if applicable **S**ignal peptides (**S**)

Further, MP topology predictions requires an additional step of refining the residue preferences after the first level prediction in order to obtain the correct location and number of TMH.

A schematic representation showing the secondary structure and membrane protein topology prediction as a functional relationship to be estimated by a standard back-propagation network is shown in Figure 1.6. The figure is adapted from [20].

The figure shows an estimation of the underlying functional relationship between an input sequence of amino acids and the structures it adopts. The structures are represented one-dimensionally. An input sequence is read in a sliding window scheme in order to represent the interactions of its neighbours in the sequence. In the figure, the size of the window, $w = 7$. The sliding window scheme is used to convert the input sequence into a feature matrix, which will be described in detail in section 3.3. Each

Figure 1.6: Schematic of SS and MP prediction. The schematic of an input protein sequence and learning of its secondary structures (or membrane residue preferences) by a standard backpropagation neural network is shown. The feature encoding process is done using a sliding window scheme, with the neighbours' information being read in.

An input sequence is read in a sliding window, w (rectangle in heavy black lines), with the central residue, x^t outlined as a square in heavy grey lines within the window. After being read in as a matrix of features, x^t is passed as input to the network which learns from class labels y^t . The network weights are updated in the backpropagation step. The number of input neurons is usually a multiple of the number of features. In the case of secondary structure prediction, the classes represent the secondary structures Helix (H), Sheet (E) and Coil (C). In the case of membrane proteins, two levels of predictions occur. In the first level, membrane protein residue environment preferences such as within Membrane (M), Inside membrane (I) or Outside membrane (O) are first learnt. In the second level, the predicted residue preferences are refined with other models like smoothing functions, before the final predicted topology is obtained. The number of output neurons is typically three, one each for (H, E and C) or (M, I and O). If Signal peptide (S) predictions are also included in the model, the number of output neurons is 4 (M, I, O and S). The figure is adapted from [20, p. 335].

residue, t in a sequence is represented as a feature vector $\mathbf{x}^t$. The network weights are updated during backpropagation, by learning from the labelled secondary structures (or membrane residue preferences) $\mathbf{y}^t$.

During testing, the saved network parameters from the training are used to estimate the class labels of an un-learned residue (sample). In the case of membrane protein topology prediction, a refinement is applied to the predicted sequence, such as Hidden Markov Models to detect the correct topology [9], or smoothing steps to ensure contiguous location of transmembrane helix residues in the sequence.

The above process to predict protein structures (or membrane residue preferences) has some challenges related to the representation of sequences as features, effect of experimental noise derived from assigning labels to sequences, as well as the appropriate sample selection to obtain a suitable model.

1.2.2 Challenges in protein SS prediction

Despite the large numbers of secondary structure (SS) predictors proposed in the literature, several challenges remain, motivating for new techniques to address them. The overall accuracy (η_o), defined as the number of correctly classified samples over the total number of samples, is also known as the Q_3 score in the secondary structure prediction literature. The Q_3 score is used as a measure of the classifier’s accuracy and is defined as

$$Q_3 \equiv \eta_o = \sum_{j=1}^3 \frac{q_{jj}}{N} \quad (1.1)$$

where q_{jj} are the number of correctly classified samples in class j and N represents the total number of samples. Equation 1.1 is also equivalent to equation 5.10.

As discussed in chapter 2, the highest Q_3 accuracy achieved by current methods range from at least 80% [5] to 86% [3, 21]. However, the estimated theoretical upper limit for SS prediction is expected to be 88% to 90% [2, 22], suggesting further room for improvement.

Experimental Variation and Labelling Noise:

The theoretical upper limit for protein secondary structure prediction of 90% was estimated from an observation that 5% to 15% of secondary structure points can differ among structures depending on the experimental methods used, even for the same protein [2]. Moreover, different SS assignment programs can lead to a difference in SS labels of residues. The secondary structure assignment process relies on the correct detection of hydrogen bonds of specific atomic types, including having the correct distance and criteria of geometric planarity to be satisfied. Depending on the choice of distance cut-offs and quality of the PDB structures, there may exist differences in assigned SS [23]. These noises resulting from experimental variation can lead to a small loss in performance of classifiers, even though machine learning classifiers are robust in the face of such small variations.

While existing homology based SS predictors such as [24–26], reported Q_3 accuracies close to 82%, they could fail for sequences that do not have homologous hits, or if poor fitting templates to sequences are given high scores resulting in wrong predictions for the SS. For methods that obtained close to 86% accuracy, they are often very computationally intensive with a complex architecture needing a long time to classify [3], or they obtain the consensus of several methods to reach a high score. More powerful neural networks or classifiers may be needed to address these concerns.

Appropriate Sample Selection for Model Development:

Conventionally, cross-validation approaches are used to develop predictive models for protein secondary structures. Unlike the case of MP topology prediction, there are many structures available for globular proteins with large datasets containing thousands of protein chains [5, 6, 27].

Typically, the model is trained on hundreds of proteins while testing is conducted on a much smaller partition (or dataset). Some potential issues with this approach include (i) over-training of the classifiers, leading to a loss of generalization and (ii) the assumption that training samples uniformly represent the input space.

However, proteins do not sample the fold space uniformly, resulting in some com-

only represented folds dominating the training model, even though researchers may carefully prune datasets to only include independent protein sequences. Moreover, the secondary structure content (ratio of Helix, Sheet and Coil residues) can affect the protein fold. Hence an appropriate choice of training samples may be important. One of the contributions in this thesis is to explore a hypothesis of detecting good training sets to obtain a suitable training model that retains its generalization capabilities.

1.2.3 Challenges in Membrane Protein Topology prediction

Membrane protein structures are even harder to obtain by experimental techniques than those of non-membrane (or globular) proteins, due to the lipid membrane that makes it difficult to express and crystallize MP in their correct conformation. Even though in recent years more attention has been paid to the discovery of new MP structures, owing to the difficulty of crystallizing MP, currently only 2% of the PDB proteins are membrane proteins [28]. However, MP are an important class of proteins that comprise 50% of all drug targets [8] and have important functional roles to play, highlighting the need for more powerful computational approaches to complement the existing experimental methods.

Experimental Variation and Labelling Noise:

Similar to the scenario for SS prediction, there exists some variation arising from the labelling of residue preferences with respect to the membrane, depending on the source. Firstly, it is difficult to experimentally determine the location of the membrane, leading to errors in determining helix boundary and orientation [29]. Secondly, depending on whether topological annotations are done from e.g. glycosylation studies reported on the protein, or by geometric algorithms (for e.g. TMDET [30]), different labels may be obtained, needing a consensus to be reached for the labels.

While various machine learning and statistics based methods have been used to predict MP topology from sequence, the accuracy of recent methods is generally lower com-

pared to older predictors. For example, earlier works like TOPPRED [31] reported greater than 90% topology scores on very small datasets (less than 30 MP), while later works reported scores like 89% [29] and 94% [9]. However in 2015, the highest reported topology score was at 80% reported by a consensus based predictor [32]. In 2016, the accuracy reported by a random forest method was 71%, averaged over several measures [33]. As more MP structures are being deposited into the PDB recently, the size of datasets and the complexity of topologies has increased, motivating for new algorithms to predict these with higher scores.

Sequence Variation and Feature Representation:

One problem plaguing both SS and MP topology predictors is the generic quality of the Position Specific Scoring Matrix (PSSM), a commonly employed feature representation scheme in both problems. The PSSM represents evolutionary information of residues, where matrix elements correspond to likelihoods of particular amino acid types occurring at given positions in the sequence. The PSSM and the techniques for obtaining it will be described in detail in chapter 2. The advantage of the PSSM is that it represents evolutionary information, providing more information rather than using a single sequence.

Effect of Sequence Variation on PSSM:

However, compared to protein structures amino acid sequences are more prone to variation resulting from evolutionary pressures, resulting in sequence space being much larger than structure space [34]. Here, sequence space refers to the set of all known sequences while structure space is defined as a set of discrete protein folds corresponding to precise spatial arrangements of secondary structures like α -helices and β -sheets [35]. Since sequence space is larger than structural space, two protein sequences may share identical secondary and tertiary structures, despite having different sequences.

Further, there exists physical pressure on sequences to adapt to their environments, such as need for flexibility of a newly translated protein to move through Golgi apparatus in the cell, introducing artefacts into the sequence composition that are unrelated to the specification of the secondary or tertiary conformations. In view

of the problem to predict one dimensional structure-based characteristics, all such information, together with sequence information coding for actual structural conformations, will be mapped to secondary structure (SS) space and membrane protein environment preference space, respectively.

For the purpose of this thesis a definition of space as relates to the prediction of secondary structures and membrane protein residue preferences from sequence is given. In the case of SS, the structure space is defined as a set of 3 consisting of Helix (H), Sheet (E) or Coil (C) structures or $SS \in \{H, E, C\}$; in the case of the MP topology prediction, it is defined as a set of 4 residue environment preferences consisting of Membrane (M), Inside (I), Outside (O) and Signal (S) peptides, or $MP \in \{M, I, O, S\}$. For a sequence of N residues, with 20 possible substitution values at each position this leads to multiple values to encode a single sequence ($N \times 20$) as input, that need to be predicted as 3 or 4 output targets (or classes) in SS or MP , respectively. Moreover, the distribution of the encoding is not clearly separable with respect to SS or MP structure space and is discussed further in the next paragraph.

Challenges of Sliding Window Based Scheme:

The sliding window and PSSM are benchmark techniques used in the literature to represent protein sequences. Previous methods that used single sequence information were not as successful as later methods that incorporated multiple sequence information in the form of PSSM as well as the sliding window [36–38]. The sliding window scheme takes properties of neighbouring residues into account and indirectly models the interaction between residues in a sequence that can better predict the final structures. However, some difficulties are associated with the sliding window scheme as discussed below.

It has been a challenge for current topology and secondary structure predictors to capture long range residue interactions using a sliding window scheme to capture the effect of long range residues on local sequence properties. Typically, the PSSM

is read in a window that slides across the protein sequence, in order to read in the information represented by adjacent, neighbouring residues. It has been a challenge to detect the optimal window size that can adequately capture long range information. In secondary structure prediction, structures like helices, that are typically separated by only three or four neighbours in the sequence are often predicted more easily than β -strand containing structures [2, 6]. β -strand structures are influenced by residues far away in the sequence, yet close in the spatial structure, which is most likely a factor for many prediction methods to have a lower accuracy for them compared to helix structures [5].

In membrane protein topology prediction, multiple window sizes have been used by works to model the distinct regions comprising transmembrane helices, inside and outside residues. Approaches include using multiple sliding window sizes to model the centre and ends of TM-helical regions, [39], a range of window sizes from 27-35 to model membrane residue preferences, [29], as well as shorter window sizes like 15 to model inside outside preferences and 31 to model transmembrane helical regions [9]. Although many works had been proposed, it remains difficult to adequately capture the long range residue information using a sliding window approach to represent the PSSM information. For a sliding window of size w , where w is odd, information of the first $\frac{w-1}{2}$ and last $\frac{w-1}{2}$ is lost since they lack neighbours. Next, the elements of the PSSM typically constitute discrete whole numbers in the range $[-10, +10]$, while the structure space is represented by numeric class labels 1, 2, 3 or 4 corresponding to various structures in the respective problems. Hence, many repeated values in the sequence space are being mapped into the same point in structure space which introduces a lack of precision into the PSSM based representation.

1.3 Objectives

Several approaches are proposed in this thesis to predict membrane protein topologies as well as secondary structures from their sequences. The following are the main objectives of the works.

1.3.1 Deep Belief Networks to extract higher level features

As discussed earlier, there are several drawbacks to the PSSM based representation of sequences as features, resulting from an overlap of values in the sequence space for many sequences. In addition, the sliding window scheme may potentially fail to capture underlying residue correlations from distantly related neighbours.

Deep learning models have the capabilities to extract high level features from input features, using the contrastive divergence algorithm [40]. In the unsupervised learning stage, the model will be able to capture underlying correlations between features and class labels, which are able to capture the underlying residue to residue interactions. It is proposed to apply Deep Belief Networks to predict protein secondary structures as well as membrane protein topologies.

1.3.2 Feature Extraction using the CABS algorithm

Due to the challenges faced by the PSSM based feature encoding process, an energy potentials based sequence representation procedure, using the CABS algorithm [41] is employed in this thesis. The C-Alpha-C-Beta-Side Group (CABS) algorithm calculates the energy of aligning a target sequence to template structures. Protein structures are represented using a reduced lattice model and the reference energy of aligning a target sequence to a template is then computed using a force field provided by the CABS model.

From the reference energies, high scoring template structures are selected and the probability of a residue in the target sequence adopting H, E or C structures is read in a sliding window. As the energetic contributions of residues are taken into account, using the CABS algorithm's energy potentials based encoding is more capable in capturing the underlying sequence to structure relationship compared to the PSSM.

Previously, datasets had been generated using the CABS algorithm [42]. Here it is proposed to use them to predict secondary structures as well as to develop an efficient training set.

1.3.3 Secondary Structure Prediction using a Complex-valued Neural Network

Current works in predicting secondary structures have employed Real-valued neural networks. However, Complex-valued neurons have been shown to have more computational power and better decision making abilities compared to Real-valued neurons [43]. This results from the orthogonal decision boundaries of the Complex plane allowing additional dimensions for separation and hence, more discriminatory power [44]. In this thesis it is proposed to apply a neural network classifier employing the Complex plane to predict the secondary structures of residues from sequence. The sequences have been encoded using the energy potentials based CABS algorithm described earlier.

1.3.4 Heuristics based approach to detect a small training set

As highlighted earlier, existing methods have mostly employed cross-validation techniques to predict protein secondary structures. Often a large ratio of training proteins compared to test proteins within a single partition is used. While Q_3 accuracies as high as 85% have been reported, cross-validation approaches may lead to loss of generalization ability in the classifier resulting from over-training.

To avoid the challenges, in this thesis, a simple, heuristics based search procedure is proposed to locate a small number of training proteins that maintains a high accuracy even when tested against a much larger number of unknown proteins. This small number of training proteins is henceforth referred to as the 'compact model' in this thesis. However, locating a minimal set of proteins that are able to maximize the predictive accuracy for any given dataset is challenging and remains a potential future work.

1.3.5 Biological Analysis of Predicted Structures

The final objective of the thesis is to analyse the predicted secondary structures and membrane proteins from a biological perspective. Based on the heuristics based ap-

proach proposed, it is possible to investigate the effect of protein folds or structural groups on the Q_3 accuracy by considering their Structural Classifications of Proteins (SCOP) [45] class. As the number of training proteins is small, structural groups with low Q_3 , or proteins whose Q_3 does not increase despite being included in training, can be identified. They present problems for current secondary structure approaches and are the limits of why the Q_3 is below theoretical upper limit proposed at 88% to 90% [2, 22].

1.4 Contributions

This section describes the main works of this thesis. Two types of one-dimensional structures were predicted from protein sequences using a Deep Ensemble Network (DEN) or Deep Belief Network (DBN). These were membrane topologies and protein secondary structures. DEN and DBN were used to predict membrane protein topologies or protein secondary structures respectively using the standard PSSM based encoding commonly used in the literature. Due to the challenges associated with extracting residue interactions from PSSM based representation an alternative energy based feature encoding known as the CABS algorithm was employed. Since Complex-valued neurons had been shown to be computationally powerful, a Fully Complex-valued Relaxation Network (FCRN) was applied to predict protein secondary structures using the alternative, energy-based feature encoding. Next, a heuristics based procedure using the FCRN was proposed in order to detect a small but efficient training model for the protein secondary structure prediction. Finally, the thesis also analyses the predicted transmembrane helices as well as protein secondary structures from a structure based perspective in order to draw biologically meaningful conclusions that highlight potential areas of improvement.

1.4.1 Predicting Protein Structures with Deep Ensemble Networks

Due to their capabilities in extracting higher level features from low level input such as the PSSM, Deep belief networks comprising of multiple Restricted Boltzmann machines were used to predict membrane protein structures and protein secondary structures. The sequences had first been represented using the PSSM feature encoding generated from FASTA files.

Predicting membrane protein topology:

A Hierarchical Deep Ensemble Network (HiDEN) was proposed to detect membrane protein topologies and to distinguish between membrane and globular proteins from several TOPCONS datasets [32]. The network was also applied to predict signal peptides. In the first level a Global Deep Ensemble Network was first applied to detect protein groups from the dataset, with an accuracy of 95%, while Local Deep Ensemble Networks in the second level were applied to predict membrane protein residue preferences of belonging to 4 classes Membrane (M), Inside (I), Outside (O) and Signal Peptide (S). A sequence fitting block was applied to the predictions to obtain a smoothed output. The overall topology score (including testing and training) obtained was 95% for a dataset of Transmembrane proteins while the topology score for ten-fold averaged cross-validation was 63%.

Predicting Protein secondary structures:

A deep belief network containing a single hidden layer was applied to predict protein secondary structures. The DBN was trained on 80% of proteins from the CB513 [38] dataset and tested on the remaining 20%. Under the PSSM based feature encoding the model obtained a Q_3 accuracy of 79.37% for training proteins. However the testing accuracy was considerably lower at 72.56% compared to at accuracies of at least 80% or more reported by recent works [27, 46]. One reason for the higher results reported by other works could be the incorporation of other sequence based properties such as the relative solvent accessibility, side chain volumes, hydrophobicity and amino acid composition among other properties that complement the evolutionary information provided in the PSSM. Additionally the secondary structure predictors in the litera-

ture have also used multiple neural networks such as a second level structure to structure prediction network [36, 37] or combination of several networks predicting other properties like backbone torsion angles [27] or relative solvent accessibility [4]. These related properties are correlated with secondary structures and help to strengthen secondary structure predictions. The incorporation of more sequence based chemical or angle based properties of interest as well as the implementation of ensemble based deep learning models is a possible future work.

To solve the problem from a different direction here a different feature encoding based on energy potentials derived from the CABS algorithm has been explored. The CABS algorithm based energy potentials had previously been shown to obtain a Q_3 accuracy of 81% used together with an Extreme Learning Machine [42]. The DBN was tested with CABS based feature encoding and obtained a five-fold cross-validation accuracy of 82.6%. This motivates the use of CABS based energy potentials in subsequent works in this thesis.

1.4.2 Predicting Protein Secondary Structures with a Fully Complex-valued Relaxation Network

The Complex plane possesses an extra decision boundary and Complex-valued neurons had been shown to be more powerful than their real-valued counterparts [43]. This motivated the use of the Fully Complex-valued Relaxation Network (FCRN) [47] to predict protein secondary structures. Another motivating factor was that the FCRN employs an analytic method to obtain the network weights and allowed for rapid simulations needed in the development of the compact model later. Here the FCRN is employed together with the energy potentials based feature encoding generated using the CABS algorithm for the CB513 and G Switch Proteins datasets [42].

Two types of five-fold studies had been performed with partitions created based on

- randomly selected residues with an average Q_3 of 82.1%
- randomly selected proteins with an average Q_3 of 82.1%

to establish the performance of the FCRN on the CB513 dataset. The results were

higher or comparable to most other results reported in the literature. These findings motivate the development of the compact model. Next the heuristics based procedure to obtain the compact model is described. To avoid the potential challenges of over training during learning on large numbers of training proteins associated with cross-validation approaches a process to search for a small set of training proteins (termed the compact model) was developed. The compact model consisting of 55 training proteins was selected to maintain a good generalization capability of the classifier. It was developed and tested on the CB513 dataset as well as validated on a blind test set of G Switch proteins (GSW25). The compact model was compared in several ways, first against the best performing cross-validation partition, including a pairwise t -test to establish the statistical significance of the results and also tested against other methods reported in the literature. Finally, an analysis with respect to SCOP groups of the predicted results was also made.

1.4.3 Biological analysis of predicted structures

Predicted Secondary Structures:

The Q_3 obtained by the compact model was further investigated by considering the Q_3 obtained according to SCOP groups. For proteins in the worst performing SCOP groups, even including them in the training set did not help the classifier to improve the accuracy.

A detailed investigation into the role of hydrogen bonds in influencing the rate of Sheet to Coil and Coil to Sheet mis-classifications was subsequently conducted on the CB513 dataset.

- with the findings that the presence of amide (or peptide) hydrogen bonds in Coil residues can cause them to be predicted as Sheets.
- two case studies of the cAMP dependant kinase inhibitor and trypsin inhibitor are investigated in their respective protein complexes.

Predicted Transmembrane Helices:

HiDEN was able to predict the locations of 99% of the Transmembrane Helices (TMH)

in the TOPCONS datasets, correctly, compared to the original method (97%). Of these predicted TMH, the capability of HiDEN to predict helices based on their hydrophobicity was investigated using the Grand Averaged Hydropathy (GRAVY) Index [48]. By comparing with TOPCONS predicted TMH, it was found that HiDEN was able to detect mildly hydrophobic helices at a better rate than TOPCONS. The extraction of higher level features from the input PSSM has likely aided in capturing underlying residue preferences of being located in the TMH.

1.5 Organization

The thesis contains a total of six chapters, inclusive of this introductory chapter. In summary, the main goal of the thesis is to predict protein structures at the one-dimensional level from primary sequence that serves as input. Specifically, protein secondary structures and membrane protein residue preferences as well as signal peptides are predicted from their respective sequences. Further, the topology or orientation of membrane proteins with respect to the cell membrane is also predicted. The remainder of the thesis chapters are organized as follows:

Chapter 2 presents a literature review of the existing computational works to predict membrane protein topology and secondary structures. The main ideas in each problem area that represent significant contributions, the types of feature encoding employed by the algorithms, as well as the current challenges faced by predictors in each of the respective areas are described. Tables summarizing the existing approaches, broadly grouped by their algorithms (e.g. NN, SVM) are presented.

Chapter 3 describes the application of Deep Ensemble Networks to predict membrane protein topology. The encoding of membrane protein sequences as features generated from PSSM is described in detail. The problem of predicting membrane protein residue preferences is formulated in terms of an input vector and output target to be obtained. The architecture and process of employing HiDEN to predict the

membrane protein topology are presented. Finally, the predictions of HiDEN are analysed with respect to the hydrophobicity of TM Helices as determined by GRAVY [48].

Chapter 4 describes the application of Deep Belief Networks (DBN) to predict protein secondary structures using the PSSM based encoding. The problem of predicting protein secondary structures is formulated as an input vector to the visible layer of the DBN and targets to be obtained by its output layer. The architecture of the DBN and the contrastive divergence learning algorithm are described. A discussion of cross-validation experiments employing the DBN with an alternative energy based feature encoding process known as the CABS algorithm, is presented.

Chapter 5 describes the application of the FCRN to predict protein secondary structures. The architecture and learning algorithm of the FCRN are described followed by the process to encode the protein sequences using energy potentials derived from the CABS algorithm. Next, several studies based on cross-validation experiments are presented and the performance of the FCRN is discussed.

The heuristics based procedure to obtain the compact model of training proteins (that is able to retain a good generalization capability of the classifier) is described and the results of predicting secondary structures using the compact model are compared with cross-validation approaches. Later, the model is validated on a blind test dataset and compared with existing approaches in the literature.

Finally, the predicted protein structures are evaluated from a structural perspective, by comparing the Q_3 of training and testing proteins in the CB513 dataset according to their SCOP groups. The findings identify challenging protein groups for prediction. Also, the influence of hydrogen bonds on the rate of residue mis-prediction rates (in particular for Coil $\leftrightarrow$ Sheet mis-predictions) and two detailed case studies highlighting these issues are discussed. From these findings, the limitations of current works and future areas of improvements are discussed.

Chapter 6 concludes by summarizing the main findings of the proposed ideas and experiments carried out in this thesis. The limitations of the proposed works are presented and areas of improvements are discussed. Finally, possible future works arising from this thesis's material, are also proposed.

Next, the literature review is discussed in chapter 2.

Chapter 2

Literature Review

The first step towards experimental determination and computational prediction of protein structures is often the prediction of one-dimensional structural characteristics from their sequence, such as the topology or transmembrane helix location in the case of Membrane Proteins (MP), or Secondary Structures (SS) in the case of non-membrane, or globular proteins. Both SS prediction and MP topology prediction shared a similar development in the form of feature encoding, commonly known as sequence profiles. Sequence profiles derived from multiple sequence alignments (MSA) [49] or Position Specific Scoring Matrices (PSSM) [50] were employed with probabilistic models such as Hidden Markov Models (HMM) [51] or by using machine learning algorithms such as Neural Networks (NNs) [36, 52, 53] and Support Vector Machines (SVM) [29]. These developments resulted in drastically improved Q_3 accuracy in the case of SS prediction and topology scores in the case of MP topology prediction, respectively. This chapter presents a review of both the areas, with the MP topology prediction being presented first. Finally, a brief introduction of deep learning networks is given and several recent deep learning methods in MP and SS prediction are discussed.

2.1 Topology Predictors for α -helical Membrane Proteins

Previous works for Transmembrane (TM) prediction may fall into three groups: (i) transmembrane helix detection methods e.g. methods to detect hydrophobic sequence segments, (ii) signal peptide prediction methods to locate signal peptides from sequences, based on e.g. amino acid composition, cleavage site prediction and (iii) TM topology prediction methods, based on models to classify residue preferences with respect to membrane orientation. The TM topology prediction methods combine prediction of transmembrane helices with inside or outside orientation relative to the membrane and are the focus of this review.

2.1.1 Resources for Membrane Proteins

Datasets of MP labelled with residue preferences with respect to the membrane are required for the TM topology prediction. Several sources of annotation are used. As it is difficult to obtain protein crystals containing the membrane by experimental methods, several authors proposed databases which provide topological annotations for TM datasets. For example, the TMDET algorithm [30] estimates the membrane planes from Protein Data Bank (PDB) structures' atomic coordinates. Using the rotational symmetry of the molecule, the likely normals to the possible membrane planes are first calculated. Next, the molecule is sliced into different sections and an objective function that takes into account the membrane exposed surface area is used to select the most likely membrane locations. The PDBTM [10] database contains annotated structures based on TMDET [54]. The Orientation of Membrane Proteins (OPM) also contains annotated MP structures based on theoretical calculations of the membrane bilayer position [55]. Besides structure based calculations, SWISS-PROT annotations of membrane protein topology are also used [56].

Most membrane protein topology prediction methods may be grouped into two branches: one branch is specialised for prediction of α -helical Transmembrane (TM) proteins and the other for β -barrel transmembrane proteins (TMB). One reason for specializa-

tion is that many TM prediction methods are based on techniques to detect hydrophobic transmembrane helices comprising an average of 15 to 35 residues in length and are not applicable to β -barrel proteins owing to the latter's different structures making residue contacts to long distant neighbours. Additionally, the number of available alpha TM protein structures is more than 6 times compared to TMB; as of December 2016, 2600 TM structures are available compared to less than 400 for TMB according to statistics released by the PDBTM database. Due to the higher number of available structures, the works proposed in the thesis have focused on TM helical proteins.

The earliest works towards the prediction of α -helical TM proteins were based on chemical and physical principles, such as hydrophobicity scales developed by measuring partition energies of individual amino acid residues inserted into apolar environments [48, 57]. These may be considered first principle based works that directly modelled chemical principles and applied them to TM helix determination. By measuring partition energies over sliding windows applied on TM sequences, later works were able to detect TM helical segments and compare the effectiveness of different scales in TM helix detection [57]. In a further development in 1992, von Heijne and colleagues proposed a work, **TOPPRED**, that employed hydrophobicity scales together with the positive inside rule to successfully predict topologies of bacterial membrane proteins [31].

The next wave of developments were made possible with an increasing number of structures, where statistics-based approaches were proposed by several researchers to handle TM helix detection as well as topology prediction.

2.1.2 Early data driven approaches using single sequences

MEMSAT [58] modelled amino acid residue preferences using statistical log-likelihood ratios. Residue distributions in five TM regions were modelled - Loop Inside, Loop Outside, Helix Inside, Helix Outside and Helix Middle. From the log-likelihood ratios, expectation minimization was applied to detect the best topological model. **TMHMM** proposed a 7-state hidden Markov model (HMM), to model various TM regions such as helix caps, helix inside, outside preferences. Maximum Likelihood

estimation was used to derive the model parameters, with a contribution of dynamic detection of helix length [59]. Later, the TMHMM model was applied to the detection of TM proteins in genomes [60] and implemented with signal peptide predictions from SignalP [61]. In subsequent methods, the use of MSA and PSI-BLAST based sequence profiles proved a significant contribution.

2.1.3 Multiple Sequence Alignment (MSA) and PSI-BLAST

Here, the **M**ultiple **S**equences **A**lignment (MSA) and **P**osition **S**pecific **I**terated-**B**last or **PSI**-BLAST, two commonly employed feature encoding techniques (or sequence profiles), are described [50]. MSA is a technique to align three or more sequences together to identify regions of similarity between the sequences. The alignment may involve the whole sequence, known as global alignment or on localized segments of the sequence known as local alignments. The alignment programs typically consist of a dynamic programming algorithm to locate matches to a target sequence based on k -tuple word searches, given a database of sequences and a scoring matrix that assigns points based on the goodness of the match. Examples of MSA algorithms include the CLUSTAL suite of software [49] and MUSCLE [62]. Figure 2.1 shows an example of an MSA on a sequence. Since there are 20 possible amino acids that could occur at position of the sequence, each residue of the sequence is encoded by a feature vector of dimension (1×20) . A single element of the vector holds the frequency of observed amino acids of that type corresponding to one column of the alignment.

In PSI-BLAST, distantly related sequences are detected using a local alignment algorithm [50]. Similar to the MSA, a profile matrix known as the Position Specific Scoring Matrix (PSSM) is generated. However compared to MSA, PSI-BLAST detects more distantly related sequences, ensuring the incorporation of more structurally similar sequences into the profile, even if their sequences are distant. This is important, as sequences evolve at a faster rate than structures; for proteins lacking homologues, PSI-BLAST may be better at detecting similarities of more distantly related sequences. The PSSM is similar to an MSA based sequence profile. However

Query protein	...	K	R	L	E	H	G	G	G	V	A	Y	A	I	A	K	A	C	A	G	D	A	G	L	...
YP_002995377	...	K	Y	L	E	H	G	G	G	V	A	Y	A	I	A	K	A	A	S	G	D	V	R	E	...
YP_002958591	...	K	Y	L	E	H	G	G	G	V	A	Y	A	I	A	K	A	A	A	G	N	V	A	E	...
YP_003418650	...	S	Y	L	Q	H	G	G	G	V	A	Y	A	I	V	K	K	G	G	-	-	-	-	-	...
YP_002828572	...	S	Y	L	Q	H	G	G	G	V	A	Y	A	I	V	K	K	G	G	-	-	-	-	-	...
ZP_04861702	...	G	M	L	K	H	V	G	G	V	A	A	A	I	V	K	K	G	G	-	-	-	-	-	...
ZP_05391340	...	G	A	L	K	H	G	G	G	A	A	A	A	I	V	K	A	G	G	-	-	-	-	-	...
YP_003345806	...	E	Y	L	K	H	G	G	G	V	A	G	A	I	V	R	A	G	G	-	-	-	-	-	...
YP_003496764	...	S	H	L	K	M	G	G	G	V	A	G	A	I	R	R	A	G	G	-	-	-	-	-	...

Figure 2.1: Diagram of a Multiple Sequence Alignment. The figure shows a query protein sequence which is a fragment of Chain A of the AF1521 protein from the NCBI RefSeq database [12]. The query protein has been aligned to eight other proteins in the database. The first column contains the protein names, $\dots$ refers to continuation of the chain sequence and - to gaps in the alignment that could not be matched. The boxed column at the 11th index of the displayed sequence indicates the generation of a 20-dimensional frequency vector for that column. Each of the 20 dimensions corresponds to one of the 20 possible amino acids. For this particular column, the amino acids Y, A and G would have frequencies of 5/9, 2/9 and 2/9 respectively while the rest of the vector would be 0. The figure was reproduced from [63] by permission of the publisher, Springer for use in this thesis.

instead of directly representing the observed frequencies from an alignment, elements of the matrix are computed by dividing the amino acid residue substitution scores from pre-defined substitution matrix (like BLOSUM62 [64]) and normalized against a background frequency. Moreover, the values are dynamic and updated with each iteration of the alignment process. Figure 2.2 shows the example of a PSSM generated using the query sequence in Figure 2.1. In the MSA, the alignment is completed and the frequencies computed only from the final alignment.

MSA and PSSM are traditionally used feature encoding in both membrane protein topology prediction and protein secondary structure prediction. MSA was favoured by older methods and was later replaced by PSSM based encoding, with advantages such as detecting more distantly related sequence as described earlier. Further, in the case of SS prediction additional structure based properties like the relative solvent accessibility (see section 2.2.2, under *Feature Encoding in Existing Works*) were used to strengthen predictions. However, membrane protein topology predictors used mostly MSA or PSSM and are described in detail in the following sections.

2.1.4 Early data driven approaches using multiple sequences

Multiple sequence alignments (MSA) which provided homology information, were first implemented with NN, to improve prediction accuracies by Rost and colleagues [52].

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	...
1	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-3	5	-2	-3	-1	0	-1	-3	-2	-3	...
2	R	-2	2	-2	-3	-3	-1	-2	-3	1	-2	-1	0	-1	2	-3	-2	-2	2	7	-2	...
3	L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1	...
4	E	-1	0	0	2	-4	2	5	-2	0	-4	-3	1	-2	-4	-1	0	-1	-3	-2	-3	...
5	H	-2	0	1	-1	-3	0	0	-2	8	-4	-3	-1	-2	-1	-2	-1	-2	-3	2	-3	...
6	G	0	-3	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-3	-3	-4	...
7	G	0	-3	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-3	-3	-4	...
8	G	0	-3	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-3	-3	-4	...
9	V	0	-3	-3	-4	-1	-2	-3	-4	-3	3	1	-3	1	-1	-3	-2	0	-3	-1	4	...
10	A	4	-2	-2	-2	0	-1	-1	0	-2	-1	-2	-1	-1	-2	-1	1	0	-3	-2	0	...
11	Y	-2	-2	-3	-4	-2	-2	-2	-4	1	0	1	-2	0	3	-3	-2	-2	2	6	-1	...
12	A	4	-2	-2	-2	0	-1	-1	0	-2	-1	-2	-1	-1	-2	-1	1	0	-3	-2	0	...
13	I	-1	-3	-4	-3	-1	-3	-4	-4	-4	5	2	-3	1	0	-3	-3	-1	-3	-1	3	...
14	A	4	-2	-2	-2	0	-1	-1	0	-2	-1	-2	-1	-1	-2	-1	1	0	-3	-2	0	...
15	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-3	5	-2	-3	-1	0	-1	-3	-2	-3	...
16	A	4	-2	-2	-2	0	-1	-1	0	-2	-1	-2	-1	-1	-2	-1	1	0	-3	-2	0	...
17	C	2	-3	-3	-3	9	-2	-3	-2	-3	-1	-1	-2	-1	-3	-2	0	-1	-3	-2	-1	...
18	A	4	-1	-1	-1	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	2	0	-3	-2	-1	...
19	G	0	-3	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-3	-3	-4	...
20	D	-2	-2	1	6	-4	0	2	-1	-1	-3	-4	-1	-3	-4	-2	0	-1	-5	-3	-4	...
21	A	3	-2	-2	-2	-1	-1	-1	-1	-2	0	-1	-1	-1	-2	4	0	0	-3	-2	1	...
22	G	0	3	0	-1	-3	3	0	2	-1	-3	-3	1	-2	-3	-2	1	0	-3	-2	-3	...
23	L	-1	0	-1	0	-3	1	4	-2	-1	-1	1	2	0	-2	-2	-1	-1	-3	-2	-1	...
...																						

Figure 2.2: Diagram of a PSSM. A Position Specific Scoring Matrix (PSSM) of part of the AF1521 query protein obtained by running PSI-BLAST is shown. The first two columns contain the residue number and type in the query protein (in bold). The first row indicates the 20 possible amino acid types (in bold). Let $p_{i,j}$ denote the probability an amino acid type j (in columns) occurs at position i at the query sequence (in rows), as determined by values of a substitution matrix (such as BLOSUM62). An element in this PSSM, $m_{i,j}$ is then computed as the following log-likelihood: $m_{i,j} = \log(p_{i,j}/b_{i,j})$, where $b_{i,j}$ is the background frequency of amino acid type j . The figure was reproduced from [63] by permission of the publisher, Springer for use in this thesis.

Following this seminal work, MSA or homology information began to increasingly be incorporated by various authors. **HMMTOP** applied a HMM with multiple sequence information, based on the principle of maximizing the divergence between different amino acid compositions found within various structures in TM sequence and whole proteins [65]. **PRODIV-TMHMM**, combined the maximal divergence principle applied in HMMTOP with a sequence profile built from MSA, also using HMM [51] to predict TM topology.

2.1.5 Modern data driven approaches combining signal peptide (SP) with TM prediction

The previous works discussed, were methods for TM protein topology or helix prediction from TM protein sequences. In addition to these structures, signal peptides are important segments of proteins that contain information about protein localization. Located at the N-terminal segment of a protein sequence, signal peptides (SP) share similar levels of hydrophobicity to N-terminal TM helices (TMH) [60]. One difference between SP and TMH is that a signal peptide signals that the protein should be transported to a membrane and is typically cleaved off once the protein has been inserted into the membrane, while a TMH plays a structural role supporting the protein function (e.g. as a pore for ion transport). However, it is difficult to distinguish SP and TMH from sequences alone. Firstly, not all signal peptides are cleaved off and therefore remain within the protein sequence, requiring that the algorithm distinguish them from other TMH within the sequence. Secondly both TMH and SP share long stretches of hydrophobic residues, which makes it difficult for prediction algorithms to detect their difference based on sequence alone, leading to cross-classification of SP and N-terminal TM helices [53, 60, 66]. Thirdly it has also been shown that N-terminal TM helices are even more hydrophobic than other types of TMH [67]. As most predictors rely on sequence similarity based measures to distinguish TMH from SP, this means that N-terminal TM helices are even more likely to be predicted as signal peptides compared to other types of TM helices. Early signal peptide (SP) prediction works, such as **SignalP** [61, 68] and **TargetP** [69, 70] had been proposed to handle detection of SP from genomic sequences and also to predict the cleavage sites. However, to characterize an unknown sequence, output from topology predictors had to be combined with SP predictors, resulting in overlapping, contradictory predictions since SP could be mistaken as TMH, or vice-versa.

To address the challenges, works employing data driven approaches to handle SP and TM topology predictions have been proposed. **Phobius** [71] employed HMM to predict signal peptides as well as MP topology using single sequence information, with **PolyPhobius** [71] implementing a multi sequence model incorporating homol-

ogy information. Employing information from single sequences only, **TMMOD** [72] proposed a HMM using Bayesian posterior estimator, to predict signal peptide and transmembrane topologies in genomes. **TOPCONS** [73] employed a consensus prediction of output from five different individual predictors by creating a topology profile as input. A HMM was applied to the topology profile to obtain the best topology from the highest scoring state path. Recently, the method was extended to include signal peptide (S) prediction in a new server [32].

Besides the HMM, many neural network (NN) based works were proposed to handle the challenge of SP and TM topology prediction in a single predictor. NN based works devised different architectures to handle membrane residue preferences, employing a sliding window scheme to 'read' neighbour residues information from PSSM. In the last step, HMM or other dynamic programming algorithms were applied to predicted residue preferences from neural network outputs, to obtain the grammatically correct topology. **MEMSAT3** [53] employed a three-layered NN with PSSM as feature representation, where a dynamic programming algorithm was applied to calculate the optimal topology from NN predictions. **OCTOPUS** [9] applied 5 separate neural networks to predict transmembrane topology with PSSM based feature encoding. Four NN handled membrane residue preferences such as Inside, Loop, Membrane and Globular and the 5th was used to model the ratio of positively charged residues within a window of 31 residues. Later, signal peptide prediction was integrated in a new model, **SPOCTOUS** [74].

Recently, **TMSEG** [33] combined Random Forests (RF) in a hierarchical manner with NN output to obtain membrane topology and signal peptide predictions. Although the method could detect TM helices with a high accuracy, the topology score at 71%, was lower than that obtained by many modern predictors.

2.1.6 Other approaches applied to membrane protein predictions

While HMM and NN based approaches were commonly used to handle the membrane topology and signal prediction problem, several authors proposed other techniques

such as SVM. **MEMSAT-SVM** [29] used 4 binary SVM classifiers together with PSSM based feature representation to predict MP topology as well as signal peptides, obtaining a topology score of 89%. The method also predicts re-entrant helices, which are helices that do not fully span the membrane, entering and exiting from the same side. **Philius** [75] implemented Dynamic Bayesian Networks on single sequence information, providing TM topology prediction and signal peptide detection.

The works discussed had implemented machine learning models such as NN and SVM in combination with statistical models such as HMM to predict the final topology. In purely HMM based works, model parameters were often residue states such as loop, inside, outside, or helix/membrane.

Besides the purely data driven approaches discussed, first principle works have been proposed, where HMM parameters directly reflect biological principles. Previously, von Heijne and colleagues had implemented the hydrophobicity rule based on TM helices being more hydrophobic than other portions of the protein sequence. The hydrophobic TM rule was used together with the positive inside rule specifying that the net positive charge resulting from residues on 'inside' membrane loops, is higher than outside loop residues. In [76] a HMM based work **SCAMPI**, is presented in which a hydrophobicity parameter was used to detect TM helices, with a positive inside bias parameter to model the ratio of charges on opposite sides of the membrane. Recently, based on an observation of N-terminal helices being more hydrophobic than middle and C-terminal helices, **SCAMPI2** [67] was proposed with two additional parameters modelled in their HMM to account for their hydrophobicity. The topology score obtained on their dataset of TM proteins was 73%.

2.1.7 Non-topology Prediction methods

The previous discussion of existing works pertained to full topology prediction methods, whether SP prediction was incorporated or otherwise. Many predictors have also been developed, that focus on properties such as TM helix detection alone, or membrane residue preference prediction together with secondary structure prediction. These methods are difficult to compare with the works proposed in the thesis since

they do not directly provide topology, but their helix detection accuracies are above 90%; they may be used to detect TM helices or other hydrophobic segments for correct placement in 3-D models. Some of these works are discussed below.

MemBrain [39] proposed a K - nearest neighbour algorithm with optimized neighbours, to predict TM helices of differing lengths as well as helix boundaries correctly, together with PSSM based feature encoding. Recently, **DCRF-TM** [77] implemented a deep learning neural network towards the prediction of TM Helices. Besides TM helix detection methods, one interesting work combining both Membrane preferences and SS prediction, has been proposed. **BCL-JUFO9D** [78] combines membrane and SS preferences of residues. The membrane environment and SS preferences of residues, are an interrelated problem. The TM helices are predominantly α -helices and a strong Helix prediction from an SS classifier can strengthen a TM helix prediction when the Helix is only mildly hydrophobic. Three-state membrane environment preferences (Membrane core, Solution and Transition) were combined together with three-state SS preference (Helix, Sheet and Coil) to give 9-dimensional predictions. Several three-layered NN employing the sigmoidal activation function with learning by back-propagation were employed, resulting in 95% accuracy for identifying residue membrane preferences and 73%, for SS preference.

A summary of the TM prediction methods discussed in this chapter, is given in Table 2.1. The available SS predictors are next reviewed.

Table 2.1: Summary of main Transmembrane (TM) Topology Predictors in the Literature.

Name	[Reference]	(Year)	Model	Properties + Feature Encoding	Mode
Phdhtm_ref	[52]	(1996)	TM	NN+MSA	Multiple
HMMTOP	[65]	(1998)	TM	HMM+MSA	
PRODIV-TMHMM	[51]	(2004)	TM	HMM+MSA	
PolyPhobius	[71]	(2005)	TM+SP	HMM +MSA	
MEMSAT3	[53]	(2007)	TM+SP	NN + PSSM	
OCTOPUS	[9]	(2008)	TM	NN +PSSM	
SPOCTOPUS	[74]	(2008)	TM+SP	NN +MSA	
MemBrain	[39]	(2008)	TMH+SP	K-nearest neighbour +PSSM	
SCAMPI	[76]	(2008)	TM	First Principles and HMM * + MSA	
TOPCONS	[73]	(2009)	TM	HMM + consensus Topology profile	
MEMSAT-SVM	[29]	(2009)	TM+SP	SVM+PSSM	
TOPCONS	[32]		TM+SP	HMM + consensus Topology profile	
TMSEG	[33]	(2016)	TM+SP	NN and RF + PSSM	
DCRF-TM	[77]	(2016)	TMH+SP	Deep NN + PSSM	
SCAMPI2	[67]	(2016)	TM	First Principles and HMM *	
TOPPRED	[31]	(1992)	TM	First Principles	Single
MEMSAT	[58]	(1994)	TM	Information based	
Phobius	[79]	(2004)	TM+SP	HMM	
TMMOD	[72]	(2005)	TM+SP	HMM	
Philius	[75]	(2008)	TM+SP	DBN	
TMHMM	[59, 60]	(2001)	TM	HMM	

Methods are grouped by their mode (single vs. multiple sequences) and the type of input profile used by their learning algorithm. TM refers to TM topology predictors, SP refers to Signal Peptide predictions and TMH refers TM Helix detection without inside/outside orientation prediction.

Abbreviations: NN Neural Network, SVM - Support Vector Machine,

HMM - Hidden Markov Models, RF - Random Forest, DBN - Dynamic Bayesian Network,

MSA - Multiple Sequence Alignment and PSSM - Position Specific Scoring Matrix.

* Method also provides single sequence mode

2.2 Modelling and Predicting Protein Secondary Structures

The existing works to predict secondary structures (SS) from an input primary sequence can fall into three categories: HMM based works, NN-based works and SVM-based works. Similar to the developments seen in membrane protein topology predictors, the use of MSA-derived and PSSM -based profiles together with a machine learning approach like neural networks, drastically improved the prediction accuracies, compared to single sequence information. For example in 1996, Rost and Sander proposed the PHD method [36], combining the use of multiple sequence information together with neural networks to improve prediction accuracy by 10%.

2.2.1 Secondary Structure prediction Based on Information Theory

As discussed, the incorporation of multiple sequence information in the form of evolutionary profiles was a significant development leading to improved Q_3 accuracy. However before this development, early authors had proposed works based on Information theory, to predict the protein secondary structures of a given sequence. In 1978, Garnier, Osguthorpe and Robson proposed to use an information function to predict secondary structures from a sequence. The information function was first proposed in [80]. Based on Bayesian statistics, the method estimated the joint probability density $P(S, R)$ of observing a residue R, in a conformation S and was called the **GOR** technique [81]. The information function is given as

$$I(S; R) = \log[P(S|R) / P(R)] \quad (2.1)$$

The conditional and posterior probabilities $P(S|R)$ and $P(R)$ respectively, were estimated from a database of protein sequences that had experimentally assigned secondary structures. In the **GOR I** [81] approach, the function was modified to use

a sliding window of 17 residues, with information from the eight neighbours to the left and right to estimate the secondary structure propensity of the middle residue. Later, the **GOR II** [82] approach applied the function to a new database. In the **GOR III** [83] approach, pairwise occurrences of the central residue in the window, with every other residue in the window was incorporated in the information function. In a later development, **GOR IV** [84] further improved accuracy over the GOR III approach by considering all possible pairs within the window of 17 residues. They obtained a Q_3 accuracy of 64.4% on a database of 267 proteins.

In 1993, Rost and Sander proposed multiple sequence alignments to be used as input to neural networks [36], improving the prediction accuracy up to 72.3%, an increase of 8% points compared to GOR IV. In a later work in 1999, Jones first trained neural networks with input sequences represented by PSSM to further improve the prediction accuracy [37], obtaining 76% Q_3 after three-fold cross-validation. The inclusion of MSA information and PSSM (commonly referred here as 'sequence profiles'), were developments that represented a break-through in protein secondary structure prediction and were subsequently incorporated in several other works. **GOR V** [85] implemented PSSM based feature encoding together with the information function that had been proposed in the earlier GOR methods, reporting a Q_3 of 73.4%. A variable window size and residue triplet propensities within the window were also incorporated. However, a drawback of the method is low accuracy for Sheet class, with only 50% of Sheet residues being predicted correctly.

Later, the Consensus Data Mining (**CDM**) [86] method was proposed, in which the previous GOR V predictor was combined with a Fragment Database Mining (FDM) [87] technique to obtain secondary structures from matching structural fragments. Finally, **PROF-King** [88] combined the GOR algorithms and neural networks together with linear discriminants in a cascade to build a classification pipeline, reporting a Q_3 accuracy of 76.7%.

2.2.2 Secondary Structure Prediction based on Neural Networks

Besides the information theory approaches presented, a large number of SS predictors using neural network classifiers had been proposed. As described earlier, the PHD method and PSIPRED were amongst the earliest neural network classifiers developed to predict protein secondary structures. The first level contains a sequence to structure prediction network, whose output is fed into the second level consisting of a structure to structure prediction network, followed by a third level containing independent networks forming a jury.

First proposed in 1993, the **PHD** method [89] employed three levels of feed forward neural networks using the gradient descent algorithm, to predict protein secondary structures using single sequences, obtaining an accuracy of 70.8%. Later, using a similar network architecture and using MSA based profile matrices as input instead, it was found that the prediction accuracy increased to 72.3% [36]. Inspired by the successes of neural networks for SS prediction, the authors of **PSIPRED** [37] proposed PSSM as feature encoding using a two level neural network architecture, similar to the PHD method. As a result, PSIPRED showed an improved accuracy of up to 76.0%, averaged over three-fold cross-validation. Recently PSIPRED obtained an accuracy of 78.2% on the CASP8 dataset [6]. The PHD method is currently available as part of the 'Predict Protein Server', which provides other sequence predictions such as the solvent accessibility and location of low complexity regions [90] while a similar server for PSIPRED is also available [91]. Later, **JNet** [92] compared the effect on SS predictive performance by using various alignment profiles as input, such as MSA, PSSM and Hidden Markov Model profiles and found that PSI-BLAST based PSSM input provided the highest Q_3 (76.9%) accuracy. Subsequently an updated server of the method, (**Jpred 3**) was developed [93].

Several other neural network predictors whose architecture was influenced by the PHD method were proposed. **SABLE** [94] proposed Elman-type recursive neural networks in a two level protocol (sequence to structure prediction followed a second level of structure to structure prediction), together with a regression model to pre-

dict the Relative Solvent Accessibility (RSA) in a server. SABLE reported a Q_3 of 76.8%. **SPINE** [4] obtained a high accuracy of 79.64 % on the CB513 dataset, using a consensus of two neural network systems, with each system implementing the two -level protocol, similar to the PHD method. Later, in addition to SS prediction, the method was developed to incorporate prediction of backbone torsion angles and made available as an integrated web server known as SPINE X [95]. More recently, **SCORPION** [96] implemented context based scores by calculating pseudo potentials of residues adopting H, E or C structural states, reporting about 83% accuracy on nearly 8000 protein chains.

Several works also proposed bidirectional recurrent neural networks (BRNN) to predict protein secondary structures. A single BRNN contains a network component to capture information from the central residue in a window, flanked on the left and right by forward and backward networks to provide context to the central window [97]. **SSPro** [98] developed three and eight state secondary structure predictors (SSPro 2.0 and SSPro8, respectively) based on an ensemble of 11 BRNN, obtaining a Q_3 score of 78.1%.

The incorporation of homology modelling algorithms in protein SS predictors strengthened the prediction accuracy by 1-2% [5]. Homology modelling aligns sequences to known structures from databases like the PDB. The SS of residues from high scoring structures is assigned to the query sequence. With at least 30% sequence similarity of target to template sequences, the method can transfer *a priori* information from a structure database, but predictions become more unreliable if homologues are lacking for the given sequence. Recently, SSPro 4.0 was designed with homology modelling as part of the SCRATCH prediction server [24] and was one of the top performing methods in a large scale study [5], performing at 82% accuracy. The study also found that the performance of ab-initio methods like SPINE X was close to those of homology modelling methods (80.5%).

Further developments using BRNN were implemented in the **Porter method** [99], like using a very long sliding window of 225 amino acid residues in order to account for long range interactions. In 2007, a homology modelling version, **Porter_H** [26] was

proposed. The authors reported that, for sequences that had at least 30% similarity or higher hits to existing structures, the theoretical maximum of secondary structure prediction of 90% could be reached. It remains difficult for ab-initio methods to obtain such high accuracies. However, one drawback of Porter_H is the need to estimate about 6000 free parameters in the model. **Proteus** [25] proposed a combination of consensus based prediction approach together with homology prediction, obtaining a Q_3 accuracy of 81.3%. They developed TRANSSEC, a two-level neural network classifier, combining its output together with JNet and PSIPRED with SS assigned from any known homologues to obtain a consensus prediction.

In 2014, Yan *et al.* [21] combined SS predictions of five methods (SSPro, SPINE X, Porter, Proteus and PSIPRED) in a majority vote consensus and enhanced prediction by training a Radial Basis Function (RBF) kernel SVM on the predicted SS states. They achieved a high Q_3 of 85.6%, an accuracy that is difficult to obtain by any single predictor for a large dataset of nearly 2000 chains. Recently, several works implementing deep learning neural networks in a hierarchical manner [27, 46], have been proposed. They are discussed further in section 2.3.

Feature Encoding in Existing works:

All of the neural networks approaches discussed in this chapter so far, had relied on input feature encoding in the form of various sequence profile consisting of sequence alignments (MSA), or substitution matrices derived from homology alignments (PSSM from PSI-BLAST). As described, several works had proposed more features or incorporated structural information to enhance the level of information present in sequence profiles. For example, TRANSSEC [25] obtained predicted SS from a first level network and utilized the predicted SS to produce a structure guided alignment of homologous sequences as input to the 2nd level network. Porter represented gaps and non-standard amino acids in their MSA [99]. SPIDER 2 [27] used physicochemical properties like hydrophobicity, isoelectric point as well as predicted structural properties such as ASA and backbone angles to improve prediction accuracy.

Saras *et al.* presented an Extreme Learning Machine (ELM) called **FLOPRED**,

trained on energy potentials based feature encoding rather than the PSSM or other types of sequence profiles [41,42]. Using reference energies computed from alignments of target sequences to template structures, the probability of each residue in a window adopting a specific structural state was calculated and encoded in a profile. ELM parameters were selected through particle swarm optimization (PSO), obtaining 81% accuracy on the CB513 dataset. A similar type of encoding is proposed in experiments in this thesis.

2.2.3 Secondary Structure Prediction based on Support Vector Machines

Besides the NN based predictors, several authors have proposed Support Vector Machines (SVM) based classification systems to handle SS prediction, all using PSSM based feature encoding. **YASSPP** [100] presented a two-level cascaded classifier with each level containing 3 binary SVM classifiers and reported 81% accuracy. **DIS-SPred** [101] combined 3 binary SVM classifiers to predict the secondary structures and also employed a k -means clustering algorithm to detect residue dihedral angles. More recently, the Compound Pyramid Model (**CPM**) [3] proposed a three layer prediction system, consisting of two level SVM classifiers and back-propagation NNs in the first layer, followed by rules based classification in the second and third layer. The method reported a very high accuracy of 85.6% on the CB513 dataset, the highest reported thus far. However, the model had a high computational complexity resulting in a long run-time [3].

2.2.4 Secondary Structure Predictors based on Hidden Markov Models

Besides the SVM and NN based works, Hidden Markov Model (HMM) was developed to predict protein secondary structures. YASPIN [102] proposed a two-level approach, with first-level neural networks to predict local SS states, followed by a HMM to model transition probabilities to obtain the final secondary structures.

While YASPIN had used PSSM encoding, MSA alignments weighted with HMM

predictions had also been used by later works, described here: **Optimal Secondary Structure Prediction Hidden Markov Model (OSS-HMM)** [103] developed 36 states in a model that allowed detection of structural transitions to different states favoured by residues. **Protein Secondary Structure Predictor using HMM (P.S. HMM)** [104] used genetic algorithms to evolve the topology of HMM states in their model. OSS-HMM reported 75.5% Q₃ accuracy and YASPIN, 77.06% on their respective datasets.

Role of HMM in Secondary Structure and Membrane Topology Prediction

HMM had been popular in structure prediction algorithms and especially in the case of membrane topology prediction, drove the development of probabilistic models for sequence representation and subsequent topology prediction [65]. This was because HMM were particularly useful as the membrane residue preferences could be explicitly modelled as various state symbols representing inside, outside loops and transmembrane helix structures [65]. As the protein topology is direction dependent with fixed grammar or rules (see Figure 3.4), it is a problem that was well modelled by HMM since they may be also thought of as directed graphical models [59, 65]. It was therefore effective to use either single sequence or MSA to probabilistically learn the target states to be modelled.

With the implementation of PSSM based input features, it became possible to represent evolutionary information in the form of amino acid substitution probabilities. Residue interactions could also be modelled from the PSSM with a sliding window scheme [9]. Although the sliding window scheme contained its own challenges as discussed under section 2.4.3, the PSSM representation popularized the use of neural networks since they could easily be fed as input features to neural networks to be classified into the membrane residue preferences [9, 53, 74]. Subsequently, the final topology of the predicted model could be determined by a HMM or through a dynamic programming algorithm [32].

In the case of protein secondary structure prediction, HMM were less popularly used

compared to membrane protein topology prediction. **OSS-HMM** used a purely HMM based model since transitions between different secondary structures may be modelled explicitly by the HMM states, leading to more interpretability than provided by NN based methods. Other HMM based methods as mentioned earlier, used HMM in combination with neural networks [102,104] and were not standalone HMM based works. However, in later years, in terms of predictive performance there was a clear indication that neural network based predictors were much more effective compared to HMM based methods in two wide scale studies on thousands of protein chains [5,6]. For instance, SSPro [24] employing neural networks obtained a Q_3 score of 82% vs **YASPIN** the highest performing HMM based predictor which obtained about 75% [5]. The two wide scale studies indicated that the use of PSSM based representation together with neural networks based classifiers were more effective in terms of Q_3 accuracy. It is suggested that the two-level structure to structure network architecture employed by many neural network predictors (as discussed in section 2.2.2) serves as a hierarchy of classifiers that reinforces correct classifications and penalizes the wrongs ones. Subsequently, the two-level structure to structure network together with PSSM based representation, may be the reason behind effective prediction of protein secondary structures, leading to NN outperforming other models based on HMM or SVM.

Table 2.2 summarizes the secondary structure predictors discussed in this chapter.

Table 2.2: Summary of Main SS Predictors in the Literature

Learning Method	Name (Year)[ref]	Properties
NN-based	PHD (2004) [90]	2 level feed forward NN + filter network
	PSIPRED (1999) [37]	2 level feed forward NN
	JPred (2000) [92]	2 level feed forward NN + filter network
	SABLE (2005) [94]	Elman-type recursive NN
	SSPro (2002) [98] and Porter (2005) [99]	Bidirectional recurrent NN (BRNN)
	SSPro (2005) [24] and Porter_H (2007) [26]	Ensemble of BRNN+ Homology Modelling
	SPINE (2007) [4]	Feed-forward NN
	Proteus (2006) [25]	2 level feed forward NN + filter network
	FLOPRED (2012) [42]	ELM + Particle Swarm Optimization
	SCORPION (2014) [96]	Feed-forward NN
	Sscon (2014) [21]	Consensus of NN predictors + SVM layer
	SPIDER 2 (2015) [27]	9 Deep NN in three levels
	DNSS (2015) [46]	Deep NN with RBM
SVM-based	CPM (2011) [3]	SVM + NN in level 1 of 3-level complex architecture
	DISSPred (2009) [101]	3 SVM + k-means clustering
	YASSPP (2006) [100]	2-level cascade with 3 SVM classifiers in each level
HMM-based	YASPIN (2005) [102]	Feed forward NN + 7 state HMM
	OSS-HMM (2006) [103]	36 state HMM
	P.S.HMM (2007) [104]	2 levels. GA + HMM in level 1 combined with NN in level 2
Information-Theory based	GOR V (2002) [85]	Bayesian statistics
	CDM (2006) [86]	Consensus of GOR V and Structure mining method [87]
	PROF-King (2000) [88]	4 level cascade Information theory, NN + linear discriminants

NN, SVM and HMM refer to Neural Networks, Support Vector Machines and Hidden Markov Model, respectively. All methods have used sequence profiles like MSA or PSSM except FLOPRED which employed energy potentials based encoding. BRNN, ELM and PSO refer to Bidirectional Recurrent Neural Networks, Extreme Learning Machine and Particle Swarm Optimization, respectively. GA refers to Genetic Algorithms. Other abbreviations refer to SS classifiers which are described in the text.

2.3 Deep Learning Neural Networks and their applications in Structure Prediction

Deep Belief (learning) Networks (DBN), initially proposed as *autoencoders*, comprise of multiple Restricted Boltzmann Machines (RBM) stacked together [105]. An RBM possesses a visible layer and a hidden layer, with the visible layer of one RBM serving as the hidden layer to the next, in a top-down approach. The training of a DBN typically comprises an unsupervised pre-training step to obtain network weights followed by a supervised training step that refines the weights further. For the unsupervised learning step, the contrastive divergence learning rule is used to generate higher level abstractions of the input features [40]. Network weights in the DBN are learnt in a layer wise manner. Backpropagation combined with a gradient descent approach is used in the supervised training step to fine-tune the weights.

It has been shown in the literature that the pre-training step assists generalization to unseen data [106, 107]. DBN have been successful in applications such as image reconstruction [105] and 3-dimensional object recognition [108] where they showed a good improvement over other classifiers like SVM.

More recently, works applying deep learning algorithms to predict TM helices and secondary structures have been proposed. In 2016, **DCRF-TM** [77] implemented conditional Random Fields in a deep learning neural network and applied their model to the prediction of TM helices. Besides the standard 20 features from PSSM, they also included the following 5 raw features from properties calculated from the protein chain - elasticity, hydrophilicity, hydrophobicity, free energy and polarity. While not a direct topology prediction method, it was applied to 24 GPCR structures to evaluate TM helix placement. In 2015, **SPIDER 2** [27] applied three levels of deep learning networks (DLN) to predict SS, accessible surface area and backbone dihedral angles amongst others. Training their DLN on all the predicted structural properties, they reported a Q_3 of nearly 81% on an independent test set. In another predictor **DNSS**, [46], outputs from two independently trained Tier 1 DLN were combined and sent to a third DLN (Tier 2) to refine SS predictions. Each DLN contained Restricted

Boltzmann Machine (RBM) arranged in layers, with the study concluding that up to 4 RBM layers gave the best Q_3 accuracy of nearly 81%.

2.4 Challenges in Secondary Structure Prediction and Membrane Topology Predictions

Recent works in SS prediction had shown a high Q_3 of nearly 86% [3, 21], demonstrating favourable results. For the MP topology prediction, earlier methods had reported very high accuracies of up to 89% [29] on smaller datasets, but the current best performance is at about 80% [32].

2.4.1 Theoretical upper Limit for Accuracy & Noise from Variation

The theoretical accuracy for SS prediction is estimated to be 88% to 90% [2, 22] and is yet to be reached. There are several reasons for the difficulty in reaching the theoretical upper limit. Sequences are more prone to variation while structures are more conserved evolutionarily [34]. This results in possible, minor sequence variations observed even in identical structures, needing SS and MP topology predictors to be sensitive. By representing the input encoding in sequence profiles such as MSA and PSSM, distinct values in sequence space are reduced to the same class in the structure space, a similar problem that also affects membrane residue preference and topology prediction. The loss of information is hard to recapture. Additionally, secondary structures, themselves may contain small experimental variations depending on the 3D structural models and SS assignment programs used [2].

Similarly for membrane proteins (MP), the difficulty in detecting membrane orientation and in crystallizing the MP structure leads to different topological annotations dependent upon the source [29].

2.4.2 Effects of PSSM based Representation

The representation of a sequence as PSSM incorporates homology information from multiple sequences. As structures like TM helices tend to be heavily conserved in sequence families, the PSSM has been a commonly used feature encoding to model the sequence to structure relationship for the one-dimensional topology prediction. Since the PSSM is a sequence profile based on substitution matrices, its elements comprise discrete whole numbers belonging in ranges (for e.g. $[-10, 10]$ or $[-14, 13]$). Hence it contains many repeated values being mapped into the same structure space which results in overlapping signals for a given region or class. For instance, Inside (I) and Outside (O) regions in TM proteins tend to be less hydrophobic and share similar sequence characteristics with each other. Due to the shared properties, the PSSM representation of I and O regions may contain similarities which makes separation difficult. Similarly in the case of SS prediction, many repeated values in the PSSM could be mapped into the three structures of Helix (H), Sheet (E) and Coil (C), making discrimination difficult.

2.4.3 Modelling Structural Properties in Sliding Windows

Another challenge is the difficulty of capturing the effect of long range interactions by a sliding window approach. In the case of SS predictors, structures like α -Helix that contain sequentially and spatially close residues can be predicted more easily than β -strand structures that are spatially close in the 3D model, but are distant neighbours in a residue. Often, Helix classification accuracy in an SS predictor is higher than Sheet which contains β -strands.

Particularly for MP predictors, since the topology score penalizes wrong lengths and numbers of the predicted TM regions, it has been a challenge to determine the appropriate window size that captures the residue preferences correctly. TM helical regions (M) are approximately 8 to 35 residues long. However, the I and O regions are subject to much more variability in sequence length. The sliding window aids in capturing the interactions of amino acid residues between neighbours. A longer window length captures the effect of more distant neighbours. In current works, the M, I and O regions

are modelled using various window sizes like 39, 31 [9] or 33, 35 [29]. Moreover, even for a single region there exists no standard window size in the literature despite all methods employing the same PSSM encoding. The above mentioned drawbacks with PSSM representation as well as the influence of window size in modelling introduce uncertainties at the level of features that make classification difficult.

SS prediction and MP topology prediction are the first level input to higher order structure prediction and template alignment algorithms [28, 63]; even small errors at the primary level can result in a wrong model being selected downstream of the modelling process [19]. The challenges outlined motivate the need for new approaches to be designed. In this thesis, two approaches are proposed for the SS prediction problem and MP topology prediction, respectively. In the first approach, an energy potentials based feature encoding is used to represent protein sequences for SS prediction, instead of the commonly used PSSM or other types of sequence profiles. The energy based representation was derived from the CABS algorithm [41] and consists of alignments of target sequences to template structure, taking into account spatially close neighbours [42]. Moreover, to allow for the efficient separation of SS structural states, a Complex-valued neural network classifier is used, as the orthogonal decision boundaries provided by the Complex plane give an extra surface for classification compared to the Real plane [43]. Finally, a compact training model is proposed to overcome challenges resulting from over training the SS classifiers.

In the second approach, deep belief networks (DBN) are proposed to predict SS as well as MP topology. In DBN, higher order information is extracted from the raw PSSM or sequence profiles due to the layer by layer learning proposed by the contrastive divergence algorithm [40]. This promotes learning of the underlying residue interactions within the window that may otherwise be difficult to extract by existing learning algorithms used by current networks. While DBN had been used to predict TMH (by DCRF-TM), they have yet to be applied to topology prediction or the detection of signal peptides. For the MP topology prediction, a prediction model based on an ensemble of DBN classifiers, is proposed. Deep learning approaches had been

previously used to predict protein secondary structures, as reviewed in section 2.3. The previous works employed PSSM based features as input and adopted hierarchical DBN architectures. While structure based properties had been used to enhance the prediction in previous works, they mainly relied on the PSSM based encoding. Here, a single DBN is applied to the SS prediction problem using two different types of feature encoding - the traditional PSSM and CABS energy potentials. The performance of the two types of encoding are compared, with the construction of an ensemble DBN being the scope of a future work further discussed in chapter 6.

2.5 Summary

This chapter presented a review of existing works that have been proposed to predict two types of structures from protein primary sequences: membrane protein (MP) regions (including topology) as well as protein secondary structures (SS).

Most of the algorithms proposed may be divided into statistical approaches such as Hidden Markov Models or Information theory based methods (in the case of SS predictors), or machine learning approaches like neural networks (NN) and SVM. For MP topology predictors, predominantly HMM based works have been proposed, owing to the need to obtain the correct topology according to biologically acceptable rules. Even NN based works for topology prediction first classify membrane residue preferences and then add a layer with HMM or other dynamic programming techniques to build the final topological model. For SS prediction, NN based methods have proven to be the most successful predictors in terms of the Q_3 accuracy, although HMM approaches may assist interpretation of structural transitions favoured by residues in different segments of the protein sequence.

A review of Deep Belief (learning) Networks and especially their applications in the prediction of protein SS and MP regions was also given. Several recently proposed works reported similar Q_3 accuracies to existing works.

The representation of evolutionary information in the form of sequence profiles derived from alignments of homologous sequences (e.g. PSSM), has been an important development in both MP topology and SS prediction, greatly improving prediction

accuracy. However as discussed, the theoretical accuracy for SS prediction is not yet reached and some structures like β -strands are harder to model, as some long distance residue interactions remain difficult to capture by the existing works. For the MP topology prediction, the PSSM may not be adequate to sufficiently capture the diverse sequence properties exhibited by the transmembrane (TM) regions and non-TM regions, needing new techniques to deal with the PSSM effectively. Additionally, the topology scores of current predictors is lower than the older ones, which may be attributed to newly discovered MP with more complex topologies that must be handled correctly. The remainder of the thesis chapters propose works to contribute some developments in these above mentioned areas.

Chapter 3

Membrane Protein Topology

Prediction using Deep Belief

Networks

3.1 Introduction

The existing works in transmembrane (TM) topology prediction had shown improvements in the topology scores as discussed in chapter 2 . Similar to that of protein secondary structure prediction works, the employment of machine learning approaches together with evolutionary information in the form of multiple sequence alignments and position specific scoring matrices resulted in high accuracies achieved. Recently, TOPCONS implemented a topology profile that resulted from the combined outputs of five separate predictors [32]. Using the topology profile as input, the topology score obtained was 80% on a dataset of *alpha*-helical transmembrane proteins. For the first-principles based works, the recently developed SCAMPI 2 method employed the N-terminal hydrophobic helices rule to effectively distinguish transmembrane and non-transmembrane proteins and aid topology prediction [67]. Nevertheless, their overall topology score when considering TM proteins alone, remained low at 61.5%. The prediction of signal peptides together with transmembrane topology has been implemented in various works [29, 74, 75, 109, 110]. However, because signal peptides share hydrophobic sequence characteristics together with the N-terminal TM regions,

it has been a challenge to model signal peptide prediction accurately without an accompanying drop in topology prediction [29, 67].

Membrane protein sequences are usually represented by Position Specific Scoring Matrices (PSSM) [50]. However, it has been a challenge to model different regions of the TM sequences using PSSM, as they contain many repeated values to represent various residues whose membrane environment preferences such as Membrane (M), Inside (I) or Outside (O) must be predicted. This leads to noise in the PSSM based feature encoding. Further, it has been a challenge to determine the optimal window size to capture the neighbouring residue interactions; different works have used various window lengths to model the regions to be predicted [9, 29]. These challenges have been discussed in detail in sections 2.4.2 and 2.4.3.

Here, Deep Belief Networks [105] are proposed for their ability to model a higher level of feature abstraction, applied to PSSM input. A Hierarchical Deep Ensemble Network (HiDEN) comprising of multiple Deep Belief Networks (DBN) connected in an ensemble, was used to predict membrane protein topologies from sequence, with the PSSM generated from PSI-BLAST [50] serving as input features. The network employs contrastive divergence learning [40] with a backpropagation algorithm to update the weights. The contrastive divergence algorithm provides a high level abstraction of the PSSM, enabling the underlying interactions between residues in the window to be captured as correlations within the reduced feature set. The higher level features extracted from the raw PSSM input provide increased discrimination and are shown to effectively distinguish between membrane preferences of residues. To obtain a rapid convergence using gradient descent, a risk sensitive hinge-loss error was implemented [111]. An ensemble combining the outputs of multiple deep belief networks was implemented to increase the strength of the prediction. A sequence fitting block was implemented to detect the correct topology of the protein. HiDEN is applied on recently developed datasets of transmembrane, signal peptide and globular protein sequences [32] that are next described.

3.2 Datasets

Four datasets provided by TOPCONS are used to assess the performance of the HiDEN method developed [32]. These are the (i) transmembrane (TM), (ii) transmembrane proteins containing N-terminal cleavable signal peptides (SP+TM), (iii) Globular proteins (Globular) and (iv) Globular proteins with a signal peptide (Globular + SP). The Globular+SP dataset does not contain any TM helices. The resultant numbers of proteins in the TM and SP+TM dataset were 313 and 752 respectively. The Globular and Globular+SP datasets contained 3597 and 2194 proteins, respectively. The total number of proteins was 6856. Topology assignments had been done by combining annotations provided by various sources such as OPM [55], PDBTM [10] and UniProt [56]. The final topology contained 4 class labels comprising Membrane (M), Inside (I), Outside (O) and Signal peptide (S). The number of proteins in each of the four datasets and the percentage of residues belonging to each class are given in Table 3.1.

Table 3.1: TOPCONS Datasets. Brackets indicate percentage of residues labelled M, I, O or S in the respective datasets.

Dataset	Proteins	Residues	Classes (%)
TM	313	99404	M (34.16), I (40.48), O (25.36)
SP+TM	752	581336	M (5.24), I (19.74), O (71.74), S (3.27)
Globular	3597	2057090	I (100)
Globular+SP	2194	864877	O (94.04), S (5.96)

This section describes the generation of features from protein sequences. **Position Specific Iterative - Basic Local Alignment Search Tool** (PSI-BLAST) which is a widely used alignment algorithm to detect similar sequences, was used to generate Position Specific Scoring Matrices (PSSM) given a FASTA format sequence as input [50].

PSI-BLAST is a heuristics based sequence alignment algorithm that, given a query sequence as input, builds an alignment of related sequences that are selected from a pre-existing database of sequences. Matching substrings to the query sequence are located using a dynamic programming algorithm and the alignments scored to ascertain the best matching sequences, also termed as hits. Homologous proteins

often share similar sequences since they are derived from a common ancestor. It is assumed that closely related sequences will produce high scoring matches and the reverse relation, that is, high scoring matches indicate closely related sequences is also true. Regardless, as the size of the search database increases, the chance of detecting a poorly related but high-scoring alignment also increases. To give an idea of the confidence of the reported hit, the PSI-BLAST program also reports the Expected value or E-value. The E-value indicates the expectation that an alignment of a given score could occur randomly when searching a database of the given size. A low E-value corresponds to significant alignments with low chances of randomly being selected. Typically, cut-off E-values up to 10^{-3} are used [29].

Apart from building an alignment of related sequences and reporting the score of the hit, a sequence profile known as a Position Specific Scoring Matrix (PSSM) is also built from the iterative database searches. Given a query sequence containing M amino acid residues, the elements of the PSSM are log-likelihood ratios of residue substitutions at each position of the query sequence compared against a naturally occurring baseline frequency. The baseline is provided by a pre-defined block substitution matrix such as BLOSUM62 [64]. For instance, a native sequence may require the first 15 amino acids to be highly hydrophobic in order that it may be targeted to the membrane for localization. A sequence from the database that has the first 15 residues to be highly hydrophilic would be penalized and given a low scoring alignment by PSI-BLAST. Since there are 20 amino acids that are naturally occurring, the size of PSSM for a given query is $M \times 20$.

A summary of steps to generate the PSSM based feature encoding is given:

- Given a query sequence of length M , run PSI-BLAST against database to generate the Position Specific Scoring Matrix (PSSM) of dimensions $M \times 20$
- For each residue $i \in M$, read the PSSM in a sliding window of size 31 to obtain its neighbours' information. For a central residue i , read $i - 15$ neighbours to the left and $i + 15$ neighbours to the right, to obtain the final feature matrix of dimensions $M \times 620$
- Use the feature matrix as input to train HiDEN or obtain predictions from test

sequences

In the next section, the details of the generation of PSSM derived features from sequences are presented.

3.3 Feature Encoding: Generating PSSM for Sequences

The process of encoding a sequence as features to be used as input for the Deep Ensemble Network classifier was as follows. Firstly the PSSM were generated using PSI-BLAST and secondly the PSSM were converted to feature matrices by reading in the neighbour residue information using a sliding window. In the final step, the generated features are represented as input to the HiDEN classifier. A schematic illustrating the representation of sequences as features that serve as input to the deep ensemble classifier is given in Figure 3.1. The PSI-BLAST program was downloaded

Figure 3.1: Flow Diagram of Membrane Residue Prediction. The figure shows HiDEN prediction of membrane sequence residue preferences. M is the number of residues in a sequence. The output, $\hat{y}^t$ represents residue level predictions.

from the BLAST+ toolbox provided by the National Center for Biotechnology Information (NCBI) [112]. To obtain high quality, relevant hits, the Protein Reference Sequence (*RefSeq*) database [12] was used instead of the non-redundant database that is commonly used. Transmembrane sequences are prone to false positive matches, due to their highly hydrophobic segments needed in a membrane environment. Even unrelated membrane proteins can be detected as hits due to the physical pressures imposed by the membrane requiring the presence of hydrophobic residues [113]. As

described earlier, the PSSM itself is a generic feature representation method that introduces elements of overlap into the classes. The inclusion of spurious sequences resulting from false hits into the PSSM can skew the profile of a query sequence considerably, causing it to be predicted wrongly. To avoid these and other problems, the *RefSeq* database containing carefully curated protein sequences gathered from multiple primary records is used instead of the non-redundant database commonly used. Further, a strict E-value cut-off of 10^{-5} is employed.

The protein *RefSeq* database was downloaded from an FTP site ¹ on June 2015. The number of iterations was set to 3. SEG filtering was done to remove hits arising from low complexity regions in sequences. All other settings were left as the default program values. Under these settings, all sequences had hits with the exception of 57 sequences for which the settings proved too strict and no hits were obtained. For these remainder sequences, the E-value was increased to 1000. For 14 sequences within the remainder, SEG filtering was set to off in order to detect hits. In this work, the size of the sliding window had been set to 31. Various other works had employed different window sizes, depending on the region being modelled. OCTOPUS employed window sizes of 29 for transmembrane regions and 31 for Inside/Outside residues. Their final output network further used a window size of 39 and 51 for smoothing predictions of Membrane (M) and Globular (G) residues, respectively [9]. MEMSAT-SVM found 33 to be the optimal window size for predicting TM regions [29]. Here, it was found that varying the window size did not have a significant effect on improving the accuracy, although using too small a window size (e.g. below 19) could lead to a drop in the accuracy. In this work a fixed window size of 31 was employed for all the classes - membrane, inside, outside regions or signal peptides. The process of PSSM and feature matrix generation was repeated for all sequences from the four datasets employed in this study.

¹<ftp://ftp.ncbi.nlm.nih.gov/blast/db>

3.4 Architecture of HiDEN

In this section the architecture of the Hierarchical Deep Ensemble Neural Network (HiDEN) used to predict membrane protein and signal peptide topologies is presented. Here, the transmembrane helical structures (M), Inside (I) and Outside (O) regions as well as signal peptides (S) are collectively referred to as residue preferences or locations with respect to the membrane. The prediction of residue preferences into one of four possible classes, M, I, O or S is formulated as a decision function to be approximated by HiDEN. Let $\mathbf{x}^t \in \mathfrak{R}^N$ be the features representing an amino acid residue t and $N = 620$ corresponds to the features obtained from the PSSM. Then the set of observations to predict the protein topology is given by $\{(\mathbf{x}^1, \mathbf{y}^1), \dots, (\mathbf{x}^t, \mathbf{y}^t), \dots, (\mathbf{x}^M, \mathbf{y}^M)\}$, where $\mathbf{y}^t \in \mathfrak{R}^q$ are the coded class labels describing membrane location preferences. Since there are 4 classes, $q = 4$. The coded class labels $\mathbf{y}^t$ are obtained from class labels c^t as follows:

$$y_j^t = \begin{cases} 1 & \text{if } c^t = j \\ -1 & \text{otherwise} \end{cases} \quad j = 1, \dots, q \quad (3.1)$$

where $c^t \in \{1, 2, 3, 4\}$ are the numeric class labels representing M, I, O and S in order, respectively. The problem of predicting residue preferences with respect to the membrane is given as the following functional approximation: $F : \mathbf{x}^t \in \mathfrak{R}^N \rightarrow \mathbf{y}^t \in \mathfrak{R}^q$, where $\mathbf{x}^t$ represents the input feature vector and $\mathbf{y}^t$, the predicted network output. In the first level of the network, the predicted residue preferences are used to classify an input sequence into four groups corresponding to the four datasets TM, SP+TM, Globular and Globular + SP. This step distinguishes membrane and non-membrane proteins and detects the presence of signal peptides in the sequence. The second level of the network focuses solely on membrane proteins (TM and SP+TM groups) to obtain their correct topologies. The network architecture is described in detail as follows. HiDEN consists of two levels, each comprising multiple Deep Ensemble Networks (DEN) that predict membrane protein residue preferences. A diagram showing the overall architecture of HiDEN is presented in Figure 3.2. The first level contains a Global Deep Ensemble Network (GDEN) that predicts residue preferences amongst four possible classes M, I, O and S ($q = 4$). Within the first level, the GDEN residue

predictions are sent to a decision block that determines if the sequence is a membrane or non-membrane protein. The decision block applies the following two rules. Firstly, if 8 or more consecutive residues are predicted as M by the GDEN, the sequence is taken to be a membrane protein. Secondly, if 11 or more consecutive residues are predicted as S in the first 45 residues of an input sequence, a signal peptide is predicted. Together, the GDEN and decision block perform a group level classification to classify an input sequence into TM, TM proteins containing signal peptides (SP+TM), globular proteins (Globular) or globular proteins with signal peptides (Globular+SP).

The second level contains two independent, Local Deep Ensemble Networks (LDEN)

Figure 3.2: Schematic of Hierarchical Deep Ensemble Neural Network (HiDEN). HiDEN applies a hierarchical, two-level approach to predict membrane protein topology. In the first level, the Global Deep Ensemble (GDEN) makes residue predictions given an input sequence represented by PSSM. The decision block classifies it into one of the 4 groups - Transmembrane (TM), Transmembrane with signal peptide (SP+TM), Globular (G) and Globular + SP (G+SP). In the second level, sequences that had been classified as TM or SP+TM are sent to their respective Local Deep Ensemble Network (LDEN) to obtain residue predictions. Next, the predictions are sent to a Sequence Fitting Block to obtain the topology grammar. Both GDEN and LDEN are presented with the primitive, PSSM based features as input. Dotted arrows indicate residue predictions flowing from the first to second level networks.

to predict residue preferences of proteins in the TM and SP+TM groups, respectively. Their network architecture is identical except for the number of output neurons, where $q = 3$ for the LDEN predicting TM group and $q = 4$ for the second LDEN predicting the SP+TM group proteins. The PSSM derived features represented by $\mathbf{x}^t$ are again presented as input to the respective LDEN, depending on the group classification by

the first level ensemble network. For example, if the group classification determined the sequence to be TM, the sequence is sent to the first LDEN for a second level of residue preference prediction.

Before describing the sequence fitting block, the architecture of the Deep Ensemble Networks is first described. Both the GDEN and LDEN contain 3 Deep Belief Networks (DBN) - DBN_1 , DBN_2 and DBN_3 connected in an ensemble. A DBN contains multiple Restricted Boltzmann Machines (RBM) stacked together in a layer-wise manner [105]. An RBM consists of a two layered neural network that contains no intra-layer connections. The hidden layer of one RBM serves as the visible layer to the next, with the final, bottom-most RBM being connected to the output layer. The output layers of the three DBNs are fully connected to the final output layer of the DEN. The architecture of DBNs connected in an ensemble network is given in Figure 3.3.

Figure 3.3: Architecture of the Deep Ensemble Network (DEN). A DEN comprises of three Deep Belief Networks (DBN). The output of 3 DBN are combined in an ensemble output to improve residue prediction accuracy. All DBN receive 620 primitive features from the PSSM input, with an additional bias neuron. The number of hidden and output nodes varies and is given in Table 3.2.

The RBM weights within each DBN, are pre-trained with the contrastive diver-

gence learning algorithm in an unsupervised manner [40]. The algorithm employs a greedy, layer by layer learning. The hidden neurons of the RBM employ a logistics activation function and the output layer employs a linear activation function. The Deep Ensemble Network combines the outputs of Deep Belief Networks and employs a back-propagation learning algorithm. To increase the rate of convergence, the risk sensitive hinge loss error [111] is employed. Given actual output y and predicted output y^t , the risk sensitive hinge-loss error is defined by the following equation:

$$E = \begin{cases} m * (y - \hat{y}) & y\hat{y} < 0 \\ y - \hat{y} & 0 \leq y\hat{y} < 1 \\ 0 & y\hat{y} \geq 1 \end{cases} \quad (3.2)$$

where m is the risk factor and is set to a value in the interval [1.2, 1.3]. A single DBN, when given low-level features such as the PSSM, is capable of extracting higher-level features that capture the underlying correlations between features, thereby enabling efficient class-wise discrimination. Here, the underlying correlations likely arise from neighbours from within a window influencing the structure of the residue to be predicted. All DBN receive 620 features as input from the primitive PSSM and contain an additional bias neuron. DBN used with multiple architectures obtain a different high-level abstraction of the input features, allowing different distributions to be learnt. The number of hidden neurons employed in each of the DBN is given in Table 3.2. The use of multiple DBN in the Deep Ensemble Network helps to generate dif-

Table 3.2: HiDEN Settings. No. of neurons used by each DBN in HiDEN. The input layer consists of 621 neurons and the output layer consists of either 3 or 4 hidden neurons. A colon separates the input, hidden and output layers.

	GDEN	LDEN-TM	LDEN-SPTM
DBN_1	621:450:4	621:400:3	621:400:4
DBN_2	621:450:50:4	621:450:3	621:450:4
DBN_3	621:450:40:4	621:400:40:3	621:450:40:4

ferent underlying abstractions of the input primitive PSSM, contributing to a slightly higher residue prediction accuracy than if a single DBN were used. Next, the residue predictions of the LDEN are sent to a sequence fitting block to determine the global

topology of the sequence.

3.5 Sequence Fitting Block

Owing to the properties of transmembrane helical segments that span the membrane, inside and outside regions alternate with one another, allowing only two biologically permissible topologies for a standard integral TM protein anchored to the cell membrane (see chapter 1). In the case of a TM protein containing an even number of M regions, the non-membrane regions at the beginning and end regions must be of the same type. An example of a TM protein with two membrane (M) regions is shown in Figure 3.4. A beginning Inside region must be matched by an ending Inside region; similarly a beginning Outside region must be matched with an ending Outside region. For a sequence with an odd number of M, the begin and ending Inside or Outside region types will be opposite as shown in the figure.

Figure 3.4: Topology of α -helical transmembrane (TM) protein. The typical topology of two hypothetical TM proteins is shown. The Outside, Membrane and Inside regions are indicated. *Membrane* regions (M) are grey rectangles, *Outside* regions are solid white shapes while *Inside* regions are crossed white shapes, respectively. The N and C termini indicating the beginning and end of sequences respectively, are also labelled. For an odd number of M regions (panel a), the N and C termini must be opposite region types. For a TM protein with an even number of M regions (panel b), the beginning and ending regions must be of the same type.

Atypical membrane spanning structures deviating from this rule may be observed for outer membrane bacterial proteins such as the Wza capsule export protein in *E. Coli* [114]. However such cases are rare. These two biologically permissible topologies are used to guide the topology made by the sequence fitting block.

After having obtained the residue preference predictions from the LDEN, the membrane regions (M) and signal peptide regions (S) are fixed, allowing the non-membrane containing regions (stretches of I and O residues) to be variable. The sequence fitting block fits the correct topology in two steps. Firstly, the optimal non-membrane regions are determined based on the frequency of predictions in a given non-membrane segment. Secondly, if one of the two permissible topologies are not obtained, it implies that membrane regions were not predicted correctly and a membrane region rectification step is performed. According to the first step, the sequence fitting block calculates the conditional probability of a given non-membrane region being I or O, dependent on the predicted frequency of O or I residues in that segment. For instance, a non-membrane segment of length 10 containing 7 Inside residues and 3 Outside residues is predicted as I, since $0.7 > 0.3$. Figure 3.5 provides an illustration. In the

Figure 3.5: Sequence Fitting Block. An example of topology prediction performed by the Sequence Fitting Block in the 2nd level of HiDEN for an input sequence of length N . The first 10 residues indicate non-membrane segment, k . Since the conditional probability, $P(I|k) = 0.7 > P(O|k) = 0.3$, the segment will be set to Inside. Residues $N - 9$ to N indicate a non-membrane segment, $k + 1$. The segment is read in a window of 8 residues and the conditional probability $P(M|k + 1)$, obtained. If three consecutive windows have probabilities higher than 0.75, the segment is deemed to be a membrane segment. Since $P(M|k + 1) = 7/8, 7/8$ and $6/8$ for residues $N - 9, N - 8$ and $N - 7$ respectively, all residues in segment $k + 1$ will be predicted as M.

second step, a split membrane region is merged and checked if a permissible topology is obtained. If the obtained topology is still not permissible, it indicates that the LDEN predictions have missed membrane regions. To detect the missed membrane regions, non-membrane regions are read in sliding windows of size 8 and the number of M predictions in each window, counted. If there are at least three such consecutive windows containing 6 or more M residues, the region is predicted as M. Finally, if the two steps performed by the sequence fitting block still do not yield a permissible topology, the allowed topology closer to the prediction is selected in the final topology.

More details of the sequence fitting block are presented in [115].

The final predicted topology serves as the output of the second level LDEN networks. To train the HiDEN model, 75% of residues from each of the four datasets are randomly selected as training samples and the remainder 25% are used as test. To provide a stringent measure of HiDEN performance an additional 10 fold cross-validation study was also performed on the TM dataset by randomly selecting 90% of proteins for training with the remaining 10% being used as test. The performance of HiDEN on the TM dataset is then compared with other classifiers reported in the literature.

3.6 Performance Measures

This section describes the measures used to assess the performance of HiDEN. The topology score is a sequence wide score that measures the orientation of the predicted membrane and non-membrane regions with respect to the cellular membrane. The commonly used definition in the literature is described. For a single protein sequence to have a correct topology, the following 4 criteria should be satisfied.

1. The first residue in the sequence is predicted correctly (Correct N-terminus)
2. The last residue in the sequence is predicted correctly (Correct C-terminus)
3. The number of membrane helical regions is predicted correctly (M regions)
4. The overlap between actual and predicted membrane helices is at least 5 residues

For the SP+TM dataset, in addition to the 4 criteria listed above, a signal peptide should be predicted in the N-terminal of the sequence. For a protein in the Globular dataset, no signal peptides or membrane regions should be predicted in order for its topology to be considered correct. In the case of a protein from the Globular + SP dataset, only an N-terminal signal peptide should be predicted, to be considered as a correct topology [32]. The topology score η_T is reported as the percentage of proteins

in a dataset with correct topologies:

$$\eta_T = \frac{\text{Proteins with Correct Topology}}{\text{Total proteins}} * 100 \quad (3.3)$$

Besides the topology score, global performance measures such as the overall accuracy (η_o), class-wise accuracy (η_j) and averaged class-wise accuracy (η_a) are calculated from the confusion matrix. The measures are defined in the following equations:

$$\eta_o = \frac{\sum_{j=1}^4 q_{jj}}{\sum_{j=1}^4 N_j} \times 100 \quad (3.4)$$

$$\eta_j = \frac{r_{jj}}{N_j} \times 100 \quad (3.5)$$

$$\eta_a = \frac{1}{4} \sum_{j=1}^4 \eta_j \quad (3.6)$$

$$(3.7)$$

Here, measures for the SP+TM dataset have been presented, where $j \in M, I, O, S$. For the TM dataset, S residues are absent, giving a total of three classes. r_{jj} refers to the total number correctly classified residues in class c_j and N_j denotes the total number of residues in class j .

Next, section 3.7 compares the performance of HiDEN with linear classification using SVM as well as a meta-cognition based learning model known as Projection-based meta-cognitive learning Radial Basis Function Network (PBL-McRBFN). Subsequently in section 3.8 various aspects of HiDEN performance on the four datasets are discussed. The group-level prediction results of the GDEN as well as the overall topology score obtained by the network are also given. Lastly the method's performance is compared to several other existing topology predictors in the literature.

3.7 Comparison of HiDEN Performance

This section describes a study in which the performance of HiDEN is compared with a Projection-based Meta-cognitive learning Radial Basis Function Network (PBL-

McRBF) [116] as well as an SVM classifier from the LIBSVM package [117]. The meta-cognitive learning model has been inspired from human learning studies, where aspects of learning such as *what to learn*, *when to learn* and *how to learn* are regulated by strategies [118, 119]. The PBL-McRBFN consists of a cognitive component that approximates the optimum network weights and a meta-cognitive component that controls the behaviour of the cognitive component. In the meta-cognitive component, various learning strategies, such as sample deletion, neuron growth, neuron pruning, parameter update and sample reserve are employed. The choice of the appropriate learning strategy of a particular sample, is made based on several knowledge measures derived from known information regarding the sample. The meta-cognitive component dynamically models the aspects of the cognitive component such as the number and center of neurons. The cognitive component contains a single hidden layer employing Gaussian activation function, that is combined linearly with an output layer. In the cognitive component, a projection based learning algorithm is employed to analytically obtain the weight matrix. In a prior work, the PBL-McRBFN was applied to problems such as Alzheimer’s disease (AD) detection from MRI scans. Details of the PBL-McRBFN such as the architecture and projection based learning algorithm, information regarding the learning strategies and results of application to AD detection, are available in [116].

HiDEN was compared with the PBL-McRBFN and SVM classifiers on the TM dataset of proteins. All training and testing was conducted on three machines each with 16G of RAM. Depending on the machine used the processor speed was 3.6 GHz or 3.2 GHz each. Both the PBL-McRBFN and SVM classifiers were trained on 12,000 residues which comprise approximately 12% of the residues in the TM dataset. These 12,000 residues comprise 37 proteins and 276 proteins were left for testing. In the case of HiDEN 75% of residues were used for training and the remainder 25% used for testing. After testing was completed for the PBL-McRBFN and SVM classifiers, the same sequence fitting block algorithm described under HiDEN was applied to their respective predictions in order to obtain the topology score. The performance was

computed according to the overall accuracy, η_o and topology score, η_T which had been defined in section 3.6.

In the case of HiDEN since the model was also developed to predict other protein groups, any TM proteins predicted to belong to other groups were first assumed as correct predictions. Next, the individual residue predictions are measured using the overall accuracy η_o . Although the group prediction may be wrong, many individual residues within the sequence may be predicted correctly and it allows a residue-wise assessment to be made. The number of proteins with correct topology is then determined (described in section 3.6 and equation 3.3).

Table 3.3 shows overall train and test results obtained on the whole TM dataset which contains 313 proteins comprising 99,404 residues. The table indicates that HiDEN obtained the highest single residue accuracy ($\eta_o = 95.45\%$) followed by the PBL-McRBFN ($\eta_o = 78.88\%$). The SVM classifier had obtained the lowest accuracy at 77.1%. HiDEN also obtained the highest topology score ($\eta_T = 94.89\%$) followed by the SVM and PBL-McRBFN classifiers, which had obtained below 70%.

HiDEN obtains the correct topology for 86 more proteins than the SVM classifier and 91 more correct topologies than the PBL-McRBF Network, respectively. These results were from training the SVM classifier and PBL-McRBFN on 12,000 residues, while HiDEN had been trained on approximately 74,500 residues which is the most likely reason for the higher performance. However, as deep learning networks are suited for extracting information in layer wise learning, they usually need large numbers of samples to detect the correct parameters before convergence. Hence it was difficult achieve convergence in HiDEN using small numbers of samples. Using too many samples may result in over fitting for SVM and PBL-McRBFN. In a ten-fold cross-validated approach containing approximately 89,460 training residues per partition, HiDEN obtained a topology score of about 63% which is similar to those reported for the PBL-McRBFN and SVM here. The ten-fold cross-validated results on the TM dataset are discussed in detail in section 3.8.1.

As the SVM classifier tested was not designed with meta-cognitive learning principles, a larger number of training samples may improve performance. To check this,

Table 3.3: HiDEN Comparison with other Classifiers. Performance on TM dataset containing 313 proteins is shown. Bold font indicates highest score in the column. HN or SV denote number of Hidden Neurons or Support Vectors, respectively.

Classifier	HN or SV	η_o (%)	η_T (%)
PBL-McRBFN	3800	78.88	65.81(206/313)
SVM	8346	77.1	67.41 (211/313)
HiDEN	1290	95.45	94.89 (297/313)

four-fold cross-validation experiments were also performed on the TM dataset using the SVM classifier. 25% of proteins were randomly selected as training sequences and the remainder used for testing. On average, the number of training residues in a partition was approximately 24,850. Under these conditions, the SVM classifier achieved an averaged four-fold overall accuracy (η_o) of 77.38% with a standard deviation (s.d.) of 0.62. While $\eta_o = 77.38\%$ is 0.28% higher than obtained with training with 12,000 residues, the PBL-McRBF network performance remains higher ($\eta_o = 78.88\%$).

The ratio of hidden neurons (or support vectors) to training samples is 0.017, 0.32 and 0.7 for HiDEN, PBL-McRBFN and SVM classifiers respectively, indicating a better generalization performance for HiDEN compared to other models. Although HiDEN was trained on a larger number of samples, the model needs much fewer hidden neurons than PBL-McRBFN or the number of support vectors employed by SVM.

The next section investigates HiDEN performance further by considering the ability to detect protein groups and analyses the performance on all four TOPCONS datasets. The ten-fold cross-validated performance on the TM dataset is also discussed with respect to sequence similarities in the datasets.

3.8 HiDEN Performance Discussion

The group classification accuracy indicates the ability of the network to distinguish between the four types of proteins in the datasets which were the TM proteins, TM proteins with an N-terminal signal peptide, Globular proteins and Globular proteins with signal peptides. It can be applied to bacterial proteome identification studies where it is used to detect information such as predicted membrane proteins to be potentially sent for drug target screens.

The results in this section are reported on the entire dataset (including train and test residues) and compared with the TOPCONS method [32].

The group classification confusion matrix is given in Table 3.4. The percentage of correctly classified proteins for the TM and SP+TM datasets are 97.76% and 96.28% respectively while for the globular and Globular+SP datasets it is 96.72% and 95.67%, respectively.

Table 3.4: Confusion matrix for Protein Group classification.

Data set	TM	SP+TM	Globular	Globular+SP
TM	97.76%	1.92%	0.32%	0%
SP+TM	3.18%	96.28%	0.27%	0.27%
Globular	1.59%	0.08%	96.72%	1.61%
SP	0.14%	3.23%	0.96%	95.67%

A majority of the mis-classified proteins from the TM dataset had been predicted as SP+TM sequences (1.92%). Likewise a majority of the SP+TM mis-classifications (3.18%) had been predicted as TM proteins, with the N-terminal signal peptides being mistaken as Inside (I) or Outside (O) regions in a majority of the wrongly predicted cases. While it can be expected that N-terminal helices are highly hydrophobic this is not always true and it is possible to have mildly hydrophobic N-terminal helices, as has been discussed in the literature regarding the difficulty of correctly distinguishing membrane helices and other hydrophobic regions [29, 67]. A mildly hydrophobic TM helix can be mistaken as Inside or Outside regions. For the Globular group of proteins a majority of mis-classifications were evenly spread into TM (1.59%) or Globular+SP (1.61%) groups, indicating the propensity of Inside residues to be confused for membrane structures or signal peptides respectively. A majority of the wrong predictions in the Globular+SP dataset fell into the SP+TM group (3.23%). These mis-predictions arose from Outside residues in the Globular+SP dataset that had been wrongly predicted as membrane regions (M). Overall a total of 6626 of 6856 sequences had correct group predictions resulting in an average correct classification of 96.65% compared with 92.25% reported by TOPCONS for the same four datasets [32].

The topology score is calculated according to the performance measures described in Section 3.6. The performance of six topology prediction methods on these same datasets are also obtained from a recent study [32] and compared with HiDEN, in order of highest to lowest overall topology scores obtained. The topology scores are given in Table 3.5. All methods considered obtained higher topology scores for non-membrane proteins relative to their own performance on the membrane proteins.

The results indicate that HiDEN gives the highest scores, with an overall topology score of 95% averaged for the four protein groups considered. The second highest overall topology is obtained by TOPCONS at 87%. The individual topology scores for each of the datasets obtained by HiDEN are 95% (297 of 313 proteins) and 92% (691 of 752 proteins) from the TM and SP+TM datasets, respectively. The topology score for Globular and Globular+SP groups are 97% and 96%, respectively. HiDEN performance was 15% higher on the TM dataset compared to the next best predictor, TOPCONS which reported a score of 80%. For the SP+TM dataset, HiDEN performance was 9% higher compared to 83% obtained by Phobius, the second best method for this dataset. In the case of the Globular and Globular+SP datasets, HiDEN performance was either on par with TOPCONS (97%), or higher by 5%, respectively.

Table 3.5: HiDEN Topology Prediction compared with Reported Methods, obtained from TOPCONS [32]

Method	MSA	TM (%)	SP+TM (%)	Globular (%)	Globular+SP (%)
HiDEN	+	95	92	97	96
TOPCONS	+	80	80	97	91
Philius	-	70	75	94	94
Phobius	-	55	83	95	94
PolyPhobius	+	68	64	95	85
SPOCTOPUS	+	71	78	78	79
MEMSAT-SVM	+	67	52	88	0

It has been a challenge in the literature for a single predictor to obtain high accuracies for both signal peptide prediction and transmembrane protein prediction. TOPCONS reported high accuracies of up to 80% on both the TM and SP+TM

datasets. However the method employed a topology profile that consisted of the combined output obtained from five other predictors. The next highest accuracy obtained by a single predictor (after HiDEN) is by SPOCTOPUS [74], with a topology score of 71% on the TM dataset and Phobius [109] with a score of 83% on the SP+TM dataset, respectively. Besides methods employing Multiple Sequence Alignments (MSA) based techniques or PSSM, predictors employing a single sequence directly such as Philius [75], had been considered. Nevertheless, the topology scores obtained remained below 80%, with TM and SP+TM datasets scores at 70% and 75%, respectively.

3.8.1 Sequence similarities and Cross-Validation

Ten-fold cross-validation was also performed using HiDEN. A topology score of 63% (198/313) was obtained on the TM dataset, averaged for the ten testing partitions. The topology score obtained on the training partitions was much higher, with the ten-fold average being 99%. Upon noting the cross-validated topology score of HiDEN at 63% (for testing only), it seems significantly lower than the overall train and test score of 95% (reported in Table 3.5). Other methods in the literature had reported scores of 80% (TOPCONS) or near 70% (Philius, PolyPhobius and SPOCTOPUS). However, the higher performance reported by other methods may not be derived from stringent criteria and some caution may be needed in interpretation of the reported scores as explained further.

Previous authors had undertaken efforts to ensure a homology reduction of 30% in the TOPCONS datasets by clustering and subsequent removal of sequence similarities [32]. Theoretically, after the homology reduction, the maximal sequence identity (or similarity) between any two sequences in the combined datasets should not exceed 30%. However, here the findings were different. Sequences from the TM dataset were clustered using the CD-HIT program [120] at an identity threshold of 30% and the number of similar sequences counted. The TM dataset had 23 similar sequences at the 30% or higher identity level, of which 7 of them had 95% or higher identities.

Similarly, the SP+TM, Globular and Globular+SP datasets also had many similar proteins.

Further, the original predictor applied to the TOPCONS datasets [32] was a consensus method built from the topological profiles of five individual predictors (OC-TOPUS [9], PolyPhobius [71], SPOCTOPUS [74], Philius [75] and SCAMPI [76]) and hence had also not performed cross-validation. Four of the five individual methods [71, 74–76] had been trained on comprehensive datasets initially developed in 2004, known as the Phobius datasets [79]. By clustering similar proteins from the TOPCONS and Phobius datasets using CD-HIT, large numbers of similar proteins between the two datasets were detected. For instance, at an identity level of 95% or higher, the Phobius datasets contained 43 similar proteins to the TM dataset, 14 similar proteins to the SP+TM dataset, 432 similar proteins to the Globular+SP dataset and 54 similar proteins to the Globular dataset. Also, the fifth individual predictor [9] had been trained on a separate dataset that contained 54 proteins with 95% or higher sequence identities to the TM dataset.

If only a strictly cross-validated score was considered for HiDEN the topology score is much lower as compared to other predictors. However, this would be unfair to HiDEN as other predictors contain many similar (or even identical proteins) to the TOPCONS datasets as explained. Hence, the thesis has presented discussions on both the train and test overall, as well as cross-validated scores.

Recently a first-principle based method SCAMPI 2 was also developed that employs the terminal hydrophobic helices rule to predict topology of TM proteins, without considering signal peptides [67]. The method was applied on a dataset of 285 transmembrane proteins, achieving a topology score of 73%.

As discussed in the literature, the incorporation of homology information in employing multiple sequence alignments greatly improves the topology prediction [74, 110]. HiDEN also relies on information contained in the PSSM and extends it in two ways. Firstly, the unsupervised training step of the contrastive divergence learning algorithm employed in the Deep Belief Networks extracts higher level features from the

input PSSM itself. Secondly as an ensemble of DBN is employed, various DBN are able to learn different aspects of the underlying correlations shown by neighbouring residues within the window. The combined ensemble output, such as by the GDEN or LDENs, is able to achieve higher accuracies than a single DBN could give.

Comprising one of the components of correct topology, the detection of the number and location of transmembrane helices is an aspect of biological significance. Wrong predictions here may provide some clues to obtain a better topology prediction model or to establish what correct topology predictions mean in a biological sense. Specifically, the next section presents a simple analysis based on the hydrophobicity of the predicted TM helices in the TM and SP+TM datasets. The predicted TMH from the overall results (including train and test proteins) are compared and discussed.

3.9 Effectiveness of TM Helix Prediction

Previously, the HiDEN performance and topology score were discussed and compared to existing predictors in the literature as well as with other types of classifiers. The prediction of the correct number and location of transmembrane helices, or M residues, is important in obtaining a correct topology. In this section, the results of HiDEN are evaluated from a biological perspective; the predicted TMH are analysed with respect to their hydrophobicity and compared with those of TOPCONS. A brief explanation of the hydrophobicity of a TM helix is first given to provide a biological background. The hydrophobicity of a TM helix is an important property that has been used by first principle methods to predict membrane spanning regions [31,67,121]. TM helices are often highly hydrophobic, with the energy of membrane insertion, ΔG typically peaking at -2 to -0.3 kcal/mol [67]. However, it is difficult to implement a ΔG cut-off value that enables a clear separation of TM helices and other hydrophobic sequence segments such as signal peptides; weakly hydrophobic helices may be wrongly predicted as Inside or Outside regions. One factor for the performance shown by HiDEN is the ability to better detect TM helices compared to other methods. It is shown that HiDEN correctly identifies less hydrophobic helices compared to TOPCONS for the TM and SP+TM datasets by comparing the hydropathy values of predicted helices

of the two methods.

To quantify the hydrophobicity of a given TM helix, the Grand Averaged Hydropathy (GRAVY) index is used [48]. The GRAVY index calculates a value for a given protein sequence based on averaging the hydropathy values of individual amino acid residues within windows of the sequence. The hydropathy values are obtained from a pre-determined scale that assigns values to each of the 20 possible amino acid based on their hydrophobic or hydrophilic characteristics. The values of the scale range from the most hydrophobic residue, Isoleucine (+4.5) to the least hydrophobic, Arginine (-4.5). To obtain the GRAVY index of a protein sequence, the average hydropathic value of amino acids in the sequence is obtained.

To calculate the GRAVY values of the TM regions, in this work the ProtParam server, was used [122]. TM helical regions (labelled as M residues) from the TM and SP+TM datasets are sent to the ProtParam server and their GRAVY values were retrieved. Next, the number of correctly predicted TM regions is determined.

To determine the number of correctly determined TM helices, a correct TM helix prediction must first be defined. However, there exists no standard definition in the literature of a correctly predicted TM helix. Previous definitions include MemBrain, which considered nine overlapping residues between an observed and predicted helix as correct [39], as well as Philius which considered an overlap of 5 residues to be correct [75]. In the definition of a correct topology, a minimum overlap of 5 residues is also followed, as described in section 3.6. Therefore in the following discussion, a TM helix is determined to be predicted correctly if there is an overlap of at least 5 residues in the prediction.

An error at the group level prediction from the GDEN described in section 3.4 may result in some TM or SP+TM proteins being discarded although the method may have predicted some helices correctly. For instance a TM protein predicted as SP+TM could still have correctly detected TM helices, despite erroneously having had a signal

peptide predicted. To purely assess the ability to detect helices, in the following discussion the group prediction results are ignored and only the final topology prediction is considered. It is assumed that a given sequence is already correctly predicted as a TM or SP+TM protein. A similar procedure discounting the group predictions is allowed for TOPCONS to perform a fair comparison. As explained in section 3.8, there exists a chance that underlying individual predictors forming the TOPCONS profiles have been trained on similar sequences to the transmembrane datasets. Here the transmembrane predictions of HiDEN are obtained from the overall prediction results (of which 95% was obtained on the TM dataset and 92% on the SP+TM dataset). Given an observed TM helix, it is considered as having been predicted correctly as long as there is a predicted TM helix with at least 5 residues of overlap. Each predicted helix is only counted once towards a correct prediction

The TM and SP+TM datasets contain 1761 and 1442 TM helices, respectively. Table 3.6 reports the average GRAVY values of TM helices and their standard deviation (s.d.). TM helices in these datasets are grouped into four categories based on HiDEN and TOPCONS prediction results. These are the helices that (i) both TOPCONS and HiDEN predict correctly (Both Correct), (ii) only HiDEN predicts correctly (HiDEN only), (iii) only TOPCONS predicts correctly (TOPCONS Only) and (iv) neither method is correct (Neither Correct). From table 3.6, it is observed that 97.84% of helices in the TM dataset and 94.73% from the SP+TM dataset are predicted correctly by both methods. However, HiDEN Only obtains 29 correct helices from the TM dataset as well as 56 helices in the SP+TM dataset that TOPCONS misses.

To detect if there is a significant difference in the average hydropathy values between helices in the Both Correct and HiDEN only groups, a two sample *t*-test was performed assuming equal variances. If there is no significant difference, it may be concluded that the ability of HiDEN and TOPCONS to detect TM helices has little to do with their Hydropathy values. If there is a significant difference, since HiDEN

Table 3.6: GRAVY values of Transmembrane Helices. Average Hydropathy Values (GRAVY) of TM Helices predicted by HiDEN and TOPCONS. GRAVY values of TM helices are averaged and the standard deviation reported in braces. The helices are grouped into 4 categories. Both correct - Helices that both HiDEN and TOPCONS predict correctly. HiDEN Only: HiDEN predicts correctly and TOPCONS, wrongly. TOPCONS Only: Helices that TOPCONS predicts correctly and HiDEN, wrongly. Neither Correct: Helices that both methods fail to predict. The number of helices falling into each type of category and the total number in each dataset is indicated.

Dataset	Both Correct	HiDEN Only	TOPCONS Only	Neither Correct
TM	1.69 (0.53)	1.23 (0.57)	1.38 (0.96)	1.75 (0.82)
Total TMH = 1761	1723	29	6	3
SP+TM	2.12 (0.58)	1.46 (0.95)	1.81 (0.57)	0.15 (1.31)
Total TMH = 1442	1366	56	14	6

Only group has GRAVY values that are lower than that of helices in the Both Correct group, it indicates that HiDEN can detect the weakly or mildly hydrophobic helices effectively compared to TOPCONS.

In the case of the TM dataset, for the helices in the Both Correct group, the average hydropathy value was 1.69 and for the HiDEN only group, it was 1.23. At a significance level of $\alpha = 0.05$, the p -value was 4.5×10^{-6} , with a t -statistic of 4.60 and d.f. = 1750, indicating that the hydropathy values of the two groups are significantly different. A similar t -test was also conducted for the SP+TM dataset. The p -value obtained was 1.7×10^{-17} , with a t -statistic of 8.05 (d.f. = 1420). The findings indicate that the average hydropathy values of helices from both categories are significantly different.

In contrast, when considering the helices that TOPCONS obtained correctly but HiDEN failed to detect (TOPCONS only), it is difficult to attribute the reason to the hydropathy values. A similar t -test conducted comparing GRAVY values of (Both correct) and (TOPCONS only) gave p -values of 0.16 and 0.043 for the TM and SP+TM datasets, respectively. At $\alpha = 0.05$, there is no significant difference in case of the TM dataset while for the SP+TM dataset, there is a weakly significant difference. As the numbers of sample sizes are small (6 and 14 for TM and SP+TM, respectively), it is difficult to draw a definite conclusion.

Hence, both HiDEN and TOPCONS are capable of detecting highly hydropho-

bic TM helices (Both Correct : GRAVY values of 1.69 and 2.12 for the TM and SP+TM datasets, respectively). However, HiDEN detects moderately hydrophobic helices better (HiDEN Only: GRAVY values 1.23 vs. 1.46 for TM and SP+TM datasets, respectively). The findings indicate that HiDEN has the ability to detect mildly hydrophobic helices correctly. The next section concludes this chapter by summarizing these and earlier findings.

3.10 Summary

In this chapter, a Hierarchical Deep Ensemble (HiDEN) model was developed and applied to the problem of predicting residue preferences in membrane protein sequences and to determine their correct topologies. An input sequence was represented using a Position Specific Scoring Matrix (PSSM) read in a sliding window to generate the features. Besides membrane proteins (TM and TM with signal peptides (SP+TM)), the HiDEN model was also applied to the detection of Globular (G) and Globular proteins containing signal peptides (G+SP). In total, four protein group classifications were tested. HiDEN consists of a two level, hierarchical approach that combines a Global Deep Ensemble Network (GDEN) applied to detecting the four protein groups, followed by the second level LDEN to predict membrane protein residues and to determine their correct topology in a sequence fitting block.

HiDEN obtained a ten-fold cross-validated topology score of 63% on the TM datasets. If the train and test proteins were included the performance of HiDEN was much higher than other methods for the membrane datasets while for the non-membrane datasets the performance was on par, or slightly higher. Additionally, HiDEN performance was also compared to an existing Projection based learning, Meta-cognitive Radial Basis Function Network (PBL-McRBFN) and an SVM classifier for the TM dataset. The results indicated that HiDEN obtained a much better generalization needing fewer neurons (or support vectors) than the former two methods, although the PBL-McRBFN performance based on the single residue accuracy was better than that of SVM.

In HiDEN an unsupervised training step that employs a contrastive divergence learning algorithm was applied to extract higher level representation of the input PSSM. Additionally an ensemble model was employed in HiDEN to strengthen predictions by individual Deep Belief Nets (DBN). The current testing topology score at 63% (with training topology score of 99%) indicates that the peak in training accuracy had already been reached with the current features. However the model may need more information to have a higher result on the test sequences. The other existing predictors in the literature and classifiers compared also used the PSSM directly, which is not a precise representation of the sequence in the feature space. This is due to the variety of possible sequences that can adopt a particular structure, contributing to overlapping values in the feature space that make it difficult to discriminate between different classes.

Finally, evidence was presented for the biological significance of the HiDEN predictions. It was shown that while both methods were able to detect a large proportion of highly hydrophobic TM helices accurately (95% or more), HiDEN was able to correctly predict TM helices that were mildly hydrophobic compared to TOPCONS. The findings indicate that the higher level abstraction of the input features generated in the unsupervised training stage of the DBN, were able to sufficiently capture underlying correlations between features.

Chapter 4

Protein Secondary Structure

Prediction using Deep Belief

Networks (DBN)

4.1 Introduction

In chapter 3 Deep Belief Networks (DBN) were applied to predict membrane topology and to detect signal peptides as well as globular proteins. In this chapter a DBN is used to predict protein secondary structures. DBN comprise of multiple stacked Restricted Boltzmann Machines (RBM) that learn from the input features in a layer-wise manner [105]. In the literature, the contrastive divergence (CD) learning algorithm has been used to capture the underlying abstractions given a set of features [40]. Here the DBN are used together with the CD learning algorithm to detect underlying long range residue interactions given a set of features derived from the PSSM. Unlike the Hierarchical Deep Ensemble Neural Network (HiDEN) approach presented previously, here a single Deep Belief Network (DBN) is first used to predict protein secondary structures. The DBN is trained by a random selection of 80% of proteins in the CB513 dataset [38] and tested on the remaining 20%. The CB513 dataset is briefly described below.

4.2 Dataset

The CB513 dataset contains protein chains labelled with secondary structures (SS) such as Helix (H), Sheet (E) or Coil (C), corresponding to numeric class labels 1, 2 and 3, respectively. The assignment of SS from structure files had been done using the Define Secondary Structure of Proteins (DSSP) program [23]. The homology reduced version of the CB513 dataset containing 385 proteins is divided into train and test sets, forming 80% and 20% of proteins, respectively. The 308 training proteins are used for model development and contain 50,712 residues, while the remainder 77 proteins comprising 12,367 residues serve as testing proteins. Further details of the CB513 dataset such as descriptions of homology removal with respect to sequence and structural similarity are given in 5.2.

4.3 PSSM based Feature Encoding and Performance Measures

The feature encoding uses the Position Specific Scoring Matrix (PSSM) derived from the PSI-BLAST algorithm [50] that has been previously described in chapter 3. Similar to the generation of PSSM for membrane protein sequences given in section 3.3, the sequences from the CB513 datasets are queried against the *RefSeq* [12] database using PSI-BLAST to generate PSSM [50]. To avoid hits to low complexity regions, transmembrane and coiled coil regions, the *RefSeq* database was first filtered with the *pfilt* program ¹ available from the PSIPRED [91] server to predict protein secondary structures. The PSI-BLAST program parameters were set to yes for SEG filtering, with the number of iterations set to three and had E-value cut offs at 10^{-5} for all three datasets. All other values were set to the program default (substitution matrix = BLOSUM62) However, about 15 sequences had no hits detected under these settings; for these sequences, the E-value was increased to a higher limit of 100. In the case of CB513 dataset proteins that had short sequences (e.g. less than 100 residues long), other substitution matrices (e.g. PAM70) were used. In the case of protein

¹<http://www.cs.ucl.ac.uk/fileadmin/bioinf/pfilt/downloads/pfilt.c>

secondary structure prediction the window size was set as $w = 21$ rather than $w = 31$ used for membrane protein structures. To model the secondary structures, various window sizes have been reported in the literature, such as PSIPRED [37] $w = 17$ and JNet [92], $w = 19$. Further SPINE [4] reported that there was a negligible effect on the Q_3 for a range of window sizes between 15 and 25 with a weak peak suggested at $w = 21$. To obtain the optimal window size, more experiments need to be conducted on the CB513 dataset in this study. Here, the study is conducted for a window size of $w = 21$.

Next the performance measures used in this chapter are described. The overall accuracy (η_o) is the percentage of correctly classified residues over the total number of residues and is also known as the Q_3 measure. The class-wise accuracy (η_j) is the percentage of correctly classified residues in class j , where $j \in \{H, E, C\}$. The average class accuracy (η_a) is the average of the class-wise accuracy (η_j). A complete definition of these performance measures is available in section 5.4.

4.4 Architecture and Learning Algorithm of Deep Belief Network

The section describes the architecture of a Deep Belief Network containing a single hidden layer. Similar to the first Deep Belief Networks (DBN₁) within a Deep Ensemble Network (LDEN) described in chapter 3, the DBN comprises of one or more Restricted Boltzmann Machines (RBM) stacked together in a layer-wise manner. Figure 4.1 shows the simple architecture of a DBN containing a single RBM, used to predict protein secondary structures. The input layer consists of 420 input neurons for single residue t , $\mathbf{x}^t \in \mathfrak{R}^N$, where N is the number of features. The features were obtained by reading the $M \times 20$ PSSM in a sliding window of 21 residues, where $N = 20 \times 21 = 420$. The input layer presents the features to the hidden layer neurons which employ a sigmoidal activation function. The third and last layer consists of three output neurons employing the linear activation function. According to the output, a residue is classified into one of the three secondary structures H, E or C,

Figure 4.1: Architecture of Deep Belief Network (DBN). It consists of three layers. The first layer consists of N input neurons, the second layer contains K hidden neurons and the third layer contains n output neurons. For SS prediction, the number of input neurons, $N = 420$, which take in the PSSM based sliding window features as input. The input was normalized to lie in $[0, 1]$. The hidden layer employs the sigmoidal activation function to calculate the hidden response, h_i^t . Finally, the output layer employs a linear function to obtain the predicted output, $\hat{y}_i^t$. In the output layer, $n = 3$ corresponding to the secondary structures to be predicted, H, E or C. w_n^K represents the weight connecting the K^{th} hidden neuron to the n^{th} output neuron. The diagram was modified from Figure 3.3.

represented by $c^t \in \{1, 2, 3\}$. The coded class labels $\mathbf{y}^t$ are generated as follows:

$$y_j^t = \begin{cases} 1 & \text{if } c^t = j \\ -1 & \text{otherwise} \end{cases} \quad j = 1, \dots, q \quad (4.1)$$

In a restricted Boltzmann machine the encoded sequences are represented as visible units $\mathbf{x}$ in the input layer. The hidden hidden unit layers are represented as $\mathbf{h}$. The joint configuration is described by the energy function given in equation 4.2.

$$E(\mathbf{x}, \mathbf{h}) = - \sum_k \sum_j h_k x_j W_{kj} - \sum_j c_j x_j - \sum_k b_k h_k \quad (4.2)$$

$$E(\mathbf{x}, \mathbf{h}) = -\mathbf{h}^T \mathbf{W} \mathbf{x} - \mathbf{c}^T \mathbf{x} - \mathbf{b}^T \mathbf{h} \quad (4.3)$$

where x_j, h_k are the binary states of visible unit j and hidden unit k , respectively. W_{kj} represents the weight connecting hidden unit k to visible unit j while b_k and c_j are their respective biases. Equation 4.3 gives the matrix equivalent of 4.2. The joint probability $p(\mathbf{x}, \mathbf{h})$ of every possible pair of values for the visible and hidden units is given by equation 4.4, where Z is the partition function. However the partition

function Z makes it intractable and a gradient descent approach is used to minimize the negative log likelihood error function.

$$p(\mathbf{x}, \mathbf{h}) = \frac{\exp(-E(\mathbf{x}, \mathbf{h}))}{Z} \quad (4.4)$$

The negative log-likelihood error is given in Equation 4.5 where N is the number of training samples. Estimators that minimize the negative log-likelihood may be derived with respect to network parameters such as the weights or biases. However obtaining the partial derivative of equation 4.5 with respect to network parameters will result in intractable terms in the derivative. The CD algorithm employed Gibbs sampling to obtain point estimates of the intractable term.

$$\frac{1}{N} \sum_{t=1}^N -\log p(\mathbf{x}^{(t)}) \quad (4.5)$$

To train an RBM the average negative log-likelihood is minimized using the stochastic gradient descent approach. The partial derivative of 4.5 with respect to weights W_{kj} is given. For example, for network parameter $\theta = W_{kj}$, the partial derivative of the negative log-likelihood function with respect to W_{kj} is given as equation 4.6

$$\frac{\partial}{\partial W_{kj}}(-\log p(\mathbf{x}^{(t)})) = \left\langle \frac{\partial E(\mathbf{x}^{(t)}, \mathbf{h})}{\partial W_{kj}} \middle| \mathbf{x}^{(t)} \right\rangle_h - \left\langle \frac{\partial E(\mathbf{x}, \mathbf{h})}{\partial W_{kj}} \right\rangle_{x,h} \quad (4.6)$$

$$\mathbf{W} = \mathbf{W} + \alpha(\mathbf{h}(\mathbf{x}^{(t)})\mathbf{x}^{(t)T} - \mathbf{h}(\tilde{\mathbf{x}})\tilde{\mathbf{x}}^T) \quad (4.7)$$

$$\mathbf{b} = \mathbf{b} + \alpha(\mathbf{h}(\mathbf{x}^{(t)}) - \mathbf{h}(\tilde{\mathbf{x}})) \quad (4.8)$$

$$\mathbf{c} = \mathbf{c} + \alpha(\mathbf{x}^{(t)} - \tilde{\mathbf{x}}) \quad (4.9)$$

The CD algorithm is applied during the unsupervised training to reconstruct the input $\mathbf{x}^{(t)}$ by minimizing the divergence between the observed data and (Gibbs) sampled data. The negative phase of the stochastic gradient equation 4.6 that takes the expectation over x, h is intractable. In order to calculate $-\left\langle \frac{\partial E(\mathbf{x}, \mathbf{h})}{\partial W_{kj}} \right\rangle_{x,h}$ the CD algorithm estimates it by replacing $\frac{\partial E(\mathbf{x}, \mathbf{h})}{\partial W_{kj}}$ with a Gibbs sampled point estimate, $\frac{\partial E(\tilde{\mathbf{x}}, \tilde{\mathbf{h}})}{\partial W_{kj}}$. $\tilde{\mathbf{x}}$ is obtained as follows.

1. For each training sample $\mathbf{x}^{(t)}$
 - (a) Perform k steps of Gibbs sampling to obtain $\tilde{\mathbf{x}}$, with $\mathbf{x}^{(t)}$ as initial point
 - (b) Update parameters using equations 4.7, 4.8, 4.9
2. Repeat for i epochs

From 4.6 it may be shown that the weight update rule may be derived to obtain equation 4.7, where α denotes the learning rate. A similar process may be used to derive the update rule for the bias terms $\mathbf{b}$ and $\mathbf{c}$, given in equations 4.8 and 4.9 respectively. The algorithm description and equations were based on Dr. Hugo Larochelle’s video lectures [123–125] and Geoffrey Hinton’s paper [40]. Details of the CD algorithm and derivation of the weight update rules may also be found therein. To measure the error between the actual output y^t and predicted output $\hat{y}^t$, the hinge loss error [111] was used to enable faster convergence of the network. The DBN was trained with 80% of the proteins in the CB513 dataset and tested on 20%. In the unsupervised training stage, the momentum was set to 0.05 and the learning rate $\alpha = 0.1$. During the supervised training the learning rate and α were set to 10^{-6} and 5×10^{-5} respectively. Section 4.5 presents the performance of the network.

4.5 Performance of Deep Belief Network

The performance studies reported in this section were carried out on MATLAB version R2012b on Windows 7 running on a machine with 8G RAM and 3.60 GHz processor speed. Table 4.1 presents the training and testing accuracies of the DBN on the CB513 dataset. The network obtained a training accuracy of $Q_3 = 79.37\%$ and a test accuracy of 72.56%. In comparing the testing class accuracies, Helix was the highest with $\eta_H = 77.74\%$ followed by Coil at 75.43% and lastly Sheet at 61.04% (similar to observations for other predictors).

Five-fold cross-validation had also been performed on the CB513 dataset using the DBN described in this chapter, employing 3000 neurons in the hidden layer. The results were a Q_3 accuracy of 72.80% averaged over the testing partitions with similar

confusion matrices to the single partition results presented in Table 4.1. For comparison, previous methods (not based on deep networks) had reported cross-validated Q_3 accuracies of 73.5% [85], 76.4% [92] and 80% [86]. In general most methods report Q_3 scores of about 80%. However, a compound pyramid model reported a significantly higher Q_3 of 85.6% [126], although the higher accuracy was at the cost of more computational complexity. Further detailed discussions of the performance of these predictors (that were not based on deep learning models) are presented in Table 5.4. While the training accuracy of the DBN ($Q_3 = 79.37\%$) is comparable to other methods, the test accuracy ($Q_3 = 72.56\%$) is much lower and needs improvement. Typically the number of neurons needed by a single hidden layer is fewer than the input neurons, which allows the contrastive divergence learning algorithm to extract the higher level features from the raw PSSM input. However, 2500 hidden neurons have been employed which suggests a challenging dataset from which it is difficult to extract underlying structural correlations. Even increasing the number of hidden neurons to 3000 did not give a significantly higher Q_3 score.

Various DBN architectures had also been attempted to predict the secondary structures in this chapter. For example, up to six hidden layers containing different numbers of hidden neurons per layer such as [400:380:400:380:400:380] were attempted (a colon is a layer separator). Another instance of the DBN with two hidden layers [800:50] was also attempted. Besides the single DBN presented in Table 4.2, a hierarchical approach consisting of a second level structure to structure network was also attempted here, similar to other protein secondary structure prediction works [4, 36, 37, 94]. The predicted values for the three output nodes ($\hat{y}_l^t$, where $l \in \{1, 2, 3\}$), were read in a window of 15 residues, giving a second level feature matrix of size $M \times (15 * 3)$, where M is the number of residues in a sequence. The $M \times 45$ feature matrix was then presented as input to a second level structure to structure prediction DBN. However, the test accuracy remained below 70% for these and similar architectures attempted. Interestingly it was found that adding more hidden layers seemed to decrease the performance on this dataset. The results of the single DBN applied here indicate that approximately $(2500/420 \approx 6)$ hidden neurons were needed to adequately recon-

struct a sample that is similar to the input of a single neuron. Hence, it is suggested that mapping the input neurons to higher dimensions may be more effective on this dataset.

Table 4.1: DBN Performance on PSSM derived features. Performance of a Deep Belief Network (DBN) using PSSM based encoding. The DBN comprises a single hidden layer with 2500 neurons. No. of train and test proteins are given in brackets.

	observed $j \rightarrow$	H	E	C	η_j	Q_3	η_a
	predicted $j \downarrow$						
Train (308)	H	14631	617	2447	82.69		
	E	682	8075	2635	70.89	79.37	78.24
	C	2144	1940	17541	81.12		
Test (77)	H	3170	227	681	77.74		
	E	328	1903	887	61.04	72.56	71.41
	C	659	612	3900	75.43		

4.5.1 Motivation for using CABS energy potential based Feature Encoding with DBN

This subsection compares the performance of the DBN proposed here with other deep learning networks in the literature. Some limitations of the current approaches are identified, motivating the need for an energy potential based feature encoding.

Table 4.2 presents two deep network based models developed in 2015 and compares the performance with the DBN applied in this chapter. The Q_3 of 72.8% (referring to results of five-fold cross-validation) obtained by the DBN is lower than the other two methods which reported nearly 81%. It was found that despite changing various factors such as the number of hidden layers, adopting a second level structure to structure network or changing the parameters (number of training epochs, window and batch size) the Q_3 accuracy for test proteins remained low (around 63% or lower). These are very different findings from other works like SPIDER 2 [27] or DNSS [46] which had implemented hierarchical deep belief networks and obtained Q_3 accuracies of nearly 81%.

Table 4.2: Comparison of Deep Networks based SS Predictors. PSSM - Position Specific Scoring Matrix, PP - Physicochemical Properties. No. of test proteins are given in round brackets.

Method	Architecture	Input Features	Q ₃ (%)
DNSS [46] (198)	Two - Tier DBN	PSSM + Atchley Factors	80.7
SPIDER 2 [27] (72)	Three Tier DBN	PSSM + 7 PP	80.8
DBN (385)	Single DBN	PSSM only	72.8

One factor for the high performance reported by SPIDER 2 and DNSS, besides hierarchical network architecture or the number of training proteins used, is the inclusion of more information in their input features in addition to the commonly used PSSM. SPIDER 2 incorporated various physicochemical and structural properties such as the solvent accessibility, volumetric contributions of amino acid residue side chains and backbone dihedral angles as part of their features to improve their predictions. DNSS employed the Atchley factors [127] that provided a numerical representation of amino acid variability in aligned sequences.

Moreover, while DNNS and SPIDER 2 had implemented hierarchical deep learning architectures to obtain predictive accuracies close to 81%, this score had already been reached by older methods which were not based on deep learning models such as PSIPRED and SPINEX in 2011 [5].

It was found that it is challenging to achieve accuracies higher than 75% based on only the PSSM, despite employing different network parameters or architectures. It was therefore desirable to explore a different type of feature representation to extract more information than is currently provided by PSSM based features alone. After a sufficiently high accuracy (more than 75%) has been obtained by a single DBN, an ensemble model (that is proposed as a future work in subsection 6.2.2) may prove much more effective to improve the accuracy further.

Since the PSSM alone is insufficient for the DBN to extract the underlying residue correlations it was proposed to use an alternative sequence representation derived from energy potentials that had previously been generated from the CB513 dataset [42]. The energy potentials were derived from the C α -C β -Side (CABS) algorithm [41]. The CABS algorithm represents protein sequences as models of atoms on a lattice. The

alignment of target sequences with template structures may be energetically computed under the model and the highest scoring templates obtained. The feature encoding process represents a single residue $\mathbf{x}^t$ as a 27-dimensional feature vector whose elements contain probabilities of the t^{th} residues and its four neighbours to the right and left adopting Helix, Sheet and Coil secondary structure states respectively. Details of the CABS algorithm based feature generation are given in chapter 5.

A DBN with a three layered architecture like the one shown in Figure 4.1 was applied to predict secondary structures given the CB513 sequences that had been represented as energy potentials. In this dataset the number of input or visible units for the DBN was $N = 27$ and the number of hidden units $K = 100$. The total number of residues was 63,079. Five-fold cross-validation was conducted in which the DBN obtained a Q_3 of 82.6% averaged over five partitions with a standard deviation (s.d.) of 0.63. The average accuracy was $\eta_a = 81.9\%$ with s.d. = 0.4. The results of the DBN using the CABS algorithm based features encoding and PSSM derived features were compared using the same 77 test proteins and presented in Table 4.3.

Table 4.3: Comparison of DBN Performance on different features. DBN performance for two types of feature encoding is shown. Confusion matrices are from 77 test proteins comprising 12,367 residues in total. Brackets indicate the number of neurons in the hidden layer of respective DBNs.

		observed $j \rightarrow$	H	E	C	η_j	η_o	η_a
		predicted $j \downarrow$						
PSSM (2500)	H		3170	227	681	77.74		
	E		328	1903	887	61.04	72.56	71.41
	C		659	612	3900	75.43		
CABS (100)	H		3678	1	399	90.2		
	E		13	2383	722	76.43	81.59	81.51
	C		561	582	4028	77.9		

The results indicate that the energy potentials based encoding of the CABS algorithm has a significantly better performance across all the measures compared. The Q_3 or η_o obtained using the CABS encoding is 81.59% which is about 9% points higher than that obtained with the PSSM derived features. The average class-wise performance was 81.51% under the CABS based encoding (10.1% more than PSSM derived features were able to provide). The highest difference is seen in the results of

the Helix (η_H) and Sheet (η_E) classes. Under the CABS based encoding $\eta_H = 90.2\%$ which is higher by nearly 12.5% compared to that of PSSM based encoding. Similarly $\eta_E = 76.43\%$ which is 15.39% greater in comparison to PSSM. In comparison to these large differences, for the Coil class the performance is rather similar. CABS based feature encoding had $\eta_C = 77.9$ compared to 75.43% obtained under the PSSM based encoding.

The results suggest that the CABS based feature encoding was able to represent the underlying correlations of neighbouring residues more effectively and by using much smaller number of features compared to the PSSM derived features. Both feature encoding processes rely on the sliding window scheme to capture interactions from neighbouring residues. Under the CABS algorithm based feature encoding process structural information from templates is not directly used. Instead, the probability of adopting the structural states H, E or C is calculated. A more detailed description is presented in section 5.3.1 and is also available from [42]. As discussed in the Introduction and previous chapters, the PSSM contains artefacts from sequence variation that may pose a challenge in the discrimination of features. Hence, the CABS algorithm based feature detection had proven to be more effective.

The improvement in DBN performance using the CABS algorithm based feature encoding (9%) over that obtained for the PSSM based encoding may be attributed to the better residue interactions captured by the former. Currently it remains challenging to raise the Q_3 accuracy further using the single DBN architecture presented in this chapter.

Choice of Complex-valued classifier:

As discussed earlier in this section, works implementing hierarchical or ensemble DBN together with PSSM based feature encoding had been reported in the literature in 2015 [27, 46]. While the Q_3 accuracies reported were around 81%, earlier methods in 2011 had already obtained up to 82% (using PSSM based features) without using deep learning networks [5]. Ideally, before the construction of an ensemble DBN the single DBN results should be sufficiently high. However in the experiments reported

here using the PSSM based feature encoding, it remained challenging obtain a test accuracy higher than 75%, despite adding more layers, or more neurons.

This motivates a search for different network models that have not been applied to this problem. Currently, all existing NN predictors are based on Real-valued neural networks. However, Complex-valued classifiers have some advantages not possessed by Real-valued networks. In particular, Complex-valued neurons have been shown to be more powerful than Real-valued neurons due to their orthogonal decision boundaries [44]. Therefore, they are efficient and need a lesser number of neurons to perform the classification task. Such a property would allow for rapid prediction of secondary structures which is advantageous to develop enhanced training models. Hence, in the next chapter it is proposed to apply a Fully Complex-valued Relaxation Network (FCRN) [47] to predict protein secondary structures. The FCRN computes its network weights analytically unlike gradient descent based approaches and has been shown to need a lesser number of neurons on Real-valued classification tasks [128]. Due to these properties it is a suitable classifier to rapidly predict protein secondary structures. Besides cross-validated studies, the FCRN is also used in a heuristics based procedure to develop an efficient training set of proteins (known as the compact model) that is shown to have a good generalization capability when tested on a blind dataset.

4.6 Summary

In this chapter a deep belief network (DBN) was applied to predict protein secondary structures from their sequences using two types of encoding - PSSM based features and energy potential based features derived from the CABS algorithm. The DBN model was trained on 80% of the dataset and tested on the remainder comprising 20%. Under the PSSM derived encoding the accuracy of the method needed significant improvement as the information from the PSSM proved challenging to extract into higher level features. The results could be significantly improved by building a deep ensemble, as well as by incorporating additional properties of interest in the prediction, such as the residue solvent accessibility (RSA), amino acid composition amongst others.

Under the energy potential based encoding derived from the CABS algorithm, a simple DBN consisting of a single hidden layer of 100 neurons obtained a better performance on all measures. As several deep learning models had recently been proposed in the literature, it was of interest to develop a model using a Complex-valued neural network with the CABS algorithm based feature encoding. The next chapter presents the use of a Fully Complex-valued Relaxation Network (FCRN) for secondary structure prediction together with the CABS algorithm based feature encoding process that had been shown to be effective. In addition, a heuristics based procedure to obtain a small training model known as the compact model, aimed at preserving the generalization ability of the classifier is proposed.

Chapter 5

Protein Secondary Structure

Prediction using a Compact Model

5.1 Introduction

As protein structure prediction is a complex problem, it is often broken into four levels which are 1-D, 2-D, 3-D and 4-D [129]. Here, the focus is on the prediction of secondary structures (SS) from an input protein sequence, regarded as a 1-D problem. This chapter investigates the prediction of protein secondary structures using a compact model together with an energy-based feature encoding known as the **C-Alpha C-Beta Side-chain** (CABS) algorithm [41]. As discussed in chapter 2, the prediction of secondary structures from protein sequences has been an important area with many works proposed to address the problem. Such works include machine-learning methods like neural networks, support vector machines as well as statistical methods such as hidden Markov models. A large scale survey that compared the performance of 12 SS predictors on nearly 2000 protein chains found that the highest three-state prediction accuracy (Q_3) for a single method was 82% [5]. According to this study, SSPRO [24] and SPINEX [95] obtained an accuracy of 82% and 80.5%, respectively. A majority voted consensus of at least three separate predictors being used, led to an accuracy of 84% [5]. More recently, the C3-SCORPION server implementing neural networks learning from context dependent potentials based on neighbour residues, achieved nearly 83% on a dataset of approximately 8000 chains [96]. In the machine

learning literature, a Compound Pyramid Model (CPM) employing neural networks in conjunction with SVM classifiers achieved an accuracy of 85.6% [3]. Besides these, a consensus based prediction implemented with SVM, also reported a Q_3 accuracy of 85.6% [21].

While advances in recent methods as described above have led to increased Q_3 accuracies, the theoretical upper limit for the problem, estimated at 88-90% [2,22] has yet to be reached. With the exception of results reported by large scale reviews such as [5,6], most of the SS predictors, report Q_3 scores averaged over k -fold cross-validation.

5.1.1 Compact Model

The traditional view in developing SS prediction methods is to expose the learning algorithm or model to as many protein chains as possible, in order to better increase the chances of predicting a new or unseen protein chain correctly. The average of cross-validated partitions is reported to prevent over-estimation of the classifier’s ability and to have a consistent measure across the entire dataset of protein chains. Following this traditional view, large datasets that contained training sequences numbering in the thousands were used. For example, SPINE used 2640 chains and reported ten-fold cross-validation results while OSS-HMM used nearly 3000 chains in four-fold cross-validation [4,84]. In comparison, smaller numbers (in hundreds) have been used for testing. Examples of train to test ratios used by various methods are 8:1 for YASPIN and 5:1 for SPINE and SSPro [98]. However, learning from a large number of chains may result in over training, leading to a loss of the classifier’s ability to generalize to new sequences.

In this study, it is proposed to detect a small number of proteins within a dataset that is able to provide a high accuracy when trained with an efficient classifier. This subset of small training proteins is termed the ‘compact model’. The compact model is a subset of training proteins within a dataset that comprises only a small portion of the dataset, yet produces accuracies that are comparable to or higher than those reported by cross-validation. It is complex to obtain the minimal number of training proteins

that give the optimum accuracy for any given dataset. To locate the minimal set, one solution is that the problem must be formulated as an objective function whose minima can be derived or approximated, a task beyond the scope of this work. Here, the objective is to find a compact model of proteins using a proposed heuristics based procedure. Ideally, the compact model should have a high generalization capability. It is difficult to quantify the generalization capability of a classifier; however, for the purpose of this work the generalization capability of a classifier (or model) is defined as its performance on a set of unseen test proteins. Given two different models developed on the same training dataset, the model with a higher predictive accuracy on an independent test dataset is deemed to have the better generalization capability.

The location of a compact model has several advantages. Firstly, the use of a small set of proteins enables quick training and allows efficient secondary structure prediction while preventing over fitting. It thus retains generalization ability to new and unseen sequences. Secondly, because a limited number of proteins will be present in the training model, it allows for easy identification of poorly performing protein folds. For example, a protein with poor performance despite being included as a training sequence identifies a fold that is difficult to model correctly using existing feature encoding techniques. Such a protein’s structure could be analysed for any chemical features that pose challenges (e.g. particular bonds that are difficult to detect.)

Besides the compact model proposed, the protein sequences in this work are represented as features using the energy potentials provided by the CABS algorithm, rather than the PSSM based encoding or MSA based features commonly used by Secondary Structure Prediction (SSP) methods. The CABS algorithm is an energy based feature representation of sequences as matrices, where matrix elements correspond to probabilities of residues adopting each of Helix (H), Sheet (E) or Coil (C) states. The CABS algorithm had been previously used to generate energy potential based features for the CB513 [38] and G Switch Proteins datasets [42]. The procedure is described in detail in section 5.3.1. Next, a brief review and motivation for using a

Complex-valued neural network classifier are given.

Complex-valued Neural Networks:

The neural networks employed by secondary structure predictors in the literature have been Real-valued networks. In comparison, here a fully Complex-valued neural network (CVNN) is employed to predict protein secondary structures. CVNN were initially proposed for pattern recognition tasks involving Complex signals in applications such as imaging and telecommunications [44]. Further, it has been shown that a Complex-valued neuron has a more powerful decision making ability compared to its Real-valued counterpart, due to orthogonal decision boundaries provided by the Complex plane [43, 44]. To perform classification, the Complex signal was treated as two Real-valued components in split networks as initially there was a lack of Complex-valued activation functions. Subsequently, a class of Complex-valued activation functions known as elemental transcendental functions were employed together with a Complex-valued backpropagation algorithm [130] in a Fully Complex-valued Multi-Layer Perceptron (FC-MLP) [131]. Besides being fully Complex in its input, hidden and output layers, the FC-MLP also employed the gradient descent algorithm. In a later development, the Improved fully Complex- MLP (IC-MLP) implemented a logarithmic error function in which the phase of the error was accounted for in addition to its magnitude [132].

To better emulate the capabilities of Radial Basis Function (RBF) networks in CVNN, the Gaussian-like *sech* activation was implemented in Complex-RBF (CRBF) networks [133, 134] and later the Fully Complex-valued Radial Basis Function (FC-RBF) network [132] with Complex-valued weights. Subsequently a Meta-cognitive approach to self-regulate aspects of learning was implemented in the Meta-cognitive FC-RBF (McFC-RBF) [135]. In the McFR-RBF network generalization capabilities were enhanced due to strategies such as sample selection.

Besides these networks, a Fully Complex-valued Relaxation Network (FCRN) [47] that employed a projection based-learning algorithm to analytically obtain its weights as well as a Circular Complex-valued Extreme Learning Machine (CC-ELM) [136] that

converted real input to the Complex plane by a circular transformation ($\Re \rightarrow \mathbb{C}$) were also proposed.

Sequential CVNN (distinct from the batch learning CVNN discussed thus far) are capable of learning online and adjusting their network architectures by growing or pruning neurons, or by making sample selection decisions as required. Proposed networks include the Complex-valued Minimal Resource Allocation Network (CMRAN) [137] and Self-regulatory Resource Allocation Network (CSRAN) [138]. As CVNN are computationally powerful and have shown good performance on classification tasks from different applications [128], it is motivating to apply CVNN towards the prediction of protein secondary structures and in the development of the compact model.

The FCRN [47] transforms Real-valued input to the Complex domain using a circular transformation, with the analytically obtained weights of the learning algorithm making it a rapid classifier needing only a small number of hidden neurons. Further, the training time was several orders of magnitude lower than other CVNN while performance was high [128]. Hence in this work the FCRN is used for the prediction of protein secondary structures and in the development of the compact model due to its efficient classification. It has previously been shown that datasets could contain noise arising from uncertainties in experimental structures [22]. To better approximate the decision function, the FCRN employs a logarithmic error function that accounts for both the magnitude and phase of the error. Additionally, the FCRN calculates the network optimum weights analytically, avoiding local minima traps common in gradient descent based approaches. The CB513 dataset is used to detect the compact model proteins while the G Switch Proteins dataset serves as a validation set. Results of using the FCRN classifier on the CB513 and G Switch Proteins datasets are also presented and discussed.

The remainder of this chapter, is organized as follows. Section 5.2 presents the CB513 and G Switch Proteins datasets used in this work. Section 5.3.1 describes the CABS algorithm based feature encoding procedure, the architecture and learn-

ing algorithm of the Fully Complex-valued Relaxation Network (FCRN) classifier. In addition, the heuristics based procedure of deriving training proteins in the compact model is also presented. Section 5.6 presents the performance results of FCRN on the CB513 dataset using the compact model training proteins and cross-validated results. A comparative study using other classifiers such as SVM is also presented. Next, the compact model is validated on the G Switch proteins dataset and compared with existing secondary structure predictors in the literature. In section 5.7, an analysis of SCOP groups present in the CB513 dataset is done to see the effect on Q_3 accuracy, with poorly performing chains identified. Next, the structural reasons behind poorly performing protein chains are investigated through a case study of the trypsin and cAMP dependent protein kinase inhibitors. Based on the case study, the effect of hydrogen bond type on mis-predicted Coil and Sheet residues is studied for the CB513 dataset, leading to the conclusion that current methods may not capture some types of hydrogen bonded residues effectively.

5.2 Datasets

This section describes the CB513 dataset used to develop the compact model and the G Switch proteins set used as validation.

5.2.1 CB513

The CB513 dataset was developed in 1999 by Cuff and Barton and is a standard benchmark dataset used widely in protein secondary structure prediction [38]. The authors combined sequences from the 3Dee database of structural domains [139] together with the RS126 dataset [89] and removed similar sequences with an SD score greater than 5 in order to obtain a non-redundant dataset. The total dataset contained 513 non-redundant sequences comprising 84,107 residues [38]. The CABS algorithm based encoding used in this chapter relies on structural templates to calculate the energy of an alignment of a target protein sequence to a template. To ensure a fair study, the homologues in the CB513 dataset were further removed with respect to CATH templates [140] used in this work, resulting in a final of 385 proteins. Details

of the homology removal are presented in section 5.3.1.

5.2.2 G Switch Proteins (GSW25)

The G Switch Proteins dataset contains a set of protein sequences derived from the G_A and G_B domains of the *Streptococcus* G protein [141]. The G_A domain proteins contain the $3\text{-}\alpha$ fold and the G_B domain proteins contain the $4\text{-}\beta + \alpha$ fold. A series of mutation experiments discovered that a single residue mutation from Lysine to Tyrosine at position 45 (L45Y), causes a change in secondary structures adopted by neighbouring residues [142]. Consequently, the protein is able to switch between the $3\text{-}\alpha$ and $4\text{-}\beta + \alpha$ structural folds. The G_A proteins bind human serum albumin while G_B proteins bind to the Fc region of Immunoglobulin G; thus for a small change in sequence space, a large change in structural and function space is observed.

This dataset contains 25 protein sequences of which 12 were G_A and 13 were G_B chains. The set serves as a blind test to validate the compact model and is not used in training. A list of the 25 sequences is available in Appendix B.

5.2.3 Secondary Structure Assignments

DSSP secondary structure assignments [23] were performed on both the CB513 and GSW25 datasets. Following the commonly used eight to three state reduction in most secondary structure prediction works such as [2, 4], states H, G and I corresponding to α , 3_10 and π helices respectively, were reduced to Helix (H). States E and B corresponding to extended and single residue β -strands respectively, were reduced to Sheet (E). Lastly, states T and S corresponding to β -turns and β -bends respectively, as well as loops and irregular structures denoted by blank states, were reduced to Coils (C). Table 5.1 shows the percentage of residues belonging to each of Helix (H), Sheet (E) or Coil (C).

Table 5.1: Secondary Structure Datasets. Brackets indicate percentage of residues belonging to each of the classes H, E and C.

Dataset	Proteins	Residues	Classes (%)
CB513	385	63,079	H (35), E (23), C (42)
GSW25	25	1400	H (52), E (39), C (9)

5.3 Methods

This section describes the CABS algorithm based feature extraction procedure that was applied to the CB513 and GSW25 datasets. The data were generated in a previous work [42]. The learning algorithm and architecture of the FCRN as well as the heuristics based procedure for detecting the compact model proteins are also described.

5.3.1 Feature Generation using the CABS algorithm

The CABS algorithm is a reduced lattice based representation of protein structures that allows a target sequence to be aligned to a template and the reference energy of the alignment, to be computed [41]. To ensure that performance results are unbiased, homology removal between target sequences of interest and templates is performed in order to provide fair results. However, since high scoring templates are detected using energy of alignment in the threading process, the method indirectly incorporates structural information unlike a sequence profile based method like the PSSM as discussed in section 5.3.1.

Using the knowledge-based statistical potentials provided by the CABS algorithm, a residue from a target sequence was encoded as a vector of 27 features. The first 9 features contain the probability that it adopts a helix, the next 9 the propensity to form Sheet (E) structures and the final 9, the propensity of adopting a Coil (C) structure. Figure 5.2 illustrates the process of feature generation. The process consists of two broad steps - template removal and energy potentials generation which are described next. In the following description, target refers to an amino acid sequence for which features are being generated.

Removal of Similar Templates: A library of 1000 CATH structural templates was firstly downloaded [140]. In order to conduct a fair study, Needleman-Wunsch [143] global alignment of CB513 targets to template sequences was performed and homologous templates were removed. This step resulted in 513,000 pairwise alignments with 97% of alignments having sequence similarity between 10 to 18%. Next, structural homologues were removed by querying the target names against Homology-derived

Secondary Structure of Proteins (HSSP) [144] data for template structures. Additionally, some templates had to be discarded due to PDB file errors and computational resource concerns. This resulted in a final number of 422 CATH templates and 385 CB513 target sequences that were non-redundant and available for use. DSSP secondary structure assignments were then performed for the template structures and the contact maps for heavy atoms such as C, O and N were computed in preparation for the next step. Any two heavy atoms were regarded to be in contact if their distance (computed from atomic coordinates) was below a threshold of $4.5 \text{ \AA}$.

Threading and Computation of Reference Energy: In this step, the target to template alignment was performed and the reference energy computed using the CABS algorithm. A brief description of the CABS algorithm is first given. The CABS algorithm represents a protein structure using a coarse-grained side chain model. This reduced representation saves computational cost. Figure 5.1 illustrates the reduced representation, where a single amino acid residue is represented as four connected atoms that serve as interaction centres. These are the C-Alpha ($C\alpha$), C-Beta ($C\beta$), Side chain centroid and a virtual $C\alpha$ to $C\alpha$ bond modelling the peptide bond. Using this reduced representation, the template structure is projected onto a cubic lattice with grid points spaced $0.61\text{\AA}$ apart such that $C\alpha$ atoms are constrained to lie on the grid points while other atoms are allowed to occupy any point in the lattice.

Figure 5.1: CABS Lattice. Reduced representation in the CABS lattice model. The figure was reproduced from [41].

The reference energy of interacting atoms is computed by a force field and the Replica Exchange Monte Carlo technique is used to locate structures with lower energy minima. The interaction energies modelled by the force field are briefly presented here:

- Hydrogen bonds, E_H
- Sequence independent short range interactions, E_g
- Sequence dependent short range interactions, E_s
- Repulsive interactions, ϵ_r
- Long range pairwise interactions, $E_{i,j}$

These force field terms were derived from careful analysis of structural regularities from a representative set of crystallographic structures. The detailed force field equations may be found in [41]. Using a sliding window of size 17, a target sequence was threaded onto each template structure and the reference energy (or energy potentials) computed with the CABS algorithm that has been described. The best matching templates were detected using a scoring function (unpublished) and the best fit residues having the lowest energy were retained.

Next, DSSP assignments from high scoring templates were read in a centre of 9 residues within the window of 17. In order to represent the most likely SS adopted by a target residue, the probability of the 9 residues adopting each of Helix, Sheet or Coil states was computed using the hydrophobic cluster similarity [145], resulting in a feature vector of 27 elements per residue. The process is illustrated by Figure 5.2. The 27 features serve as input to the FCRN classifier next described.

5.3.2 Fully Complex-valued Relaxation Network Classifier

This section firstly presents the problem of secondary structure prediction to be approximated by the Fully Complex-valued Relaxation Network (FCRN) [47]. The architecture and learning algorithm of the FCRN are then described. Let a set of observed residues and their corresponding labels be described by $\{(\mathbf{x}^1, \mathbf{c}^1) \dots (\mathbf{x}^t, \mathbf{c}^t) \dots (\mathbf{x}^N, \mathbf{c}^N)\}$,

Figure 5.2: CABS algorithm based Feature Encoding. A target residue, t in the input sequence is represented as a 27-dimensional feature vector. The input sequence is read in a sliding window (w) of 17 residues (grey). The central residue (t) and several of its neighbours to the left and right are shown. CATH templates were previously assigned SS using DSSP. Target to template threading was done using $w = 17$ and the reference energy computed with the CABS algorithm. The SS are read in from best fit template sequences that have the lowest energy for the central 9 residues within w . Since multiple SS assignments will be available for a residue, t and its neighbours from templates, the probability of each SS state is computed using a hydrophobic cluster similarity score. $P(H)$, $P(E)$ and $P(C)$ denote probabilities of t and its four neighbours to the left and right, adopting Helix, Sheet and Coil structures respectively. CATH templates are homology removed and independent with respect to the CB513 dataset.

where $\mathbf{x}^t \in \Re^m$ are m -dimensional input features ($m = 27$) and $\mathbf{c}^t \in \Re^n$ are the 3-dimensional class labels. N is the total number of residues in a dataset. Here, $\mathbf{x}^t$ is a single residue with $m = 27$, that has been represented as a 27 dimensional feature vector using the CABS based feature encoding described earlier. Also, $\mathbf{c}^t$ are the **H**elix, **S**heet (**E**) or **C**oil states numerically represented as 1, 2 and 3, respectively.

The real valued input features $\mathbf{x}^t$ and class labels $\mathbf{c}^t$ are made Complex using a circular transformation and coded class labels respectively, in order to convert the problem from the Real to Complex domain. The circular transformation for the j -th feature for the t -th residue, or sample is given by

$$z_j^t = \sin(ax_j^t + ibx_j^t + \alpha_j); j = 1, \dots, m \quad (5.1)$$

where $\mathbf{z}^t = [z_1^t \dots z_j^t \dots z_m^t]^T$ are m -dimensional Complex-valued input features. $a, b \in [0, 1]$ are randomly chosen scaling constants. $\alpha_j \in [0, 2\pi]$ is the origin shift term, allowing the origin of the Complex plane to be shifted such that all 4 quadrants are used effectively for classification. The coded class labels in the Complex domain for each residue t are given by $\mathbf{y}^t = [y_1^t, \dots, y_l^t, \dots, y_n^t]^T \in \mathbb{C}^n$, where

$$y_l^t = \begin{cases} 1 + 1i & \text{if } c^t = l \\ -1 - 1i & \text{otherwise} \end{cases} \quad l = 1, \dots, n; \quad (5.2)$$

The circular transformation and coded class labels map the real valued set $\{(\mathbf{x}^1, \mathbf{c}^1), \dots, (\mathbf{x}^t, \mathbf{c}^t), \dots, (\mathbf{x}^N, \mathbf{c}^N)\}$ to the Complex-valued set $\{(\mathbf{z}^1, \mathbf{y}^1), \dots, (\mathbf{z}^t, \mathbf{y}^t), \dots, (\mathbf{z}^N, \mathbf{y}^N)\}$. The problem of predicting secondary structures given the set of observed residues can be written as approximation of the functional relationship $F : \mathbf{x}^t \in \Re^m \rightarrow \mathbf{y}^t \in \mathbb{C}^n$, where $m = 27$ and $n=3$ since there are three secondary structure classes of interest. The FCRN is a feed forward neural network comprising of three layers: an input layer; a hidden layer; and an output layer. The architecture of the classifier is shown in Figure 5.3.

The input layer consists of m transformation neurons employing the circular transformation given in equation (5.1) to convert the Real-valued input into the Complex

Figure 5.3: Architecture of FCRN. The FCRN consists of a first layer of m input neurons, a second layer of K hidden neurons and a third layer of n output neurons. For the SS prediction problem presented in this work, $m = 27$, $n = 3$ and K is allowed to vary. The hyperbolic secant (*sech*) activation function computes the hidden response (h_l^t) and the predicted output $\hat{y}_l^t$ is given by the exponential function. w_{nK} represents the weight connecting the K th hidden neuron to the n th output neuron.

domain. The hidden layer contains K neurons employing the Gaussian-like, hyperbolic secant activation function, $\text{sech}(\cdot)$. The output layer comprises n output neurons employing an exponential activation function. The response of the k th neuron in the hidden layer h_k^t , is given by

$$h_k^t = \text{sech}(\mathbf{v}_k^T(\mathbf{z}^t - \mathbf{u}_k)); k = 1, \dots, K \quad (5.3)$$

where $\mathbf{v}_k = [v_k^1, \dots, v_k^m]^T \in \mathbb{C}^m$ is the scaling factor of the k -th hidden neuron and $\mathbf{u}_k = [u_k^1, \dots, u_k^m]^T \in \mathbb{C}^m$ is the centre of the k -th hidden neuron. The hyperbolic secant function is defined as $\text{sech}(z) = \frac{2}{e^z + e^{-z}}$.

The response of the l -th neuron in the output layer, employing an exponential activation function, is given by

$$\hat{y}_l^t = \exp\left(\sum_{k=1}^K w_{lk} h_k^t\right); l = 1, \dots, n \quad (5.4)$$

where $w_{lk} \in \mathbb{C}$ is the weight connecting the k -th hidden neuron and the l -th output neuron. The input weights connecting the input and hidden layers, w_{kj} are randomly selected and the output weights w_{lk} are analytically estimated using a logarithmic error function.

The logarithmic error function employed by the FCRN [47] takes both the phase and the magnitude of the error into account. It estimates the error between the network output $\hat{\mathbf{y}}^t$ and actual output $\mathbf{y}^t$ and is given by

$$J(\mathbf{W}) = \frac{1}{2} \sum_{t=1}^N J_t = \frac{1}{2} \sum_{t=1}^N \sum_{l=1}^n \left(\ln \left(\frac{\hat{r}_l^t}{r_l^t} \right)^2 + \left(\hat{\phi}_l^t - \phi_l^t \right)^2 \right) \quad (5.5)$$

where

$r_l^t = \|y_l^t\|$ is the magnitude of y_l^t

$\hat{r}_l^t = \|\hat{y}_l^t\|$ is the estimated magnitude

$\phi_l^t = \arctan\left(\frac{\text{Im}(y_l^t)}{\text{Re}(y_l^t)}\right)$ is the phase of y_l^t

$\hat{\phi}_l^t = \arctan\left(\frac{\text{Im}(\hat{y}_l^t)}{\text{Re}(\hat{y}_l^t)}\right)$ is the estimated phase

It can be shown that the optimal weight matrix ($\mathbf{W}^*$) corresponds to the minimum energy point of equation (5.5) [47]. Using a projection based learning algorithm, the output weights corresponding to the minimum of the energy function are analytically estimated, as given by

$$\mathbf{W}^* = \bar{\mathbf{B}} \bar{\mathbf{A}}^{-1} \quad (5.6)$$

Further, the projection matrix $\mathbf{A} \in \mathbb{C}^{K \times K}$ is given by

$$A_{pk} = \sum_{t=1}^N h_p^t \bar{h}_k^t; \quad p = 1, \dots, K; \quad k = 1, \dots, K \quad (5.7)$$

the output matrix $\mathbf{B} \in \mathbb{C}^{n \times K}$ is

$$B_{lp} = \sum_{t=1}^N \ln \bar{y}_l^t h_p^t; \quad l = 1, \dots, n; \quad p = 1, \dots, K \quad (5.8)$$

Since the projection matrix $\mathbf{A}$ is square and therefore non-singular, it allows the optimal weight matrix $\mathbf{W}^*$ to be calculated easily. Given a set of N training samples,

randomly chosen scaling factors $\mathbf{u}_k$ and hidden neuron centres, $\mathbf{v}_k$, the hidden layer response are computed. Next, the projection matrix A and output matrix B , are computed. Using Equation (5.6), the optimal weights W are calculated. Pseudocode 1 (obtained from [47]) describes the sequence of steps in the learning algorithm. Further details of the learning algorithm and derivation of the error function are available in [47]. The complexity of this algorithm is estimated to be in the order of $O(K^3)$. Due to the analytical method of determining the optimum weights, the FCRN can rapidly classify the secondary structures. In the next step, the FCRN is used to derive the compact model proteins.

Pseudocode 1 Pseudo code for the FCRN algorithm

Given the training dataset:
 $(\mathbf{z}^1, \mathbf{y}^1), \dots, (\mathbf{z}^t, \mathbf{y}^t), \dots, (\mathbf{z}^N, \mathbf{y}^N)$
and the random hidden layer
parameters: $(\mathbf{u}_j, \mathbf{v}_j)$, $j = 1, \dots, K$
START
 Compute the hidden layer responses h_j^t
 using Eq.(5.3).
 Compute the square matrix A
 according to Eq. (5.7).
 Compute the matrix B using Eq. (5.8).
 Estimate the optimum output weights
 using Eq. (5.6).
END

5.3.3 Heuristics based Approach to derive the Compact Model

The CB513 dataset proteins are used to develop the compact model. The first four and last four residues of a protein sequence lack information of neighbouring residues to the left and right, since the energy potentials based feature encoding employs a sliding window of 9 residues. Hence, the representation of these residues in the feature space lacks adequate information and may introduce a high variability in the sequence to structure signal, making them difficult to model. In addition the first and last residues correspond to the N and C termini of a protein sequence, respectively. It is known that the termini of a protein are flexible and subject to evolutionary pressures resulting from physical constraints rather than sequence specific structures.

For example, during translation the nascent protein needs to move through the Golgi apparatus which necessitates high flexibility regardless of sequence. Such instances could introduce much variation into the sequence to secondary structure relationship $F : \mathbf{x}^t \in \mathbb{R}^m \rightarrow \mathbf{y}^t \in \mathbb{C}^n$ that is to be estimated. Since the number of training proteins needed to achieve a good accuracy for a given a dataset is unknown, it was first estimated using randomized trials as follows. The 385 proteins from the CB513 dataset were sequentially numbered from 1 to 385. Next, 385 unique numbers were generated using the MATLAB *rand* function based on a uniform random distribution in the range $[1, 385]$. Trials were conducted by adding randomly chosen 5 training sequences to the training set and the Q_3 achieved for a particular training set, was estimated by testing on the remainder of the protein sequences. The total number of hidden neurons (K) in the hidden layer of the FCRN was allowed to vary, but capped at 100 neurons to avoid over-training. All experiments were conducted using MATLAB R2012b on a Windows 7 platform, with a processor speed of 3.6 GHz and 8GB RAM. Figure 5.4 shows the Q_3 as a function of increasing the number of training sequences. From the figure, it can be seen that the Q_3 peaks at 82% for 50 training sequences.

Figure 5.4: Plot of Q_3 vs no. of training sequences (N). The accuracy achieved by FCRN as a function of increasing N is shown. Highest Q_3 is observed at 82% for 50 sequences. Maximum allowed hidden neurons = 100.

Beyond this number, adding more training proteins does not contribute much to Q_3 ; instead, a slight drop in performance is observed with $Q_3 = 81.72\%$ for 55 proteins. Hence, with 50 as an approximate guideline of the number of proteins needed to obtain a maximal accuracy, various protein sets are constructed such that comparable

accuracies to cross-validated Q_3 (e.g. around 80%) are obtained. First, the CB513 proteins are visualized according to the ratio of H, E and C structures in their respective sequences as shown in Figure 5.5. Each data point in the figure represents a

Figure 5.5: Secondary Structure Ratio in CB513 Sequences. Plot of CB513 proteins by their secondary structure content. One circle represents a single protein sequence. SSP_{50} proteins are represented as yellow circles while the remainder of the CB513 dataset are green circles. The compact model, SSP_{55} proteins are spread out in a similar fashion to the SSP_{50} proteins shown here. Axes show the proportion of Helix, Coil and Sheet residues divided by the sequence length. For instance, a hypothetical 30 residue protein comprised of only Helix residues, would be represented at the bottom-right most corner of the plot.

single protein sequence and the axes indicate the ratio of each SS type. From here, a training group of 50 proteins labelled SSP_{50} is manually selected such that combinations of various secondary structure ratios are included. For example, proteins containing only Helix residues, proteins with similar proportions of all three SS states and so on are selected such that structure space is sampled evenly. SSP_{50} proteins are shown as yellow circles in Figure 5.5. To test the effect of individual residue based rather than residues from protein based sampling, an additional training set (labelled $SSP_{balanced}$) was created. It contains randomly selected residues in equal proportion from the three classes. Besides SSP_{50} and $SSP_{balanced}$, another randomly selected dataset of 50 proteins (labelled $SSP_{sampled}$) that is distinct from the Q_3 vs. N experiments presented earlier was also selected to serve as a control set. The numbers

of proteins and residues in the three sets as well as the ratios of secondary structure classes are given in Table 5.2 and Figure 5.6 respectively.

Table 5.2: Summary of Protein Training Sets

Training	Residues	Proteins
SSP ₅₀	7974	50
SSP _{sampled}	7192	50
SSP _{balanced}	7974	-

Figure 5.6: Secondary Structure Composition of training sets. The pie charts indicate the ratio of residues from different classes in the SSP₅₀, SSP_{sampled} and SSP_{balanced} training sets.

By testing on the remainder of the CB513 proteins, it was observed that the Q_3 accuracies obtained for the three sets were similar at about 81%. The training sets from the Q_3 vs N experiments as well as the three sets obtained above were tested against the G Switch proteins (GSW25) dataset. Next, a training set containing 55 proteins that gave the highest accuracy on GSW25, was selected. This set was termed the compact model, SSP₅₅ and is given in Appendix B. A similar technique may be applied to other datasets. A generic procedure containing 3 steps is given below.

1. Estimate the number of training proteins, P needed to achieve a good accuracy on a given dataset by randomized trials increasing the number of proteins.
2. Using P as a guideline, construct various other training sets according to a desired distribution, such as manual selection, residues based selection, etc..
3. Test all derived training sets against a blind test dataset and select training set with highest score as the compact model.

A pseudocode describing the procedure is given in Procedure 1. The next sec-

```

Input: Main dataset of protein chains,  $D_1$  and
          Blind Test Dataset,  $D_2$ 
Output: Compact Model,  $M_1 \subset D_1$ 
// Initialization.  $Q_3$  is the single residue accuracy.
 $Q_3 \leftarrow 0$ ;
 $M_1 \leftarrow 5$  randomly selected training sequences from  $D_1$ ;

// Stage 1: Generate training sets,  $M_i$  by adding 5 random sequences
          from  $D_1$  at each iteration
repeat
    Train FCRN classifier with train dataset  $M_i$ ;
    Obtain  $Q_3$  on test dataset  $D_1 - M_i$ ;
     $M_{i+1} \leftarrow M_i + 5$  random sequences from  $D_1$ ;
    Update best  $Q_3$ ;
    Save all  $M_i$  and their FCRN parameters that give  $Q_3$  within desired error;
    Save  $P$ , the no of training sequences giving best  $Q_3$  ;
until  $Q_3$  does not increase significantly;

// Stage 2: Generate more training sets,  $M_i$  from  $D_1$ 
Using  $P$  as guideline, select training sets based on various characteristics (See text);

// Stage 3: Select the Compact Model,  $M_1$  by choosing the best
          performance against  $D_2$ 
 $BEST \leftarrow 0$ ;
foreach  $M_i$  do
    Test  $M_i$  against  $D_2$  using FCRN;
    Save  $Q_3$  ;
    if  $BEST < Q_3$  then
         $BEST \leftarrow Q_3$ ;
         $M_1 \leftarrow M_i$ ;
    end
end

```

Procedure 1: Obtain Compact Model

tion compares the performance results of the compact model to results from cross-validation. Both the cross-validated and compact models were developed using separate instances of the FCRN classifier. First, the performance measures used in this study are presented.

5.4 Performance Measures

Similar to chapter 3, global and class-wise performance measures are used. Here they are applied to Helix (H), Sheet (E) and Coil classes rather than Membrane, Inside or Outside residues. The measures given below are computed from the confusion matrix of classification.

The class-wise performance is measured as the percentage of correctly classified residues per class (η_j) which is defined as:

$$\eta_j = \frac{q_{jj}}{N_j} \times 100\% \quad (5.9)$$

where q_{jj} is the total number of correctly classified samples in the class c_j and N_j is the total number of samples in that class for $j \in \{H, E, C\}$.

The global performance measures used are the overall accuracy, η_o which is the overall number of correctly classified samples in the dataset, as well as the average class-wise accuracy, η_a , defined as

$$\begin{aligned} \eta_a &= \frac{1}{3} \sum_{j=1}^3 \eta_j \\ \eta_o &= \frac{\sum_{j=1}^3 q_{jj}}{\sum_{j=1}^3 N_j} \times 100\% \end{aligned} \quad (5.10)$$

The overall accuracy (η_o) is the same measure as the Q_3 score and the two are used interchangeably. Further, Q_j corresponds to class-wise accuracy η_j which is defined in Equation 5.9.

Besides the single residue accuracy, a segment based prediction measure known as the Segment Overlap Score (SOV) is computed. The SOV compares predicted SS

segments with native SS segments and is defined by Zemla [146] as

$$SOV = 100 \times \frac{1}{N} \sum_{i \in H, E, C} \frac{\minov(s_1, s_2) + \delta(s_1, s_2)}{\maxov(s_1, s_2)} \times \text{len}(s_1) \quad (5.11)$$

where

s_1 and s_2 are observed and predicted secondary structure segments respectively, in state i

$\minov(s_1, s_2)$ is the length of actual overlap between segments s_1 and s_2

$\maxov(s_1, s_2)$ is the total extent for which either of the segments s_1 and s_2 has a residue in state i

$$\delta(s_1, s_2) = \min\{(\maxov(s_1, s_2) - \minov(s_1, s_2)); \minov(s_1, s_2) \\ \text{int}(\text{len}(s_1)/2); \text{int}(\text{len}(s_2)/2)\}$$

N is a normalized sum over all states

Individual segment overlaps scores for each of the three states, SOV_H , SOV_E and SOV_C are also computed in a similar manner, except that the normalized sum N , is calculated only for segments in that particular state, following Zemla's definition.

Finally, the Matthews Correlation Coefficient (MCC), following the definition in [6]

$$\text{is also computed. It is defined as } MCC_j = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

where $j \in \{H, E, C\}$. Since the MCC is a binary classification based measure, here it is adapted to a three-state classification as follows. If $j = H$ (for the Helix state), True Positives (TP) denote the number of correctly predicted positives in that class, i.e., the number of native helices predicted as helices. True Negatives (TN) are the number of non-helices predicted as non-helices (e.g. Sheets predicted as Sheets or Coils). False Positive (FP), is any non-native Helix predicted as a Helix (e.g. Sheet or Coil predicted as Helix). False Negatives are native Helices predicted as either Sheets or Coils. A similar interpretation is followed for the Sheet and Coil classes.

The next section presents the performance results.

5.5 Performance Study of FCRN

In this section, a residue-wise performance study of the FCRN using the CB513 dataset is presented. For the results in the section, the full number of residues, 63,079 is used to conduct the studies, i.e. the initial and ending four residues are included. The results of this section first serves to investigate the performance of the FCRN using the traditional cross-validated models training on a large number of protein and residue samples. Cross-validation was performed by randomly selecting 75% of residues for training and the remainder used as test. Each training partition contained 47,309 residues and each testing partition, 15,770 testing residues. The performance is compared with an Extreme learning machine ELM [147] and Support Vector Machine (SVM) package obtained from LIBSVM [117]. The SVM parameters (c, γ) were optimized with a grid search. The training and testing efficiencies are reported in Table 5.3, with brackets showing the standard deviation.

Table 5.3: Performance of FCRN Predictor in Comparison with other classifiers

Method	K	Training (%)		Testing (%)	
		η_o	η_a	η_o	η_a
FCRN	82	82.0 (0.05)	81.0 (0.08)	82.1 (0.09)	81.25 (0.25)
ELM	77	81.56 (0.12)	80.76 (0.16)	81.48 (0.17)	80.69 (0.24)
SVM	21351 ¹	81.86 (0.05)	80.99 (0.1)	81.81 (0.19)	80.94 (0.29)

Table 5.3 indicates that the testing overall accuracy (η_o) of all three classifiers is very similar at about approximately 82%. However, η_o for the FCRN at 82.1% was the highest compared to the ELM and SVM which obtained 81.48% and 81.81%, respectively. Of these results the FCRN reported the lowest standard deviation (s.d.) of 0.09, compared to ELM (0.17) and SVM (0.19), with the brackets showing the standard deviation. The average class accuracy (η_a) for testing obtained by the FCRN was 81.25% which was higher by 0.56% than the ELM and 0.31% higher than the SVM. The training performance was very similar, also showing close performances near 82%. FCRN obtained the highest result of ($\eta_o = 82.0\%$) while the second highest is the SVM at 81.86% followed by the ELM classifier at 81.56%.

Overall, the FCRN showed a slightly better performance compared to both the ELM

¹Number of support vectors

and SVM, with s.d. of η_o and η_a being either the lowest, or on par with the lowest reported across the 4 partitions.

The results suggest that the FCRN performance is more efficient as only 82 hidden neurons are employed compared to 21,351 support vectors needed by the SVM. It indicates that despite using a much smaller number of neurons, the FCRN was able to generate the highest performance of the three classifiers. As 47,309 residues are used in the training, the ratio of training instances to support vectors is estimated to be 2.2. This is very low in comparison to the FCRN, with a ratio of training instance to hidden neuron at 577 and 614 for the ELM correspondingly. Hence, FCRN needs much lesser parameters than SVM to classify the same amount of samples.

The protein secondary structure prediction literature rarely report residue-wise training results, as the Q_3 accuracy is often computed on a per protein basis. Nevertheless, the experiment indicates that given a sufficiently large number of training residues, e.g. greater than 40,000, the classifier may be able to achieve Q_3 accuracies of 80% or higher.

Next, five-fold cross-validation was performed where the partitions consist of randomly selected proteins rather than residues as is conventionally done. Here each partition contains 308 proteins with the remaining 77 serving as test and the number of hidden neurons is set to a maximum of 200. The overall accuracy (η_o) averaged across 5 partitions is 82.14%, with a standard deviation (s.d.) of 0.64; the average class accuracy (η_a) is 81.3% (s.d. = 0.42). The overall accuracy or Q_3 score is compared with several results that have been reported for the CB513 dataset in Table 5.4.

Table 5.4: FCRN overall accuracy comparison with other protein secondary structure prediction methods

Method	$Q_3(\%)$	Total proteins	Total residues	Training
GOR V [85]	73.5	513	84,107	Jack-knife
JNet [92]	76.4	480	-	7-fold CV
PSIPRED [86], [37]	80	513	84,107	3-fold CV
CDM [86]	80.3 ²	513	84,107	-
CPM [126]	85.6	513	84,119	7-fold CV
FLOPRED [42]	81.3	387	63,383	5-fold CV
FCRN	82.14	385	63,079	5-fold CV

²Average of actual accuracies obtained based on availability of structural fragments in the PDB

As the number of proteins and cross-validation partitions differs depending on the method used, these are also reported in Table 5.4 to provide a context for interpretation of the Q_3 scores. The results indicate that the FCRN is the second highest method, with a Q_3 of 82.14%. The highest Q_3 was 85.6% obtained by the Compound Pyramid Model or CPM [126]. The model contained a hierarchical classification system with three layers employing a combination of neural networks and support vector machines, thereby increasing the predictive power. Nevertheless, one of the drawbacks was a high computational cost since the model also involved querying databases for rule associations [126]. The older secondary structure prediction methods had accuracies of about 80% or lower such as GOR V, JNet (also known as Jpred) and PSIPRED. The Consensus Data Mining, or CDM method reported Q_3 scores in a range between 67.5% to 93.2%, since in their method Q_3 is dependent on the availability of matching structural fragments in the PDB. Here, an average of the two scores has been reported in Table 5.4. All the methods so far discussed, represent protein sequences using PSSM, or Position Specific Scoring Matrices as described in 2. However, both FLOPRED and FCRN used similar, energy potentials based feature encoding that was described in section 5.3.1. The FCRN performance is slightly better by 0.84% in comparison to 81.3% reported by FLOPRED. The results indicate that the CABS algorithm based feature encoding employed in this chapter, is able to provide comparable or better performance to those reported literature in most of the cases.

5.6 Performance Study of Compact Model Using FCRN

The previous section presented a performance study and a comparative study of the FCRN in using cross-validation together with energy potentials based feature encoding to predict the SS. The results indicated that for the CB513 dataset, the FCRN demonstrated a reasonably high Q_3 of about 82%. In this chapter, the FCRN is used to explore the idea of a compact model of proteins as described in section 5.3.3.

Compact Model vs. Cross-validation on CB513: A five-fold cross-validation study was again performed for a basis of comparison, with each training partition containing randomly selected 308 proteins and the remaining 77 used as test. All experiments were again conducted using MATLAB 2012b running on a Windows 7 platform. Under these settings, the average five-fold overall accuracy (η_o) was 81.9% and class accuracy (η_a) was 81.11%. The average number of hidden neurons was $K = 98$. The parameters of the five partitions were saved and the best performing partition was selected as a model. This model was the Cross-validated Model, a training set of 308 proteins (SSP_{CV}) and compared with the Compact Model, a training set of 55 proteins (SSP_{55}), with results of the comparison presented in section 5.3.3. Separate instances of the FCRN are used to train SSP_{CV} and SSP_{55} and each instance is allowed 100 hidden neurons for a fair test. The compact model contains 55 training proteins and the remaining 330 serve as testing. For a fair comparison, the SSP_{CV} performance for these same 330 test proteins, comprising 51,634 residues is reported in Table 5.5.

Table 5.5: Results on CB513 (51,634 residues)

Model	Observed j	Predicted j			$Q_j(\%)$	$Q_3(\%)$	$SOV_j(\%)$	$SOV(\%)$	MCC_j
SSP_{CV}		H	E	C					
	H	16469	48	1840	89.72	82.03	83.14	79.46	0.82
	E	92	8804	2955	74.29		72.24		0.71
	C	2313	2032	17081	79.73		75.46		0.64
SSP_{55}	H	16333	62	1962	88.98	81.72	82.19	78.93	0.81
	E	87	9001	2763	75.96		73.43		0.71
	C	2288	2279	16859	78.69		74.5		0.63

The Q_3 obtained by SSP_{CV} and SSP_{55} are very similar at 82.03% and 81.72% respectively, although SSP_{CV} is slightly higher by 0.31% or about 160 residues. For the Helix and Coil accuracies, SSP_{CV} again had a slightly better accuracy (of about 1%) with Q_H and Q_C at 89.72% and 79.73%, respectively. However, for the Sheet class SSP_{55} obtained a Q_E of 75.96%, better than SSP_{CV} by approximately 1.7%. Consistent with findings from earlier studies [4, 5], the Helix class was easiest to predict, followed by Coil and Sheet. Under the SSP_{55} model most of the wrongly predicted Helices were Coils rather than Sheet. Similarly a majority of wrong Sheet predictions were Coils (96.9%) rather than Helices. Coil residues however, may have equal

propensity to be falsely predicted either Helices, or Sheet (50%). A similar trend is observed for the SSP_{CV} model.

In considering the SOV scores, the overall SOV is again close for both models, with SSP_{CV} being slightly higher at 79.46% compared to 78.93% obtained by SSP₅₅, a difference of about 0.5%. Again, SSP_{CV} obtained slightly higher SOV scores compared to SSP₅₅ for the Helix and Coil classes but SSP₅₅ predicted the strand segments better by 1.2% with SOV_E = 73.43%. The Matthews Correlation Coefficient, or MCC scores indicate that the performance for both models is extremely close. In considering the entire dataset of 385 proteins (including performance on training proteins), similar performances were obtained for both the SSP_{CV} and SSP₅₅ models. Since the results of both models are close, statistical tests were performed to detect if there is any significant difference in performance.

Statistical Tests: Two statistical tests were conducted to detect if performance per sequence as measured by Q₃ and overall SOV scores were significantly different depending on the model used. For the SSP_{CV} model, average scores of 5 partitions were used. First, a test for normality of the data is applied by performing the Shapiro-Wilk test [148]. The results indicated that at $\alpha = 0.05$, neither the Q₃ nor SOV scores were normally distributed with P values being very low ($<< 10^{-5}$). Since the data are not normal, the non-parametric Wilcoxon signed-rank test [149] was next applied to detect if paired scores per sequence were significantly different. The P values obtained for the Q₃ and SOV scores were 0.00012 and 0.015, respectively. At $\alpha = 0.005$, the test indicates that the scores from SSP_{CV} and SSP₅₅ are models are significantly different.

It was expected that employing only 55 training proteins would give a low accuracy. While cross-validated accuracies may be higher and significant according to statistical tests, it is remarkable that using only 55 training proteins, high scores of Q₃ ($Q_3 \geq 80\%$) were obtained. The ratio of number training proteins in the SSP_{CV} to SSP₅₅ models is 5.6, indicating that an increase of more than 5 times of training proteins only resulted in 0.31% difference in the Q₃ score.

It has been previously reported that increasing the number of training proteins results in a higher Q_3 by up to 4% [4]. Other factors between the results reported here and the study differ, such as the type of feature encoding and Real vs. Complex-valued neural networks employed. Hence, these factors likely influence the results reported here, resulting in only a small improvement to Q_3 despite adding more training proteins in the cross-validated model. The results suggest that the 55 training proteins have captured most of the structural fold information in the remainder of the CB513 test proteins; otherwise it would be difficult to achieve comparable accuracies to the cross-validated model with a much smaller number of training proteins.

Blind Tests on GSW25: Blind test results of the SSP_{55} model on the GSW25 dataset are presented and compared with the SSP_{CV} model. Here, the GSW25 dataset contains the first and last four residues of a protein sequence, unlike for the CB513 dataset where they had been excluded (see section 5.3.3). However, for a fair study that contains no knowledge of the validation set the normalization of the GSW25 potentials data was performed with respect to the maxima and minima for the full CB513 dataset (including the first and last four residues). For the SSP_{CV} model, parameters from the best performing partition are used in the prediction. Table 5.6 gives the performance of the two models on the GSW25 dataset.

Table 5.6: Results for G Switch Proteins (1400 residues)

Model	Observed j	Predicted j			$Q_j(\%)$	$Q_3(\%)$	$SOV_j(\%)$	SOV (%)	MCC_j
SSP_{CV}		H	E	C					
	H	682	1	39	94.46		100		0.79
	E	58	352	136	64.47	76.65	66.04	59.07	0.65
	C	53	40	39	29.55		62.75		0.13
SSP_{55}	H	680	0	42	94.19		100		0.83
	E	25	384	137	70.33	80.36	63.68	62.44	0.73
	C	51	20	61	46.22		78.91		0.25

The results of the table indicate that overall, the SSP_{55} model performed better than SSP_{CV} for this dataset across several measures. For the Q_3 accuracy, SSP_{55} obtained 80.36%, higher by 3.71% compared to SSP_{CV} . For the class-wise accuracies Q_H , Q_E and Q_C , SSP_{55} obtained 94.19%, 70.33% and 46.22% respectively. For the SSP_{CV} model the Q_H obtained is higher than SSP_{55} at 94.46%. However, a large drop

is observed in comparison to SSP₅₅ for the Sheet and Coil classes ($Q_E = 64.47\%$ and $Q_C = 29.55\%$ respectively). Both SSP₅₅ and SSP_{CV} obtained perfect Helix segment predictions with SOV_H at 100%, but Sheet and Coil classes proved difficult with SOV_E and SOV_C remaining lower than 80% for both models. Although SSP_{CV} obtained a higher Sheet segment prediction compared to SSP₅₅ ($SOV_E = 66.04\%$ vs. $SOV_E = 63.68\%$ respectively), the increased performance in this class resulted in a trade-off in the Coil class; the SOV_C of 78.91% is 16.16% higher than SSP_{CV}. Consequently, the overall SOV measure for SSP₅₅ is also higher by 3.37% compared to SSP_{CV}. The MCC scores for each class under the SSP₅₅ model are also higher. In general, the performance trend reflects that SSP₅₅ is a model that has much better generalization ability with respect to the G Switch proteins dataset; despite being trained with only 55 proteins, the results indicated a better performance in comparison with SSP_{CV} (trained with 308 proteins).

However, there exists much to be improved in the performance of SSP₅₅. In particular, the ability to detect Coil residues is poor with Q_C only being 46.22% and $Q_E = 70.33\%$ could also be improved further. As discussed earlier, in general methods show a high accuracy for the Helix class but find it difficult to detect Sheet and Coil residues at a similar level of performance. The use of the compact model, while efficient at predicting SS by learning from fewer structures, maintains the generalization capabilities well. However, there may be underlying difficulties imposed by either the feature representation or assignment of SS labels, that makes it difficult to distinguish the Coil and Sheet classes. In section 5.7, this thesis provides some structure based evidence as to why separation of the E and C classes remains difficult for the current techniques.

Next, the SSP₅₅ performance on the GSW25 dataset is compared with five widely used predictors in the literature. These are the *ab initio* methods, PSIPRED [91], SPINEX [95] and PORTER [99] as well as the homology model based methods SSPro [24] and PROTEUS [25]. These methods had been assessed in a comparative study of 2000 protein chains, of which they were the top performers in the respective categories of homology-based and *ab initio* predictors, having obtained Q_3 accuracies between 80-

82%. Here they are also applied to the GSW25 dataset. Since the comparative study, updated versions of the methods have been made available and are employed here. These are PORTER 4.0 [150] and Proteus 2 ³. For the SPINE method a server that employs deep learning to predict the secondary structures, termed SPIDER 2 ⁴ [27] was used. Besides these 5 methods which employed the PSSM based feature encoding, a 6th method called FLOPRED that employs the same energy potentials based feature encoding derived from the CABS algorithm was also compared [42]. FLOPRED applied an Extreme Learning Machine with particle swarm optimization to predict the secondary structures. The performance is presented in Table 5.7, ordered from highest to lowest Q₃. For consistency, all method names have been capitalized in the following discussion. From the table, SSP₅₅ obtained the highest Q₃ of 80.36%, with

Table 5.7: Methods Comparison on G Switch Proteins

Method	Observed j	Predicted j			Q _j (%)	Q ₃ (%)
SSP ₅₅	H	H	E	C	94.19	80.36
	E	680	0	42	70.33	
	C	25	384	137	46.22	
FLOPRED	H	51	20	61		
	E	665	19	38	92.11	78.72
	C	41	380	125	69.6	
PROTEUS 2	H	49	26	57	43.19	
	E	556	50	116	77.01	61.72
	C	17	302	227	55.32	
PSIPRED	H	2	124	6	4.55	
	E	519	99	104	71.89	57.36
	C	167	243	136	44.51	
PORTER 4.0	H	5	86	41	31.07	
	E	405	99	218	56.1	51.08
	C	22	267	257	48.91	
SPIDER 2	H	0	89	43	32.58	
	E	473	95	154	65.52	50.79
	C	112	213	221	39.02	
SSPRO	H	0	107	25	18.94	
	E	368	162	192	50.97	50.43
	C	13	312	221	57.15	
		1	105	26	19.7	

FLOPRED being the second highest at 78.72%. The performance of FLOPRED was very close to SSP₅₅, but the latter performed slightly better across the remainder of

³<http://wishart.biology.ualberta.ca/proteus2/index.jsp>

⁴<http://sparks-lab.org/yueyang/server/SPIDER2/>

the measures considered. The largest margin was observed for the Coil class, with $Q_C = 46.22\%$ for SSP_{55} being 3.03% higher than FLOPRED. The Q_H and Q_E scores indicated that SSP_{55} was able to predict Helix and Sheet structures better by 2.08% and 0.73%, respectively. For the remainder of the methods, the Q_3 was below 70%, with the third highest method being PROTEUS 2 at 61.72% and the lowest being SSPRO at 50.43%. The general trend indicates again, that Helix is the easiest class to predict for all methods. However, unlike performance on the CB513 observed in Table 5.5, the second easiest to predict class was Sheet structures rather than Coil, with the lowest rate of Coil prediction obtained by Proteus 2 being 4.55%. Since PROTEUS 2 is a homology based method that assigns SS states from matching templates directly as one of its steps, it may be possible that a high scoring but poor matching template structure resulted in wrong assignments for this particular dataset. Again, most of the misclassified Sheets were Coils and vice versa.

The higher results obtained by SSP_{55} and FLOPRED indicates that the choice of feature encoding can contribute to a good performance. The CABS based feature encoding employs potentials that are based on averages from a wide collection of crystallographic structures. Unlike homology based models, secondary structure states are not assigned to templates directly, but encoded as probabilities of various secondary structures adopted by the neighbour residues. While the method may indirectly incorporate structural information, it is homology removed with respect to CATH and CB513 templates; further it is ensured that the SSP_{55} compact model itself, contains no similar structural folds or sequences to the GSW25 proteins. The results presented hence indicate the generalization ability of the compact model. The other methods compared in the table, trained on large datasets that likely contain related sequences to the GSW25 dataset. Yet, the performance remained below 70%, with the exception of FLOPRED. It indicates that over training may have resulted in loss of ability to recognize structures, even though they had been including in training.

While here the heuristics based procedure to obtain the compact model has been used together with an FCRN employing the energy potentials based feature encoding,

other types of neural networks or feature encoding may, in theory be applied. However, the effect on the results could be different than what has been presented here. A fast or efficient classifier would enable the quick detection of a compact model. The current technique has several drawbacks, the improvement of which is the scope of future works. In particular, the feature encoding process is computationally intensive since a target sequence must be aligned across all templates and the reference energies, computed. The time taken is estimated to be 2 hours per 100 residues on a 2.3 GHz processor using 8G of RAM. The low accuracy of the Coil class (e.g. $MCC_C = 0.25$) leaves much room for improvement.

As seen from the class-wise prediction accuracies for Sheet and Coil classes in Table 5.7, in general they remain the main source of errors for the SS prediction problem regardless of type of classifier, training model or feature encoding used. A similar conclusion may also be drawn for the CB513 dataset. The remainder of this chapter, investigates the predictions of the compact model on the CB513 dataset with a goal towards identifying underlying structural reasons for the poor performance seen in the Coil class.

5.7 Structure-based insights from Analysis of Predictions

SCOP Analysis: To detect if a protein’s structural class has any effect on the predictive accuracy, the Q_3 scores on the SSP_{55} training proteins are compared with the remainder of the CB513 dataset. The compact model proteins and test proteins are grouped under their Structural Classification of Proteins (SCOP) classes [45] and the Q_3 averaged over the number of structures for each class is obtained. The CB513 proteins were first matched with sequences derived from ATOM records in their respective PDB files. The PDB identifiers were next used to identify structural domains from parse-able files of version 2.03 of the SCOPe database. For a majority of the proteins, SCOP domain sequences also directly matched the CB513 sequences; however, some entries could only obtain partial or gapped matches, most likely due to updated

versions of the domain definitions from the older PDB files. Nevertheless, as long as a partial match was obtained, the SCOP domain definition was taken as correct. At the end of this procedure, a total of 11 structures that had either missing SCOP domains or multiple domain matches were excluded from the analysis below. Table 5.8 gives the SCOP structural classes available. Figure 5.7 presents the Q_3 accuracies averaged

Table 5.8: List of SCOP Structural Classes

Class	Description
<i>a</i>	All alpha proteins
<i>b</i>	All beta proteins
<i>c</i>	Interspersed alpha and beta proteins (a/b)
<i>d</i>	Segregated alpha and beta proteins (a+b)
<i>e</i>	Multi domain proteins with known homologues
<i>f</i>	Membrane and cell surface proteins
<i>g</i>	small proteins
<i>h</i>	Coiled coil structures
<i>i</i>	Low resolution proteins
<i>j</i>	Peptides
<i>k</i>	Designed proteins

per SCOP structural group obtained by the compact model proteins (SSP_{55}). Also, the average Q_3 for the remainder of the CB513 dataset that served as test proteins is presented. It can be seen that overall, there is not much of a difference in the Q_3 obtained between the train and test tests. The four main structural classes *a* to *d*, have Q_3 scores ranging from about 80% (class *b*) to 84% (class *a*); the lowest performing class (excluding classes with absent structures) is *g*, small proteins with a low Q_3 of 74% averaged over 19 structures. Given that in earlier studies Sheet had proved to be the most difficult class to detect, the lowest performing class surprisingly turned out to be small proteins instead of classes with a high content of strand segments (e.g. beta proteins, *b*). It would be expected that, a classifier would show a higher accuracy for training proteins in the SSP_{55} compact model rather than the test proteins. However, the results indicate a different outcome. The largest difference between the average Q_3 for train and test sets, is rather small at only 2%. Additionally, two of the SCOP classes *b* and *e* had higher accuracies on the test set instead.

The results indicate that, information regarding the underlying structural folds, has been captured by the SSP_{55} compact model proteins and that they are able to gener-

alize to the remainder of the dataset. As for the low accuracy observed for class g at 74% and relative success for the all beta proteins (class b) at 80%, it is suggested that, some structural features inherent to certain protein classes, makes them difficult to predict, regardless if the protein has been learnt by the classifier, or not. For example while extended β -strand structures are found in larger proteins, they form a regular geometry that may make them easier to predict. In contrast, smaller proteins such as those in class g may contain single residue β bridges that may be mistaken as coils or loops.

Figure 5.7: Q_3 accuracies per SCOP class in CB513. The Q_3 for SCOP classes a-k is shown. Two types of Q_3 are presented below the classes. 1. Tests on the SSP₅₅ compact model proteins, which had been used in training (shaded bars). 2. Tests on the remainder of CB513 dataset NOT used in training (white bars). The Q_3 for SSP₅₅ is not necessarily higher than the remainder. A Q_3 of 0 does not indicate failure to predict the structure, but rather that no structures were found in that category (absent bar). For example, a value of 0.85 on the x-axis corresponds to $Q_3 = 85\%$. Excluding classes with absent structures, class g (small proteins) is the worst performing. The no. of structures present in each class is indicated above columns.

Case Study: The experiments in this chapter, have shown that most of the errors of the model are because of an inability to distinguish between Sheet and Coil residues and Helix and Coil residues, respectively. To investigate the structural reasons behind the poor performance of small proteins presented by the SCOP analysis earlier,

a case study of two inhibitors from this group is given. In particular, the hydrogen bonds that are difficult to detect by the current SS prediction methods resulting in Coil to Sheet mis-classification are described. Coil residues participating in hydrogen bonds with peptide backbone atoms are often wrongly predicted as Sheet while those participating in hydrogen bonds with waters were correctly classified as Coils. These conclusions were drawn by comparing two inhibitor peptides from class *g* - the trypsin inhibitor peptide (PDB: 1MCT) and cAMP dependent protein kinase (PDB 1ATP). The trypsin inhibitor contains 28 residues, of which 20 are considered as the predicted region after having excluded the first and last four residues. Although it is a small peptide, it is a highly interesting case that does not respond well to any of the current predictors regardless of feature encoding, underlying training model or type of classifier used.

The Q_3 does not exceed 40% using either the SSP_{55} compact mode or SSP_{CV} model; it also remains low even if included as a training protein. Additionally, none of the methods compared were able to obtain a Q_3 higher than 60%. The Q_3 accuracies obtained on this inhibitor, are PORTER (60%), PSIPRED (45%), PROTEUS 2 (45%), SSP_{55} (40%) and SSPRO (30%). Table 5.9 gives the inhibitor sequence, from which it can be seen that 70% of the observed residues are Coils (excluding the first and last four residues). Of these, more than 50% are wrongly predicted as Sheets; a similar trend was observed for predictions by other methods.

Table 5.9: Observed and Predicted SS in two Inhibitors by SSP_{55}

Trypsin Inhibitor, $Q_C = 42.8\%$	
AA	R I C P R I W M E C T R D S D C M A K C I C V A G H C G
OB	C C C C E C C C H H H C C C C C C E E C
PRED	E E E E E C C C C C C C C C C E E E E C E
Kinase Inhibitor, $Q_C = 100\%$	
AA	T T Y A D F I A S G R T G R R N A I H D
OB	H H H C C C C C C C C C
PRED	H H C C C C C C C C C C

Coil residues mis-predicted as Sheets are in bold.

The trypsin inhibitor was compared to the protein kinase inhibitor, a similar molecule that is 20 residues long (a predicted region of 12 residues is considered).

It contains approximately 75% Coils in the predicted region, but unlike the trypsin inhibitor, it makes mostly water based hydrogen bond contacts. SSP₅₅ predicts all Coils here correctly, with $Q_C = 100\%$. The Q_C of other methods were also discovered to be high, such as PORTER (100%), PSIPRED (88.9%), PROTEUS 2 (100%) and SSPPRO (88.9%). Table 5.9 gives the inhibitor sequences with predictions made by SSP₅₅; Coils wrongly predicted as Sheets are the residues in bold. Upon a comparison of the structures of both molecules, it was discovered that, the trypsin inhibitor made a high number of contacts with carbonyl oxygen (CO) and amide hydrogens (NH) of the peptide backbone, while the kinase inhibitor made predominantly hydrogen bonded contacts with water molecules. These findings are presented in Figure 5.8.

Figure 5.8: Case Study of Two Inhibitors. Detailed views of Coil prediction in inhibitors. **A**, Porcine trypsin inhibitor (PDB entry: 1MCT). **B**, cAMP dependent protein kinase inhibitor (PDB entry: 1ATP) with partially visible ATP in yellow. Correct predictions are in light purple and wrong predictions are in magenta. First and last four terminal residues are light brown and are not predicted. N marks the N-terminal. 1ATPI has more correct predictions than 1MCTI. Residues **RIWM** (5-8) and **KCI** (19-21) of 1MCTI are Coils wrongly predicted as Sheets. Residues **ASGRTGRRN** (8-16) of 1ATPI are correct Coil predictions. Waters are red and white sticks in **A** and red spheres in **B**. Putative hydrogen bonds (h-bonds) are indicated with dashed black lines, identified by inhibitor polar atom centres within 3.6Å of any O, N atoms. Italics denote the respective enzyme residues (green). The trypsin inhibitor residues make several h-bonds with peptide backbone O, N atoms and the kinase inhibitor, none. Examples in **A**, ARG5 CO with GLY193 NH; ILE6 NH with PHE41 CO. The kinase inhibitor prefers side-chain and water molecule contacts. Examples in **B**, SER9 N with ASP241 OD1; THR12 CO with ARG133 NH1; ARG14 CO with two waters. Not all h-bonds are shown; see text for more.

The detailed hydrogen based contacts listed below are inferred from PyMOL [151] distance based contacts to heavy atoms N and O. Some of these contacts have been illustrated as dashed black lines in Figure 5.8. Three-letter amino acid residue names in capital letters indicate the kinase and trypsin inhibitors; capitalized italics indicate the trypsin and protein kinase main chain enzymes that the respective inhibitors were complexed with. The numbers following the three letter names, correspond to residue numbers according to ATOM records in their PDB files, 1MCT and 1ATP, respectively. Observed hydrogen bonds of the peptide backbone for the two inhibitors are listed below.

- Trypsin inhibitor

- Carbonyl oxygen (CO) of ARG5 in bifurcated hydrogen bonds with the amide (NH) of *SER195* and *GLY193*
- NH of ARG5, hydrogen bonded with CO of *SER195*
- NH of TRP7 with CO of *PHE41*
- CO of MET8 with NH of CYS27
- NH of LYS19 with CO of ILE2;
- CO of ILE21 with NH of GLY28
- NH of CYS20 with CO of MET17 and so forth.
- CO of ILE6 in bifurcated hydrogen bonds with two waters
- CO of TRP7 with one water atom
- NH of MET8 with one water atom
- NH of MET17 with one water atom
- CO of CYS22 with one water atom

- Kinase inhibitor

- NH of SER13 with CO of PHE10
- SER9 CO, with water atom
- GLY10 CO, with water atom
- THR12 N with water atom

- ARG14 CO, with water atom
- ARG15 CO with water atom

From the list as well as Figure 5.8, it can be seen that the trypsin inhibitor Coil residues participate in a large number of hydrogen bonds involving the peptide main chain atoms, such as carbonyl oxygen (CO) from one residue to amide hydrogen (NH) in another residue. Conversely, the kinase inhibitor Coil residues form a majority of main chain hydrogen bond contacts to hetero atom water molecules rather than with other peptide main chain heavy atoms CO, or N. In fact, only one main chain to main chain hydrogen bond, NH of SER13 with CO of PHE10, was observed in the structure of 1ATP. From the findings of the case study above, it was identified that Coil residues preferring hydrogen bonds with water had better prediction rates, compared to those preferring peptide backbone atoms. Hence the effect of type of hydrogen bond on residue mis-prediction rates is further analysed by observing the hydrogen bond preferences across the entire CB513 dataset.

Effect of Hydrogen Bond Type on Residue Mis-prediction: First, the HBPLUS software [152] is used to detect all hydrogen bonds made by main chain atoms of residues in the 385 protein sequences from CB513. Of these, 9 sequences had to be discarded because PDB files matching their sequences could not be obtained, resulting in a total of 376 PDB files. According to HBPLUS, a hydrogen bond is formed between two atoms known as **Donor** and **Acceptor (DA)**, if certain chemical specifics, geometric angles as well as a maximum distance criteria are satisfied. Here, the distance cut-off was set at $3.6\text{\AA}$ and all other properties were kept at the default values. Following the notation of HBPLUS, main chain (M) to main chain (M) hydrogen bonds are denoted as MM and main chain to water hetero atoms (H) are denoted as MH. Here, the DA role is disregarded, as the direction of the bond (whether main chain atom is donor or acceptor) is not of interest. Using these settings, the MM and MH hydrogen bonds for a given residue, R_{ij} , was counted, where i denotes the true SS state and j is the predicted state under the SSP₅₅ compact model proteins. Table 5.10 presents the bond counts for Coil and Sheet residues in the dataset.

It can be observed from the table that, correctly predicted Coil residues (R_{CC}) have a larger proportion of hydrogen bonds with water atoms (58.7%) compared to those mis-predicted as Sheets at 53.0%. Conversely, the proportion of main-chain to main-chain (MM) bonds preferred by R_{CC} is 41.3%, compared to 47.0% in the case of R_{CE} . For the Sheet residues, the difference is even more significant. Correctly predicted Sheet residues R_{EE} , have 72.5% of their main chain atoms participating in hydrogen bonds with other main chain atoms (MM) and only 27.5% from hetero atom waters (MH). However for the Sheets predicted as Coils R_{EC} , MM contacts comprise only 63.3%. A higher number of atoms from these residues also participate in hydrogen bonds with water atoms (36.7%).

A clear picture emerges that, Coil residues participating heavily in main chain hydrogen bonding could be wrongly predicted as Sheets, since main chain hydrogen bonding is similar to the regular hydrogen bonds observed in secondary structures. Also, Sheet residues participating more in hydrogen bonds with waters, tend to be predicted as Coils instead. One reason could be that they are more exposed and tend to be flexible rather than following an ordered geometrical conformation typically observed in β -sheet or β -strand secondary structure elements. Here, the discussion has only been made with respect to main-chain atoms and water hetero atoms. However, other hydrogen bonds involving side chains, bound waters or ligand molecules could also greatly influence the molecular structure and the predicted results. To provide some context, the total no. of hydrogen bonds made by residues divided into their respective R_{ij} categories, is also provided in the table. The next section summarizes

Table 5.10: Detected hydrogen bonds of Sheet and Coil residues

	MM	MH	MM + MH	$\frac{MM}{(MM+MH)}$ (%)	$\frac{MH}{(MM+MH)}$ (%)	All	No. of Residues
R_{CC}	10345	14690	25035	41.3	58.7	78700	19182
R_{CE}	1685	1898	3583	47.0	53.0	10652	2584
R_{EC}	3972	2303	6275	63.3	36.7	17052	3193
R_{EE}	15143	5732	20875	72.5	27.5	51286	10370

Types of hydrogen bond contacts considered are from **M**ain-chain to **M**ain-chain (MM) atoms and **M**ain-chain to **H**etero-atom Water (MH) atoms. MM + MH is their sum. All indicates all hydrogen bonds including those involving side chains. R_{ij} denotes a residue in native state i predicted as j .

the findings presented in this chapter.

5.8 Summary

In this chapter, the idea of a compact model of training proteins to predict secondary protein structures, was explored, in conjunction with an energy potentials based feature encoding scheme to represent the protein sequences that are to be predicted. A heuristics based procedure was employed together with a Fully, Complex-valued Relaxation Network (FCRN) on the CB513 dataset to locate a small group of training proteins that was able to obtain good overall accuracy when tested on the remainder of the CB513 dataset. This small training set was termed the compact model SSP₅₅. When compared with a blind testing dataset of G Switch proteins (GSW25) the compact model was able to demonstrate a better performance than traditional cross-validated approaches, known as the SSP_{CV} model. Also, a comparison with several other existing secondary structure predictors in the literature indicated that the compact model obtained a better performance on the GSW25 dataset despite being trained on a much smaller number of proteins. The results suggest that information pertaining to protein folds has been captured adequately by the SSP₅₅ training while avoiding the problems of over training that can result in loss of performance when presented with new sequences. The compact model could be an efficient way to rapidly predict protein sequences, as theoretically only a small number of training proteins, have shown to be sufficient.

The predictions made by the compact model were also analysed from a structural perspective, from which it was discovered that some SCOP folds, like small proteins, had proteins that were inherently difficult to predict; even including them in the training did not improve the Q₃ accuracy to be predicted. A detailed analysis of hydrogen bond preferences in the mis-predicted Coil residues indicated that residues preferring main-chain to main-chain hydrogen bonding networks had a higher rate of being predicted as Sheets instead; this is likely because, Sheet residues have regular patterns of main-chain to main chain hydrogen bonding patterns. Conversely, Coil residues that participated in more hydrogen bonds with water atoms, were correctly predicted. In particular for the 2 structures analysed in the case studies, the Coil

residues participating in more hydrogen bonds with water molecules had consistently higher prediction rates for most methods tested. This indicates that the inability to accurately distinguish between Coil and Sheet is a difficult problem for the current SS predictors and is an area of improvement.

The future improvements of the compact model, are firstly, a need for a better algorithm in the encoding of the CABS algorithm based energy potentials which is a computationally intensive stage in the method. Secondly, in order to generalize the location of the compact model for any dataset in general, an automated procedure could be implemented.

Chapter 6

Conclusions and Future Works

The final chapter of this thesis summarizes the main findings from the protein secondary structure and membrane protein topology prediction experiments that were conducted. In the case of secondary structure prediction a hypothesis related to learning models was investigated. The hypothesis was that a small training set (known as the compact model) could be sufficient to capture the underlying structural relationship between train and testing proteins. The findings indicated that the hypothesis holds true. The resultant future directions of these works are also discussed.

6.1 Conclusions

The works proposed in this thesis had focused on the problem of predicting protein structures from sequences which is also known as the one-dimensional prediction from sequence. Due to the different structural characteristics and organization of globular and membrane proteins, secondary structures are predicted for globular proteins while residue preferences, transmembrane helix locations and topologies are predicted for α -transmembrane proteins. This is because for membrane proteins their placement within a cell membrane orders the residues in a direction dependant manner. Hence, the prediction of membrane protein residue preferences (as well as subsequent topology prediction) and protein secondary structure prediction from sequence are considered as separate but related problems.

6.1.1 Summary of Findings in Membrane protein Topology Prediction

In predicting membrane protein topology, a Hierarchical Deep Ensemble Network (HiDEN) was applied to (i) distinguish membrane protein sequences from globular proteins (i.e., protein group prediction) and (ii) to predict topologies of membrane proteins.

An input sequence was first represented as features generated from a Position Specific Scoring Matrix (PSSM) of dimensions $M \times (20 * w)$, where $w = 31$ and M denote the window size and number of residues in the sequence, respectively. HiDEN consists of two levels of Deep Ensemble Networks (DEN) combined hierarchically. The targets to be predicted were the membrane protein residue preferences (Membrane (M), Inside (I), Outside (O) and Signal peptides (S)), represented as integer class labels [1, 4] respectively.

In the first level, a Global DEN classifies proteins into their respective groups. In the second level, two Local DEN (LDEN) are applied to predict the topologies of transmembrane proteins (TM) and transmembrane proteins containing signal peptides (SP+TM), respectively. Each DEN combines the output of three individual Deep Belief Networks (DBN). In the unsupervised learning stage, the contrastive divergence learning algorithm [40] was used to extract higher level features from the primitive PSSM based features. During the supervised learning stage, backpropagation was used to update network weights, with a risk sensitive hinge-loss error algorithm [111] applied to enable faster network convergence instead of gradient descent based approach.

The results of applying HiDEN on four types of datasets transmembrane (TM), transmembrane with signal peptides (SP+TM), Globular (G) and Globular with Signal peptides (G+SP) were presented in chapter 3 and are summarized below.

HiDEN Performance:

- In the protein group classification, (i.e. to predict whether an input sequence belongs in TM, SP+TM, G or G+SP), HiDEN obtained an overall score of $\frac{6626}{6856} = 96.65\%$ across all groups, higher by 4.4% than the original method applied on the datasets (TOPCONS) [32].

- Most of the wrongly classified TM proteins (1.92%) were misclassified as SP+TM proteins due to false positive predictions of signal peptide regions. Conversely, a majority of the mis-classified SP+TM sequences (3.18%) were predicted to be TM sequences, having had their signal peptides mistaken for Inside and Outside regions (Results from Table 3.4).
- In topology detection HiDEN obtained 95% correct topologies on the TM dataset (containing 313 proteins) and 92% on the SP+TM dataset (containing 752 proteins) (Results from Table 3.5). These scores represent an overall of training and testing proteins.
- However, while the model obtained a very high topology score (95%) if overall scores are considered, a topology score averaged over ten-fold cross-validation is 63% on the TM dataset (Results from section 3.8.1).
- Currently, it is challenging for a single predictor to obtain topology scores greater than 75% on the TM and SP+TM datasets; the highest topology score reported was through the consensus method (TOPCONS) combining the predictions of five individual methods into a single topology profile [32]. The need for a higher topology score for test proteins is an area of improvement to be discussed in future works.

Comparison with TOPCONS Transmembrane Helices:

The predicted TMH of HiDEN and TOPCONS have been analysed with respect to the hydrophobicity of the helices, as measured by the Grand Averaged Hydropathy (GRAVY) Index [48]. It is challenging to detect weakly hydrophobic transmembrane helices (TMH) as they may be mistaken for less hydrophobic segments such as 'Outside' regions. For example, a few residues within the weakly hydrophobic TMH may be predicted as membrane residues (M), with the rest being classified as 'Inside' or 'Outside'. Since the number of predicted 'M' residues is important in detecting TMH, too few 'M' residues within a considered a window can lead to missed TMH.

The HiDEN model detected TMH with slightly weaker hydrophobicity compared to TOPCONS and the findings were statistically significant using a *t*-test on the TM

and SP+TM datasets. The results suggest the biological significance of the predicted TMH by the deep ensemble networks employed in HiDEN by highlighting the sensitivity of the model in detecting TMH. It further suggests that underlying residue interactions had been captured correctly. Next, the results of protein secondary structure prediction are summarized.

6.1.2 Summary of Findings in Protein Secondary Structure Prediction

Deep belief Networks (DBN) were also applied to predict protein secondary structures and the results were discussed in chapter 4. DBN were tested on two types of feature encoding: (i) features read in a sliding window from PSSM that contains evolutionary information in the form of multiple alignments and (ii) energy potentials based feature encoding derived from the CABS algorithm [41]. Employing the given features as input the contrastive divergence learning algorithm used in the DBN extracts higher level features, hence capturing the underlying residue interactions present. The overall accuracy (Q_3) was 81.59% for the energy based encoding, which was about 9% higher than that obtained for the PSSM (results from Table 4.3).

Comparison of PSSM and CABS energy potential based Feature Encoding:

It is challenging to obtain higher accuracies in the case of PSSM by using DBN, unlike in the prediction of membrane protein structures discussed earlier. As the PSSM possesses some limits arising from varied sequence characteristics being mapped onto similar values in the elements of the matrix, it remains difficult to capture using DBN even if additional hidden layers are used. Other works in the literature applying the PSSM based encoding incorporated additional sources of information with the PSSM or built second level structure prediction networks to obtain Q_3 accuracies of about 81% [27, 46].

However, the performance of the DBN at 82.6% (in a five-fold cross-validated study) using the CABS algorithm motivates the choice of CABS algorithm based feature encoding to be used in further studies of secondary structure predictions. Although

a drawback of the CABS algorithm feature encoding is the difficulty of generating the features due to being computationally resource intensive, the features proved successful in the prediction of secondary structures. Hence, the CABS algorithm based feature encoding is explored with a different type of neural network classifier and under varied training conditions in this thesis.

Motivation for Development of the Compact Model:

The theoretical upper limit for the SS prediction problem had been estimated at 88-90% [2,22]. As this accuracy remains a challenge, it would be advantageous to explore other models as various DBN architectures had already been implemented by existing works [27,46].

The current SS predictors trained their models on thousands of protein chains using cross-validation. However, in the machine-learning field there remains a trade-off between training and generalization; too many samples intended to increase learning may lead to over training, reducing the sensitivity of the classifier to new test proteins. Due to the potential challenges of over training associated with the cross-validation approach, a hypothesis that a small training set is sufficient to achieve a good performance was investigated. In order to test the hypothesis, a heuristics based procedure was developed to detect a small set of training proteins known as the compact model. The compact model is defined as a small group of training proteins that gives rise to comparable accuracies to cross-validated models on the training dataset, yet maintains a good generalization to proteins from new datasets. The development of the compact model was motivated by the search for select proteins that share special structural characteristics to other proteins in the dataset, enabling a classifier to maintain the good generalization performance to a blind test dataset.

Previous studies had shown Complex-valued neurons to be more computationally powerful than Real-valued neurons, which is a motivation to the Complex-valued neural networks in the problem of protein secondary structure prediction [43,44]. A Fully Complex-valued Relaxation Network, (FCRN) [47] was used to develop the compact model as it could be rapidly trained due to its analytical weight update algorithm that allows rapid prediction of secondary structures. Next the compact model was tested

with the FCRN to establish if the generalization capability of the classifier could be maintained.

The performance results of the FCRN under various training conditions were presented in chapter 5 and the findings are summarized below.

- Cross-validation performance studies conducted on the CB513 [38] dataset indicated that the FCRN obtained about 82%, higher than most other classifiers in the literature, with the exception of the Compound Pyramid Model (CPM) [3]. However, the latter has high computational complexity due to a cascaded architecture of several levels of Support Vector Machines combined with neural networks (results from Tables 5.3 and 5.4).
- The compact model containing 55 proteins (SSP_{55}) was obtained and training the FCRN on SSP_{55} led to a much higher performance compared to cross-validated models as well as other predictors on a blind test dataset (Results from Table 5.7).
- The results indicate that selecting training samples carefully has the potential to preserve the generalization ability of the classifier and hints at special underlying structural properties relating to the train and test proteins. The implications are that they have been captured in the model.

Findings from Structure-based Analysis:

A breakdown of Q_3 obtained under the compact model's training by SCOP classes identified some poorly performing classes. In some cases even including the proteins from the poorly performing classes to the training set did not help to increase the training accuracy. Certain proteins displayed a low Q_3 when tested on existing predictors in the literature that had used different training models and feature encoding schemes from those reported in this thesis, even if the proteins were older sequences that had already been included in the training models. The findings indicate that such cases of proteins present the current limits of the secondary structure prediction capabilities.

Subsequently, detailed case studies mapping the predicted secondary structures onto the actual structures, identified the role of hydrogen bonds in Sheet $\leftrightarrow$ Coil mis-prediction rates, with a finding that

- Coil residues with hydrogen bonding involving the peptide backbone had a higher propensity to be mistaken as Sheet residues, whereas those that had predominantly hydrogen bonded with free waters had a higher propensity to be correctly classified as coils (Results from Figure 5.8)
- Sheet residues that had been correctly predicted, had a higher proportion of hydrogen bonds involving the peptide backbone compared to Sheets that had been erroneously predicted as Helix residues (Results from Figure 5.8)

A further study of Sheet $\leftrightarrow$ Coil mis-classifications conducted on the whole CB513 dataset had similar findings to those observed in the case studies. Most of the erroneous predictions in secondary structures were Sheets mistaken for Coils (or vice versa) particularly with the Sheet class being the hardest to detect out of the three secondary states [5]. A similar difficulty was also observed in FCRN predictions on the secondary structure prediction datasets even though a different feature encoding had been used compared to the PSSM normally used for this problem.

In light of the findings reported in this thesis, the current secondary structure prediction methods may not be sensitive in distinguishing Sheet or Coil residues. Sheet residues with a high number of hydrogen bonded contacts to water molecules and Coil residues with stabilizing hydrogen bonds to peptide backbone atoms (amide nitrogen or carbonyl carbon) are particularly prone to mis-classification. This highlights the need for more methods that can accurately detect Sheet from Coil residues and is a future work resulting from these findings.

6.2 Future Works

In this thesis, several models were explored and applied to predict one-dimensional protein structures from sequence. Specifically, the problems of predicting membrane

protein residue preferences and topology as well as the prediction of secondary structures were addressed. As discussed in the previous chapters, the one-dimensional structural features have applications in both computational and experimental determination of structures and in the prediction of other residue properties from the sequence. The findings from the models applied in this work may be extended in several directions and form the scope of future works.

6.2.1 Improvements to Sequence representation and Feature Encoding

Two types of feature encoding had been employed which were (i) Position Specific Scoring Matrix (PSSM) derived from PSI-BLAST [50] and (ii) energy potentials based encoding derived from the CABS algorithm [41, 42].

Evaluation of PSSM and CABS algorithm based features:

Currently the CABS algorithm based feature encoding for protein secondary structures demonstrated higher accuracies than the PSSM conventionally employed in the datasets tested. In particular the force-field based model provides indirect structural information in the form of context-dependent energy potentials that are calculated during the target to template alignment step. In chapter 4, DBN were tested on both PSSM based and CABS energy potential based encoding with the latter having a much higher Q_3 of 81.59% compared to 72.6% obtained by PSSM. The energy potentials based encoding contributed to SSP₅₅ (or compact model's) favourable results on the blind test dataset while other predictors based on the PSSM could not achieve similar levels of performance despite having been trained on much larger numbers of proteins.

Drawbacks of using CABS algorithm based feature encoding:

However, the energy potentials based feature representation is computationally complex. The process needs the alignments of target sequences to template structures, following which reference energies must be computed in a 17-residue window. Within this 17 residue window probabilities for the 9 central residues adopting each of Helix, Sheet or Coil states, $P(H)$, $P(E)$ and $P(C)$ must also be calculated with scoring

functions. The FCRN classification is rapid due to the analytically computed weight matrix but generation of the CABS-based features proves computationally intensive.

Evaluation of Compact Model approach:

The search for compact model proteins could be implemented with the PSSM or other rapidly generated feature encodings such as Markov based models to represent sequences. The combinations of other types of feature encoding in conjunction with the heuristics based procedure currently remain unexplored. At present the compact model has only been demonstrated with the CABS algorithm based feature encoding. If the compact model is shown to work on other types of encoding, it indicates a finding that can be generalized.

One of the advantages of the compact model is that it can predict Sheet residues better than existing approaches. Improvements in the prediction of the Sheet class were seen using the compact model in the G Switch Proteins (GSW25) dataset [42] compared to a cross-validated approach. However the compact model's performance on the CB513 dataset shows that the accuracy of correctly predicted Sheets is still the lowest of the three classes, suggesting the challenge of capturing long-range interactions of the peptide backbone atoms.

Finally, the most interesting aspect of the compact model is that it presents a small group of training proteins whose folds are able to capture a wide range of structural properties effectively. A deeper investigation of such properties could uncover previously unknown structural relationships with implications for protein structure prediction and template recognition algorithms.

Inclusion of more Sequence Specific Features together with PSSM:

A similar difference in class-wise accuracy is observed in the prediction of membrane protein residue preferences using the HiDEN model. Although the ten-fold cross-validated topology score of HiDEN was lower than that reported by TOPCONS [32] HiDEN had a much higher topology score if an overall performance (including training and testing proteins) was taken into account. As TOPCONS' score itself was not cross-validated and was the consensus of several predictors that were highly likely to have been trained on proteins similar to the TOPCONS datasets, in the analysis of

predicted TM helices given below the overall prediction results had been considered. The comparison of predicted transmembrane helices' hydropathy values with TOPCONS has indicated that HiDEN has successfully captured slightly less hydrophobic helices. It strongly suggests that the underlying correlations between raw PSSM features had successfully been extracted during the layer by layer learning procedure employed during deep training. However, the Outside (O) residues had often been mistaken for Inside, resulting in the lowest accuracies being reported for that class. This also likely contributes to the low topology score; even though most TM helices had been detected correctly the N and C termini locations were likely identified wrongly.

Besides a deeper analysis that clearly identifies structural factors contributing to wrong topologies, the model could benefit from additional sources of information together with the PSSM to be included in the training matrices that serve as input to HiDEN. The HiDEN model used could be developed further by training on more sequence properties, such as the predicted solvent accessible surface area, residue molecular weight or chemical composition and so forth.

Membrane residue environment preferences: Redefining structures included in Class Labels:

Another issue in the prediction of membrane residue preferences is that currently the annotation of the class labels M, I, O and S is generic and contains many different structural features collapsed under any one class. Many fine grained structural features such as re-entrant helices, kinks would all be grouped together with transmembrane helices under the category Membrane. However, the prediction of these structures at a finer resolution has several biological applications. It would be valuable to search for new groupings for these structures in the annotation of residue preferences (or labels), in order to provide a finer grained prediction besides the three or four state membrane residue preferences.

Moreover other structural features such as loops and external helices would be reduced into one category, 'Outside'. This motivates for existing class labels in membrane protein residue preferences to be re-defined with more precise definitions.

6.2.2 Ensemble DBN to Predict Secondary Structures

It remains a challenge to obtain Q_3 accuracies higher than 82% using PSSM derived features for the secondary structure prediction problem for datasets other than CB513. Using a single deep belief network containing a single RBM, it was difficult to obtain further improvements to testing accuracy beyond 72 or 73 % by adjusting parameters like the learning rate, or number of hidden layers. Increasing the number of hidden layers had decreased prediction accuracy. As the PSSM is a widely available feature encoding to represent protein sequences, methods that are able to raise Q_3 using PSSM may be very desirable.

It is proposed to build an ensemble of three DBN to strengthen predictions with each individual DBN to have a similar architecture to the single DBN that has been presented earlier in the thesis. Previous experiments have indicated that an individual DBN may perform very well on one class but fail to discriminate the other two successfully. Best performing DBN for each of classes Helix (H), Sheet (E) and Coil (C) could be selected and their combined output used to determine the final output class. Another potential strategy is to treat the three-category classification problem as three separate binary classification tasks, with each DBN predicting $(H|\overline{H})$, $(E|\overline{E})$ or $(C|\overline{C})$, respectively. The DBN specialized for each class may be better able to detect residues belonging in each class for an improved prediction. Finally, the ensemble DBN could be combined in a second level structure to structure prediction network as is commonly done [36, 37, 46] to raise the accuracy further.

6.2.3 Automated Procedure to Detect Compact Model

A heuristics based procedure for detecting a good set of training proteins was proposed. In the future, a process to derive a compact model could be automated making it generic and applicable to any training dataset. Although such a work would involve high computational complexity, such as needing to locate a minimal set of proteins (or derive an estimate close to the minimal set), it would be an exciting work.

The principle behind the compact model is the selection of the most informative protein chains that contribute to learning as a training set. This idea could be imple-

mented more systematically by combining with principles of Meta-cognitive learning. Meta-cognitive learning regulates aspects of learning such as what to learn and when to learn [119]. Examples of classifiers employing meta-cognition in the literature are the Sequential Resource Allocation Network (SRAN) [153] and Projection based, Meta-Cognitive learning Radial Basis Function Network (PBL-McRBFN) [116]. The PBL-McRBFN has been currently applied to the prediction of membrane protein residue preferences in chapter 3 and its application could be extended to detect a better set of training proteins in the secondary structure prediction problem. Additionally the heuristics based procedure at present has only been implemented in conjunction with the CABS algorithm based feature encoding (see results in chapter 5). It could be integrated with the PSSM based features making it a general model applicable to a wider range of algorithms.

6.2.4 Relation of Sequence to Structure Space

The findings of the heuristics based approach indicate that there is a special relationship between the training proteins and the testing proteins, even if they are from independent datasets or if sequence or structural homologies had been removed. Previous work has shown that fold space is much smaller than sequence space with one protein structure being related to another within 7 transitive steps if certain constraints relating to domain and chain lengths are satisfied [35]. The level of similarity of structural folds between the compact model proteins and the remainder of the CB513 datasets is low; for the example TM-align score [154] that quantifies related structural folds is lower than 0.5 between compact model proteins and the remainder of the dataset. A score of 0.5 and above indicates related structural folds. It is difficult to meaningfully quantify the structural relationships between the compact model training proteins and the remainder of the CB513 dataset. Yet, underlying structural relationships that transcend current known measures of sequence similarities or fold based similarities must be present or the model could not have efficiently captured a high predictive accuracy on the remainder of the dataset as well as the G Switch proteins. A deeper investigation could have more interesting findings and remains the

scope of future works.

Appendix A

Twenty Amino Acids

Nonpolar, aliphatic side groups

Polar, uncharged side groups

Aromatic side groups

Positively charged side groups

Negatively charged side groups

Figure A.1: The 20 standard amino acids. The amino acids are presented with their side chains in grey rectangles. They are organized into five groups according to the chemical properties of their side chains which give different properties. The standardized three-letter and one-letter abbreviations separated by a comma are given below the names of each amino acid. A protein sequence 'APGDAG' indicates that the amino acids Alanine, Proline, Glycine, Aspartate, Alanine and Glycine have been bound together with peptide bonds. Amino acids in a protein chain are often called amino acid residues or simply 'residues.' All three terms have been used interchangeably in the report. The figure was downloaded from [155].

Appendix B

Protein Information

No.	Sequence
G_A 3α fold	
1	TTYKLILNLK QAKEEEAIKEL VDAGIAEKYI KLIANAKTVE GVWTLKDEIL KFTVTE
2	TTYKLILNLK QAKEEEAIKEL VDA AIAEKYI KLIANAKTVE GVWTLKDEIL KATVTE
3	TTYKLILNLK QAKEEEAIKEL VDAGTAEKYI KLIANAKTVE GVWYKDEIL KATVTE
4	TTYKLILNLK QAKEEEAIKEL VDAGIAEKYF KLIANAKTVE GVWTLKDEIL KATVTE
5	TTYKLILNLK QAKEEEAIKEL VDAGIAEKYI KLIANAKTVE GVWTLKDEIL TATVTE
6	TTYKLILNLK QAKEEEAIKEL VDAGIAEKYI KLIANAKTVE GVWYKDEIL KATVTE
7	TTYKLILNLK QAKEEEAIKEL VDAGIAEKYI KLIANAKTVE GVWTLKDEIK KATVTE
8	TTYKLILNLK QAKEEEAIKEL VDAGIAEKYI KLIANAKTVE GVWTLKDEIL KATVTE
9	TTYKLILNLK QAKEEEAIKEL VDAGIAEKYI KLIANAKTVE GVWTLKDEIL TFTVTE
10	TTYKLILNLK QAKEEEAIKEL VDAGTAEKYI KLIANAKTVE GVWTLKDEIL TFTVTE
11	TTYKLILNLK QAKEEEAIKEL VDAGTAEKYI KLIANAKTVE GVWTLKDEIK TFTVTE
12	TTYKLILNLK QAKEEEAIKEL VDAGTAEKYF KLIANAKTVE GVWTLKDEIK TFTVTE
G_B $4\beta + \alpha$ fold	
13	TTYKLILNGK QLKEEAITEA VDAATAEKYF KLIANAKTVE GVWYKDETK TFTVTE
14	TTYKLILNGK QLKEEAITEA VDAGTAEKYF KLIANAKTVE GVWYKDETK TFTVTE
15	TTYKLILNLK QAKEEAITEA VDAATAEKYF KLIANAKTVE GVWYKDETK TFTVTE
16	TTYKLILNGK QLKEEAIKEA VDAATAEKYF KLIANAKTVE GVWYKDETK TFTVTE
17	TTYKLILNGK QLKEEAITEA VDA AIAEKYF KLIANAKTVE GVWYKDETK TFTVTE
18	TTYKLILNGK QLKEEAITEA VDAATAEKYF KLIANAKTVE GVWYKDEIK TFTVTE
19	TTYKLILNGK QLKEEAITEA VDAATAEKYF KLIANAKTVE GVWYKDETK TFTVTE
20	TTYKLILNGK QLKEEAITEA VDAATAEKYF KLIANAKTVE GVWYKDETK TFTVTE
21	TTYKLILNLK QAKEEEAIKEL VDAATAEKYF KLIANAKTVE GVWYKDETK TFTVTE
22	TTYKLILNLK QAKEEAITEA VDAGTAEKYF KLIANAKTVE GVWYKDEIK TFTVTE
23	TTYKLILNLK QAKEEAIKEA VDAGTAEKYF KLIANAKTVE GVWYKDEIK TFTVTE
24	TTYKLILNLK QAKEEAIKEA VDAGTAEKYF KLIANAKTVE GVWYKDEIK TFTVTE
25	TTYKLILNLK QAKEEEAIKEL VDAGTAEKYF KLIANAKTVE GVWYKDEIK TFTVTE

Table B.1: GSW25 sequences

Protein	PDB	CB513 ID	SCOP Class	Fold	Residues	Q3 (%)
CELLULOSE CELA (1,4-BETA-D-GLUCAN-GLUCANOHYDROLASE)	1cemA	1cem-1-GJB	a.102.1.2	alpha/alpha toroid	363	85
INTERLEUKIN-2	3inkD	3inkd-1-DOMAK	a.26.1.2	4-helical cytokines	122	84
INTERLEUKIN-10	1ilkA	1ilk-1-AS	a.26.1.3	4-helical cytokines	106	88
RECOVERIN	1recA	1rec-1-DOMAK	a.39.1.5	EF Hand-like	83	84
TAQ DNA POLYMERASE	1taqA	1taq-2-AS	a.60.7.1	SAM domain-like	69	82
ANNEXIN V	1avhB	1avhb-4-AS	a.65.1.1	Annexin	74	77
GLUTAMYL-TRNA SYNTHETASE	1gnA	1gn-4-AS	a.97.1.1	An anticodon-binding domain of class I aminoacyl-tRNA synthetases	98	86
RIBONUCLEOTIDE REDUCTASE PROTEIN R1	1rhA	1rh-1-JAC	a.98.1.1	R1 subunit of ribonucleotide reductase, N-terminal domain	211	88
CD2	1hufA	1huf-1-AS	b.1.1.1	Immunoglobulin-like beta-sandwich	101	69
T CELL SURFACE GLYCOPROTEIN CD4	3cd4A	3cd4	b.1.1.3	Immunoglobulin-like beta-sandwich	178	74
HUMAN VASCULAR CELL ADHESION MOLECULE-1	1vcaB	1vcab-2-AUTO.1	b.1.1.3	Immunoglobulin-like beta-sandwich	110	75
CYCLODEXTRIN GLYCOSYL-TRANSFERASE	1eguA	1egu-3-GJB	b.1.18.2	Immunoglobulin-like beta-sandwich	84	83
HUMAN TISSUE FACTOR	2hftA	2hft-1-AS	b.1.2.1	Immunoglobulin-like beta-sandwich	102	82
HUMAN RHINOVIRUS 14 COAT PROTEIN (SUBUNIT VP4)	4rhv4	4rhv4	b.121.4.1	Nucleoplasmin-like/VP	40	81
HUMAN RHINOVIRUS 14 COAT PROTEIN (SUBUNIT VP3)	4rhv3	4rhv3	b.121.4.1	Nucleoplasmin-like/VP	236	82
HEMAGGLUTININ	3hmgA	3hmgA	b.19.1.2	Viral protein domain	328	78
XYLANASE	1bcxA	1bcx-1-DOMAK	b.29.1.11	Concanavalin A-like lectins/glucanases	185	82
THERMONUCLEASE PRECURSOR	2snsA	2sns	b.40.1.1	OB-fold	141	74
TOXIC SHOCK SYNDROME TOXIN-1	2tsbB	1tsb-2-DOMAK	b.40.2.2	OB-fold	73	86
ALPHA-LYTIC PROTEASE	2alpA	2alp	b.47.1.1	Trypsin-like serine proteases	198	83
TENDAMISTAT	3aitA	3ait	b.5.1.1	Alphac-Amylase inhibitor tendamistat	74	79
NITRITE REDUCTASE	2afnC	2afnc-1-AUTO.1	b.6.1.3	Cupredoxin-like	149	79
METHYLAMINE DEHYDROGENASE (HEAVY SUBUNIT)	1mdaJ	1mdaj-1-GJB	b.69.2.1	7-bladed beta-propeller	282	76
ALANINE RACEMASE	1sftB	1sftb-2-AS	c.1.6.1	TIM beta/alpha-barrel	230	86
T4 RNASE H	1ftfA	1ftf-1-GJB	c.120.1.2	PIN domain-like	283	85
17-BETA-HYDROXYSTEROID-DEHYDROGENASE	1fdtA	1fdt-1-AS	c.2.1.2	NAD(P)-binding Rossmann-fold domains	285	87
SUCCINYL-COA SYNTHETASE BETA SUBUNIT	1scuE	1scue-3-AS	c.23.4.1	Flavodoxin-like	149	88
FERRDOXIN-NADP+ REDUCTASE	1fndA	1fnd	c.25.1.1	Ferredoxin reductase-like, C-terminal NADP-linked domain	296	86
PYRUVATE OXIDASE	1powB	1powb-4-DOMAK	c.36.1.9	Thiamin diphosphate-binding fold (THDP-binding)	44	81
THIOLTRANSFERASE	1kteA	1kte-1-AS	c.47.1.1	Thioredoxin fold	105	86
RESTRICTION ENDONUCLEASE	1cfrA	1cfr-1-GJB	c.52.1.7	Restriction endonuclease-like	283	85
HEAT-SHOCK COGNATE 70 kD PROTEIN	1ngaA	1nga-2-AS.1	c.55.1.1	Ribonuclease H-like motif	190	85
LIPASE	1trhA	1trh-1-AS	c.69.1.17	alpha/beta-Hydrolases	534	81
ALPHA-D-GLUCOSE-1,6-BISPHOSPHATE	3pngB	3pngb-1-AS	c.84.1.1	Phosphoglucosyltransferase, first 3 domains	188	83
PHOSPHOGLYCERATE KINASE	3pgkA	3pgk-2-AS	c.86.1.1	Phosphoglycerate kinase	210	83
PHOSPHOFRUCTOKINASE	4pfkA	4pfk	c.89.1.1	Phosphofructokinase	319	85
CYTIDINE DEAMINASE	1ctuA	1ctu-1-AUTO.1	c.97.1.1	Cytidine deaminase-like	164	76
CYTIDINE DEAMINASE	1ctuA	1ctu-2-AUTO.1	c.97.1.1	Cytidine deaminase-like	130	82
RIBONUCLEASE T1	3ntA	3nt	d.1.1.4	Microbial ribonucleases	104	78
PROFILIN	1cqaA	1cqa-1-AUTO.1	d.110.1.1	Profilin-like	123	84
MALATE DEHYDROGENASE	2cmdA	2cmd-2-GJB	d.162.1.1	LDHC-terminal domain-like	166	86
STEFIN B (CYSTATIN B)	1stfI	1stf-1-DOMAK	d.17.1.2	Cystatin-like	98	79
SUBTILISIN BPN' PROSEGMENT	1spbP	1spb-1-AS	d.58.3.2	Ferredoxin-like	71	75
SONIC HEDGEHOG	1vhA	1vhb-1-AS	d.65.1.2	Hedgehog/DD-peptidase	157	84
PROTEIN-TYROSINE PHOSPHATASE SYP (N-TERMINAL SH2 DOMAIN)	1ayaB	1ayab-1-GJB	d.93.1.1	SH2-like	101	78
HISTIDINE-CONTAINING PHOSPHOCARRIER PROTEIN HPR	2hprA	2hpr-1-DOMAK	d.94.1.1	HPr-like	87	86
D-AMINO ACID AMINOTRANSFERASE	1daaB	1daab-1-AS	e.17.1.1	D-amino acid aminotransferase-like PLP-dependent enzymes	119	76
BETA-LACTAMASE	3blmA	3blm	e.3.1.1	beta-lactamase/transpeptidase-like	257	80
MEDIUM CHAIN ACYL-COA DEHYDROGENASE	3mddb	3mddb-1-AS	e.6.1.1	Acyl-CoA dehydrogenase NM domain-like	120	87
GRAMBIN	1crnA	1crn	g.13.1.1	Crambin-like	46	84
GLUCOCORTICOID RECEPTOR	1latB	1lab-1-AUTO.1	g.39.1.2	Glucocorticoid receptor-like (DNA-binding domain)	74	76
RUBREDOXIN	4rxnA	4rxn	g.41.5.1	Rubredoxin-like	54	54
NEUROTOXIN I	1sh1A	1sh1	g.9.1.1	Defensin-like	48	88
COIL-VALD	1coiA	1coi-1-AS	k.40.1.1	Designed trimeric coiled-coil	29	100
HOLO-D-GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE	1gd1O	1gd1o	Multiple SCOP	Multiple SCOP	334	81

Table B.2: SSP₅₅ compact model proteins and their properties

Appendix C

Author's Publications

- S. Rashid, S. Saraswathi, A. Kloczkowski, S. Sundaram and A. Kolinski, "Protein secondary structure prediction using a small training set (compact model) combined with a Complex-valued neural network approach," *BMC Bioinformatics* vol. 17, no. 1, p.362, 2016.
- R. Shamima and S. Suresh, "Prediction of membrane protein structures using a Projection based Meta-cognitive Radial Basis Function Network," *IEEE International Joint Conference on Neural Networks (IJCNN)*, Jul. 2016, pp. 1229-1235.
- B. Shamima, R. Savitha, S. Suresh and S. Saraswathi, Protein secondary structure prediction using a fully complex-valued relaxation network," *IEEE International Joint Conference on Neural Networks (IJCNN)*, Aug. 2013, pp. 1-8.
- J. Abeegithan, R. Shamima, S. Suresh and D. Shirin, "Hierarchical approach to separate membrane, non-membrane, signal peptide and to improve prediction of transmembrane protein topology," *submitted to IEEE Transactions on Computational Biology and Bioinformatics*, Jan 2018

Bibliography

- [1] D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, 6th ed. W. H. Freeman and Company, 2013.
- [2] D. Kihara, “The effect of long-range interactions on the secondary structure formation of proteins,” *Protein Science*, vol. 14, no. 8, pp. 1955–1963, Aug. 2005.
- [3] B. Yang, Q. Wu, Z. Ying, and H. Sui, “Predicting protein secondary structure using a mixed-modal SVM method in a compound pyramid model,” *Knowledge-Based Systems*, vol. 24, no. 2, pp. 304–313, Mar. 2011.
- [4] O. Dor and Y. Zhou, “Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training,” *Proteins*, vol. 66, no. 4, pp. 838–845, Mar. 2007.
- [5] H. Zhang, T. Zhang, K. Chen, K. D. Kedariseti, M. J. Mizianty, Q. Bao, W. Stach, and L. Kurgan, “Critical assessment of high-throughput standalone methods for secondary structure prediction,” *Briefings in Bioinformatics*, vol. 12, no. 6, pp. 672–688, Nov. 2011.
- [6] L. Kurgan and F. M. Disfani, “Structural protein descriptors in 1-dimension and their sequence-based predictions,” *Current Protein & Peptide Science*, vol. 12, no. 6, pp. 470–489, Sep. 2011.
- [7] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, “The protein data bank,” *Nucleic acids research*, vol. 28, pp. 235–242, Jan. 2000.

- [8] M. A. Yildirim, K.-I. Goh, M. E. Cusick, A.-L. Barabási, and M. Vidal, “Drug-target network,” *Nature Biotechnology*, vol. 25, no. 10, pp. 1119–1126, Oct. 2007.
- [9] H. Viklund and A. Elofsson, “OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar,” *Bioinformatics (Oxford, England)*, vol. 24, no. 15, pp. 1662–1668, Aug. 2008.
- [10] D. Kozma, I. Simon, and G. E. Tusndy, “PDBTM: Protein Data Bank of trans-membrane proteins after 8 years,” *Nucleic Acids Research*, vol. 41, no. Database issue, pp. D524–529, Jan. 2013.
- [11] T. Schwede, “Protein Modeling: What Happened to the Protein Structure Gap?” *Structure*, vol. 21, no. 9, pp. 1531–1540, Sep. 2013.
- [12] K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott, “NCBI reference sequences (RefSeq): current status, new features and genome annotation policy,” *Nucleic Acids Research*, vol. 40, no. D1, pp. D130–D135, Nov. 2011.
- [13] C. B. Anfinsen, “Principles that govern the folding of protein chains,” *Science (New York, N.Y.)*, vol. 181, no. 4096, pp. 223–230, Jul. 1973.
- [14] H. Zhang, T. Zhang, K. Chen, S. Shen, J. Ruan, and L. Kurgan, “Sequence based residue depth prediction using evolutionary information and predicted secondary structure,” *BMC Bioinformatics*, vol. 9, p. 388, Sep. 2008.
- [15] V. Pulim, J. Bienkowska, and B. Berger, “LTHREADER: prediction of extracellular ligand-receptor interactions in cytokines using localized threading,” *Protein Science*, vol. 17, no. 2, pp. 279–292, Feb. 2008.
- [16] J. D. Fischer, C. E. Mayer, and J. Sding, “Prediction of protein functional residues from sequence by probability density estimation,” *Bioinformatics*, vol. 24, no. 5, pp. 613–620, Mar. 2008.

- [17] J. Pei and N. V. Grishin, “PROMALS: towards accurate multiple sequence alignments of distantly related proteins,” *Bioinformatics*, vol. 23, no. 7, pp. 802–808, Apr. 2007.
- [18] F. Cicconardi, D. Di Marino, P. P. Olimpieri, W. Arthofer, B. C. Schlick-Steiner, and F. M. Steiner, “Chemosensory adaptations of the mountain fly *Drosophila nigrosparsa* (Insecta: Diptera) through genomics’ and structural biology’s lenses,” *Sci Rep*, vol. 7, p. 43770, Mar. 2017.
- [19] B. E. Weiner, N. Woetzel, M. Karakas, N. Alexander, and J. Meiler, “BCL::MP-fold: folding membrane proteins through assembly of transmembrane helices,” *Structure (London, England: 1993)*, vol. 21, no. 7, pp. 1107–1117, Jul. 2013.
- [20] W. Pirovano and J. Heringa, “Protein secondary structure prediction,” in *Data Mining Techniques for the Life Sciences*, ser. Methods in Molecular Biology, O. Carugo and F. Eisenhaber, Eds. Humana Press, Jan. 2010, no. 609, pp. 327–348.
- [21] J. Yan, M. Marcus, and L. Kurgan, “Comprehensively designed consensus of standalone secondary structure predictors improves Q3 by over 3%,” *Journal of Biomolecular Structure & Dynamics*, vol. 32, no. 1, pp. 36–51, 2014.
- [22] B. Rost, C. Sander, and R. Schneider, “Redefining the goals of protein secondary structure prediction,” *Journal of Molecular Biology*, vol. 235, no. 1, pp. 13–26, Jan. 1994.
- [23] W. Kabsch and C. Sander, “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features,” *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.
- [24] J. Cheng, A. Z. Randall, M. J. Sweredoski, and P. Baldi, “SCRATCH: a protein structure and structural feature prediction server,” *Nucleic Acids Research*, vol. 33, no. suppl 2, pp. W72–W76, Jul. 2005.

- [25] S. Montgomerie, S. Sundararaj, W. J. Gallin, and D. S. Wishart, “Improving the accuracy of protein secondary structure prediction using structural alignment,” *BMC Bioinformatics*, vol. 7, p. 301, 2006.
- [26] G. Pollastri, A. J. Martin, C. Mooney, and A. Vullo, “Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information,” *BMC Bioinformatics*, vol. 8, no. 1, p. 201, Jun. 2007.
- [27] 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.
- [28] J. Koehler Leman, M. B. Ulmschneider, and J. J. Gray, “Computational modeling of membrane proteins,” *Proteins: Structure, Function, and Bioinformatics*, vol. 83, no. 1, pp. 1–24, Jan. 2015.
- [29] T. Nugent and D. T. Jones, “Transmembrane protein topology prediction using support vector machines,” *BMC Bioinformatics*, vol. 10, no. 1, p. 159, 2009.
- [30] G. E. Tusndy, Z. Dosztanyi, and I. Simon, “Transmembrane proteins in the Protein Data Bank: identification and classification,” *Bioinformatics (Oxford, England)*, vol. 20, no. 17, pp. 2964–2972, Nov. 2004.
- [31] G. von Heijne, “Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule,” *Journal of Molecular Biology*, vol. 225, no. 2, pp. 487–494, May 1992.
- [32] K. D. Tsirigos, C. Peters, N. Shu, L. Kll, and A. Elofsson, “The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides,” *Nucleic Acids Research*, pp. W401–407, May 2015.
- [33] M. Bernhofer, E. Kloppmann, J. Reeb, and B. Rost, “TMSEG: Novel prediction of transmembrane helices,” *Proteins*, vol. 84, no. 11, pp. 1706–1716, Nov. 2016.

- [34] C. Bystroff, K. T. Simons, K. F. Han, and D. Baker, “Local sequence-structure correlations in proteins,” *Current Opinion in Biotechnology*, vol. 7, no. 4, pp. 417–421, Aug. 1996.
- [35] J. Skolnick, A. K. Arakaki, S. Y. Lee, and M. Brylinski, “The continuity of protein structure space is an intrinsic property of proteins,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 37, pp. 15 690–15 695, Sep. 2009.
- [36] B. Rost, “PHD: predicting one-dimensional protein structure by profile-based neural networks,” *Methods in enzymology*, vol. 266, pp. 525–539, 1996.
- [37] D. T. Jones, “Protein secondary structure prediction based on position-specific scoring matrices,” *Journal of molecular biology*, vol. 292, no. 2, pp. 195–202, Sep. 1999.
- [38] J. A. Cuff and G. J. Barton, “Evaluation and improvement of multiple sequence methods for protein secondary structure prediction,” *Proteins*, vol. 34, no. 4, pp. 508–519, Mar. 1999.
- [39] H. Shen and J. J. Chou, “Membrain: Improving the accuracy of predicting transmembrane helices,” *PLoS ONE*, vol. 3, no. 6, 2008.
- [40] G. E. Hinton, “Training Products of Experts by Minimizing Contrastive Divergence,” *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, aug 2002.
- [41] A. Kolinski, “Protein modeling and structure prediction with a reduced representation,” *Acta Biochimica Polonica*, vol. 51, no. 2, pp. 349–371, 2004.
- [42] S. Saraswathi, J. L. Fernandez-Martnez, A. Kolinski, R. L. Jernigan, and A. Kloczkowski, “Fast learning optimized prediction methodology (FLOPRED) for protein secondary structure prediction.” *Journal of Molecular Modelling*, vol. 18, no. 9, pp. 4275–89, 2012.
- [43] T. Nitta, “The computational power of complex-valued neuron,” *Lecture Notes in Computer Science*, vol. 2714, pp. 993–1000, 2003.

- [44] ———, “Orthogonality of decision boundaries of complex-valued neural networks,” *Neural Computation*, vol. 16, no. 1, pp. 73–97, 2004.
- [45] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, “SCOP: a structural classification of proteins database for the investigation of sequences and structures,” *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536–540, Apr. 1995.
- [46] 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*, vol. 12, no. 1, pp. 103–112, Jan. 2015.
- [47] S. Suresh, R. Savitha, and N. Sundararajan, “A fast learning fully complex-valued relaxation network (FCRN),” *IEEE International Joint Conference on Neural Networks, 2011. (IJCNN 2011)*, pp. 1372–1377, 2011.
- [48] J. Kyte and R. F. Doolittle, “A simple method for displaying the hydropathic character of a protein,” *Journal of Molecular Biology*, vol. 157, no. 1, pp. 105–132, May 1982.
- [49] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson, “Multiple sequence alignment with the clustal series of programs,” *Nucleic acids research*, vol. 31, no. 13, pp. 3497–3500, Jul. 2003.
- [50] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, “Gapped BLAST and PS I-BLAST: a new generation of protein database search programs,” *Nucleic Acids Res*, vol. 25, no. 17, pp. 3389–3402, 1997.
- [51] H. Viklund and A. Elofsson, “Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information,” *Protein Science: A Publication of the Protein Society*, vol. 13, no. 7, pp. 1908–1917, Jul. 2004.

- [52] B. Rost, P. Fariselli, and R. Casadio, “Topology prediction for helical transmembrane proteins at 86% accuracy.” *Protein Science : A Publication of the Protein Society*, vol. 5, no. 8, pp. 1704–1718, Aug. 1996.
- [53] D. T. Jones, “Improving the accuracy of transmembrane protein topology prediction using evolutionary information,” *Bioinformatics (Oxford, England)*, vol. 23, no. 5, pp. 538–544, Mar. 2007.
- [54] G. E. Tusndy, Z. Dosztanyi, and I. Simon, “PDB.tm: selection and membrane localization of transmembrane proteins in the protein data bank,” *Nucleic Acids Research*, vol. 33, no. Database issue, pp. D275–278, Jan. 2005.
- [55] M. A. Lomize, A. L. Lomize, I. D. Pogozheva, and H. I. Mosberg, “OPM: orientations of proteins in membranes database,” *Bioinformatics (Oxford, England)*, vol. 22, no. 5, pp. 623–625, Mar. 2006.
- [56] UniProt Consortium, “Activities at the Universal Protein Resource (UniProt),” *Nucleic Acids Research*, vol. 42, no. Database issue, pp. D191–198, Jan. 2014.
- [57] D. M. Engelman, T. A. Steitz, and A. Goldman, “Identifying Nonpolar Transbilayer Helices in Amino Acid Sequences of Membrane Proteins,” *Annual Review of Biophysics and Biophysical Chemistry*, vol. 15, no. 1, pp. 321–353, 1986.
- [58] D. T. Jones, W. R. Taylor, and J. M. Thornton, “A model recognition approach to the prediction of all-helical membrane protein structure and topology,” *Biochemistry*, vol. 33, no. 10, pp. 3038–3049, Mar. 1994.
- [59] E. L. Sonnhammer, G. von Heijne, and A. Krogh, “A hidden Markov model for predicting transmembrane helices in protein sequences,” *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, vol. 6, pp. 175–182, 1998.
- [60] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer, “Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes,” *Journal of Molecular Biology*, vol. 305, no. 3, pp. 567–580, Jan. 2001.

- [61] H. Nielsen and A. Krogh, “Prediction of signal peptides and signal anchors by a hidden Markov model,” *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, vol. 6, pp. 122–130, 1998.
- [62] R. C. Edgar, “MUSCLE: multiple sequence alignment with high accuracy and high throughput,” *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.
- [63] K. Chen and L. Kurgan, “Computational prediction of secondary and super-secondary structures,” in *Protein Supersecondary Structures*, ser. Methods in Molecular Biology, A. E. Kister, Ed. Humana Press, Jan. 2013, no. 932, pp. 63–86.
- [64] S. Henikoff and J. G. Henikoff, “Amino acid substitution matrices from protein blocks,” *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, pp. 10 915–10 919, Nov. 1992.
- [65] G. E. Tusndy and I. Simon, “Principles governing amino acid composition of integral membrane proteins: application to topology prediction,” *Journal of Molecular Biology*, vol. 283, no. 2, pp. 489–506, Oct. 1998.
- [66] T. N. Petersen, S. Brunak, G. von Heijne, and H. Nielsen, “SignalP 4.0: discriminating signal peptides from transmembrane regions,” *Nature Methods*, vol. 8, no. 10, pp. 785–786, Oct. 2011.
- [67] C. Peters, K. D. Tsirigos, N. Shu, and A. Elofsson, “Improved topology prediction using the terminal hydrophobic helices rule,” *Bioinformatics (Oxford, England)*, vol. 32, no. 8, pp. 1158–1162, Apr. 2016.
- [68] J. Dyrly Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak, “Improved Prediction of Signal Peptides: SignalP 3.0,” *Journal of Molecular Biology*, vol. 340, no. 4, pp. 783–795, Jul. 2004.
- [69] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, “Predicting subcellular localization of proteins based on their N-terminal amino acid sequence,” *Journal of Molecular Biology*, vol. 300, no. 4, pp. 1005–1016, Jul. 2000.

- [70] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen, “Locating proteins in the cell using TargetP, SignalP and related tools,” *Nature Protocols*, vol. 2, no. 4, pp. 953–971, 2007.
- [71] L. Käll, A. Krogh, and E. L. L. Sonnhammer, “An HMM posterior decoder for sequence feature prediction that includes homology information,” *Bioinformatics (Oxford, England)*, vol. 21 Suppl 1, pp. i251–257, Jun. 2005.
- [72] R. Y. Kahsay, G. Gao, and L. Liao, “An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes,” *Bioinformatics (Oxford, England)*, vol. 21, no. 9, pp. 1853–1858, May 2005.
- [73] A. Bernsel, H. Viklund, A. Hennerdal, and A. Elofsson, “TOPCONS: consensus prediction of membrane protein topology,” *Nucleic Acids Research*, vol. 37, no. Web Server issue, pp. W465–468, Jul. 2009.
- [74] H. Viklund, A. Bernsel, M. Skwark, and A. Elofsson, “SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology,” *Bioinformatics (Oxford, England)*, vol. 24, no. 24, pp. 2928–2929, Dec. 2008.
- [75] S. M. Reynolds, L. Kll, M. E. Riffe, J. A. Bilmes, and W. S. Noble, “Transmembrane topology and signal peptide prediction using dynamic bayesian networks,” *PLoS computational biology*, vol. 4, no. 11, p. e1000213, Nov. 2008.
- [76] A. Bernsel, H. Viklund, J. Falk, E. Lindahl, G. v. Heijne, and A. Elofsson, “Prediction of membrane-protein topology from first principles,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 20, pp. 7177–7181, May 2008.
- [77] H. Wu, K. Wang, L. Lu, Y. Xue, Q. Lyu, and M. Jiang, “Deep conditional random field approach to transmembrane topology prediction and application to gpcr three-dimensional structure modeling,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 5, pp. 1106–1114, Sept 2017.

- [78] J. K. Leman, R. Mueller, M. Karakas, N. Woetzel, and J. Meiler, “Simultaneous prediction of protein secondary structure and transmembrane spans,” *Proteins*, vol. 81, no. 7, pp. 1127–1140, Jul. 2013.
- [79] L. Käll, A. Krogh, and E. L. L. Sonnhammer, “A combined transmembrane topology and signal peptide prediction method,” *Journal of Molecular Biology*, vol. 338, no. 5, pp. 1027–1036, May 2004.
- [80] R. M. Fano, *Transmission of Information*, 1st ed. New York: John Wiley & Sons, Inc., 1961.
- [81] J. Garnier, D. J. Osguthorpe, and B. Robson, “Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins,” *Journal of Molecular Biology*, vol. 120, no. 1, pp. 97–120, Mar. 1978.
- [82] J. Garnier and B. Robson, “The GOR Method for Predicting Secondary Structures in Proteins,” in *Prediction of Protein Structure and the Principles of Protein Conformation*, G. D. Fasman, Ed. Springer US, 1989, pp. 417–465.
- [83] J. F. Gibrat, J. Garnier, and B. Robson, “Further developments of protein secondary structure prediction using information theory: New parameters and consideration of residue pairs,” *Journal of Molecular Biology*, vol. 198, no. 3, pp. 425–443, Dec. 1987.
- [84] J. Garnier, J. F. Gibrat, and B. Robson, “GOR method for predicting protein secondary structure from amino acid sequence,” *Methods Enzymology*, vol. 266, pp. 540–553, 1996.
- [85] A. Kloczkowski, K.-L. Ting, R. L. Jernigan, and J. Garnier, “Combining the GOR v algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence,” *Proteins*, vol. 49, no. 2, pp. 154–166, Nov. 2002.
- [86] T. Z. Sen, H. Cheng, A. Kloczkowski, and R. L. Jernigan, “A consensus data mining secondary structure prediction by combining GOR v and frag-

- ment database mining,” *Protein Science : A Publication of the Protein Society*, vol. 15, no. 11, pp. 2499–2506, Nov. 2006.
- [87] H. Cheng, T. Z. Sen, A. Kloczkowski, D. Margaritis, and R. L. Jernigan, “Prediction of protein secondary structure by mining structural fragment database,” *Polymer*, vol. 46, no. 12, pp. 4314–4321, May 2005.
 - [88] M. Ouali and R. D. King, “Cascaded multiple classifiers for secondary structure prediction,” *Protein science: a publication of the Protein Society*, vol. 9, no. 6, pp. 1162–1176, Jun. 2000.
 - [89] B. Rost and C. Sander, “Prediction of protein secondary structure at better than 70% accuracy,” *Journal of molecular biology*, vol. 232, no. 2, pp. 584–599, Jul. 1993.
 - [90] B. Rost, G. Yachdav, and J. Liu, “The PredictProtein server,” *Nucleic acids research*, vol. 32, no. Web Server issue, pp. W321–326, Jul. 2004.
 - [91] K. Bryson, L. J. McGuffin, R. L. Marsden, J. J. Ward, J. S. Sodhi, and D. T. Jones, “Protein structure prediction servers at university college london,” *Nucleic Acids Research*, vol. 33, no. Web Server issue, pp. W36–38, Jul. 2005.
 - [92] J. A. Cuff and G. J. Barton, “Application of multiple sequence alignment profiles to improve protein secondary structure prediction,” *Proteins*, vol. 40, no. 3, pp. 502–511, Aug. 2000.
 - [93] C. Cole, J. D. Barber, and G. J. Barton, “The jpred 3 secondary structure prediction server,” *Nucleic acids research*, vol. 36, no. Web Server issue, pp. W197–201, Jul. 2008.
 - [94] R. Adamczak, A. Porollo, and J. Meller, “Combining prediction of secondary structure and solvent accessibility in proteins,” *Proteins*, vol. 59, no. 3, pp. 467–475, May 2005.
 - [95] E. Faraggi, Y. Yang, S. Zhang, and Y. Zhou, “Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free

- protein structure prediction,” *Structure*, vol. 17, no. 11, pp. 1515–1527, Nov. 2009.
- [96] A. Yaseen and Y. Li, “Context-Based Features Enhance Protein Secondary Structure Prediction Accuracy,” *Journal of Chemical Information and Modeling*, vol. 54, pp. 992–1002, Mar. 2014.
 - [97] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri, “Exploiting the past and the future in protein secondary structure prediction,” *Bioinformatics (Oxford, England)*, vol. 15, no. 11, pp. 937–946, Nov. 1999.
 - [98] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, “Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles,” *Proteins*, vol. 47, no. 2, pp. 228–235, May 2002.
 - [99] G. Pollastri and A. McLysaght, “Porter: a new, accurate server for protein secondary structure prediction,” *Bioinformatics*, vol. 21, no. 8, pp. 1719–1720, Apr. 2005.
 - [100] 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, p. 575586, 2006.
 - [101] 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, Dec. 2009.
 - [102] K. Lin, V. A. Simossis, W. R. Taylor, and J. Heringa, “A simple and fast secondary structure prediction method using hidden neural networks,” *Bioinformatics (Oxford, England)*, vol. 21, no. 2, pp. 152–159, Jan. 2005.
 - [103] J. Martin, J.-F. Gibrat, and F. Rodolphe, “Analysis of an optimal hidden markov model for secondary structure prediction,” *BMC structural biology*, vol. 6, p. 25, 2006.

- [104] K.-J. Won, T. Hamelryck, A. Prgel-Bennett, and A. Krogh, “An evolutionary method for learning HMM structure: prediction of protein secondary structure,” *BMC Bioinformatics*, vol. 8, no. 1, p. 357, Sep. 2007.
- [105] G. E. Hinton and R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [106] R. Salakhutdinov and G. Hinton, “Deep boltzmann machines,” *Journal of Machine Learning Research Workshop and Conference Proceedings*, vol. 5, pp. 448–455, 2009.
- [107] D. Erhan, A. Courville, and P. Vincent, “Why Does Unsupervised Pre-training Help Deep Learning ?” *Journal of Machine Learning Research*, vol. 11, pp. 625–660, 2010.
- [108] V. Nair and G. E. Hinton, “3d Object Recognition with Deep Belief Nets,” in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Curran Associates, Inc., 2009, pp. 1339–1347.
- [109] L. Kall, A. Krogh, and E. L. L. Sonnhammer, “A combined transmembrane topology and signal peptide prediction method,” *Journal of Molecular Biology*, vol. 338, no. 5, pp. 1027–1036, May 2004.
- [110] ———, “An HMM posterior decoder for sequence feature prediction that includes homology information,” *Bioinformatics (Oxford, England)*, vol. 21 Suppl 1, pp. i251–i257, Jun. 2005.
- [111] S. Suresh, N. Sundararajan, and P. Saratchandran, “Risk-sensitive loss functions for sparse multi-category classification problems,” *Information Sciences*, vol. 178, no. 12, pp. 2621–2638, Jun. 2008.
- [112] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, “BLAST+: architecture and applications,” *BMC bioinformatics*, vol. 10, p. 421, 2009.

- [113] W.-C. Wong, S. Maurer-Stroh, and F. Eisenhaber, “More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology,” *PLoS Computational Biology*, vol. 6, no. 7, p. e1000867, 2010.
- [114] C. Dong, K. Beis, J. Nesper, A. L. Brunkan-LaMontagne, B. R. Clarke, C. Whitfield, and J. H. Naismith, “Wza the translocon for E. coli capsular polysaccharides defines a new class of membrane protein,” *Nature*, vol. 444, no. 7116, pp. 226–229, Nov. 2006.
- [115] J. Abeegithan, R. Shamima, S. Suresh, and D. Shirin, “Hierarchical approach to separate membrane, non-membrane, signal peptide and to improve prediction of transmembrane protein topology,” *Submitted*.
- [116] G. Babu and S. Suresh, “Sequential Projection-Based Metacognitive Learning in a Radial Basis Function Network for Classification Problems,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 2, pp. 194–206, Feb. 2013.
- [117] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.
- [118] W. P. Rivers, “Autonomy at All Costs: An Ethnography of Metacognitive Self-Assessment and Self-Management among Experienced Language Learners,” *The Modern Language Journal*, vol. 85, no. 2, pp. 279–290, Aug. 2001.
- [119] R. M. Isaacson and F. Fujita, “Metacognitive Knowledge Monitoring and Self-Regulated Learning: Academic Success and Reflections on Learning,” *Journal of Scholarship of Teaching and Learning*, vol. 6, no. 1, pp. 39–55, Aug. 2006.
- [120] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, “CD-HIT Suite: a web server for clustering and comparing biological sequences,” *Bioinformatics*, vol. 26, no. 5, pp. 680–682, Mar. 2010. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2828112/>

- [121] A. Bernsel, H. Viklund, J. Falk, E. Lindahl, G. v. Heijne, and A. Elofsson, "Prediction of membrane-protein topology from first principles," *Proceedings of the National Academy of Sciences, USA*, vol. 105, no. 20, pp. 7177–7181, May 2008.
- [122] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. Wilkins, R. Appel, and A. Bairoch, "Protein Identification and Analysis Tools on the ExPASy Server," in *The Proteomics Protocols Handbook*, J. Walker, Ed. Humana Press, Jan. 2005, pp. 571–607.
- [123] H. Larochelle, "Neural networks [5.4]: Restricted boltzmann machine - contrastive divergence," Nov. 2013, video lectures. Date Accessed: 25 Apr 2017. [Online]. Available: <https://www.youtube.com/watch?v=MD8qXWucJBY>
- [124] —, "Neural networks [5.5]:restricted boltzmann machine - contrastive divergence (parameter update)," Nov. 2013, video lectures. Date Accessed: 25 Apr 2017. [Online]. Available: <https://www.youtube.com/watch?v=wMb7cads0go>
- [125] —, "Neural networks [5.6] : Restricted boltzmann machine - persistent contrastive divergence," Nov. 2013, video lectures. Date Accessed: 25 Apr 2017. [Online]. Available: <https://www.youtube.com/watch?v=S0kFFiHzR8M>
- [126] B. Yang, H. Wei, Z. Zhun, and Q. Huabin, "KAAPRO An approach of protein secondary structure prediction based in KDD* in the compound pyramid prediction model," *Expert Systems with Applications*, vol. 36, no. 5, pp. 9000–9006, 2009.
- [127] W. R. Atchley, J. Zhao, A. D. Fernandes, and T. Drke, "Solving the protein sequence metric problem," *Proceedings of the National Academy of Sciences*, vol. 102, no. 18, pp. 6395–6400, 2005.
- [128] S. Suresh, N. Sundararajan, and R. Savitha, "Performance study on real-valued classification problems," in *Supervised Learning with Complex-valued Neural*

Networks, ser. Studies in Computational Intelligence. Springer Berlin Heidelberg, Jan. 2013, no. 421, pp. 125–133.

- [129] J. Cheng, A. N. Tegge, and P. Baldi, “Machine learning methods for protein structure prediction,” *IEEE reviews in biomedical engineering*, vol. 1, pp. 41–49, 2008.
- [130] H. Leung and S. Haykin, “The complex backpropagation algorithm,” *IEEE Transactions on Signal Processing*, vol. 39, no. 9, pp. 2101–2104, 1991.
- [131] T. Kim and T. Adali, “Fully complex multi-layer perceptron network for non-linear signal processing,” *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 32, no. 1-2, pp. 29–43, Aug. 2002.
- [132] R. Savitha, S. Suresh, N. Sundararajan, and P. Saratchandran, “A new learning algorithm with logarithmic performance index for complex-valued neural networks,” *Neurocomputing*, vol. 72, no. 16-18, pp. 3771–3781, 2009.
- [133] S. Chen, S. McLaughlin, and B. Mulgrew, “Complex-valued radial basic function network, part i: Network architecture and learning algorithms,” *Signal Processing*, vol. 35, no. 1, pp. 19–31, Jan. 1994.
- [134] —, “Complex-valued radial basis function network, part II: application to digital communications channel equalisation,” *Signal Processing*, vol. 36, no. 2, pp. 175–188, Mar. 1994.
- [135] R. Savitha, S. Suresh, and N. Sundararajan, “Metacognitive learning in a fully complex-valued radial basis function neural network,” *Neural Computation*, vol. 24, no. 5, pp. 1297–1328, Dec. 2011.
- [136] —, “Fast learning circular complex-valued extreme learning machine (CC-ELM) for real-valued classification problems,” *Information Sciences*, vol. 187, no. 1, pp. 277–290, 2012.
- [137] D. Jianping, N. Sundararajan, and P. Saratchandran, “Communication channel equalization using complex-valued minimal radial basis function neural net-

- works,” *IEEE Transactions on Neural Networks*, vol. 13, no. 3, pp. 687–696, 2002.
- [138] S. Suresh, R. Savitha, and N. Sundararajan, “A sequential learning algorithm for complex-valued self-regulating resource allocation network-CSRAN,” *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, vol. 22, no. 7, pp. 1061–1072, Jul. 2011.
- [139] A. S. Siddiqui, U. Dengler, and G. J. Barton, “3Dee: a database of protein structural domains,” *Bioinformatics (Oxford, England)*, vol. 17, no. 2, pp. 200–201, Feb. 2001.
- [140] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, “CATH—a hierarchic classification of protein domain structures,” *Structure*, vol. 5, no. 8, pp. 1093–1108, Aug. 1997.
- [141] P. A. Alexander, Y. He, Y. Chen, J. Orban, and P. N. Bryan, “A minimal sequence code for switching protein structure and function,” *Proceedings of the National Academy of Sciences, USA*, vol. 106, no. 50, pp. 21 149–21 154, Dec. 2009.
- [142] P. N. Bryan and J. Orban, “Proteins that switch folds,” *Current Opinion in Structural Biology*, vol. 20, no. 4, pp. 482–488, Aug. 2010.
- [143] S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, Mar. 1970.
- [144] C. Sander and R. Schneider, “Database of homology-derived protein structures and the structural meaning of sequence alignment,” *Proteins*, vol. 9, no. 1, pp. 56–68, 1991.
- [145] P. J. Silva, “Assessing the reliability of sequence similarities detected through hydrophobic cluster analysis,” *Proteins*, vol. 70, no. 4, pp. 1588–1594, Mar. 2008.

- [146] A. Zemla, C. Venclovas, K. Fidelis, and B. Rost, “A modified definition of sov, a segment-based measure for protein secondary structure prediction assessment,” *Proteins*, vol. 34, no. 2, pp. 220–223, Feb. 1999.
- [147] G. B. Huang, Q. Y. Zhu, and C. K. Siew, “Extreme learning machine: a new learning scheme of feedforward neural networks,” *IEEE International Joint Conference on Neural Networks. (IJCNN 2004)*, vol. 2, pp. 985 – 990, 25-29 2004.
- [148] S. S. Shapiro and M. B. Wilk, “An analysis of variance test for normality (complete samples),” *Biometrika*, vol. 52, no. 3-4, pp. 591–611, Dec. 1965.
- [149] F. Wilcoxon, “Individual comparisons by ranking methods,” *Biometrics*, vol. 1, no. 6, p. 80, Dec. 1945.
- [150] C. Mirabello and G. Pollastri, “Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility,” *Bioinformatics*, vol. 29, no. 16, pp. 2056–2058, Aug. 2013.
- [151] Schrödinger, LLC, “The PyMOL molecular graphics system, version 0.99,” 2006.
- [152] I. K. McDonald and J. M. Thornton, “Satisfying Hydrogen Bonding Potential in Proteins,” *Journal of Molecular Biology*, vol. 238, no. 5, pp. 777–793, May 1994.
- [153] S. Suresh, K. Dong, and H. J. Kim, “A sequential learning algorithm for self-adaptive resource allocation network classifier,” *Neurocomputing*, vol. 73, no. 1618, pp. 3012–3019, Oct. 2010.
- [154] Y. Zhang and J. Skolnick, “TM-align: a protein structure alignment algorithm based on the TM-score,” *Nucleic Acids Research*, vol. 33, no. 7, pp. 2302–2309, 2005.
- [155] D. University of California, “Chemwiki lab 7,” 2013, date accessed: December 2013. [Online]. Available: http://chemwiki.ucdavis.edu/Wikitexts/UCD_Chem_115_Lab_Manual/Lab_7%3A_Electrospray_Mass_Spectrometry