Applying Conditional Random Fields into Bioinformatics

Wei Liu

SID: 307035077

Master of Science by Research
School of Information Technologies
Supervisor: Dr Sanjay Chawla
The University of Sydney
5 November 2007
Abstract

In this thesis, we investigate Conditional Random Fields (CRFs) to predict the orientations and locations of alpha helices in transmembrane proteins. CRFs are discriminative undirected graphical models which are known to have several advantages over other classification models. We have selected 18 features to capture the structural information within the amino acids sequences. Most of these features are based on domain knowledge in protein science. We used a standard benchmark server to assess the performance of CRFs, and received the highest score in both the per-segment and per-residue accuracy among twenty eight other methods. We have also set up a CRFs prediction web server, which includes all the optimised parameters obtained from training process. Furthermore, we have carried out detailed experiments comparing CRFs against Maximum Entropy Markov Models (MEMMs), and have shown that CRFs are significantly better than MEMMs in predicting the location of helix segments in transmembrane proteins.
Acknowledgements

I would like to thank my supervisor, Sanjay Chawla, for his support and encouragement throughout the duration of this project.

The thesis is an extension of the work of Lior Lukov, a past MSc research student in the School of IT. I want to acknowledge that I gained immense knowledge from reading his thesis and the software code that he had written as part of his thesis.
## Contents

1 Introduction .................................................. 1
   1.1 Motivation ............................................. 1
   1.2 Problem Definition .................................... 2
   1.3 Key Contributions ...................................... 3
   1.4 Road Map ............................................... 5

2 Proteins and Protein Structures .......................... 6
   2.1 Amino Acids: Elements in Primary Sequences of Proteins . . . . 6
   2.2 Protein Structure ....................................... 11
      2.2.1 Primary Structure .................................. 11
      2.2.2 Secondary Structure ................................ 11
      2.2.3 Tertiary Structure .................................. 12
      2.2.4 Quaternary Structure ............................... 13
   2.3 Transmembrane Helix Prediction ........................ 13
   2.4 Summary ................................................ 15

3 Literature Review in Transmembrane Helix Prediction .... 16
   3.1 Hydrophobicity Scale Based Methods ................... 16
      3.1.1 Kyte and Doolittle’s Method ....................... 17
      3.1.2 Wimley-White (WW) index ......................... 19
      3.1.3 TopPred method .................................... 19
      3.1.4 SOSUI ............................................. 22
   3.2 Advanced Methods ....................................... 24
3.2.1 Neural Network Based Method .................................. 24
3.2.2 DAS Method ......................................................... 25
3.2.3 Hidden Markov Model Based Method ......................... 27
    3.2.3.1 TMHMM method ............................................. 27
    3.2.3.2 HMMTOP method ........................................... 29
3.2.4 Maximum Entropy Markov Model .............................. 30
3.3 Summary ............................................................. 32

4 Conditional Random Fields (CRFs) .................................. 33
4.1 Model Definition .................................................. 34
4.2 Feature Integration with the Model ............................... 35
    4.2.1 Start, End and Edge Features ............................... 35
    4.2.2 Basic Amino Acid Features ................................. 36
    4.2.3 Hydrophobic Window Features ............................. 37
    4.2.4 Hydrophilic Window Features ............................. 38
    4.2.5 Single Side Neighboring Amino Acid Features .......... 38
    4.2.6 Single Side Shuffled Neighboring Amino Acid Features . 39
    4.2.7 Single Side Hydrophobic Neighbouring Amino Acid Features ................................................. 39
    4.2.8 Single Side Hydrophilic Neighbouring Amino Acid Features ..................................................... 40
    4.2.9 Double Side Neighboring Amino Acid Features .......... 40
    4.2.10 Double Side Shuffled Neighboring Amino Acid Features ....................................................... 41
    4.2.11 Double Side Hydrophobic Neighbouring Amino Acid Features .................................................. 42
    4.2.12 Double Side Hydrophilic Neighbouring Amino Acid Features .................................................. 42
    4.2.13 Amino Acid Property Features ........................... 42
    4.2.14 Border Features ............................................. 43
    4.2.15 Short Loop Features ......................................... 43
# List of Figures

1.1 Examples of Protein Sequence Pairs .......................... 3

2.1 The General Structure of an Amino Acid ......................... 7
2.2 Amino Acids Sidechains ........................................... 8
2.3 The Formation of a Covalent Bond ............................... 10
2.4 Formation of a Polypeptide ....................................... 10
2.5 Polypeptide Backbone .............................................. 10
2.6 Proteins Four Levels Structures ................................. 11
2.7 Representation of the $\alpha$-helix Structure .................... 13
2.8 Membrane and Soluble Proteins ................................. 14

3.1 Example of Hydrophobicity Plot ................................. 18
3.2 Example of Positively Charged Residues ........................ 20
3.3 Sliding window in TopPred ......................................... 21
3.4 Three regions of a transmembrane helix in SOSUI ............ 23
3.5 Architecture of PHDhtm’s Neural Network ....................... 25
3.6 Architecture of TMHMM ............................................. 28
3.7 Structure of Submodel for Loop Regions in TMHMM .......... 28
3.8 Structure of Submodel for Helix Core Regions in TMHMM .... 28
3.9 Structural states in HMMTOP ....................................... 30
3.10 MEMM graphical framework ....................................... 31
4.1 CRFs graphical structure ........................................... 34
4.2 Segments Hydrophobic Value Comparison .......................... 37
5.1 Accuracy of the eight Experiments ............................... 53
5.2 Training Time Comparison ........................................ 55
5.3 Training Time Comparison ........................................ 56
5.4 CRFs vs. MEMM on Experiment 1 .............................. 57
5.5 CRFs vs. MEMM on Experiment 2 .............................. 57
5.6 CRFs vs. MEMM on Experiment 3 .............................. 58
5.7 CRFs vs. MEMM on Experiment 4 .............................. 58
5.8 CRFs vs. MEMM on Experiment 5 .............................. 59
5.9 CRFs vs. MEMM on Experiment 6 .............................. 59
5.10 CRFs vs. MEMM on Experiment 7 ............................ 60
5.11 CRFs vs. MEMM on Experiment 8 ............................ 60
6.1 Conform Test Data ..................................................... 64
6.2 Example of Server Output .......................................... 65
## List of Tables

3.1 Amino Acids Hydrophobic Index ........................................ 17  
4.1 Amino Acid Electronic Property ........................................ 44  
4.2 Chemical Component in Side Chains ................................... 45  
4.3 Classification by Side Chains Component ............................. 46  
5.1 Enabled and disabled feature combination ............................ 52  
5.2 Prediction score comparison between 29 methods ................. 54  
5.3 Accuracy Comparison between CRFs and MEMM .................... 61  
5.4 P-Values of CRFs vs. MEMM in each Experiment ................... 62
Chapter 1

Introduction

1.1 Motivation

Membrane proteins are crucial for many biological functions in living organisms and are dominant targets for pharmacological agents. They control key components for the regulation of ion and metabolite fluxes, maintaining appropriate contact between cells, connective tissues, and extracellular matrices. Membrane-bound receptors are the preferable targets of pharmacological industry. The G protein-coupled receptor (GOCR), for instance, is an example of such a membrane protein which is of interest to the pharmaceutical industry as they present novel targets for drugs[6].

In order to determine the functions and behaviours of membrane proteins, it is necessary to know their structures. However, detailed structural information available for membrane proteins is limited. We only have low-resolution experimental information for less than 500 helical membrane proteins, and even the PDB [3] contains no more than 50 sequence-unique protein chains with high-resolution helical structures[4].

The fact that we can hardly get enough data on the structures of membrane pro-
teins from traditional biophysical methods is mainly due to the difficulty of X-ray crystallography and NMR spectroscopy on these proteins. New methods of mass spectroscopy have begun to assist topology assignment in rapid product identification in limited proteolytic cleavage experiments, but its efficiency is limited [23]. Therefore, there is a high demand for a method that can efficiently and accurately deliver membrane proteins structure.

1.2 Problem Definition

In this thesis we will use a conditional probability model, Conditional Random Field, to solve the transmembrane helix prediction problem.

We considered this prediction problem as a binary sequential classification problem. Classification problems are well known in various fields such as computational linguistics, speech recognition, and computational biology. All of these problems share a common task: given a set of observation sequences, identify a corresponding label sequence for these observations. The observation sequence can be taken from any domain, such as words in a document, different values of stocks over time, nucleotides in a DNA sequence, or amino acids in a protein sequence. The label sequence is the classification of the items in the observation sequence [25].

In our scenario, the classification problem is to classify amino acids inside a sequence into two categories: one category in \textit{Helix Segment} and the other one in \textit{Non-helix Segment}. Suppose we have a set of membrane proteins, the structures of which are already known, then each protein in the set is represented by a pair of sequences: the observation sequence represented by $x$, and the structure sequence represented by $y$. The protein observation sequence consists of amino acids, in the form of 20 different letters (every amino acid is exhibited by one letter format). The structure sequence consists of protein’s helical or non-helical structure at the position of corresponding amino acids. As mentioned earlier, we consider
this structure as a binary sequence, so in this thesis we notate the helical state in sequence y by "1" and the non-helical state by "0". The amino acid’s helical structure at position \( i \) in sequence y corresponds to the amino acid at the same position \( (i) \) in sequence x. An example set of membrane protein sequence pairs is shown in Figure 1.1.

\[
\text{MF INRMLFSWNHDKIGTLYLLFGAMAGTVGTALSSLRLRAELGQPGLLGDIQTVNYTVTAHAFU 000000000011111111111111111111111111000000000011111111111111111111111111111}
\]

Figure 1.1: Examples of Protein Sequence Pairs [25]

Given a set of proteins in the form of sequence pairs, known as the \textit{training data sets}, the classification problems are how to capture the structural information displayed in the training data set, and how to use those information to predict transmembrane helix on other proteins’ amino acids sequences, known as the \textit{test data sets}. These are the main problems we are going to solve in this thesis. We use various features populated in CRFs to capture the information in training data, and then train these features using a optimization mechanism to make predictions on test data.

### 1.3 Key Contributions

The main contributions we made in this thesis are listed as follows:

1. Conditional Random Fields for Transmembrane Helix Prediction:
   The performance of CRFs have been tested in a standard TMH benchmark server [2], using a new set of features that have been derived and extracted from the "bioinformatics literature". On segments prediction, 88% of the 2247 proteins have their segments predicted correctly, and 99% of the entire helices in test data are accurately predicted. On residues prediction, 84% of all the amino acids in the test data have their helical states predicted.
Chapter 1. Introduction

correctly. We compared our results with other twenty eight methods on the
same test data (2247 protein sequences). We obtained the highest score in
both per-segment and per-residue accuracy. In this respect we improve upon
the results by Lior Lukov [25] (especially on per-segment accuracy). Details
of these prediction accuracies and comparisons are presented in Chapter 5.

2. Features selection:

Extensive tests have been carried out in this thesis to select the most ap-
propriate features to populate the CRFs model. We selected 18 effective
features, most of which are based on domain knowledge from within pro-
tein science. A detailed features evaluation is presented in Chapter 4.1. The
set of features that we have deployed are much more biologically complex
than those presented by Lukov. The selection of the features and their inte-
gration into the CRF is one of the main distinguishing feature of this work
compared to Lukov.

3. CRFs Web Server:

We have set up a CRFs transmembrane helices prediction web server. This
server contains all the optimised parameters, which are the outputs of train-
ing process and are the main factors in labelling process. An introduction
of the CRFs server is presented in Chapter 6.

4. Comparison between CRFs and MEMM on Transmembrane Prediction:

We have carried out a more detailed comparison between CRFs and MEMM
compared to what was reported in [25]. We compared not only the accuracy
differences between these two models, but also the differences in length of
training time and labelling time. CRFs outperform MEMM on most of the
prediction metrics. Details of these comparisons are presented in Chapter 5.
1.4 Road Map

The organization in the rest of the thesis is as follows: In Chapter 2 we give an introduction about protein science and protein structures. Chapter 3 presents a review of literature on previous computational methods, which characterize and predict the locations of transmembrane helices or the topology of protein sequences. In Chapter 4 we introduce the Conditional Random Fields model, highlight its main strengths, and give a detailed explanation of the features used in the model. In Chapter 5 we describe the experiments, the evaluation of our results on a benchmark data set [2], and the comparisons of CRFs against MEMM and 28 other methods. Chapter 6 gives an introduction of the CRFs transmembrane helices prediction web server. In Chapter 7 we conclude with a summary of the results and directions for future research.
Chapter 2

Proteins and Protein Structures

Proteins consist of complex molecules which mediate most of the significant functions in living organism. The functionalities of protein molecules involved in almost all biological reactions — the signal transduction, transport of material and defence of self from foreign bodies are all carried out by thousands of different proteins in the body. Proteins also play structural roles in an organism such as forming tissues and muscular fibre. These proteins’ functions are mediated by their structures which inherit the designated functionality from the genetic information of the cells, known as the DNA (Deoxyribonucleic Acid) sequence[15]. In this chapter, we present an introduction to the protein science and proteins structures.

2.1 Amino Acids: Elements in Primary Sequences of Proteins

Proteins are made of 20 types of amino acids, which are linked to each other in a linear fashion like beads in a chain. These 20 different amino acids share a common chemical composition: a carbon atom (called $C_\alpha$) at the center, covalently linked to four groups: (i) the amino group ($-NH_2^+$), (ii) the carboxyl group
Chapter 2. Proteins and Protein Structures

Figure 2.1: The General Structure of an Amino Acid: there is a carbon atom at the center that forms 4 covalent bonds with: the amino group, the carboxyl group, one hydrogen atom and the side chain \( R \). \[47\]

\[ \text{(-COO}^{-}) \], (iii) one hydrogen atom \( (H) \) and (iv) the side chain group \( (R) \) \[15\]. The general structure of an amino acid is shown in Figure 2.1. The first three components are common to all amino acids, while the side chain \( R \) consists of one or several chemical groups that can identify one amino acid from the others. Figure 2.2 shows the side chains of the 20 amino acids along with their names and the 3-letter and 1-letter codes that are usually used to represent them. Since in most of the computational methods the amino acids are commonly represented by one-letter format, we also adopt this fashion when we deal with amino acids sequences in this thesis. In addition, the term "residue" is commonly used to refer to any amino acid in a protein sequence, so we also adopt this term in the thesis.

Because of the differences in their side chains, the 20 amino acids have distinct chemical and physical properties. These properties have been proposed as a basis of classifying amino acids. One of the common used ways of classification, defined by Sternberg \[38\], is to classify the amino acids into nine groups:

1. Aromatic (F,W,Y,H)
Chapter 2. Proteins and Protein Structures

Figure 2.2: Sidechains of the 20 Amino Acids. There are 20 various side chains giving rise to 20 amino acids. The name of the 20 amino acids and their 3-letter and 1-letter codes are shown here. [1]

3. Positive (H,K,R)


5. Charged (H,K,R,E,D)

6. Negative (E,D)

7. Aliphatic (I,L,V)


9. Tiny (A,G,S)

This classification will be used in Chapter 4.1 to define one of the features (Amino Acid Property Feature) for the task of learning training data.

Two amino acids can join to each other through reaction between amino group of one and the carboxyl group of the other. As shown in Figure 2.3, one oxygen atom ($O$) from the carboxyl group on the left amino acid and two hydrogen atoms ($H$) from the amino group on the right amino acid form a water molecule ($H_2O$), leading to the formation of a covalent bond between the carbon ($C$) and nitrogen ($N$) atoms of the carboxyl and amino groups respectively. If this reaction iterates several times between different combinations of amino acids, a polypeptide or a protein will be produced, as shown in Figure 2.4.

In a protein sequence, the main chain or backbone of the sequence is made of the $C_\alpha$ along with the atoms $N$, $C$, $O$ and $H$ that participate in the polypeptide. Conventionally, a protein chain is written from one end to the other, beginning with the amino group ($-NH_3^+$) from the left and ending with the carboxyl group ($-COO^-$) to the right. Thus, the left end of a protein is usually called $N$-terminus and right end is called $C$-terminus. Figure 2.5 shows an example of a polypeptide backbone.
Chapter 2. Proteins and Protein Structures

Figure 2.3: The Formation of a Covalent Bond between two Amino Acids: a covalent bond can form between the carboxyl group of one amino acid and amino group of the other, releasing one water ($H_2O$) molecule.[47]

Figure 2.4: Formation of a Polypeptide with the amino acid sequence: Asp-Lys-Gln-His-Cys-Arg-Phe[17].

Figure 2.5: An Example of a Polypeptide Backbone. [17].
2.2 Protein Structure

Protein functions mainly depend on their three-dimensional structures. Four hierarchical levels of protein structure organization are commonly used in protein science, which are the primary, secondary, tertiary, and quaternary structures.

2.2.1 Primary Structure

In brief, proteins are composed of amino acid residues connected by covalent peptide bonds, which are planar and rigid (Figure 2.6), and the protein’s primary structure is simply the amino acid sequence, which is linear and hence in one-dimension [7].

2.2.2 Secondary Structure

As discussed above, proteins are polymers of amino acids, which specify the primary structure of a protein. However, the function of a given protein depends not only on the amino acid sequences, but also on its overall three dimensional structure. So the protein’s conformation and its functionality is mainly determined by the fragile balance between numbers of forces made by electrostatic bonds,
hydrophobic interactions, van der Waals forces and in particular the Hydrogen 
Bonds [24], and these interactions impart the secondary structure of proteins.

Hydrogen bonds are non-covalent bonds form between two electronegative atoms 
that share one Hydrogen atom (H), which can be normally found in between the 
carbonyl oxygen and amide proton of adjacent peptide groups. Two basic types 
of structures may be formed by the hydrogen bonds: $\alpha$-helix and $\beta$-sheet.

Helix is formed due to the hydrogen bond formation between the carboxyl group 
of $i_{th}$ residue and the amino group of the $(i + n)_{th}$ residue, where the $n$ determines 
whether it is a $3_{10}$, $\alpha$, or $\pi$ helix for $n=3, 4, 5$ respectively. Therefore, the interac-
tion between the amino acids that lead to the formation of a helix are short-range 
to the residues in the helix. Sheet on the other hand is formed due to long-range 
interactions between amino acids. There is another basic secondary structure, 
known as coil. A coil is a segment of protein that does not conform to any of the 
above secondary structure types [14].

In this thesis, it is the $\alpha$-helical segment in proteins secondary structure that we 
concern about. We may use the term "transmembrane helix" to represent "$\alpha$ – 
helix", since the transmembrane domain usually denotes a single alpha helix of 
a transmembrane protein. $\alpha$-helix is a helical structure that consists of 3.6 amino 
acids in average to form one helix turn. A turn is defined as a short segment that 
causes the protein to bend. Each amino acid along the $\alpha$-helix extends about 1.5Å, 
so one helical turn is approximately 5.4Å in length. Figure 2.7 shows an example 
of $\alpha$-helix structure.

2.2.3 Tertiary Structure

The tertiary structure refers to the complete information of the three-dimensional 
positions of all atoms in the protein. The regions consisting of secondary struc-
tures are folded into a specific compact structure for the entire polypeptide chain, 
and the fold usually gives proteins the most stable structure. Figure 2.6 presents
2.2.4 Quaternary Structure

Protein may be composed of several separate polypeptide chains, known as the subunits. The associations of these proteins subunits in three-dimensional complexes constitutes the protein’s quaternary structure [15]. An example of proteins quaternary structure is also shown in Figure 2.6.

2.3 Transmembrane Helix Prediction

Most of the proteins in an organism are present inside a cell, which is made of mostly water. Proteins that are completely in water are called soluble proteins (Figure 2.8C). In contrast, some proteins are embedded in the membrane enveloping the cell, and these proteins are called membrane proteins (Figure 2.8B). Mem-
brane proteins have a significant portion which is embedded in the cell membrane, and experience a different environment, as shown in Figure 2.8A [14].

The task of transmembrane helix prediction is to find the orientation and location of transmembrane helices from amino acids sequences. Transmembrane domain usually denotes a single transmembrane alpha helix of a transmembrane protein. It is called “domain” because an alpha-helix in membrane can be folded independently on the rest of the protein.

Transmembrane prediction may also refers to the topology prediction, which aims to find not only the location of transmembrane helices, but also the orientation of the loops between helix segments. The topology predictions tend to find out whether a loop is in the cytoplasmic side of membrane or in the extracellular side.

In this thesis, we focus on the predictions of the locations of transmembrane helices.
2.4 Summary

The membrane proteins play a significant role in almost every aspect of functions in living organism. They are made up of amino acids which are identified by their unique side chains. The function of a given protein depends not only on the amino acid sequence, but on its overall three dimensional structure. Several hydrogen bonds among the peptide chain form the protein’s secondary structure, of which the basic structures are helix, sheet, and coils.

At the end of this chapter, we introduced the transmembrane prediction problem, which we will solve in this thesis. In the next chapter, we will give a literature review about previous methods in transmembrane helix prediction.
Chapter 3

Literature Review in Transmembrane Helix Prediction

Several methods have been developed for predicting the locations of transmembrane helices. Briefly, there are two main approaches: one uses hydrophobicity scale as basis of their method \[12, 16\], and the other employs statistical models such as the neural network \[31\] and the hidden markov model\[21, 37, 40, 41\] to solve the prediction problem. In this chapter, we will review some of the well known and representative existing methods, analyse their working mechanisms, compare their prediction accuracy and discuss their strengths and weaknesses. Some of the approaches they used are converted into features in our CRFs model, which will be explained in chapter 4.2.

3.1 Hydrophobicity Scale Based Methods

The biochemical nature of cytoplasmic membranes have been extensively studied and described to consist of a lipid bilayer, containing a hydrophobic interior. By definition transmembrane domains are protein segments which span a cytoplasmic membrane and are subsequently exposed to the hydrophobic interior. To exist
in a stable and energetically favourable manner, the protein’s transmembrane do-
mains must also be hydrophobic [4]. So in the earliest transmembrane prediction
methods, simple hydrophobicity plots were used to detect probable transmem-
brane segments [18]. Hydrophobicity of an amino acid was first defined in terms
of its chemical structure. Many other scales have been computed later through
statistical propensities of amino acids in transmembrane and non-transmembrane
regions. The most commonly used hydrophobicity scales are those derived by
Kyte and Doolittle[16, 27].

3.1.1 Kyte and Doolittle’s Method

Kyte and Doolittle [16] defined a hydrophobicity scale for the hydrophilic and
hydrophobic properties of the 20 amino acids. As shown in Table 3.1, each amino
acid is given a hydrophobicity score between 4.6 and -4.6. A score of 4.6 is the
most hydrophobic and a score of -4.6 is the most hydrophilic.

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Hydrophobic Value</th>
<th>Amino Acid</th>
<th>Hydrophobic Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lysine (K)</td>
<td>-3.9</td>
<td>Proline (P)</td>
<td>-1.6</td>
</tr>
<tr>
<td>Arginine (R)</td>
<td>-4.5</td>
<td>Glycine (G)</td>
<td>-0.4</td>
</tr>
<tr>
<td>Histidine (H)</td>
<td>-3.2</td>
<td>Alanine (A)</td>
<td>1.8</td>
</tr>
<tr>
<td>Glutamic(E)</td>
<td>-3.5</td>
<td>Methionine (M)</td>
<td>1.9</td>
</tr>
<tr>
<td>Glutamine (Q)</td>
<td>-3.5</td>
<td>Cysteine (C)</td>
<td>2.5</td>
</tr>
<tr>
<td>Aspartic acid (D)</td>
<td>-3.5</td>
<td>Phenylalanine (A)</td>
<td>2.8</td>
</tr>
<tr>
<td>Asparagine (N)</td>
<td>-3.5</td>
<td>Leucine (L)</td>
<td>3.8</td>
</tr>
<tr>
<td>Trptophan (W)</td>
<td>-0.9</td>
<td>Valine (V)</td>
<td>4.2</td>
</tr>
<tr>
<td>Tyrosine (Y)</td>
<td>-1.3</td>
<td>Isoleucine (I)</td>
<td>4.5</td>
</tr>
<tr>
<td>Serine (S)</td>
<td>-0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Threonine (T)</td>
<td>-0.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Kyte et al developed a program, which uses a moving-segment approach that
continuously determines the average hydrophobicity value with in a segment of
predetermined length as it advances through the sequences. This segment is the
so-called “window”. As we move with the window, consecutive hydrophobicity scores are plotted to amino acids from one end of a sequence to the other end [16].

At the same time of moving the windows, a midpoint line that corresponds to the grand average of the hydrophobicity of the amino acid compositions found in most of the sequenced proteins is drawn. For the transmembrane helices in proteins, the portions of their sequences that are located within the lipid bilayer are clearly delineated by large uninterrupted areas on the hydrophobic side of the midpoint line. As such, the membrane-spanning segments of these proteins can be identified by this procedure.

For example, if we have an amino acid sequence (every amino acid is represented by one letter): "NIPFSVENKWLAMMTLFFGSGFAAPFIHRHQLLK”, and we plot them by using the above hydrophobicity values, then we will get a graph as shown in Figure 3.1 [30]. The right part of blue line, which is higher than 1.6, means a potential transmembrane helix according to Kyte’s paper [16].

Although a window of any length can be chosen, ordinarily spans of odd numbers
were employed by their program, so that a given average value could be plotted above the middle residue of the segments. Setting window size to 5-7 is suggested to be a good value for finding putative surface-exposed regions, whereas a window size of 19-21 yields a plot in which transmembrane domains stand out sharply, with values of at least 1.6 at their centres [42]. The example given in Figure 3.1 uses a window size of 21 amino acids, and uses 1.6 as threshold.

These hydrophobicity scales will be used in defining hydrophobic-based features (e.g. Single Side Hydrophobic Neighbouring Amino Acid Features) in chapter 4.2.

### 3.1.2 Wimley-White (WW) index

A later study [46] suggests that improvement in prediction can be accomplished through the use of an alternative hydrophobicity index. This study, by University of California, Irvine, reports that predictions based on Wimley-White (WW) [34] hydrophobicity index result in more analysis of putative transmembrane domains than traditional hydrophobicity indices, such as the Kyte and Doolittle index. The reason for this increased accuracy is that the WW index includes factors which account for thermodynamic instabilities resulting from both the amino acid side chains and the peptide backbone. Use of the WW index in transmembrane domain identification is reported to have an accuracy greater than 99% [46].

### 3.1.3 TopPred method

Typically, the method used in Kyte and Doolittle’s paper will identify some segments of such high average hydrophobicity that they can be considered “certain” transmembrane segments, but it will also produce one or more candidate segments of intermediate average hydrophobicity that cannot be confidently predicted as transmembrane. So in Gunnar’s paper [8, 43] a complementary approach was suggested: to “bootstrap” the output from a standard hydrophobicity analysis by a
Figure 3.2: The Cytoplasmic Loops Tending to be enriched by Positively Charged Residues ("+" in the figure) in Comparison with the Extracellular Loops. [13]

subsequent step of charge-bias analysis that ranks all possible structures on the basis of their conformity with the positive-inside rule. This new prediction approach is named "TopPred", which is short for "Topology Prediction".

The positive-inside rule is based on the observation that the loop segments in the cytoplasmic side and the helical segments that are adjacent to the cytoplasm are often enriched in the positively charged amino acids, Lysine (K) and Arginine (R), when compared with the extracellular loops, as shown in Figure 3.2 [9]. So the "K + R" bias can serve as a rule for predicting topology, by requiring that more positively charged residues face the cytoplasm [13].

In TopPred, hydrophobicity analysis was carried out using the hydrophobicity scale and a trapezoid sliding window composed of a central, 11-residue rectangular section and 2 flanking wedge-like sections, each 5 residues long (Figure 3.3), so there are 21 amino acids in the window. This window-shape was chosen to combine the favourable noise-reduction of a triangular window [9] with a physically more realistic rectangular window representing the central apolar part of a lipid bilayer. In this way, one obtains a physically reasonable, soft transition from
the apolar interior to the polar surface of the membrane [8].

From the hydrophobicity profile, candidate transmembrane segments were extracted automatically by a procedure that identifies the highest peak, records its average hydrophobicity (AH), and removes the part of the profile that corresponds to the 21 residues in the window. In TopPred’s program, this is repeated until either no stretch longer than 20 amino acids remains, or until no peak higher than the lower cutoff (AH = 0.5) remains. All peaks with AH higher than 1.0 are considered “certain” transmembrane segments, and all with 0.5 < AH < 1.0 are considered “putative candidates” [43].

Finally, all possible topologies that include the certain transmembrane segments and either include or exclude each of the candidate segments were generated by program, and the difference in the number of positively charged amino acid residues (K+R) between the 2 sides of each structure was calculated. The possible topologies were then ranked in decreasing order of charge bias, and their orientation was predicted as the one with the more highly charged side facing the cytoplasm.

Authors claimed that TopPred can predict the correct topology for 23 out of 24 inner membrane proteins and correctly identifies 135 transmembrane segments
with only one overprediction [8].

The sliding window theory used in TopPred has been combined into the definition of our features (Hydrophobic Window Features and Hydrophilic Window Features) in Chapter 4.2

### 3.1.4 SOSUI

SOSUI is a physicochemical properties-based system that achieves very high accuracy in classifying soluble and membrane proteins and also very high segment prediction accuracy [39].

There are three basic assumptions in the SOSUI system:

1. Membrane proteins are characterized by at least one, particularly hydrophobic, primary transmembrane helix.

2. Hydrophilic transmembrane helices may also exist in multi-spanning membrane proteins, even though their hydrophobicity is in fact similar to the hydrophobic segments of soluble proteins. The possible role of secondary transmembrane helices is the formation of active sites of proteins.

3. The primary transmembrane helices are stabilized by a combination of amphiphilic side chains at the helix ends as well as high hydrophobicity in the central region.

Four physicochemical parameters were used in the SOSUI program:

1. the hydrophobicity index of Kyte and Doolittle

2. an amphiphilicity index

3. an index of amino acid charges

4. the length of each sequence
Figure 3.4: Three regions of a transmembrane helix in SOSUIpic (helix centre, helix end, and loop end) [43]

The second parameter (amphiphilicity index) expresses the amphiphilicity of polar side chains and was devised by the calculation of the transfer energy of the hydrocarbon part of a polar side chain. The values of this parameter are finite for large polar residues: 3.67 (Lys); 2.45 (Arg); 1.45 (His); 1.27 (Glu); 1.25 (Gln); 6.93 (Trp); 5.06 (Tyr) [39].

As shown in Figure 3.4, in SOSUI’s program, transmembrane helices were divided into three regions: two end regions "Pe" (five residues long) and a central region "Pc" (all other residues in the middle). SOSUI program calculated the propensities of amino acids for these three regions of helices and a loop region ten residues long. Propensities for the intra-helical regions were then normalized with the corresponding propensities for loop segments, in order to compare the position preference of each amino acid.

Transmembrane helices were identified by plotting the preceding four parameters into the calculation of the amino acids propensities in each of the three regions [39]. Author claimed that the accuracy of transmembrane helix prediction was as high as 97% [43].

The theory of considering the ends of helices and ends of loops as special regions
in protein sequences has been converted into CRFs features (Border Features) in Chapter 4.2

### 3.2 Advanced Methods

#### 3.2.1 Neural Network Based Method

Burkhard Rost [33] created a neural network system, called PHDhtm, which was trained to recognize the regions of high hydrophobicity and make transmembrane domain predictions. The neural network identifies not only transmembrane helices, but also strand, which is another type of secondary structure. The architecture of their network was shown in Figure 6. In PHDhtm’s framework, a sequence profile of a protein family, rather than just a single sequence, is used as input to a neural network for structure prediction. Each sequence position is represented by the amino acid frequencies derived from multiple sequence alignments, which is taken from the HDSP (Homology Derived Structure of Proteins) database [35].

In a protein sequence, as mentioned in 2, the end with the amine group (−\(NH_2\)) is called N-terminus, and the other end with the carboxyl group (−\(COOH\)) is called C-terminus. In the neural network, the residue frequencies for the 20-residue types are represented by 3 bits each. To code the N-terminus and C-terminus ends an additional 3 bits are used. These 63 bits originating from one sequence position are mapped onto 63 input units for the neural network [31].

The input signal is propagated through a network with three layers (Figure 3.5): one input layer, one hidden layer and one output layer. The output layer has three units corresponding to the three secondary structure states: helix, strand and loop (the "loop" here represents a segment between helices or strands).

We can see from the architecture of the network that there is a net cascade inside its structure (Figure 3.5). The first network (from input layer to hidden layer) is followed by a second network (from hidden layer to output layer) to learn the
structural context. The input to the second network is the three output bits for helix, strand and loop from the first network, plus a fourth spacer unit, for each position in a 17-residue window. From the 17(3+1) = 68 input nodes the signal is propagated via a hidden layer to three output nodes for helix, strand and loop, as in the first network. In prediction mode, a 13-residue sequence window is presented to the network, and the secondary-structure state of the central residue is chosen, according the output unit with the last signal [31].

The network system was trained on the full set of 151 sequence families of known structure and then tested on 26 protein families for which a x-ray or three-dimensional structure was available at that time. In this test, 72% of the observed helical and 68% of strand residues were predicted correctly [35].

### 3.2.2 DAS Method

A considerable constraint of the above PHDhtm method is that it fails to produce significant predictions for proteins lacking significant homologous sequences [32].
The Dense Alignment Surface (DAS) [10, 11] method of transmembrane protein prediction was developed to address the weakness of PHDhtm.

The DAS method is based on a traditional dot-plot of two proteins. If two segments of a certain length of the two proteins have a similarity score which is significance higher than a certain cut-off, that region is marked on the dot-plot. RReM scoring matrix [10] was used in DAS. It is based on the "neighbourhood selectivity" of amino acids pairs (up to 10 residues distant from each other in the sequence). The RReM matrix is a measure of the similarity of the "neighbourhood selectivity" values of the various amino acids to each other.

DAS circumnavigated the need for homologous protein sequences by creating a global transmembrane protein profile. The DAS profile consists of an average of individual cumulative score profiles, calculated for all pair wise comparisons of 44 non-redundant, transmembrane proteins with known topologies. The RReM scoring matrix and default parameters (window size = 10 residues, cut-off value = 1.0) were used in the calculation. By comparing the query sequence to the global profile, consensus local alignment sequence windows were created and global sequence statistics were calculated without requiring a multi-sequence alignment with a set of homologous sequences.

In DAS’s prediction, techniques in other methods were used. It used a sliding window with trapezoid size which is similar to the one in TopPred. To match the window size used in the DAS profile calculation, the window core was set to 9-residue long and full size was 11. It then used TopPred algorithm to predict transmembrane helices. These predictions were transformed to topology profiles by setting a value of "1" for transmembrane regions and "0" for the rest regions. At last, these helices were predicted using PHDhtm server [32]. From the predicted topologies, PHDhtm profiles with a value of 1 for predicted transmembrane regions and 0 elsewhere were generated.

Authors compared the performance of DAS with TopPred and PHDhtm methods, and claimed that DAS method was on par with the PHDhtm method and TopPred
method in predicting transmembrane segments [32].

3.2.3 Hidden Markov Model Based Method

3.2.3.1 TMHMM method

TMHMM is a transmembrane domain identification and characterization programs, which use another form of data-trained method—the Hidden Markov Models—to do their predictions. The basic principle of TMHMM is to define a set of states, each corresponding to a region or specific site in the proteins being modelled. In the simplest case, a model for a transmembrane protein may consist of three states: one for inside loops, one for transmembrane regions, and one for outside loops. Each state has an associated probability distribution over the 20 amino acids, characterising the variability of amino acids in the region it models. The states are connected to each other in a biologically reasonable way, so for instance the state for inside loop is connected to itself (because loops may be longer than 1), and to the transmembrane helix state (because after an inside loop a helix begins). When one state transits to another along the amino acids, a transition probability is associated with this transition. The amino acid probabilities and the transition probabilities are learned by a standard inference technique that computes the maximum posterior probabilities given a prior and the observed frequencies [21].

The architecture of TMHMM model is shown in Figure 3.6. Each box in the drawing corresponds to a submodel designed to model a specific region of a membrane protein. There are seven sub-systems been defined: the helix core, inside loop, outside loops (short and long), helix caps (C-terminus and N-terminus), and globular domains (no distinction between various globular types is made).

To capture the topological signal of proteins, the residues close to the membrane were modelled into the submodels labelled ”loop” and ”cap”, which are shown in Figure 3.7. Loops of lengths up to 20 residues are modelled by the loop model,
Chapter 3. Literature Review in Transmembrane Helix Prediction

Figure 3.6: Architecture of TMHMM) [37]

Figure 3.7: The structure of the submodel for loop regions in TMHMM) [37]

Figure 3.8: The structure of the submodel for helix core regions in TMHMM) [37]
whereas longer loops have to use the "globular" state (Figure 3.6). The "cap" sub-models simply model the five first or last residues of the transmembrane regions.

The submodel for the "core" region of the transmembrane helices is shown in Figure 3.8. It is an array of 25 identical states with the possibility of "jumping" from one state to many of the others down-stream. This topology models sequences of lengths between five and 25, which translates to helix lengths between 15 and 35 when the "caps" are included [21].

The TMHMM was trained on two data sets. One contains 83 proteins and the other contains 160 proteins. Prediction of the transmembrane helices is done by finding the most probable topology given the HMM. Accuracy of the prediction was compared against other two models, PHDhtm and Memsat, and was relatively higher than them. The advantage of this type of prediction method is that it not only provides for the prediction of transmembrane domain regions, but also characterizes the topology of proteins.

The theory of dividing amino acids sequence into several regions, as used in TMHMM, has been converted into features in our CRFs model (Sequence State Features) which is introduced in Chapter 4.2.

### 3.2.3.2 HMMTOP method

HMMTOP [40, 41] differs from TMHMM in the type and number of associated structures that is assigns to non-helix core peptides. HMMTOP uses five states, including the helix core, an inner loop, an inner helix tail, an outer loop and an outer helix tail. HMMTOP does not provide for the differentiation between the outer loops and large globular domains. The layout of HMMTOP model is shown in Figure 3.9.

The HMMTOP model built a Hidden Markov Model architecture which is very similar to TMHMM, but the model used for prediction is different. A model regularizer is estimated from a set of known transmembrane proteins. And for predic-
The model is estimated from the query sequence and then used for predicting the structure of that sequence [43].

The reported topology prediction accuracy of HMMTOP is 78% [40], which is roughly the same as that of TMHMM.

### 3.2.4 Maximum Entropy Markov Model

A conditional model, named Maximum Entropy Markov Model (MEMM) has been introduced into the transmembrane helix prediction problem by Yohan Kim [19].

While the above models use hidden markov model as basis to define transition functions which calculate the joint probability distribution of the observation (amino acids sequences) and label sequences (helical structure sequences), MEMM finds the label distribution conditioned by the observation. In other words, the label assigned at position $n$ depends only on the observation at position $n$ and the previous label at position $n-1$ (Figure 3.10) [26].

The MEMM model in Yohan’s paper [19] is a collection of separately trained tran-
transition functions, \( P_{s'}(s|x) \) and a vector of parameters. The trained function \( P_{s'}(s|x) \) outputs a probability value of seeing a state transition from \( s' \) to \( s \) given observation \( x \). When \( P_{s'}(s|x) \) is summed over \( s \), the sum should be 1 by the definition of the probability function.

The work in this paper did not explicitly incorporate the state transitions. In other words, no functions of \( P_{s'}(s|x) \) are constrained. Instead, only one probability function \( P(s|x) \) is used, and the observation \( x \) is characterized by a set of 300 feature functions. The feature functions in MEMM implicitly capture the dependence on previous states.

The MEMM program was implemented within Matlab environment. Training was stopped after 4 iterations (parameters converged within second decimals). In the result part, author declared that the prediction of MEMM has achieved 58.55% accuracy [19].
3.3 Summary

There are two main approaches of predicting transmembrane helices: one uses hydrophobicity scales, and the other uses advanced model such as neural network and hidden markov model. In this Chapter, we gave a review of both these approaches. Some of their ways of capturing relationships between amino acids and corresponding helical structures have been converted into features in CRFs. These features are introduced in the next Chapter.
Chapter 4

Conditional Random Fields (CRFs)

Conditional Random Fields (CRFs) is a sequence modelling framework which has all the advantages of MEMM (Maximum Entropy Markov Model, introduced in previous chapter). We can think of a CRF as a form of undirected graphical state models that define a single log-linear distribution for the entire label sequence conditioned on the observation sequence [45]. Figure 4.1 demonstrates the undirected graph structure of CRFs. The critical difference between CRFs and MEMM (Maximum Extropy Markov Model) is that a MEMM uses per-state exponential models for the conditional probabilities of next state given the current state, while a CRF has a single exponential model for the joint probability of the entire sequence of labels given the observation sequence [22]. In this chapter, we will give the definition of CRFs model, analyze the parameter optimization approach, and explain the features populated in the model. A large part of this chapter is based on the work of Lior Lukov [25]. However, as we explain below, we will use a much richer set of features.
Figure 4.1: CRFs graphical structure. [25]

The edges between the label nodes $Y$ correspond to the cliques between these random variables. The observation nodes $X$ do not share edges as the model does not assume any independence assumptions among the observation.

### 4.1 Model Definition

A CRF model is an undirected Markov model and the resulting probability distribution is equivalent to a conditional Gibbs distribution [22].

Using the notation of Lafferty [22], the conditional probability density function of a CRF can be described as

$$p(y|x) \propto \exp \left( \sum_{e \in E, j} \lambda_j f_j(e, y_e, x) + \sum_{v \in V, j} \mu_j g_j(v, y_v, x) \right)$$

where $x$ is a data sequence, $y$ a label sequence, $y_S$ is the set of components of $y$ associated with the vertices in subgraph $S$. The vectors $f$ and $g$ represent the local features with corresponding weight vectors $\lambda$ and $\mu$. We assume that the feature functions $f_k$ and $g_k$ are given and fixed [22].

Using the notation of Wallach [44], the joint distribution can be expressed in a slightly different form:

$$p(y|x) \propto \exp \left( \sum_{j} \lambda_j f_j(y_{i-1}, y_i, x, i) + \sum_{k} \mu_k g_k(y_i, x, i) \right) \quad (4.1)$$

where $f_j(y_{i-1}, y_i, x, i)$ is a transition feature function of the entire observation sequence and the labels at positions $i$ and $i-1$ in the label sequence. $g_k(y_i, x, i)$ is a
state feature function of the entire observation sequence and the label at position $i$ in the label sequence. $\lambda_j$ and $\mu_k$ are estimated from the training data [44].

The parameters of the CRFs model are calculated using the principal of the Maximum Entropy theory. The Maximum Entropy theory and Parameter Estimate process are explained in [25].

### 4.2 Feature Integration with the Model

As a statistical framework, CRFs model can be considered as a template which needs to be populated with scenario dependent features. In our context, the term “features” represents functions on the combination of the primary sequences and labels. The features used in our experiments are presented in this section. These features are defined from both statistical and biological perspectives, and are divided into eighteen different groups. Eight features among them, defined in Chapter 4.2.1, Chapter 4.2.2, Chapter 4.2.5, Chapter 4.2.6, Chapter 4.2.9, Chapter 4.2.10, Chapter 4.2.13, and Chapter 4.2.14, are based on the work of Lior Lukov [25]. The other ten more biologically complex features are extracted from previous bioinformatics literature.

#### 4.2.1 Start, End and Edge Features

These features capture the probability of starting/ending a sequence with a given label or the transition probability for moving from one state to the consecutive state [25]. For instance, the start unigram feature has the form:

$$u_{\text{start}}(x, i) = \begin{cases} 
1 & \text{if the Amino Acid in sequence } x \\
& \text{at position } i \text{ is the first in the sequence} \\
0 & \text{otherwise}
\end{cases}$$
The relationship between the observation and the two possible structures, helix membrane/non-helix membrane, is described in the feature:

\[
    f_{\text{start}}(y_i, x, i) = \begin{cases} 
    u_{\text{start}}(x, i) & \text{if } y_i = \text{Helix membrane} \\ 
    0 & \text{otherwise}
    \end{cases}
\]

\[
    f_{\text{start}}(y_i, x, i) = \begin{cases} 
    u_{\text{start}}(x, i) & \text{if } y_i = \text{Non-Helix membrane} \\ 
    0 & \text{otherwise}
    \end{cases}
\]

The Edge feature in contrast, is a bigram feature which depends on two consecutive labels:\(^1\):

\[
    f_{\text{edge}}(y_{i-1}, y_i, x, i) = \begin{cases} 
    u_{\text{edge}}(x, i) & \text{if } y_{i-1} = \text{Helix membrane}
    \quad \text{and } y_i = \text{Helix membrane} \\ 
    0 & \text{otherwise}
    \end{cases}
\]

### 4.2.2 Basic Amino Acid Features

There twenty different kinds of amino acids in protein primary sequence, so we have 20 different unigram features from this type [25]. The unigram feature of amino acid \(n\) in position \(i\) is:

\[
    u_n(x, i) = \begin{cases} 
    1 & \text{if the Amino Acid in sequence } x \\
    & \text{at position } i \text{ is from type } m \\
    0 & \text{otherwise}
    \end{cases}
\]

Using this unigram, a feature for describing the relationship between the observation and the two possible structures has the form:

\[
    f_{nH}(y_i, x, i) = \begin{cases} 
    u_n(x, i) & \text{if } y_i = \text{Helix membrane} \\ 
    0 & \text{otherwise}
    \end{cases}
\]

\[
    f_{nNH}(y_i, x, i) = \begin{cases} 
    u_n(x, i) & \text{if } y_i = \text{Non-Helix membrane} \\ 
    0 & \text{otherwise}
    \end{cases}
\]

\(^1\)This is an example of \(H − H\) edge feature. The other three edges features, \(H − NH\), \(NH − H\) and \(NH − NH\) are all similar to this one [25].
4.2.3 Hydrophobic Window Features

As discussed in chapter 3.1, hydrophobicity analysis has been proved to be an efficient way to detect transmembrane helices[43]. Due to Kyte and Doolittle’s work[16], each amino acid has been endowed with a unique hydrophobic value. The values of hydrophobic index are shown in Table 3.1.

In this feature a sliding window, which consists of 19 residues, is used to test if the average hydrophobic value of the 19 amino acids inside the window is greater than a certain "threshold". In order to find the most effective threshold, 400 helix-segments and 400 loop-segments are randomly selected from the training dataset (The training dataset is introduced in chapter 5). We calculate and compare the average hydrophobic values of these two types of segments.

As shown in Figure 4.2, most of the average hydrophobic values of helix-segments are higher than 1.0, while in contrast the average hydrophobic values of loop-segments are mostly lower than that value. So we decide to use “1.0” as the threshold of this feature.

Figure 4.2: Comparison of Average Hydrophobic Values between Helix Segments and Non-helix Segments
Chapter 4. Conditional Random Fields (CRFs)

In conclusion, the hydrophobic window feature of amino acid n in position i is:

\[
\begin{align*}
\mathbf{f}_{\text{hydrophobic-window}}(y_i, x, i) &= \begin{cases} 
1 & \text{if amino acid } i \text{ is in a helix segment and} \\
& \text{in a sliding window which has an average} \\
& \text{hydrophobic value higher than 1.0} \\
0 & \text{otherwise}
\end{cases}
\end{align*}
\]

4.2.4 Hydrophilic Window Features

This time it is the hydrophilic amino acids which present in loop segments (a loop is the segment between two helix segments) which we care about. We use the same sliding window to test if the average hydrophobic value is lower than the preceding "threshold". This feature can be described as following:

\[
\begin{align*}
\mathbf{f}_{\text{hydrophilic-window}}(y_i, x, i) &= \begin{cases} 
1 & \text{if amino acid } i \text{ is in a loop segment and} \\
& \text{in a sliding window which has an average} \\
& \text{hydrophobic value lower than 1.0} \\
0 & \text{otherwise}
\end{cases}
\end{align*}
\]

4.2.5 Single Side Neighboring Amino Acid Features

By using this feature, we try to capture the tendency of the same amino acid given its adjacent neighbors [25]. Suppose our current amino acid is \(A_i\), a feature is created to capture the tendency of \(A_i\) to appear in a helix given the previous acid \(A_{i-1}\). In the same way, we also created additional features of \(A_i\) given \(A_{i-1}, A_{i-2}, \ldots, A_{i-k}\). Similar features are also applicable for the opposite single side of \(A_i\) given \(A_{i+1}, \ldots, A_{i+k}\) [25].
4.2.6 Single Side Shuffled Neighboring Amino Acid Features

The motivation of creating the shuffled features is based on the hypothesis that the locations of the transmembrane regions are determined by the difference in the amino acid distribution in various structural parts of the protein rather than by specific amino acid composition of these parts [25]. We are interested in capturing the tendency of the same amino acid given its adjacent neighbors without concerning their order.

Suppose our current amino acid is \( A_i \), a shuffled feature is created to capture the tendency of \( A_i \) to appear in a helix given \( A_{i-1} \). In the same way we also created additional features of \( A_i \) given \( A_{i-1} \cup A_{i-2} \), up to \( k \)-feature of \( A_i \) given \( A_{i-1} \cup, ..., \cup A_{i-k} \). Similar features are also applicable for the opposite single side of \( A_i \) given \( A_{i+1} \cup, ..., \cup A_{i+k} \).

4.2.7 Single Side Hydrophobic Neighbouring Amino Acid Features

Similar to Single Side Shuffled Neighbouring Amino Acid Features, but this time we want to capture not only the composition of the adjacent amino acids, but also the hydrophobic tendency of these neighbouring residues. Since the hydrophobic value of a certain segment may affect the formation of a transmembrane helix[43], the hydrophobicity of neighbouring residues may also determine a single amino acid to be in a helix or not.

Suppose our current amino acid is \( A_i \), and its hydrophobic value is \( H_i \), we create this feature to capture the tendency of \( A_i \) to appear in a helix given \( A_{i-1} \) and its hydrophobic value \( H_{i-1} \). In the same way we have also created additional features of \( A_i \) given \( A_{i-1} \cup A_{i-2} \) and \( H_{i-1}, H_{i-2} \), up to \( k \)-feature of \( A_i \) given \( A_{i-1} \cup, ..., \cup A_{i-k} \), and \( H_{i-1}, ..., H_{i-k} \). We use the same “threshold” (1.0) as in Hydrophobic
4.2.8 Single Side Hydrophilic Neighbouring Amino Acid Features

This time we care about the neighbouring hydrophilic amino acids which present in loop-segments. The same "threshold" is used here again, and the Single Side Hydrophilic Neighbouring Amino Acid Feature here is only active when the average hydrophobic value of these k neighbouring residues is lower than 1.0.

4.2.9 Double Side Neighboring Amino Acid Features

Besides the single side features, we are also interested in capturing the tendency of an amino acid given its adjacent neighbors from both sides together. Suppose our current amino acid is $A_i$, a double side feature is defined to capture the tendency of $A_i$ to appear in a helix segment given $A_{i-1}, A_{i-2}, \ldots, A_{i-k}$ and $A_{i+1}, A_{i+2}, \ldots, A_{i+k}$ [25].

The multigram feature of amino acid $n$ in position $i$ is:

$$m_{i-1,i+1}(x, i) = \begin{cases} 
1 & \text{if the Amino Acid in sequence $x$ at position $(i - 1), i, (i + 1)$ is from type $m_{i-1}, m_i, m_{i+1}$ respectively} \\
0 & \text{otherwise}
\end{cases}$$

Using this multigram, a feature for describing the relationship between the observation and the two possible structures has the form:

$$f_{mn}(y_i, x, i) = \begin{cases} 
m_{i-1,i+1}(x, i) & \text{if } y_i = \text{Helix membrane} \\
0 & \text{otherwise}
\end{cases}$$
4.2.10 Double Side Shuffled Neighboring Amino Acid Features

Similar to Double Side Neighboring Amino Acid Features, the double side shuffled feature is defined to capture the tendency of an amino acid given it’s adjacent neighbors from both sides together without concerning their order [25].

Suppose our current amino acid is $A_i$, we define a feature which captures the tendency of $A_i$ to appear in a helix given $A_{i-1} \cup \ldots \cup A_{i-k}$ and $A_{i+1} \cup \ldots \cup A_{i+k}$.

The multigram feature of amino acid $n$ in position $i$ is:

$$m_{i-1,i+i}(x, i) = \begin{cases} 1 & \text{if the Amino Acid in sequence } x \text{ at position } (i - 1), i, (i + 1) \text{ is from type } m_{i-1} \cup m_{i} \cup m_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

Using this multigram, a feature for describing the relationship between the observation and the two possible structures has the form:

$$f_{m_{NH}}(y_i, x, i) = \begin{cases} m_{i-1,i+i}(x, i) & \text{if } y_i = \text{Helix membrane} \\ 0 & \text{otherwise} \end{cases}$$

$$f_{m_{NH}}(y_i, x, i) = \begin{cases} m_{i-1,i+i}(x, i) & \text{if } y_i = \text{Non-Helix membrane} \\ 0 & \text{otherwise} \end{cases}$$
4.2.11 Double Side Hydrophobic Neighbouring Amino Acid Features

Similar to Double Side Shuffled Neighbouring Amino Acid Features, but this time we want to capture not only the composition of the adjacent amino acids, but also the hydrophobic tendency of these neighbouring residues from both sides. Suppose our current amino acid is $A_i$, this feature is active only when the $2k + 1$ neighbouring residues $A_{i-k} \cup \ldots, A_i \cup A_{i+1} \cup \ldots, A_{i+k}$ have an average hydrophobic value higher than 1.0.

4.2.12 Double Side Hydrophilic Neighbouring Amino Acid Features

Similar to Double Side Hydrophobic Neighbouring Amino Acid Features, but this time we care about the hydrophilic amino acids which present in loop-segments. Suppose our current amino acid is $A_i$, this feature is active only when the $2k + 1$ neighbouring residues $A_{i-k} \cup \ldots, A_i \cup A_{i+1} \cup \ldots, A_{i+k}$ have an average hydrophobic value lower than 1.0.

4.2.13 Amino Acid Property Features

As discussed in chapter 2.1, the twenty kinds of amino acids are classified into nine groups [38]. The fact that amino acids from the same classification group often appear in similar locations motivated us to create special features for the amino acids properties [25]. Each group is described by a unigram feature. For example, the hydrophobic property is described in the feature:

$$u_{\text{Hydrophobic}}(x, i) = \begin{cases} 
1 & \text{if the Amino Acid in sequence } x \text{ at position } i \in \{M,I,L,V,A,G,F,W,Y,H,K,C\} \\
0 & \text{otherwise}
\end{cases}$$
The relationship between observation and label for hydrophobicity property is:

\[
\begin{align*}
  f_{\text{hydrophobic}}(y_i, x, i) &= \begin{cases} 
    u_{\text{Hydrophobic}}(x, i) & \text{if } y_i = \text{Helix} \\
    0 & \text{otherwise}
  \end{cases} \\
  f_{\text{hydrophobic}}(y_i, x, i) &= \begin{cases} 
    u_{\text{Hydrophobic}}(x, i) & \text{if } y_i = \text{Non-Helix} \\
    0 & \text{otherwise}
  \end{cases}
\end{align*}
\]

**4.2.14 Border Features**

We define this feature to capture the border between a segment of amino acids labelled with one structure and an adjacent segment labelled with another [25]. The relationship between the observation sequence labelled with the two possible structures, helixtonon−helix and non−helixtohelix are captured by the following two features:

\[
\begin{align*}
  f_{\text{border}_{\text{H}}-\text{NH}}(y_i, x, i) &= \begin{cases} 
    1 & \text{if } y_{i-j}, ..., y_{i-1} = \text{Helix segment} \\
    \quad \text{and } y_i, ..., y_{i+j} = \text{Non-helix segment} \\
    0 & \text{otherwise}
  \end{cases} \\
  f_{\text{border}_{\text{NH}}-\text{H}}(y_i, x, i) &= \begin{cases} 
    1 & \text{if } y_{i-j}, ..., y_{i-1} = \text{Non-helix segment} \\
    \quad \text{and } y_i, ..., y_{i+j} = \text{Helix segment} \\
    0 & \text{otherwise}
  \end{cases}
\end{align*}
\]

**4.2.15 Short Loop Features**

This feature aimed at capturing the composition of short loops, which appear to be very difficult to predict especially when they are shorter than 7 residues[5]. In this feature, we don’t concern about the order of the amino acids in a short loop,
instead, it is the composition in the short loops that we care about. This feature
can be described as:
\[
f_{\text{short-loops}}(y_i, x, i) = \begin{cases} 
1 & \text{if the Amino Acid in sequence } x \text{ at position } i \\
& \text{is from a loop segment which is shorter than } n \text{ residues. (} n = 1, 2, 3, \ldots, 7) \\
0 & \text{otherwise}
\end{cases}
\]

4.2.16 Electron Transport Chain Features

Electron carriers and biochemical reactions are associated by electron transport
chains, in which amino acids in protein sequence also take part [14]. From this
point of view, we classify the 20 amino acids into five groups, as shown in Table
4.1, and create this feature to capture the influence of electron transport in helices
formation.

<table>
<thead>
<tr>
<th>Electronic Property</th>
<th>Amino Acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong donor</td>
<td>A, D, E, P</td>
</tr>
<tr>
<td>Weak donor</td>
<td>I, L, V</td>
</tr>
<tr>
<td>Neutral</td>
<td>C, G, H, S, W, M</td>
</tr>
<tr>
<td>Weak acceptor</td>
<td>F, Q, T, Y</td>
</tr>
<tr>
<td>Strong acceptor</td>
<td>K, N, R</td>
</tr>
</tbody>
</table>

The unigram feature of amino acid \( n \) in position \( i \) is:
\[
u_n(x, i) = \begin{cases} 
1 & \text{if the Amino Acid in sequence } x \\
& \text{at position } i \text{ is from Electronic Property } e \\
0 & \text{otherwise}
\end{cases}
\]

Using this unigram, a feature for describing the relationship between the observa-
tion and the two possible structures has the form:
\[
f_{\text{elec}}(y_i, x, i) = \begin{cases} 
u_n(x, i) & \text{if } y_i = \text{Helix membrane} \\
0 & \text{otherwise}
\end{cases}
\]
\[ f_{\text{elec}}(y_i, x, i) = \begin{cases} u_n(x, i) & \text{if } y_i = \text{Non-Helix membrane} \\ 0 & \text{otherwise} \end{cases} \]

### 4.2.17 Chemical Groups Features

As discussed in chapter 2.1, twenty different side chains (see Figure 2.2) bring on twenty various amino acids. We tried to analyse the twenty sidechains (see Table 4.2) and defined corresponding features to capture the information displayed by those side chains.

<table>
<thead>
<tr>
<th>AminoAcids</th>
<th>SideChainComponents</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>-COO(^-), -CH(_2)(^-)</td>
</tr>
<tr>
<td>E</td>
<td>-COO(^-), -CH(_2)(^-)</td>
</tr>
<tr>
<td>L</td>
<td>-CH(_3), -CH(^-), -CH(_2)(^-)</td>
</tr>
<tr>
<td>V</td>
<td>-CH(_3), -CH(^-)</td>
</tr>
<tr>
<td>C</td>
<td>-SH(^-), -CH(_2)(^-)</td>
</tr>
<tr>
<td>W</td>
<td>-CH(_2)(^-), = C(\text{aromatic})(^-), -NH(\text{ring})(^-), = C(\text{aromatic}) (-)</td>
</tr>
<tr>
<td>A</td>
<td>-CH(_3)</td>
</tr>
<tr>
<td>P</td>
<td>-CH(_2)(\text{ring})(^-), -CH(\text{ring})(^-), -NH(\text{ring})(^-)</td>
</tr>
<tr>
<td>I</td>
<td>-CH(_3), -CH(_2)(^-), -CH(^-)</td>
</tr>
<tr>
<td>S</td>
<td>-OH(^-), -CH(_2)(^-)</td>
</tr>
<tr>
<td>G</td>
<td>-H</td>
</tr>
<tr>
<td>K</td>
<td>-NH(_3)(^+), -CH(_2)(^-)</td>
</tr>
<tr>
<td>N</td>
<td>-C = O, -NH(_2), -CH(_2)(^-)</td>
</tr>
<tr>
<td>R</td>
<td>= NH(_2)(^+), = NH(<em>2), -C = - , = NH(</em>-), -CH(_2)(^-)</td>
</tr>
<tr>
<td>T</td>
<td>-OH(^-), -CH(_3), -CH(^-)</td>
</tr>
<tr>
<td>Q</td>
<td>= C = O, = NH(_2)(^-), = CH(_2)(^-)</td>
</tr>
<tr>
<td>F</td>
<td>= CH(\text{aromatic})(^-), = C(\text{aromatic}) (-), = CH(_2)(^-)</td>
</tr>
<tr>
<td>H</td>
<td>= NH(\text{ring})(^-), = N(^-), = CH(\text{aromatic}) (-), = C(\text{aromatic}) (-), = CH(_2)(^-)</td>
</tr>
<tr>
<td>Y</td>
<td>= C(\text{aromatic}) (-), = CH(\text{aromatic}) (-), = CH(_2)(^-), = OH</td>
</tr>
<tr>
<td>M</td>
<td>-CH(_3), -CH(_2)(^-)</td>
</tr>
</tbody>
</table>

Based on these chemical components in each sidechain, we classify the twenty amino acids into 18 groups, as shown in Table 4.3:
### Table 4.3: Classification of Amino Acids by the Chemical Component in Their Side Chains

<table>
<thead>
<tr>
<th>GroupNumber</th>
<th>ChemicalComponents</th>
<th>AminoAcids</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>− − C − −</td>
<td>R</td>
</tr>
<tr>
<td>2</td>
<td>= C_{aromatic} − −</td>
<td>Y, F, H, W</td>
</tr>
<tr>
<td>3</td>
<td>− − CH−</td>
<td>L, V, I, T</td>
</tr>
<tr>
<td>4</td>
<td>−CH$_2$−</td>
<td>K, N, D, E, L, C, W, S, I, R, Q, F, H, Y</td>
</tr>
<tr>
<td>5</td>
<td>−CH$_2^{ring}$−</td>
<td>P</td>
</tr>
<tr>
<td>6</td>
<td>−CH$_3$</td>
<td>L, V, I, A, T, M</td>
</tr>
<tr>
<td>7</td>
<td>= CH$_{aromatic}$−</td>
<td>W, F, Y, H</td>
</tr>
<tr>
<td>8</td>
<td>− − CH$_2^{ring}$</td>
<td>W, P</td>
</tr>
<tr>
<td>9</td>
<td>− − C = O</td>
<td>N, Q</td>
</tr>
<tr>
<td>10</td>
<td>−COO−</td>
<td>D, E</td>
</tr>
<tr>
<td>11</td>
<td>= N−</td>
<td>H</td>
</tr>
<tr>
<td>12</td>
<td>−NH−</td>
<td>R</td>
</tr>
<tr>
<td>13</td>
<td>−NH$_2$</td>
<td>N, R, Q</td>
</tr>
<tr>
<td>14</td>
<td>= NH$_2$+</td>
<td>R</td>
</tr>
<tr>
<td>15</td>
<td>−NH$_3$+</td>
<td>K</td>
</tr>
<tr>
<td>16</td>
<td>−OH</td>
<td>S, T, Y</td>
</tr>
<tr>
<td>17</td>
<td>−SH</td>
<td>C</td>
</tr>
<tr>
<td>18</td>
<td>−NH$_3^{ring}$−</td>
<td>P, H, W</td>
</tr>
</tbody>
</table>

The unigram feature of amino acid $n$ in position $i$ is:

$$u_n(x, i) = \begin{cases} 
1 & \text{if the Amino Acid in sequence } x \\
& \text{at position } i \text{ belongs to chemical group } c \\
0 & \text{otherwise}
\end{cases}$$

Using this unigram, a feature for describing the relationship between the observation and the two possible structures has the form:

$$f_{c_H}(y_i, x, i) = \begin{cases} 
1 & \text{if } y_i = \text{Helix membrane} \\
0 & \text{otherwise}
\end{cases}$$

$$f_{c_{NH}}(y_i, x, i) = \begin{cases} 
1 & \text{if } y_i = \text{Non-Helix membrane} \\
0 & \text{otherwise}
\end{cases}$$
4.2.18 Sequence States Features

The sequence states mechanism in TMHMM [21, 37] have been adopted here. We divide protein sequences into three states: Helix Core (8 residues in the centre of a transmembrane helix segment), Helix Ends (5 residues at the ends of a transmembrane helix segment), Loops Ends (5 residues at the ends of a loop segment). This feature is based on the hypothesis that the differences between the amino acids distributions in the various structural parts are one of the driving forces in the formation of the transmembrane helices. The unigram feature of amino acid \( n \) in position \( i \) is:

\[
u_n(x, i) = \begin{cases} 
1 & \text{if the Amino Acid in sequence } x \\
& \text{at position } i \text{ belongs to state } s \\
0 & \text{otherwise}
\end{cases}
\]

Using this unigram, a feature for describing the relationship between the observation and the two possible structures has the form:

\[
f_{sH}(y_i, x, i) = \begin{cases} 
u_n(x, i) & \text{if } y_i = \text{Helix membrane} \\
0 & \text{otherwise}
\end{cases}
\]

\[
f_{sNH}(y_i, x, i) = \begin{cases} 
u_n(x, i) & \text{if } y_i = \text{Non-Helix membrane} \\
0 & \text{otherwise}
\end{cases}
\]

4.3 Summary

Conditional Random Fields (CRFs) are a probabilistic framework from the family of conditional models for labelling sequential data. The main principal in CRFs is to create a model that satisfies all known constraints. The parameter estimation problem is to determine each feature’s parameters from the training data. The most important aspect of specifying the model is selecting a set of features that captures the relationships among the observation and the label sequences. In this chapter we have created 18 types of features. Integrating these features into CRFs and carrying out experiments on them are the main topics of the next chapter.
Chapter 5

Experiments, Evaluation and Analysis

In this chapter, we present the experiments that we carried out to assess the performance of the CRFs model and the efficacy of the features we defined in the previous chapter. Similar to the work of Lukov [25] we have carried out two sets of experiments.

The first set of experiments is the comparison of different feature selection strategies and their effect on the prediction accuracy. The CRFs with the best set of features were then compared with twenty eight other models at the “Static Benchmarking” web server [2]. The second set of experiments are to compare CRFs model against MEMM (Maximum Entropy Markov Model) from the perspective of not only prediction accuracy, but also the differences in the length of training time and the labelling time etc.
5.1 Data Set

5.1.1 Training Data

The CRFs model were trained on a data set consisting of a set of benchmark sequences with experimentally confirmed transmembrane regions compiled by Möller et al. [28]. We have only included proteins with a high level of trust (assigned with transmembrane annotation trust level of A to C, as was suggested by Möller et al.) and which are significantly different, based on pairwise similarity clustering. The resulting set consists of 148 transmembrane protein sequences.

5.1.2 Test Data

The CRFs model were test on a data set consisting of 2248 observation sequences, which was download from the "Static benchmarking of membrane helix predictions" website hosted in Columbia University [2].

5.2 Prediction Metrics

The metrics used in this thesis are those commonly used in the previous transmembrane helix prediction literature. There are two main approaches to rank a prediction model for membrane helices: per-residue and per-segment accuracy [4, 6].

---

1The data set can be accessed via ftp://ftp.ebi.ac.uk/pub/databases/testsets/transmembrane/sequences
2The data set can be accessed via http://cubic.bioc.columbia.edu/cgi/var/kernytsky/tmh/advanced.cgi
5.2.1 Segment Level Accuracy

$Q_{ok}$: Percentage of proteins for which all membrane helices are predicted correctly. It is computed as follows:

$$Q_{ok} = 100 \times \frac{\text{number of proteins which have their helices all predicted correctly}}{\text{total number of tested proteins}}$$

$Q_{\%obs \_htm}$: Percentage of all observed helices that are predicted correctly. It is computed as follows:

$$Q_{\%obs \_htm} = 100 \times \frac{\text{number of correctly predicted helices}}{\text{number of observed helix segments}}$$

$Q_{\%prd \_htm}$: Percentage of all predicted helices that are predicted correctly. It is computed as follows:

$$Q_{\%prd \_htm} = 100 \times \frac{\text{number of correctly predicted helices}}{\text{number of predicted helix segments}}$$

5.2.2 Residue Level Accuracy

$Q_2$: Percentage of correctly predicted residues. It is computed as follows:

$$Q_2 = 100 \times \frac{\text{number of correctly predicted residues}}{\text{total number of residues}}$$

$Q_{\%obs \_2T}$: Percentage of all observed helix residues that are correctly predicted. It is computed as follows:

$$Q_{\%obs \_2T} = 100 \times \frac{\text{number of residues predicted correctly as helices}}{\text{total number of residues that are actually helices}}$$

$Q_{\%prd \_2T}$: Percentage of all predicted helix residues that are correctly predicted. It is computed as follows:

$$Q_{\%prd \_2T} = 100 \times \frac{\text{number of residues predicted correctly as helices}}{\text{total number of residues predicted as helices}}$$
\( Q_{2N}^{\text{obs}} \): Percentage of all observed non-helix residues that are correctly predicted. It is computed as follows:

\[
Q_{2N}^{\text{obs}} = 100 \times \frac{\text{number of residues predicted correctly as Non-helices}}{\text{total number of residues that are actually Non-helices}}
\]

\( Q_{2N}^{\text{prd}} \): Percentage of all predicted non-helix residues that are correctly predicted. It is computed as follows:

\[
Q_{2N}^{\text{prd}} = 100 \times \frac{\text{number of residues predicted correctly as Non-helices}}{\text{total number of residues predicted as Non-helices}}
\]

### 5.3 Transmembrane Helix Prediction

The development of our codes which we used to carry out our experiments was based on the JAVA CRFs implementation package of Conditional Random Fields for sequential labelling, developed by Sunita Sarawagi [36].

Using different selections of features, we set up eight experiments to test the performance of CRFs. The eight different combinations of features are described in Table 5.1.

We send our prediction results of the eight experiments to the "Static benchmarking of membrane helix predictions" website [2], and obtained detailed accuracy of their performances, which are shown in Figure 5.1.

As we can see from these accuracies, there are three main kinds of features which contribute the most in capturing useful information from training data: Hydrophobic-based features, neighbouring-acids-based features and sequence-state-based features.

After adding single and double side neighbouring features, \( Q_{ok} \) increases from 27% to 63% and \( Q_2 \) increases from 65% to 80%. After adding single and double shuffle neighbouring feature, \( Q_{ok} \) increases to 75%. After adding hydrophobic neighbouring features, \( Q_{ok} \) increases from 75% to 80% and \( Q_2 \) increases from
Table 5.1: Enabled and disabled feature combination.

<table>
<thead>
<tr>
<th>Features</th>
<th>Exp 1</th>
<th>Exp 2</th>
<th>Exp 3</th>
<th>Exp 4</th>
<th>Exp 5</th>
<th>Exp 6</th>
<th>Exp 7</th>
<th>Exp 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Properties</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Hydrophobic Windows</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Hydrophilic Windows</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Single</td>
<td>-</td>
<td>+2</td>
<td>+5</td>
<td>+3</td>
<td>+5</td>
<td>+5</td>
<td>+5</td>
<td>+5</td>
</tr>
<tr>
<td>Double</td>
<td>-</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>+3</td>
<td>+3</td>
<td>+3</td>
<td>+3</td>
</tr>
<tr>
<td>Single Shuffled</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+3</td>
<td>+6</td>
<td>+6</td>
<td>+6</td>
<td>+6</td>
</tr>
<tr>
<td>Double Shuffled</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+1</td>
<td>+3</td>
<td>+3</td>
<td>+3</td>
<td>+3</td>
</tr>
<tr>
<td>Single Hydrophobic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+3</td>
<td>+6</td>
<td>+6</td>
</tr>
<tr>
<td>Double Hydrophobic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+1</td>
<td>+3</td>
<td>+3</td>
</tr>
<tr>
<td>Single Hydrophilic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+1</td>
<td>+3</td>
</tr>
<tr>
<td>Double Hydrophilic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+1</td>
<td>+3</td>
</tr>
<tr>
<td>Border</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Short Loops</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Electronic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Groups</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>States</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

Basic = Basic Amino Acid Features, Properties = Amino Acid Property Features, Hydrophobic Windows = Hydrophobic Window Features, Hydrophilic Windows = Hydrophilic Window Features, Single = Single Side Neighboring Amino Acid Features (with 1 to 5 neighbors from left or right), Double = Double Side Neighboring Amino Acid Features (with 1 to 3 neighbors from both sides), Single Shuffled = Single Side Shuffled Neighboring Amino Acid Features (with 1 to 6 neighbors from left or right), Double Shuffled = Double Side Shuffled Neighboring Amino Acid Features (with 1 to 3 neighbors from both sides), Single Hydrophobic = Single Side Hydrophobic Neighbouring Amino Acid Features (with 1 to 6 neighbors from left or right), Double Hydrophobic = Double Side Hydrophobic Neighbouring Amino Acid Features (with 1 to 3 neighbors from both sides), Border = Border Features, Short Loops = Short Loop Features, Electronic = Electron Transport Chain Features, Groups = Chemical Group Features, States = Sequence States Features.
This phenomena may indicate three rules of the micro-level behaviour of proteins:

1. Hydrophobic values of amino acids is an significant factor of locating helical segments [43];

2. The helical structure of an particular amino acid is greatly influenced by its neighbouring amino acids;

3. Differences between the amino acids distributions in the various structural parts are one of the driving forces in the formation of the transmembrane helices [21, 37].

We also compared our best performance (Experiment 8) with other twenty eight methods on the “Benchmark Server”. As shown in Table 5.2 the CRFs ranks the first at both the per-segment ($Q_{ok}$) and per-residue ($Q_2$) accuracy.
Table 5.2: Prediction score comparison between 29 methods.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Per-Segment</th>
<th>Per-Residue</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Qok</td>
<td>Q_{obs}^{%}</td>
</tr>
<tr>
<td>CRFs</td>
<td>88</td>
<td>99</td>
</tr>
<tr>
<td>PHDpsihm08</td>
<td>84</td>
<td>99</td>
</tr>
<tr>
<td>HMMTOP2</td>
<td>83</td>
<td>99</td>
</tr>
<tr>
<td>DAS</td>
<td>79</td>
<td>99</td>
</tr>
<tr>
<td>TopPred2</td>
<td>75</td>
<td>90</td>
</tr>
<tr>
<td>TMHMM1</td>
<td>71</td>
<td>90</td>
</tr>
<tr>
<td>SOSUI</td>
<td>71</td>
<td>88</td>
</tr>
<tr>
<td>PHDhtm07</td>
<td>69</td>
<td>83</td>
</tr>
<tr>
<td>KD</td>
<td>65</td>
<td>94</td>
</tr>
<tr>
<td>PHDhtm08</td>
<td>64</td>
<td>77</td>
</tr>
<tr>
<td>GES</td>
<td>64</td>
<td>97</td>
</tr>
<tr>
<td>PRED-TMR</td>
<td>61</td>
<td>84</td>
</tr>
<tr>
<td>Ben-Tal</td>
<td>60</td>
<td>79</td>
</tr>
<tr>
<td>Eisenberg</td>
<td>58</td>
<td>95</td>
</tr>
<tr>
<td>Hopp-Woods</td>
<td>56</td>
<td>93</td>
</tr>
<tr>
<td>WW</td>
<td>54</td>
<td>95</td>
</tr>
<tr>
<td>Roseman</td>
<td>52</td>
<td>94</td>
</tr>
<tr>
<td>Av-Cid</td>
<td>52</td>
<td>93</td>
</tr>
<tr>
<td>Levitt</td>
<td>48</td>
<td>91</td>
</tr>
<tr>
<td>A-Cid</td>
<td>47</td>
<td>95</td>
</tr>
<tr>
<td>Heijne</td>
<td>45</td>
<td>93</td>
</tr>
<tr>
<td>Bull-Breese</td>
<td>45</td>
<td>92</td>
</tr>
<tr>
<td>Sweet</td>
<td>43</td>
<td>90</td>
</tr>
<tr>
<td>Radzicka</td>
<td>40</td>
<td>93</td>
</tr>
<tr>
<td>Nakashima</td>
<td>39</td>
<td>88</td>
</tr>
<tr>
<td>Fauchere</td>
<td>36</td>
<td>92</td>
</tr>
<tr>
<td>Lawson</td>
<td>33</td>
<td>86</td>
</tr>
<tr>
<td>EM</td>
<td>31</td>
<td>92</td>
</tr>
<tr>
<td>Wolfenden</td>
<td>28</td>
<td>43</td>
</tr>
</tbody>
</table>

Obtained from the "Static benchmarking of membrane helix predictions" server (sorted by $Q_{ok}$)\cite{2}. 
The codes which we used to carry out experiments on Maximum Entropy Markov Model (MEMM) were based on the java MaxEnt implementation package, developed by Dan Klein [20].

In this section, we compare CRFs against MEMM from three perspectives: training process, labelling process and prediction accuracy. We rebuilt the eight preceding experiments on MEMM, and analyzed the differences of these two models on training time (Figure 5.2) and labelling time (Figure 5.3). From these comparisons, we can see that CRFs take much longer time in the training process, but takes shorter time in labelling stage.

After we rebuilt these experiments, we also submitted the prediction results to the "Static benchmarking of membrane helix predictions" server. The MEMM prediction accuracy generated from this server demonstrated that CRFs outperform MEMM in all of our experiments.

3 Machine Configuration: Pentium 4 Processor (3.0GHz), 1GB RAM
Figure 5.3: Comparison between CRFs and MEMM on Labelling Time

Accuracy comparison between these two models are shown in Figure 5.4 to 5.11, and summarized in Table 5.3.

In order to do the comparison we will test the null hypothesis that the difference of the performance metrics between CRFs and MEMMs is less than or equal to zero. This can be stated as:

\[
\mu^d = \mu^\text{CRFs} - \mu^\text{MEMMs}
\]

\[H_0: \mu^d \leq 0 \quad (5.1)\]

\[H_a: \mu^d > 0 \quad (5.2)\]

where

\[d_1 = Q^\text{CRFs}_{ok} - Q^\text{MEMMs}_{ok}, \ldots, d_8 = Q^{%\text{prd}^\text{CRFs}}_{2N} - Q^{%\text{prd}^\text{MEMMs}}_{2N}\]

We took a \( t \)–test on each pair of the prediction results in Table 5.3, and obtained the \( p \)–Values for each Experiment, as shown in Table 5.4.
Chapter 5. Experiments, Evaluation and Analysis

Figure 5.4: Accuracy Comparison between CRFs and MEMM on Experiment 1

Figure 5.5: Accuracy Comparison between CRFs and MEMM on Experiment 2
Chapter 5. Experiments, Evaluation and Analysis

Figure 5.6: Accuracy Comparison between CRFs and MEMM on Experiment 3

Figure 5.7: Accuracy Comparison between CRFs and MEMM on Experiment 4
Figure 5.8: Accuracy Comparison between CRFs and MEMM on Experiment 5

Figure 5.9: Accuracy Comparison between CRFs and MEMM on Experiment 6
Chapter 5. Experiments, Evaluation and Analysis

Figure 5.10: Accuracy Comparison between CRFs and MEMM on Experiment 7

Figure 5.11: Accuracy Comparison between CRFs and MEMM on Experiment 8
Table 5.3: Prediction Accuracy Comparison between CRFs and MEMM

<table>
<thead>
<tr>
<th>Experiments</th>
<th>( Q_{\text{obs}} )</th>
<th>( Q_{\text{obs}}^{% \text{emb}} )</th>
<th>( Q_{\text{emb}}^{% \text{emb}} )</th>
<th>( Q_2 )</th>
<th>( Q_{2T}^{% \text{obs}} )</th>
<th>( Q_{2T}^{% \text{emb}} )</th>
<th>( Q_{2N}^{% \text{obs}} )</th>
<th>( Q_{2N}^{% \text{emb}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1\text{CRFs}</td>
<td>27</td>
<td>46</td>
<td>69</td>
<td>65</td>
<td>42</td>
<td>57</td>
<td>88</td>
<td>60</td>
</tr>
<tr>
<td>1\text{MEMM}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>48</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>48</td>
</tr>
<tr>
<td>2\text{CRFs}</td>
<td>63</td>
<td>88</td>
<td>95</td>
<td>80</td>
<td>76</td>
<td>81</td>
<td>81</td>
<td>77</td>
</tr>
<tr>
<td>2\text{MEMM}</td>
<td>0</td>
<td>3</td>
<td>18</td>
<td>49</td>
<td>3</td>
<td>15</td>
<td>97</td>
<td>48</td>
</tr>
<tr>
<td>3\text{CRFs}</td>
<td>69</td>
<td>86</td>
<td>87</td>
<td>81</td>
<td>72</td>
<td>79</td>
<td>90</td>
<td>77</td>
</tr>
<tr>
<td>3\text{MEMM}</td>
<td>61</td>
<td>80</td>
<td>78</td>
<td>80</td>
<td>68</td>
<td>81</td>
<td>89</td>
<td>72</td>
</tr>
<tr>
<td>4\text{CRFs}</td>
<td>72</td>
<td>89</td>
<td>90</td>
<td>82</td>
<td>74</td>
<td>82</td>
<td>89</td>
<td>78</td>
</tr>
<tr>
<td>4\text{MEMM}</td>
<td>63</td>
<td>69</td>
<td>78</td>
<td>79</td>
<td>63</td>
<td>72</td>
<td>91</td>
<td>76</td>
</tr>
<tr>
<td>5\text{CRFs}</td>
<td>75</td>
<td>94</td>
<td>91</td>
<td>83</td>
<td>76</td>
<td>87</td>
<td>90</td>
<td>78</td>
</tr>
<tr>
<td>5\text{MEMM}</td>
<td>61</td>
<td>62</td>
<td>63</td>
<td>76</td>
<td>57</td>
<td>59</td>
<td>94</td>
<td>73</td>
</tr>
<tr>
<td>6\text{CRFs}</td>
<td>80</td>
<td>90</td>
<td>92</td>
<td>83</td>
<td>74</td>
<td>84</td>
<td>90</td>
<td>79</td>
</tr>
<tr>
<td>6\text{MEMM}</td>
<td>66</td>
<td>72</td>
<td>76</td>
<td>77</td>
<td>59</td>
<td>69</td>
<td>88</td>
<td>72</td>
</tr>
<tr>
<td>7\text{CRFs}</td>
<td>83</td>
<td>93</td>
<td>96</td>
<td>83</td>
<td>76</td>
<td>87</td>
<td>90</td>
<td>79</td>
</tr>
<tr>
<td>7\text{MEMM}</td>
<td>69</td>
<td>72</td>
<td>80</td>
<td>80</td>
<td>68</td>
<td>72</td>
<td>91</td>
<td>76</td>
</tr>
<tr>
<td>8\text{CRFs}</td>
<td>88</td>
<td>99</td>
<td>98</td>
<td>84</td>
<td>68</td>
<td>90</td>
<td>88</td>
<td>80</td>
</tr>
<tr>
<td>8\text{MEMM}</td>
<td>72</td>
<td>78</td>
<td>80</td>
<td>80</td>
<td>68</td>
<td>72</td>
<td>91</td>
<td>76</td>
</tr>
</tbody>
</table>

Obtained from the "Static benchmarking of membrane helix predictions" server[2].
Chapter 5. Experiments, Evaluation and Analysis

Table 5.4: P-Values of CRFs vs. MEMM in each Experiment

<table>
<thead>
<tr>
<th>Experiments</th>
<th>$p$-Value</th>
<th>Experiments</th>
<th>$p$-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp 1</td>
<td>0.005300</td>
<td>Exp 5</td>
<td>0.004705</td>
</tr>
<tr>
<td>Exp 2</td>
<td>0.001893</td>
<td>Exp 6</td>
<td>0.000055</td>
</tr>
<tr>
<td>Exp 3</td>
<td>0.010066</td>
<td>Exp 7</td>
<td>0.001253</td>
</tr>
<tr>
<td>Exp 4</td>
<td>0.006360</td>
<td>Exp 8</td>
<td>0.011213</td>
</tr>
</tbody>
</table>

Obviously, all of these $p$-Values are very small, that is, they disagree rather strongly with the null hypothesis $H_0$, and favor $H_a$. These results suggest we accept the alternative hypothesis that the CRFs model is significantly better than MEMM.

5.5 Summary

In this chapter, we explained the data sets and the prediction matrices we used. Detailed explanations of every experiment were also presented. Our experiment results showed that CRF reached 88% at per-segment accuracy ($Q_{ok}$) and 84% at per-residue accuracy ($Q_2$), which made CRFs the most accurate prediction method compared to the twenty eight other methods hosted at the Benchmarking website. We also set up several comparisons between CRFs and MEMM, and took $t$-test on these two methods, which proved that CRFs outperform MEMMs significantly for transmembrane helix prediction using the set of features we defined. In the next chapter, we introduce the web server which makes our CRFs model available to anyone to use.
We have set up a web server to make our CRFs method available to the larger research community. This server contains the parameter (weights) of every feature after training of experiment 8 (details of the experiment are given in previous chapter), so CRFs model on the server have already been populated by all of our features extracted from training data and hence are able to give the best transmembrane helix prediction.

Users can access the server by visiting http://www-test.it.usyd.edu.au/wliu4500/. To use the CRFs prediction server, users have to follow three steps: submit test data file, conform the data and then get the prediction results.

### 6.1 Submit Test Data

To predict the transmembrane (alpha) helices by the server, users have to store their proteins primary sequences into a plain text file (.txt), and upload it onto this server.

The format of the uploaded data should be as follows: The "protein ID line" should begin with "#", users can add their command in this line. Below this line
Figure 6.1: Conform Test Data

should be the corresponding protein primary sequence. Every amino acid in the
sequence has to be represented by 1-Letter format (the 1-letter format of amino
acids are presented in chapter 2.1). The following is an example (two sequences):

#Sequence 0
MENLNMDLLYMAAAVMMGLAIGAIGIGILGGKLEGARQPDL
IPLLRTQFFIVMGLVDAIPMIAVGLGLYVMFAVA
#Sequence 1
VQLAHHFSEPIILIFGVMAGVIGTILLISYGIRLLIKK

On the server, we also provide the test data introduced in chapter 5.1.2 which was
used in our experiments. These data has already been converted into our format.

6.2 Conform Test Data

After submitting the test data file, the file name and file size will be shown to
users. For example, if we submit the file that we provided on the server, some
information will be shown as in Figure 6.1. If there is something wrong with the
information shown, users can go back and submit their file again.
6.3 Obtain Results

The prediction and labelling process may take about five to ten minutes. In prediction results, the server use ”1” to represent Helix Segments, and ”0” to represent Non-helix Segments. For instance, if we used the sample sequences in section 6.1, the prediction result will be given as shown in Figure 6.2.

6.4 Summary

In this chapter, we introduced the CRFs prediction server, which has all of the optimized parameters from our best experiment. Through three steps on the server,
users can get the predictions of transmembrane helices from their own protein primary sequences.
Chapter 7

Conclusion

In this thesis we have investigated the use of Conditional Random Fields (CRFs) as a sequential classifier to predict the locations of transmembrane helical regions in membrane proteins. We have defined 18 various features, which are used to capture the structural information displayed by amino acids. We compared our approach with 28 other methods available through the "Statistic benchmarking of membrane helix predictions" [2], and achieved the highest places at both per-segment and per-residue accuracy. We have shown experimentally that CRFs are significantly better than Maximum Entropy Markov Model (MEMM).

We believe that CRFs have a great potential power which can be employed in not only the transmembrane helix prediction problem but many other interesting problems in bioinformatics.
Bibliography


[18] David T. Jones. Improving the accuracy of transmembrane protein topology


