Tran et al. BMC Genomics 2012, 13(Suppl 2):S5
http://www.biomedcentral.com/1471-2164/13/S2/S5

BMC Genomics

RESEARCH Open Access

A graph-theoretic approach for classification and structure prediction of transmembrane β-barrel proteins

Van Du T Tran*, Philippe Chassignet, Saad Sheikh, Jean-Marc Steyaert

From First IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS 2011)
Orlando, FL, USA. 3-5 February 2011

Abstract

Background: Transmembrane β-barrel 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 transmembrane β-barrel 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 transmembrane β-barrel proteins. Most of these methods emphasize on homology search rather than any biological or chemical basis.

Results: We present a novel graph-theoretic model for classification and structure prediction of transmembrane β-barrel 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 41 proteins gathered from existing databases on experimentally validated transmembrane β-barrel 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.

Conclusions: We show that it is possible to design models for classification and structure prediction for transmembrane β-barrel proteins which do not depend essentially on training sets but on combinatorial properties of the structures to be proved. These models are fairly accurate, robust and can be run very efficiently on PC-like computers. Such models are useful for the genome screening.

Background

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 (TMB). 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

* Correspondence: vandu@lix.polytechnique.fr
INRIA AMIB Team, Laboratory of Computer Science (LIX), Ecole Polytechnique, 91128, Palaiseau CEDEX, France

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.

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 tiny. Thus, it is arguable whether these approaches can 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 methods allow for permutations in protein structures. The TMB structures are not merely a series of β-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 [PDB:3L48]. This level of structure may be described as a permutation on the order of the bonded strands.

In this paper, we present a novel *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 well in structure prediction and the results are comparable to those of the existing algorithms. Ours is the first model that actually gives an 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.

Results and discussion

Folding

The folding prediction results are presented in Table 1 and Figure 1. Figure 1 plots the Matthews Correlation Coefficient for our approach BBP (Beta-Barrel Predictor) and TMBpro for different proteins along the x-axis. The results of our approach are comparable to those of TMBpro but more consistent as we do not rely on training for folding. We note that, in the cases the program predicts an optimal structure with a wrong number of strands, the optimal energy is really close to the energy of the topologically right structure.

The TMBETAPRED-RBF web-server predicted non-TMB for 24 over 41 proteins of PDBTM40, or 58.5%. The structures for correctly identified proteins were completely accurate. This might be because they were included in the training set.

Evaluation of shear numbers

We study the energy distribution of 17 TMB structures (ECOLI40) in E. coli taken from PDBTM40 (including [PDB: 1AF6_A, 1BXW_A, 1BY3_A, 1FEP_A, 1ILZ_A, 1PNZ_A, 1QJ8_A, 1TLW_A, 2F1T_A, 2GSK_A, 2HDF_A, 2IWW_A, 2J1N_A, 2R4P_A, 2WJQ_A, 3AEH_A, 3GP6_A]) with regards to the slant angle, hence the shear number (see Figure 2). Most optimal structures incline with an angle of 41° - 49°, as observed in databases. This suggests that our model performs well the physicochemical properties of TMB structures. It should be also noted that there is no natural way to define the shear number *a priori*.

Influence of the filtering threshold

We apply the filtering thresholds $\rho = \frac{1}{3}, \frac{1}{2}$ and $\frac{2}{3}$ on ECOLI40. These thresholds ensure that on average, considering 3-residue blocks as subunits, each segment is accepted as a β-strand if its propensity to be β-strand is at most 3, 2, 1.5 times, respectively, less than its propensity to be other structure (α-helices or turns/loops). The observed minor difference in accuracy with such considerably distinguished thresholds reinforces the fair independence of our approach from the training data. The results in Table 2 show the strong predicting ability of BBP from a poor known database. The lower the parameter ρ, the more independent to the training the predictor. This reduced the prediction performance of the

Table 1 Comparison of prediction accuracy on PDBTM40

		Residues				Strands			
Method	**Q_2**	**Specificity**	**Sensitivity**	**F-score**	**MCC**	**Specificity**	**Sensitivity**	**F-score**	**MCC**
TMBpro	81.2 ± 6.1*	79.3 ± 7.9	84.2 ± 11.2	0.76 ± 0.1	0.61 ± 0.14	90.1 ± 15.0	94.2 ± 12.5	0.93 ± 0.12	0.85 ± 0.26
BBP	79.2 ± 5.4	78.4 ± 6.3	80.4 ± 9.9	0.74 ± 0.1	0.57 ± 0.12	91.4 ± 12.0	91.4 ± 11.3	0.92 ± 0.11	0.83 ± 0.22

*Standard Deviation

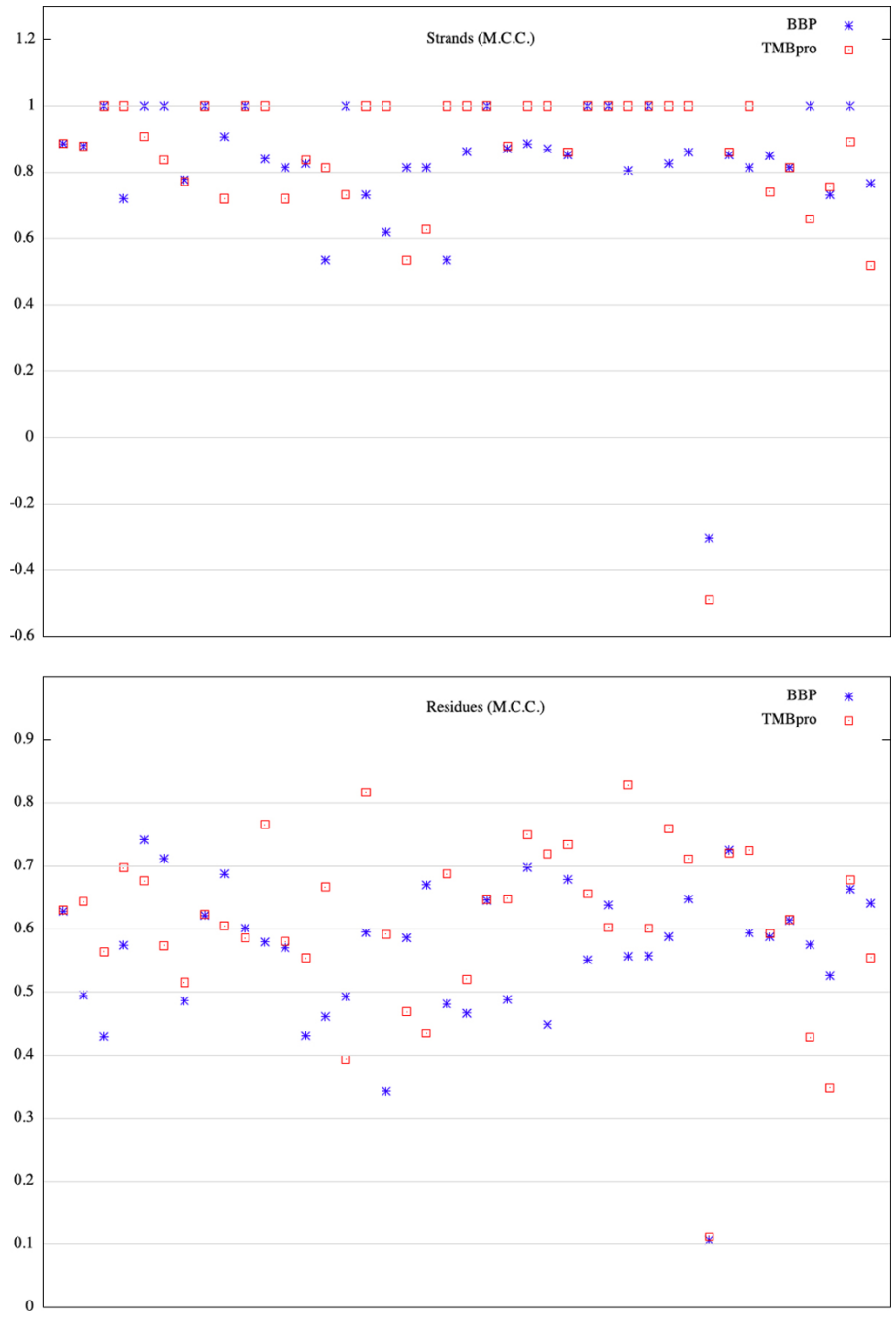

Figure 1 Comparison of BBP and TMBpro on structure prediction results.

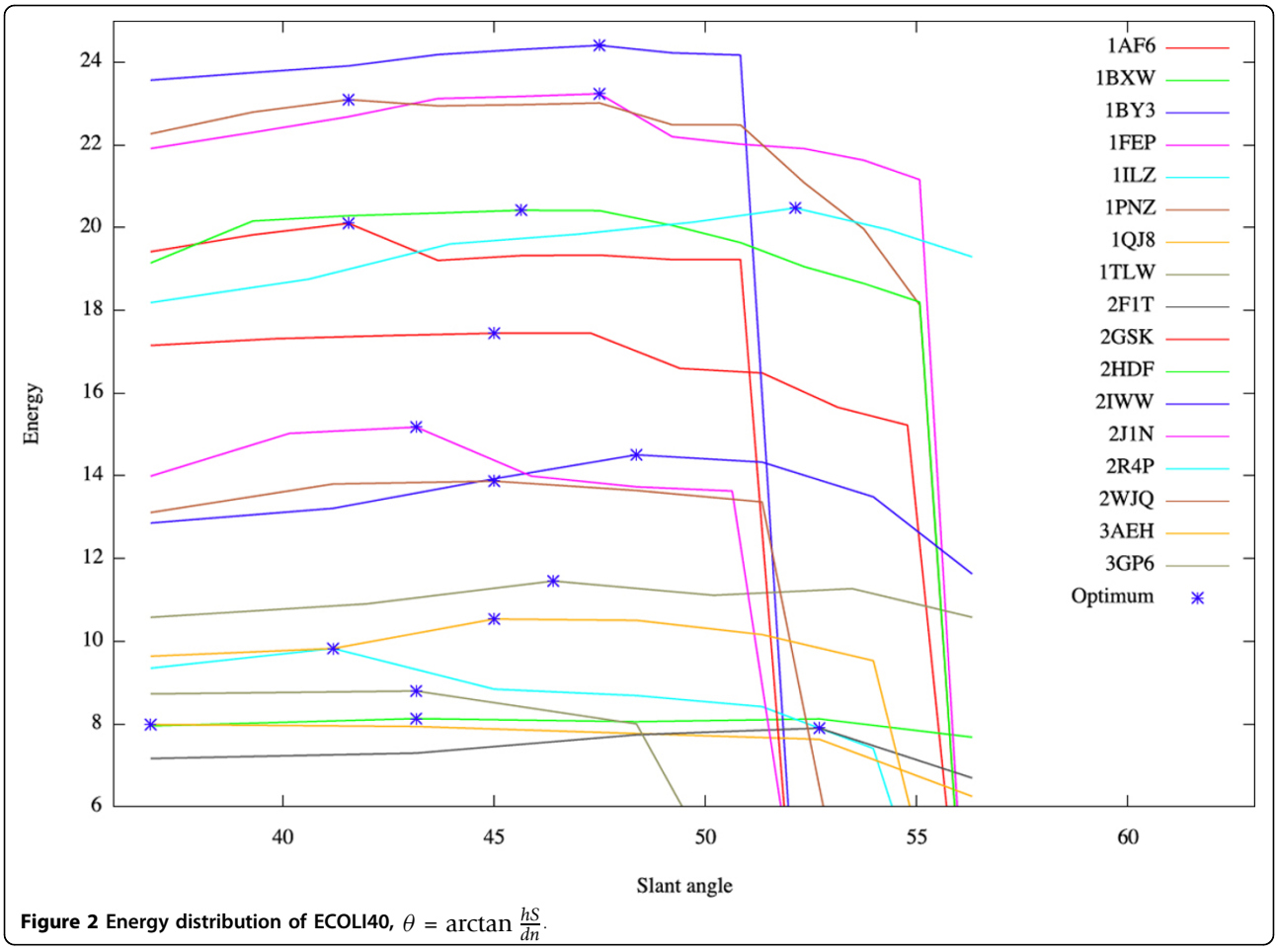

Figure 2 Energy distribution of ECOLI40, $\theta = \arctan \frac{hS}{dn}$.

model on the known structures, however, it may be useful to discover new TMB proteins.

Evaluation on mutated sequences

We generate the mutated sequences from ECOLI40 by substituting the amino acids at turns or loops using the PAM250 substitution matrix [22]. Each sequence in ECOLI40 is mutated up to 5% of amino acids into 10 new sequences. Figures 3 and 4 show the Matthews Correlation Coefficient and F-score for residues and β-strands. We observe from these results the stability of our predictions. It also suggests that the TMB proteins are stable against these mutations at their turns and loops. The difference in structures of those mutated proteins may merely come from the shift of membrane spanning β-strands when their two extremities are mutated.

Permuted structures

For [PDB: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 σ = (1, 4, 3, 2, 5, 6, 7) and is predicted with an accuracy (Q_2) of 70.2% at ρ = 0.2.

Following the experimental observations that were published previously on the efficiency of the *in vivo* membrane assembly of OmpA variants [23], we test our algorithm with different given permutations. OmpA

Table 2 Comparison of prediction accuracy on ECOLI40 with different thresholds

		Residues				Strands			
ρ	Q_2	Specificity	Sensitivity	F-score	MCC	Specificity	Sensitivity	F-score	MCC
2/3	80.9 ± 4.8*	80.4 ± 5.2	82.7 ± 8.4	0.77 ± 0.04	0.61 ± 0.08	94.8 ± 5.7	93.3 ± 5.9	0.94 ± 0.05	0.88 ± 0.1
1/2	79.7 ± 6.0	78.5 ± 5.1	82.4 ± 8.6	0.76 ± 0.05	0.58 ± 0.11	96.1 ± 4.8	95.4 ± 5.3	0.96 ± 0.05	0.91 ± 0.09
1/3	77.7 ± 5.6	75.6 ± 6.5	81.1 ± 8.6	0.74 ± 0.05	0.55 ± 0.11	91.7 ± 9.2	94.9 ± 6.5	0.94 ± 0.07	0.87 ± 0.07

*Standard Deviation

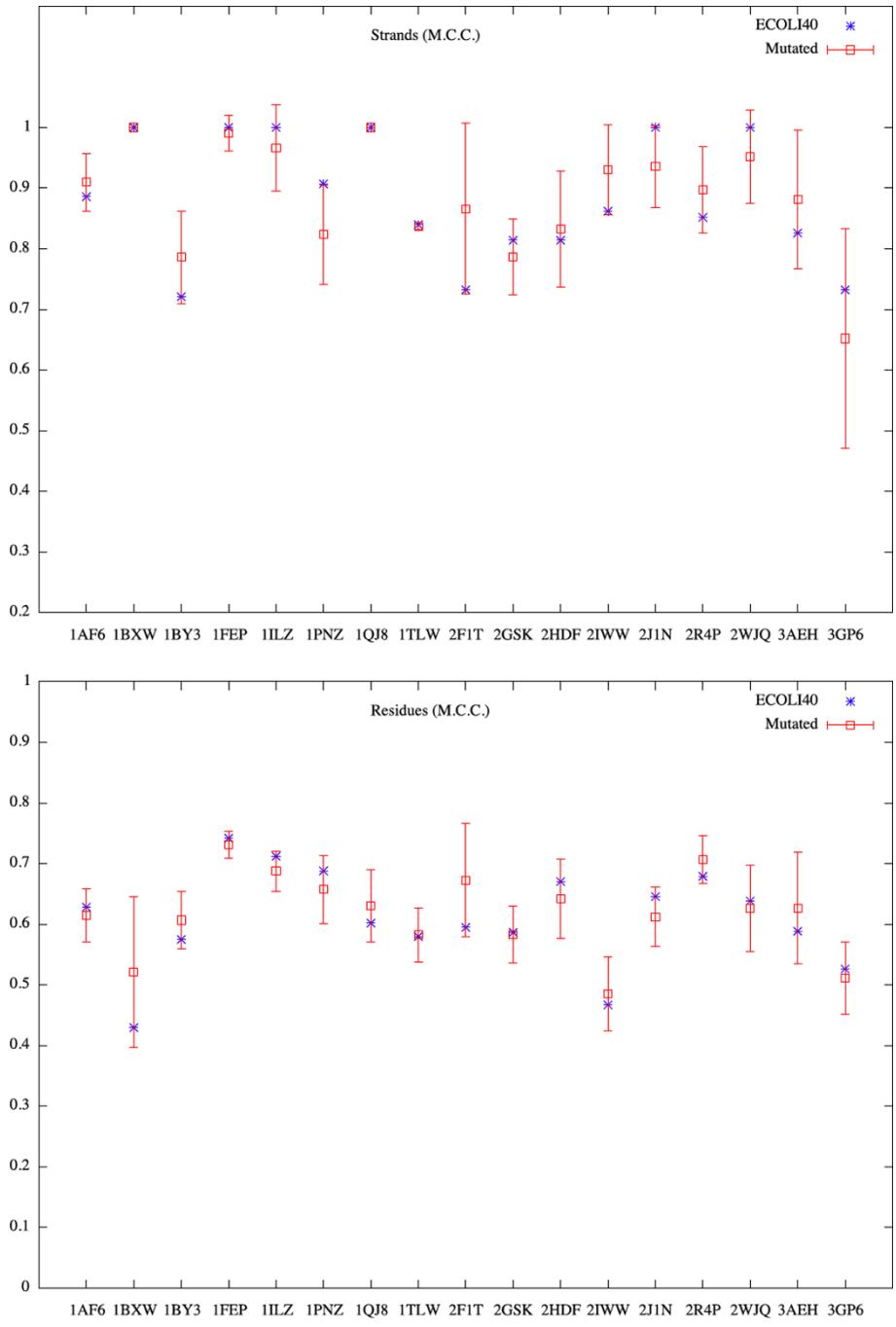

Figure 3 MCC of mutated ECOLI40.

[PDB:1BXW] consists of eight β-strands, thus without feasibility being taken into account, there are (8-1)! = 5040 circular permutations to check (see Figure 5). 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 ± 0.3). A ratio of about 1.31% is found in the case of OmpX [PDB:1QJ8] (see Figure 6). 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 the 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

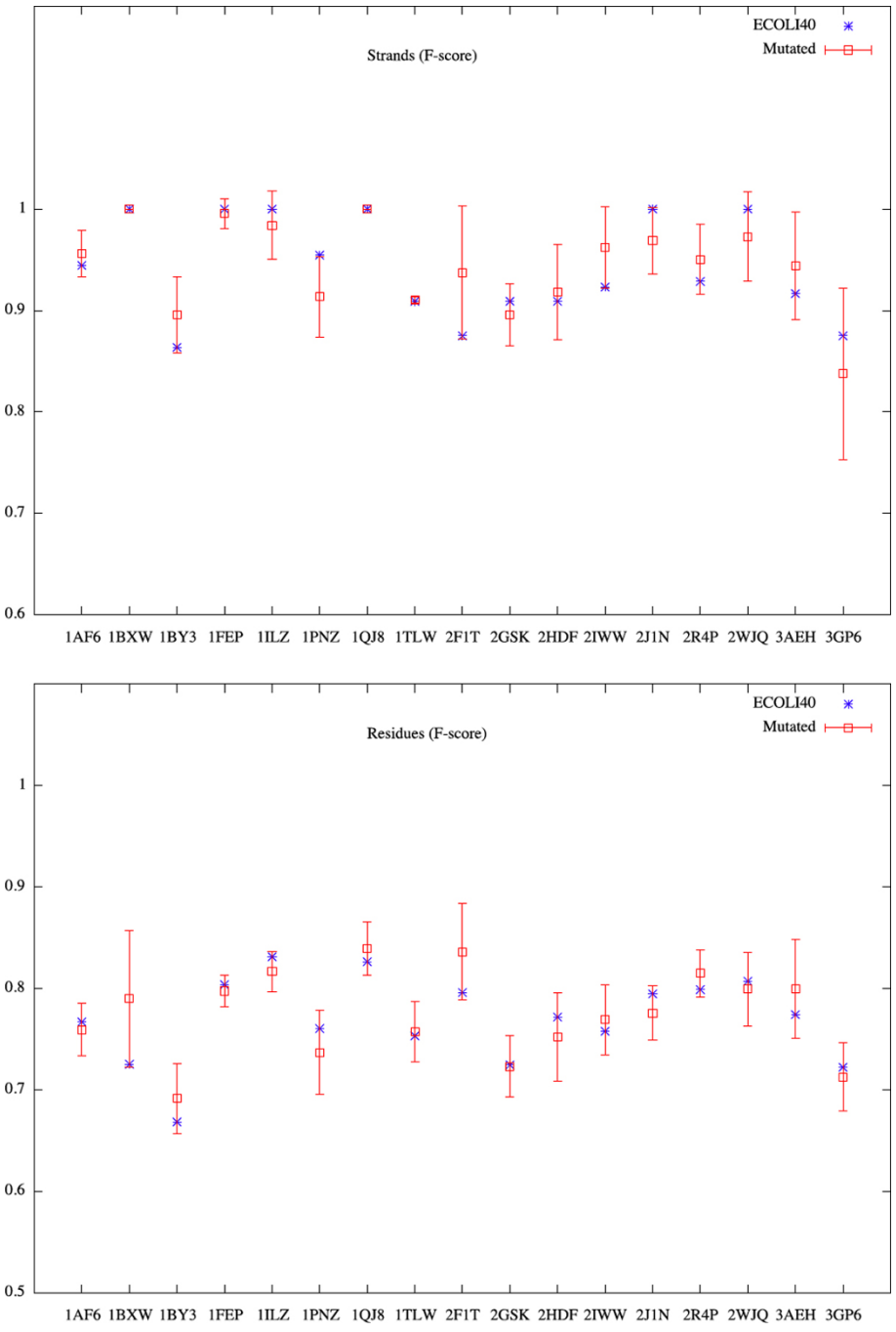

Figure 4 F-score of mutated ECOLI40.

conditions on permutations. This set might be much smaller be refining the biologically plausible permutations. Other proposed solutions in this set may be the candidates for *in vivo* and *in vitro* studies.

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 (the dataset is available at [24]). Though lipocalins are also β-barrels which reverse the TMB pattern with a hydrophobic core, the environmental effects on both sides of the barrel are still different. Our pseudo-energy model yields unfavorably on such structures and discriminates considerably better than the learning-

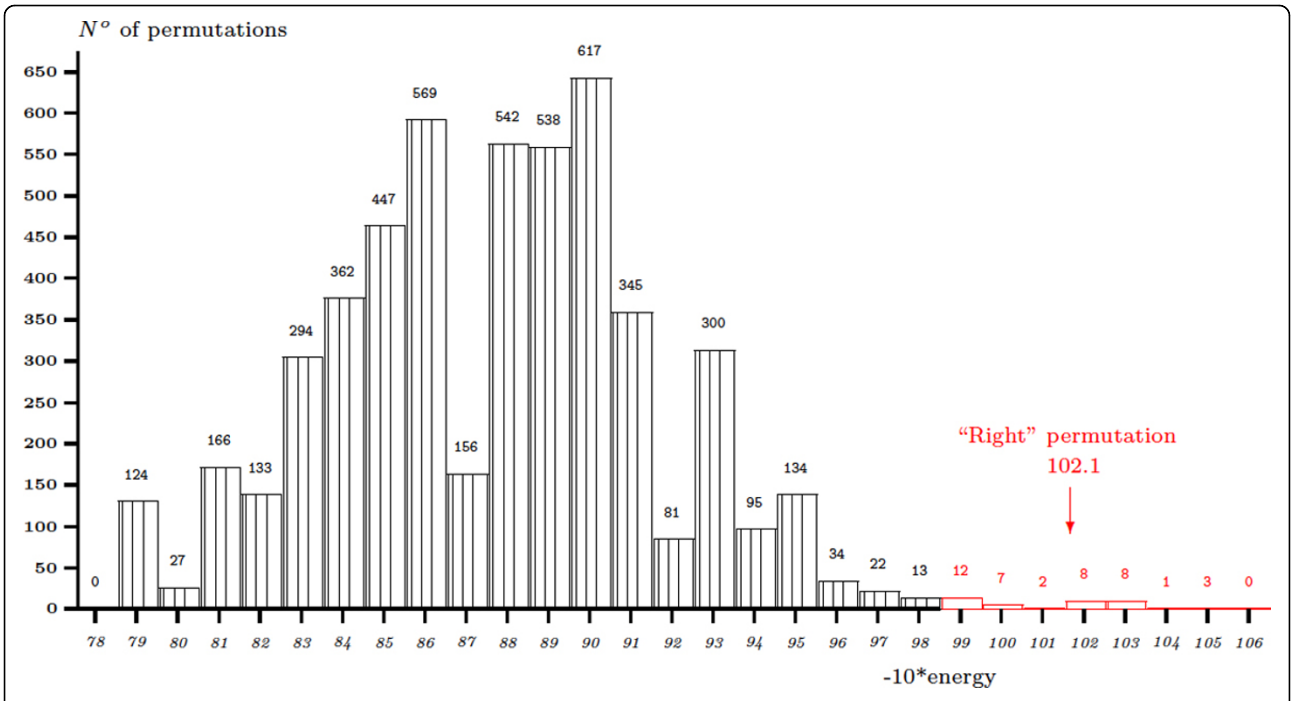

Figure 5 Predicting for 7! permutations on E. Coli OmpA [PDB:1BXW] 8-strand barrel.

based methods like Freeman-Wimley [6], TMBpro [9], PRED-TMBB [18] and TMBETAPRED-RBF [10], but also of transFold [12].

Conclusions

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.

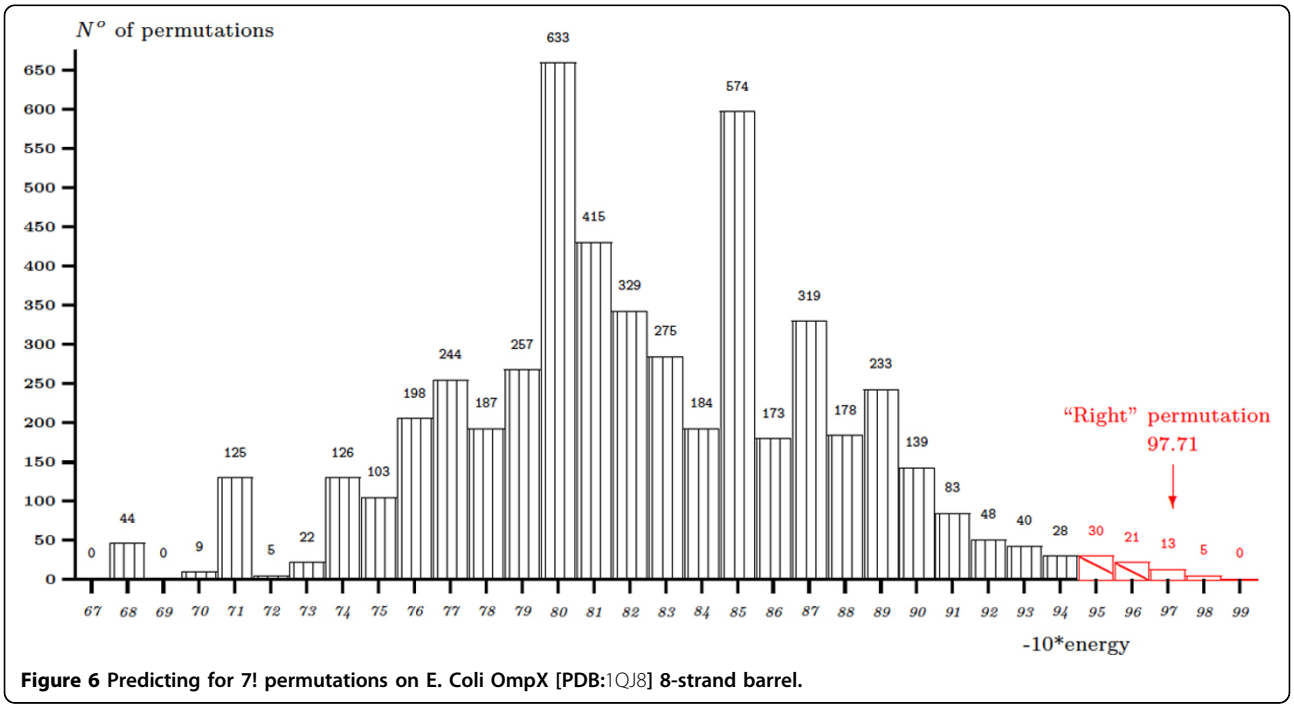

Figure 6 Predicting for 7! permutations on E. Coli OmpX [PDB:1QJ8] 8-strand barrel.

score on strands; and over 74% 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 a very small role. 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 statistical model for classifying β-barrel strands because that would risk overfitting and reliance on the training dataset. It is also interesting to note that our approach performs very well for identification of TMB proteins, rejecting all the α-helical bundles. The Freeman and Wimley [6] approach is more accurate on some datasets. However, it risks overfitting and does not predict the structure. Therefore, our approach provides the best overall 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.

Methods

We now present the methods developed for classification and structure prediction of TMB proteins (a preliminary version of this work appeared as a short paper in [25,26]). 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 β-barrel structures as a membrane spanning β-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 β-barrels according to the energy minimization framework, the protein is rejected and classified as a non-TMB protein.

Before presenting the simple model that we used for filtering the transmembrane β-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.

Geometric framework for β-barrels

For a regular β -barrel [27-29], the backbone geometry is entirely determined by n, the number of strands composing the barrel, and by S, the shear number, which is defined below.

Definition 1 Shear number of a β-barrel *In a regular β-barrel, the shear number S is unambiguously defined as the ordinal distance between an amino acid A and an amino acid B that is located on the same strand as A and linked to A through a path of hydrogen bonds. B is the projection of the "copy" of A after one turn on the first strand of the barrel.*

Structural constants are h ($\approx$ 3.3Å), the jump per amino acid along a strand, and d ($\approx$ 4.4Å), the mean distance between adjacent strands, given respectively by the peptide bond and hydrogen bond geometries. The other geometric characteristics, such as θ, the slant angle of the strands relative to the z barrel axis, are given from n, S, h and d [30]:

$$\tan\theta = \frac{hS}{dn}$$

Angle θ, in association with a given membrane thickness, is involved in the energetic rules and restricts the membrane spanning β-strand length. Then, n and S have to be fixed as parameters.

Definition 2 Relative shear number *Given a shear number S, the* relative shears *between adjacent strands remain as n - 1 degrees of freedom. As a convention, we consider the* relative shears *on the extracellular side of the barrel. So, $\forall i > 1$, s_i, the* relative shear *of strand i + 1 with respect to strand i (strand n + 1 being identified with 1), is measured on strand i as the ordinal distance*

between the undermost amino acid of strand i and the one that is directly bound to the undermost amino acid of strand i + 1.

On the example of Figure 7, the sequence of *relative shears* (s_i) is (1 1 1 2 1 1 1 2). The sum of consecutive *relative shears* naturally defines the *shear* between two extreme strands, thus we have the constraint for the β-barrel, where the two extreme strands are strand 1, for instance, and itself after a round on the barrel:

$$\sum_{1 \le i \le n} s_i = S$$

We define the shear number, by extension, for the case of a β-sheet (i.e. an open β-barrel) to make our algorithms capable of dealing with the structure of β-sheets.

Definition 3 Shear number of a β-sheet *The shear number of a n-strand β-sheet is defined as the sum of relative shears on consecutive pairs of adjacent strands:*

$$S = \sum_{1 \le i \le n-1} s_i$$

where s_i is the relative shear of strand i + 1 with regard to strand i.

Each β-strand is directed with respect to the sequence order from N-terminal to C-terminal. A strand is said to be *upward* if it is oriented from the extracellular environment to the periplasmic space, i.e. the N-terminal of the strand is located on the extracellular side and its C-terminal is on the periplasmic side. Inversely, the strand is said to be *downward*. The *upward*/*downward* orientation of the strand, relatively to the barrel axis, defines another degree of freedom.

Finally, considering a β-strand as a ribbon where the amino acids direct their side-chains alternatively on both sides, toward the barrel interior (channel) or toward the surrounding lipid (membrane), we will distinguish two ways of facing, neglecting small swivel adjustments. A strand is said to be *odd inward* if the odd indexed amino acids face to the channel and *odd outward* if those face to the membrane. We have one more degree of freedom.

Physicochemical constraints. On the amphipathic β-strand of TMB proteins, the side-chains of amino acids are directed towards the membrane and the channel alternatively. Hydrophilic and polar side-chains orient towards the aqueous interior while hydrophobic ones contact the hydrophobic bilayer [1]. We use the Kyte-Doolittle scale [31] to measure the hydrophobicity $H(r)$ of each amino acid r. In this scale, a higher value represents higher hydrophobicity, and vice versa. The necessary condition for a segment r_i .... r_j to be a potential membrane spanning β-strand is that one side is hydrophobic and the other side is hydrophilic. Formally, we define

$$H^e_{i,j} = \langle H(r_{2k}) \rangle, i \le 2k \le j$$
$$H^o_{i,j} = \langle H(r_{2k+1}) \rangle, i \le 2k+1 \le j, k \in \mathbb{N}$$

as the average hydrophobicity on the respective even and odd numbered sides. Hence, the constraints

$$\max\{H^e_{i,j}, H^o_{i,j}\} > \zeta^- \text{ and } \min\{H^e_{i,j}, H^o_{i,j}\} < \zeta^+$$

are necessary for a segment of j - i + 1 consecutive amino acids r_i .... r_j to be a potential membrane spanning β-strand, where ζ^- is a lower bound for the hydrophobic side and ζ^+ is an upper bound for the hydrophilic side. We use the values ζ^- = -1 and ζ^+ = 1, which were obtained through an statistical data analysis

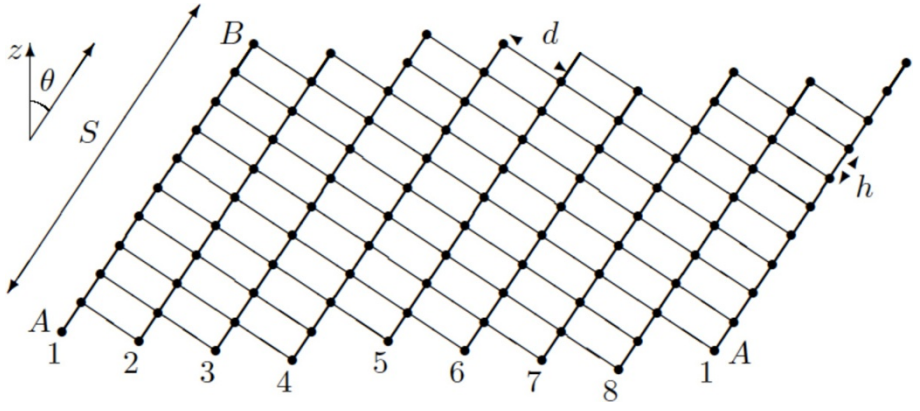

Figure 7 The schematic planar view of a 8-β -strand barrel (strand 1 is duplicated for clarity). Thick lines represent the peptide bonds between consecutive amino acids along their strand. Thin lines represent the hydrogen bonds between the amino acids in adjacent strands. In this example, the *shear number* is S = 10, which is the ordinal distance between amino acids *A* and *B*. We note that all known β -barrels have a positive *shear number* [43] and are slanted "to the right", as illustrated here.

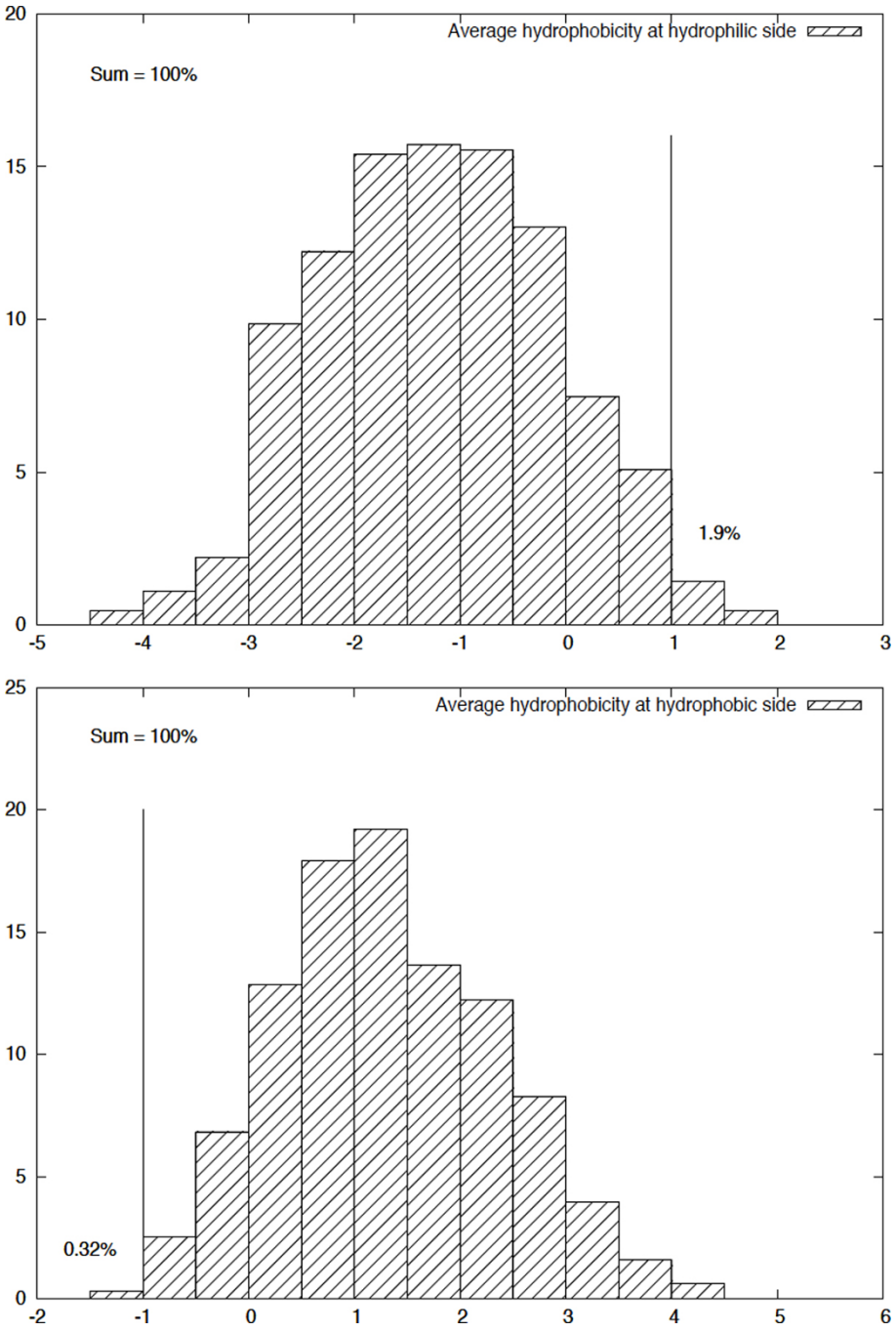

Figure 8 The distribution of average hydrophobicity index of the hydrophilic and hydrophobic side of the membrane spanning β -strands from PDBTM40.

on known TMB structures (see Figure 8). Then, with respect to the TMB structure, the segment $r_i....r_j$ is defined as *odd inward* oriented if $H^o_{i,j} < H^e_{i,j}$ and *odd outward* oriented if $H^e_{i,j} < H^0_{i,j}$.

Classification filtering. In order to identify substrings as potential membrane spanning β-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 in our implementation). Let r denote the occurrence of a given block ($r = r_1 r_2 \ldots r_l$) and let τ be the event that a block

is found in a given conformation (β-strand or turn/loop). The information that τ gets from r is defined as:

$$I(\tau; r) = \log \frac{P(\tau|r)}{P(\tau)} = \log \frac{f_{\tau,r}/f_{\cdot,r}}{f_{\tau,\cdot}/f_{\cdot,\cdot}},$$

where $f_{\tau,r}$ represents the frequency observed in the training dataset for a block r to be found in conformation τ and we denote for short [32]:

$$f_{\cdot,r} = \sum_{\tau} f_{\tau,r},\ f_{\tau,\cdot} = \sum_{r} f_{\tau,r},\ f_{\cdot,\cdot} = \sum_{\tau}\sum_{r} f_{\tau,r}$$

Thus, $I(\tau\ ;\ r)$ measures the influence of r on the occurrence of τ. If $I(\tau\ ;\ r) = 0$, there is no influence; whereas $I(\tau\ ;\ r) > 0$ indicates that r is favorable to the occurrence of τ and vice versa. Formally, the preference of r in favor of τ as opposed to $\bar{\tau}$, any conformation different from τ [33], is:

$$I(\tau : \bar{\tau};\ r) = I(\tau;\ r) - I(\bar{\tau};\ r) = \log \frac{f_{\tau,r}/f_{\tau,\cdot}}{f_{\bar{\tau},r}/f_{\bar{\tau},\cdot}}$$

A simple measure is associated to each segment $r_1\ r_2\ \ldots\ r_p$ that helps determine if it is likely a β-strand or a turn/loop. It is defined as the sum of informations on all the l-residue blocks:

$$\tilde{I}(\tau : \bar{\tau}; r_1\ r_2 \ldots r_p) = \sum_{i=1}^{p-l+1} \frac{I(\tau : \bar{\tau};\ r_i\ r_{i+1} \ldots r_{i+l-1}) - \log\rho}{p-l+1}$$

The segment is then considered as a candidate for conformation τ if $\tilde{I}(\tau\ :\ \bar{\tau}; r_1\ r_2 \ldots r_p) > 0$.

The non-redundant training set PDBTM40 of 41 TMB proteins is used to learn this probabilistic model. Due to the small size of the training set, we apply the filter with a relatively low threshold at $\rho = \frac{2}{3}$ to avoid overfitting. This ensures that on average, each block r is accepted in conformation τ if the propensity for τ to be in τ (i.e. $f_{\tau,r}/f_{\tau,\cdot}$) is at most 1.5 times less than the propensity to be in $\bar{\tau}$ (i.e. $f_{\bar{\tau},r}/f_{\bar{\tau},\cdot}$). Only substrings that pass these very stringent criteria are considered to be putative strands.

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

Definition of the graph structure

Dynamic programming approach. Let $\mathcal{S}$ be the sequence of the N amino acids constituting the primary structure of a given protein. We will consider $\mathbf{G}(\mathbf{V}, \mathbf{E}, \mathcal{E}_{\text{intr}}, \mathcal{E}_{\text{adj}}, \mathcal{E}_{\text{loop}})$, the weighted directed acyclic graph (DAG) [34] built from S as follows: **Vertices** Let $\mathbf{V} = \mathbf{V}^* \cup \{\top, \bot\}$ be the set of vertices. Each vertex of $\mathbf{V}^*$ represents a candidate secondary structure item as a β-strand associated with a given set of parameters. It corresponds to a contiguous part (a substring, defined by its starting and ending indices $1 \le \nu < k \le N$) of $\mathcal{S}$ that satisfies given conformational constraints (such as length, propensity to be a β-strand, . . .). The associated parameters provide information about the discretized spatial laying of this part relatively to the whole structure. So, combining the *upward*/*downward* and *inward*/*outward* degrees of freedom previously introduced, we consider 4 different orientations for each given candidate β-strand. We could also consider the different instances of *relative shear* to multiply the number of vertices, but we do not for reasons to be clarified later. A canonical order is defined on $\mathbf{V}^*$ as the lexicographic order on tuples formed by the respective starting/ending indices in $\mathcal{S}$ and the associated parameters. The length constraint implies that the number of candidate substrings and thus $|\mathbf{V}|$, the number of vertices, are bounded above by kN for a small value k. To simplify further definitions, a dummy vertex $\top$ will be used to represent an empty substring at the start of $\mathcal{S}$ and, similarly, $\bot$ will represent an empty substring at the end of the sequence. To extend the order on all of the vertices, we set $\top < \nu < \bot$, $\forall \nu \in \mathbf{V}^*$ (see Figure 9).

Edges

Let $\mathbf{E} \subset \mathbf{V} \times \mathbf{V}$ be the set of directed edges. Intuitively, an edge corresponds to a turn or a loop that connects two consecutive β-strands. To be more precise, $\forall \nu, w \in \mathbf{V}^*$, with ν_ν, κ_ν, ν_w, κ_w denoting their respective starting and ending indices, (ν, w) is an edge, if $\kappa_\nu < \nu_w - 2$ and the substring of amino acids from $\kappa_\nu + 1$ to $\nu_w - 1$ satisfies the constraints that allow to form a turn or a loop (such as conditions on length, flexibility, propensity, . . .) also depending on the relative laying of the two substructures. We have the elementary property:

$$\forall \nu,\ w \in \mathbf{V}^*,\ (\nu,\ w) \in E \Rightarrow \nu < w$$

for the lexicographic order, and this ensures the DAG structure.

The set $\mathbf{E}$ also contains edges of the form $(\top, \nu)$ that define the subset of starting vertices - the leading substrings satisfying specific constraints. Similarly, $\mathbf{E}$ contains edges of the form $(\nu, \bot)$ that define the subset of ending vertices, with a satisfactory trailing substring. Again, the length constraints applied to the substrings associated to edges imply that $|\mathbf{E}|$, the number of edges, is $\mathcal{O}(|\mathbf{V}|)$ or $\mathcal{O}(N)$.

Figure 9 gives a small example of such a graph (to simplify, only one orientation has been considered). An edge like (ν_1, ν_2) is forbidden, since the two corresponding substrings overlap. Edges like (ν_2, ν_3) or (ν_2, ν_6) are also forbidden, since the inserted substrings are respectively too short for a turn or too long for a loop.

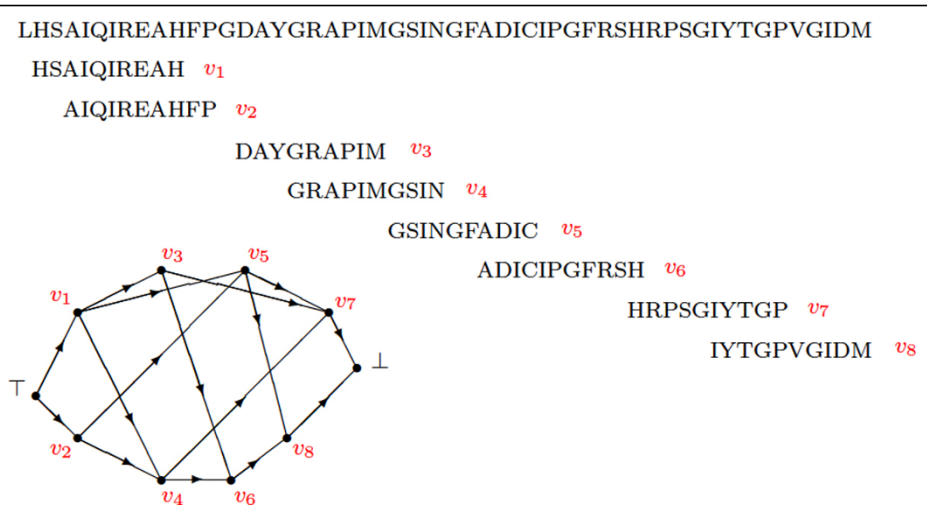

Figure 9 A short example of the graph structure.

Energy attributes

The attributes that complete the definition of the graph **G** are pseudo-energy functions defined as follows:

- $\forall v \in \mathbf{V}^*, \mathcal{E}_{\text{intr}}(v)$ represents the intrinsic energy of the given strand in the given orientation. This term is the sum of both the internal energy of the substructure, i.e. the interactions between its own amino acids, and the interaction energy with the environment (e.g. membrane and channel) apart from the rest of the considered protein.
  Note that $\mathcal{E}_{\text{intr}}(\top) = \mathcal{E}_{\text{intr}}(\bot) = 0$.
- $\forall (v, w) \in \mathbf{V}^* \times \mathbf{V}^*, \mathcal{E}_{\text{adj}}(v, w, s)$ represents the interaction energy of the pair (v, w) when the two corresponding strands are placed side by side along the barrel, with respect to the respective orientation parameters associated to the vertices and accordingly to the *relative shear* s. The energy will take into account the number of contacts and different side-chain interactions such as the packing of hydrophobic cores and bonding abilities. Then, $\forall (v, w) \in \mathbf{V}^* \times \mathbf{V}^*, \mathcal{E}_{\text{adj}}(v, w) = \min_s \mathcal{E}_{\text{adj}}(v, w, s)$is the interaction energy of the pair (v, w) for an optimal relative shear. It is further assumed that $\mathcal{E}_{\text{adj}}$ is defined over a superset of **E**, since we will consider the case where two adjacent strands are not consecutive along the sequence.

We also introduce the particular values:

$$\mathcal{E}_{\text{adj}}(\top, v) = \mathcal{E}_{\text{adj}}(v, \bot) = 0, \forall v \in \mathbf{V}.$$

- An associated function s_{adj} is defined such that:
- $\forall (v, w) \in \mathbf{V}^* \times \mathbf{V}^*, \mathcal{E}_{\text{adj}}(v, w, s_{\text{adj}}(v, w)) = \mathcal{E}_{\text{adj}}(v, w)$, which is a *relative shear* that leads to the optimal interaction energy.

An arising question is why the orientation degrees of freedom are described as a multiplicity of nodes but the *relative shear* degrees of freedom are considered when calculating the $\mathcal{E}_{\text{adj}}$ terms. A first answer comes from the fact that wrong orientations are rather absolute and will result in pruning the sets **E** and **V** while the *shear* parameters are not so discriminative. The main reason is that we will consider "floating" parts in which adjacencies are already set, while a *relative shear* between any two parts is not yet known. In such a situation, attaching the *relative shears* to node pairs allows a significant factorization.

- $\forall (v, w) \in \mathbf{E}, \forall t \in \{1, 2, \ldots, n - 1\}$ and $\forall$ s - a *relative shear*, $\mathcal{E}_{\text{loop}}(v, w, t, s)$ is related to the intrinsic energy of the turn/loop between the strands v and w (consecutive along the sequence) when they are placed at a distance t along the barrel with a *relative shear* s. The distance $t = 1$ corresponds to the case where the strands are placed consecutively on the barrel, while an integer value $t > 1$ will correspond to the case where $t - 1$ other strands are interleaf.
  To simplify, we will also use $\mathcal{E}_{\text{loop}}(\top, v)$ or $\mathcal{E}_{\text{loop}}(v, \bot)$ for denoting the intrinsic energy of the outer fragment attached respectively to a starting or an ending vertex v. As such a fragment has a free side, the position parameters may be dropped.

Then, in the usual case of two β-strands that fold as a hairpin, the related energy is considered to be $\mathcal{E}_{\text{adj}}(v, w) + \mathcal{E}_{\text{loop}}(v, w, 1, s_{\text{adj}}(v, w))$. It is supposed a relative flexibility for turns and loops, so, when a fold is feasible, $\mathcal{E}_{\text{loop}}$ is weak compared to $\mathcal{E}_{\text{adj}}$ and the relative placement of the two β-strands is enforced to be close to s_{adj}. Nevertheless, $\mathcal{E}_{\text{loop}}$ will result in a strong penalty

in the case of an unfeasible turn or loop, for example a loop with a majority of hydrophobic residues.

Protein folding problem

Given a graph $G(V, E, \mathcal{E}_{intr}, \mathcal{E}_{adj}, \mathcal{E}_{loop})$ defined as above, two integers n, S, and a permutation σ as 3 parameters, we look for the path $\mathcal{P}$ in $\mathbf{G}$ that maximizes the following objective function:

$$\mathcal{E} = \sum_{v\in\mathcal{P}} \mathcal{E}_{intr}(v) + \sum_{(v,w)\in\mathcal{P}} \mathcal{E}_{loop}(v,\ w) + \sum_{(v,w)\in\sigma(\mathcal{P})} \mathcal{E}_{adj}(v,w)$$

$$\text{such that } \sum_{(v,w)\in\mathcal{P}} s_{adj}(v,\ w) = S.$$

Such a path $\mathcal{P}$ whose vertices are arranged onto a circle is called a *circle-attached path*. The adjacent vertices in the path are not necessarily successive on the circle. This order of succession is determined by the given permutation σ (see Figure 10).

Solving as the longest path problem

We will first consider an open structure, as a β-sheet, where the adjacency of strands follows their natural order along the amino acid sequence, i.e. σ is an identity permutation. We involve here the constraint $\Sigma_{1<i\leq n}\ s_i = S$. Hence, solving such a structure will result in finding a path $\mathcal{P}$ in $\mathbf{G}$ whose overall "energy" is given by the sum:

$$\mathcal{E} = \Sigma_{v\in\mathcal{P}}\mathcal{E}_{intr}(v) + \Sigma_{(v,w)\in\mathcal{P}}[\mathcal{E}_{adj}(v,\ w) + \mathcal{E}_{loop}(v,\ w,\ 1,\ s_{adj}(v,\ w))]$$

Aiming at minimizing $\mathcal{E}$, the protein folding problem will turn into finding the path from $\top$ to $\bot$ that maximizes the criterion $\mathbf{C} = -\mathcal{E}$. Let $\mathbf{C}_v^h$ be the maximum value for $\mathbf{C}$ over all the paths from $\top$ to v, with a shear number of h of the corresponding β-sheet, then $\mathbf{C}_\top^0 = 0$ and $\forall v \in \mathbf{V}\backslash\{\top\}$, $\forall h$, $\mathbf{C}_v^h$ is defined as:

$$\mathbf{C}_v^h = \max_{u\in\mathbf{V},(u,v)\in\mathbf{E}} [\mathbf{C}_u^{h-s_{adj}(u,v)} - \mathcal{E}_{intr}(v) - \mathcal{E}_{adj}(u,\ v) - \mathcal{E}_{loop}(u,\ v,\ 1,\ s_{adj}(u,\ v))]$$

Since the graph is a DAG, the longest path problem is solved with a well known dynamic programming scheme [34] of complexity $\mathcal{O}(|\mathbf{V}|)$ in space and $\mathcal{O}(|\mathbf{V}| + |\mathbf{E}|)$ in time, that is also $\mathcal{O}(N)$ for both, from the structural constraints that relate $|\mathbf{V}|$, $|\mathbf{E}|$ and N. The objective is the computation of $\mathbf{C}_\bot^S$ and the optimal structure is then reconstructed by a usual traceback post-processing. Note that, for each path, we only have to consider its last vertex, so, we have to track single index states.

For a barrel secondary structure, we have to consider a closing spatial adjacency between the last and the first strands. σ is still an identity permutation. The constraint on the shear number becomes $\Sigma_{1<i\leq n+1}\ s_i = S$. The dynamic programming scheme is almost the same as previously, except that we also have to keep track of the first vertex of any path. So, $\forall v \in \mathbf{V}^*$, such that $(\top, v) \in \mathbf{E}$, let $\mathbf{C}_{(v,v)}^0 = -\mathcal{E}_{intr}(v) - \mathcal{E}_{loop}(\top,\ v)$, then the general recurrence is: $\forall v, w \in \mathbf{V}^*$, $\forall h$, such that $(\top, v) \in \mathbf{E}$,

$$\mathbf{C}_{(v,w)}^h = \max_{u\in\mathbf{V},(u,w)\in\mathbf{E}} [\mathbf{C}_{(v,u)}^{h-s_{adj}(u,w)} - \mathcal{E}_{intr}(w) - \mathcal{E}_{adj}(u,\ w) - \mathcal{E}_{loop}(u,\ w,\ 1,\ s_{adj}(u,\ w))]$$

and a special closing step is needed: $\forall\ v \in \mathbf{V}^*$, $\forall h$, such that $(\top, v) \in \mathbf{E}$,

$$\mathbf{C}_{(v,\bot)}^h = \max_{u\in\mathbf{V},(u,\bot)\in\mathbf{E}} \left[\mathbf{C}_{(v,u)}^{h-s_{adj}(u,v)} - \mathcal{E}_{adj}(u,\ v) - \mathcal{E}_{loop}(u,\ \bot)\right]$$

The goal is to calculate $\max_{v,(\top,v)\in\mathbf{E}}\mathbf{C}_{(v,\bot)}^S$. Thus the scheme is of complexity $\mathcal{O}(|\mathbf{V}|^2)$ in space and $\mathcal{O}(|\mathbf{V}|\cdot|\mathbf{E}|)$ in time, that is also $\mathcal{O}(N^2)$ for both, from the structural constraints. This may produce paths of any length and the constraint of n strands is applied as a cut in the recurrence.

Generalization

In a more general case, we consider permutations to deal with the fact that the arrangements of the strands along the barrel do not necessarily follow their order along the sequence. This usually occurs with Greek key motifs or more rarely with Jelly roll motifs. Hence, the protein folding problem becomes finding the longest

Figure 10 A permuted β-barrel with a Greek key motif 3654, σ = 1 2 3 6 5 4. Left part is a 3D synthetic view of such a protein. Right part is a schematic view from above, looking down along the axis of the barrel.

path $\mathcal{P}$ in a graph with respect to a given permutation σ, i.e. the vertices of $\mathcal{P}$, seen on a circle as in Figure 10 are permuted according to σ.

Let σ be a circular permutation of {1, 2, . . . , n}. When 1, 2, . . . , n are numbering the positions along the barrel, values σ (1), σ (2), . . . , $\sigma(n)$ will give the respective ranks of the strands in the sequence order. A position of reference along the barrel is fixed by setting $\sigma(1) = 1$. Figure 10 shows a first example of a structure with a Greek key motif, which is described by the permutation σ = (1, 2, 3, 6, 5, 4).

Hereafter, we will illustrate the presentation of our algorithm by following the example σ = (1, 2, 5, 4, 3, 6), which is a bit trickier situation. This example is now said the *current example*. The corresponding structure and the dynamic programming process are illustrated in Figures 11 and 12.

The dynamic programming scheme now consists in building a barrel, by adding a next strand, taken in sequence with respect to the graph edges, but that is inserted at the position defined by the given permutation. Useful values are the ranks (in the sequence order) of the two strands between which a given one will be inserted. For instance, with the current example, the 5th strand will be inserted between the 2nd and the 4th strands.

Let now k denote the level of construction ($1 \leq k \leq n$), that is the number of strands already placed.

Proposition 4 *The k^{th} strand (in the sequence order) is inserted between the two strands whose ranks (in the sequence order) are* $\mathbf{left}_k$ *and* $\mathbf{right}_k$*, defined as:*

$$\mathbf{left}_k = \begin{cases} \sigma(\sigma^{-1}(k)-1) & \text{if } \sigma^{-1}(k) > 1 \\ \sigma(n) & \text{otherwise} \end{cases}, \quad \mathbf{right}_k = \begin{cases} \sigma(\sigma^{-1}(k)+1) & \text{if } \sigma^{-1}(k) < n \\ 1 & \text{otherwise} \end{cases}$$

With the current example, we get (see Figure 11):

$\mathbf{left}_1 = 6$ $\mathbf{left}_2 = 1$ $\mathbf{left}_3 = 4$ $\mathbf{right}_1 = 2$ $\mathbf{right}_2 = 5$ $\mathbf{right}_3 = 6$

$\mathbf{left}_4 = 5$ $\mathbf{left}_5 = 2$ $\mathbf{left}_6 = 3$ $\mathbf{right}_4 = 3$ $\mathbf{right}_5 = 4$ $\mathbf{right}_6 = 1$

An important piece of information to store for the dynamic programming scheme is the set of "active" indices, i.e. ranks of the strands (in the sequence order) that are not definitively bonded on both sides, along the barrel, and also not linked along the sequence and thus have to be kept as degrees of freedom. So, in the current example (see Figure 12), we have to keep in memory as many solutions (to subproblems) as valid instances of the 2nd and 4th strands, until an optimal choice for these is recorded as a solution for each instance of the 5th strand. At that time, any instance as the 5th strand is kept as a candidate for a link with the 6th, by a turn or loop, while the different instances as the 3rd and 1st are kept for proceeding to an insertion in between.

Definition 5 *Two ranks i and j, which refer to the sequence order, are said "adjacent" if:*

$$|\sigma^{-1}(i) - \sigma^{-1}(j)| \in \{1, n-1\},$$

where the case n - 1 is intended for the adjacency that will close the barrel.

Proposition 6 *The set of "active" indices (in the sequence) at level k is defined by:*

$$\mathbf{conf}_k = \{k\} \cup \{i | 1 \leq i < k \text{ and } (\exists j : k < j \leq n | i, j \text{ are "adjecent"})\}$$

With the current example of Figures 11 and 12, we get:

$\mathbf{conf}_1 = \{1\}$ $\mathbf{conf}_2 = \{1, 2\}$ $\mathbf{conf}_3 = \{1, 2, 3\}$
$\mathbf{conf}_4 = \{1, 2, 3, 4\}$ $\mathbf{conf}_5 = \{1, 3, 5\}$ $\mathbf{conf}_6 = \{6\}$

Thus, for this example, the maximal complexity in space, $\mathcal{O}(N^4)$, is reached for the set of solutions to the subproblem with 4 strands. Then looping over this set, for computing the set of solutions to the subproblem with 5 strands, will also cost $\mathcal{O}(N^4)$ in time, since the choice for the 5th strand is bounded by the structural constraints embedded as edges in the graph. It is a difference with most of the dynamic programming schemes where the complexity in time is expressed with an additional $\mathcal{O}(N)$ factor compared the complexity in space. As an other example, in the case of Figure 10, we obtain the complexity $\mathcal{O}(N^2)$ in both time and space,

Figure 11 A permuted β-barrel with a Greek key motif 5436, σ = 1 2 5 4 3 6. Left part is a "flattened" side view, stands 6 and 1 being also adjacent. Right part is a schematic view from above, looking down along the axis of the barrel.

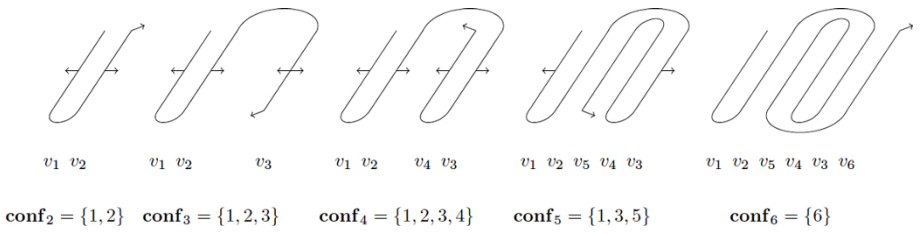

Figure 12 Successive steps of the dynamic programming scheme for a Greek key motif 5436, σ = 1 2 5 4 3 6. Arrows show the "waiting" constraints. As all the (loop or adjacency) constraints have not been satisfied by determining the linked segments, a segment should be considered as a variable with the memory of a partial solution for each possible instance. The corresponding ranks are said "active" and are listed in the **conf** set.

which is similar to the case where σ is an identity permutation.

Now we have to decide at which minimal level k each term $\mathcal{E}_{\text{adj}}$ or $\mathcal{E}_{\text{loop}}$ is determined and can be integrated in the dynamic programming scheme. For the $\mathcal{E}_{\text{adj}}$ terms, it is simply asserted that the previous or the next strand along the barrel is already placed when $\mathbf{left}_k < k$ or $\mathbf{right}_k < k$, respectively.

Proposition 7 *For all k, we have:*

$$\mathbf{left}_k \text{ ¡ } k \Leftrightarrow \mathbf{left}_k \in \mathbf{conf}_{k-1},$$
$$\mathbf{right}_k < k \Leftrightarrow \mathbf{right}_k \in \mathbf{conf}_{k-1}$$

This results from the definition of the "active" indices of $\mathbf{conf}_{k-1}$. To simplify the further energy expression, we use the following notation for an "ifelse" function:

$$\mathbf{if}_k(i,\ \mathcal{E}) = \begin{cases} \mathcal{E} \text{ if } i < k \\ 0 \text{ otherwise} \end{cases}$$

For the $\mathcal{E}_{\text{loop}}$ terms, the problem is to wait until the *relative shear* between the two ends of a turn or loop is solved by the interleaf adjacencies. So, in the given example, the energy of the loop between the 2nd and 3rd strands can only be evaluated when the 5th strand has been laid and the optimal *relative shear* $s^*_{\text{adj}}(v_2, v_3) = s_{\text{adj}}(v_2, v_5) + s_{\text{adj}}(v_5, v_4) + s_{\text{adj}}(v_4, v_3)$ is known.

Definition 8 *Let $\mathcal{A}_k$ be the relation on positive integers, defined as: ∀i, j,*

$$i\mathcal{A}_k j \Leftrightarrow \begin{cases} i = j \\ or \\ i \le k \text{ and } j \le k \text{ and } i, j \text{ are "adjacent"} \end{cases}$$

then let $\mathcal{A}^*_k$ denote the equivalence relation defined by the transitive closure of $\mathcal{A}_k$ and let $\mathbf{A}_k = \{i < k | i\mathcal{A}^*_k(i+1)\}$.

Thus, $i \in \mathbf{A}_k$ means that the i^{th} and $(i+1)^{\text{st}}$ strands are geometrically linked by adjacencies when the k^{th} substructure is laid and we can compute by composition an optimal *relative shear* s^*_{adj}.

We will now focus on the set $\delta\ \mathbf{A}_k = \mathbf{A}_k - \mathbf{A}_{k-1}, \forall\ k > 1$.

Proposition 9 *For all k, we have:*

$$(k-1) \in \delta\mathbf{A}_k \Leftrightarrow \mathbf{left}_k \mathcal{A}^*_{k-1}(k-1) \text{ or } \mathbf{right}_k\ \mathcal{A}^*_{k-1}(k-1)$$

Proposition 10 *For all i < k - 1,*

$$i \in \delta\mathbf{A}_k \Leftrightarrow \begin{cases} i \notin \mathbf{A}_{k-1} \\ \text{and} \\ \mathbf{left}_k\ \mathcal{A}^*_{k-1}\ i \quad \text{and } \mathbf{right}_k\ \mathcal{A}^*_{k-1}(i+1) \\ \text{or} \\ \mathbf{right}_k\ \mathcal{A}^*_{k-1}\ i \text{ and } \mathbf{left}_k\ \mathcal{A}^*_{k-1}(i+1) \end{cases}$$

Definition 11 *Let $\mathbf{T}_k \subset \mathbf{V}^{*|\mathbf{conf}_k|}$ denote the set of all tuples of $|\mathbf{conf}_k|$ vertices such that there is at least one path (of k edges) starting from ⊤ and passing through these vertices in order.*

For any instance $\mathbf{z} \in \mathbf{T}_k$ of such a tuple and, $\forall i \in \mathbf{conf}_k$, let $\mathbf{z}[i]$ *denote the i^{th} vertex of a corresponding path.*

This notation (not to be confused with $\mathbf{z}_i$, the i^{th} component of tuple $\mathbf{z}$) is not ambiguous since, from definition, the vertex $\mathbf{z}[i]$ is in common to any path associated to $\mathbf{z}$. Particularly, $\mathbf{z}[k]$ is the last vertex of any path associated to $\mathbf{z}$.

Proposition 12 *For all $\mathbf{z} \llcorner \mathbf{T}_k$, the set of tuples corresponding to paths of length k - 1 that can be extended to a path corresponding to $\mathbf{z}$ is defined as:*

$$\mathbf{pre}(\mathbf{z}) = \{\mathbf{y} \in \mathbf{T}_{k-1} | (\mathbf{y}[k-1], \mathbf{z}[k]) \in \mathbf{E} \text{ and } \forall i \in \mathbf{conf}_k \cap \mathbf{conf}_{k-1}, \mathbf{y}[i] = \mathbf{z}[i]\}$$

Let $\mathbf{C}^h_{k,\mathbf{z}}$ be the maximum value for $\mathbf{C}$ over all paths starting from ⊤ and leading in order through the vertices of a given tuple $\mathbf{z} \in \mathbf{T}_k$ with a shear number of h of the corresponding β-barrel. The general recurrence relation is: $\forall \mathbf{z} \in \mathbf{T}_k$,

$$\mathbf{C}^h_{k,\mathbf{z}} = \max_{\mathbf{y} \in \mathbf{pre}(\mathbf{z})} \Big(\mathbf{C}^{h - s_{\text{adj}}(\mathbf{y}[|\mathbf{left}_k|], \mathbf{z}[k]) - s_{\text{adj}}(\mathbf{z}[k], \mathbf{y}[|\mathbf{right}_k|]) + s_{\text{adj}}(\mathbf{y}[|\mathbf{left}_k|], \mathbf{y}[|\mathbf{right}_k|])}_{k-1, \mathbf{y}} - \mathcal{E}_{\text{intr}}(\mathbf{z}[k]) - \mathbf{if}_k(\mathbf{left}_k, \mathcal{E}_{\text{adj}}(\mathbf{y}[\mathbf{left}_k], \mathbf{z}[k] - \mathbf{if}_k(\mathbf{right}_k, \mathcal{E}_{\text{adj}}(\mathbf{z}[k], \mathbf{y}[\mathbf{right}_k] - \sum_{i \in \delta\mathbf{A}_k} \mathcal{E}_{1\text{oop}}(\mathbf{y}[i],\ \mathbf{y}[i+1],\ \sigma^{-1}(i+1) - \sigma^{-1}(i),\ s^*_{\text{adj}}(\mathbf{y}[i],\ \mathbf{y}[i+1])) \Big)$$

Note that, from proposition 7, $\forall \mathbf{y} \in \mathbf{T}_{k-1}$, if $\mathbf{left}_k < k$ then the vertex $\mathbf{y}[\mathbf{left}_k]$ is defined (and the same is worth for $\mathbf{right}_k$). We can check that each $\mathcal{E}_{\text{adj}}$ term is finally counted exactly once in the sum, at the level corresponding to the position of its further vertex in the sequence order. The optimum is found at $k = n$ and $h = S$.

Corollary 13 *The complexities are* $\mathcal{O}\left(N^{\max_k ||\mathbf{conf}_k||}\right)$ *in space and time.*

For any permutation, we have

$$||\mathbf{conf}_{n-k}|| \leq \min\{1 + 2k,\ n - k\},\ \forall k = 0,\ \ldots,\ n-1$$

Hence, $\max_k ||\mathbf{conf}_k|| \leq 1 + (2n - 2)/3$. For a permutation that only differs from the identity permutation by disjoint Greek key motifs [35], i.e. $\sigma = (1,\ 2,\ \ldots,\ i_1,\ \mathcal{G}_1,\ i_1 + 5, \ldots,\ i_2, \mathcal{G}_2,\ i_2 + 5, \ldots,\ \mathcal{G}_j,\ \ldots,\ n)$ where $\mathcal{G}_j = i_j + 3$, $i_j + 2$, $i_j + 1$, $i_j + 4$ or $\mathcal{G}_j = i_j + 1$, $i_j + 4$, $i_j + 3$, $i_j + 2$, it is easy to prove that $\max_k ||\mathbf{conf}_k|| \leq 4$ by a discrete analysis on different configurations. The complexities are thus at most $\mathcal{O}(N^4)$ for such a permutation.

In short, it is possible to compute the optimum in $\mathcal{O}(N^2)$ running time for structures corresponding to the identity permutation and from $\mathcal{O}(N^2)$ (for instance, example of Figure 10) to $\mathcal{O}(N^4)$ (for instance, example of Figure 11) for structures containing disjoint Greek key motifs, where N is the input sequence length. These computation costs might be further improved by a tree decomposition-based algorithm that we are currently working on.

Implementation details

The number of strands n and the shear number S determine the geometry of the barrel, particularly the membrane spanning part of the segments, and are thus involved in the computation of energy terms. If known, the algorithm can enforce these 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]. The values for S, are even and included between n and $2n$ [28,29]. The problem is then solved by the constrain dynamic programming with the constraints of given n and S. A small number of couples (n, S) have to be explored and our algorithm is fast enough for that.

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. We use here by default a parameter of 3 nm for the membrane thickness, thus 8 residues thick [36,37]. The features on size, polarity [38], and flexibility [39] of turns and loops are also taken into consideration, i.e. turns and loops satisfy threshold constraints on their polarity and flexibility indices and their length. Their energies are approximated by hydrophobicity [31].

We use the Dunbrack backbone-dependent rotamer library [40] and the partial charges from GROMOS force field [41] 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 $e_{uv} = <e_{uv|rotamers}>$. 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 β-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.

Experimental setup

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.

Datasets

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

- Folding: We used CD-HIT [42] to constrain the redundancy in proteins. A threshold of 40% similarity was applied 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 41 sequences [PDB: 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, 3FHH_A, 3FID_A, 1ILZ_A, 1BY3_A, 2GSK_A, 1BH3_A, 2HDF_A, 2J1N_A, 2IAH_A, 3JTY_A, 1BXW_A, 2VDF_A, 1PNZ_A, 3GP6_A, 1AF6_A, 3NJT_A, 2O4V_A, 2ODJ_A, 1QJ8_A, 1P4T_A, 2POR_ , 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 over-fitting 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.

Acknowledgements

The authors would like to thank all the INRIA AMIB Team members, especially Mireille Régnier, Yann Ponty, Julie Bernauer and Balaji Raman. This article has been published as part of *BMC Genomics* Volume 13 Supplement 2, 2012: Selected articles from the First IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS 2011): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/13/S2.

Authors' contributions

VDTT and PC designed the algorithm. JMS, SS and VDTT designed the experiments. VDTT performed the experiments and analyzed the results. VDTT, PC and SS wrote the manuscript. JMS conceived and supervised the study and revised the manuscript. All authors read and approved the paper.

Competing interests

The authors declare that they have no competing interests.

Published: 12 April 2012

References

1. Tamm LK, Hong H, Liang B: **Folding and assembly of β -barrel membrane proteins.** *Biochim Biophys Acta* 2004, **1666**:250-263.
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28**:235-242.
3. Arora A, Tamm LK: **Biophysical approaches to membrane protein structure determination.** *Curr Opin Struct Biol* 2001, **11**:540-547.
4. Casadio R, Fariselli P, Martelli PL: **In silico prediction of the structure of membrane proteins: Is it feasible?** *Brief Bioinform* 2003, **4(4)**:341-348.
5. Taylor PD, Toseland CP, Attwood TK, Flower DR: **Beta-barrel transmembrane proteins: Enhanced prediction using a Bayesian approach.** *Bioinformation* 2006, **1(6)**:231-233.
6. Freeman TCJ, Wimley WC: **A highly accurate statistical approach for the prediction of transmembrane beta-barrels.** *Bioinformatics* 2010, **26(16)**:1965-74.
7. Gromiha M, Ahmad S, Suwa M: **Neural network-based prediction of transmembrane β -strand segments in outer membrane proteins.** *J Comput Chem* 2004, **25**:762-767.
8. Gromiha MM, Ahmad S, Suwa M: **TMBETA-NET: discrimination and prediction of membrane spanning β -strands in outer membrane proteins.** *Nucleic Acids Res* 2005, **33**:W164-W167.
9. Randall A, Cheng J, Sweredoski M, Baldi P: **TMBpro: secondary structure, β -contact and tertiary structure prediction of transmembrane β -barrel proteins.** *Bioinformatics* 2008, **24**:513-520.
10. Ou YY, Chen SA, Gromiha MM: **Prediction of membrane spanning segments and topology in β -barrel membrane proteins at better accuracy.** *J Comput Chem* 2010, **31**:217-223.
11. Bagos P, Liakopoulos T, Hamodrakas S: **Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method.** *BMC Bioinformatics* 2005, **6**:7.
12. Waldispühl J, Berger B, Clote P, Steyaert JM: **Predicting transmembrane β -barrels and interstrand residue interactions from sequence.** *Proteins* 2006, **65**:61-74.
13. McGuffin LJ, Bryson K, Jones DT: **The PSIPRED protein structure prediction server.** *Bioinformatics* 2000, **16**:404-405.
14. Jacoboni I, Martelli PL, Fariselli P, Pinto VD, Casadio R: **Prediction of the transmembrane regions of β -barrel membrane proteins with a neural network-based predictor.** *Protein Sci* 2001, **10**:779-787.
15. Martelli P, Fariselli P, Krogh A, Casadio R: **A sequence-profile-based HMM for predicting and discriminating β -barrel membrane proteins.** *Bioinformatics* 2002, **18(Suppl 1)**:S46-S53.
16. Ahn C, Yoo S, Park H: **Prediction for beta-barrel transmembrane protein region using HMM.** *KISS* 2003, **30(2)**:802-804.
17. Bigelow HR, Petrey DS, Liu J, Przybylski D, Rost B: **Predicting transmembrane beta-barrels in proteomes.** *Nucleic Acids Res* 2004, **32**:2566-2577.
18. Bagos PG, Liakopoulos TD, Spyropoulos IC, Hamodrakas SJ: **PRED-TMBB: a web server for predicting the topology of β -barrel outer membrane proteins.** *Nucleic Acids Res* 2004, **32**:W400-W404.
19. Natt NK, Kaur H, Raghava G: **Prediction of transmembrane regions of β -barrel proteins using ANN-and SVM-based methods.** *Proteins* 2004, **56**:11-18.
20. Tusnády GE, Dosztányi Z, Simon I: **PDB_TM: selection and membrane localization of transmembrane proteins in the Protein Data Bank.** *Nucleic Acids Res* 2005, **33**:D275-D278.
21. Bagos P, Liakopoulos T, Spyropoulos I, Hamodrakas S: **A Hidden Markov Model method, capable of predicting and discriminating β -barrel outer membrane proteins.** *BMC Bioinformatics* 2004, **5**:29.
22. Dayhoff MO, Schwartz RM, Orcutt CB: **A model of evolutionary change in proteins.** *Atlas of Protein Sequence and Structure* 1978, **5(Suppl 3)**:345-352.
23. Koebnik R, Krämer L: **Membrane assembly of circularly permuted variants of the E. coli outer membrane protein OmpA.** *J Mol Biol* 1995, **250**:617-626.
24. **Beta-Barrel Predictor Web Server.** [http://www.lix.polytechnique.fr/Labo/Van-Du.Tran/bbp/].
25. Tran VD, Chassignet P, Steyaert JM: **Prediction of permuted super-secondary structures in beta-barrel proteins.** *Proceedings of the 2011 ACM Symposium on Applied Computing SAC'11, ACM Digital Library* Taichung, Taiwan; 2011, 110-111.
26. Tran VD, Chassignet P, Sheikh S, Steyaert JM: **Energy-based classification and structure prediction of transmembrane beta-barrel proteins.** *Proceedings of the 2011 IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), IEEE Xplore* Orlando, FL, USA; 2011, 159-164.
27. Marsh D: **Infrared dichroism of twisted beta-sheet barrels. The structure of E. coli outer membrane proteins.** *J Mol Biol* 2000, **297**:803-808.
28. Murzin AG, Lesk AM, Chothia C: **Principles determining the structure of β -sheet barrels in proteins I. A theoretical analysis.** *J Mol Biol* 1994, **236**:1369-1381.
29. Murzin AG, Lesk AM, Chothia C: **Principles determining the structure of β -sheet barrels in proteins II. The observed structures.** *J Mol Biol* 1994, **236**:1382-1400.
30. Chou KC, Carlacci L, Maggiora GM: **Conformational and geometrical properties of idealized beta-barrels in proteins.** *J Mol Biol* 1990, **213**:315-326.
31. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein.** *J Mol Biol* 1982, **157**:105-132.
32. Fano R: *Transmission of Information* Wiley, New York; 1961.
33. Gibrat JF, Garnier J, Robson B: **Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs.** *J Mol Biol* 1987, **198**:425-443.

34. Cormen TH, Leiserson CE, Rivest RL, Stein C: *Introduction to Algorithms*. 3 edition. MIT Press; 2009.
35. Zhang C, Kim SH: **A comprehensive analysis of the Greek key motifs in protein β -barrels and β -sandwiches.** *Proteins* 2000, **40**:409-419.
36. Lewis BA, Engelman DM: **Lipid bilayer thickness varies linearly with acyl chain length in fluid phosphatidylcholine vesicles.** *J Mol Biol* 1983, **166(2)**:211-217.
37. Rawicz W, Olbrich K, McIntosh T, Needham D, Evans E: **Effect of Chain Length and Unsaturation on Elasticity of Lipid Bilayers.** *Biophys J* 2000, **79**:328-339.
38. Grantham R: **Amino Acid Difference Formula to Help Explain Protein Evolution.** *Science* 1974, **185**:862-864.
39. Bhaskaran R, Ponnuswamy P: **Amino acid scale: average flexibility index.** *Int J Pept Protein Res* 1988, **32**:242-255.
40. Dunbrack RL, Cohen FE: **Bayesian statistical analysis of protein side-chain rotamer preferences.** *Protein Sci* 1997, **6(8)**:1661-1681.
41. van Gunsteren WF, Billeter SR, Eising AA, Hünenberger PH, Krüger P, Mark AE, Scott WRP, Tironi IG: **Biomolecular simulation: the GROMOS96 manual and user guide.** *vdf Hochschulverlag AG an der ETH Zürich and BIOMOS b.v.: Zürich, Groningen* 1996.
42. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22(13)**:1658-1659.
43. Liu WM: **Shear numbers of protein β -barrels: definition, refinements and statistics.** *J Mol Biol* 1998, **275**:541-545.

doi:10.1186/1471-2164-13-S2-S5
Cite this article as: Tran *et al.*: A graph-theoretic approach for classification and structure prediction of transmembrane β-barrel proteins. *BMC Genomics* 2012 **13**(Suppl 2):S5.