# Syntactic Approach to Predict Membrane Spanning Regions of Transmembrane Proteins

Koliya Pulasinghe[1] and Jagath C. Rajapakse[2]

[1] Sri Lanka Institute of Information Technology, Sri Lanka
[2] BioInformatics Research Centre, Nanyang Technological University, Singapore
koliya@sliit.lk, asjagath@ntu.edu.sg

**Abstract.** This paper exploits "biological grammar" of transmembrane proteins to predict their membrane spanning regions using hidden Markov models and elaborates a set of syntactic rules to model the distinct features of transmembrane proteins. This paves the way to identify the characteristics of membrane proteins analogous to the way that identifies language contents of speech utterances by using hidden Markov models. The proposed method correctly predicts 95.24% of the membrane spanning regions of the known transmembrane proteins and correctly predicts 79.87% of the membrane spanning regions of the unknown transmembrane proteins on a benchmark dataset.

## 1 Prediction of Membrane Spanning Regions of the Transmembrane Proteins

Transmembrane Proteins (TMPs), which traverse the phospholipid bi-layer of the membrane one to many times, as illustrated in Fig. 1 and Fig. 2, are integral membrane proteins, i.e., proteins which are attached to the cell membrane to keep their hydrophobic regions intact with aqueous cytosol. Thus, they make a channel between cytosome and extracellular environment, which transports various ions and proteins to and from cytosol. In addition, TMPs take part in vital cell functions such as cleavage of substances for metabolic functions, functioning as receptors, recognition and mediation in specific cell signaling, and participation in intercellular communication. Therefore, they are good therapeutic targets and the knowledge of the topography of the TMPs is of paramount importance to the design new drugs.

TMPs with experimentally verified structures are limited to about 1% of the total entries in most of the protein databanks though they amount to 20-30% of all open reading frames of the genomic sequences of several organisms [1][2]. Verifying TMP structures using experimental methods, such as X-ray crystallography and nuclear magnetic resonance spectroscopy, is not only expensive but also requires a lot of efforts due to the difficulties in protein expression, purification, and crystallization. Especially, TMPs have hydrophobic regions, which are buried inside the membrane, i.e., membrane spanning regions (MSRs), to

F. Rothlauf et al. (Eds.): EvoWorkshops 2005, LNCS 3449, pp. 95–104, 2005.

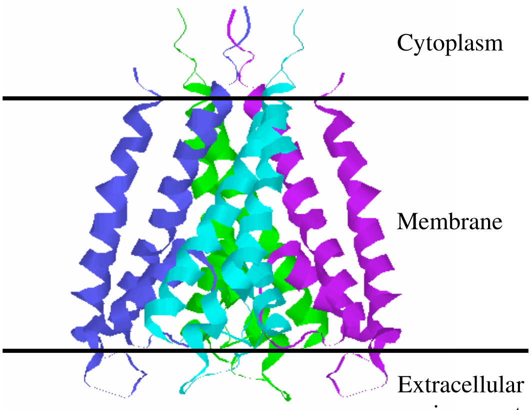

Image: Protein Explorer (PDB ID: 1BL8)

**Fig. 1.** Typical structure of a transmembrane protein: Potassium channel protein (KCSA) from *Streptomyces lividans*

keep the hydrophobic residues intact with aqueous cytosol and extracellular environment, and do not dissolve properly in aqueous solvents in the process of purification. Consequently, the prediction of MSRs of the TMPs became a classical problem in bioinformatics. Experimentally verified TMPs have two different motifs: membrane spanning $\alpha$-helix bundles and $\beta$-barrels. Usually the $\alpha$-helix bundles (Fig. 1) are predominant [3]. This paper focuses its attention on predicting the $\alpha$-helix bundles of the TMPs.

Early MSR prediction methods of the TMPs were based on the hydrophobicity analysis of the constituent amino acids [4][5][6]. Because, hydrophobicity values of the amino acids in MSRs are relatively high compared to the other regions. As illustrated in Fig. 2, high presence of Isoleucine, Valine, and Leucine,

```
MLYGF SGVIL QGAIV TLELA LSSVV LAVLI GLVGA GAKLS
ooooo ooooo ooMMM MMMMM MMMMM MMMMM MMMii iiiii

QNRVT GLIFE GYTTL IRGVP DLVLM LLIFY GLQIA LNVVT
iiiii iiiii iiiii iiiMM MMMMM MMMMM MMMMM MMMMo

DSLGI DQIDI DPMVA GIITL GFIYG AYFTE TFRGA FMAVP
ooooo ooMMM MMMMM MMMMM MMMMM MMMii iiiii iiiii

KGHIE AATAF GFTHG QTFRR IMFPA MMRYA LPGIG NNWQV
iiiii iiiii iiiii iiiii iiiii iiiii iiMMM MMMMM

ILKAT ALVSL LGLED VVKAT QLAGK STWEP FYFAV VCGLI
MMMMM MMMMM MMMoo ooooo ooooo ooooo ooooM MMMMM

YLVFT TVSNG VLLLL ERRYS VGVKR ADL
MMMMM MMMMM MMMMM iiiii iiiii iii
```

**Fig. 2.** Amino acid sequence and topography information of Histidine transport system permease protein (hisQ) of *Salmonella typhimurium*: `o, M, i` indicate outer (extracellular), membrane, and inner (cytoplasmic) residues respectively

which are relatively high hydrophobic amino acids according to Kyte-Doolittle hydrophobicity indices [4], can be observed in MSRs (denoted by M's). Accordingly, frequent occurrences of highly hydrophobic amino acids is a good guess for detecting MSRs. This technique is employed in hydropathy plots, an early technique that is still popular in recognizing the MSRs in TMPs [4][7]. Among more recent methods, which predict the topography of transmembrane proteins, hidden Markov model (HMM) based methods claims the highest accuracy [8]. Among them, TMHMM [2][9], HMMTOP [10], and MEMSAT [11] can predict the membrane bounded region of the transmembrane proteins upto 65% to 80% accuracy [8]. The above methods are different due to the structure of the HMMs, i.e., the domains and segments that the HMM represents, and the training method used.

Among the non-HMM methods, PHDhtm predicts MSRs of TMPs by using an artificial neural network (ANN) [12]. A special feature of the PHDhtm is that the ANN learns the patterns of the evolutionary information (homology). In Toppred, the approach combines hydrophobicity analysis and positive inside rule to predict the putative transmembrane helices [6]. A general dynamic programming-like algorithm, MaxSubSeq (stands for Maximal Sub-Sequence), optimizes the MSRs predicted by other methods [13]. An evaluation of methods for the prediction of MSRs can be found in [8]. Protein sequences of the TMPs verified by the imperial methods can be found in several databases such as MPtopo database [14], TMPDB [15], and TMHMM site [2]. In TMHMM, a state was designed to absorb the properties of one residue except in self-looping globular state. All other states are designed without self-transition probabilities. Contrary to that, in HMMTOP, each characteristic region is represented by a self-looping single state. Approach taken in the proposed method used moderate number of states to represent various characteristic regions. This approach is motivated by the fact that each turn of the helix in MSR consist of 3-4 residues. Accordingly, a state is designed to represent one turn of a $\alpha$-helix rather than a one residue, as in TMHMM, or one characteristic segment, as in HMMTOP, of the TMPs. Length of an MSR is ranging from 15 residues to 30 residues.

The our approach to MSR predication of TMP is also based on HMMs. Unlike previous approaches, in our HMM model (see Fig. 4), self transitions and transition between every other states can align different length MSRs in the training process as well as in the recognition process. The proposed method correctly predicts 95.24% of the MSRs of the known TMPs and correctly predicts 79.87% of the MSRs of the unknown TMPs on a bench mark dataset.

The organization of this paper is as follows. In Section 2, the syntactic rules derived by observing the various segments of the TMPs are described as a syntactic network where each HMM model is aimed at recognizing an allowable segment combinations of the TMPs. In Section 3, we described the HMMs and training algorithm based on Viterbi segmentation. Section 4 describes the data used in training and testing the proposed method along with results obtained. Finally, a brief discussion about the proposed method and the future directions are given in Section 5.

## 2 Syntactic Rules of the "Biological Grammar" of TMPs

The presence of alternate sequences of *i*nner (i.e., inside or cytoplasmic), *m*embrane spanning, and *o*uter (outside or extra-cytoplasmic) regions of the transmembrane proteins follows a simple rule of grammar [16]. These regions have unique features inherent to them and do not occur randomly. The syntactical rules of these occurrences of different regions are derived and given below in Fig. 3. The establishment of these rules has great importance to our study, which follows a similar approach that used to identify the language contents of the unknown speech utterances. TMPs with unknown regional boundaries are analogous to unknown speech utterances.

The following symbols lay down the syntactic rules in the biological grammar:

| | denotes alternatives
[ ] encloses options
{ } denotes zero or more repetitions
⟨ ⟩ denotes one or more repetitions
$var denotes a variable word.

The two different orientations, outer-membrane-inner (i.e., *omi*) and inner-membrane-outer (i.e., *imo*), with respect to the cell membrane can be observed in the helix core of the MSRs and are defined them as separate literals. Inner and outer residue sequences can be observed in different lengths. They are categorized into three groups, each according to their length. As an example, in inner loops, the literal "*i*" represents a protein sequence with 1-6 residues. The literal "*ii*" represents protein sequences with 7-20 residues, while the literal "*iii*" represent the very long protein sequences with more than 20 residues. Same procedure is applied in defining literals "*o*", "*oo*", and "*ooo*". Accordingly syntactical rules governing on possible TMP configuration can be symbollically described as follows:

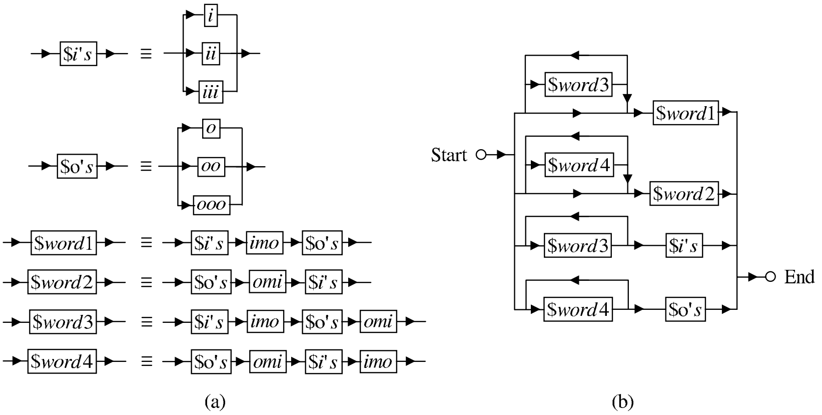

**Fig. 3.** A graphical illustration of characteristic regions of transmembrane proteins

```
$i's = i | ii | iii ;
$o's = o | oo | ooo ;
$word1 = $i's imo $o's ;
$word2 = $o's omi $i's ;
$word3 = $i's imo $o's omi ;
$word4 = $o's omi $i's imo ;
```

A pictorial view of these syntactic rules is shown in Fig. 3(a). As illustrated in the Fig. 3(b), any continuous path from start node to end node, along the direction of arrows, generate a possible segment sequence of any TMP. According to our symbolic notations, syntax of any TMP can be given by:

```
( ⟨ $word3 ⟩ $i's | ⟨$word4 ⟩ $o's | {$word3} $word1 | {$word4} $word2 )
```

These grammatical rules can generate numerous syntactically correct TMPS, as illustrated below. To derive the following sequences, the word network shown in Fig. 3(b) can be used. According to our interpretation, these are possible topological structures of the TMPs, where each literal represents a characteristic feature of a TMP segment:

- *i imo o omi i imo oo omi ii*
- *ooo omi iii*
- *ii imo ooo omi i imo oo omi iii imo oo omi i*
- *oo omi i imo o omi ii imo ooo*
- *i imo o omi i imo o*

A set of HMMs to represent these literals are described in the Section 3.

# 3 Methodology

Several HMMs are defined, in which each HMM represent a literal, e.g., *imo*, described in the previous section. A special kind of HMM called left-to-right HMM is defined as shown in the Fig. 4 with the intention that all HMMs can be tied parallelly by using first state and last state, to make a single large HMM. The motive behind is that the combination of a giant HMM and syntactical networks described above can be used to recognize unknown segments of a TMP by training and using testing algorithm as described in [18].

## 3.1 Definition of HMMs

In our design, each literal is designed by a separate HMM; all HMMs share the same configuration as illustrated in Fig. 4. In this type of HMMs, no transitions are allowed to the states whose indices are lower than the current state. In what follows, we give a definition for the HMM to be used, by using the same notation used in Rabinar's seminal paper [19].

1) $N$: the number of states in the model. We denote the set of individual states as $S = \{S_1, S_2, \ldots, S_N\}$, and the state at site $t$ (or $t$th observation or $t$th residue) as $q_t$.

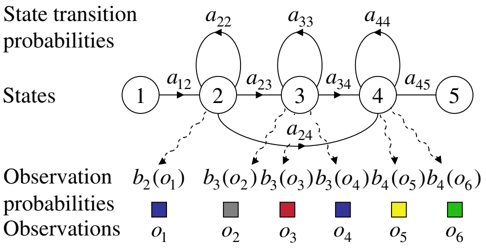

**Fig. 4.** A left-to-right hidden Markov model

2) $M$: the number of distinct observation symbols per state, i.e., the number of amino acids (#20) in our study. Though we are not interested about the physical output of the system, to model the system, this must be taken into consideration. We denote the set of individual residues as $R = \{r_1, r_2, \ldots, r_M\}$.

3) State transition probability distribution $A = \{a_{ij}\}_{N \times N}$ where,

$$a_{ij} = P\big[q_{t+1} = S_j \mid q_t = S_i\big], \qquad 1 \leq i, j \leq N. \tag{1}$$

In left-to-right HMM model, the state transition coefficients have the property

$$a_{ij} = 0, \qquad j < i \tag{2}$$

That is, no transitions are allowed to states whose indices are lower than the current state. It should be also noted that, for the last state in a left-to-right model, the state transition coefficients are specified as

$$a_{NN} = 1 \tag{3}$$

$$a_{Ni} = 0, \quad i < N. \tag{4}$$

4) The observation symbol probability distribution in state $j$, $B = \{b_j(k)\}$, where

$$b_j(k) = P\big[r_k \mid q_t = S_j\big], \qquad \begin{array}{c} 1 < j < N \\ 1 \leq k \leq M \end{array} \tag{5}$$

5) The initial state distribution $\pi = \{\pi_i\}$ where

$$\pi_i = \begin{cases} 0, & i \neq 1 \\ 1, & i = 1 \end{cases} \tag{6}$$

Once parameters are estimated using a proper algorithm, this HMM can generate observation sequence $O = (O_1 O_2 \ldots O_T)$, where each observation $O_t$ is the residue at site $t$, and $T$ is the number of observations in the sequence.

### 3.2 HMM Parameter Estimation

Parameter estimation of the HMMs is done by Viterbi alignments [20].To initialize the model parameters Viterbi training is replaced by a uniform segmentation, i.e., each training observation is divided into $N$ equal segments. In Viterbi training, each training sequence is segmented using a state alignment procedure which results from maximizing

$$\phi_N(T) = \max_i \{\phi_i(T) a_{iN}\} \tag{7}$$

for $1 < i < N$ where

$$\phi_j(t) = \max_i \{\phi_i(t-1) a_{ij}\} b_j(r_t) \tag{8}$$

with initial conditions given by

$$\phi_1(1) = 1 \tag{9}$$

$$\phi_j(1) = a_{1j} b_j(r_1). \tag{10}$$

for $1 < j < N$.

If $A_{ij}$ represents the total number of transitions from state $i$ to state $j$ and $b_i(k)$ represents the observation probabilities of emitting symbol $k$ in state $i$, by performing the above maximization, the transition probabilities can be estimated from the relative frequencies:

$$\hat{a}_{ij} = \frac{A_{ij}}{\sum_{k=2}^{N} A_{ik}} \tag{11}$$

$$\hat{b}_i(k) = \frac{\sum^{N}_{\substack{k=2 \\ s.t. O_t = r_k}} A_{ik}}{\sum_{k=2}^{N} A_{ik}} \tag{12}$$

As a by-product of above calculation the maximum likelihood $\hat{P}(O|M)$ is given by Eq. (7). The above process can be iteratively carried out until the change of the maximum likelihood between two consecutive iteration reached to an acceptable level.

## 4 Experiments

In this section, we demonstrate the accuracy and efficacy of the proposed approach, using the dataset that used in training TMHMM [2]. And for the testing, 73 TMPs unknown to the system is extracted from dataset C, which contribute maximum number of unknown proteins to the comparison of different methods including TMHMM 2.0, TMHMM 1.0, HMMTOP, and MEMSAT 1.5 [21]. The labeled data was used to estimate the parameters of each HMM separately. The number of states in each literal, which denotes an HMM, is given in the Table 1.

After training separately, all HMMs are tied parallelly by using first state and the last state to make a single large HMM. The combination of this giant HMM and a syntactical network described in Section 2 above is used to recognize unknown segments of TMPs by using a token passing algorithm described in [18].

**Table 1.** The number of states in each literal in the training dataset

| Literal | Number of States |
|---|---|
| *imo* | 7 |
| *omi* | 7 |
| *i* and *o* | 2 |
| *ii* and *oo* | 6 |
| *iii* and *ooo* | 9 |

**Table 2.** The performance of the present method in the prediction of topology of both training and testing dataset

| | Training Set | | Testing Set | |
|---|---|---|---|---|
| | Number | Percentage | Number | Percentage |
| Number of Proteins | 159 | | 73 | |
| MSRs found | 132 | 83% | 46 | 63% |
| Additionally correct sidedness | 110 | 85% | 34 | 74% |
| | | | | |
| Total number of helices | 694 | | 328 | |
| Predicted helices | 661 | 95.24% | 262 | 79.87% |
| Over-predicted helices | 20 | 2.88% | 21 | 6.40% |
| Under-predicted helices | 20 | 2.88% | 53 | 16.15% |
| Shifted helix prediction | 13 | 1.87% | 14 | 4.27% |
| Falsely merged helices | 24 | 3.46% | 21 | 6.40% |

Tools provided with HTK toolkit, a toolkit primarily designed for modeling and manipulating HMMs in speech processing, was used in training process as well as in testing process [20].

Results of the prediction can be found in the Table 2, which shows the performance of the proposed method for the training dataset as well as for the test dataset. Performance of the method is evaluated on two different bases, firstly as a complete topography predictor and secondly as an MSR predictor. The present method predicts all the MSRs of 46 TMPs out of 73 unknown TMPs. In addition, it predicts correct positioning of start region in 34 TMPs out of 46 TMPs. As an MSR predictor, it predicts the 95.24% of MSRs (true positive predictions) from the total number of 694 MSRs in training data set, 79.87% of MSRs from the total number of 328 MSRs in test data set. It reported about 3% of over-predicted helices (false positives) and under-predicted helices (false negatives) in training data set, while those values were 6.4% and 16.15% in the test data set respectively. Shift helix prediction represents the regions, which share less than 9 residues with the reference annotation's MSRs. Falsely merged helices shows the regions, where adjacent helices are predicted as a single helix. Here, an MSR to be evaluated as predicted, it must share at least nine residues with the reference annotation's MSR. The other methods compared in Table 3 was evaluated on this basis in [8]. A test data set consists of 73 TMPs retrieved from the same data set that is used to evaluate the other methods. Table 3

**Table 3.** Comparison of performance of the present method compared to the previous approaches

| Method | No. of proteins | All MSRs found | Additionally correct sidedness |
|---|---|---|---|
| Prposed Method | 73 | 46 (63%) | 34 (74% of 44) |
| TMHMM 2.0 | 108 | 64 (59%) | 40 (63% of 64) |
| TMHMM 1.0 | 108 | 57 (53%) | 21 (37% of 57) |
| HMMTOP | 106 | 54 (51%) | 42 (78% of 54) |
| MEMSAT 1.5 | 159 | 80 (50%) | 58 (73% of 80) |

compares the proposed method with previous approaches to TMP topology prediction. The performance figures of the TMHMM 2.0, TMHMM 1.0, HMMTOP, and MEMSAT 1.5 were obtained from [8]. The present method showed the best performance on the tested dataset.

## 5 Discussion and Future Directions

We have trained and have tested a new algorithm to predict the membrane spanning regions ($\alpha$-helices) of the transmembrane proteins by looking at the protein in a syntactic point of view. The proposed model is a dynamic one which adjusts to the protein structure according to the characteristics of its segments. The hidden Markov models of the proposed method contain states which represent properties of small segments rather than a single residue and automatically adjust to the segment lengths.

On the tested dataset, the present method showed better performance over the reported accuracy measures of previous methods in both identification of MSR and description of their sidedness. The methods predicting protein topology with high accuracy has high pharmaceutical applications as membrane proteins are good therapeutic targets.

The syntactic rule set is flexible to absorb new characteristics such as sequences belong to the signal peptides which hamper the prediction accuracy, when they are inserted in the transmembrane proteins. The performance of the present method can be improved either by removing the signal peptides before the prediction process or by introducing new HMM model trained with signal peptide data.

## References

1. Wallin, E., von Heijne, G.: Genome-wide analysis of integral membrane proteins from eubacterial archaean, and eukaryotic organisms. Protein Sci. **7** (1998) 1029–1038
2. Krogh, A., et al.: Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J. Mol. Biol. **305** (2001) 567–580

3. White, H. W., Wimley, C. W.: Membrane protein folding and stability: Physical principles. Annu. Rev. Biophys. Biomol. Struct. **28** (1999) 319–365
4. Kyte, J., Doolittle, R. F.: A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157** (1982) 105–132
5. von Heijne, G.: The distribution of positively charged residues in bacterial inner membrane proteins correlates with the transmembrane topology. EMBO J. **5** (1986) 3021–3027
6. von Heijne, G.: Membrane protein structure prediction: Hydrophobicity analysis and the positive inside rule. J. of Mol. Bio. **225**(2) (1992) 487–494
7. Claros, M. G., von Heijne, G.: TopPred II: An improved software for membrane protein structure predictions. Computer Applications in the Biosciences. **10**(6) (1994) 685–686
8. Moller, S., et al.: Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics **17**(7) (2001) 646–653
9. Sonnhammer, E. L. L., et al.: A Hidden Markov Model for Predicting Transmembrane Helices in Protein Sequences. Proc. on Intelligent Systems in Mol. Biology. **6** (1998) 175–182
10. Tusnady, G. E., Simon, I.: Principles governing amino acid composition of integral membrane proteins: Applications to topology prediction. J. of Mol. Bio. **283** (1998) 489–506
11. Jones, D. T., et al.: A model recognition approach to the prediction of all helical membrane protein structure and topology. Biochemistry **33** (1994) 3038–3049
12. Rost, B., et al.: Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Science. **5** (1996) 1704–1718
13. Fariselli, P., et al.: MaxSubSeq: an algorithm for segment-length optimization. The case study of the transmembrane spanning segments. Bioinformatics. **19**(4) (2003) 500–505
14. Jayasinghe S. et al.: MPtopo: A database of membrane protein topology. Protein Science. **10** (2001) 455–458
15. Ikeda, M., et al.: TMPDB: a database of experimentally-characterized transmembrane topologies. Nucleic Acids Research, **31**(1) (2003) 406–409
16. Melen, K., et al.: Reliability measures for membrane protein topology prediction algorithms. J. Mol. Biol. **327** (2003) 735–744
17. Grundy, W. N.: Modelling Biological Sequences Using HTK. Technical Report, prepared for Entropic Research Laboratory, Inc. (1997)
18. Young, S. J., et al.: Token Passing: a conceptual model for connected speech recognition systems. CUED Technical Report F_INFENG/TR38, Cambridge University (1989)
19. Rabinar, L. R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2) (1989) 257–286
20. Young, S., et al.: The HTK book (for HTK version 3.2.1). Cambridge university Engineering Department. (2002)
21. Moller, S., et al.: A collection of well characterised integral membrane proteins. Bioinformatics **16**(12) (2000) 646–653