BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Prediction Model

Volume 7(8)

Hidden markov model for the prediction of transmembrane proteins using MATLAB

Navaneet Chaturvedi[1*], Sudhanshu Shanker[2], Vinay Kumar Singh[3], Dhiraj Sinha[4] & Paras Nath Pandey[5]

[1,2,4]Center of Bioinformatics, University of Allahabad, Allahabad, India; [3]Bioinformatics Center, School of Biotechnology, Banaras Hindu University, Varanasi, India; [5]Department of Mathematics, University of Allahabad, Allahabad, India; Navaneet Chaturvedi - Email: bioinfonavneet@gmail.com; * Corresponding author

Received November 24, 2011; Accepted December 07, 2011; Published December 21, 2011

Abstract:
Since membranous proteins play a key role in drug targeting therefore transmembrane proteins prediction is active and challenging area of biological sciences. Location based prediction of transmembrane proteins are significant for functional annotation of protein sequences. Hidden markov model based method was widely applied for transmembrane topology prediction. Here we have presented a revised and a better understanding model than an existing one for transmembrane protein prediction. Scripting on MATLAB was built and compiled for parameter estimation of model and applied this model on amino acid sequence to know the transmembrane and its adjacent locations. Estimated model of transmembrane topology was based on TMHMM model architecture. Only 7 super states are defined in the given dataset, which were converted to 96 states on the basis of their length in sequence. Accuracy of the prediction of model was observed about 74 %, is a good enough in the area of transmembrane topology prediction. Therefore we have concluded the hidden markov model plays crucial role in transmembrane helices prediction on MATLAB platform and it could also be useful for drug discovery strategy.

Availability: Matlab script is available upon request to bioinfonavneet@gmail.com, vinaysingh@bhu.ac.in

Keywords: Hidden Markov Model, Transmembrane Proteins, MATLAB

Background:
Accurate predictive success of transmembrane proteins by applying hidden markov model [HMM] is frequently used in biological research. This is fully machine learning approach in which genome structure and proteins topology prediction are the fascinating and most demanding subject in bioinformatics. The body of a HMM is closely compatible to the biological entities which is being simulated by the model. Looking the feasibility of HMM by going through the recent research articles, is totally statistical approach that compiled by the set of states which have potentially able to emit symbols on the basis of probability **[1, 6]**. These states are estimated by model parameters. Model parameters consist of three probabilities *i.e.* initial probability, transition probability and emission probability.

In the context of accurate prediction method of transmembrane topology many workers had obtained results with more than five helices were predicted at a significantly lower accuracy than proteins with five or fewer and in addition the estimation of the standard procedure to resolve the prior work and presented novel trends that may impact the analysis of entire proteomes **[2]**. Furthermore, some workers were addressed a method to reduce the number of false positives, i.e., proteins falsely predicted with membrane helices **[3].** Previously the workers had worked on the effectiveness of model

regularization, dynamic model modification and optimization strategies of model and validated through experimentally **[4]**. Several workers were developed five different prediction method, these are TMMOD **[14]**, PHD, HMMTOP, MEMSAT, and TOPPRED for comparatively better result in transmembrane topology prediction **[5]**. Although the prediction of transmembrane helices can be analyzed with good enough score plot but this idea was not successful for other helices. Recently the hydrophobic property of transmembrane helices was used for the prediction **[6]**. But after some time one feature was taken into account i.e. abundance of positive charge residues **[7, 13, 15]**. Due to better incorporation of helix length, compositional bias and grammar constraints, HMM is suitable for the prediction of transmembrane helices. Helical membrane proteins are specified a "grammar", in which cytoplasmic and non-cytoplasmic loops have occurred alternate fashion. Therefore this feature provides the efficient information, even performs better result in prediction **[8, 14].** TMHMM approach is applied here because these days mostly membrane protein is predicted through this method and result is more accurate. For better understanding of model we applied here MATLAB. MATLAB is widely used to visualize and analyze the biological data. MATLAB provides a bioinfo tool box in which various bioinformatics tools are available for technical computing and scripting. This language is a high-level matrix/array language with control flow statements, functions, data structures, input/output, and object-oriented programming features **[9]**. It was experienced that MATLAB shows better feasibility with HMM rather than other bio-statistical packages.

In this work we present our model performance, based on hidden markov model, by taking approach from previous research on TMHMM **[10]**. We have reviewed the previous TMHMM model architecture which has specialized modeling of different regions of membrane proteins like inner, middle and outer region **[8]**. These regions are collectively divided by 7 locations. The states are connected to each other and transitions may be possible between adjacent states. Best possible transition of a state recruits on the basis of high transition probability value, known as transition probability and further 20 amino acids are emitted by probability distribution of each state.

Methodology:

Dataset

We have downloaded two types of dataset from the TMHMM website in which second dataset has 160 transmembrane sequences. Most protein pattern of these dataset was determined by experimentally. The dataset is labeled by three main locations, these are transmembrane helix (m), inside (i) and outside loops (o) on the basis of the existence of deferent amino acids pattern within a transmembrane region. 10-fold cross validation was applied for the validation of model **[8, 10].**

Method

In this work, we have introduced the probabilistic framework of the HMM for transmembrane helix prediction. Well known architecture of TMHMM was considered and three main locations in dataset were further divided into seven different super states as shown in TMHMM model architecture **[12] (Figure 1).** Conversion from three to seven super states was decided on the basis of position and length of strings location. In the context of outer region of transmembrane topology, five residues length of cap cytoplasmic was used to fix the boundary between helix and loop region. These seven super states were also labeled. For outer loop, either it was short (S) or long (O) region of non-cytoplasmic, was considered the length of 20 residues and similarly the loop (L) of cytoplasmic side was counted as 20 residues. The cap non-cytoplasmic (N) was considered the length of 5 residues **(Figure 2)**. The length of helix core (M) region was, determined previously, and considered 25 residues. **(Figure 3)** The seventh super state, rest part of the transmembrane architecture was treated as 1 or may be more than 1. Same text of the state was considered same region. Each super state has an associated probability distribution over the 20 amino acids characterizing the variability and pattern of amino acids in the region. Here the every single position of a super state was considered as individual state of the model.

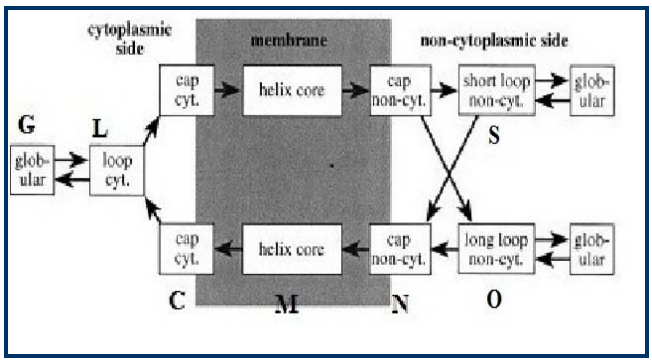

Figure 1: Architecture of TMHMM consists of 7 boxes and each box indicates the particular region. Same text box treats as same region. Here we have denoted each box by symbols to its corresponding region. Each box may be one or more states in HMM with same parameters.

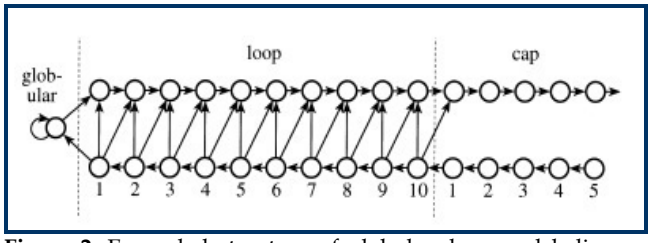

Figure 2: Expended structure of globular, loop and helix cap cytoplasmic region. Possible transitions between the states are shown. Therefore we have considered maximum 10 states for loop on the basis of length of amino acid and same maximum 5 states as for helix cap region.

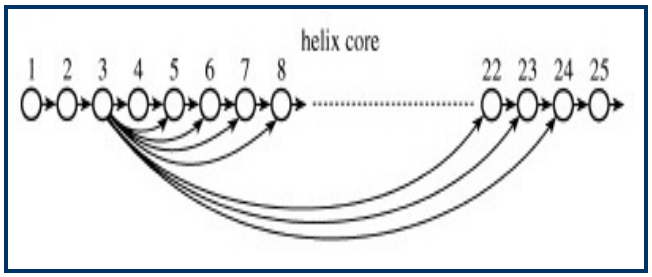

Figure 3: Here representation the detail structure of helix core and all possible transitions between 25 states.

In this work we have considered the number of states of a region on the basis of number of amino acid positions **[8]**. For similar reasons, a loop's first and last 10 residues are explicitly modeled, and the each residue corresponds to an individual state in the model. All other residues in the middle of a loop are collectively represented by one "globular" state, which has a transition back to itself thus can repeat as many times as the loop length dictates. But in the case of loop apart of other here we have considered all loop regions for 20 residues including long and short loop. Short and long loops were separated here on the basis of the prediction of adjacent regions by possible state transitions. Since long loops outside the membrane appear to have divergent properties than short loops, two separate chains of states are introduced, as expressed in figure1. The transmembrane helix model was constructed exactly like TMHMM **[8, 12]**, by two cap regions of 5 residues each, including a core region of length 25 residues. Therefore the total length for membrane helices was 35 residues, if we were considered the length including cap regions, covering the real size range observed for transmembrane domains. The state diagram for the overall transmembrane topology collectively was considered minimum 96 states by converting these 7 super states on the basis of length corresponding to its transitions. The entire states model is thus 20+5+25+5+20+20+1 = 96. Therefore the IP (Initial probability), STP (State transition probability) and REP (Residue emission probability) matrices were 96x1, 96x96 and 96x20 respectively. Each state should be emitted single amino acids.

HMM Training and Scoring

In the first stage the aim is to fix the boundaries of transmembrane helices. Previously many worker were suggested for boundary correction **[8]** in labeled dataset, like M, I and O. 'M' represents membrane region, 'I' is for inside region (cytoplasmic) and 'O' means outer region (non-cytoplasmic). Therefore we were compiled the MATLAB script for converting these 3 labeled dataset sequences to 7 labeled sequences on the basis of TMHMM model architecture **(Figure 1)**. New model estimation was done using the relabeled sequences. In Second step for estimating the HMM model parameters, posterior probability method were applied. Each state of the model has been predicted by forward-backward algorithm, called as posterior probability method. Previously the scoring method described in the article was implemented using N-best algorithm **[16],** was also very effective but here it was found that the posterior probability method is more convenient and compatible for the estimation. This HMM model parameter is estimated in 7 super states and the training was performed by estimating the maximum likelihood **[11]**. As usual we have also applied the traditional method (posterior probability) for estimating the HMM and its transmembrane model parameter θ. Here the joint probability of a set of sequence $P(x^1,\ldots\ldots,x^n)$ being emitted the symbol using a particular state of path **[1, 8]** $P(x^1,x^2,\ldots\ldots, x^n/\theta)$.

The parameter θ contains all the STP (State Transition Probability), REP (Residue Emission Probability) and IP (Initial probability) matrices. Probability of a residue for a particular state has been evaluated by posterior probability method **[8, 1].** Baum-Welch algorithm is standard method for maximum likelihood estimation of HMMs, in which posterior probabilities were performed by utilizing both forward and backward algorithms. These algorithms were used to compile the STP and REP matrices. The program was set to 15 iteration time to converse the values. Initial probability was arbitrary taken for the initializing the parameter θ calculation. Here we were used dataset sequences for the model estimation and training. The procedure used for training a model was the same regardless of the scoring method used and in most aspects identical to the procedure used by Krogh et al. (2001) **[8].** Here, both training and scoring was done using the original data set of 160, which we have divided it into subset of 16x10.

Validation

Since 160 dataset have been reported. For ten cross fold validation we were divided 160 sequences set into 10 subset. Created 10x16 data and put newly estimated HMM parameter on this 10 subset. In the result, MATLAB calculated the scores for each sequences of every subset**.** We have found the score for each sequences of all subset. Test set were evaluated by original test structure taken from dataset **[8]**.

Discussion:

Labeled 160 sequences were recruited, measuring the accuracy of model whose topology and locations of transmembrane helices are correctly predicted by TMHMM. The score of each sequence was calculated by posterior probability method of HMM for each sequence of 160 dataset. The 16x10 score plot is showing in **Figure 4.** Legend of figure shows the 160 stars of different sequences. Each subset of 16 sets of amino acid sequences possesses 10 sequences. Plot was represented the scores of 16x10 data. Position of a star indicates the score of a sequence and the aggregate prediction was calculated by following equation on MATLAB. We were obtained approximately 74% accuracy on average. The higher score value is observed at 98.6. In addition, the 63 times have observed the score values above 80.

Aggregate prediction = mean (sum (score, 2)/16)

This HMM-based method embodies many conceptual and methodological aspects of previous methods. The main realities are that the model architecture closely to the transmembrane pattern and that everything is done in the probabilistic framework of HMMs, so we do not have to develop a specialised dynamic programming algorithm or posterior probability method. Our model prediction is helpful for better understanding of existing model.

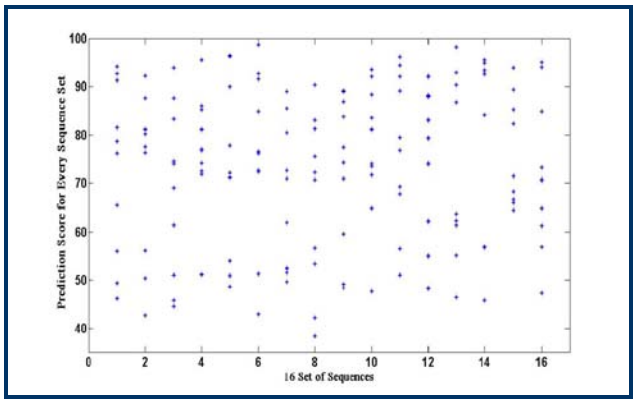

Figure 4: Prediction Score plot for each 10 sequence of 16 set.

Dataset sequences were helpful for the validation of model and we have treated them as a unified set in order to find general principles.

Conclusion:

All transmembrane sequences for validation of model were accessed from TMHMM site indicating that they are all recruited to the transmembrane helices. Specific transmembrane topology prediction is helpful for protein functional annotation. Along with other documented programming languages, MATLAB was also taken into consideration because the MATLAB performed better compatibility with biological entities. Therefore the hidden markov model could perform crucial role in transmembrane helices prediction on MATLAB platform and it could also be useful for drug discovery strategy. Since there are many various statistical approaches are using by several workers in biological sciences but HMM is well known for higher accuracy result in the area of protein topology prediction. Meanwhile there should be more work yet to be explored for higher result accuracy and low level of redundancy for transmembrane proteins prediction.

References:

[1] Juang BH & Rabiner LR, *Technometri.* 1991 **33**: 251
[2] Chen PC *et al. Protein Sci.* **11**: 2774 [PMID: 12441377].
[3] Rost B *et al. Protein Sci.* **5**: 1704 [PMID: 8844859].
[4] Hughey R & Krogh A, *Comput App Biosci.* 1996 **12**: 95. [PMID: 8744772].
[5] Kall L & Sonnhammer E L, *FEBS Lett.* 2002 **18**: 415. [PMID: 12482603].
[6] Argos P *et al. Eur J Biochem.* 1982 **15**: 565 [PMID: 7151796].
[7] Heijne G & Manoil C, *Protein Eng.* 1990 **4:** 109 [PMID: 2075184].
[8] Krogh A *et al. J Mol Biol.* 2001 **19**: 567 [PMID: 11152613].
[9] MATLAB technical documentation. Www.mathworks.com. Retrieved 2010-06-07.
[10] Sonnhammer E L *et al. Proc Int Conf Intell Syst Mol Biol.* 1998 **6**:175 [PMID: 9783223].
[11] Krogh A, *Proc Int Conf Intell Syst Mol Biol.* 1997 **5**:179 [PMID: 9322033].
[12] Durbin R *et al. Cambridge University Press.* 1998 Cambridge, UK.
[13] Heijne G, *EMBO J.* 1986 5: 3021 [PMID: 16453726].
[14] Robel Kahsay RY *et al. Bioinformatics.* 2005 **21**:1853 [PMID: 15691854].
[15] Von Heijne G. *J Mol Biol.* **225**: 487 [PMID: 1593632].
[16] Kall L *et al. J Mol Biol.* 2004 **338**: 1027 [PMID: 15111065].

Edited by P Kangueane
Citation: Chaturvedi *et al.* Bioinformation 7(8): 418- 421 (2011)