Supplemental Data

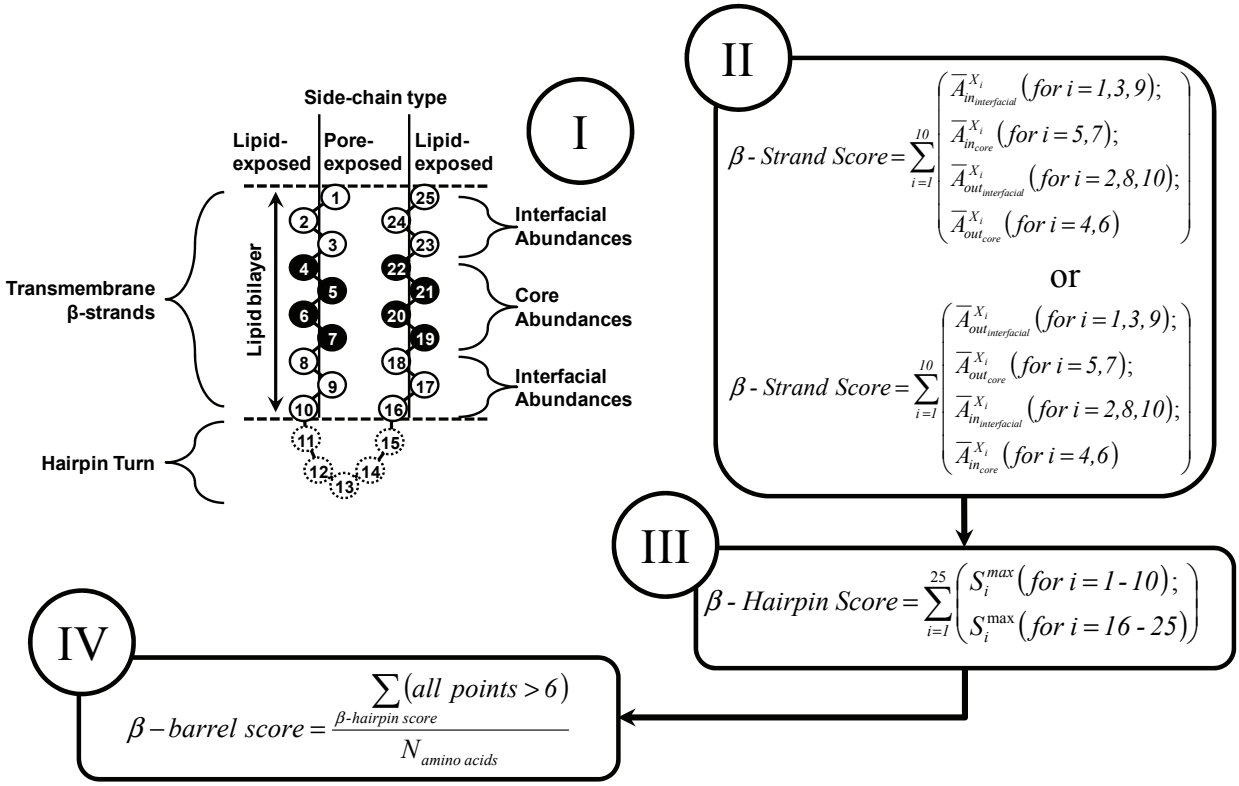

FIGURE S1 The Freeman-Wimley algorithm. The schematic shows a β-hairpin, the major structural subunit of a TMBB, spanning a lipid bilayer. The transmembrane strands are the target of the initial analysis, which are predicted by the β-strand score, and the β-strand score is used to predict β-hairpins. I) The abundance values are assigned to each residue in a 10-residue sliding window of the sequence. The abundances are assigned in an internal/external dyad repeat pattern, where ($\bar{A}_{in}^{Xi}$) and ($\bar{A}_{out}^{Xi}$) are the respective abundances (see Table S2). Also note that the 6 terminal residues in the 10-residue window are given interfacial abundances while the central 4 residues are given core abundances. II) The β-strand score is the sum of the abundances in the 10-residue sliding window, which increments through the sequence one residue at a time. There are two assignments with opposite registers of the dyad in/out pattern, and whichever window has the greatest sum is taken as the β-strand score for that particular position. The β-strand score is given for the median (5.5^{th}) residue in the window, where a peak in the score plot for a given sequence represents the middle of a β-strand (see Figure 2). III) The β-hairpin score is a secondary analysis of the β-strand score where the maximum β-strand score value (S_i^{max}) found in positions 1-10 is added to the S_i^{max} found in positions 16-25 of a 25-residue sliding window of the β-strand score. The purpose is to look for two peaks in the β-strand score (predicted TM β-strands), which are separated by five residues (hairpin turn). In this way it is possible to identify any potential TM β-hairpins, which are the structural subunit of TMBBs, in a sequence. IV) Finally, a summary value of the entire topology prediction is given by the β-barrel score where all points in the β-hairpin score whose value is greater than 6 are summed and divided by the length of the sequence. This is a means of expressing the density of predicted β-hairpins in a sequence, which is much greater in TMBBs than in other proteins (see Figure S2).

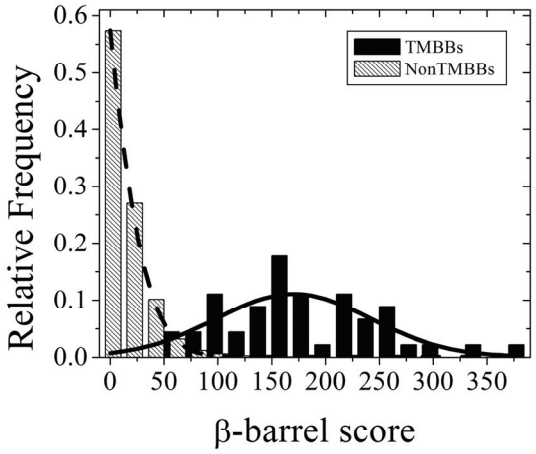

FIGURE S2 β-barrel score distributions of TMBBs and non-TMBBs from the NRPDB. There is little relative overlap between the scores of the two classes of proteins, which suggests that the β-barrel score from the Freeman-Wimley algorithm can be used to distinguish TMBBs from other types of proteins.

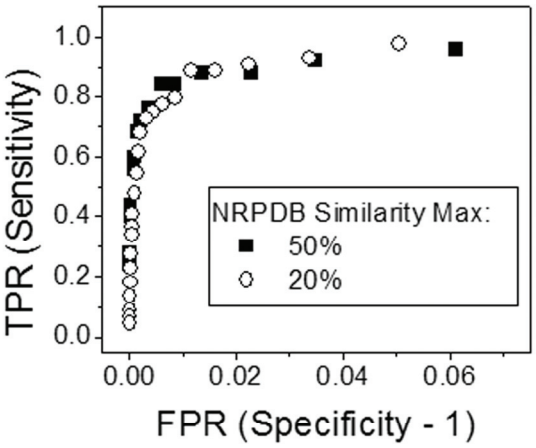

FIGURE S3 Comparison of ROC curves from the analysis of NRPDBs with different sequence similarity cutoffs. The AUC (area under the curve) for the 50% NRPDB is 0.974 and for the 20% NRPDB is 0.983. The difference between the two is insignificant (p = 0.67). The p-value was calculated using the "Significance of the Difference between the Areas under Two Independent ROC Curves" tool found at the following web site: http://faculty.vassar.edu/lowry/VassarStats.html.

Table S1. New TMBB structures analyzed for abundance data

PDB ID[a]	*Protein*	*Organism*	*Architecture*	*Strands*	*Reference*
1i78	OmpT	*Escherichia coli*	Monomer	12	Vandeputte-Rutten, *et al.*, 2001
1k24	OpcA	*Neisseria meningitidis*	Monomer	10	Prince, *et al.*, 2002
1kmo	FecA	*Escherichia coli*	Monomer	22	Ferguson, *et al.*, 2000
1mm4	PagP	*Escherichia coli*	Monomer	8	Hwang, *et al.*, 2002
1nqe	BtuB	*Escherichia coli*	Monomer	22	Chimento, *et al.*, 2003
1p4t	NspA	*Neisseria meningitidis*	Monomer	8	Vandeputte-Rutten, *et al.*, 2003
1t16	FadL	*Escherichia coli*	Monomer	14	van den Berg, *et al.*, 2004
1tlw	TSX	*Escherichia coli*	Monomer	12	Ye and van den Berg, 2004
1wp1	OprM	*Pseudomonas aeruginosa*	Trimeric single barrel[b]	4	Akama, *et al.*, 2004
1xkh	FpvA	*Pseudomonas aeruginosa*	Monomer	22	Cobessi, *et al.*, 2005a
1xkw	FptA	*Pseudomonas aeruginosa*	Monomer	22	Cobessi, *et al.*, 2005b
2erv	PagL	*Pseudomonas aeruginosa*	Monomer	8	Rutten, *et al.*, 2006
2f1c	OmpG	*Escherichia coli*	Monomer	14	Subbarao and van den Berg, 2006
2f1t	OmpW	*Escherichia coli*	Monomer	8	Hong, *et al.*, 2006
2hdf	Cir	*Escherichia coli*	Monomer	22	Buchanan, *et al.*, 2007
2o4v	OprP	*Pseudomonas aeruginosa*	Trimer	16	Moraes, *et al.*, 2007
2odj	OprD	*Pseudomonas aeruginosa*	Trimer	18	Biswas, *et al.*, 2007
2qdz	FhaC	*Bordatella pertussis*	Monomer	16	Clantin, *et al.*, 2007
2qom	EspP	*Escherichia coli*	Monomer	12	Barnard, *et al.*, 2007
2qtk	Opdk	*Pseudomonas aeruginosa*	Trimer	18	Biswas, *et al.*, 2008
2vqi	PapC	*Escherichia coli*	Dimer	24	Remaut, *et al.*, 2008
3bry	TbuX	*Ralstonia pickettii*	Monomer	14	Hearn, *et al.*, 2008

[a] The PDB ID is the four-character code used to identify each structure in the PDB.
[b] The architecture of OprM consists of three monomers each with four β-strands that form a single-barreled homo-trimer.

Table S2. Updated relative amino acid abundances

	*Abundance on external surface**			*Abundance on internal surface**		
Amino Acid	*Bilayer*[a]	*Interface*[b]	*Core*[c]	*Bilayer*[a]	*Interface*[b]	*Core*[c]
Ala	1.05	0.69	1.42	0.74	0.79	0.69
Arg	0.17	0.33	0.01	1.19	1.39	0.99
Asn	0.56	0.92	0.20	1.82	1.48	2.16
Asp	0.21	0.35	0.08	1.28	1.56	1.00
Cys[*]	0.02	0.02	0.02	0.02	0.02	0.02
Gln	0.30	0.51	0.10	1.27	1.48	1.07
Glu	0.07	0.13	0.01	1.04	1.02	1.07
Gly	0.81	0.53	1.09	2.00	1.51	2.48
His	0.66	0.94	0.39	0.59	0.78	0.40
Ile	1.19	1.18	1.20	0.27	0.30	0.24
Leu	1.66	1.14	2.17	0.26	0.29	0.23
Lys	0.20	0.39	0.02	1.03	1.04	1.01
Met	0.70	0.53	0.86	0.71	0.80	0.61
Phe	2.56	3.18	1.94	0.58	0.62	0.54
Pro	0.53	0.51	0.54	0.33	0.52	0.13
Ser	0.42	0.52	0.31	1.97	1.76	2.18
Thr	0.87	0.80	0.95	1.82	1.90	1.74
Trp	3.39	5.61	1.18	0.53	0.53	0.54
Tyr	4.42	5.59	3.25	1.81	1.60	2.03
Val	1.76	1.39	2.14	0.39	0.41	0.37

[a] Bilayer refers to the entire thickness of a lipid bilayer, which is 27 Å.
[b] Interface refers to the portion of the bilayer plane that is greater than 6.5 Å and up to 13.5 Å from the center of the plane.
[c] Core refers to the portion of the bilayer plane that is between 0 and 6.5 Å.
[*] There were no cysteine residues found in any of the β-strands so an arbitrary value of 0.02 was assigned.

Table S-2 shows the relative amino acid abundances which were updated with the latest structural information. Relative abundance is calculated as described by Wimley, 2002. Briefly, the expected abundance for a given amino acid, $\bar{f}_x^i$, was calculated as a weighted average of genomic abundances, f_x^i.

$$\bar{f}_x^i = \sum_i w_i f_x^i$$

The weights, w_i, were calculated for organism, i, by dividing the number of amino acids contributed by organism i, n_i, by the total number of amino acids, n_{total}, in the database.

$$w_i = \frac{n_i}{n_{total}}$$

The observed or raw abundances were divided by the expected abundances to calculate the relative amino acid abundances.

Table S3. Prediction efficiency comparison among several algorithms [a]

Algorithm	*Sensitivity*	*Specificity*	*Accuracy* [b]	*MCC* [b]	*Reference*
Freeman-Wimley [c]	0.711	0.999	0.998	0.667	This study
Original [c]	0.489	0.998	0.997	0.477	This study
Bayesnet [d]	0.760	0.935	0.901	(0.686)	Gromiha and Suwa, 2006
Naïve Bayes [d]	0.832	0.892	0.881	(0.661)	Gromiha and Suwa, 2006
Logistic function [d]	0.668	0.943	0.891	(0.635)	Gromiha and Suwa, 2006
Neural network [d]	0.793	0.938	0.910	(0.715)	Gromiha and Suwa, 2006
RBF network [d]	0.793	0.930	0.903	(0.699)	Gromiha and Suwa, 2006
k-nearest neighbor (k-NN) [d]	0.851	0.850	0.850	(0.611)	Gromiha and Suwa, 2006
Bagging meta learning [d]	0.639	0.947	0.888	(0.620)	Gromiha and Suwa, 2006
Classification via regression [d]	0.688	0.934	0.887	(0.630)	Gromiha and Suwa, 2006
Decision tree J4.8 [d]	0.673	0.945	0.893	(0.643)	Gromiha and Suwa, 2006
NBTree [d]	0.692	0.941	0.893	(0.648)	Gromiha and Suwa, 2006
Partial decision tree [d]	0.678	0.945	0.894	(0.647)	Gromiha and Suwa, 2006
Class variance	0.855	0.925	(0.921)	(0.564)	Liu, *et al.*, 2003
k-NN	0.864	0.988	0.971	0.876	Hu and Yan, 2008
BOMP (with BLAST search) [d]	0.798	0.974	0.950	0.787	Hu and Yan, 2008
TMB-Hunt (with evolutionary information) [d]	0.815	0.957	0.937	0.747	Hu and Yan, 2008
PRED-TMBB [d]	0.891	0.604	0.643	0.342	Hu and Yan, 2008
PROFtmb [d]	0.714	0.960	0.926	0.684	Hu and Yan, 2008

[a] Comparison is not statistically stringent due to the analysis of different data sets
[b] Values in parentheses were calculated from the available data in the reference
[c] NRPDB analyzed to obtain performance measurements
[d] Algorithms were developed by other groups but performance was tested by listed reference

Table S4. Structural classification of NRPDB and analysis results*

		Threshold[a]		
Structural classification	N_{NRPDB}[b]	*45*	*90*	*134*
TMBBs[c]	48	46 / 37	38 / 36	31 / 30
All α[d]	1577	17 / 2	3 / 0	1 / 0
All β[d]	1813	170 / 50	42 / 12	8 / 3
α/β[d]	2561	113 / 13	17 / 1	4 / 0
α+β[d]	2768	77 / 32	6 / 3	1 / 1
Multi-domain[d]	218	12 / 2	1 / 1	0 / 0
Membrane[d]	153	28 / 5	6 / 0	0 / 0
Small proteins[d]	397	4 / 2	0 / 0	0 / 0
Coiled-coil[d]	101	4 / 0	0 / 0	0 / 0
Low-resolution structures[d]	284	14 / 6	1 / 0	0 / 0
Peptides[d]	12	0 / 0	0 / 0	0 / 0
Designed proteins[d]	15	1 / 0	0 / 0	0 / 0
Unclassified	4291	159 / 49	28 / 9	4 / 1
Total non-TMBBs	14190	599 / 161	104 / 26	18 / 5

*Results from the F-W algorithm (left of slash) and the F-W algorithm with MRS (right of slash) are compared
[a]Positive prediction threshold (β-barrel score)
[b]Number of each type of sequence in the NRPDB
[c]Number of TMBBs identified at each threshold using a given algorithm. The TMBBs are the true positives predicted and all others are false positives.
[d]Structural class identified in SCOP database version 1.75 (Murzin, *et al.*, 1995).

Table S5. Impact of training TMBBs on prediction results

	50%NRPDB[a]			20%NRPDB[a]		
	Test Only[b]	Test+Training[c]	P-value[d]	Test Only[b]	Test+Training[c]	P-value[d]
ROC-AUC[e]	0.915	0.975	0.285	0.961	0.984	0.531
$Accuracy_{90}$[f]	0.992	0.992	ND[g]	0.991	0.990	ND[g]

[a]Non-redundant protein database with similarity cutoff of either 50% or 20%
[b]Sequences tested were not included in the training set
[c]Sequences include some members which were a part of the training set
[d]P-value determined for the significance of the difference between including and excluding training sequences from the NRPDB analysis
[e]Receiver Operating Characteristic-Area Under the Curve
[f]Accuracy of all predictions made in each data set with a β-barrel score threshold of 90
[g]ND = not determined

References

Akama, H., M. Kanemaki, *et al.* (2004). "Crystal structure of the drug discharge outer membrane protein, OprM, of Pseudomonas aeruginosa: dual modes of membrane anchoring and occluded cavity end." J Biol Chem **279**(51): 52816-9.

Barnard, T. J., N. Dautin, *et al.* (2007). "Autotransporter structure reveals intra-barrel cleavage followed by conformational changes." Nat Struct Mol Biol **14**(12): 1214-20.

Biswas, S., M. M. Mohammad, *et al.* (2008). "Crystal structure of the outer membrane protein OpdK from Pseudomonas aeruginosa." Structure **16**(7): 1027-35.

Biswas, S., M. M. Mohammad, *et al.* (2007). "Structural insight into OprD substrate specificity." Nat Struct Mol Biol **14**(11): 1108-9.

Buchanan, S. K., P. Lukacik, *et al.* (2007). "Structure of colicin I receptor bound to the R-domain of colicin Ia: implications for protein import." Embo J **26**(10): 2594-604.

Chimento, D. P., A. K. Mohanty, *et al.* (2003). "Substrate-induced transmembrane signaling in the cobalamin transporter BtuB." Nat Struct Biol **10**(5): 394-401.

Clantin, B., A. S. Delattre, *et al.* (2007). "Structure of the membrane protein FhaC: a member of the Omp85-TpsB transporter superfamily." Science **317**(5840): 957-61.

Cobessi, D., H. Celia, *et al.* (2005a). "The crystal structure of the pyoverdine outer membrane receptor FpvA from Pseudomonas aeruginosa at 3.6 angstroms resolution." J Mol Biol **347**(1): 121-34.

Cobessi, D., H. Celia, *et al.* (2005b). "Crystal structure at high resolution of ferric-pyochelin and its membrane receptor FptA from Pseudomonas aeruginosa." J Mol Biol **352**(4): 893-904.

Ferguson, A. D., V. Braun, *et al.* (2000). "Crystal structure of the antibiotic albomycin in complex with the outer membrane transporter FhuA." Protein Sci **9**(5): 956-63.

Gromiha, M. M. and M. Suwa (2006). "Discrimination of outer membrane proteins using machine learning algorithms." Proteins **63**(4): 1031-7.

Hearn, E. M., D. R. Patel, *et al.* (2008). "Outer-membrane transport of aromatic hydrocarbons as a first step in biodegradation." Proc Natl Acad Sci U S A **105**(25): 8601-6.

Hong, H., D. R. Patel, *et al.* (2006). "The outer membrane protein OmpW forms an eight-stranded β-barrel with a hydrophobic channel." J Biol Chem **281**(11): 7568-77.

Hu, J. and C. Yan (2008). "A method for discovering transmembrane β-barrel proteins in Gram-negative bacterial proteomes." Comput Biol Chem.

Hwang, P. M., W. Y. Choy, *et al.* (2002). "Solution structure and dynamics of the outer membrane enzyme PagP by NMR." Proc Natl Acad Sci U S A **99**(21): 13560-5.

Liu, Q., Y. Zhu, *et al.* (2003). "Identification of β-barrel membrane proteins based on amino acid composition properties and predicted secondary structure." Comput Biol Chem **27**(3): 355-61.

Moraes, T. F., M. Bains, *et al.* (2007). "An arginine ladder in OprP mediates phosphate-specific transfer across the outer membrane." Nat Struct Mol Biol **14**(1): 85-7.

Murzin, A. G., S. E. Brenner, *et al.* (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures." J Mol Biol **247**(4): 536-40.

Prince, S. M., M. Achtman, *et al.* (2002). "Crystal structure of the OpcA integral membrane adhesin from Neisseria meningitidis." Proc Natl Acad Sci U S A **99**(6): 3417-21.

Remaut, H., C. Tang, *et al.* (2008). "Fiber formation across the bacterial outer membrane by the chaperone/usher pathway." Cell **133**(4): 640-52.

Rutten, L., J. Geurtsen, *et al.* (2006). "Crystal structure and catalytic mechanism of the LPS 3-O-deacylase PagL from Pseudomonas aeruginosa." Proc Natl Acad Sci U S A **103**(18): 7071-6.

Subbarao, G. V. and B. van den Berg (2006). "Crystal structure of the monomeric porin OmpG." J Mol Biol **360**(4): 750-9.

van den Berg, B., P. N. Black, *et al.* (2004). "Crystal structure of the long-chain fatty acid transporter FadL." Science **304**(5676): 1506-9.

Vandeputte-Rutten, L., M. P. Bos, *et al.* (2003). "Crystal structure of Neisserial surface protein A (NspA), a conserved outer membrane protein with vaccine potential." J Biol Chem **278**(27): 24825-30.

Vandeputte-Rutten, L., R. A. Kramer, *et al.* (2001). "Crystal structure of the outer membrane protease OmpT from Escherichia coli suggests a novel catalytic site." Embo J **20**(18): 5033-9.

Wimley, W. C. (2002). "Toward genomic identification of β-barrel membrane proteins: composition and architecture of known structures." Protein Sci **11**(2): 301-12.

Ye, J. and B. van den Berg (2004). "Crystal structure of the bacterial nucleoside transporter Tsx." Embo J **23**(16): 3187-95.