Energy-based Classification and Structure Prediction of Transmembrane Beta-Barrel Proteins

Van Du Tran, Philippe Chassignet, Saad Sheikh, Jean-Marc Steyaert
INRIA AMIB Team, Laboratory of Computer Science (LIX)
Ecole Polytechnique
Palaiseau CEDEX, France
{vandu, chassignet, sheikh, steyaert}@lix.polytechnique.fr

Abstract—Transmembrane β-barrel (TMB) proteins are a special class of transmembrane proteins which play several key roles in human body and diseases. Due to experimental difficulties, the number of TMB proteins with known structures is very small. Over the years, a number of learning-based methods have been introduced for recognition and structure prediction of TMB proteins. Most of these methods emphasize on homology search rather than any biological or chemical basis. We present a novel graph-theoretic model for classification and structure prediction of TMB proteins. This model folds proteins based on energy minimization rather than a homology search, avoiding any assumption on availability of training dataset. The ab initio model presented in this paper is the first method to allow for permutations in the structure of transmembrane proteins and provides more structural information than any known algorithm. The model is also able to recognize β-barrels by assessing the pseudo free energy. We assess the structure prediction on 42 proteins gathered from existing databases on experimentally validated TMB proteins. We show that our approach is quite accurate with over 90% F-score on strands and over 74% F-score on residues. The results are comparable to other algorithms suggesting that our pseudo-energy model is close to the actual physical model. We test our classification approach and show that it is able to reject α-helical bundles with 100% accuracy and β-barrel lipocalins with 97% accuracy.

Keywords—transmembrane proteins; β-barrels; protein structure prediction; super-secondary structure; permuted structure; Greek key; ab initio modeling; pseudo-energy model;

I. INTRODUCTION

Transmembrane proteins play several key roles in the human body including inter-cell communication, transportation of nutrients, and ion transport. They also play key roles in human diseases like depression, hypertension, cancer, thus are targeted by a majority of pharmaceuticals being manufactured today. The transmembrane proteins are divided into two main types according to their conformation: α-helical bundles and β-barrels. The TMB proteins, which are much less abundant than helical bundles, are found in the outer membrane of Gram-negative bacteria, mitochondria and chloroplasts. They perform diverse functions such as porins, passive or active transporters, enzymes, defensive or structural proteins [1]. Thus, the structure of TMB proteins is very important for both biological and medical sciences.

These proteins, which span the membrane entirely, make up 20 – 30% of identified proteins in most whole genomes. However, due to difficulties in determination of their structures, solved TMB structures constitute only a meagre 2% of the RCSB Protein Data Bank (PDB) [2]–[5]. This is mainly due to experimental difficulties and complexity of the TMB structure [6]. Consequently, various learning-based techniques have been developed for discriminating TMB proteins from globular and transmembrane α-helical proteins [6]–[8], and for predicting TMB secondary structures [7]–[12]. We first discuss these methods and their potential shortcomings in detail, and then proceed with describing our approach.

A. Existing Approaches

Ou et al. [10] proposed a method based on radial basis function networks to predict the number of β-strands and membrane spanning regions in β-barrel outer membrane proteins. Randall et al. [9] tried to predict the TMB secondary structure with 1D recursive neural network using alignment profiles. Gromiha et al. [7], [8] used the amino acid compositions of both globular and outer membrane proteins (OMPs) to discriminate OMPs and developed a feed forward neural network-based method to predict the transmembrane segments. Bagos et al. [11] produced a consensus prediction from different methods based on hidden Markov models, neural networks and support vector machines [8], [13]–[19]. Tractability has been an issue for some of these approaches. In order to overcome this limitation, Waldispühl et al. [12] used a structural model and pairwise interstrand residue statistical potentials derived from globular proteins to predict the supersecondary structure of TMB proteins. Freeman et al. [6] have introduced a statistical approach for recognition of TMB proteins based on known physicochemical properties.

Most of these rely on the learning assumptions in the underlying models as well as the sampling of proteins in their training set. However, the number of TMB proteins known today is very small. Thus, it is arguable whether these approaches will work well for recognizing and folding TMB proteins which are not homologous to those currently known. It is also important to note that none of these meth-
ods allow for permutations in protein structures. Generally, the TMB structures are not only a series of \( \beta \)-strands where each is bonded to the preceding and succeeding ones in the primary sequence, but they may contain Greek key or Jelly roll motifs as well, for instance, the C-terminal domain of the PapC usher (3L48 in PDB). This level of structure may be described as a permutation on the order of the bonded strands.

B. Our Approach

In this paper, we present a novel \textit{ab initio} model for classification and structure prediction of TMB proteins based on minimizing free energy in a graph-theoretic framework. It is able to deal with permuted TMB structures. The prediction accuracy is evaluated on known TMB proteins available in popular protein databases [20], and compared with existing software [9], [10], [12], [21]. Our approach also performs reasonably well in structure prediction and the results are comparable to those of the existing algorithms. Ours is the first model that actually gives a good insight into the physicochemical model rather than merely classifying or predicting TMB proteins. The results show that our approach is also good at discriminating TMB proteins. The rest of this paper is organized as follows: We present our approach including classification and folding models in section II. The experimental setup is discussed in section III. We present the results in section IV, followed by conclusions in section V.

II. METHOD

We now present the methods developed for classification and structure prediction of TMB proteins\(^1\). TMB proteins are hard to identify, however, it is relatively easy to identify a majority of other proteins which are not TMB. We use physicochemical properties and a simple probabilistic model based on a sliding window for filtering amino acid segments that are obviously not involved in any \( \beta \)-barrel structures as a membrane spanning \( \beta \)-strand. Proteins that are considered to be putative TMB proteins by this initial phase are then further analyzed. Next, we try to fold the given protein, treating it as a TMB protein, using the pseudo-energy minimization model. If the protein cannot be folded into \( \beta \)-barrels according to the energy minimization framework, the protein is rejected and classified as a non-TMB protein.

\(^{1}\)A preliminary version of this work will appear as a short paper in [22].

Before presenting the simple model that we used for filtering the transmembrane \( \beta \)-strands, we discuss some physicochemical constraints that a protein must obey to be a TMB protein. We enforce these constraints in both the filtering and folding steps of our algorithm.

A. Physicochemical Constraints

On the amphipathic \( \beta \)-strand of TMB proteins, the side-chains of amino acids are directed toward the membrane and the channel alternatively. Hydrophilic and polar side-chains orient toward the aqueous interior while hydrophobic ones contact the hydrophobic bilayer [1]. We have used the Kyte-Doolittle scale [23] to evaluate the hydrophobicity \( H(r) \) of each amino acid \( r \). In this scale a higher value represents higher hydrophobicity, and vice versa. A segment \( r_1,...,r_j \) is a potential membrane spanning \( \beta \)-strand if one side is hydrophobic and the other side is hydrophilic. Formally, we define:

\[
H_{i,j}^e = \langle H(r_{2k}) \rangle, i \leq 2k \leq j \\
H_{i,j}^o = \langle H(r_{2k+1}) \rangle, i \leq 2k + 1 \leq j , k \in \mathbb{N}
\]

the average hydrophobicity on the respective even and odd numbered sides. A segment \( r_1,...,r_j \) is a potential membrane spanning \( \beta \)-strand if

\[
\max \{ H_{i,j}^e, H_{i,j}^o \} > \zeta^- \quad \text{and} \quad \min \{ H_{i,j}^e, H_{i,j}^o \} < \zeta^+
\]

where \( \zeta^- \) is a lower bound for the hydrophobic side and \( \zeta^+ \) is an upper bound for the hydrophilic side. We use the values \( \zeta^- = -2 \) and \( \zeta^+ = 2 \), which were obtained through training. Then, with respect to the TMB structure, the segment \( r_i,...,r_j \) is defined as odd inward oriented if \( H_{i,j}^e < H_{i,j}^o \) and odd outward oriented if \( H_{i,j}^e > H_{i,j}^o \).

B. Classification Filtering

In order to identify substrings as potential membrane spanning \( \beta \)-strands (the vertices) or turns/loops (the edges), we introduce a simple probabilistic model that acts as a primary filter. We use a sliding window (segment) as a sequence of consecutive \( l \)-residue subsegments (or blocks) \((l = 3 \text{ in our implementation})\). Let \( r \) denote the occurrence of a given block \((r = r_1r_2...r_l)\) and let \( s \) be the event that a block is found in a given conformation (\( \beta \)-strand of coil). The information that \( s \) gets from \( r \) is defined as:

\[
I(s; r) = \log \frac{P(s|r)}{P(s)} = \log \frac{f_{s|r}}{f_{s}} = \log \frac{f_{s|r}}{f_{s}}
\]

where \( f_{s|s} \) represents the frequency observed in the training dataset for a block \( r \) to be found in conformation \( s \) and we denote briefly \( f_{s|s} = \sum f_{s|s}, \quad f_s = \sum f_{s|s}, \quad f_{s|r} = \sum f_{s|s} \) and \( f_{s} = \sum f_{s} \) [24]. Thus, \( I(s; r) \) measures the influence of \( r \) on the occurrence of \( s \). If \( I(s; r) = 0 \) there is no influence, whereas \( I(s; r) > 0 \) indicates that \( r \) is favorable to the occurrence of \( s \) and vice versa. Formally, the preference of
r in favor of s as opposed to $\pi$, where $\pi$ is any conformation different from s [25] is:

$$I(s : \pi ; r) = I(s ; r) - I(\pi ; r) = \log \frac{f_{s,r}}{f_{s,\pi}}.$$  

A simple measure is associated to each segment $r_1 r_2 ... r_p$ which helps determine if it is likely a $\beta$-strand or a coil. It is defined as the sum of information on all l-residue blocks:

$$I(s : \pi ; r_1 r_2 ... r_p) = \sum_{i=1}^{p-l+1} \left( I(s : \pi ; r_i r_{i+1} ... r_{i+l-1}) - \log \rho \right).$$

The segment is then considered as a candidate for conformation $s$ if $I(s : \pi ; r_1 r_2 ... r_p) > 0$.

A non-redundant training set of TMB proteins (described below) is used to learn this probabilistic model. Due to the small size of training set, we apply the filter with a relatively low threshold at $\rho = \frac{1}{2}$ to avoid overfitting. This ensures that on average, each block $r$ is accepted in conformation $s$ if the propensity for $r$ to be in $s$ $(f_{s,r}/f_{s,\pi})$ is at most 1.5 times less than the propensity to be in $\pi$ $(f_{\pi,r}/f_{\pi,\pi})$.

Only substrings that pass these very stringent criteria are considered to be putative strands in the graph that follows.

Now we present a graph-theoretic energy minimization model for recognizing and folding TMB proteins.

C. Graph-theoretic Pseudo-energy Model

We model the folding problem of TMB proteins with minimum energy as a graph-theoretic optimization problem: finding the longest path $P$ in a weighted graph $G(V, E, E_{\text{intr}}, E_{\text{adj}}, E_{\text{loop}})$, with respect to a given permutation $\sigma$, i.e. the vertices of $P$ located on a circle are permuted according to $\sigma$ (see Figure 1). We first describe how the graph is built given a protein sequence.

Graph

Each vertex of $V$ represents a potential $\beta$-strand in a given configuration. It corresponds to a particular layout of a substring of the amino acid sequence that satisfies given conformational constraints. An edge in $E$ corresponds to a turn or a loop that connects two consecutive $\beta$-strands. Between the two $\beta$-strands connected by an edge, there is a substring of amino acids validated by the constraints described previously. For every vertex $v \in V$, $E_{\text{intr}}(v)$ represents the intrinsic energy of $v$ (strand in a given layout). This includes the internal energy of the substructure, i.e. the interactions between its own amino acids, and the interaction energy with the environment (membrane and channel) apart from the rest of the considered protein. For every pair of vertices $(v, w) \in V \times V$, $E_{\text{adj}}(v, w)$ is the interaction energy of the pair $(v, w)$. This takes into account the number of contacts and different side-chain interactions as packing of hydrophobic cores and bonding abilities. For every edge $(v, w) \in E$, $E_{\text{loop}}(v, w)$ is related to the intrinsic energy of the turn/loop between the strands $v$ and $w$. The precise energy calculation are described below in Implementation Details.

Optimization Objective

Given this graph, the optimal protein structure can be determined by maximizing the following objective function:

$$E = \sum_{v \in P} E_{\text{intr}}(v) + \sum_{(v, w) \in \sigma(P)} E_{\text{adj}}(v, w) + \sum_{(v, w) \in P} E_{\text{loop}}(v, w).$$

It is possible to compute the optimum in $O(n^2)$ running time for structures corresponding to the identity permutation and at most $O(n^3)$ for structures containing Greek key motifs, where $n$ is the input sequence length. We skip the algorithmic details due to space constraints.

Implementation Details

The number of strands $n$ determines the geometry of the barrel, particularly the membrane spanning part of the segments, and is involved in the computation of energy terms. If known, the algorithm can enforce this value and fold the protein accordingly. The values for $n$, which are usually even, are governed by the consideration on the length of the sequence, the thickness of membrane and the length of turns or loops and vary between 8 and 22 [1].

Side-chain interactions between contiguous residues along a segment on the same side and interactions with the environment of channel or bilayer define the intrinsic energy of the corresponding vertex. The pairing energy of two adjacent segments in the barrel is computed by optimizing the relative positions between constituent amino acids. These energies involve hydrogen bonds in main chains, electrostatic interactions between side-chains, hydrophobic effect as well as environmental effect. More specifically, the extracellular and intracellular environments with distinct hydrophobicity indices can have significantly different hydrophobic effects. In addition, the membrane thickness gives constraints on segment size and helps identify the interactions inside or outside the membrane region. The features on size, polarity and flexibility of turns and loops are taken into consideration, their energies are approximated by hydrophobicity.

We use the Dunbrack backbone-dependent rotamer library [26] and the partial charges from GROMOS force field [27] to compute pairwise interaction energies. The hydrophobic interaction between two side-chains $u, v$ is assessed by the amount of contacts between non-polar groups, calculated by taking the average on all rotamer pairs of the two side-chains $e_{uv} = < e_{uv[\text{rotamers}]>$. Each side-chain plays a role of a group of partial charges in the electrostatic interaction. The main-chain hydrogen bond is measured by the electrostatic potential energy between peptide CO and NH groups.

The probabilistic model and the constraints on hydrophobicity help discard the unlikely membrane spanning $\beta$-strands. A threshold on overall energy can also be involved to enhance the discrimination. We studied the per-strand
energy value for a variety of TMB proteins including the training dataset and other TMB proteins. Even though this value is always higher than 0.9 for these proteins, we chose 0.85 as a threshold to avoid overfitting. Note that this does not affect the prediction results, and is only used for classification.

III. EXPERIMENTAL SETUP

A. Software

We compare our folding prediction accuracy to TMBpro [9] and TMBETAPRED-RBF [10]. We compare our classification results to Freeman et al. [6], TMBETAPRED-RBF [10], PRED-TMBB [18] and transFold [12]. TMBpro and TMBETAPRED-RBF results are executed from their web-server.

B. Datasets

We used TMB proteins from the PDBTM database [20] to train and test our approaches.

Folding

We used CD-HIT [28] to constrain the redundancy in proteins. A threshold of 40% similarity was used to reduce the dataset, resulting in 49 sequences (PDBTM40). We retain only the monomeric barrels, i.e. the sequences that form a unique complete barrel. Thus, PDBTM40 contains 42 sequences 1OH2_Q, 3A2R_X, 3AEH_A, 3BRZ_A, 3CSL_A, 2R4P_A, 3DWO_X, 2FGQ_X, 3EFM_A, 3EMN_X, 2ERV_A, 2IWW_A, 2F1T_A, 1FEP_A, 3FHJ_A, 3FID_A, 1ILZ_A, 1BY3_A, 2GSK_A, 1BH3_A, 2HDZ_A, 2J1N_A, 2IAH_A, 3JTY_A, 1BXW_A, 2VDF_A, 1PNZ_A, 3KYN_X, 3GP6_A, 1AF6_A, 3NJT_A, 2O4V_A, 2ODJ_A, 1QJ8_A, 1P4T_A, 2POR_A, 1TLW_A, 1UXF_A, 1UYN_X, 2WJQ_A, 2X4M_A, 1XKW_A. It is important to note that both TMBPro and our method use the entire dataset to train. While this may result in overfitting for a learning-based approach, the effect on our approach should be very small.

Classification

We used a set of 177 α-helical transmembrane proteins of length from 140 to 800 residues, at 40% redundancy reduction, from PDBTM and 32 non-redundant lipocalins taken from PDB.

IV. RESULTS

A. Folding

The folding prediction results are presented in Table I and Figure 2. Figure 2 plots the Matthews Correlation Coefficient for our approach and TMBpro for different proteins along the x-axis. Our results are comparable to those of TMBpro but more consistent as we do not rely on training for folding. Especially, for 3CSL, a 753-residue-long sequence composed of 22 β-strands, where the N-terminal creates another domain that plugs into the channel, it can be observed that our MCC is much better than in TMBpro.

The TMBETAPRED-RBF web-server predicted non-TMB for 25 over 42 proteins of PDBTM40, or 59.5%. However, the structures for correctly identified proteins were completely accurate, we suspect this might be because they were included in the training set.

B. Permuted Structure

For 3L48, the C-terminal domain of the PapC usher in E. coli, the observed structure topology containing a Greek key motif corresponds to the permutation $\sigma = (1, 4, 3, 2, 5, 6, 7)$ and is predicted with $Q_2$ of 70.2% at $\rho = 0.2$.

Following the experimental observations that were published previously on the efficiency of the in vivo membrane assembly of OmpA variants [29], we test our algorithm with different given permutations. OmpA (1BXW) consists of eight β-strands, thus without feasibility taken into account, there are $(8-1)! = 5040$ circular permutations to verify (see Figure 3). The pseudo-energy 10.21 of the observed permutation is found in the lowest energy zone. 41 permuted structures, or 0.81%, reach an energy of $10.21 \pm 0.3$. A ratio of about 1.31% is found in the case of OmpX 1QJ8. These results are not surprising since a protein may be folded into more than one spatial conformation. In both cases, a Poisson-like distribution is found. This observation may help to discriminate most of infeasible conformations with the use of a threshold on global energy. Hence, the method is expected to rapidly find a small set containing the right structure within a threshold of, for instance, 2% from the lowest energy and with structural feasibility conditions on permutations. Other proposed solutions in this set may be the candidates for in vivo and in vitro studies.

C. Classification

100% of the non-redundant set of 177 α-helical transmembrane proteins of length from 140 to 800 residues in PDBTM are rejected, whereas 31 out of 32 non-redundant lipocalins taken from PDB are predicted as non-TMB. Though lipocalins are also β-barrels which reverse the TMB

\[ \text{domain that plugs into the channel, it can be observed that our MCC is much better than in TMBpro.} \]

\[ \text{The TMBETAPRED-RBF web-server predicted non-TMB for 25 over 42 proteins of PDBTM40, or 59.5%.} \]

\[ \text{However, the structures for correctly identified proteins were completely accurate, we suspect this might be because they were included in the training set.} \]

\[ \text{B. Permuted Structure} \]

\[ \text{For 3L48, the C-terminal domain of the PapC usher in E. coli, the observed structure topology containing a Greek key motif corresponds to the permutation } \sigma = (1, 4, 3, 2, 5, 6, 7) \text{ and is predicted with } Q_2 \text{ of 70.2% at } \rho = 0.2. \]

\[ \text{Following the experimental observations that were published previously on the efficiency of the in vivo membrane assembly of OmpA variants [29], we test our algorithm with different given permutations. OmpA (1BXW) consists of eight } \beta-\text{strands, thus without feasibility taken into account, there are } (8-1)! = 5040 \text{ circular permutations to verify (see Figure 3). The pseudo-energy 10.21 of the observed permutation is found in the lowest energy zone. 41 permuted structures, or 0.81%, reach an energy of } (10.21 \pm 0.3). \]

\[ \text{A ratio of about 1.31% is found in the case of OmpX 1QJ8. These results are not surprising since a protein may be folded into more than one spatial conformation. In both cases, a Poisson-like distribution is found. This observation may help to discriminate most of infeasible conformations with the use of a threshold on global energy. Hence, the method is expected to rapidly find a small set containing the right structure within a threshold of, for instance, 2% from the lowest energy and with structural feasibility conditions on permutations. Other proposed solutions in this set may be the candidates for in vivo and in vitro studies.} \]

\[ \text{C. Classification} \]

\[ \text{100% of the non-redundant set of 177 } \alpha-\text{helical transmembrane proteins of length from 140 to 800 residues in PDBTM are rejected, whereas 31 out of 32 non-redundant lipocalins taken from PDB are predicted as non-TMB. Though lipocalins are also } \beta-\text{barrels which reverse the TMB}\]
We have presented a new pseudo-energy minimization method for the classification and prediction of transmembrane protein super-secondary structure based on a variety of potential structures. Our approach takes into account many physicochemical constraints and minimizes the free energy. It also accounts for permuted structures, thus giving more complete information on the folded structure. Our method is quite accurate with more than 90% sensitivity and F-score, over 80% M.C.C. score on strands; and over 70% accuracy and F-score on residues. The results are comparable to those given by TMBpro and TMBETAPRED-RBF, which are both learning based methods. Moreover, our results are more consistent and have a significantly less variation across different TMB proteins. This is especially interesting given that our algorithm is based mainly on pseudo-energy minimizations, and the probabilistic model only plays the role of a filter for potential β-strands. While the model presented here is only for TMB proteins, it can be easily extended to accommodate α-helical bundles. We did not use a more sophisticated model for classifying β-barrel strands because that would risk overfitting and reliance on the training dataset. It is also interesting to note our approach performs very well for identification of TMB proteins, rejecting all the α-helical bundles. While the Freeman and Wimley [6] approach may be more accurate on some datasets, it risks overfitting and does not predict the structure. Therefore, our approach provides the best classification results amongst the methods that try to predict structures. Our model does learn the probabilistic model from training dataset, but it is mainly to screen out obvious non-TMB strands. Therefore, there are no concerns about the size of the training data or overfitting.

Even though the results presented in this paper are comparable to other methods, the methodology presented here is novel and gives insight into the actual physicochemical constraints and energy. Moreover, our approach should be able to predict TMB proteins which are significantly different from known proteins. Finally, our approach provides more information than the current approaches by providing the permutations of the strands.

Future Work

We are working on energy models for TM α-helical bundles and β-barrels with broken strands, as well as globular β-barrels like lipocalins or membrane targeting proteins (C2 domain) where permuted structures are usually found. Nevertheless, similar to the other methods, we only propose single-domain protein structures.

We are also currently working on refinements in structural constraints and hydrophobicity, which may help to improve the accuracy of our predicted structure. Finally, it will be interesting to investigate more sophisticated statistical models for the initial screening, both to improve the results and understand how effective a mixed approach can be.

ACKNOWLEDGMENT

The authors would like to thank all the INRIA AMIB Team members, especially Mireille Régnier, Yann Ponty, Julie Bernauer and Balaji Raman.