Transmembrane Region Prediction with Hydropathy Index/Charge Two-Dimensional Trajectories of Stochastic Dynamical Systems

T. Kaburagi, D. Muramatsu, S. Hashimoto, M. Sasaki and T. Matsumoto

Department of Electrical Engineering and Bioscience
Waseda University
Tokyo 169-8555, Japan

{kaburagi02, daigo, hashimoto02, sasaki }@matsumoto.elec.waseda.ac.jp,
takashi@mse.waseda.ac.jp

Abstract

A new algorithm is proposed for predicting transmembrane regions from two dimensional vector trajectories consisting of hydropathy index and charge of a test amino acid sequence by stochastic dynamical system models. The prediction accuracy of a preliminary experiment is 96.09%. Since no fine-tuning is done, this appears encouraging.

Keywords: Bioinformatics; HMM; Transmembrane Proteins; Transmembrane Region Prediction; Hydropathy Index

1 Introduction

The importance of transmembrane protein structure prediction problems is well documented in [1], [2], [6] and [9], among others. Roughly speaking, there are two ways of looking at protein structure predictions. One is solely based on the construction principles of proteins associated with physico-chemical properties of amino acids. No concept of training is involved. The other is to collect data sets with known structures, extract features and use machine learning algorithms for predictions. In many of the prediction problems for protein structure prediction in general, and transmembrane protein structure prediction in particular, prediction accuracies still call for improvements.

This paper considers a restricted class of transmembrane protein structure prediction problems instead of general classes of problems. Specifically, the paper assumes that a given amino acid sequence is a transmembrane protein, even though we can attempt to predict if the sequence is water soluble or transmembrane protein. A primary reason for this is that there are several very good tools for such prediction problems. (see e.g., [1]). This paper attempts to perform the following:

Goal: Predict transmembrane regions of a given amino acid sequence.

To the best of our knowledge, this class of problems is non trivial. One reason is that the number of transmembrane proteins whose structure is known is severely limited because of the difficulties associated with using X-ray crystallography for transmembrane proteins. However, for this very reason, transmembrane protein structure prediction is a great challenge for a machine learning approach.

This paper proposes a new algorithm for predicting transmembrane regions along with the line formulated for the transmembrane counts predictions reported in [8]. The algorithm incorporates two-dimensional trajectories consisting of the hydropathy index and charge of amino acids associated with stochastic dynamical systems. This paper also reports a preliminary experimental result. Prediction accuracy is 96.09%. Since no fine-tuning is done, the result appears encouraging.

We conclude this section by remarking that the references cited are far from exhaustive.

2 Algorithm

Given a protein primary structure, this paper considers the two dimensional vector trajectory

$$\{O_t := (O_t^1 = hydropathy \quad index\ , O_t^2 = charge\ )\}_{t=1}^{T} \qquad (1)$$

associated with amino acids instead of the 20-letter symbol sequence. One way of taking into account the sequential nature of the problem is to consider an auxiliary sequence $\{Q_t\}$, which is a trajectory of an *inner stochastic dynamical system* indexed by a one dimensional parameter t and look at O_t as an *output* with uncertainty where the *joint* probability distribution is described by:

$$\begin{aligned}
&P(\{O_t^1 = v_{k_1}^1, O_t^2 = v_{k_2}^2\}_{t=1}^T, \{Q_t = q_i\}_{t=1}^T \mid w, \mathcal{H}) \\
&= \prod_{t=1}^{T} P(O_t^1, O_t^2 \mid Q_t, w, \mathcal{H}) P(Q_{t+1} \mid Q_t, w, \mathcal{H}) P(Q_1 \mid w, \mathcal{H}) \\
&= \prod_{t=1}^{T} b^1_{Q_t=q_i, O_t^1=v_{k_1}^1} b^2_{Q_t=q_i, O_t^2=v_{k_2}^2} a_{Q_{t+1}=q_l, Q_t=q_i} \pi_{Q_1=q_j} \\
&w := \{\{a_{ij}\}, \{b_{ik_1}^1, b_{ik_2}^2\}, \{\pi_i\}\}
\end{aligned} \qquad (2)$$

where $\mathcal{H}$ stands for the underlying model structure. The first and the second equations in (2) are in general form, whereas the last equation is a Hidden Markov Model, which will be used in this paper. Observe that even though the hydropathy index is real valued, there are only a finite number of index values, e.g., 17 for the Kyte-Doolittle index [3]. Similarly, there is only a finite number of charge values +1, 0, and –1;

$$O_t^1 = v_{k_1}^1,\ k_1 = 1,....,K_{1,} \quad O_t^2 = v_{k_2}^2,\ k_2 = 1,....,K_2$$

Figure 1 illustrates the typical trajectories of a K-D index and charge of transmembrane proteins. A major consequence in considering these physico-chemical indices instead of 20 symbols, is the fact that *nearness* can be taken into account between different amino acids. Namely, two amino acids with a similar hydropathy index can be considered close to each other with respect to this particular metric. This enables to perform "smoothing" to avoid over-fitting problem. This formulation was used to infer transmembrane region counts [8].

Schemes described by equation (2) are sometimes successful for nonlinear time series prediction problems, where the inner dynamical system has an infinite number of states [4], hand writing recognition problems [10], and online signature verification problems [7], where the inner dynamical system has a finite number of states. In these three problem classes, index parameter *t* is *time*, while in protein primary sequence, *t* stands for spatial *position*. Specification (2) assumes that the inner stochastic dynamical system is the first order, and the observation mechanism is independent (with respect to probabilities involved) of the inner dynamical system, although generalizations are possible.

2.1 Structure

The model structure $\mathcal{H}$ is crucial for successful applications of (2) so that model structure must be carefully designed taking into account the specific purpose(s) of the prediction problem as well as the available data sets. Although someone may want to design a model structure as detailed as possible taking into account the many facets of transmembrane proteins, the number of transmembrane proteins with known structure is severely limited and detailed models with many delicate parameters to be tuned would be infeasible. This is one aspect of the data fitting *vs.* simplicity dilemma (Ockham's razor).

Model proposed in this paper consists of the following. Let *m* denote the number of transmembrane regions.

(i) For each *m*, models $\mathcal{H}_m(n)$ are constructed, $n = 1,\ldots, n_m$, where n_m will be defined later;

(ii) Each model $\mathcal{H}_m(n)$ has an open loop structure consisting of alternative connections of loop region submodels $\mathcal{H}_{m\lambda_u}(n)$, $u = 1,\ldots, m+1$ and transmembrane region submodels $\mathcal{H}_{m\mu_v}(n)$, $v = 1,\ldots, m$. (Fig. 2(a));

(iii) Within a submodel $\mathcal{H}_{m\mu_v}(n)$ for transmembrane region, there are τ states where a simple left-to-right topology with self loop is built in as indicated by Fig. 2(b) . τ is defined as the average residue length of transmembrane regions given the whole training data sets;

(iv) The submodel for loop region $\mathcal{H}_{m\lambda_u}(n)$ has only one state with a self loop (Fig. 2(c));

(v) The first component of output O_t is the Kyte-Doolittle index; the second component is the charge associated with an amino acid.

2.2 Learning

Recall that for each *m*, the model $\mathcal{H}_m(n)$ proposed has 2*m*+1 submodels and *m*(τ+1)+1 states. Let

$$\{O_t^{m_k}\}_{t=1}^{T_{m_k}}, \qquad k = 1,....,|O^m|$$

be the training data sets for a particular *m*, where $|O^m|$ denotes the number of available data sets with *m* transmembrane regions. The proposed algorithm attempts to construct one model from one training data set so that n_m in (i) is equal to $|O^m|$.

[Step 1 K-D Index Emission Probabilities]

[*Learning* $b_{ik_1}^1$]

Step 1.1 (Flooring)

For each state q_i of $\mathcal{H}_{m\mu_v}(n)$ *set*

$$\tilde{b}_{i_l k_1}^1(\mu_v) := \frac{n(\{KD\}, k_1; \mu_v) + \beta_\mu}{\sum\limits_{K-D\ index\ k_1\ within\ \mathcal{H}_{m\mu_v}(n)} (n(\{KD\}, k_1; \mu_v v) + \beta_\mu)},$$

uniformly with respect to $i_l,\ l = 1,...,\tau$

For state q_i of $\mathcal{H}_{m\lambda_u}(n)$ *let*

$$\tilde{b}_{i_l k_1}^1(\lambda_u) := \frac{n(\{KD\}, k_1; \lambda_u) + \beta_\lambda}{\sum\limits_{K-D\ index k_1\ within\ \mathcal{H}_{m\lambda_u}(n)} (n(\{KD\}, k_1; \lambda_u) + \beta_\lambda)},$$

where

$n(\{KD\}, k_1; \mu_v) :=$ *number of residues with K-D index k_1 within transmembrane region* $\mathcal{H}_{m\mu_v}(n)$

$n(\{KD\}, k_1; \lambda_u) :=$ *number of residues with K-D index k_1 within transmembrane region* $\mathcal{H}_{m\lambda_u}(n)$

β_μ *and* β_λ *are hyperparameters*

Step 1.2 (Smoothing)

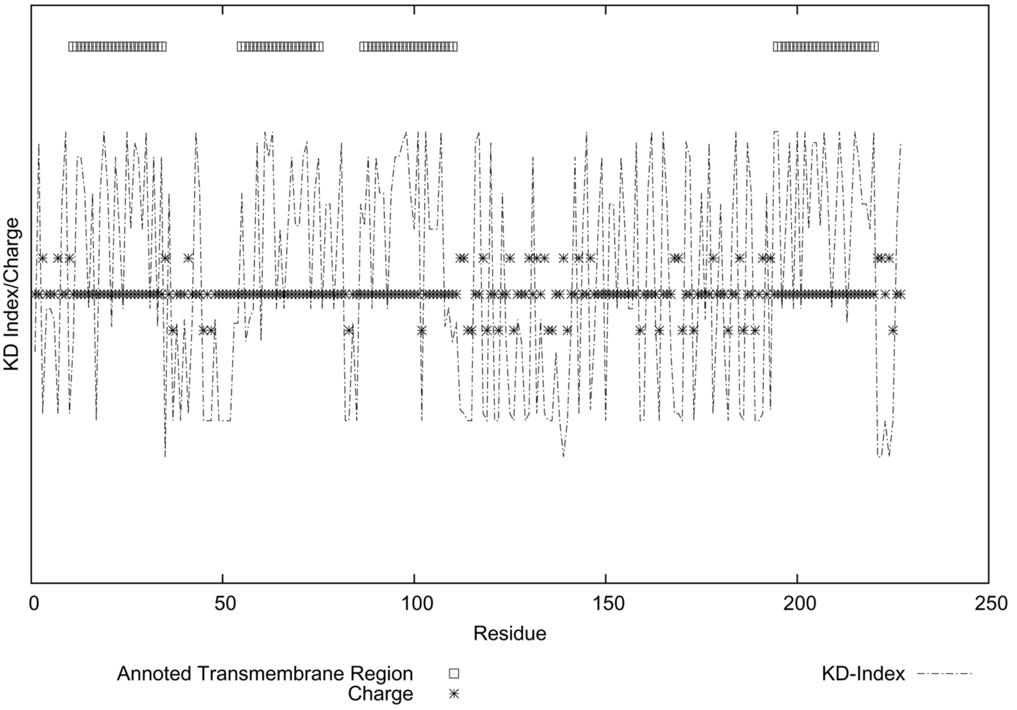
$$\hat{b}_{ik_1}^1(\mu_v) := \frac{1}{z_i} \sum_{j:\ |k_1 - k_j| \le 1} v_j\ \tilde{b}_{ik_j}^1(\mu_v)$$

$$v_j := \int_{x=|k_1-k_j|-1/2}^{x=|k_1-k_j|+1/2} \exp(-x^2 / 2\pi\sigma^2)dx$$

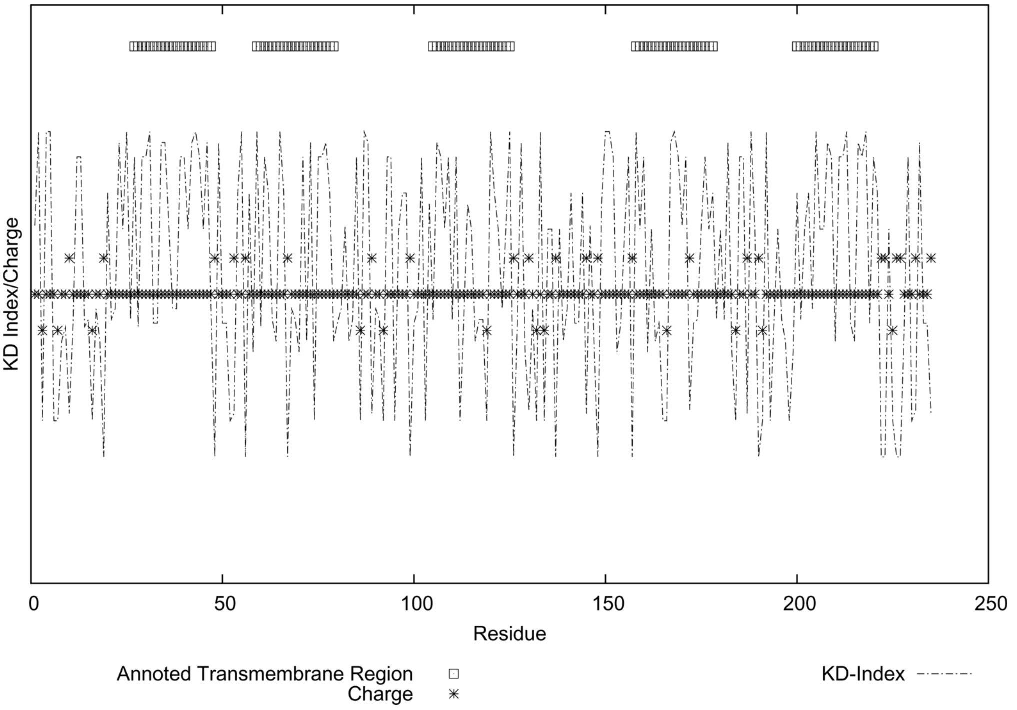
where σ *is a hyperparameter and* z_i *is a normalization constant.*

Similar smoothing is done for $\hat{b}_{ik_1}^1(\lambda_u)$

In this algorithm, the emission probabilities are the same within individual submodels. Over-fitting is avoided by hyper parameters chosen (in this paper) in an empirical manner even though Bayesian inference [4] is possible (one of our next projects). Note that Step 1.2 would have been impossible if the nearness between two amino acids were not defined. Note also that there are four amino acids out of 20, which have the same K-D index (-3.5): ASP, ASN, GLU, and GLN.

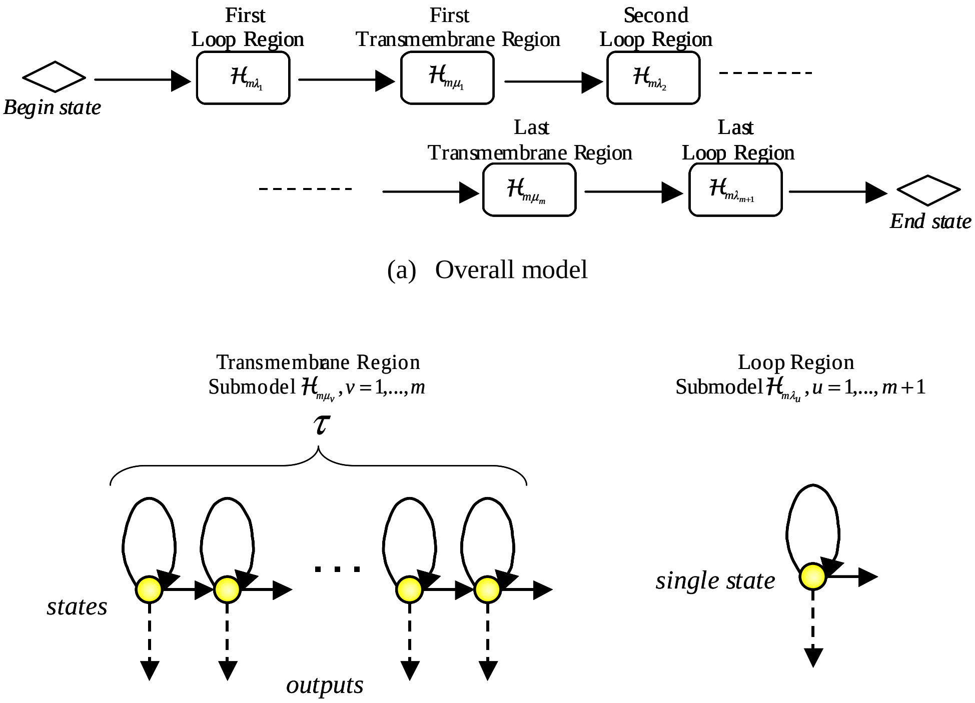


(a) (AC: P21926 ID: CD9_HUMAN)



(b) (AC P02913 ID: HISQ_SALTY)

Fig. 1 Typical K-D Index and charge trajectories
AC stands for Accession number and ID is the entry name of the sequence defined in Swiss-Prot.



(a) Overall model

(b) Transmembrane region submodel

(c) Loop region submodel

Fig. 2 Transmembrane protein model $\mathcal{H}_m$ with m transmembrane regions

[Step 2 Charge Emission Probabilities]

[*Learning* b_{ik}^2]

For each state q_i *of* $\mathcal{H}_{m\mu_v}(n)$ *set*

$$\hat{b}_{i_l k2}^2(\mu_v) := \frac{n(\{Charge\}, k_2; \mu_v) + \gamma_\mu}{\sum_{Charge\ \ k_2\ \ within\, \mathcal{H}_{m\mu_v}(n)} (n(\{Charge\}, k_2; \mu_v) + \gamma_\mu)},$$

uniformly with respect to $i_l,\ l = 1,\ldots,\tau$

For state q_i *of* $\mathcal{H}_{m\lambda_u}(n)$ *let*

$$\hat{b}_{i_l k2}^2(\lambda_u) := \frac{n(\{Charge\}, k_2; \lambda_u) + \gamma_\lambda}{\sum_{Charge\ \ k_2\ \ within\, \mathcal{H}_{m\lambda_u}(n)} (n(\{Charge\}, k_2; \lambda_u) + \gamma_\lambda)}$$

where

$n(\{Charge\}, k_2; \mu_v)$:= *number of residues with Charge* k_2 *within transmembrane region* $\mathcal{H}_{m\mu_v}(n)$

$n(\{Charge\}, k_2; \lambda_u)$:= *number of residues with Charge* k_2 *within loop region* $\mathcal{H}_{m\lambda_u}(n)$

γ_μ *and* γ_λ *are hyperparameters.*

Remark:

Histidine can assume two possible charge values depending on its pH. In the experiment reported below, the Histidine charge will be assumed to be +1. The two different possibilities should be examined in our future works. Since the number of Histidines appears to be small in the data sets used in our experiment, our tentative assumption does not appear to have a significant effect on prediction performance.

[Step 3 State Transition Probabilities]

Given residue sequence, consider the decomposition;

$$\{O_t^{m_k}\}_{t=1}^{\eta_{\lambda_1}(m_k)},\ \{O_t^{m_k}\}_{t=\eta_{\lambda_1}(m_k)+1}^{\eta_{\lambda_1}(m_k)+\eta_{\mu_1}(m_k)},\ \{O_t^{m_k}\}_{t=\eta_{\lambda_1}(m_k)+\eta_{\mu_1}(m_k)+1}^{\eta_{\lambda_1}(m_k)+\eta_{\mu_1}(m_k)+\eta_{\lambda_2}(m_k)},$$

$$\ldots\ldots\ldots, \{O_t^{m_k}\}_{t=\eta_{\lambda_1}(m_k)+\eta_{\mu_1}(m_k)+\eta_{\lambda_2}(m_k)+\ldots\ldots+1}^{T_{m_k}}$$

[*Learning* a_{ij}]

For each state q_i *of* $\mathcal{H}_{m\lambda_u}(n)$ *set*

$$\hat{a}_{ij}(\lambda_u) := \begin{cases} 1 - 1/\eta_{\lambda_u}(m_k), & j = i \\ 1/\eta_{\lambda_u}(m_k), & j = i+1 \\ 0, & otherwise \end{cases}$$

For state q_i *of* $\mathcal{H}_{m\mu_v}(n)$ *set*

$$\hat{a}_{ij}(\mu_v) := \begin{cases} \alpha_{\mu_v i}(m_k), & j = i \\ 1 - \alpha_{\mu_v i}(m_k), & j = i+1 \\ 0, & otherwise \end{cases}$$

where $\alpha_{\mu_v i}(m_k)$ *is a parameter to be tuned.*

Remarks:

(i) Given m, each submodel $\mathcal{H}_{m\mu_v}(n)$ has the same number of states as well as the same topology. Therefore a_{ij} are the same for all n, however, *emission probabilities* will be different because in our learning rule, each data set has a different number of K-D indices and charge. Thus, $n_m = m_k$, i.e., each data set creates one model, though considerations should be given in our future works, to redundant models being absorbed into a representative model.

(ii) In the above formulation, emission probabilities $\{\hat{b}_{ik}^1\}$ and $\{\hat{b}_{ik}^2\}$ are assumed to be independent for the sake of simplicity while in reality, they are not.

(iii) We chose not to use Baum-Welch for several reasons. First, it often suffers from local minima. Second, we wanted to test the degree of reasonableness of our first trial parameter values so that our proposed structure makes sense. Of course, the learning scheme must be improved in various ways, which will be our future and current research topics.

(iv) There could be better hydropathy indices other than the Kyte-Doolitle index. Actually as many as eighty different hydropathy indices have been proposed.

(v) Observe that each model in our scheme carries a fixed value of *m*, the number of transmembrane regions, and is entirely "open loop" except for those self loops associated with each state within individual submodels. In contrast, transitions between submodels are allowed in [2] so that *m* is not fixed.

2.3 Predictions

Let $D_{test} \coloneqq \{O_t^{test}\}_{t=1}^{T_{test}}$ be a test sequence. In the prediction phase *m* as well as the associated state sequence $\{Q_t\}$ is *unknown*. Observe that given a model $\mathcal{H}_m(n)$, each state q_i is associated with a unique submodel $\mathcal{H}_{m\mu_v}(n)$ or $\mathcal{H}_{m\lambda_u}(n)$. Let us first recall our prediction of transmembrane count $\hat{m}$ reported in [8]:

$$\hat{m} \coloneqq \arg\max_{m} \left(P\left(D_{test} \mid \hat{w}, \mathcal{H}_m(\hat{n}_m)\right)\right)$$

$$= \arg\max_{m} \left(\sum_{\text{all}\,\{Q_t\}_{t=1}^{T_{Test}-1}} P(\{O_t^{Test}\}, \{Q_t\}, Q_{T_{Test}} = q_{m(\tau+1)+1} \mid \hat{w}, \mathcal{H}_m(\hat{n}_m)) \right)$$

where

$$\hat{n}_m \coloneqq \arg\max_{n \text{ with } m \text{ fixed}} \left(P\left(D_{test} \mid \hat{w}, \mathcal{H}_m(n)\right)\right)$$

$$\hat{w} \coloneqq \{\{\hat{a}_{ij}\}, \{\hat{b}_{ik_1}^1, \hat{b}_{ik_2}^2\}, \{\hat{\pi}_i\}\}$$

is the learned parameter vector.

[Prediction of Transmembrane Regions]

Amino acid associated with O_t^{test} *is predicted to be in the* v *- th transmembrane region* $\mathcal{H}_{\hat{m}\mu_v}$ *if*

$$Q_t^* \in \mathcal{H}_{\hat{m}\mu_v}$$

where

for $t = 1,\ Q_1^* \coloneqq q_1$;

whereas for $t > 1$

$$Q_t^* \coloneqq \arg\max_{q_i \in \{q_j, q_{j+1}\}} P(O_{t+1}^{test}, \ldots\ldots, O_T^{test} \mid Q_t = q_i, \hat{w}, \mathcal{H}_{\hat{m}}(\hat{n})) \quad (3)$$

with $Q_{t-1}^* = q_j$

Amino acid associated with O_t^{test} *is predicted to be in the* u *- th Loop Region* $\mathcal{H}_{\hat{m}\lambda_u}$ *if*

$$Q_t^* \in \mathcal{H}_{\hat{m}\lambda_u}$$

Remarks

(i) Observe that

$$P(O_{t+1}^{test}, \ldots\ldots, O_T^{test} \mid Q_t = q_i, \hat{w}, \mathcal{H}_{\hat{m}}(\hat{n}))$$

is the likelihood of state Q_t being at q_i, where parameter *w* and model $\mathcal{H}$ are fixed. This can be used to write the equation as

$$P(Q_t = q_i \mid O_{t+1}^{test}, \ldots\ldots, O_T^{test}, \hat{w}, \mathcal{H}_{\hat{m}}(\hat{n}))$$

$$\propto P(O_{t+1}^{test}, \ldots\ldots, O_T^{test} \mid Q_t = q_i, \hat{w}, \mathcal{H}_{\hat{m}}(\hat{n})) \qquad (4)$$

under uniform prior for $P(\mathcal{H}_{\hat{m}}(\hat{n}))$.

(ii) In words, the left hand side of equation (4) is the probability of state Q_t is at q_i, given the test primary sequence for *t+1, t+2,….,T.*

(iii) Equation (3) is not the only way of predicting transmembrane regions. In our next project, we will refine the prediction scheme, which will be reported elsewhere.

3 EXPERIMENT

3.1 Data Sets

One of the very difficult issues in protein structure prediction problems in general, and transmembrane protein structure prediction problems in particular, seems to be the difficulty associated with acquiring appropriate data sets for experiments. The amino acid sequences for our experiment reported below were downloaded from the ftp site mentioned in [5] (ftp://ftp.ebi.ac.uk/databases/testsets/transmembrane). Of the downloaded amino acid sequences, those with the following clear annotations were used for our experiment:

DOMAIN CYTOPLASMIC, DOMAIN MATRIX, DOMAIN EXTRACELLULAR, DOMAIN INTERMEMBRANE, DOMAIN PERIPLASMIC, and TRANSMEM for which we have interpreted CYTOPLASMIC, MATRIX, EXTRACELLULAR, INTERMEMBRANE, and PERIPLASMIC as loop segments with TRANSMEM as a transmembrane segment.

There is an important issue to be examined with a considerable amount of care. After performing prediction experiment, we naturally want to compare the prediction performance with the best existing algorithms or tools. To accomplish this, we need to know which data sets have or have not been used to train a particular existing tool. Within the authors' present environment, it is extremely difficult, if not impossible, to identify such data sets.

There are four different classes of data sets in the above ftp site, A, B, C, and D. This classification is in accordance with the degree of reliabilities of the proteins structures. Structures of the data sets classified as A, B, and C, appear to be reasonably well analyzed even though the degree of reliability differs. Therefore we suspect that many of the data sets in A, B, and C, if not all, have already been used for training the existing tools. This implies that it would be appropriate to use data sets in A, B, and C as training data sets but not test data sets for comparing performance of different algorithms. It is extremely difficult, if not impossible, to perform new training with new training data sets for the existing prediction tools developed by other researchers. To remedy this, we chose the data sets in files A, B, and C in the above site for training, and used data sets in files D for testing. The number of data sets in file D used for training well-known tools could be small, and it may put different tools including ours on a reasonably equal footing. A natural drawback of this choice is the fact that the test data sets in D are less reliable. This dilemma will continue to be important issues in comparing different protein structure prediction algorithms.

3.2 Result

One of the important parameters, among others, is τ, the number of states in each transmembrane region. With several preliminary experiments, τ =21 was chosen. Table 1 illustrates some details of the data sets, i. e., the number of data sets in A, B, C, and D.

m	Number of Data Sets			
	Training Data Sets			Test Data Sets
	Class A	Class B	Class C	Class D
1	13	3	20	13
2	4	2	5	0
3	2	1	10	1
4	0	1	20	13
5	4	1	3	3
6	0	3	7	1
7	3	0	2	17
8	0	2	4	0
9	0	0	1	0
10	0	0	4	1
11	0	1	0	0
12	2	7	12	1
14	0	0	1	0
15	0	0	1	0
Total	28	21	90	50
	139			

Table 1: Training and test data sets

There were altogether 230 transmembrane regions. Performance evaluation criterion follows that of [6]. In order to define performance criterion, consider

(i) True Positive (TP) predictions. A TP must share *at least nine residues* with transmembrane region of the reference annotation. The following gives a schematic of this concept where “T” stands for amino acid within transmembrane region, while “–“ stands for amino acid in a loop region.

```
Annoted   -------TTTTTTTTTTTTTTT--------
Predicted -----TTTTTTTTTTTTTTTT---------
```

(ii) False Negative (FN) Predictions. FN’s are those transmembrane regions that are not predicted which is schematically described by;

```
Annoted   -------TTTTTTTTTTTTTTT--------
Predicted ------------------------------
```

(iii) False Positive (FP) Predictions. An FP stands for the predicted transmembrane region that is not existent as transmembrane region in the reference protein test set. This is schematically described by;

```
Annoted   ------------------------------
Predicted -------TTTTTTTTTTTTTTT--------
```

Remark:

Each predicted transmembrane region should correspond to only one reference transmenbrane region. This excludes the possibility of double counting TP's. For instance, the following prediction has one TP and one FN instead of two TP’s:

```
Annoted   -----TTTTTTTTT-TTTTTTTTTT-----
Predicted -----TTTTTTTTTTTTTTTTTTTT-----
```

Performance criterion is defined by

$$Percent\ Correct := \left(1 - \frac{n(FN) + n(FP)}{n(TP) + n(FN)}\right) \times 100\,(\%)$$

where n(TP), n(FN) and n(TP) denote the True Positive counts, False Negative counts, and False Possitive counts.

which we suspect is the criterion in [6], however, the equation is not explicitly written out. The proposed algorithm gave:

n(*TP*)=*224*, *n*(*FN*)=*6*, *n*(*FP*)=*3*,

Percent Correct=96.09%

Figure 3 illustrates some of the prediction results. Figure 3(a), 3(b) and 3(c) are examples of predictions where all transmembrane regions were True Positive, while (d) shows an example, which contains one False Negative prediction although all other predictions are correct, according to the above definitions.

Exact comparisons with other prediction algorithms are difficult because the data sets that are used for training could be different. For comparison purposes, however, we used the same test data set against TMHMM [2] , one of the algorithms that are most often referred to in transmembrane protein structure predictions, and evaluated it by accessing
http://www.cbs.dtu.dk/services/TMHMM/.
The algorithm gave:

n(*TP*)=*219*, *n*(*FN*)=*11*, *n*(*FP*)=2,

Percent Correct = 94.35%

Performance comparisons of various prediction algorithms up to year 2000 are reported in [6].

4 CONCLUSIONS

The experimental results reported in Section 3 appear encouraging since no fine-tuning has been done. Naturally there are many facets of the algorithm to be improved:

(i) Parameters, hyperparameters as well as states can be inferred via Bayesian framework where Monte Carlo can be utilized;

(ii) Sidedness (interior/exterior) can be predicted where charge trajectories could be more important than the present problem;

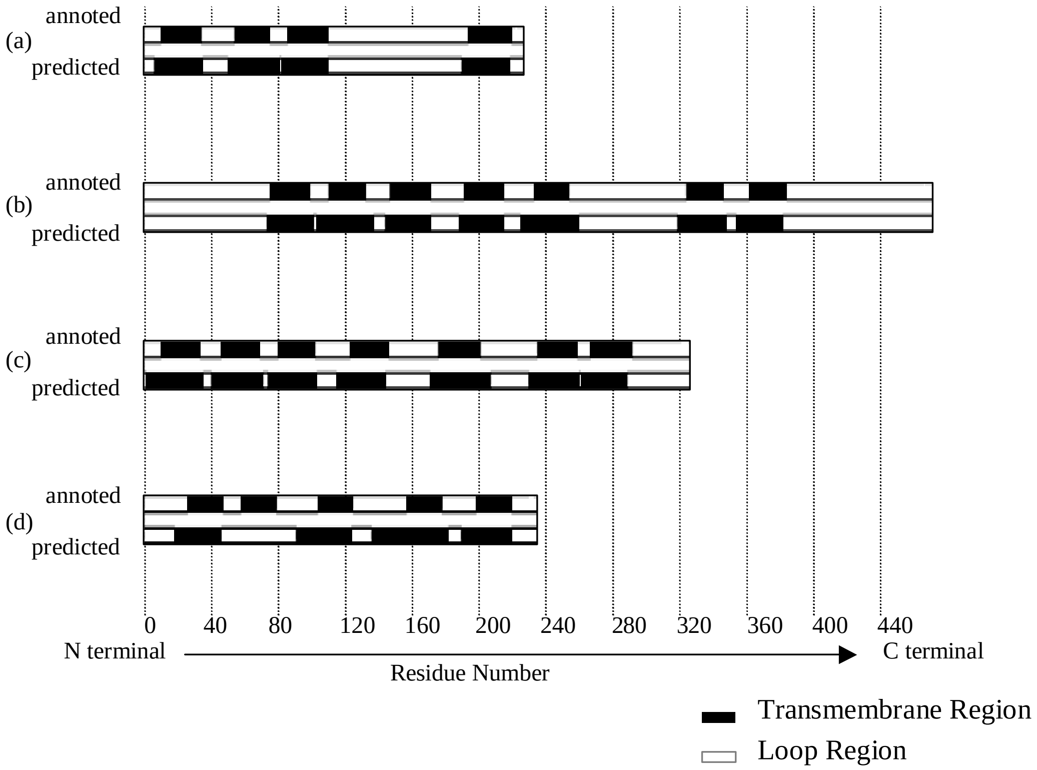
(iii) Definition of true positive predictions [6] seems rather weak, so that further elucidation will be necessary.

(iv) It will be interesting to see the prediction capabilities of the proposed algorithm on larger data sets including human proteins.

(v) A more detailed structure should be considered, for instance, by considering more than one state in the loop regions and by incorporating boundary regions between transmembrane and loop regions;

(vi) Other physico-chemical quantities could be taken into account for improvements;

(vii) Three-dimensional structure predictions, which are challenging.



(a) AC: P21926 ID:CD9_HUMAN , (b) AC: P18599 ID: 5H2A_CRIGR,
(c) AC: P11616 ID: AA1R_CANFA, (d) AC: P02913 ID:HISQ_SALTY

Fig. 3 Typical Prediction Results. (a), (b) and (c) are examples of predictions where all transmembrane regions were True Positive, while (d) shows an example which contains one False Negative prediction although all other predictions are correct, according to the definitions described in the text.

5 References

[1] T. Hirokawa, S. Boon-Chieng, and S. Mitaku, "SOSUI: classification and secondary structure prediction system for membrane proteins," Bioinformatics, vol. 14, pp. 378-379, 1998.

[2] A. Krogh, B. Larsson, G. von Heijne, and E. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes," J. Mol. Biol., vol. 305, pp. 567-580, 2001

[3] J. Kyte, and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," J. Mol. Biol., vol. 157, pp. 105-132, 1972

[4] T. Matsumoto, Y. Nakajima, M. Saito, J. Sugi, and H. Hamagishi, "Reconstructions and predictions of nonlinear dynamical systems: A Hierarchical Bayesian Approach," IEEE Trans. Signal Processing, vol. 49, pp. 2138-2155, 2001

[5] S. Möller, E. Kriventseva, and R.Apweiler, "A collection of well characterized integral membrane proteins," Bioinformatics, vol. 16, pp. 1159-1160, 2000

[6] S. Möller, M. Croning, R.Apweiler, "Evaluation of methods for the prediction of membrane spanning regions", J. Mol. Biol., vol. 17, pp. 646-653, 2001

[7] D. Muramatsu, and T. Matsumoto, "An online HMM signature verifier incorporating signature trajectories," International Conference on Document Analysis and Recognition, Vol. 1, pp. 438-442, Edinburgh, Scotland, August 2003

[8] D. Muramatsu, S. Hashimoto, T. Tsunashima, T. Kaburagi, M. Sasaki, and T. Matsumoto, "Inferring transmembrane region counts with Hydropathy Index/Charge two dimensional trajectories of stochastic dynamical systems", IEEE NNSP International Workshop, Special Session on Bioinformatics, pp101-110, Toulouse, France, Sept. 2003

[9] B. Rost, R. Casadio, P. Fariselli, and C. Sander, "Transmembrane helices predicted at 95% accuracy", Protein Science, vol. 4, pp. 521-533, 1995

[10] H. Yasuda, K. Takahashi, and T. Matsumoto, "A discrete HMM for online handwriting recognition," Int. J. Pattern Recognition and Artificial Intelligence, vol. 4, pp. 675 – 688, 2000

6 Acknowledgements

Discussions with T. Hirokawa of Computational Biology Research Center, AIST Tokyo, H. Sakai and T. Tsuji of Tokyo University of Agriculture and Technology, T. Ohnishi of Tokyo Medical Dental University, and S. Moeller of European Bioinformatics Institute, greatly helped us in preparing this manuscript. Thanks are also due to the reviewers for their comments.