OPEN ACCESS Freely available online

PLoS one

MemBrain: Improving the Accuracy of Predicting Transmembrane Helices

Hongbin Shen, James J. Chou*

Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, United States of America

Abstract

Prediction of transmembrane helices (TMH) in α helical membrane proteins provides valuable information about the protein topology when the high resolution structures are not available. Many predictors have been developed based on either amino acid hydrophobicity scale or pure statistical approaches. While these predictors perform reasonably well in identifying the number of TMHs in a protein, they are generally inaccurate in predicting the ends of TMHs, or TMHs of unusual length. To improve the accuracy of TMH detection, we developed a machine-learning based predictor, MemBrain, which integrates a number of modern bioinformatics approaches including sequence representation by multiple sequence alignment matrix, the optimized evidence-theoretic K-nearest neighbor prediction algorithm, fusion of multiple prediction window sizes, and classification by dynamic threshold. MemBrain demonstrates an overall improvement of about 20% in prediction accuracy, particularly, in predicting the ends of TMHs and TMHs that are shorter than 15 residues. It also has the capability to detect N-terminal signal peptides. The MemBrain predictor is a useful sequence-based analysis tool for functional and structural characterization of helical membrane proteins; it is freely available at http://chou.med.harvard.edu/bioinf/MemBrain/.

Citation: Shen H, Chou JJ (2008) MemBrain: Improving the Accuracy of Predicting Transmembrane Helices. PLoS ONE 3(6): e2399. doi:10.1371/journal.pone.0002399

Editor: Bostjan Kobe, University of Queensland, Australia

Received February 28, 2008; **Accepted** April 27, 2008; **Published** June 11, 2008

Funding: This work was supported by the NIH and the Pew Scholars Program in the Biomedical Sciences awarded to J.J.C.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: james_chou@hms.harvard.edu

Introduction

Motivation for a more accurate TMH predictor

Membrane-embedded α helical, polytopic proteins constitute the majority of ion channels, transporters, and receptors in living organisms. This class of proteins, which account for ~40% of all membrane proteins, are difficult targets for high resolution structural studies. Although experimentally determined structures of integral membrane proteins have been increasing at a fast rate in recent years, they only sum to less than 1% of the structures in the Protein Data Bank (PDB). Probably the first analysis that researchers perform when studying a helical membrane protein, whether it is for functional or structural characterization, is prediction of TMHs from the protein amino acid sequence. Knowledge of TMHs is very useful in initial elucidation of the overall topology of the protein, as well as in the rational design of protein constructs for structural studies.

Computational tools for TMH prediction are widely available. In this paper and in previous papers on TMH prediction, TMH is defined as a segment of helix that is embedded in the membrane. Hence, TMH sequence ends when the transmembrane region ends, although the helix can continue beyond the membrane. In general, residues of TMHs are mostly hydrophobic. Hence, earlier TMH prediction programs, such as TOP-PRED [1], compute sequence hydrophobicity from amino acid hydrophobicity scales assigned by biophysical and chemical measurements [2–4], and predict TMH propensity based on the average hydrophobicity score of a sliding prediction window of N successive residues along the sequence. Later predictors use more statistics-based, machine learning techniques. For example, PHDhtm [5] is based on neural networks, and TMHMM [6] and Phobius [7] are based on the hidden Markov model. The available TMH predictors are used routinely in membrane protein characterization and, in most cases, are sufficiently reliable in providing descriptive information about TMHs [8].

However, as more high resolution structures of helical membrane proteins become available, we learn that TMH has a wide length distribution. About 5% of the TMHs in the known structures are very short (<15 residues) and only span the membrane partially. These helices are known as the 'half TMHs' (see an example in the structure of the glycerol-conducting channel [9]). Very long TMHs (>40 residues) have also been found in the membrane proteins, e.g., the metalloenzyme particulate methane monooxygenase protein [10]. None of the existing TMH predictors perform satisfactorily in detecting TMHs of irregular lengths. For example, TOP-PRED [1] predicts all the TMHs to be 21 residues long, TMHMM [6] cannot predict TMHs shorter than 16 residues or longer than 35 residues, and SOSUI [11] cannot predict TMHs longer than 25 residues.

We developed a TMH prediction method, named MemBrain, which aims to improve the accuracy of TMH prediction. MemBrain was trained using the standard training dataset that was used by many other predictors, yet performed ~20% better than others when tested with a benchmark testing dataset. The improvement came mainly from the capability of MemBrain to predict accurately the ends of TMHs and therefore to detect TMHs of irregular lengths. Such capability was realized by applying the powerful **o**ptimized **e**vidence-**t**heoretic **K**-**n**earest **n**eighbor (OET-KNN) prediction algorithm [12–14] to protein sequence representations that include sequence evolution infor-

mation, and by merging results from prediction sequence windows of different sizes. Our results show that, with the fast expanding database of experimental membrane protein structures, there is still much room for improving the accuracy of TMH prediction using a pure statistics-based protocol.

Results

The algorithm

A flowchart of the MemBrain predictor is shown in Figure 1. We represented a protein sequence of N residues by the **p**osition-**s**pecific **s**coring **m**atrix (PSSM) (N rows and 20 columns), generated using the PSI-BLAST program [15] (see Methods section). The PSSM contains sequence evolution information from multiple sequence alignment against the SWISS-PROT protein database, and therefore provides a more complete description of the characteristics of a protein sequence. The propensity of a residue at positions i for being a part of a TMH was predicted based on a sequence segment of length L centered on i, where L is an odd number that represents the prediction window size. The prediction window size has a profound effect on the prediction outcome. Large window size, e.g., $L=17$ (used in the PHDhtm predictor [5]), is more effective for predicting residues in the middle of a long TMH due to higher content of neighborhood information. However, it performs poorly for residues near the ends of TMHs, and is incapable of predicting half TMHs shorter than 15 residues. On the other hand, if L is too small, the prediction accuracy generally suffers as a result of losing the neighborhood sequence information. In the MemBrain predictor, we combined two window sizes to minimize the bias caused by the use of only one window size. We found that the fusion of two window sizes, 13 and 15, gave the best prediction results.

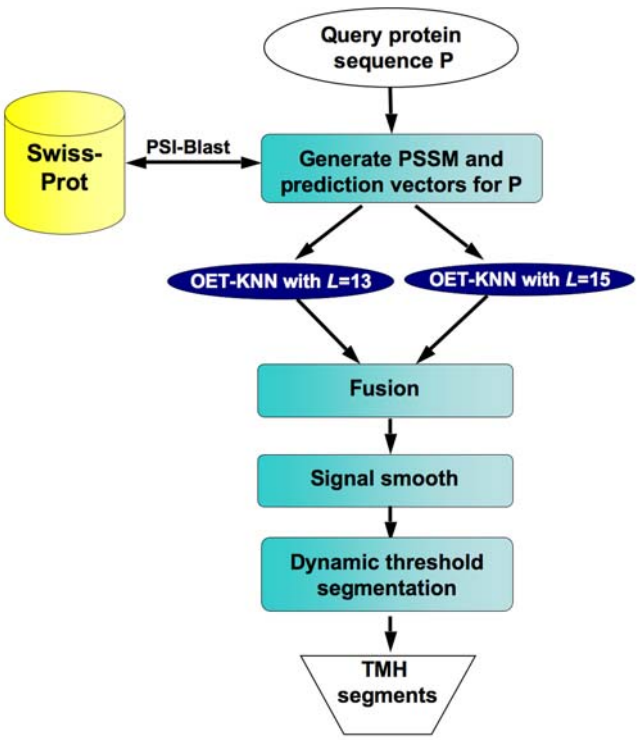

Figure 1. A flowchart diagram of the MemBrain protocol.
doi:10.1371/journal.pone.0002399.g001

For TMH prediction, we used the standard training dataset which was used by most other TMH predictors, including TMHMM [6], Phobius [7], THUMBU [16] and SVMtm [17]. This dataset includes 50 helical membrane proteins of known TMH regions (see Supplementary Table S1). For each of the 50 proteins, the PSSM was generated using the PSI-BLAST program. From the PSSM, the matrix elements ($L\times20$) for various sequence segments of $L=13$ or 15 were extracted and stored in the training vectors $\mathbf{t}_i^{13}$ or $\mathbf{t}_j^{15}$, respectively (see Methods section for details of constructing these vectors). These training vectors were labeled as 'TMH' if the residue j at the middle of the sequence segment is a part of a TMH, and were otherwise labeled as 'NOT TMH'. From the 50 PSSMs, we built a training set of 14,531 vectors of $L=13$ and 14,431 vectors of $L=15$. These vectors were used as statistical rulers for making predictions on the target protein.

Given a query protein, the PSSM was constructed and the query vector for sequence segment centered on residue i ($\mathbf{q}_i^L$) was defined. To predict the TMH propensity of residue i, denoted here as $\mathbf{E}_i$, we applied the OET-KNN algorithm for which the inputs are the query vector $\mathbf{q}_i^L$ and all $\mathbf{t}_j^L$s in the training set with the same dimension. The OET-KNN algorithm is a classification tool which has proven to be powerful in pattern recognition [12,14] as well as in the prediction of sub-cellular locations of proteins [13,18]. In the OET-KNN calculation (described in details in the Methods section), the Euclidean distances between $\mathbf{q}_i^L$ and all $\mathbf{t}_j^L$s were calculated, and the 50 closest matches were used to calculate $\mathbf{E}_i$, which ranges from 0 to 1, where 0 and 1 are zero and unity probability of TMH, respectively. The TMH propensity obtained for $L=13$, $\mathbf{E}_i^{13}$, was merged with that obtained for $L=15$, $\mathbf{E}_i^{15}$, by simple averaging. Thus the combined TMH propensity for residue i is $\mathbf{E}_i=\left(\mathbf{E}_i^{13}+\mathbf{E}_i^{15}\right)/2$, ranging from 0 to 1. The procedure was repeated to cover all residues, $(L\text{-}1)/2\leq i\leq N-(L\text{-}1)/2$, in the query protein.

For a query protein, the E_i vs. i plot gives an overview of the residue-specific TMH propensity. We used the median filter technique [19] to smooth the TMH propensity profile, and at the same time, to reduce noise. The final step is to determine the TMHs based on the smoothened propensity profile. In most other predictors, a fixed threshold is used to segment the scores, i.e., residues having scores larger than the threshold are assigned as TMH [11,17,20]. However, the optimal threshold for defining two TMHs separated by long loops is very different from the threshold required for identifying TMHs separated by short loops or tight turns. High-resolution structures show that two consecutive TMHs are often connected by very short loops or turns. In these cases, since the loop residues only represent a small region of the prediction window, the TMH propensity calculated for the short loops are higher than those of long loops. To solve this problem, we used a dynamic threshold method in which a base threshold propensity of 0.4 was first used to define TMH fragments. Then we raised the threshold according to the shape of the local propensity profile for identifying short loops or helical breaks in these fragments (see Methods section for details).

Finally, in some membrane proteins, the first N-terminal TMH is a N-terminal signal peptide. We included an extra module in the MemBrain program to detect potential N-terminal signal peptide in a membrane protein using methods described in ref. [21].

Performance

To test the MemBrain predictor and compare its performance with the existing TMH predictors, we constructed a testing dataset consisting of 70 helical membrane proteins of known high resolution structures which do not overlap with the training dataset (see Supplementary Table S2). There are a total of 378

TMHs in the testing dataset. The performances of the TMH predictors were evaluated with four different scores.

1. *The TMH prediction success rate (V_{TMH}).* V_{TMH} is simply the fraction of TMHs in the testing set that are correctly predicted [22]; it is defined as

$$V_{TMH} = \frac{\text{number of correctly predicted TMHs}}{\text{total number of TMHs}}, \quad (1)$$

where a TMH is considered predicted correctly if it has an overlap of at least 9 residues with the prediction. However, we note that such definition is not robust, and in some other studies, different lengths of residue overlap were used [22,23].

2. *The protein prediction success rate (V_P).* V_P is the fraction of helical proteins in the testing set that are correctly predicted [22]; it is defined as

$$V_P = \frac{\text{number of correctly predicted proteins}}{\text{total number of proteins}}, \quad (2)$$

where a protein is considered predicted correctly if all the TMHs in this protein are correctly predicted (as defined in V_{TMH} above) and the number of predicted TMHs is equal to the observed number of TMHs in the protein.

3. *The N and C scores.* These two scores evaluate the accuracy of predicting the ends of TMHs [22]. N and C scores are the number of N- and C-terminal residues that do not match when aligning the predicted and observed TMHs. In the best case, if the predicted and observed TMHs are completely matched, the N and C scores equal to 0.

4. *The normalized RMSD.* Finally, we calculated the normalized distance between the predicted and known TMH representation vectors, denoted by $\mathbf{p} = [p_1, p_2, \ldots, p_N]$, in which p_i is assigned to 1 if residue i is a part of a TMH and is otherwise assigned to 0. The normalized distance, or $RMSD_N$, is defined as

$$RMSD_N = \frac{\|\mathbf{p}-\mathbf{p}^0\|}{\|\mathbf{p}^0\|} = \frac{\left(\sum_{i=1}^{N}(p_i - p_i^0)^2\right)^{1/2}}{\left(\sum_{i=1}^{N}(p_i^0)^2\right)^{1/2}}, \quad (3)$$

where $\mathbf{p}$ and $\mathbf{p}^0$ are the predicted and known TMH representation vectors of a protein, respectively. The normalized RMSD is less subjective than the definition of V_{TMH} and V_P above.

Table 1 compares the performances of MemBrain and other TMH predictors as judged by the four different scorings described above. MemBrain performs significantly better than other predictors in all four scoring categories. The V_{TMH} and V_P scores have been widely used in evaluation of TMH predictors. MemBrain V_{TMH} and V_P scores are 97.9% and 87.1%, respectively, which are about 6–16% better than Phobius (the best performer in this scoring category among the published predictors). MemBrain also has an improved capability to predict correctly the ends of TMHs as shown by the mean N and C scores of 3.2 and 3.1, which are about 20% better than the best published predictor for this scoring category. Finally the MemBrain mean normalized rmsd is 0.35, also about 20% better than the second-best performing predictor Phobius. The observed and predicted TMHs for the 70 membrane proteins in the testing dataset are given in Supplementary Data S1.

Table 1. Performance comparison of various TMH predictors[a].

Predictor	V_{TMH}	V_P	N-score	C-score	$RMSD_N$
THUMBU[16] [b]	85.5%	47.1%	6.9±4.9	6.7±4.9	0.58±0.19
SOSUI[11] [c]	89.1%	57.1%	5.0±4.1	5.0±4.2	0.44±0.21
DAS-TMfilter[20] [d]	90.7%	64.3%	6.5±5.0	5.5±5.3	0.58±0.16
TOP-PRED[1] [e]	92.6%	60.0%	4.5±3.8	4.6±3.9	0.45±0.15
TMHMM[6] [f]	91.0%	65.7%	4.5±3.8	4.5±3.9	0.44±0.15
Phobius[7] [g]	91.8%	71.4%	4.6±4.0	4.4±4.1	0.44±0.19
MemBrain [h]	**97.9%**	**87.1%**	**3.2**±3.0	**3.1**±2.8	**0.35±0.14**

[a]The testing dataset consists of 378 TMH segments from 70 proteins (see Supplementary Table S2).
[b]http://sparks.informatics.iupui.edu/Softwares-Services_files/thumbup.htm [16].
[c]http://bp.nuap.nagoya-u.ac.jp/sosui/ [11].
[d]http://mendel.imp.ac.at/sat/DAS/DAS.html [20].
[e]http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html [1].
[f]http://www.cbs.dtu.dk/services/TMHMM/ [6].
[g]http://phobius.cgb.ki.se/ [7].
[h]http://chou.med.harvard.edu/bioinf/MemBrain/.
doi:10.1371/journal.pone.0002399.t001

Discussion

The above prediction scores obtained from a fairly complete testing dataset show that MemBrain is the best TMH predictor to date. Probably the most attractive feature of MemBrain is the improved ability in correctly identifying the ends of TMHs. This capability is important because there is a wide distribution of TMH length amongst the 70 helical polytopic membrane proteins in the testing dataset (Fig. 2a), e.g., TMH can be as short as 10 residues. Most TMH predictors cannot detect TMHs shorter than 15 residues (e.g., Figures 2b&c show that the shortest TMH predicted by TMHMM and Phobius, the predictors which gave the second best N and C scores in Table 1, is 17 residues). However the length distribution of TMHs predicted by MemBrain matches most closely to that of the observed dataset (Fig. 2d). We also noticed that MemBrain shows similar improvements in prediction when considering only TMHs that are longer than 15 residues (see Supplementary Table S3).

The improvement came from a combination of the steps used in our protocol shown in Figure 1. First, the PSSM representation contains sequence evolution information, which provides more complete sampling for statistical prediction methods. The advantage of a pure statistical approach over hydrophobicity-based prediction methods is that the prediction outcome does not depend on our interpretation of amino acid sequence in TMH formation, which could introduce bias. Second, the OET-KNN algorithm is a powerful classification method that can combine many different evidences and deal with the uncertainty to reach the optimal decision. Third, the fusion of two prediction window sizes provides more flexibility in accounting for length variation of TMHs, and thus reduces the bias towards a fixed TMH length introduced by using only one window size (as treated in all the previous predictors). Finally, assignment of TMHs using the dynamic threshold method further refines the prediction by detecting short loops and turns that separate TMHs.

A somewhat unsatisfying aspect of the TMH-only prediction is the complete absence of amphipathic, extramembrane helices that are common in helical membrane protein structures. In both the training and testing datasets, the TMH sequences are defined to end when the transmembrane regions end. However, according to many high resolution structures, a considerable portion of transmembrane helices extend well beyond the lipid bilayer and

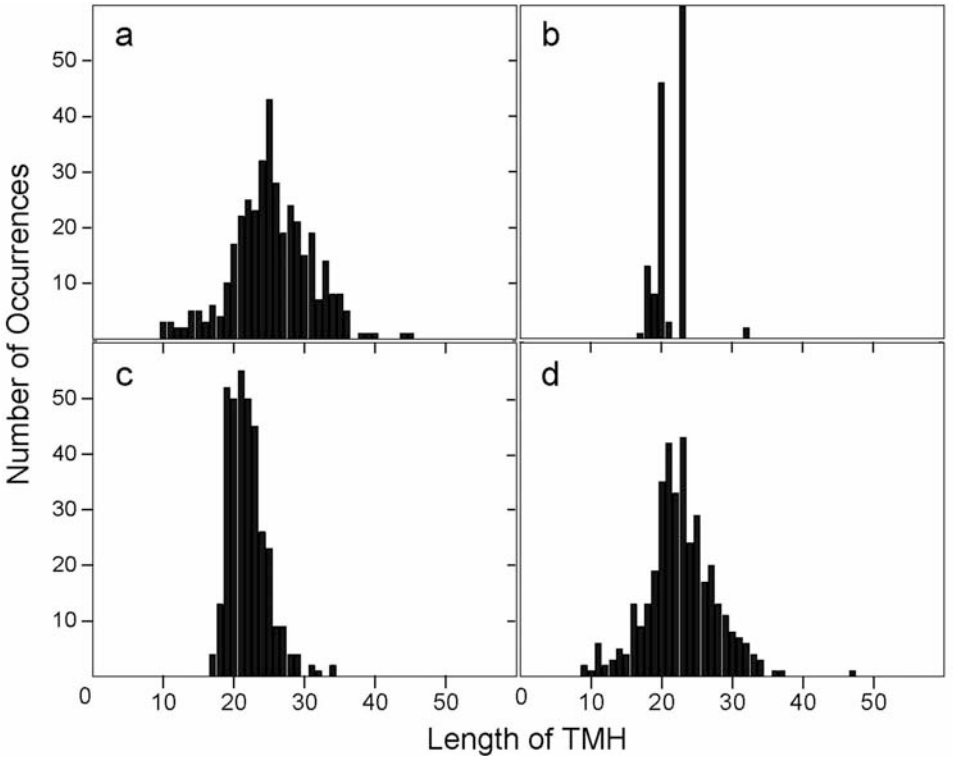

Figure 2. TMH length distribution in (a) 70 known membrane protein structures in the testing dataset, (b) TMHs predicted by TMHMM [6], (c) TMHs predicted by Phobius [7], and (d) TMHs predicted by MemBrain.
doi:10.1371/journal.pone.0002399.g002

become hydrophilic. Therefore, TMH predictors cannot predict the extramembrane portions of helices. Our future direction is to develop methods to predict both transmembrane and extramembrane helical segments in helical polytopic membrane proteins.

Methods

Construction of query and training vectors

The PSSM matrix of a protein $\mathbf{P}$ of N residues, which contains sequence evolution information, is defined as

$$\mathbf{P}^{\mathrm{PSSM}}=\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,20} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,20} \end{bmatrix}, \tag{4}$$

where $a_{i,j}$ denotes the probability of residue i of the protein being changed to amino acid type j as determined from multiple sequence alignments [15]. The matrix elements in Eq. 4 were generated using the PSI-BLAST [15], which searches the SWISS-PROT database (version 52.0 released on 6-March-2007) against the sequence of the protein. For prediction studies, a residue at position i of the protein can be represented by a query vector, $\mathbf{q}_i^L$, composed of the PSSM matrix elements of the query protein corresponding to a sequence segment of length L centered on i, e.g.,

$$\begin{aligned}\mathbf{q}_i^L=&\left[\left(a_{i-(L-1)/2,1},\ldots,a_{i-(L-1)/2,20}\right)\right.\\&\left(a_{i-(L-1)/2+1,1},\ldots,a_{i-(L-1)/2+1,20}\right)\cdots,\\&\left.\left(a_{i+(L-1)/2,1},\ldots,a_{i+(L-1)/2,20}\right)\right],\end{aligned} \tag{5}$$

where L is an odd number. Eq. 5 is also used to construct training vectors, $\mathbf{t}_j^L$, from their corresponding PSSM matrices of proteins in the training dataset.

Calculation of TMH propensity

Consider the problem of predicting the propensity of residue i of the query protein belonging to a structural pattern, denoted by ϕ, where

$$\phi=\begin{cases}1 & \text{TMH}\\ 0 & \text{NOT TMH}\end{cases}. \tag{6}$$

We represent the residue by a query vector $\mathbf{q}_i^L$ (see Eq. 5 above), constructed for prediction window size L. The knowledge basis used for the prediction is given by the training dataset, T^L, e.g.,

$$\mathrm{T}^L=\left\{\left(\mathbf{t}_1^L,\phi_1\right),\left(\mathbf{t}_2^L,\phi_2\right),\ \ldots,\ \left(\mathbf{t}_M^L,\phi_M\right)\right\}, \tag{7}$$

where vectors $\mathbf{t}_j^L$s were also constructed as in Eq. 5 for window size L, and their corresponding patterns ϕ_j's are known.

Let $\mathrm{S_K}$ be a set of vectors consisting of K $\mathbf{t}_j^L$s in T^L that have the shortest Euclidean distances to $\mathbf{q}_i^L$, referred to here as the K nearest neighbors of $\mathbf{q}_i^L$. For any $\mathbf{t}_j^L\in\mathbf{S}_\mathrm{K}$, the knowledge that $\mathbf{t}_j^L$ has a pattern ϕ is a piece of evidence which increases our belief that $\mathbf{q}_i^L$ also has the pattern ϕ. This evidence is quantified, as in refs. [24,25], by an *evidence function*

$$\mathrm{E}\left(\mathbf{q}_i^L\middle|\mathbf{t}_j^L,\phi\right)=\exp\left[-C_\phi^L\ D^2\left(\mathbf{t}_j^L,\mathbf{q}_i^L\right)\right]\delta\left(\phi_j,\phi\right), \tag{8}$$

where $D\left(\mathbf{t}_j^L,\mathbf{q}_i^L\right)$ is the Euclidean distance between $\mathbf{t}_j^L$ and $\mathbf{q}_i^L$, and

the parameter C_ϕ^L is associated with a particular pattern ϕ; the delta function in Eq. 8 is

$$\delta(\phi_j,\phi)=\begin{cases}1, & \text{if } \phi_j=\phi\\ 0, & \text{if } \phi_j\neq\phi\end{cases}. \quad (9)$$

In OET-KNN, C_ϕ^L is optimized by maximizing the prediction accuracy of every sample in T^L. Using the detailed optimization protocol described in ref. [14], we found the following values of C_ϕ^L: $C_1^{13}=0.105$, $C_0^{13}=0.094$, $C_1^{15}=0.096$, and $C_0^{15}=0.085$.

Combining the knowledge of the K nearest neighbors in S_K, the *evidence* of $\mathbf{q}_i^L$ belonging to the pattern ϕ is

$$\mathrm{E}(\mathbf{q}_i^L,\phi)=1-\prod_{j=1}^{K}\left(1-\mathrm{E}\left(\mathbf{q}_i^L\middle|\mathbf{t}_j^L,\phi\right)\right). \quad (10)$$

The final *evidences* $\mathbf{E}(\mathbf{q}_i^L,\phi)$ are then normalized as in

$$\mathbf{E}(\mathbf{q}_i^L,\phi)=\frac{\mathrm{E}(\mathbf{q}_i^L,\phi)}{\mathrm{E}(\mathbf{q}_i^L,\phi=1)+\mathrm{E}(\mathbf{q}_i^L,\phi=0)} \quad (11)$$

to satisfy the normalization condition $\mathbf{E}(\mathbf{q}_i^L,\phi=1)+\mathbf{E}(\mathbf{q}_i^L,\phi=0)=1$.

Finally, after merging the prediction results obtained using two different window sizes, $L=13$ and 15, the propensity of residue i belonging to TMH is

$$\mathbf{E}(\mathbf{q}_i,\phi=1)=\left[\mathbf{E}(\mathbf{q}_i^{13},\phi=1)+\mathbf{E}(\mathbf{q}_i^{15},\phi=1)\right]/2. \quad (12)$$

Dynamic threshold segmentation

To assign TMH fragments based on the propensity profile, we used a dynamic threshold segmentation approach. First, residues with propensity greater than or equal to 0.4 were considered as TMH. The base threshold, $\lambda=0.4$, was selected by optimizing the self-consistency test performance as was done in refs. [11,17,20]. A TMH is initially assigned when λ intersects the propensity profile at two consecutive points. For example, given $\lambda=0.4$, the N-terminal residue of a TMH is residue n0 if $E_{n0-1}<\lambda$ and $E_{n0}>\lambda$. Moving along the sequence, the next encounter of $E_{c0}>\lambda$ and $E_{c0+1}<\lambda$ defines the C-terminal residue of the TMH to be residue c0. Hence, the initial assignment of TMH is from residues n0 to c0. The value of λ was then increased by increment of 0.05 until λ intersects the profile within the initial TMH at four points. In this case, the original TMH was split into two TMH segments. The first TMH is from residues n0 to c1, where $E_{c1}>\lambda$ and $E_{c1+1}<\lambda$, and the second TMH is from residues n1 to c0, where $E_{n1-1}<\lambda$ and $E_{n1}>\lambda$. A TMH shorter than 5 residues was not segmented out and remained as a part of the original TMH. Figure 3 shows an example of dynamic threshold assignment of TMHs in the protein lactose permease of *Escherichia coli* (PDB code: 1PV7) [26]. Note that the short loops between the 3rd and 4th TMHs, and between the 9th and 10th TMHs were successfully detected using this method.

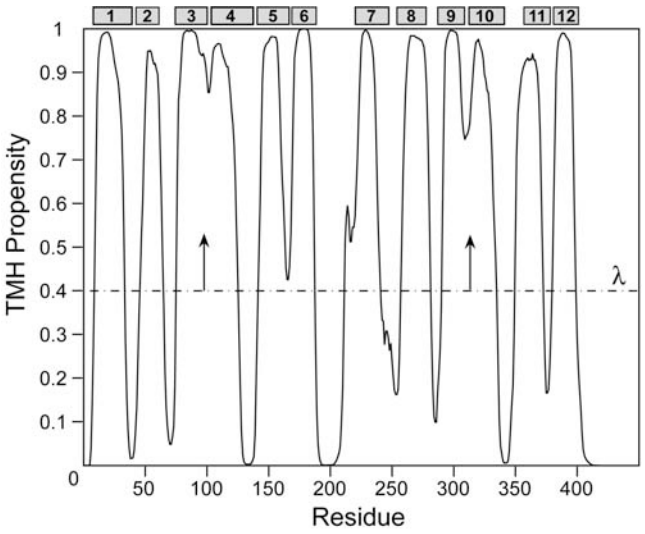

Figure 3. The residue-specific TMH propensity of lactose permease of *Escherichia coli* (PDB code: 1PV7) [26], illustrating the method of assignment of TMHs by dynamic threshold segmentation. The observed TMHs, assigned in ref. [26], are shown as the gray boxes.
doi:10.1371/journal.pone.0002399.g003

All algorithms used in MemBrain were implemented in the C programming language and executed in the Linux operating system.

Supporting Information

Table S1
Found at: doi:10.1371/journal.pone.0002399.s001 (0.02 MB DOC)

Table S2
Found at: doi:10.1371/journal.pone.0002399.s002 (0.02 MB DOC)

Table S3
Found at: doi:10.1371/journal.pone.0002399.s003 (0.05 MB DOC)

Data S1
Found at: doi:10.1371/journal.pone.0002399.s004 (0.10 MB DOC)

Acknowledgments

We thank Kirill Oxenoid and Matthew Call for useful discussion.

Author Contributions

Conceived and designed the experiments: HS JC. Performed the experiments: HS. Analyzed the data: HS. Wrote the paper: JC.

References

1. Claros MG, von Heijne G (1994) TopPred II: an improved software for membrane protein structure predictions. Comput Appl Biosci 10: 685–686.
2. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157: 105–132.
3. Chamberlain AK, Lee Y, Kim S, Bowie JU (2004) Snorkeling preferences foster an amino acid composition bias in transmembrane helices. J Mol Biol 339: 471–479.
4. Wimley WC, White SH (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat Struct Biol 3: 842–848.
5. Rost B, Fariselli P, Casadio R (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Sci 5: 1704–1718.
6. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567–580.

7. Kall L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338: 1027–1036.
8. White SH (2004) The progress of membrane protein structure determination. Protein Sci 13: 1948–1949.
9. Fu D, Libson A, Miercke LJ, Weitzman C, Nollert P, et al. (2000) Structure of a glycerol-conducting channel and the basis for its selectivity. Science 290: 481–486.
10. Lieberman RL, Rosenzweig AC (2005) Crystal structure of a membrane-bound metalloenzyme that catalyses the biological oxidation of methane. Nature 434: 177–182.
11. Hirokawa T, Boon-Chieng S, Mitaku S (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics 14: 378–379.
12. Shen HB, Chou KC (2006) Ensemble classifier for protein fold pattern recognition. Bioinformatics 22: 1717–1722.
13. Chou KC, Shen HB (2006) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-Nearest Neighbor classifiers. J Proteome Res 5: 1888–1897.
14. Zouhal LM, Denoeux T (1998) An evidence-theoretic K-NN rule with parameter optimization. IEEE Transactions on Systems, Man and Cybernetics 28: 263–271.
15. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, et al. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29: 2994–3005.
16. Zhou H, Zhou Y (2003) Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-Markov-model-based method. Protein Sci 12: 1547–1555.
17. Yuan Z, Mattick JS, Teasdale RD (2004) SVMtm: support vector machines to predict transmembrane segments. J Comput Chem 25: 632–636.
18. Chou KC, Shen HB (2007) Recent progress in protein subcellular location prediction. Anal Biochem 370: 1–16.
19. Makivirta A, Koski E, Kari A, Sukuvaara T (1991) The median filter as a preprocessor for a patient monitor limit alarm system in intensive care. Comput Methods Programs Biomed 34: 139–144.
20. Cserzo M, Eisenhaber F, Eisenhaber B, Simon I (2004) TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. Bioinformatics 20: 136–137.
21. Shen HB, Chou KC (2007) Signal-3L: A 3-layer approach for predicting signal peptides. Biochem Biophys Res Commun 363: 297–303.
22. Cuthbertson JM, Doyle DA, Sansom MS (2005) Transmembrane helix prediction: a comparative evaluation and analysis. Protein Eng Des Sel 18: 295–308.
23. Jones DT (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics 23: 538–544.
24. Shafer G (1976) A mathematical theory of evidence. Princeton, N.J.: Princeton University Press.
25. Denoeux T (1995) A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Transactions on Systems, Man and Cybernetics 25: 804–813.
26. Abramson J, Smirnova I, Kasho V, Verner G, Kaback HR, et al. (2003) Structure and mechanism of the lactose permease of Escherichia coli. Science 301: 610–615.