Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W189–W193***
doi:10.1093/nar/gkl205

transFold: a web server for predicting the structure and residue contacts of transmembrane beta-barrels

J. Waldispühl[1,2], Bonnie Berger[2,3], Peter Clote[1,4,*] and Jean-Marc Steyaert[5]

[1]Department of Biology, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA, [2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA, [3]Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Avenue Cambridge, Cambridge, MA 02139, USA, [4]Department of Computer Science (courtesy appointment), Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA and [5]Laboratoire d'Informatique (LIX), École Polytechnique, 91128 Palaiseau, France

Received February 14, 2006; Revised and Accepted March 22, 2006

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 makes them an important protein class. At the present time, very few non-homologous TMB structures have been determined by X-ray diffraction because of the experimental difficulty encountered in crystallizing transmembrane (TM) proteins. The transFold web server uses pairwise inter-strand residue statistical potentials derived from globular (non-outer-membrane) proteins to predict the super-secondary structure of TMB. Unlike all previous approaches, transFold does not use machine learning methods such as hidden Markov models or neural networks; instead, transFold employs multi-tape S-attribute grammars to describe all potential conformations, and then applies dynamic programming to determine the global minimum energy supersecondary structure. The transFold web server not only predicts secondary structure and TMB topology, but is the only method which additionally predicts the side-chain orientation of transmembrane β-strand residues, inter-strand residue contacts and TM β-strand inclination with respect to the membrane. The program transFold currently outperforms all other methods for accuracy of β-barrel structure prediction. Available at http://bioinformatics.bc.edu/clotelab/transFold.

INTRODUCTION

Gram-negative bacteria are surrounded by two radically different membranes, themselves separated by a region called the periplasm. The composition of the outer membrane is asymmetric and radically different than that of the inner membrane. The architecture of proteins embedded in each membrane is strikingly different—inner membrane proteins generally form transmembrane (TM) α-helical structures, while outer membrane proteins (omps) generally form transmembrane β-barrel (TMB) structures. Such TMB proteins are not exclusively found in Gram-negative prokaryotes; indeed, it is believed that in eukaryotes, omps in mitochondria and chloroplasts adopt the same architecture.

In the last few years, various methods have addressed TM β-barrel structure prediction (1–6). Nevertheless, with the exception of transFold, all other current approaches use machine learning, hence (i) they do not account for inter-strand residue interactions and (ii) are required to train on the extremely sparse set of available omp structures.

In this paper, we describe the transFold web server, which implements a novel method (7) to predict TMB architecture. The transFold program extends a method introduced previously by Waldispühl and Steyaert (8) for TM α-bundle proteins, and employs statistical potentials developed for the program BETAWRAP (9,10). Major advantages of transFold over other methods are (i) an improved predition accuracy for TM β-barrels, including residue side-chain orientations, inter-strand residue contact and strand inclination, and (ii) the definition of a folding pseudo-energy, which depends on inter-strand statistical potentials, which latter do not depend on experimentally determined omp structures. Additionally the user can obtain a prediction of inter-strand residue contacts and strand inclination, not otherwise available. Moreover, when performing *in silico* experiments, energy values

*To whom correspondence should be addressed. Tel: +1 617 552 1332; Fax: +1 617 552 2011; Email: clote@bc.edu
*Correspondence may also be addressed to Bonnie Berger. Email: bab@mit.edu

computed by transFold for a given peptide reflect the relative folding stability, hence functionality, of the peptide; indeed, as described in (7), *in silico* folding experiments performed using transFold were shown to qualitatively agree with *in vivo* experimental results for pointwise mutagenesis and domain permutations of the protein OmpA from *Escherichia coli*.

MATERIALS AND METHODS

Method

TMB prediction is realized using an underlying energy model and an abstract description of all potential structures. Our software, named transFold, applies grammars to describe all potential β-barrel supersecondary structures and then computes the global minimum energy structure by dynamic programming.

The space of all possible TMB structures is described using *multi-tape S-attribute grammars* (7)—a framework which extends that of classical context-free grammars (Figure 1a). Using grammars, we describe the following:

- the secondary structure,
- the topology (location of loops and number of TM β-strands),
- the side-chain orientation of TM β-strands residues (side-chain directed toward the membrane or toward the pore),
- the inter-strand residue contacts,
- the TM β-strand inclination with respect to the membrane plane (using shear number),

Any particular TMB structure is represented by a *parse tree* for a multi-tape S-attribute grammar, whose leaves yield the input peptide sequence. For an example of a parse tree for a small peptide see Waldispühl *et al*. (7).

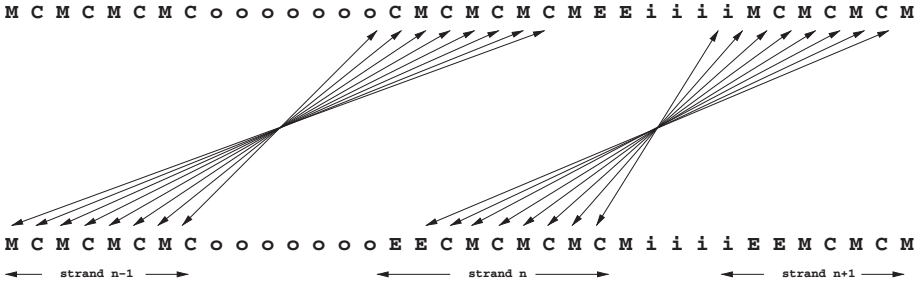
The first technical advance supported by transFold is to account for inter-strand residue contact energy, which is known to be essential for any realistic energy model for TM β-barrel structures (11). The set of possible TMB structures can be constrained by setting several biologically motivated parameters. In particular, transFold allows the user to fine-tune all of the following parameters: (i) number of TM β-strands in the barrel, (ii) length of TM β-strands, which can be interpreted as the membrane thickness, (iii) strand inclination with respect to membrane plane (shear number), (iv) size of periplasmic and extra-cellular loops and (v) hydrophobic profile of TM β-strands.

Since very few TMB structures are actually known, the inter-strand residue contact potentials are derived from an analysis of globular proteins, independent of omps.

Contact potentials are taken directly from the program BETAWRAP (9,10), developed for β-helix and β-trefoil structure prediction. Residue contact energies are differentiated according to the environment where interactions occur. The transFold web server uses contact potentials obtained from the former to compute energy contributions for TMB residue pairs, whose side-chains are directed toward the pore, while contact potentials obtained from the latter are used when side-chains are directed toward the lipid bilayer. Note that, at present, energy terms are not associated with loop regions, but only with inter-strand residue interactions. As well, in the present version of transFold, energy units are not given in kcal/mol. These shortcomings will be removed in the next version of the software.

Implementation

Current hardware supporting the transFold web server consists of a Beowulf style cluster comprising 6 Dell 1650, 2 × 1300 MHz Pentium III, 2 GB RAM with 4 Apple XServe, 2 × 1333 MHz G4, 2 GB RAM and finally 12 Dell 1850, 2 × 2800 MHz Xeon EM64T, 2 GB RAM. Interconnect is 1 Gb



(a) 2-tape representation of TM -barrel

1	A	C	31	148
2	T	M	30	147
3	S	C	29	146
4	T	M	28	145
5	V	C	27	144
6	T	M	26	143
7	G	C	25	142
8	G	M	24	141
9	Y	C	23	140
10	A	M	22	139
11	Q	C	21	138
12	S	M	20	137
13	D	C	19	136
14	A	o	0	0
15	Q	o	0	0

(b) text file output

Figure 1. (**a**) Linear representation of three consecutive and paired TM β-strands (standard output). Inter-strand residue contacts are indicated by arrows. (**b**) Sketch of a four-column tab-delimited text file summarizing the transFold prediction.

Ethernet. Pentium III nodes are running 32-bit CentOS 4.2, Xeon EM64T nodes are running 64-bit CentOS 4.2 and G4 nodes are running MacOS 10.2.8.

Worst-case time complexity for transFold is $O(n^6)$ (7), where n denotes protein length; however, during benchmark experiments to compute accuracy, we observed time complexity of $O(n^3) - O(n^4)$ on average for real proteins. Similarly, in contrast to a worst-case space upper bound of $O(n^4)$, in practice transFold appears to require memory resource of $O(n^3)$. Given an input protein of 300 residues, transFold typically uses ~1 min of CPU time and 1 GB of memory, when run on a Xeon 2.4 GHz dual processor with 2 GB of memory, running Fedora 3.0 Linux.

Performance

A brief overview of prediction accuracy measured on known TMB is given in Table 1. Our test dataset is composed of 14 structures, extracted from the PDB and filtered at 30% of sequence identity. The dataset is divided into two different subsets according to the width of the channel. Proteins with a non-water-filled pore (tight pore) are referred to as NWF. Proteins with a water-filled channel (large pore) belong to a dataset denoted WF. Formally, NWF consists of (PDB id) 1QJP, 1QJ8, 1THQ, 1P4T, 1I78, 1K24 and 1QD6, and WF of 1A0S, 1AF6, 1PRN, 2OMF, 1E54, 1TLY and 2POR.

For these datasets, the sensitivity and specificity are computed to estimate the per-segment and per-residue accuracy. Sensitivity gives the percentage of true structure which is correctly predicted, and specificity gives the percentage of predicted structure which is true structure. In the context of secondary structure assignment, sensitivity of β-strand residue assignment is denoted by $Q_{\mathrm{TM}}^{\%\mathrm{obs}}$, and specificity by $Q_{\mathrm{TM}}^{\%\mathrm{pred}}$ ($Q_{\mathrm{N}}^{\%\mathrm{obs}}$ and $Q_{\mathrm{N}}^{\%\mathrm{pred}}$ for non-TM residues, respectively). We summarize these scores by Q_2 which represents the rate of correct secondary structure assignment.

A segment is correctly predicted if the observed segment intersects one and only one predicted segment, and vice-versa. Here, 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. Finally, contact predictions were compared with H-bonded pairs of residues extracted from PDB files (an error of ±2 in index was allowed), and sensitivity ($Q_{\mathrm{ct}}^{\%\mathrm{obs}}$) and specificity ($Q_{\mathrm{ct}}^{\%\mathrm{pred}}$) scores were computed.

Real performance may not be entirely reflected by this benchmark, especially in case of porins (NWF dataset) (7).

Note also that these rates are reported for contact potentials and constraints given in (7). Using an alternate energy model or different constraint sets may change the accuracy and could in fact increase it.

INPUT AND OUTPUT FORMATS

Input

Standard submission form. The standard input form or home page of transFold uses default parameters discussed in (7). To run transFold on a given protein, the user must enter the amino acid sequence in FASTA format (single-character IUPAC codes) and should indicate the type of the pore—either water-filled (large pore) or non-water-filled (tight pore). Water-filled porins are generally much larger than non-water-filled TM β-barrel proteins, and default parameters differ according to whether the pore is filled with water.

Advanced submission form. The advanced submission form of transFold will be of most interest to structural biologists, who can thus set a number of parameters—see Figure 2.

This form allows a user to control all of the parameters mentioned so far. In particular, the user can specify upper and lower bounds for the number of β-strands, strand length, shear number, periplasmic loop size and extra-cellular loop size, and can choose constraints to be satisfied by the hydrophobicity profile. In addition to a choice of 10 standard hydrophobicity scales (Kyte and Doolittle, Eisenberg, etc.), the user can upload an arbitrary hydrophobicity scale. As a default, the transFold web server uses contact pair potentials derived from BETAWRAP (9,10). Nevertheless, the user can alternatively upload his or her own contact potentials for membrane-oriented residues and for pore-oriented residues, and thus harness the computational strength of transFold to predicte minimum energy TMB structures for a custom energy model.

By using the advanced form, a user can predict TMB structures using constraints for specific environments that might be encountered in experiments. This should help an experimental biologist to understand the folding properties of a polypeptide and to compare *in vitro* or *in vivo* experiments with computational experiments.

Output

As previously mentioned, transFold makes five types of prediction:

- the secondary structure residue assignment,
- the number of TM β-strands and loop location,
- the side-chain orientation of TM β-strand residues,
- the inter-strand residue contacts, and
- the folding pseudo-energy.

Two kinds of output are available (Figure 1) for the first four predictions: screen output corresponding to standard output of transFold software, and text file output in a five-column format.

Table 1. Prediction accuracy (sensitivity and specificity) for TM β-strand predictions, TM residue and non-TM residue predictions and contact predictions

	Strands		2-states	TM residues		Non-TM residues		Contact	
	$Q_{\beta}^{\%\mathrm{obs}}$	$Q_{\beta}^{\%\mathrm{pred}}$	Q_2	$Q_{\mathrm{TM}}^{\%\mathrm{Obs}}$	$Q_{\mathrm{TM}}^{\%\mathrm{pred}}$	$Q_{\mathrm{N}}^{\%\mathrm{Obs}}$	$Q_{\mathrm{N}}^{\%\mathrm{pred}}$	$Q_{\mathrm{ct}}^{\%\mathrm{Obs}}$	$Q_{\mathrm{ct}}^{\%\mathrm{pred}}$
NWF	100	100	79.72	86.05	83.66	67.48	71.43	64	55
WF	92.0	78.0	63.97	76.42	64.95	48.39	62.12	32	23
all	94.9	85.2	69.91	80.44	72.16	54.44	65.47	45	35

NWF contains omp with non water-filled channel. 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).

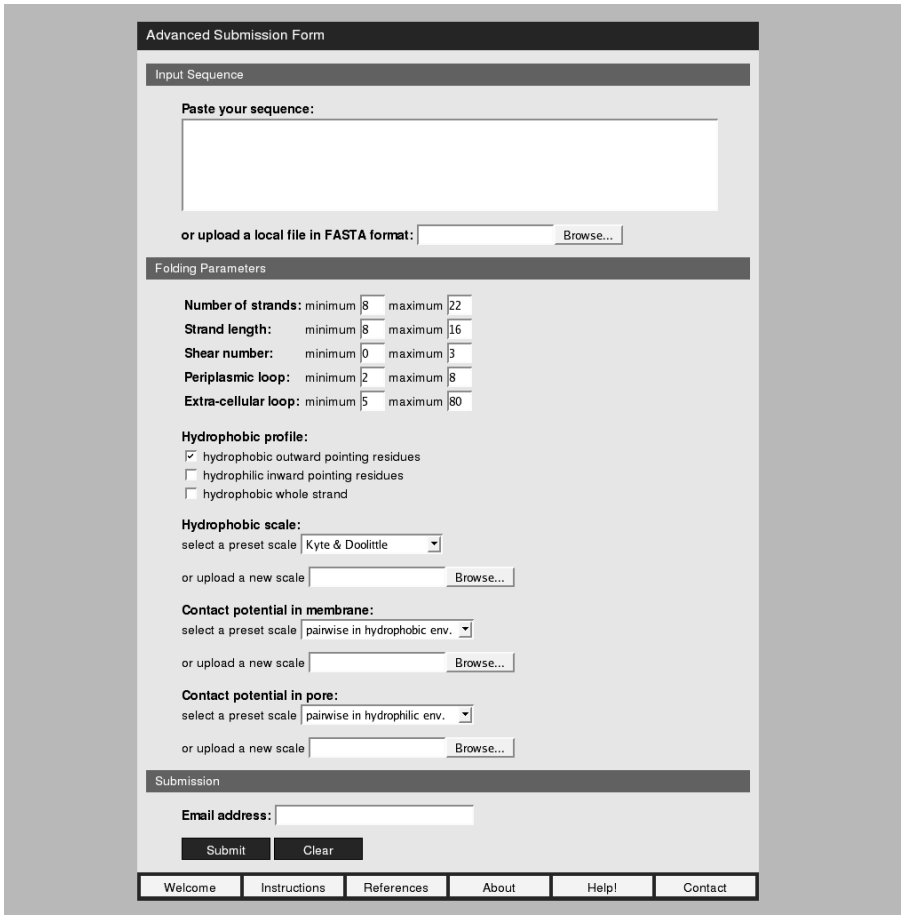


Figure 2. Input form for advanced users. The user can set bounds on the number of TM strands, bounds on strand length, shear number (inclination to plane), bounds on lengths of periplasmic and extra-cellular loops. In addition to choosing from a pull-down menu of a variety of hydrophobicity scales, the user can upload customized contact energies.

Standard output. The prediction is displayed in three lines. The first line contains the amino acid sequence input by the user (since transFold applies a lexer, all characters other than valid IUPAC single-letter amino acid codes are removed). The second and third lines contain secondary structure and topology predictions. Figure 1a illustrates this situation, although the arrows are not furnished by the software at present. There is a link 'prediction file' for an output file described later.

Residues denoted as `E`, `C` or `M` are predicted to belong to a TM β-strand (`E` denotes 'extended β-strand', `C` denotes 'channel', i.e. facing the cavity or pore of the β-barrel protein, while M denotes 'membrane', i.e. facing the outer membrane bilayer). All other positions between two consecutive strands are predicted to be loop positions—either in the extra-cellular loop or the periplasmic loop region.

The topology prediction (i.e. the orientation of the TM helices through the membrane) is given by the notation used for the amino acids located in turn regions. Residues which are predicted to be inside the periplasm milieu are marked with `i`, and those exposed to the extracellular environment are denoted with `o`. The label `E` is used to mark strand extensions.

Residue-contacts are denoted by paired residues of the same type (`C` or `M`) between the second and third lines. Pairings are articulated around a turn (denoted `i` or `o`). The first residue on the left of the third line is paired with the first residue on the right of the second line. The second residue on the left of the third line is paired with the second residue on the right of the second line. and so on. Notation is inverted for the closing pair. The folding pseudo-energy is given at the bottom.

File output. A file summarizing the prediction in a more traditional format can be dowloaded as a five-column, tab-delimitated, text file. Each row of this file corresponds to a residue. The first column contains the index of the current amino acid; the second column contains the single letter amino acid code associated with this residue (all non-IUPAC single letter amino acid codes are stripped). The third column contains the secondary structure residue assignment as well as the side-chain orientation of TM β-strand residues. Hence, a residue marked as `M` or `C` is predicted to belong to a TM β-strand, where `C` denotes 'channel' (i.e. facing the cavity of the β-barrel protein) and `M` denotes 'membrane' (i.e. facing the outer membrane bilayer). A residue marked as `i` is predicted to be in the periplasm, a residue marked as `o` to be extra-cellular, and a residue marked as '.' to be exterior to the membrane, but not in a turn. The last two columns are only used for TM β-strand residues and give the index of the amino acids interacting with the current one (including the interaction of the closing strand pairing).

ACKNOWLEDGEMENTS

B.B. is partially supported by NSF Grant ITR (ASE+NIH)-(dms)-0428715, and P.C. by NSF grant DBI-0543506. Funding to pay the Open Access publication charges for this article was provided by NSF grant DBI-0543506.

Conflict of interest statement. None declared.

REFERENCES

1. Gromiha,M.M. and Suwa,M. (2005) A simple statistical method for discriminating outer membrane proteins with better accuracy. *Bioinformatics*, **21**, 961–968.
2. Gromiha,M.M., Ahmad,S. and Suwa,M. (2005) TMBETA-NET: discrimination and prediction of membrane spanning beta-strands in outer membrane proteins. *Nucleic Acids Res*., **33**, W164–W167.
3. Bigelow,H.R., Petrey,D.S., Liu,J., Przybylski,D. and Rost,B. (2004) Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res*., **32**, 2566–2577.
4. Gromiha,M.M., Ahmad,S. and Suwa,M. (2004) Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins. *J. Comput. Chem*., **25**, 762–767.
5. Natt,N.K., Kaur,H. and Raghava,G.P. (2004) Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. *Proteins*, **56**, 11–18.
6. Liu,Q., Zhu,Y.S., Wang,B.H. and Li,Y.X. (2003) A HMM-based method to predict the transmembrane regions of beta-barrel membrane proteins. *Comput. Biol. Chem*., **27**, 69–76.
7. Waldispühl,J., Berger,B., Clote,P. and Steyaert,J.-M. (2006) Predicting transmembrane beta-barrels and Inter-strand residue contacts from sequence. *Proteins* (in press).
8. Waldispühl,J. and Steyaert,J.-M. (2005) Modeling and predicting all-alpha transmembrane proteins including helix-helix pairing. *Theor. Comput. Sci*., **335**, 67–92.
9. Cowen,L., Bradley,P., Menke,M., King,J. and Berger,B. (2002) Predicting the beta-helix fold from protein sequence data. *J. Comput. Biol*., **9**, 261–276.
10. Menke,M., King,J., Berger,B. and Cowen,L. (2005) Wrap-and-Pack: a new paradigm for beta structural motif recognition with application to recognizing beta trefoils. *J. Comput. Biol*., **12**, 777–795.
11. Tamm,L.K., Hong,H. and Liang,B. (2004) Folding and assembly of beta-barrel membrane proteins. *Biochim. Biophys. Acta*, **1666**, 250–263.