Supporting Information

Hayat et al. 10.1073/pnas.1419956112

SI Methods

In this section we provide detailed description of a few methods used in this study and have indexed the supplementary data items related to the main text. Supplementary data and source code are available at cbio.mskcc.org/foldingproteins/transmembrane/betabarrels/.

Benchmark Dataset. A total of 141 TMBs (belonging to 52 PFAM families) with 3D structures were taken from the OPM database (1). Fifty-six TMBs in 29 PFAM families were chosen after excluding multichain TMBs redundancy reduction at 30% sequence identity. Five PFAM families were discarded so that the alignment overlap between any two families is less than 20%. Of the 24 remaining PFAM families, 18 TMBs have more than five sequences per residue in their alignment and were chosen to benchmark EVfold_bb. However, two families were not folded because of failure in topology prediction by boctopus2 and LptD was added to the list of the blinded dataset after its 3D structure became available, resulting in 17 proteins that were de novo folded. Location of the β-barrel domain was obtained from the known structure. Although 36 3D structures with at most 30% sequence similarity were used in a previous study (2), here we excluded those proteins that belonged to the same PFAM domain or had an overlap of more than 20% in their multiple-sequence alignment.

Prediction of Evolutionary Couplings from Multiple-Sequence Alignments. MSAs for all proteins are generated using three iterations of jackhmmer (version 3.1) (3) against the UniProt database. For all proteins, an E-value of 10^{-2} was used to ensure the maximum number of sequences. For LptD and FecA multidomain interaction predictions, MSAs were generated at an E-value threshold of 10^{-10} and 10^{-20}, respectively, to obtain stringent alignments and ensure sequence coverage in both domains. The two criteria used for selecting the parameters for generating a MSA are the E-value and the number of columns in the multiple-sequence alignment for which sufficient sequences can be found to infer evolutionary couplings. Based on these criteria, for all proteins in our blinded dataset, an E-value of 10^{-2} and a column-inclusion threshold of 80% were used for MSA generation. Columns in MSA above this threshold were excluded from the maximum-entropy model and no contacts were predicted for them. Additionally, to predict interdomain contacts in LptD, an E-value of 10^{-10} was chosen to ensure maximum residue coverage across both domains. For estimating ECs between the FecA barrel and the plug domain a stricter E-value cutoff of 10^{-20} and column-inclusion cutoff of 50% were used as enough sequences were available at all thresholds tested (supplementary data at cbio.mskcc.org/foldingproteins/transmembrane/betabarrels/). A global statistical inference method based on pseudolikelihood maximization (4) as implemented in EVFold (evfold.org/) (5) is used to extract direct interactions from all of the observed correlations in a MSA. A ranked list of ECs is obtained by taking the average-product corrected norm of the matrix of couplings that adjusts for the phylogenetic bias (4).

Topology Prediction Using Boctopus2. An earlier nonredundant dataset (less than 30% sequence identity) of 36 TMBs with known structures along with transmembrane β-strand boundaries was curated from the OPM database (1) (Tables S4 and S5). Boctopus2 was developed using an almost identical strategy to that used when developing boctopus (2). The main difference is that all residues in the dataset were labeled as outer loop (o), inner loop (i), β-strand pore facing (p), and β-strand lipid facing (l), whereas in boctopus the "p" and "l" residues were grouped together. The position-specific scoring matrix (PSSM) obtained using three iterations of hhblits (version 2.0.13) (6) against the "nr" database (nr20_12Aug11) is used as the input to four separate support vector machines (SVMs) that were trained to predict the per-residue location. Together with secondary structure prediction using PSIPRED (7), a per-residue profile is generated and used as input to a hidden Markov model to predict the overall topology. Boctopus2 is trained in a 10-fold cross-validated manner, where all proteins belonging to the same family were put together in the training or the test set. In contrast to boctopus all transition probabilities could be set to 1, which means that the HMM architecture is not trained (Fig. S5). Within the barrel domain, boctopus2 predicts the correct β-strand arrangement for 32 of 36 proteins in the benchmark dataset and the correct number of strands for all except 1i78 and 2qdz (Tables S4 and S5). Furthermore, for OprP (Protein Data Bank [PDB] ID code 2o4v), two extra strands are predicted outside the barrel domain. Additionally, topologies for five proteins not in the initial boctopus2 dataset (3syb, 4k3c, 4e1t, 3ohn, and 2jk4) are predicted using boctopus2 (Tables S4 and S5). For InvA (PDB ID code 4e1t) and VDAC (PDB ID code 2jk4), two and one extra strands are predicted outside the barrel boundary, respectively. For BamA (PDB ID code 4k3c), FecA (PDB ID code 1kmp), and EstA (PDB ID code 3kvn), the nonbarrel/barrel boundary is predicted by adding p and l probabilities averaged over a window size of 50 residues and regions with a total probability greater than 0.6 are then classified as barrel.

Blinded Ranking of EVfold_bb Models. The de novo folded 3D structures are ranked based on a scheme that uses two criteria: number of hydrogen-bonded constraints satisfied in the generated model and the quality of torsion angles in the predicted helices and between adjacent β-strands (5).

Number of hydrogen-bonding constraints satisfied in the generated model. The rationale here is that the number of applied constraints in well-folded models will be higher than in badly folded models. For folding TMBs, we first extract residue pairs that are hydrogen bonded between two adjacent β-strands and then apply distance constraints (2.9 Å ± 0.3 Å) on the N–O and O–N atoms of those residue pairs (Table S6). To assess the quality of hydrogen-bond constraints in folded models, the percentage of constraints satisfied is calculated by dividing the number of constraints that are actually satisfied (i.e., distance between O–N atoms is in the range 2.9 Å ± 0.3 Å) in the folded model by the total number of constraints applied to fold that model. In this way, a hydrogen-bonding score is obtained for each model.

Quality of torsion angles in the generated model. Our distance-geometry–based folding protocol can generate models that satisfy the applied constraints but can still be mirror images of the correct structure. In a previous study carried out by some of us (5), it was observed that in the mirror images, the chirality of secondary structure elements, especially β-strands, is opposite of what is observed in known structure. This feature could be used to discriminate models that are mirrored from those that are not. To assess the chirality of helices, we calculated the dihedral angle formed by four consecutive Cα atoms at positions i, $i + 1$, $i + 2$, and $i + 3$. For β-strands, the value of the dihedral angle formed by Cα atoms at positions i, $i + 2$ and j, $j + 2$ (where i and j are adjacent strands) is calculated. The algorithm then calculates the proportion of dihedral angles that lie within an acceptable range with a decreasing function for antiparallel β-strands (5). These values are weighted by the proportion of predicted helices and β-strands and added to obtain a dihedral score for each model.

Combination of hydrogen-bonding and dihedral scores for final blind ranking. We normalize both the measures described above by calculating the *z*-score for the individual measure. Briefly, *z*-scores are calculated by subtracting each value in the distribution by the distribution mean and then division by the SD. For each model, the two *z*-scores are then combined into a composite score by summation. Models are ranked based on the decreasing value of the final composite score with a high value signifying a better model. Later, we validated the utility of our ranking scheme by comparing the ranking score and the TM score for each model (supplementary data item *19*).

Potential Application to Protein Families Without a Known 3D Structure. To estimate the number of transmembrane β-barrel families without a known 3D structure, we took the list of predicted transmembrane β-barrel sequences generated by Freeman and Wimley (8), extracted a subset of 15,483 sequences that passed their conservative prediction threshold (BetaBarrel_score and signal peptide score both >0.7), and mapped these to domain families in the PFAM database when possible (9). Of these, 292 domain families had no member in the PFAM alignment with a known 3D structure. However, a more sensitive search for remote homologs, using HHSearch (version 2013) (10) reduces the number of domain families without a known 3D structure to 172. Of these, 63 (~37%) have enough sequences to predict contacts using maximum-entropy analysis (supplementary data at cbio.mskcc.org/foldingproteins/transmembrane/betabarrels/). However, only three [YP_861842.1 (region 694–832), YP_001305047.1 (region 122–317), and NP_754081.1 (region 25–248)] predicted contact maps have the characteristic antiparallel β-strand pattern (supplementary data at cbio.mskcc.org/foldingproteins/transmembrane/betabarrels/). Thus, our analysis shows that current methods for identifying novel transmembrane β-barrels result in many false positives. The 109 putative transmembrane β-barrel families for which an insufficient number of homologous sequences was obtained using current sequence databases can be analyzed in the future when more sequences are available.

List of supplementary data used in this study

Serial number	Description	Location
1	Input protein sequences and β-barrel domain boundaries	barrel_domain_boundaries
2	Estimation of transmembrane β-strands using boctopus2	predicted_topologies_using_boctopus2
3	Multiple sequence alignments (MSA) used for evolutionary coupling (EC) analysis	multiple_sequence_alignments_used_for_ECs
4	Statistics on *E*-value cutoffs used for generating MSAs	Alignment_statistics
5	Raw EC output from EVFold-PLM (ranked list of predicted contacts)	Raw_predicted_ECs
6	Predicted contact maps obtained from raw ECs	Raw_predicted_contactmaps
7	ECs accuracy per protein	EC_accuracy_per_protein
8	Hydrogen-bonded constraints derived using strand shift and ECs	Hydrogenbonding_constraints
9	Predicted contact maps of derived hydrogen-bonded constraints	Hydrogenbonding_constraints_Contactmaps
10	Improvement in identification of residue pairs that are hydrogen bonded using ECs and strand-shift algorithm	Strandwise_hydrogenbondpair_prediction_comparison
11	Secondary structure predictions obtained from PSIPRED	SecondaryStructure_prediction_PSIPRED
12	Constraints applied to dihedral angles	Dihedral_angle_constraints
13	Distance constraints applied to maintain secondary structure	Distance_constraints_on_secondary_structure
14	List of side-chain atom pairs where distance constraints are applied	List_of_candidate_sidechain_atoms.txt
15	List of all distance constraints used as input to CNS for folding β-barrels	Constraints_input_to_CNS
16	All 3D models generated (barrel region only)	EVfold_bb_folded_barrelregion_models
17	All 3D models generated (full proteins)	EVfold_bb_folded_fullprotein_models
18	Mapping of UniProt sequence indexing to PDB sequence indexing for structure comparison	Uniprot_to_PDB_mapping
19	Blinded ranking of 3D models	Blindedranking_vs_tmscore
20	List of top 10 blindly ranked 3D models	List_of_top10_blindlyranked_models
21	3D structure overlay of top-ranked, best in top-5–ranked, and overall best model in the ensemble on the known structure (models and pymol sessions for barrel and full protein models)	EVFoldBB_3dmodels_superimposed_on_known_structure
22	3D structure overlay of top-ranked, best in top-5–ranked, and overall best model in the ensemble on the known structure (high-resolution images of front and top view of barrel and full protein regions)	EVFoldBB_3dmodels_superimposed_on_known_structure_figures
23	EVfold_bb model comparison with the known 3D structure (TM score and rmsd values)	Model_comparison_output_tmscore_and_RMSD
24	Cα models and topologies generated by tobmodel	tobmodel_predicted_calpha_idealized_barrel_models
25	Comparison of tobmodel and EVfold_bb models (pymol sessions)	tobmodel_EVFold_models_comparison_pymol_sessions
26	Comparison of tobmodel and EVfold_bb models (high-resolution images of front and top view of the modeled barrel region)	tobmodel_EVFold_models_comparison_figures

Cont.

Serial number	Description	Location
27	Comparison of strand-registration accuracy in tobmodel and EVfold_bb models	Correct_inregister_in_EVFoldBB_and_tobmodel_3dmodels
28	List of PFAM domains without a known 3D structure	List_of_PFAMs_without_3d_structure
29	Predicted contact maps of putative transmembrane β-barrel PFAM domains without a structure	Predicted_Contactmaps_unknown_cases
30	FecA protein: interdomain interactions and pymol session	FecA_interdomain_interactions
31	LptD: predicted topology, ECs, contact maps, and list of residues that occur in multiple evolutionary couplings	LptD_analysis
32	Residues that appear in multiple evolutionary couplings superimposed on the known 3D structure (putative functional sites)	ECs_enrichment
33	Location distribution of false positive ECs	False_positive_ECs_Location_distribution
34	Source code: estimation of transmembrane strand location boctopus2	Topology prediction using boctopus2
35	Source code: prediction of strand-registration and hydrogen-bonding constraints	Strand registration and hydrogen bonding
36	Source code: EVfold_bb folding pipeline	Folding pipeline using EVfold_bb

Description of the Essential Supplementary Data (Inputs, Intermediate Results, Outputs, and Source Code) Accompanying the EVfold_bb Pipeline That Was Used to de Novo Predict the 3D Structure of Transmembrane β-Barrel Proteins.

1) Protein sequences for the 19 proteins were obtained from UniProt (www.uniprot.org). Domain boundaries for the transmembrane β-barrel were obtained from the known 3D structures.

2) Location and number of transmembrane β-strands were predicted using a machine-learning–based method called boctopus2. For each protein in the dataset, predicted strand locations (start to end) in UniProt numbering are provided.

3) Multiple sequence alignments were generated using three iterations of jackhammer software (version 3.1) against the UniProt database of protein sequences.

4) Statistics obtained for different *E*-value cutoffs tried for searching homolog sequences that are included in the multiple-sequence alignment. A gap threshold of 80% means that all columns in the alignment that had more than 80% gaps were excluded from the maximum-entropy analysis and no contacts were predicted for those positions. A threshold of 80% was chosen to allow more loop regions to be included in the maximum-entropy model. An *E*-value cutoff of 10–2 was used for all proteins in the dataset except for predicting interactions between the two domains of FecA and LptD proteins. The *E*-value cutoff was chosen such the maximum number of sequences is included in the alignment. In the future, an empirical function to choose an optimal *E*-value cutoff and gap threshold can be used.

5) EVfold-PLM code, which uses an implementation of the plmDCA (4) method in the EVfold framework to predict coevolving residue pairs, was used to generate a ranked list of evolutionary couplings. Only medium- and long-range evolutionary couplings between pairs of residues that are separated by more than five residues in the protein sequence were used for folding.

6) Contact maps: predicted ECs (red) and observed (gray) contacts in the known 3D structure for all proteins in the blinded dataset. For FecA, FadL, OmpC, TsX, PA1, and Q9HVS0 a few long-range interactions between the first and the last β-strands were predicted. However, very few contacts are predicted for residues in the long outer loops.

7) ECs accuracy for top *X* predictions, where *X* is the number of top-ranked ECs proportional to the length of the protein sequence, is defined as the number of residue pairs that are spatially located within a threshold value from each other in the observed structure over the total number of predictions (dashed lines). Prediction accuracy improves when the predicted residue pairs are spatially located at positions that are considered structurally unviable are filtered out (solid lines) (5). Threshold values of 5 Å, 8 Å, and 10 Å were used.

8) Distance constraints were applied to NO, ON, CA–CA, and one side-chain heavy atom pair for residues on adjacent strands that were predicted to be hydrogen bonded. For example, in the following input to CNS, distance constraints of 3 ± 1 Å are applied to atoms CE1 and OD1 or residues 98 and 144, respectively.

a. assign (resid 98 and name CE1) (resid 144 and name OD1) 3 1 1 weight 2.0.

9) Contact maps showing predicted hydrogen-bonded residues (red) over contacts observed (gray) in known 3D structures.

10) The number of residue pairs correctly predicted to be hydrogen bonded increases after using ECs plus the strand shift algorithm (*Methods*).

11) Secondary structure (helixes, loops, and sheets) prediction results using PSIPRED.

12) Distance constraints were applied to intrahelix and intrasheet secondary structure elements predicted using PSIPRED.

13) Dihedral angle constraints with default values pertaining to antiparallel β-strands were applied to intrasheet secondary structure elements predicted using PSIPRED.

14) One side-chain atom from each residue type was selected and distance constraints were applied to the atom pair based on the residues predicted to be in contact.

15) Complete list of constraints used as input to CNS for folding β-barrels.

16) Realistic all-atom 3D models generated by EVfold_bb (predicted transmembrane barrel region).

17) Realistic all-atom 3D models generated by EVfold_bb (full protein including long loops).

18) Mapping of UniProt sequence to PDB sequence index for comparison of blindly folded de novo models with the known structure.

19) The top-ranked models according to our blind ranking scheme are close to the best possible model in the ensemble generated in most cases. Blindly ranked 3D models (barrel region) are compared with the known 3D structure in terms of TM score.

20) List of top 10 blindly ranked models for all proteins in the dataset.

21) Pymol sessions of full and barrel region top-ranked, best in top 5, and overall best possible 3D models generated by EVfold_bb overlaid on the known 3D structure.

22) High-resolution images (front and top views) of the overlaid 3D coordinates of the top-ranked, best in top 5, and overall best possible model (barrel and full protein) and the known 3D structure.

23) Raw output of tm-align software (*Methods*) for the model and the known 3D structure comparison in terms of TM score and rmsd values.

24) Predicted β-strand regions and Cα model generated by tobmodel.

25) Pymol sessions of the 3D models of the predicted transmembrane barrel region generated tobmodel and EVfold_bb superimposed on each other.

26) High-resolution images of front and top view of barrel region modeled by EVfold_bb and tobmodel.

27) Strand-registration accuracy (*Methods*) for all strand pairs in the 3D models generated by EVfold_bb and tobmodel.

28) List of PFAM domain families without a known structure and no close remote homologs with a known 3D structure as determined by similarity comparison of the alignments generated for the input sequence and the remote homologs with a known 3D structure found using HHsearh (10). HH_delta score of greater than 0.5 signifies that the input alignment differs significantly from that of the alignment generated for the known 3D structure (10).

29) Predicted contact maps of PFAM domain families without a known 3D structure. Only three contact maps of 63 PFAM domain families have the characteristic antiparallel β-strand pattern.

30) FecA interdomain interactions (plug domain region, 121–244; and barrel domain, 245–774): ECs are predicted with high accuracy for FecA protein (pymol session).

31) LptD analysis: Predicted transmembrane β-strand location by boctopus2 and ECs. ECs used for folding and deriving the hydrogen bond constraints and their contact maps. List of residues that occur in multiple ECs. Interdomain ECs between the N-terminal cytosolic and the C-terminal barrel domain (200–784).

32) List of residues that appear in multiple ECs and pymol sessions of those residues superimposed on the known 3D structure reveal putatively functional sites.

33) Location of false positive ECs shows that most incorrect predictions are made in loop–loop and loop–strand interactions.

34) Source code: Estimation of transmembrane strand location boctopus2.

35) Source code: Prediction of strand-registration and hydrogen-bonding constraints.

36) Source code: EVfold_bb folding pipeline.

1. Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI (2006) OPM: Orientations of proteins in membranes database. *Bioinformatics* 22(5):623–625.
2. Hayat S, Elofsson A (2012) BOCTOPUS: Improved topology prediction of transmembrane β barrel proteins. *Bioinformatics* 28(4):516–522.
3. Finn RD, Clements J, Eddy SR (2011) HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res* 39(Web Server issue):W29–W37.
4. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys Rev E Stat Nonlin Soft Matter Phys* 87(1):012707.
5. Marks DS, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* 6(12):e28766.
6. Remmert M, Biegert A, Hauser A, Söding J (2012) HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 9(2):173–175.
7. McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. *Bioinformatics* 16(4):404–405.
8. Freeman TC, Jr, Wimley WC (2010) A highly accurate statistical approach for the prediction of transmembrane beta-barrels. *Bioinformatics* 26(16):1965–1974.
9. Finn RD, et al. (2014) Pfam: The protein families database. *Nucleic Acids Res* 42(Database issue):D222–D230.
10. Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci USA* 110(39):15674–15679.

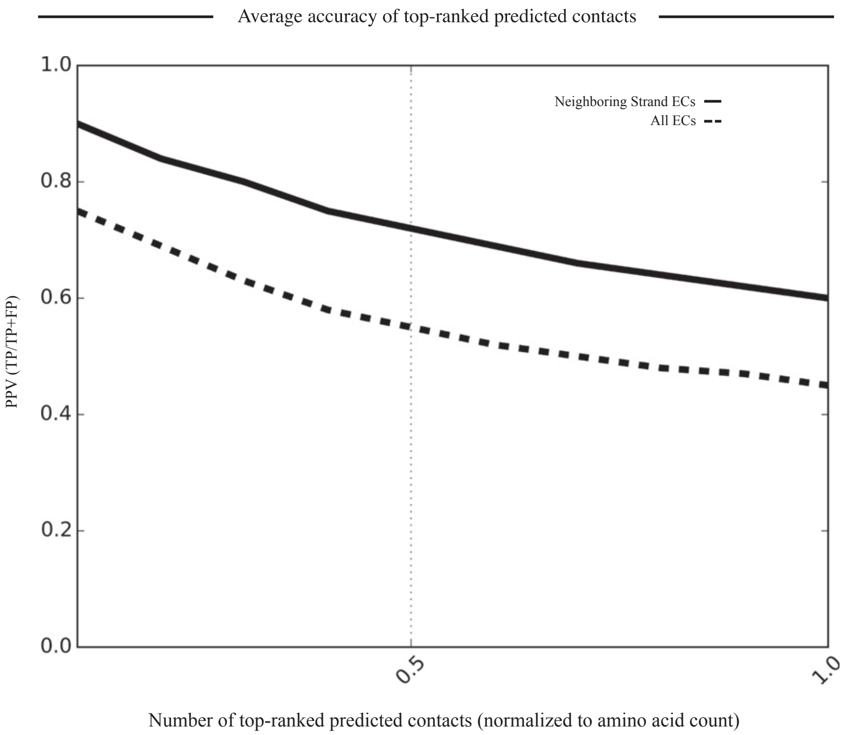

Fig. S1. Fraction of all predicted contacts that are ≤5 Å in the crystal structure over the total number of predictions made (dashed line). Shown is the fraction of ECs on adjacent β-strands that are ≤5 Å in the crystal structure over the total number of predictions made (solid) normalized to protein length. PPV, positive predictive value defined as number of correct predictions (TP) divided by the total number of predictions made; FP, false positive predictions, where all-atom minimum distance between two residues is greater than 5 Å.

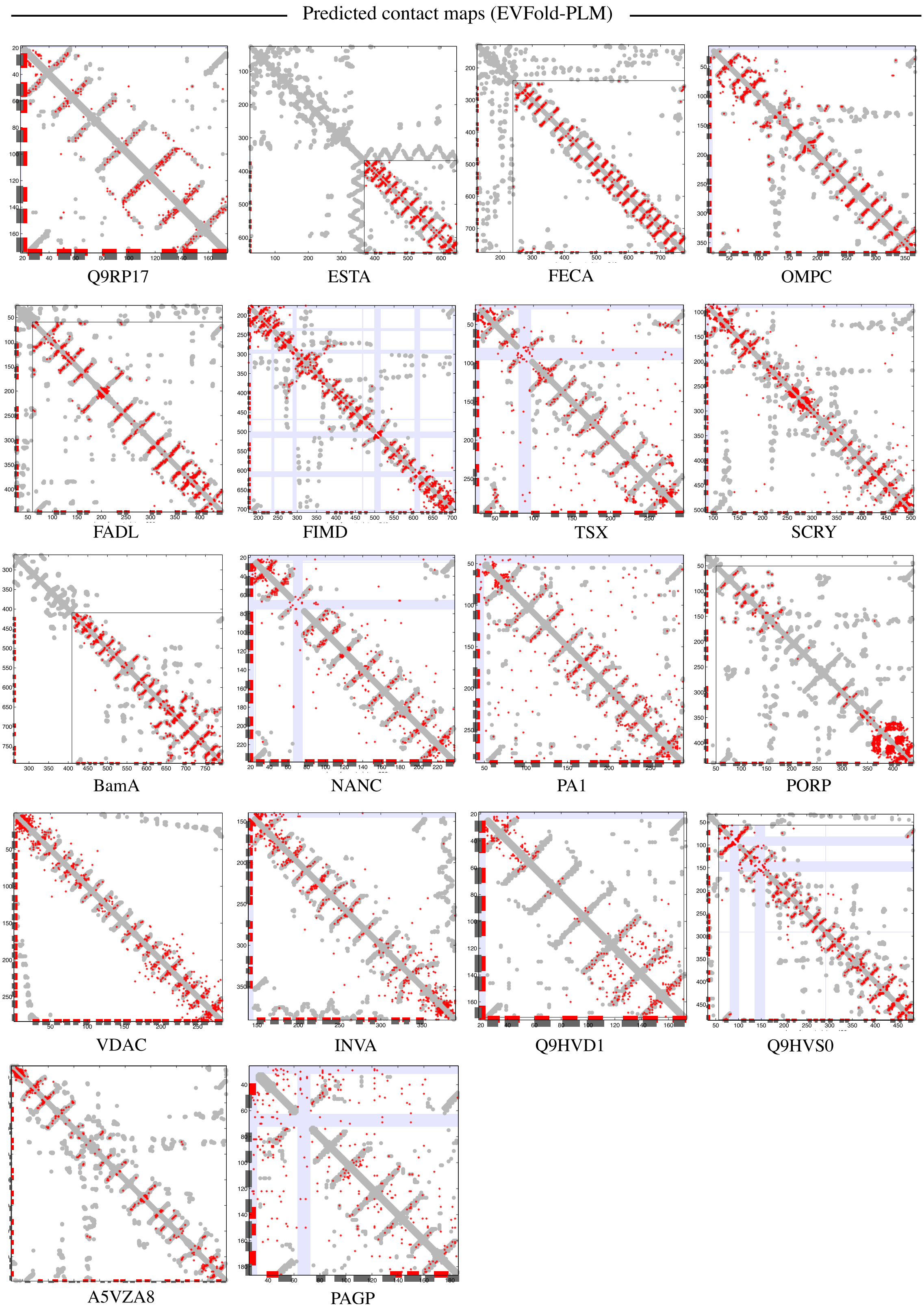

Fig. S2. Contact maps predicted using ECs obtained from EVFold-PLM. Shown are predicted ECs (red) on contacts observed (all-atom minimum distance between two residues ≤5Å) in the 3D structure (gray) and regions missing in the observed 3D structure (blue). Axes show predicted (red) and observed (black) transmembrane β-strands.

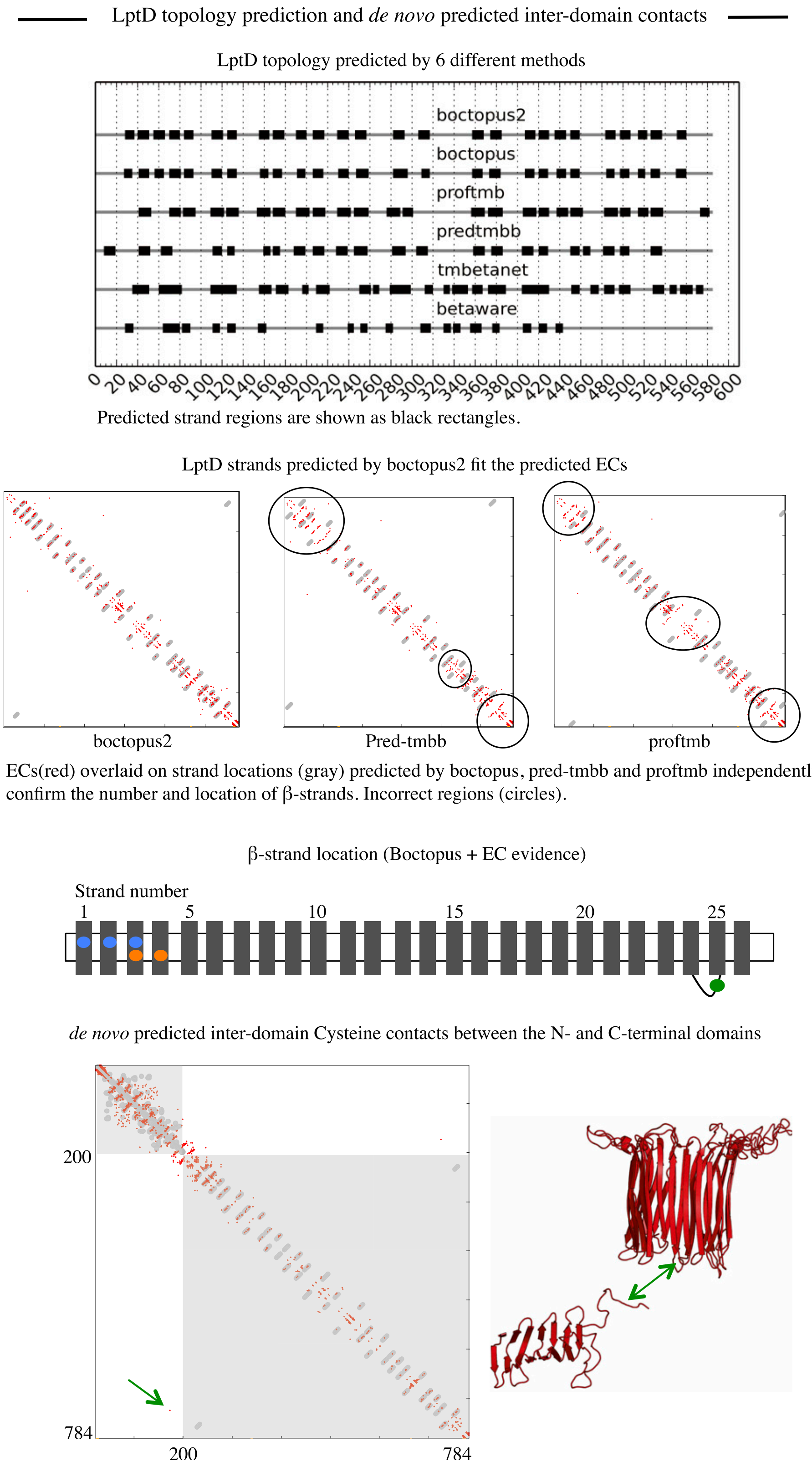

Fig. S3. Predicted LptD topologies using different prediction methods overlaid on contact maps obtained from EVFold-PLM. Shown is comparison of the de novo folded C-terminal domain of LptD (red) with the known 3D structure (gray). Twenty-six transmembrane β-strands are predicted for the LptD C-terminal domain by boctopus2 and independently validated by evolutionary couplings obtained from EVFold-PLM. Residue P246 (blue) located on strand 2 has the highest number of couplings (1) in the top *L*/2 ECs and is spatially close to P236 and P261 (blue) located on adjacent strands. Residue D256 (orange) on strand 3 has eight couplings in the top *L*/2 predictions (ranked fourth). In addition, potential salt-bridge–forming residues D256 are R277 are evolutionary coupled (ranked 66th). N- and C-terminal domain regions are shown in gray. For the N-terminal domain, ECs (red) are overlaid on predicted strand locations (gray). For the C-terminal domain, ECs (red) are overlaid on the 3D model generated by HHpred (2). Interdomain C173 and C725 interaction predicted by EVFold-PLM is highlighted (green).

[1] Finn RD, et al. (2014) Pfam: The protein families database. *Nucleic Acids Res* 42(Database issue):D222–D230.

[2] Soding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 33(Web Server issue):W244–W248.

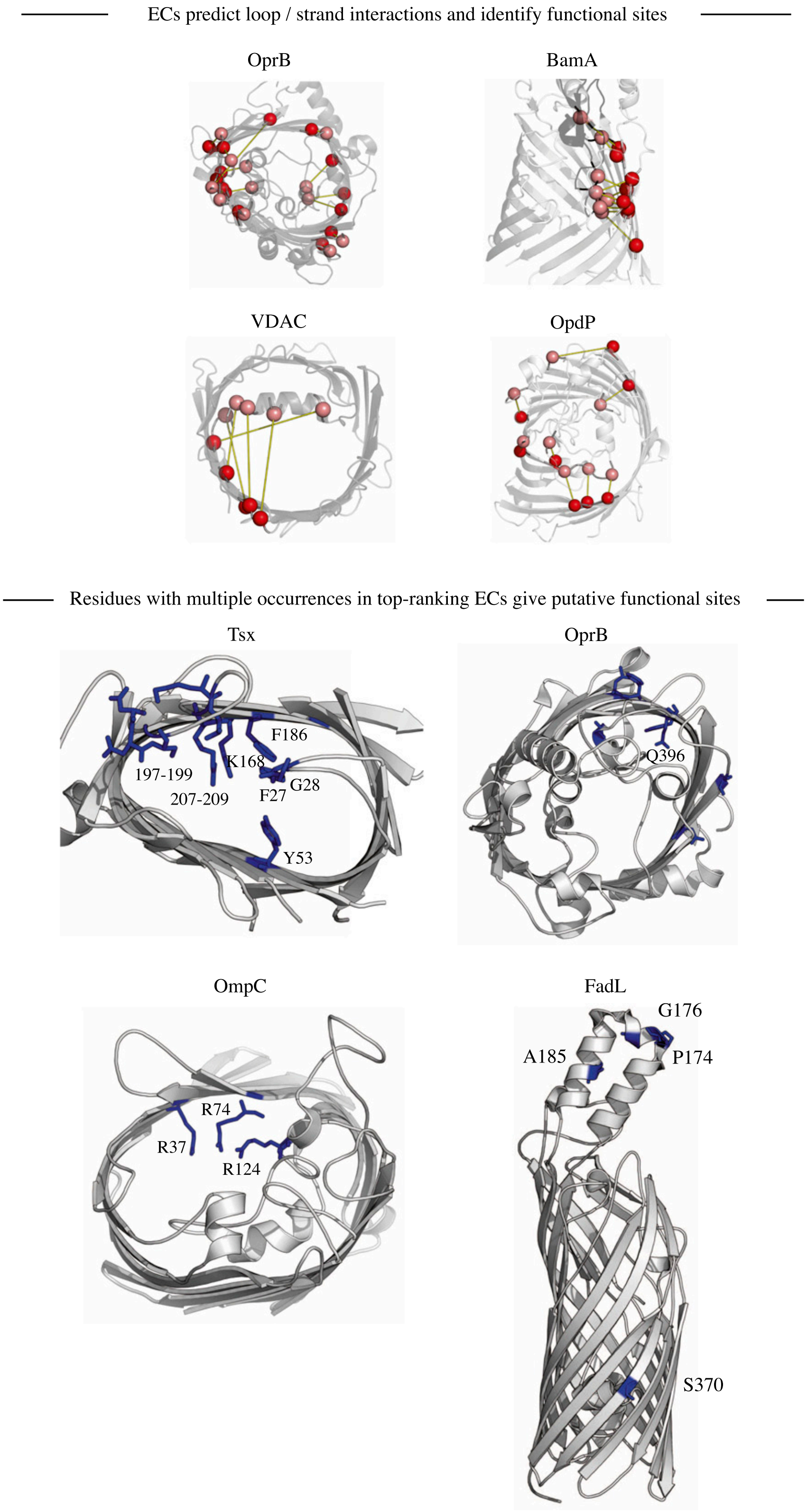
ECs predict loop / strand interactions and identify functional sites
OprB
BamA
VDAC
OpdP
Residues with multiple occurrences in top-ranking ECs give putative functional sites
Tsx
F186
K168
197-199
207-209
F27
G28
Y53
OprB
Q396
OmpC
R74
R37
R124
FadL
G176
A185
P174
S370

Fig. S4. EVFold-PLM predicts interactions between long-extracellular outer loops and β-strands and gives indication of functionally important sites.

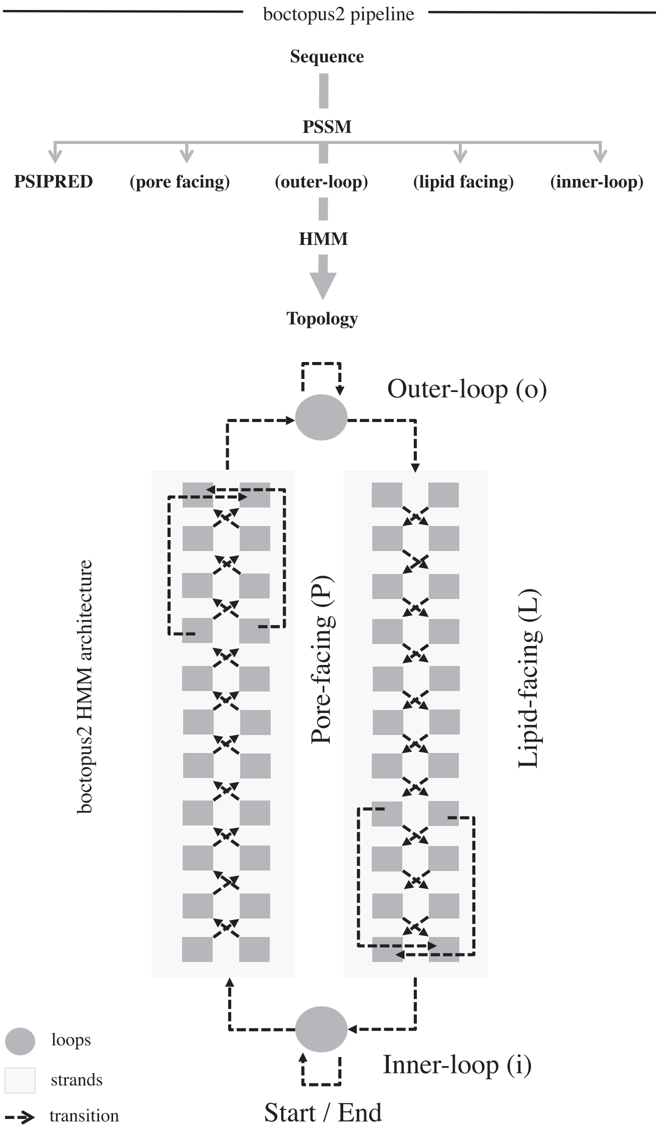

Fig. S5. Boctopus2 topology prediction pipeline and the hidden Markov model (HMM) architecture. Boctopus2 uses four separate support vector machines (SVMs) to predict the location of each residue. Output from these SVMs and PSIPRED secondary structure prediction is used as the input to the HMM.

Table S1. Prediction performance in a blinded test on proteins with a known structure (transmembrane barrel region)

Protein (no. strands)	Uniprot ID	Amino acid count in the membrane region	TM score top-ranked model	rmsd top-ranked model	TM score best in top-5–ranked model	rmsd best in top-5–ranked model	TM score best possible model in ensemble	rmsd best possible model	No. models generated
1p4tA (8)	Q9RP17_NEIME	87	**0.85**	1.59 (87)	**0.87**	1.5 (87)	**0.87**	1.50 (87)	340
1kmpA (22)	FECA_ECOLI	226	**0.67**	4.48 (217)	**0.7**	4.06 (216)	**0.7**	4.06 (216)	1100
3kvnA (12)	ESTA_PSEAE	126	**0.68**	3.12 (124)	**0.71**	3.15 (123)	**0.75**	2.50 (123)	600
2j1nA (16)	OMPC_ECOLI	160	**0.58**	4.65 (152)	**0.65**	4.05 (152)	**0.68**	3.38 (152)	740
3ohnA (24)	FIMD_ECOLI	221	0.46	4.77 (150)	**0.51**	5.41 (182)	**0.55**	5.07 (188)	1100
4k3cA (16)	Q93PM2_HAEDC	167	0.48	5.22 (141)	0.48	5.5 (146)	**0.53**	4.37 (146)	800
4e1tA (12)	INVA_YERPS	116	0.45	4.49 (93)	0.45	4.63 (97)	0.48	3.57 (83)	540
1t16A (14)	FADL_ECOLI	147	**0.66**	3.86 (147)	**0.71**	3.33 (144)	**0.74**	2.93 (144)	800
3sybA (18)	Q9HVS0_PSEAE	178	**0.62**	4.15 (162)	**0.62**	4.15 (162)	**0.63**	4.08 (162)	880
2wjrA (12)	NANC_ECOLI	120	**0.50**	4.38 (109)	**0.51**	4.43 (109)	**0.52**	4.21 (109)	480
1thq (8)*	PAGP_ECOLI	NA	NA	NA	NA	NA	NA	NA	NA
1qd6C (12)	PA1_ECOLI	122	**0.61**	3.5 (112)	**0.61**	3.5 (112)	**0.61**	3.50 (112)	520
4q35A (26)	LPTD_ECOLI	267	0.36	6.54 (174)	0.36	5.79 (153)	0.42	6.25 (185)	1200
1a0sP (18)	SCRY_SALTM	180	0.39	6.12 (139)	0.42	4.63 (111)	0.47	5.12 (144)	880
1tlyA (12)	TSX_ECOLI	125	0.41	5.06 (103)	0.46	4.87 (106)	0.46	4.45 (105)	560
2jk4A (19)	VDAC1_HUMAN	176	0.45	6.18 (157)	**0.51**	5.41 (162)	**0.53**	5.41 (161)	600
2ervA (8)	Q9HVD1_PSEAE	84	**0.67**	2.55 (79)	**0.67**	2.55 (79)	**0.67**	2.51 (79)	340
4gey (16)*	A5VZA8_PSEP1	NA	NA	NA	NA	NA	NA	NA	NA
2o4vA (16)	PORP_PSEAE	164	0.40	4.98 (109)	0.41	4.88 (108)	0.47	4.91 (124)	820

Shown are TM score and rmsd of the predicted transmembrane β-barrel regions excluding the loops. Structure comparison is shown of top-ranked, best in top-5–ranked, and best possible model in the ensemble of models generated for 17 of the 19 proteins with a known structure in a blinded test such that no known structural information is used for folding. PagP and OprB are not applicable (NA) cases for folding, as the number of strands predicted by boctopus2 is incorrect. The blindly ranked 3D models are compared with the known structure based on TM score and rmsd value. TM score ranges from 0 to 1 and the closer the TM score is to 1, the more similar the model is to the known structure. If the two structures have a TM score greater than 0.5, then they are generally considered to be in the same protein fold. rmsd is another measure of 3D structure comparison where the positional coordinates of the two structures are compared and reported in angstroms with the superimposed region for which rmsd is reported in parentheses. rmsd closer to 0 signifies high structure similarity. Proteins for which TM score is greater than 0.5 are shown in boldface.

*Incorrect topology.

Table S2. Prediction performance in a blinded test on proteins with a known structure (full protein)

Protein (no. strands)	UniProt ID	TM score top-ranked model	rmsd top-ranked model	TM score best in top-5–ranked model	rmsd best in top-5–ranked model	TM score best possible model in ensemble	rmsd best possible model
1p4tA (8)	Q9RP17_NEIME	0.72	3.36 (142)	0.73	3.19 (142)	0.76	3.52 (152)
1kmpA (22)	FECA_ECOLI	0.3	7.23 (243)	0.51	5.91 (359)	0.24	7.55 (194)
3kvnA (12)	ESTA_PSEAE	0.54	4.27 (200)	0.54	4.24 (197)	0.57	3.83 (200)
2j1nA (16)	OMPC_ECOLI	0.51	5.43 (257)	0.52	5.16 (253)	0.58	5.14 (278)
3ohnA (24)	FIMD_ECOLI	0.43	6.25 (290)	0.43	6.91 (319)	0.48	6.03 (318)
4k3cA (16)	Q93PM2_HAEDC	0.47	6.50 (287)	0.47	6.50 (287)	0.51	5.73 (285)
4e1tA (12)	INVA_YERPS	0.38	5.42 (144)	0.38	5.40 (152)	0.43	5.87 (176)
1t16A (14)	FADL_ECOLI	0.48	4.97 (248)	0.49	4.84 (247)	0.5	5.00 (253)
3sybA (18)	Q9HVS0_PSEAE	0.51	5.37 (267)	0.51	5.37 (267)	0.52	5.64 (279)
2wjrA (12)	NANC_ECOLI	0.54	5.73 (173)	0.48	5.55 (170)	0.48	5.30 (163)
1thq (8)*	PAGP_ECOLI	NA	NA	NA	NA	NA	NA
1qd6C (12)	PA1_ECOLI	0.57	4.97 (204)	0.57	4.97 (204)	0.58	5.03 (205)
4q35A (26)	LPTD_ECOLI	0.35	7.89 (324)	0.35	7.48 (307)	0.39	7.94 (376)
1a0sP (18)	SCRY_SALTM	0.36	7.14 (248)	0.35	6.90 (239)	0.41	6.48 (265)
1tlyA (12)	TSX_ECOLI	0.41	6.18 (182)	0.45	6.02 (192)	0.45	6.02 (192)
2jk4A (19)	VDAC1_HUMAN	0.49	6.12 (231)	0.54	5.65 (241)	0.54	5.65 (241)
2ervA (8)	Q9HVD1_PSEAE	0.62	3.43 (130)	0.62	3.43 (130)	0.64	3.63 (136)
4gey (16)*	A5VZA8_PSEP1	NA	NA	NA	NA	NA	NA
2o4vA (16)	PORP_PSEAE	0.32	6.78 (204)	0.32	6.95 (207)	0.38	6.33 (230)

Shown are TM score and rmsd of full protein models including the loop regions. Structure comparison is shown of top-ranked, best in top-5–ranked, and best possible model in the ensemble of models generated for 17 of the 19 proteins with a known structure in a blinded test such that no known structural information is used for folding. PagP and OprB are not applicable (NA) cases for folding, as the number of strands predicted by boctopus2 is wrong. The blindly ranked 3D models are compared with the known structure based on TM score and RMSD value. TM score ranges from 0 to 1 and the closer the TM score is to 1, the more similar the model is to the known structure. If the two structures have a TM score greater than 0.5, then they are generally considered to be in the same protein fold. rmsd is another measure of 3D structure comparison where the positional coordinates of the two structures are compared and reported in angstroms with the superimposed region for which rmsd is reported in parentheses. rmsd closer to 0 signifies high structure similarity.

*Incorrect topology.

Table S3. Comparison of EVfold_bb- and tobmodel-generated 3D models of the transmembrane barrel region in a blinded test

	Tobmodel Cα 3D models		EVfold_bb 3D models							
	Top-ranked models		Top-ranked models		Best in top-5–ranked models		Best possible models		tobmodel top-ranked model	EVfold_bb top-ranked model
Protein (no. strands)	TM score	rmsd	TM score	rmsd	TM score	rmsd	TM score	rmsd	No. residue pairs in correct registration	
1p4tA (8)	0.45	4.03 (69)	**0.85**	1.59 (87)	**0.87**	1.5 (87)	**0.87**	1.50 (87)	6	50
1kmpA (22)	0.74	2.86 (202)	0.67	4.48 (217)	0.7	4.06 (216)	0.7	4.06 (216)	100	120
3kvnA (12)	0.61	3.72 (118)	**0.68**	3.12 (124)	**0.71**	3.15 (123)	**0.75**	2.50 (123)	32	68
2j1nA (16)	0.6	3.67 (135)	0.58	4.65 (152)	**0.65**	4.05 (152)	**0.68**	3.38 (152)	0	80
3ohnA (24)	0.53	4.39 (167)	0.46	4.77 (150)	0.51	5.41 (182)	**0.55**	5.07 (188)	30	83
4k3cA (16)	0.42	3.40 (94)	**0.48**	5.22 (141)	**0.48**	5.5 (146)	**0.53**	4.37 (146)	4	57
4e1tA (12)	0.58	3.54 (102)	0.45	4.49 (93)	0.45	4.63 (97)	0.48	3.57 (83)	13	28
1t16A (14)	0.63	3.63 (134)	**0.66**	3.86 (147)	**0.71**	3.33 (144)	**0.74**	2.93 (144)	93	76
3sybA (18)	0.7	2.83 (154)	0.62	4.15 (162)	0.62	4.15 (162)	0.63	4.08 (162)	23	70
2wjrA (12)	0.35	4.49 (72)	**0.5**	4.38 (109)	**0.51**	4.43 (109)	**0.52**	4.21 (109)	4	38
1thq (8)*	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
1qd6C (12)	0.35	4.52 (77)	**0.61**	3.5 (112)	**0.61**	3.5 (112)	**0.61**	3.50 (112)	5	34
4q35A (26)	0.44	4.86 (175)	0.36	6.54 (174)	0.36	5.79 (153)	0.42	6.25 (185)	5	51
1a0sP (18)	0.56	4.15 (154)	0.39	6.12 (139)	0.42	4.63 (111)	0.47	5.12 (144)	8	52
1tlyA (12)	0.48	4.83 (110)	0.41	5.06 (103)	0.46	4.87 (106)	0.46	4.45 (105)	32	30
2jk4A (19)	0.7	2.4 (149)	0.45	6.18 (157)	0.51	5.41 (162)	0.53	5.41 (161)	36	53
2ervA (8)	0.25	4.07 (36)	**0.67**	2.55 (79)	**0.67**	2.55 (79)	**0.67**	2.51 (79)	0	36
4gey (16)*	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2o4vA (16)	0.63	3.58 (140)	0.4	4.98 (109)	0.41	4.88 (108)	0.47	4.91 (124)	2	32

Comparison of the top-ranked Cα 3D models generated by tobmodel and top-ranked all-atom models generated by EVfold_bb shows that for the barrel region, although the TM scores are comparable, EVfold_bb on average predicts more residue pairs in correct registration (958) than tobmodel (393) out of 2,172. Proteins for which the TM score of EVfold_bb top-ranked model is better than the tobmodel 3D model are shown in boldface.

*Incorrect topology. Additionally, 65% and 41% of residue pairs are within plus or minus one residue of correct registration in EVfold_bb and tobmodel models, respectively.

Table S4. Topology prediction performance of boctopus2 for proteins in the cross-validation training and test datasets

PDB ID code	No. observed strands in the known structure	No. strands predicted in the whole input sequence	No. strands predicted in barrel region	No. predicted strands at their correct location
3prn	16	16	16	16
3jty	18	18	18	18
3kvn*	12	16	12	12
3fhh	22	22	22	22
3dwo	14	14	14	14
3dzm	8	8	8	8
3bs0	14	14	14	14
3csl	22	22	22	22
3a2s	16	16	16	16
2ysu	22	22	22	22
2vqi	24	24	24	24
2wjr	12	12	12	12
2qom	12	12	12	12
2qdz†	16	22	18	16
2por	16	16	16	16
2o4v*	16	18	16	16
2k0l	8	8	8	8
2mpr	18	18	18	18
2j1n	16	16	16	16
2iah	22	22	22	22
2iww	14	14	14	14
2f1v	8	8	8	8
2grx	22	22	22	22
2erv	8	8	8	8
1tly	12	12	12	12
1uyo	12	12	12	12
1t16	14	14	14	14
1qd6	12	12	12	12
1qj8	8	8	8	8
1p4t	8	8	8	8
1k24‡	10	10	10	9
1kmp	22	22	22	22
1i78†	10	14	14	10
1e54‡	16	16	16	14
1fep	22	22	22	22
1a0s	18	18	18	18

Boctopus2 is trained on a dataset of 36 TMBs with a known structure. The results are reported in a cross-validated manner performed such that all proteins that belong to the same family are always put together in the training or the test set.

*Correct number of strands in the barrel region.

†Correct number of strands, but a few predicted strands are shifted compared with location observed in the known structure.

‡Incorrect number of strands predicted in the barrel region.

Table S5. Topology prediction performance of boctopus2 on proteins not in boctopus dataset

PDB ID code	No. observed strands in the known structure	No. strands predicted in the whole input sequence	No. strands predicted in barrel region	No. predicted strands at their correct location
3syb	18	18	18	18
4k3c*	16	20	16	16
4e1t	12	14	12	12
3ohn	24	24	24	24
2jk4	19	20	19	19

For a protein not in the boctopus2 dataset, SVMs trained on the boctopus2 dataset were used to predict the location of each residue and generate the input profile for the hidden Markov model.

*For proteins with a long nonbarrel domain, the barrel region was predicted using a postprocessing step by using probabilities obtained from support vector machines as described in *Methods*.

Table S6. Distance constraints used to fold proteins in CNS

	Raw ECs			Residue pairs predicted to be hydrogen bonded			
Residue pair location	Cα(*i*)–Cα(*j*)	Cβ(*i*)–Cβ(*j*)	Side-chain heavy atom	N(*i*)–O(*j*)	O(*i*)–N(*j*)	Side-chain heavy atom	Cα(*i*)–Cα(*j*)
Adjacent strands, membrane region only*	NA	NA	NA	2.9 ± 0.3	2.9 ± 0.3	3.0 ± 1.0	5.2 ± 0.6
Nonadjacent strand–strand, membrane region only*	NA	NA	4.5 ± 2.0	NA	NA	NA	NA
Loop–loop or loop–strand	4.0 (+4, −3)	4.0 (+4, −3)	3.0 ± 1.0	NA	NA	NA	NA

Distance constraints are applied to the residue pairs identified to be hydrogen bonded. If two residues *i* and *j* are predicted to be hydrogen bonded, then distance constraints (in angstroms) are applied to their N–O, O–N, a pair of side-chain heavy atoms, and Cα–Cα atoms. Distance constraints are applied only to the side-chain atoms in the case that residue pairs are on nonadjacent β-strands. In other cases, distance constraints are applied to Cα–Cα, Cβ–Cβ, and a pair of side-chain heavy atoms. The pairwise list of side-chain heavy atoms on which the constraints are applied is provided in the supplementary data at cbio.mskcc.org/foldingproteins/transmembrane/betabarrels/.

*Predicted transmembrane β-strand region by boctopus2.

Table S7. Intrastrand distance constraints applied to maintain the predicted transmembrane β-strand secondary structure

Atom type in residue *i*	Atom type in residue *j*	Sequence separation between residue *i* and residue *j*	Distance constraint applied, Å
O	N	1	2.25 ± 0.02
O	N	2	4.55 ± 0.35
O	N	3	7.99 ± 0.4
O	N	4	11.11 ± 0.55
O	N	5	14.41 ± 0.73
O	N	6	17.46 ± 0.97
O	N	7	20.58 ± 1.24
O	N	8	23.47 ± 1.59
O	N	9	26.19 ± 2.13
O	N	10	28.44 ± 2.94
Cα	Cα	1	3.79 ± 0.03
Cα	Cα	2	6.81 ± 0.32
Cα	Cα	3	10.23 ± 0.4
Cα	Cα	4	13.4 ± 0.68
Cα	Cα	5	16.63 ± 0.83
Cα	Cα	6	19.71 ± 1.16
Cα	Cα	7	22.74 ± 1.35
Cα	Cα	8	25.63 ± 1.79
Cα	Cα	9	28.23 ± 2.26
Cα	Cα	10	30.53 ± 3.13
Cβ	Cβ	1	5.71 ± 0.19
Cβ	Cβ	2	6.99 ± 0.78
Cβ	Cβ	3	11.01 ± 0.61
Cβ	Cβ	4	13.65 ± 1.3
Cβ	Cβ	5	17.05 ± 1.06
Cβ	Cβ	6	20 ± 1.76
Cβ	Cβ	7	23.05 ± 1.49
Cβ	Cβ	8	26.05 ± 2.1
Cβ	Cβ	9	28.63 ± 2
Cβ	Cβ	10	31.02 ± 3.39
Cα	O	1	4.86 ± 0.23
Cα	O	2	8.28 ± 0.35
Cα	O	3	11.37 ± 0.55
Cα	O	4	14.71 ± 0.71
Cα	O	5	17.76 ± 1.04
Cα	O	6	20.95 ± 1.23
Cα	O	7	23.9 ± 1.56
Cα	O	8	26.81 ± 1.92
Cα	O	9	29.4 ± 2.43
Cα	O	10	31.49 ± 3.28
N	N	1	3.52 ± 0.12
N	N	2	6.78 ± 0.26
N	N	3	10.13 ± 0.39
N	N	4	13.37 ± 0.58
N	N	5	16.58 ± 0.81
N	N	6	19.69 ± 1.07
N	N	7	22.71 ± 1.34
N	N	8	25.6 ± 1.73
N	N	9	28.25 ± 2.28
N	N	10	30.45 ± 3.16
O	O	1	4.67 ± 0.3
O	O	2	6.84 ± 0.33
O	O	3	10.58 ± 0.59
O	O	4	13.44 ± 0.64
O	O	5	16.83 ± 1.03
O	O	6	19.76 ± 1.13
O	O	7	22.92 ± 1.57
O	O	8	25.72 ± 1.8
O	O	9	28.41 ± 2.44
O	O	10	30.45 ± 3.13

Distance constraints (in angstroms) are applied to intrastrand residues to maintain the structure of predicted transmembrane β-strands. For each residue pair *i* and *j* in a predicted strand (where *j* ranges from *i* + 1 to *i* + 10), distance constraints are applied to the listed atom-type combinations. Default values for atom–atom distance between residue pairs separated from each other in sequence by 1–10 residues were calculated from known TMB structures.