PROTEINS: Structure, Function, and Bioinformatics 65:61–74 (2006)

Predicting Transmembrane β-Barrels and Interstrand Residue Interactions from Sequence

J. Waldispühl,[1,2] Bonnie Berger,[2,3]* Peter Clote,[1,4]* and Jean-Marc Steyaert[5]
[1]*Department of Biology, Boston College, Chestnut Hill, Massachusetts*
[2]*Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts*
[3]*Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts*
[4]*Department of Computer Science (courtesy appt.), Boston College, Chestnut Hill, Massachusetts*
[5]*Laboratoire d'Informatique (LIX), École Polytechnique, Palaiseau, France*

ABSTRACT **Transmembrane β-barrel (TMB) proteins are embedded in the outer membrane of Gram-negative bacteria, mitochondria, and chloroplasts. The cellular location and functional diversity of β-barrel outer membrane proteins (omps) makes them an important protein class. At the present time, very few nonhomologous TMB structures have been determined by X-ray diffraction because of the experimental difficulty encountered in crystallizing transmembrane proteins. A novel method using pairwise interstrand residue statistical potentials derived from globular (nonouter membrane) proteins is introduced to predict the supersecondary structure of transmembrane β-barrel proteins. The algorithm *transFold* employs a generalized hidden Markov model (i.e., multitape S-attribute grammar) to describe potential β-barrel supersecondary structures and then computes by dynamic programming the minimum free energy β-barrel structure. Hence, the approach can be viewed as a "wrapping" component that may capture folding processes with an initiation stage followed by progressive interaction of the sequence with the already-formed motifs. This approach differs significantly from others, which use traditional machine learning to solve this problem, because it does not require a training phase on known TMB structures and is the first to explicitly capture and predict long-range interactions. *TransFold* outperforms previous programs for predicting TMBs on smaller (≤200 residues) proteins and matches their performance for straightforward recognition of longer proteins. An exception is for multimeric porins where the algorithm does perform well when an important functional motif in loops is initially identified. We verify our simulations of the folding process by comparing them with experimental data on the functional folding of TMBs. A Web server running *transFold* is available and outputs contact predictions and locations for sequences predicted to form TMBs. Proteins 2006;65:61–74.**

Key words: outer membrane proteins; transmembrane β-barrels; residue contact; structure modeling; structure prediction; protein folding; energy model; minimum folding energy; S-attribute grammar

INTRODUCTION

Gram-negative bacteria are surrounded by two radically different membranes, themselves separated by a region called the periplasm. The inner membrane is a normal lipid bilayer, which forms an important permeability barrier and is associated with membrane-associated metabolic functions. Transmembrane (TM) α-helical proteins are typically embedded in this lipid bilayer. In contrast, the composition of the outer membrane differs significantly from that of the inner membrane. Although the inner leaflet (periplasmic side) of the outer membrane has similar composition to that of the inner membrane, the outer leaflet (extracellular side) is composed of lipopolysaccharides. In addition, the architecture of proteins embedded in the outer membrane is strikingly different from those embedded in the inner membrane. In place of TM α-helical proteins typical of the inner membrane, outer membrane transmembrane proteins are generally formed into β-strands, assembled in a self-closed β-sheet forming a barrel that spans the membrane. Outer membrane proteins (omps) are then usually assimilated into TM β-barrels (TMB). Such β-barrel membrane proteins are not exclusively found in Gram-negative prokaryotes; indeed, it is also believed that in eukaryotes, outer membrane proteins in mitochondria and chloroplasts adopt the same architecture.

In recent years, there has been an increasing interest in TM β-barrel proteins, both among experimental biologists and computational biologists. In vitro and in vivo studies

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: NSF (to B. Berger); Grant number: Grant ITR (ASE + NIH)-(dms)-0428715; Grant sponsor: NSF (to P. Clote and J. Waldispühl); Grant number: DBI-0543506.

*Correspondence to: Peter Clote, Department of Biology, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467 or Bonnie Berger, Department of Mathematics, MIT, Cambridge, MA 02139. E-mail: clote@bc.edu or bab@mit.edu

Received 6 December 2005; Revised 13 March 2006; Accepted 14 March 2006

Published online 21 July 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21046

of omps[1–4] have multiplied; nevertheless, the real nature of forces driving the folding of TM β-barrel proteins remains obscure. Most of what is now known about omps is clearly summarized in recent reviews.[5–7] At the same time, ever since the seminal hidden Markov model method introduced by Martelli et al.[8] to predict TM β-barrels, a number of other machine learning methods have been developed in the past 2 years for TM β-barrel supersecondary structure prediction.[9–13]

As far as we know, the mechanism of folding of TM β-barrel proteins is quite different from TM α-bundle proteins, which are thought to fold in a two-stage model[14,15] where the α-helices fold independently and are subsequently inserted into the bilayer. The TMB folding process is described in a four-stage model[5] where long-range interactions play a fundamental role. The unfolded molecule collapses first at the inner (periplasmic) surface of the outer membrane. A fraction of the β-structure is then formed and parts of secondary structure elements lie on the inner surface of the membrane in a state called "molten disk." Thereafter, β-hairpin loops translocate into the bilayer in a similar way as in globular protein folding. This state is naturally called "molten globule." Following this state, there is a rearrangement of β contacts resulting in the native state.

It is still extremely difficult to experimentally determine the structure of transmembrane proteins; thus, computational methods are needed. At this time, only a few structures have been crystallized (about 100 TM β-barrels, but less than 20 structures remain after removal of homologous sequences). Despite the paucity of nonredundant sequences whose TM β-barrel structure has been solved, all existing TM β-barrel prediction methods are based on classical machine learning techniques (e.g., hidden Markov models or neural networks). It follows that, even with the use of sophisticated techniques to avoid redundancy and overfitting the data, the evaluation of the reliability of any machine learning method in this context is fragile. (E.g., because advanced statistical techniques were applied, PROFtmb[10] uses only eight omp sequences for training and evaluation of its structural predictions). Hence, as was the case for TM α-helix prediction,[16] the performance of these prediction methods could be significantly overestimated.

Moreover, existing approaches using hidden Markov models (HMMs), neural networks (NNs), support vector machines (SVMs), etc., are limited to local information in fixed-size training sequences. (Indeed, all such methods, when training on size n sequences, where n is much smaller than the protein size, will invariably will miss potential long-range interactions between residues i and j, where $|j - i| > n$.) Given our state-of-the-art knowledge of the four-state physical folding process for omps, these limitations reveal a potential inability for these machine learning techniques to take into account all forces driving the folding, especially those long-range interactions between residues. (For example, the interactions involved in the pairing of the first and last TM β-strands of the barrel are difficult to handle with traditional machine-learning methods.) Linguistic methods developed previously for prediction of TM α-bundles[17] and globular β-sheets[18] enable us to circumvent the problem of an inadequate size training set and capture long-range interactions.

In this article we introduce a novel method that takes into account long-range interactions in β-barrel structure prediction. Such long-range interactions are used efficiently and without loss of generality. Moreover, to avoid using any a priori knowledge of TM β-barrel structures and to ensure the universality of our energy model, we use β-strand contact energy parameters for globular proteins taken from the program BETAWRAP.[19–22] Our model is based on an abstract physical description of omps (only the basic topology is represented; see below), and the energy parameters are defined independently of any known TM β-barrel. It follows that our method can be applied to any amino acid sequence, without preliminary analysis (e.g., one need not first determine that the protein contains β-strands, or that it is likely to be a transmembrane protein). With respect to our model and the β-strand contact energies, our algorithm then uses a multitape S-attribute grammar to describe all possible β-barrel supersecondary structures, from which the minimum free energy supersecondary structure is selected by dynamic programming.

The method described in this article is implemented in a program called *transFold*, which introduces new software (to potentially be used together with existing methods but giving alternate solutions) to study the properties of TM β-barrel proteins, predict their structure and distinguish them from other types of protein folds. We have computed the accuracy of classification (i.e., given an amino acid sequence, whether the protein is a TM β-barrel protein) and structure prediction (i.e., given an amino acid sequence, determine the minimum free energy TM β-barrel supersecondary structure). In addition to classical secondary structure predictions, *transFold* provides a list of pairs of residues that are in contact in the predicted structure. These contact predictions are examined and discussed in their biological context. Finally, we compare our performance results with existing methods, whenever possible using the same datasets. Apart from the standard performance calibrations of accuracy, sensitivity, specificity, and Mathews correlation coefficient, we include two novel measures.

The software *transFold* can be used for three types of applications: (1) prediction of omps according to structural and potentially functional characteristics (water-filled channel or not, in the present paper); (2) automatic structural classification of omps; (3) in silico folding experiments for mutants of omps. In particular, in this article we reproduce experiments on the Outer membrane protein A (OmpA) of *Escherichia coli* and compare the results computed in silico with those observed in vivo. The overall performance of *transFold* and its use in performing computational experiments for mutagenesis of outer membrane proteins argue for the potential interest of experimental biologists.

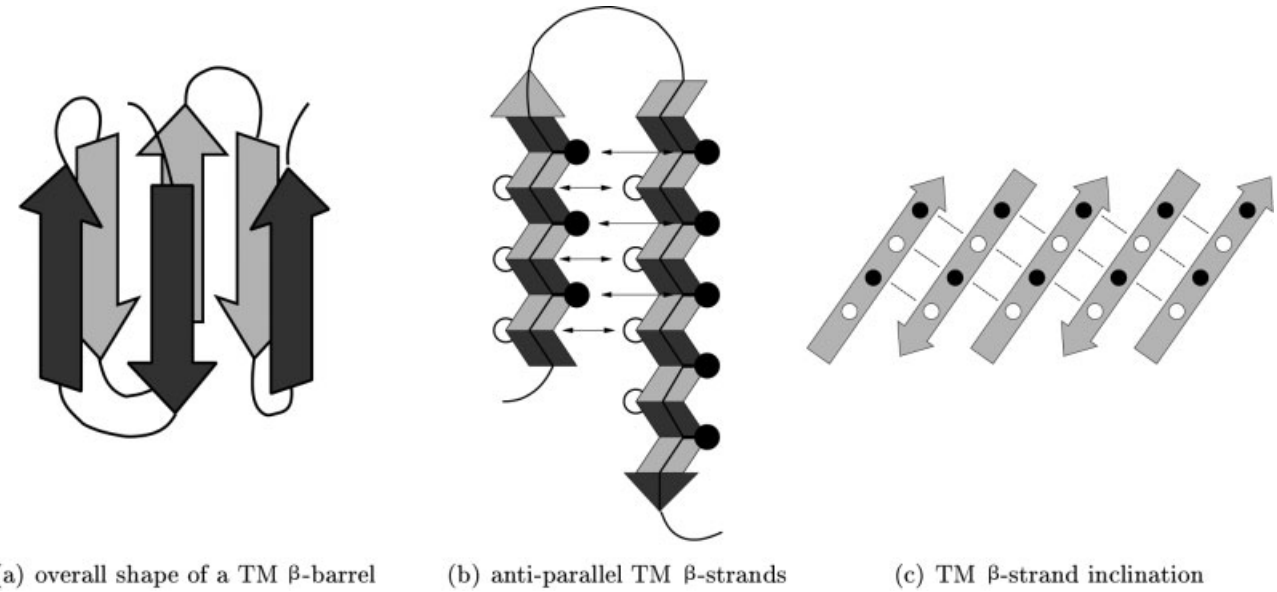


Fig. 1. Three different structural levels. **a**: Overall shape of the barrel: a barrel with 6 TM β-strands. **b**: Antiparallel strand pairing of TM β-strands with possible extension: sidechains in black are exposed to the core of the channel and those in white to the membrane. **c:** Tilt of TM β-strands through the membrane: H-bonds are represented by dotted lines and the shear number equals 1 in this example.

MATERIAL AND METHODS

Abstract Physical Model

Level decomposition of the structure

Transmembrane β-barrels can roughly be seen as channels formed from β-strands that are embedded in the outer membrane. Before precisely describing their structure, we focus on fundamental structural characteristics of TM β-barrels and their decomposition into well-defined levels of abstraction.

To a first approximation, a pore is built from an even number of β-strands, forming a single antiparallel β-sheet where the first and last strands are paired together. Figure 1(a) illustrates this configuration.

Note that β-strand lengths vary across the sequence. Hence, any formal description of antiparallel β-strand pairings should include (potential) strand extensions as shown in Figure 1(b). This description (channel and antiparallel pairing with extension) is insufficient to provide an approximate and *realistic* model of TM β-barrels. The main reason is that TM β-strands do not span the membrane at 90° (perpendicular to the membrane), but rather are usually inclined at an angle to the vertical TM axis. This implies a shift in the hydrogen-bonded (H-bonded) residues called the *shear number*. For instance, a shear number of +1 means that the H-bonded partner of the residue at position i is at position j +1 rather than j. An example is shown in Figure 1(c).

Combining these models allows us to accurately describe the most relevant characteristics of TM β-barrels. Note that, unlike the description made for a single TM α-helix in a TM α-bundle,[17] we do not give any description of local structure for single β-strands. In fact, such a description would be of no use because of the simplicity of β-strand hydrogen bonding, where side chains strictly alternate around the β-strand axis.

Strand inclination appears to be an intermediate description level. As well, strand extensions are essential for modeling at this level of abstraction. A schematic representation of a TM β-barrel is given in Figure 2(a), where the important features of strand extension and tilt are illustrated.

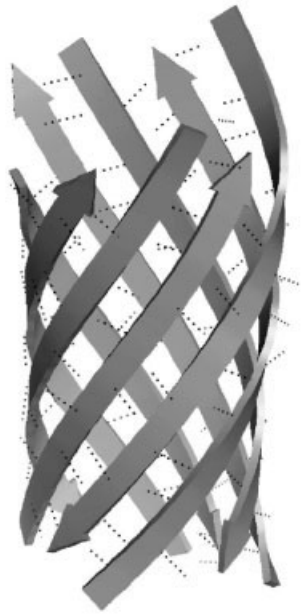
Features of transmembrane β-barrels

Most of the relevant features of TM β-barrels have been described in other articles.[5–7] For example, Schulz[7] gives a set of 10 construction rules (most of them confirmed in more recent reviews) that should be followed by all TM β-barrel proteins. These rules range from exact criteria (e.g., the number of β-strands is even) to somewhat less precise criteria (high sequence variability of external loop).

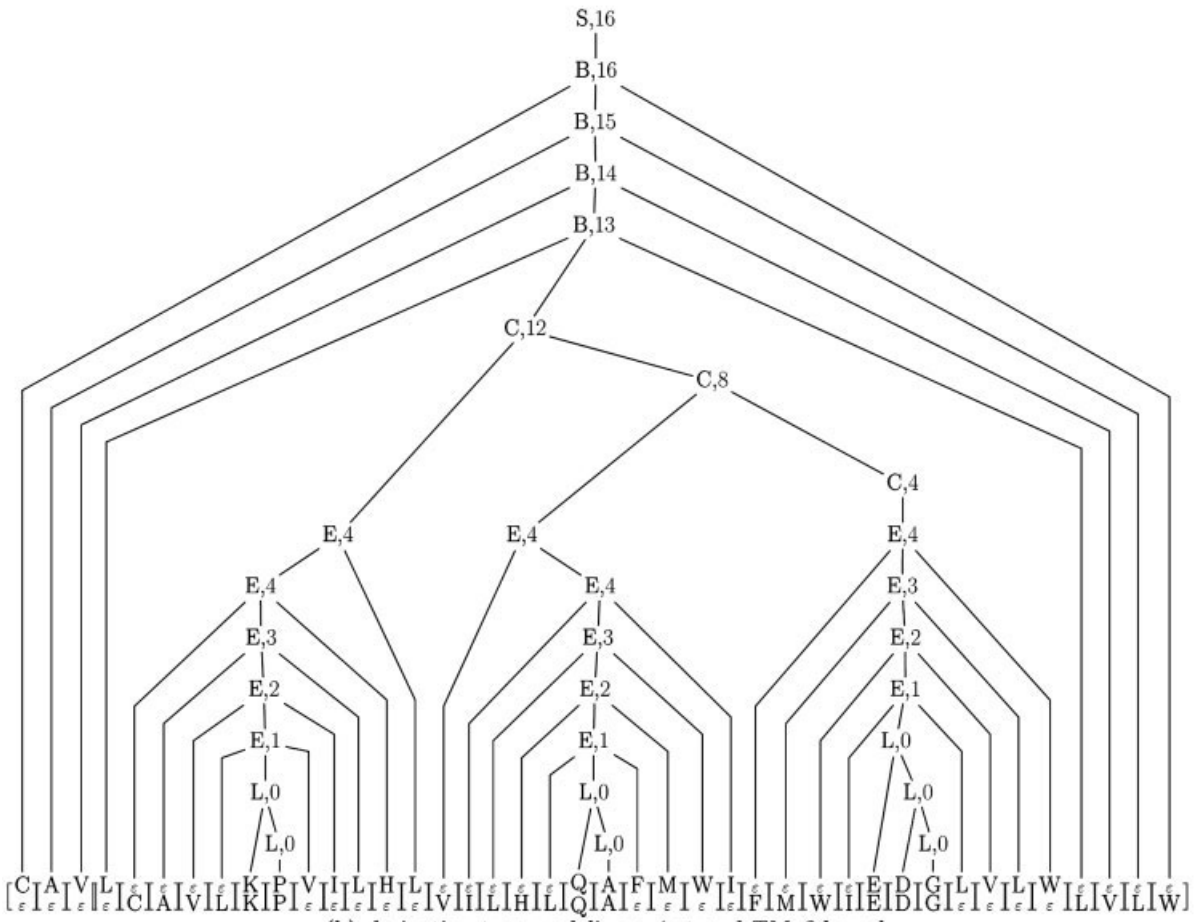
The first feature we add to our model is that the number of strands is even and that the N- and C-termini are located in the periplasmic end of the barrel. This implies that the two first β-strands are connected with an extracellular loop, and that the first and last strands are in opposite direction (with respect to the membrane) and paired in an antiparallel β-sheet.

According to Wimley,[6] the tilt of β-strands ranges from 20 to 45°. The shear number of an n-stranded barrel is thus positive and around +2. In the program *transFold*, we have implemented more flexible values, with a lower bound of +1 and upper bound of +3 (except for proteins with less than 200 residues, where a lower bound of n is set).

Strands are connected with tight turns on the periplasmic side (named T1, T2, and so on), while longer loops are observed on the extracellular side (denoted L1, L2, etc.). To deal efficiently with these features, the allowed size for periplasmic β-hairpins ranges from two to eight amino acids. In contrast, a lower bound of five residues is applied



(a) schematic representation of an 8-strand TM β-barrel



(b) derivation tree modeling a 4-strand TM β-barrel

Fig. 2. **a:** A schematic representation of an eight-strand β-barrel in our approximate physical model. H-bonds are represented by dashed lines. (Image obtained using PyMOL). **b**: A derivation tree modeling a four-strand β-barrel with pseudofolding energies stored in internal nodes. The nodes are labeled with a nonterminal determining the nature of the substructure and amino acids. Indeed, "L" represents a loop (periplasmic and extracelular), "E" represents antiparallel TM β-strand, "C" represents a TM β-strand sheet (i.e., antiparallel TM β-strands that form the unclosed β-barrel), "B" represents the closing antiparallel TM β-strand pairings, and "S" represents the whole omp structure. Here, the folding energy is simply computed as the number of interstrand residue contacts. Each nonterminal is associated with an attribute representing the energy of the substructure. Both examples describe strand extension and shear number.

TABLE I. Parameters Used for the Different Classes of Proteins

	Nonwater-filled (≤12 strands)				Water-filled (≥14 strands)	
	≤170	171–220	221–270	>270	≤320	>320
Min. length of TM strand	10	10	11	12	8	10
Max. length of TM strand	16	16	18	20	11	13
Min. shear number	0	0	1	2	1	1
Max. shear number	3	3	3	4	3	3
Min. length of periplasmic loop	2	2	2	2	2	2
Max. length of periplasmic loop	8	8	8	8	8	8
Min. length of extracellular loop	5	5	5	5	5	7
Min. hp of outward residues	0	0	0	0	0	0
Max. hp of inward residues	$+\infty$	$+\infty$	$+\infty$	$+\infty$	0	0
hydrophobic TM strands	no	yes	yes	yes	no	no

The type of constraints (physical or chemical) is listed in the left column. In this column, "hp" abbreviates hydrophobicity, while "outward residues" face the bilayer and "inward residues" face the channel interior. Central columns give values used for non water-filled channels, and the two rightmost columns give values used for proteins with water-filled channels. Each structural class is divided into subclasses with respect to the length of the sequence.

to extracellular loops. (Because loops do not contribute to the folding energy in our current model, we assign a larger value for the minimum size of an extracellular loop, to avoid ambiguity between periplasmic and extracellular turns. A threshold of five is chosen to avoid strongly structured hairpins, which should preferably occur in the periplasmic domain.)

Tamm et al.[5] note that TM β-strands span about 27–35 Å of the outer membrane. They also note a difference in their lengths in a monomeric versus multimeric protein; specifically they give an average length of 11 amino acids for trimeric porin, and 13 to 14 residues in monomeric β-barrels. The tilt of TM β-strands is generally higher in longer sequences; thus, the minimal number of residues required to span the membrane is increased.

Amino acids facing the bilayer are mostly hydrophobic, while those exposed to the channel interior are polar. However, the size of the channel varies over different TM β-barrel proteins, and the hydrophobicity of the channel interior can thus change. For example, porins have a large water-filled channel while TM β-strands of OmpA in *Escherichia coli* are tightly packed. It follows that the presence of polar residues is most important (and favored) for large pores. Because the size of the channel is naturally related to the number of strands involved, we associate these parameters together.

We then apply (slightly) different criteria according to the number of TM β-strands and the length of the protein. When the number of strands is greater than or equal to 14, we constrain the profile of residues exposed to the channel interior to be hydrophilic on each strand. Otherwise, when there are at most 12 strands, this constraint is not applied. Nevertheless, in that case, to compensate for the lack of selection constraints on TM β-strands for smaller sequences (≤170 residues), a hydrophobic profile is required for the membrane-exposed face of each TM β-strand. (This constraint has the advantage of allowing polar side chains to point toward the channel interior while at the same time limiting the overall polarity of channel interior.)

Similarly, the extracellular loops are mostly composed of polar residues. Hence, using a polarity scale,[23] we constrain these segments to be polar. Extracellular loops are also known to be quite flexible. However, at this point, this criterion seems too inexact to be used reliably. A strategy that includes the flexibility scale[24] could be used, but is intentionally omitted in the current article. High sequence variability of extracellular loop regions prevents any simple measure from being applied.

An overrepresentation of aromatic residues in found in two rings contacting the bilayer, one at either end of the β-barrel. Nevertheless, our statistical analysis of the amino acid location propensity (data not shown) does not reveal any clear signal. Some preferences might be found in subsequent analysis, but we did not succeed in designing any simple and reliable criterion.

From the preceding discussion, it emerges that the physical parameters are strongly correlated with the type of pore (water-filled or not), the length of the sequence and number of TM strands. For example, a TM β-barrel with 18 TM strands should have a water-filled pore, while a small protein with eight TM β-strands will have a less polar environment in the channel interior and its β-strands should be less inclined (lower shear number). All parameters used in this article are reported in Table I for each type of protein. Note that certain constraints have been applied to large proteins to reduce the complexity of parsing for the underlying multitape S-attribute grammar. For example, the minimum number of residues in extracellular loops is required to be seven residues for the largest proteins (more than 320 amino acids and more than 12 TM strands). In contrast, in the case of the smallest proteins, the lower bound for the shear number is relaxed to 0 and there is no overall hydrophobicity requirement (i.e., the requirement stipulating that TM β-strands are overall hydrophobic has been removed; however, we maintain the hydrophobicity requirement that side chains facing the bilayer are hydrophobic and that side chains facing the channel interior are polar).

Folding energy

It is now well known that the TM β-barrel folding process passes through different states (unfolded, molten

disk, molten globule, and native state) where long-range interactions play a dominant role.[5] Hence, a realistic energy function must be based on, or at least include, these terms.

As suggested in Figure 1, our model can describe side-chain interactions occurring between TM β-strands. [Because only local information can be used, classical hidden Markov models (HMM) and neural networks (NN) are intrinsically unable to deal with such long-range interactions.] Nevertheless, thermodynamic driving forces are still not fully understood,[1] and residue contact energy between TM β-strands remains obscure. Some values have recently been computed;[25] however, we deliberately do not use these values to ensure that our energy parameters are not learned from the dataset on which we make our predictions. As previously explained, very few TM β-barrel structures are actually known; this is a major objection against current machine-learning methods (see Ref. 10).

The BETAWRAP authors[19,20] designed an elegant strategy to compute pairwise probabilities according to their environment but independently of the context in which they are used. These values were computed using a large set of globular proteins (no TM proteins) whose tertiary structure is known. They distinguish contacts occurring in a hydrophilic environment (exposed at the surface) from those occurring in a hydrophobic environment (buried in the core of the protein), thus yielding two distinct tables, which give the contact probability according to the milieu.

Tamm et al.[5] noted that the general distribution of residues in TM β-barrel proteins is inverted with respect to that of most soluble proteins. In soluble proteins, hydrophobic amino acids are generally found buried in the core, while in TM β-barrel proteins, polar residues face the core, which is the channel interior (sometimes water-filled, and usually polar).

In our program *transFold*, we used the values provided by Refs. 19 and 20, turning them into an energy potential using the standard procedure (taking the negative logarithm of the frequencies (see pp. 223–228 of Ref. 26 for details). Finally, the folding pseudoenergy of the structure is simply computed as the sum of all contact potentials.

In our current model, extracellular loops and β-hairpins do not contribute to the folding pseudoenergy for several reasons. First, even for simple configurations, we do not have any reliable energy function. Second, extracellular loops display great structural diversity and sometimes contain strongly structured subdomains with important energy contributions that are difficult to determine. Finally, the goal of the current article is to better understand the effect of long-range interactions between β-strands in TM β-barrel folding.

Implementation

Grammatical modeling

In analogy to Ref. 17, we transcribe the previous description into a multitape S-attribute grammar (MTSAG). This approach is motivated by two reasons. First, when used in a machine learning context, this formalism generalizes other standardized approaches such as hidden Markov models (HMMs) and stochastic context free-grammars.[27] (We emphasize that our use of multitape S-attribute grammars is to permit a unified description of all possible supersecondary structures for a β-barrel protein, after which we apply dynamic programming, rather than machine learning.) Second, building on Lefebvre's work, we are able to use a meta-parser (mtsag2c) that allows us to more easily produce our *transFold* software. Moreover, modeling with multitape S-attribute grammars provides us with an elegant way to describe the structure of TM β-barrel proteins and suggests applications in future combinatorial studies.

Compared with other grammars used in computational biology (e.g., general tree grammars), the MTSAG formalism has two clear advantages. First, the expressive power of the language is better suited to the complexity of the structural description, and it offers faster parsing in practice. Second, attributes in a multitape S-attribute grammar allow us to unify the energy model together with a protein structural description.

The modeling operates in two steps. The structural description is first transcribed as a multitape context-free grammar (MTCFG). Then, we define an attribute system that computes the folding pseudoenergy of a TM β-barrel, and we associate to each production rule an attribute function.

We begin by describing each structural level with its own grammar. The translation of the overall structure of the TM β-barrel is identical to that given in Ref. 17 for the TM α-bundle. This MTCFG is called G_{barrel}, and we display its production rules in Figure 3. Similarly, we model antiparallel pairings with the grammar G_{couple} in Figure 3. In this case, a classical context-free grammar (CFG) suffices for its description. (Note that the decomposition of the structures follows the hierarchy of the languages. This could be used as a good estimator of the complexity of protein structure.)

We compose G_{barrel} with G_{couple}, resulting in a grammar which describes all TM β-barrel features except strand inclination. To restrict our model to a valid β-barrel (with a correct shear number), we constrain strand extensions. Extension of the left strand with unpaired amino acids (at most three, as explained earlier) is then required (cf. production rules $S_{\text{extend}} \rightarrow S_{\text{extend}} \cdot$ of G_{couple} in Fig. 3) when β-strands are connected with an extracellular loop. Extension of the right strand is then disallowed in such cases.

The way the grammar G_{barrrel} operates to decompose a barrel is illustrated in Figure 4(a). In these structures, each TM β-strand is paired twice: with its left and right neighbors except for the last strand, which is paired with the first strand to close the channel. However, classical linguistic tools only allow single pairings. The key idea is to duplicate the structure and make associations only from the copy to the original structure. This justifies the use of a two-tape grammar. Hence, it becomes feasible to isolate each antiparallel pairing involved in the barrel into individual blocks.

To proceed, the grammar generates two-tape words of the form given in Figure 4(b). Here, "ε," which could also be denoted "ϵ," represents the empty character, and "(" or ")"

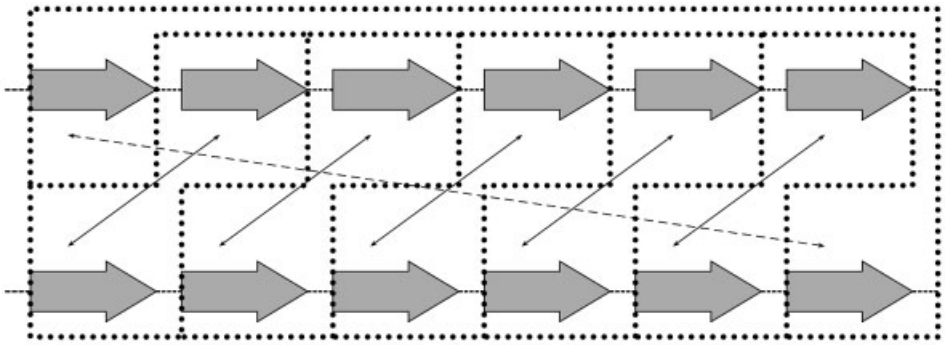
$$G_{barrel} = \begin{cases} S_{barrel} \rightarrow \left[\begin{smallmatrix}\bullet\\ \varepsilon\end{smallmatrix}\right] S_{barrel} \left[\begin{smallmatrix}\varepsilon\\ \bullet\end{smallmatrix}\right] \mid B_{seq} \\ B_{seq} \rightarrow B_{pair}\ B_{seq} \mid B_{pair} \\ B_{pair} \rightarrow \left[\begin{smallmatrix}\varepsilon\\ \bullet\end{smallmatrix}\right] B_{pair} \left[\begin{smallmatrix}\bullet\\ \varepsilon\end{smallmatrix}\right] \mid Loop \\ Loop \rightarrow \left[\begin{smallmatrix}\bullet\\ \bullet=1\end{smallmatrix}\right]\ Loop \mid \left[\begin{smallmatrix}\bullet\\ \bullet=1\end{smallmatrix}\right] \end{cases}$$

(a) Grammar modeling the overall shape of TM β-barrels

$$G_{couple} = \begin{cases} S_{couple} \rightarrow \bullet\ S_{reduce} \mid S_{extend}\ \bullet \mid S_{pairing} \\ S_{reduce} \rightarrow \bullet\ S_{reduce} \mid S_{pairing} \\ S_{extend} \rightarrow S_{extend}\ \bullet \mid S_{pairing} \\ S_{pairing} \rightarrow S_{channel} \mid S_{memb} \\ S_{channel} \rightarrow \bullet\ S_{memb}\ \bullet \mid Loop \\ S_{memb} \rightarrow \bullet\ S_{channel}\ \bullet \mid Loop \\ Loop \rightarrow \bullet\ Loop \mid \bullet \end{cases}$$

(b) Grammar modeling anti-parallel β-strands

Fig. 3. Grammatical modeling of the structural levels. Grammar G_{barrel} **a**: models the overall structure of TM β-barrel and G_{couple} **b**: the antiparallel pairings of TM β-strands with possible extension on the left (reduction: decreases length) or on the right (extension: increases length). Token ● represents any amino acid, while ● = 1 means the same amino acid as in the upper tape (see ref. 17 for technical details concerning upper and lower tape for multitape S-attribute grammars). "ϵ" represents the empty character. Grammatical modeling of strand inclination (i.e., shear number) results from combination of these two grammars.



(a) decomposition of a TM β-barrel

```
))))))εεεεεεiii))))))εεεεεεoo))))))εεεεεεii))))))εεεεεε
εεεεεε((((((iiiεεεεεε((((((ooεεεεεε((((((iiεεεεεε((((((
```

(b) 2-tape string representation of a 4-strand TM β-barrel

```
))))))iii))))))oo))))))ii))))))
((((((iii((((((oo((((((ii((((((
```

(c) 2-tape string representation after ε-suppression

Fig. 4. β-Barrel decomposition and its description as a two-tape word. **a**: Decomposition of the β-barrel into individual subunits. Each block is associated to a single β-strand antiparallel pairing. **b**: Twp-tape string modeling a four-strand β-barrel. The string includes the empty character "ε." **c**: Same structure after "ϵ"-suppression.

represent any amino acids brought into contact by the pairing. Unlike classical one-tape words, a two-tape word results from the concatenation of two-character letters. The most common way to represent these is to juxtapose characters (tokens), one on top of the other [see Fig., 4(b) and (c)]. Finally, an operation called ϵ-suppression removes all empty tokens to obtain the resulting word [see Fig. 4(c)], which represents the decomposition illustrated

in Figure 4(a). From this discussion it emerges that the parsing process (i.e., structure modeling) consists of inserting empty characters at correct positions in the original two-tape string.

The grammar G_{couple} operates in a more traditional manner. Pairs of amino acids are progressively added at each extremity of the string; in addition, amino acids can be added to only one side, to reflect the fact that beta strands of unequal lengths may be paired. The combination of grammars G_{barrel} and G_{couple} results in a system that respects both properties; that is, a system that produces and concatenates two-tape blocks following the pairing rules of TM strands.

To achieve the modeling and associate its folding energy to a given structure, we associate an attribute function to each production rule, and achieve the construction of the MTSAG G_β for TM β-barrels. Because they are the only productions that describe amino acids contacts, only production rules with S_{channel} and S_{memb} on the left-hand side contribute to the final folding pseudoenergy value. Obviously, productions $S_{\text{channel}} \rightarrow \cdot\, S_{\text{memb}} \,\cdot$ use the values computed for interactions occurring in a hydrophilic environment (residues exposed to the channel interior), while productions $S_{\text{memb}} \rightarrow \cdot\, S_{\text{channel}} \,\cdot$ use the values computed for interactions occurring in a hydrophobic environment (side chains facing the membrane core). An example of the modeling of a complete structure of a TM β-barrel with its folding pseudoenergy is given in Figure 2(b). The structure description (including strand extensions and shear number) is unified with the energy value in a unique derivation tree, representing the sequences of productions used to generate the structure. More details can be found in ref. 17.

Software

Our software, which implements the method described above, is called *transFold*. Results reported in this article have been computed with a Xeon 2.4-GHz dual processor with 2 GBytes running Fedora 3.0 Linux. The theoretical upper bound for time complexity is $O(n^6)$ (see refs. 17 and 27); however, in practice, the time complexity appears to be $O(n^3)$ or $O(n^4)$ (this assertion should be provable by careful mathematical analysis). The same remark holds true for space requirements where a complexity of approximately $O(n^3)$ is observed instead of the theoretical bound of $O(n^4)$.

A run on a sequence of 300 residues typically uses around 1 min of CPU time and 1 GB of memory. However, performance may vary according to the values assigned to structural constraints (e.g., the values defined in Table I). A web server has been implemented and is available at http://bioinformatics.bc.edu/clotelab/transFold/. and http://theory.csail.mit.edu/transFold.

Datasets

Due to the paucity of experimental data, relatively few sequences and structures are available to evaluate our structure predictions. Nevertheless, many different representative datasets (often similar) have already been used. In most cases, because all other existing methods are based on learning methods, these datasets have been designed to avoid redundancy and sequence similarity. Note that because our energy parameters come from pairwise frequencies in *globular* proteins, we need not be overly concerned with redundancy concerns necessary in a machine learning context. Nevertheless, for comparative purposes, we consider (an updated version of) the same datasets tested in previous machine-learning studies.

The two most relevant datasets have been proposed by Martelli et al.[8] and Bigelow et al.[10] The first dataset (denoted setMartelli) is composed of 15 structures filtered at 30% sequence identity (pdb id: 1A0S, 1BXW, 1E54, 1EK9, 1FCP, 1FEP, 1I78, 1K24, 1KMO, 1PRN, 1QD5, 1QJ8, 2MPR, 2OMF, 2POR), while the second dataset (denoted setTMBcomp) is smaller but is computed with more sophisticated techniques (PDB id: 1A0S, 1AF6, 1BT9, 1FEP, 1PRN, 1QD5, 1QJ9, 1QJP). Both are used in the current article.

We updated these two sets, by taking structures available in the PDB[28] release of October 2005, filtered at 30% of sequence identity. As was done in previous work,[10] we removed structures that do not fit into our description and present atypical structural features (e.g., β-barrels formed from different chains, or the presence of an important plug domain inside the pore). (These structures were removed mainly because the particular interactions, which occur between β-barrel side chains pointing toward the channel and those residues of the plug domain, are not yet described by our energy model). Moreover, due to the complexity of parsing and the reliability of the predictions, we restrict the sequence length to at most 500 residues. Because our prediction algorithm is based on long-range interactions, the combinatorial complexity increases as sequence length increases. It follows that the performance of the method is strongly correlated with input size. For this reason, we bound the length of input sequences to 500 residues, which seems to be a fairly reasonable upper bound for transmembrane proteins. Note that HMMs and neural networks are not concerned by sequence length restrictions, because they apply only local information. Secondary structure assignment has been made by DSSP[29] and is manually adjusted. (The PDB structure was examined to suppress gaps in TM β-strands or adjust strand caps.)

Our final dataset is then composed of 14 structures, divided into three different subsets according to sequence length and channel type. Proteins with a nonwater-filled pore are referred to as NWF. This set is divided into sequences having less than 200 residues (denoted NWF^-) and more than 200 residues (denoted NWF^+). Proteins with a water-filled channel belong to a dataset denoted WF. Formally, NWF $\leq$200 consists of (PDB id) 1QJP, 1QJ8, 1THQ, 1P4T, while NWF $>$200 consists of 1I78, 1K24, 1QD6, and WF of 1A0S, 1AF6, 1PRN, 2OMF, 1E54, 1TLY, 2POR.

With the same restrictions, the dataset setMartelli is restricted to 11 structures and setTMBcomp to seven structures.

Whole protein prediction and structural classification has been performed on different datasets provided by Gromiha and Suwa.[9] Sequences have also been filtered

according to their length. The remaining database contains 439 globular proteins (dataset noted GLOB), 162 helical inner membrane proteins (noted TMH), and 151 outer membrane proteins sometimes annotated as probable (noted UNK). Sequences belonging to NWF or WF are obviously removed from UNK.

Evaluation

Several different scores are usually used to evaluate the prediction accuracy. Standardized benchmarks have been established,[30,31] and roughly similar benchmarks are found in other articles. Per-segment and per-residue accuracy are the main features to consider, for which *sensitivity* and *specificity* are the two most commonly used measures.

Sensitivity gives the rate of correct prediction over the observed structure (i.e., percentage of true structure that is correctly predicted), and specificity gives the rate of correct prediction over the predicted structure (i.e., percentage of predicted structure that is true). In the context of secondary structure assignment, sensitivity of β-strand residue assignment is denoted by $Q_{TM}^{\%\mathrm{obs}}$ and specificity by $Q_{TM}^{\%\mathrm{pred}}$. Additionally, for nontransmembrane residues, we have respectively $Q_N^{\%\mathrm{obs}}$ and $Q_N^{\%\mathrm{pred}}$. Let $X \in \{TM,N\}$ be an assignment (or state). Then, formally: $Q_X^{\%\mathrm{obs}} = 100 \times$ number of residues correctly predicted in state X/number of residues observed in state X; $Q_X^{\%\mathrm{pred}} = 100 \times$ number of residues correctly predicted in state X/number of residues predicted in state X.

These scores are combined into a single score giving the rate of correct assignment: $Q_2 = 100 \times$ number of residues correctly predicted in the protein/number of residues in the protein.

Sensitivity and specificity for per-segment accuracy come from the same formulas. In that case, a segment is correctly predicted if the observed segment intersects one and only one predicted segment, and vice versa. In this study, we define intersection as an overlap of at least four amino acids. By $Q_\beta^{\%\mathrm{obs}}$ and $Q_\beta^{\%\mathrm{pred}}$, we denote the sensitivity and specificity of TM β-strand segments.

In addition to these classical scores, we consider the score Q_p, which computes the percentage of structures *correctly predicted*, where a structure is said to be correctly predicted if and only if all observed (respectively, predicted) β-strands intersect one and only one predicted (respectively observed) strand. We additionally consider the score (Q_p^{almost}), which computes the percentage of structures *almost correctly predicted*. Here, a structure is said to be almost correctly predicted if and only if each observed (respectively predicted) TM β-strand intersects *at most* one predicted (respectively observed) β-strand.

We also computed Matthew's correlation coefficient (MCC), to estimate per-residue accuracy. More details can be found in ref. 10.

We evaluated the accuracy of our contact predictions and compared them with DSSP[29] annotations. For this purpose we need first to define the notion of a *compatible pair of residues*. By *contact pair*, we mean a pair (i,j) of indices such that the residue at position i is predicted to be hydrogen bonded to the residue at position j. If δ is a given integer, then contact pairs (i,j) and (m,n) are said to be *compatible* if $(i,j) = (m \pm \delta, n \pm \delta)$. In our context, because we want to consider *only* residues with the same orientation in the barrel (e.g., side chains pointing toward the membrane or pointing toward the cavity), we choose δ to be equal to 2.

Hence, we extracted H-bonded pairs of residues from PDB files. Then, we considered a contact prediction as correct if a compatible pair is found in the observed contact list. We completed the sensitivity ($Q_{ct}^{\%\mathrm{obs}}$) and specificity ($Q_{ct}^{\%\mathrm{pred}}$) scores as defined above.

RESULTS AND DISCUSSION

Using *transFold* with parameters defined in Table I, we determine the TM β-barrel structure with the best pseudo-folding energy. In the following, we evaluate the reliability of the predictions, compare the results with two of the best current methods, and estimate the accuracy to discriminate transmembrane β-barrel proteins from other proteins. Finally, we perform in silico folding variants of the Outer membrane protein of *Escherichia coli* (OmpA), and reproduce experimental results.

Evaluation of Structure Predictions

TransFold's favorable performance when compared with existing methods for TMB prediction can be seen in Table II. As previously mentioned, there is a breakdown into different subsets according to the length and the type of protein. Sequences in NWF range in length from 148 to 297, while sequences in WF range in length from 289 to 421. As expected, Table II displays a dependence between input size and performance. Nevertheless, the performance rate obtained for proteins in NWF remains strong even for the longest sequences, where performance seems to be less affected than the performance for proteins in WF. Some structural features occurring in porins (all proteins in WF belong to the porin family) are in fact not described by our grammar; for instance, small helices and β-sheets located in extracellular loops are not represented in our current grammar. As stated in ref. 17, *the reliability of prediction is closely related to the ability of the grammar to describe important features in the native structure*.

The most exciting observation is that our prediction of TM β-strands is very accurate. Focusing on sequences in NWF, all TM β-strands were correctly predicted with no overprediction. This means that for all proteins in NWF the correct shape of the barrel has been found. According to the current consensus folding model,[5] long-range interactions drive the folding of TM β-barrel proteins. Because our predictions are based on these long-range interactions, our prediction rates are consistent with this theory, which explains the excellent results in Table II.

TM β-strand prediction for the porin family is less accurate, but still remains good in terms of per-segment accuracy. Given the percentage of almost predicted structures and the high sensitivity score, we see an overprediction of TM segments rather than a misprediction. In fact, long extracellular loops tend to cause such overpredictions. Because no constraint has been imposed on these regions, our algorithm tries to insert extra β-strands. This situation is expected to be corrected when additional

TABLE II. Prediction Accuracy

	Topology	Strands		2-states	TM residues		non-TM residues			Contact	
	Q_p	$Q_\beta^{\%\,obs}$	$Q_\beta^{\%\,pred}$	Q_2	$Q_{TM}^{\%\,obs}$	$Q_{TM}^{\%\,pred}$	$Q_N^{\%\,obs}$	$Q_N^{\%\,pred}$	MCC	$Q_{ct}^{\%\,obs}$	$Q_{ct}^{\%\,pred}$
NWF$^-$	100 (100)	100	100	84.81	92.52	86.26	68.14	80.81	0.64	83	65
NWF$^+$	100 (100)	100	100	75.57	80.40	81.20	67.02	65.86	0.48	48	44
NWF	100 (100)	100	100	79.72	86.05	83.66	67.48	71.43	0.55	64	55
WF	0 (71.4)	92.0	78.0	63.97	76.42	64.95	48.39	62.12	0.30	32	23
WF*	85.7 (100)	99.1	99.1	78.04	78.92	81.07	76.94	74.47	0.56	51	45
All	50.0 (85.7)	94.9	85.2	69.91	80.44	72.16	54.44	65.47	0.38	45	35
All*	92.9 (100)	99.4	99.4	78.68	81.90	82.19	73.95	73.57	0.56	56	49

Q_p is the rate of correctly predicted structures, while the rate of almost predicted structures Q_p^{almost} is given in parenthesis. $Q_\beta^{\%\,obs}$ and $Q_\beta^{\%\,pred}$ represent the sensitivity and specificity, respectively, of TM β-strand predictions. Q_2 is the rate of correct secondary structure assignment, $Q_{TM}^{\%\,obs}$ and $Q_{TM}^{\%\,pred}$ are the sensitivity and specificity of assignment, respectively, for β-strand residues, and $Q_N^{\%\,obs}$ and $Q_N^{\%\,pred}$ are similarly sensitivity and specificity, respectively, for non-TM residues. Matthew's correlation coefficient (MCC), as well as the sensitivity and specificity of contact predictions are given in last columns. NWF contains omps with nonwater-filled channels. It is divided into two datasets NWF$^-$ and NWF$^+$ for small proteins (≤200 res.) and larger proteins (>200 res.). WF is the dataset of proteins with a water-filled channel (porin-like). The rubric "All" indicates that the score is for the complete dataset (NWF∪WF). Rubrics with an asterisk (*) indicate results for datasets computed with constraints for unfolded subsequences (portions of loop regions where a significant structural motif has been observed).

constraints are applied in a future extension of our current grammar, in particular by including a description for small helices and β-sheets located in extracellular loops.

To confirm this hypothesis, we have constrained extracellular regions to be free of secondary structure (helices and strands), whenever a structural motif is observed in the PDB structure; for example, in the long loop L3 of porins, where the chain folds back into the lumen of the pore. By this procedure we want to: (1) add criteria for the selection of loop regions; (2) show evidence of additional structural stability caused by structured loop regions.

Results obtained with the constraint that β-strands not be allowed in loop regions are indicated in the line with rubric WF* in Table II, and additionally occur in the line with rubric All*.

The scores thus obtained show a clear increase in prediction accuracy, reaching the rates found for the WT set. From these observations, we conclude that specific constraints applied to extracellular loops should allow us to improve accuracy of our method for porins. Taking into account the flexibility of potential loop regions[24] could be a good starting point to improve the prediction of these loops. (This suggestion follows one of the rules given by Schulz;[7] however, an additional reason to consider this point is that it is believed that functional regions, mostly located in such loops, are flexible.)

Structural motifs occurring in extracellular loops play a fundamental role in structure stability and/or function. For example, the loop L2 in porins is known to contribute significantly to the trimer stability, while loop L3 has hydrophobic contacts with residues inside the barrel.[5] Because the stability of TM β-barrel proteins is rather modest (<10 kcal/mol in general), it follows that the contribution of these substructures cannot be neglected. The (grammatical) description of these structural subunits, as well as their contribution to the folding pseudoenergy, should improve prediction rates. In a sense, the behavior of our model confirms experimental observations.

Finally, the comparison of our contact predictions with H-bonded pairs of residues yields satisfying results (particularly good for small proteins), especially if we consider the difficulty of predicting them directly from sequence. Several reasons explain the apparently weak values of sensitivity $Q_{ct}^{\%obs}$ and specificity $Q_{ct}^{\%pred}$ observed for larger proteins: first, H-bonds constitute a subset of interactions occurring between residues of different strands. Moreover, the number and location of H-bonds may vary between different experimentally determined structures (e.g., 3D coordinate files) for the same protein. Hence, some contacts could have been simply missed by experimental annotations.

One can note that local features such as gaps or bulges in H-bond patterns between two paired TM β-strands are not yet modeled in *transFold*. These alterations of the structure could possibly distort the contact map (local perturbations of the structure can have a global effect). Because our current measure for the correctness of a contact does not allow us to capture this feature, the accuracy of predicted contacts is then affected.

Nevertheless, the accuracy of contact prediction is particularly good for small proteins (less than 200 residues; see supplementary data for detailed results). Hence, *transFold* could prove to be useful in annotation or reannotation efforts, specifically by indicating contacts that have been missed using classical methods. In particular, the use of *transFold* predictions together with NMR experiments (e.g., residue interactions experimentally observed) could result in a more accurate method for building complete 3D model of proteins.

Comparison with Existing Methods

To compare our method with existing methods, we ran *transFold* on the setMartelli dataset as well as setTMB-comp. According to the performance published in the literature, HMMs show better accuracy than neural network methods. For this reason, we compare *transFold* with the state-of-the-art HMM methods. Prediction rates for the HMM of Martelli et al. and for PROFtmb were obtained respectively from refs. 8 and 10. These references reported the scores Q_β (segment overlap), Q_2 (correct secondary structure assignment), $Q_{TM}^{\%obs}$ and $Q_{TM}^{\%pred}$ (sensitivity and specificity of TM β-strand residue assign-

TABLE III. Comparison of Prediction Accuracy of *transFold* with Other Methods

Method	Q_β	Q_2	$Q_{TM}^{\%\,obs}$	$Q_{TM}^{\%\,pred}$	MCC
SetMartelli					
transFold	93	70	81	70	0.35
*transFold**	99	79	83	81	0.56
Martelli	94 ± 11	84 ± 6	80 ± 14	87 ± 9	0.69 ± 11
PROFtmb	93 ± 11	83 ± 6	80 ± 14	87 ± 9	0.69 ± 11
SetTMBcomp					
transFold	93	70	82	70	0.37
*transFold**	100	80	84	82	0.58
PROFtmb	93 ± 11	83 ± 6	85 ± 14	87 ± 9	0.70 ± 11

Performance of the HMM of Martelli et al. on the dataset setMartelli is reported in Ref. 8 and performance of PROFtmb on the dataset setTMBcomp is reported in Ref. 10 Q_β is the percentage segment overlap between predicted and observed TM β-strands, Q_2 is the rate of correct secondary structure assignment, and $Q_{TM}^{\%obs}$ and $Q_{TM}^{\%pred}$ are respectively sensitivity and specificity of assignment of β-strand residues, and MCC is Matthew's correlation coefficient. Results marked with report scores computed on sequences with constrained unfolded subsequences (parts of loops where a significant structural motif has been observed).

ment), and MCC (Matthew's correlation coefficient); however, the rate of correctly predicted topology Q_p was absent from both original papers (although it could be that the score Q_p is related to Q_β).

Results are presented in Table III. Two scores for our software *transFold* are reported, denoted respectively by *transFold* and *transFold**. The latter (with asterisk) employs the additional constraint explained in the example given for the L3 loop in porins (i.e., we prevent TM strand formation in extracellular loop regions), while the former (without asterisk) imposes no such additional constraint.

Because PROFtmb as well as the HMM of Martelli et al. are machine-learning methods, which thus learn their parameters from known TM β-barrel structures, their reported accuracy is associated with a standard deviation. Scores reported in the literature are the best encountered in bootstrap experiments, and hence, should be considered to be an upper bound for the actual performance. *TransFold* is not subject to this caveat, because its contact potentials were computed from globular proteins.[19,20]

Although both HMMs give similar performance results, the scores obtained with *transFold* differ markedly. Our method significantly improves the prediction of TM β-strands; however, the per-residue prediction rate is lower. As stated above, long-range interactions play a fundamental role in the formation of TM β-strands; thus, it is not surprising that *transFold* detects TM β-strands that "local" methods have missed. It follows that *transFold* is a more reliable tool to predict the overall structure of the barrel and all transmembrane β-strands.

Nevertheless, the secondary structure assignment in *transFold* is not as accurate as that of the HMMs from refs. 8 and 10, because *transFold* has used no local information, with the exception of hydrophobicity, to evaluate the ability of a given residue to fit well in the local context. Clearly, classical learning methods (HMMs and NNs) implicitly make use of hydrophobicity, in addition to sequence information. For instance, the HMM method PROFtmb,[10] carefully describes each position that can be taken by an amino acid in the structure (periplasm, extracellular milieu, membrane core, or interfacial bilayer). It follows that given the simplicity of our local description of TM β-strands, it is not surprising that HMM methods outperform *transFold* for per-residue accuracy. Nevertheless, if we consider the standard deviation associated with learning methods, the rates of *transFold* are still in the range of those given for existing methods.

We should note that the inaccuracy of *transFold* in a secondary structure assignment is most visible in the performance on porins (water-filled barrel structures). For proteins having less than 300 residues, *transFold* obtains good secondary structure assignment scores, comparable with the best scores observed for the HMMs of Martelli et al. and of PROFtmb; and even better if we consider proteins having less than 200 residues. Results for *transFold** show an improvement over *transFold* in the case of porins, where in this case we have added the constraint that no β-strand can occur in extracellular loop regions.

From this discussion, it emerges that the techniques of HMM and *transFold* appear to be complementary. Because MTSAG generalizes the HMM method,[27] we can expect that by merging these methods, one might obtain a significant performance increase in TM β-barrel prediction.

Outer Membrane Protein Discrimination and Structural Classification

In this section, we evaluate the ability of *transFold* to discriminate outer membrane protein (omps) from globular and inner membrane proteins. In addition to this classification, we propose a method to classify omps according to structure (e.g., water-filled channel or nonwater-filled channel).

To discriminate proteins we used four parameters: sequence length, folding pseudoenergy in the nonwater-filled model (NWF model), folding pseudoenergy in the water-filled model (WF model), and overall hydrophobicity of the sequence. The folding pseudoenergies in both models are required because constraints differ significantly (see earlier); hence, the signal is different. Overall hydrophobicity is used to estimate the propensity of the sequence to be affected by strand selection constraints.

TABLE IV. Evaluation of Protein Discrimination

	ROC	TP	TN	FP	FN
TM vs Glob	0.81 ± 0.03	88 ± 2	63 ± 4	37 ± 4	12 ± 2
NWF (nonporin)	0.78 ± 0.03	83 ± 9	65 ± 6	35 ± 6	17 ± 3
WF (porin)	0.93 ± 0.02	95 ± 7	75 ± 8	25 ± 8	5 ± 7

The first line gives the performance for basic discrimination between omp and nonomp proteins. The second and third lines give the scores obtained for structural classification, where NWF refers to *nonwater-filled channel* (nonporin) and WF refers to *water-filled channel* (porin-like proteins).

Small changes in selection constraints (strand length, shear number, hydrophobicity, etc.) may have an impact on the homogeneity of the folding pseudoenergy values. We use the same settings (one for the NWF model and another one for the WF model) for all proteins regardless of their length. These parameters are determined by taking the largest constraints defined in Table I. Nevertheless, using original settings (cf. Table III) does not significantly change the scores obtained (data not shown).

Discrimination was performed using a support vector machine (SVM). For this purpose, we used the S. Noble and P. Pavlidid software Gist (http://svm.sdsc.edu). For each protein, we computed a *feature vector* $\vec{x} = (x_1,x_2,x_3,x_4)$, where x_1 is protein length, x_2 is average hydrophobicity, x_3 is the energy according to the nonwater-filled (NWF) model, x_4 is the energy according to the water-filled (WF) model. Using the default kernel of Gist, support vectors were computed on a randomly chosen training set composed of 85% non-omp proteins (GLOB and TMH datasets) and 50% TM β-barrel proteins (UNK dataset). Sequences not chosen in the training set were used to compute performance. This procedure was iterated 1000 times, to obtain a reliable estimate of classification performance.

Because two folding pseudoenergies (in NWF and WF models) are computed for each sequence, this can be done with an appropriate labeling to distinguish between both types of structure. [Only proteins of a given type (NWF or WF) are labeled as positive. The standard procedure is then applied in the same way as previously described.] We are then able to make *automatic structural classification* of omps.

Results from this benchmark are given in Table IV. Five scores are computed. We computed the receiver operating characteristic (ROC) for the test set, and report the normalized area under the ROC curve, which graphs true positive rate as a function of false positive rate. Perfect classification corresponds to a ROC area of 1.0. We also compute rates of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The mean and standard deviation for each of these scores is given as computed from 1000 runs.

Study of OmpA of *Escherichia coli*

To show the biological significance of *transFold*, we compared our predictions (structures and folding pseudoenergies) to previously published experimental results. Although only a few TM β-barrel proteins have been crystallized, the literature provides extensive experimental analysis for some of these proteins. The series of experiments on Outer membrane protein A of *E. coli* conducted by Koebnik and colleagues[2–4] is probably one of the best examples of what can be done to provide a biologically meaningful benchmark (in addition to their immediate interest, of course).

TABLE V. Predictions for Permuted Variants of OmpA [3]

Set	#*seq*	Q_p	Q_2	ΔE
Circular	3	100	70.96	−79.57
Noncircular	21	43	66.16	−36.39

#*seq* is the number of sequences in the set under consideration Q_p is the percentage of correctly predicted structures, Q_2 the rate of correct secondary structure-assignment, and ΔE is the difference of folding pseudo energy between the predicted structure (optimal structure given constraints) and randomized structure (obtained by permuting structurl units of the original structure). Dataset circular contains the circularly permuted variants of the OmpA, except that (4123) has been removed, because it has been experimentally observed to be phage-resistant. The sequence (4123) is stored, along with all non-circularly permuted variants, in dataset noncircular.

In these experiments, the authors studied the efficiency of the in vivo membrane assembly of OmpA variants. Two specific protocols were designed. Koebnik et al.[3] studied consequences of permuting four structural units 1, 2, 3, and 4 in the original sequence, while Koebnik[2] focused only on the effect of amino acids mutations in TM β-strands. In the latter study, two sets of mutagenesis experiments were performed: mutation of inward pointing residues facing the cavity, and mutation of outward pointing residues facing the bilayer.

For both studies, we have reproduced these experiments in silico; results are reported in Tables V and VI. To explain our results, we need to define several notions. By *randomized* sequence, we mean an amino acid sequence obtained from OmpA by either (1) permuting structural units 1, 2, 3, and 4 of the original structure, as in ref. 3, or (2) either by mutating inward pointing residues facing the cavity or by mutating outward point residues facing the bilayer, as in Ref. 2. (For some permutations, certain constraints, such as the hydrophobicity of membrane-exposed residues, for example, are no longer necessarily satisfied.) For each randomized sequence s', we used *transFold* to compute the minimum folding pseudoenergy structure S', using the parameters from Table I. Comparing the structure S' with the structure S for OmpA, obtained by DSSP from the PDB structure for the wild-type sequence, we then computed the percentage Q_p of randomized sequences whose structure S' is the same as S. We also computed the rate of correct secondary structure assignment Q_2, defined as the overall percentage of

TABLE VI. Predictions for OmpA Mutation Variants [2]

	Outward side chains				Inward side chains			
	#seq	Q_p	Q_2	ΔE	*#seq*	Q_p	Q_2	ΔE
8C1	3	100	70.96	−79.57	22	100	73.84	−73.87
8C2	6	33	64.72	−21.43	0	—	—	—
8C3	17	6	63.30	−8.41	1	0	60.82	11.42
6C1	1	100	71.35	−61.33	0	—	—	—
6C2	0	—	—	—	0	—	—	—
6C3	12	92	75.58	−45.95	14	71	72.89	−14.39
4C1	1	100	79.53	−43.61	5	100	76.49	−79.49
4C2	2	50	76.02	−43.97	5	100	76.49	−81.19
4C3	15	20	67.45	1.03	7	100	76.52	−82.92

#seq is the number of sequences in the set under consideration, Q_p is the percentage of correctly predicted structures, Q_2 the rate of correct secondary structure assignment, and ΔE the difference of folding pseudo-energy between the predicted structure of the mutant sequence, and the observed structure of the wild-type sequence. Lines marked with "—" mean that no sequence of this type was found in the original paper. Datasets are listed in the first column. The prefix refers to the β-strand whose residues have been mutated and the suffix refers to the observed efficiency of membrane assembly. "C1" stands for proteins with high membrane assembly efficiency, "C3" for phage-resistants and "C2" for the remaining (low but significant membrane assembly). The two groups of columns (2, 3, 4 and 5, 6, 7) distinguish the side chain orientation for the mutated residues (outward or inward pointing residues).

residue positions correctly predicted to be in a β-strand of S. Additionally, we compute the *difference of potential*, ΔE, defined as $E'(s') - E(s')$, where $E'(s')$ is the folding pseudoenergy of the structure S' predicted by *transFold* for randomized sequence s', and where $E(s')$ is the folding pseudoenergy of the structure S for s'. The latter is obtained by threading the randomized sequence s' onto the structure for OmpA, as obtained by DSSP from the PDB structure.

Concerning the first experiment, Koebnik et al.[3] separated OmpA into four structural subunits (denoted 1, 2, 3, and 4), each containing an antiparallel strand pair connected by an extracellular loop (cutpoints are located in the periplasmic loops). They then built permuted variants of OmpA and measured the efficiency of membrane assembly. Their main result was to show a reduced efficiency of plating in the order of strains 1234 > 2341/3412 > 4123/noncircular; that is, wild-type 1234 has greater plating efficiency than each of the two circular permutations 2341 and 3412, while these have greater plating efficiency than both the circular permutation 4123 as well as noncircular permutations.

We then defined a dataset, denoted Circular, containing circularly permuted variants of OmpA, with the exception of 4123 because of its lack of plating efficiency; that is, the dataset Circular consists of the permutations 1234, 2341, and 3412. All other sequences (4123 included) belong to another set denoted noncircular; that is, the dataset noncircular consists of 4!-3 = 21 many permutations. Table V presents the scores computed on these datasets. Obviously, the structure used for structure comparisons (Q_p and Q_2) are permuted variants of the wild-type structure. Note that, because no turn constrains the antiparallel pairing of first and last TM β-strands of the barrel, predictions are not automatically similar for circular variants.

In the second study, Koebnik[2] mutated the amino acids of TM β-strands 4, 6, and 8, where mutation sites were restricted either to inward pointing side chains (barrel interior) or to outward pointing side chains (membrane core). This methodology allows us to distinguish the effect of both environment constraints. A total of 114 OmpA mutants have been studied and experimentally classified into three categories. Class I represents the most efficiently assembling variant, class III contains all phage-resistant (assembly defective), and class II contains variants with a low but significant membrane assembly (see the original article for more details).

In this article, the sets of mutants affecting the eighth TM β-strand are denoted 8C1, 8C2, and 8C3 according to their membrane assembly efficiency (respectively class I, II, and III). Notations for TM β-strands 6 and 4 are similar. The scores Q_p, Q_2, and ΔE computed on these datasets are shown in Table VI.

For both experiments a clear correlation is observed in the *transFold* prediction and the experimentally observed membrane assembly efficiency. The rates of correctly predicted structure Q_p, as well as correct secondary structure assignment Q_2, show that the predicted structures differ significantly from the wild-type structure (correctly predicted by *transFold*) for OmpA variants whose membrane assembly is altered. Note that phage-sensitive clones (circular or class I) have higher values of Q_p and Q_2 than do phage-resistant (noncircular or class III) clones.

CONCLUSION

In this article, we have presented the first pseudoenergy minimization method to predict the supersecondary structure of (large) TM β-barrel proteins. Our technique is free from the limitations imposed on current machine-learning methods that use only local information and present a potential overfitting from the extremely sparse dataset of available transmembrane β-barrel structures. Our method can be applied to new biological sequences without any a priori loss of accuracy. Nevertheless, because our method assumes that the input amino acid sequence represents a

single protein domain, for large multichain proteins, one should first apply a tool to determine protein domains. Moreover, the results cited earlier on OmpA suggest further uses in in silico sequence analysis coupled with in vivo or in vitro experiments. In particular, we emphasize that *transFold* will help us to further understand the folding properties of omps.

The accuracy of our method in classification (i.e., the discrimination between TM β-barrels and other proteins) is surprisingly good if we consider that our method employs a pseudoenergy model, and that the stability of omps is rather modest (<10 kcal/mol). Performance can be compared with those of previous methods.[9,10] Although the ratio of false positives needs to be improved, the results obtained for specific subclasses of omps suggest that this can be significantly improved using a more precise discrimination. However, in its current implementation, automatic structural classification by *transFold* seems a little too fragile to be used accurately, except in the case of porin-like proteins. Nevertheless, this approach appears promising and improvements in the model and grammar underlying *transFold* will be undertaken in future work. To this end, the experimental studies on TM β-strands[32] should prove useful in refining our model.

Finally, the nature of the signal (contact interactions) used to make predictions is radically different from that used by HMMs or neural networks. Despite the lack of data, these latter techniques also give good results. It is known that MTSAGs generalize HMMs,[27] so by merging both approaches (MTSAG and HMM) we may see a great improvement in the accuracy of TM β-barrel prediction. A Gibbs sampler[33] could also be considered as a candidate for refining the per-residue accuracy. A postprocessing stage using neural networks and alignment techniques[34] could perhaps improve the accuracy of interstrand residue contact predictions.

ACKNOWLEDGMENTS

We thank M. Menke for providing the BETAWRAP data used in this work, and J. King and F. Ferrè for helpful suggestions and comments.

REFERENCES

1. Wimley WC, White SH. Reversible unfolding of beta-sheets in membranes: a calorimetric study. J Mol Biol 2004;342:703–711.
2. Koebnik R. Membrane assembly of the *Escherichia coli* outer membrane protein OmpA: exploring sequence constraints on transmembrane beta-strands. J Mol Biol 1999;2851801–1810.
3. Koebnik R, Kramer L. Membrane assembly of circularly permuted variants of the *E. coli* outer membrane protein OmpA. J Mol Biol 1995;250:617–626.
4. Ried G, et al. Membrane topology and assembly of the outer membrane protein OmpA of *Escherichia coli* K12. Mol Gen Genet 1994;243:127–135.
5. Tamm LK, Hong H, Liang B. Folding and assembly of beta-barrel membrane proteins. Biochim Biophys Acta 2004;1666:250–263.
6. Wimley WC. The versatile beta-barrel membrane protein. Curr Opin Struct Biol 2003;13:404–411.
7. Schulz GE. Beta-Barrel membrane proteins. Curr Opin Struct Biol 2000;10:443–447.
8. Martelli PL, et al. A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. Bioinformatics 2002;18(Suppl 1):S46–S53.
9. Gromiha MM, Suwa M. A simple statistical method for discriminating outer membrane proteins with better accuracy. Bioinformatics 2005;21:961–968.
10. Bigelow HR, et al. Predicting transmembrane beta-barrels in proteomes. Nucleic Acids Res 2004;32:2566–2577.
11. Gromiha MM, Ahmad S, Suwa M. Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins. J Comput Chem 2004;25:762–767.
12. Natt NK, Kaur H, Raghava GP. Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. Proteins 2004;56:11–18.
13. Liu Q, et al. A HMM-based method to predict the transmembrane regions of beta-barrel membrane proteins. Comput Biol Chem 2003;27:69–76.
14. Popot JL, Engelman DM. Helical membrane protein folding, stability, and evolution. Annu Rev Biochem 2000;69:881–922.
15. Popot JL, Engelman DM. Membrane protein folding and oligomerization: the two-stage model. Biochemistry 1990;29:4031–4037.
16. Chen CP, Rost B. State-of-the-art in membrane protein prediction. Appl Bioinformatics 2002;1:21–35.
17. Waldispühl J, Steyaert J-M. Modeling and predicting all-alpha transmembrane proteins including helix-helix pairing. Theor Comput Sci 2005;335(Pattern Discovery in the Post Genome):67–92.
18. Mamitsuka H, Abe N. Predicting location and structure of beta-sheet regions using stochastic tree grammars. Proc Int Conf Intell Syst Mol Biol 1994;2:276–284.
19. Cowen L, et al. Predicting the beta-helix fold from protein sequence data. J Comput Biol 2002;9:261–276.
20. Bradley P, et al. BETAWRAP: successful prediction of parallel beta-helices from primary sequence reveals an association with many microbial pathogens. Proc Natl Acad Sci USA 2001;98: 14819–14824.
21. Menke M, et al. Wrap-and-Pack: a new paradigm for β structural motif recognition with application to recognizing β trefoils. J Comput Biol 2005;12:777–795.
22. McDonnell A, et al. Fold recognition and accurate sequence-structure alignment of sequences directing beta-sheet proteins. Proteins Struct Funct Bioinformat 2006;63(4):976–978.
23. Grantham R. Amino acid difference formula to help explain protein evolution. Science 1974;185:862–864.
24. Bhaskaran R, Ponnuswamy PK. Amino acid scale: average flexibility index. Int J Pept Protein Res 1988;32:242–255.
25. Gromiha MM, Selvaraj S. Inter-residue interactions in protein folding and stability. Prog Biophys Mol Biol 2004;86:235–277.
26. Clote P, Backofen, R. Computational molecular biology: an introduction. New York: John Wiley & Sons; 2000. p 279.
27. Lefebvre F. A grammar-based unification of several alignment and folding algorithms. Proc Int Conf Intell Syst Mol Biol 1996;4: 143–154.
28. Berman HM, et al. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
29. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
30. Kernytsky A, Rost B. Static benchmarking of membrane helix predictions. Nucleic Acids Res 2003;31:3642–3644.
31. Zemla A, et al. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins 1999;34:220–223.
32. Jackups R Jr, Liang J. Interstrand pairing patterns in beta-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction. J Mol Biol 2005;354:979–993.
33. Mannella CA, Neuwald AF, Lawrence CE. Detection of likely transmembrane beta strand regions in sequences of mitochondrial pore proteins using the Gibbs sampler. J Bioenerg Biomembr 1996;28:163–169.
34. Cheng J, Baldi P. Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms. Bioinformatics 2005;21(Suppl 1):i75–i84.