PNAS

General Features of Transmembrane Beta Barrels From a Large Database

Daniel Montezano[a], Rebecca Bernstein[a,1], Matthew M. Copeland[a], Joanna S. G. Slusky*[,a, b]

[a] Center for Computational Biology, University of Kansas, Lawrence, KS 66045

[b] Department of Molecular Biosciences, University of Kansas, Lawrence, KS 66045

[1] Currently at University of California, Berkeley, CA 94720.

* To whom correspondence should be addressed. Email: slusky@ku.edu

Author Contributions: J.S.G.S. and D.M. designed the research; D.M., and R.B. performed the research, analyzed the data and wrote the paper; M.M.C. contributed software tools. J.S.G.S and D.M. edited the paper.

Keywords: transmembrane β-barrel, outer membrane protein, database, contact map, β-signal

Abstract

Large datasets contribute new insights to subjects formerly investigated by exemplars. We used co-evolution data to create a large, high-quality database of transmembrane β-barrels (TMBB). By applying simple feature detection on generated evolutionary contact maps, our method (IsItABarrel) achieves 95.88% balanced accuracy when discriminating among protein classes. Moreover, comparison with IsItABarrel revealed a high rate of false positives in previous TMBB algorithms. In addition to being more accurate than previous datasets, our database (available online) contains 1,938,936 bacterial TMBB proteins from 38 phyla respectively 17 and 2.2 times larger than the previous sets TMBB-DB and OMPdb. We anticipate that due to its quality and size the database will serve as a useful resource where high quality TMBB sequence data is required. We found that TMBBs can be divided into 11 types, three of which have not been previously reported. We find tremendous variance in proteome percentage among TMBB-containing organisms with some using 6.79% of their proteome for TMBBs and others using as little as 0.27% of their proteome. The distribution of the lengths of the TMBBs is suggestive of previously hypothesized duplication events. In addition, we find that the C-terminal β-signal varies among different classes of bacteria though it is most commonly LGLGYRF. However, this β-signal is only characteristic of prototypical TMBBs. The nine non-prototypical barrel types have other C-terminal motifs, and it remains to be determined if these alternative motifs facilitate TMBB insertion or perform any other signaling function.

Significance Statement

Outer membrane proteins (OMPs) are the main component of Gram negative bacterial outer membranes and are frequently vaccine targets. We created an algorithm that identifies bacterial OMPs from sequence. The quality of our algorithm allows us to identify most OMPs (>1.9 million) from prokaryotic genomes including >270,000 unrelated to previously structurally-resolved OMPs. We identify eleven types of OMPs in our database. The largest type's signal sequence—used for targeting the membrane-insertion machinery—varies by phylogenetic class. All other types of OMPs have unrelated signal sequences, raising new questions of how these proteins fold. Our web-accessible database will allow for further exploration of the varieties of outer membrane proteins to uncover new targets for vaccine development.

Introduction

Comparative studies of organisms, genomes, and protein families are dependent on the quality and size of biological databases. These studies form the basis for answering questions about the evolutionary history of different protein families and for identifying and classifying the diversity of protein structure and function. The proteins found in the outer membrane of Gram-negative bacteria, generally have a highly similar β-barrel fold (1), which has evolved mostly by replication of an ancestral hairpin motif. (2, 3) Despite the significant structural similarity, this fold emerges from a rich diversity of primary sequences (4). Attempts to classify the barrel fold into families has met with a number of challenges that can be observed in the diversity of current classifications performed in public structural databases such as SCOPe. The functional roles of transmembrane β-barrels (TMBBs) are also diverse, including nutrient uptake, membrane stabilization, catalysis, cell adhesion, cell signaling, and efflux. (5)

The hydrophobic environment of the outer membrane imposes constraints on the sequence and topology of TMBB proteins. Since the outside of barrels is in contact with the membrane and the inside of the barrel with the water-filled lumen, strand residues that point into the channel are usually hydrophilic, while residue side chains pointing outside into the membrane are hydrophobic and aromatic. Such constraints induce an imperfect hydrophobicity alternation pattern on residues of each strand. (6) Also, there are significant differences in amino acid abundance at different depths across the membrane, with more hydrophobic residues in the central region of the membrane. (7-9) Attempts have been made to computationally identify TMBBs using amino acid composition (10-12), using specific knowledge about TMBBs and their environment (7, 13, 14) and by identification of sequence motifs. (15, 16) In addition to classification within the fold, several computational methods also predict more specific features of TMBB topology such as strand number and which parts of the protein localize to the periplasm, to the outside of the cell in the extracellular milieu or are embedded in the membrane. (14, 17-21)

Homology has been explored and combined with many of these previous methods to augment available data and make explicit use of evolutionary cues thereby improving TMBB classification and prediction. (22-24) Machine learning methods have also been employed to the tasks of TMBB topology prediction and fold classification, including feed-forward neural networks and SVMs (25), recurrent neural networks (26), hidden Markov models (27, 28), extreme learning machines (29) and ensemble approaches. (26, 30)

In two instances algorithms were applied to establish large-scale homologous databases of TMBBs or to perform proteome-wide search for new TMBBs. The TMBB-DB (31) contains 1,881,712 sequences collected from 600 different bacterial proteomes with sequences ranked by their likelihood of encoding a TMBB. The OMPdb (32) contains 1,198,558 (as of Jan 2021) protein sequences predicted to be TMBBs by the PRED-TMBB2 algorithm.

These previous methods were dependent on the availability of experimentally resolved structures, which is scarce for β-barrels. Methods that used sequences were trained on a comparatively small set of putative TMBB sequences from which sequence features were extracted. Each method used different aspects of protein information attempting to extract a set of features that would provide high quality classification and topology prediction.

IsItABarrel applies heuristics for hairpin interactions and barrel closing contacts to optimally use the information embedded in contact maps. Contact maps embed several aspects of the protein such as specific ordering of amino acids in sequence, secondary structure, tertiary structure, amino acid abundances, probability distribution over contacts and multiple sequence alignments. The accuracy of IsItABarrel demonstrates that applying higher level heuristics to the informational content of contact maps provides improved prediction outcomes.

Co-evolution information is obtained from analyzing covariation of residues in an alignment of protein sequences that are homologous to a protein of interest. Strong statistical co-variation of two or more residues indicates a possible co-evolutionary event, where one set of mutations is compensated by another set of mutations. (33) A co-evolutionary event can be interpreted as evidence that the residues are in proximity, from which contact probabilities can be derived (34). If higher order interactions can be reliably extracted, contacts can be predicted with higher accuracy. (35) Evolutionary contact maps are 2D representations that show the likelihood of two residues in a protein sequence being in contact and may be used for predicting protein structure or for refining homology models. Here we use contact information to discriminate between the barrel fold and other protein folds.

Although contact information can be extracted from protein structure models, generated for example, by AlphaFold (36), this solution still does not provide a reliable way to categorize the predicted structures as barrels. We would still require an algorithm that could analyze these predicted structures in detail and decide if they represent a β-barrel structure or not. Despite recent efforts to categorize AlphaFold folds (37) TMBBs remain scattered among many fold families making it difficult to track down all instances of the fold.

Here we present a new methodology called IsItABarrel that uses co-evolution information and extracts discriminative features from contact maps for the identification of TMBBs from sequences. We apply IsItABarrel to identify 1,938,936 TMBB proteins from 14,366 bacterial organisms. In addition, we used IsItABarrel to search 2,959 bacterial proteomes spanning 78 taxonomic classes (Table S1) and 38 phyla (Table S2) for TMBBs and provide updated estimates of the TMBB content in these organisms. We also analyze general features of our new, large TMBB dataset and find evolutionary phyla-specific β-signals. Finally, we find eleven different types of TMBBs each with their own length distribution, organism preference, and C-terminal sequence motif. Three of the eleven TMBB types have not been previously documented.

Results

Algorithm for TMBB Prediction

β-barrels such as those found in bacterial outer membranes all have an up-and-down meander topology resulting in contact maps that can be distinguished from other proteins (Figs. 1 A and B). The two features that nearly definitively identify TMBBs from contact maps are strand-strand interactions (Fig. 1C) and closing contacts (Fig. 1D). Our algorithm, called IsItABarrel, identifies TMBBs by either the existence of a closing contact or the existence of at least four β-strand contacts of 16-25 contact map residues (Fig. 1E and Fig. S1). We use the Matthew's correlation coefficient (MCC) (38) to measure algorithm performance because it is less skewed by imbalanced datasets. We validated IsItABarrel on a set of 1,121 proteins with solved crystal structures, 121 of which are TMBBs and 1000 are non-TMBBs, yielding 0.921 MCC (95.88% balanced accuracy) with eight instances of false positives and nine instances of false negatives

(Table 1, left). We compared IsItABarrel with other predictors used to create previous datasets: the Freeman-Wimley algorithm that was used to create the TMBB-DB database (13), PRED-TMBB2 (20) that was used to create the OMPdb (32) and also with the BOMP predictor (15). We find that IsItABarrel has the highest MCC and the fewest false positives (Table 1 and Fig. S2), thereby allowing it to generate the most reliable dataset.

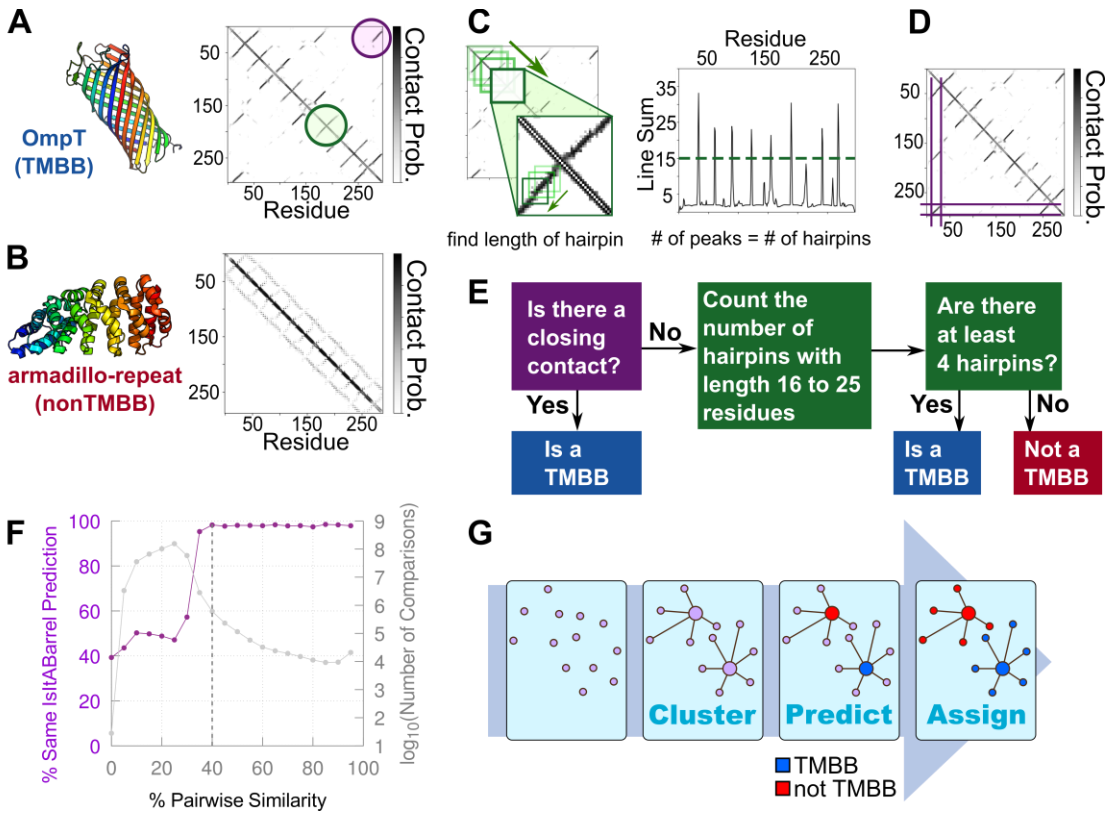

Figure 1. The IsItABarrel Algorithm. (A and B) Structures are colored from N to C-terminus using rainbows from blue to red. Probability of contact is indicated by pixel grey value, with darker pixel indicating stronger contact between residues in *x* and *y* axes. **(A)** OmpT (PDB ID: 1I78) (left) and corresponding contact map predicted by RaptorX-Contact (right). The map shows typical features of β-barrel proteins, anti-parallel contacts between neighboring strands (green circle) and the closing contact between the first and last strand of the barrel (purple circle). **(B)** Structure of the non-TMBB armadillo-repeat protein (PDB ID: 4V3O) (left) and corresponding contact map predicted by RaptorX-Contact (right) without the typical features found in maps of β-barrels. **(C)** Strand-strand interaction (β-hairpin) detection. Map is scanned along the main diagonal in search of β-hairpins. Detected β-hairpins are characterized in terms of the length of their contact network (inset). The sum of the contact probabilities along a successfully detected β-hairpin is compared against the threshold for inclusion in the final TMBB classification step. **(D)** A closing contacts line is detected that aligns with the first and last strands **(E)** Classification by IsItABarrel is a two-step process requiring either a closing contact or four hairpins. **(F)** Sequences sharing more than 40% identity tend to have the same IsItABarrel classification. Percentage of pairs that share the same IsItABarrel prediction (y-axis) for each bin of percent similarity (x-axis) (purple), total number of comparisons ($\log_{10}$) at each similarity level (gray). Dashed line at 40% pairwise sequence similarity. **(G)** Steps the algorithm used to assign IsItABarrel predictions by similarity. A large set of sequences is clustered by sequence similarity. IsItABarrel predictions are made for the cluster representatives. Cluster members are assigned the same prediction as their cluster representative.

Table 1. Comparison of predictors on a validation set.

	NONTMBB1K (121 TMBBs / 1000 non-TMBBs)				
Algorithm	**MCC**	**TP**	**TN**	**FP**	**FN**
IsItABarrel	0.926	113	992	8	8
BOMP	0.846	112	973	27	9
FW-BB Analysis	0.728	102	953	47	19
PRED-TMBB2	0.796	118	947	53	3

NONTMBB1K, includes 121 non-homologous TMBBs as positive examples and 1,000 non-homologous non-TMBBs as the negative set. IsItABarrel (top row) provides the lowest number of false positives and intermediate performance for true positives. TP – true positives; TN – true negatives; FP – false positives; FN – false negatives. Matthews' Correlation Coefficient (MCC) calculated as: MCC=(TP*TN – FP*FN)/ SQRT((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)).

Prediction of a Large Dataset with IsItABarrel

Map creation is a computationally demanding process. To increase the prediction scale, we determined the sequence similarity level that maintained consistent IsItABarrel classification. We determined that new maps do not need to be made for proteins with 40% or greater pairwise sequence similarity as homologs above that threshold have above a 95% similar IsItABarrel prediction (Fig. 1F), a success rate similar to the balanced accuracy of our algorithm overall. We determined this metric using the TMBB29183 dataset (Table S3) by evaluating 425,809,153 pairwise sequence comparisons and obtaining a distribution of sequence similarity values ranging from 4% to 99.9%. Similar IsItABarrel evaluations for proteins with high sequence similarity also indicate that our algorithm is robust for identifying bacterial TMBBs as more related proteins are likely to have the same fold.

Expanding the Dataset

By combining and evaluating the proteins in previous datasets (OMP-DB and TMBB-DB) we created a database of 514,728 bacterial TMBB proteins. We then undertook a more thorough search of genetic space to uncover previously unannotated bacterial TMBBs.

We considered 2,959 NCBI prokaryote reference and representative organisms. These proteomes span 78 classes and 35 phyla (Tables S2 and S3). Many of these organisms were not diderms and did not have any TMBBs. We found consistent representation of TMBBs (*i.e.,* TMBBs in ten or more organisms) in 20 classes. Six of the classes were in two phyla (Bacteroidetes and Firmicutes) and within those phyla they had consistent enough representation of TMBBs that we could represent them as phyla rather than individual classes. Thus, our analysis comprises 2 phyla and 14 classes (Fig. 2). TMBB-containing organismal categories included those that are known to include monoderms (Actinomycetia and Firmicutes). Other classes, such as Thermotogae and Deinococci, are diderms but usually contain atypical membrane compositions or thicker peptidoglycan layers which set them apart from other diderms. (39)

To find possible new TMBBs in those TMBB-containing proteomes, we evaluated all proteins from the largest organism (in terms of proteome size) in each class/phylum with IsItABarrel. In this search a total of 60,087 new protein sequences were found to be potential TMBBs, producing a two-fold increase over what had previously been found (Figs. 2 A and B, gray). To find the homologs of these new-found TMBBs, we performed a new search starting with the

non-redundant protein database from NCBI. This search resulted in our final set of predicted TMBBs we call IsItABarrelDB, containing 1,938,936 bacterial TMBB sequences, out of the total set of known proteins (~230 million sequences).

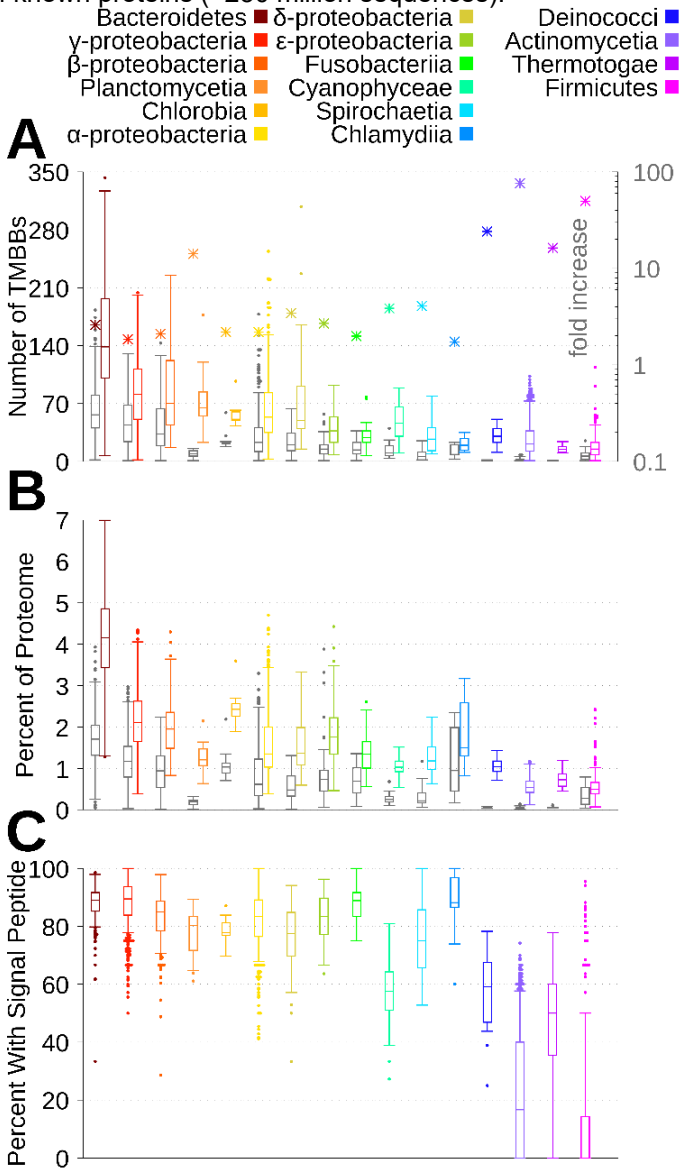

Figure 2. TMBB Content in Bacterial Proteomes and Signal Peptide Prevalence Among Organisms. Colored by class/phylum, categories are sorted by decreasing median number of TMBBs. **(A)** Distribution of number of TMBBs in 2,959 reference and representative bacterial proteomes. TMBBs predicted with IsItABarrel that were within other datasets (gray), and updated numbers after a full search with IsItABarrel (color)**.** Number of TMBB proteins per organism by class/phylum (boxplot), fold increase in number of TMBBs from the initial search (stars using $\log_{10}$ scale at right). **(B)** Distribution of percentage of proteome predicted to encode TMBB proteins (same organisms as in A). **(C)** Percent of TMBBs with signal peptide detected by SignalP 5.0 for each organism, grouped by phyla/class.

Distribution of TMBBs in Representative Organisms

A highly accurate dataset of TMBBs allows for the quantification of the TMBB distribution among different organisms (Fig. 2 A and B). We find that Bacteroidetes tend to have the most TMBBs with an average of 157.4 per organism (average of 4.18% of the proteome). Thermotogae, which is known to have an atypical outer membrane (39), has the fewest with an average of 14.9 per organism (average of 0.75% of the proteome), though we find 15.3-times more barrels in this class than previously identified.

To be inserted into the outer membrane after translation, TMBBs need to first cross the inner membrane. Many, but not all TMBBs use a signal peptide to target the protein to the inner membrane for translocation. We used SignalP 5.0 (40) to determine how many and which of our proteins contained a signal sequence (Fig. 2C). We find that the amount SignalP predicts the presence of a signal sequence varies with organismal class. Moreover, although the correlation is not perfect organismal classes with fewer TMBBs, especially Terrabacteria, tend to also have a lower percentage TMBBs predicted to have a signal peptide. Finally, there is an unusually large variation of percentage of TMBBs predicted with a signal peptide in Firmicutes.

Clustering of Sequences in IsItABarrelDB to Known TMBBs and Other Types

We performed an analysis of evolutionary relationships among the sequences in the IsItABarrelDB. We aimed to identify groupings of proteins related to each other which we will call “types” to disambiguate from other phylogenetic groupings. We identified proteins related to structurally resolved-TMBBs with an iterative sequence similarity search. Types were merged when one or more sequences matched in both TMBB types (described in Supplemental Information, *Analysis of Sequences in IsItABarrelDB*). All proteins that could not be clustered with previously known types were merged. These proteins that were unrelated to known barrels were then clustered and the largest groups containing more than 3,000 proteins were selected for downstream analysis. This clustering analysis was originally conducted with only proteins related to reference proteome proteins with signal peptides. We then identified clusters and relationships for proteins related to reference proteome proteins without signal sequences *post hoc*.

Through this process we found eleven types of barrels, five that clustered with structurally-resolved TMBBs (*Prototypicals, Fim/Usher, LptD-like, LpxR-like* and *Tsx-like)*, and six types that did not cluster with structurally-resolved TMBBs. Three of these non-structurally-resolved types had functional associations in NCBI (*NfrA-like*, *PorT-like, and TorF-like*) and three TMBB types were either associated with a domain of unknown function (DUF) or were annotated as hypothetical proteins (*DUF1302*, *Hypothetical Protein Type 1,* and *Hypothetical Protein Type 2)*. Notably, though we previously hypothesized that efflux pumps convergently evolved to the TMBB fold (3) our clustering algorithm clustered them with the prototypicals. We find no evolutionary relationships between any of the eleven types.

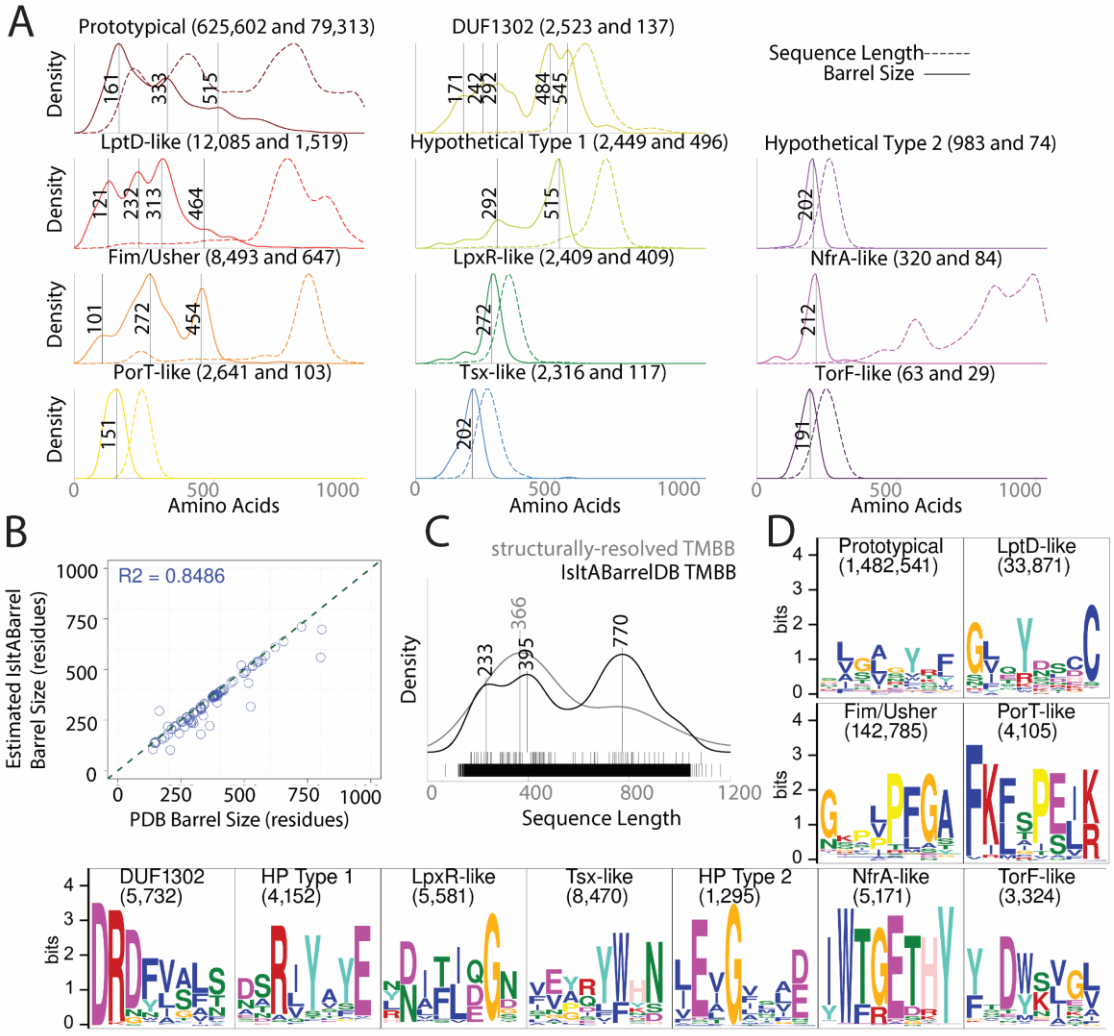

Figure 3. Analysis of IsItABarrelDB by Protein Type. (A) Distribution of sequence lengths (dashed lines) and barrel lengths (solid lines) of the eleven largest types of TMBBs in the IsItABarrelDB. Number of TMBBs used to calculate sequence and barrel lengths in parenthesis. Maxima of the barrel sizes are labeled. **(B)** Correlation between barrel size for structurally resolved barrels and barrel size estimated by IsItABarrel using the closing contact feature. **(C)** Kernel density estimate of the distribution of sequence lengths for structurally-resolved (gray) and IsItABarrel predicted (black) TMBBs, rug plot showing instances below. **(D)** C-terminal sequence motifs found by MEME. Logos plots of motifs found in the C-terminus of TMBBs from each protein type. The letters at each position are ordered from top to bottom by most common to least common. The height of each letter is proportional to its frequency in the column, while the height of the stack is adjusted to be the information content expressed in bits of the aligned sequences at that particular position. The colors of the letters indicate physicochemical properties: orange - small, magenta - acidic, blue - hydrophobic plus cysteine, cyan - tyrosine, pink - histidine, green - polar, red - basic, yellow - proline. The number of TMBBs used for the search are at top in parenthesis.

The length distribution of the protein types (Fig. 3A, dotted lines) reveals differences

among the types. In addition, our IsItABarrel algorithm can identify the number of amino acids per barrel for contact maps with the closing contact. (Fig 3B). We find that for 87 of the 100 structurally resolved proteins that have a closing contact IsItABarrel slightly underpredicts barrel size with an R^2 of 0.85. The barrel size of each protein type (Fig. 3A, solid lines) reveals differences among the types as well as evolutionary patterns. In the Prototypical type there is a three-mode distribution with the latter two approximately two and three times the length of the first. This is possibly indicative of an 8-stranded to 16-stranded duplication events of TMBBs described previously (3, 4), and possibly a less prevalent 16 to 24 stranded accretion of eight strands that has not been previously described. Doubling events are possibly also present in Fim/Usher and Hypothetical Type 1 protein types. Other domain accretion methods appear to be at play in LptD-like and DUF 1302 types. All other types are unimodal.

When comparing the IsItABarrelDB to 100 non-homologous structurally-resolved TMBBs we find that the most frequent length of TMBB sequence is 767 amino acids with another two modes for smaller barrels at 233 and 395 residues (Fig. 3C). The main peak of the length of structurally solved proteins is at 366 residues. Thus, while the smaller sized TMBBs are well represented in the structurally resolved set (peak of distribution at 366 amino acids) the larger barrels are not. In addition, though the structurally-resolved set has one mode, the full IsItABarrelDB has three modes. Notably, these modes are all roughly double each other, suggestive of two possible duplications, though only the 8-stranded barrel to 16-stranded barrel duplication has been seen previously. (3) Since our analysis only includes 120 – 1040 residues we are likely missing a population of larger TMBBs that would be less tractable to evolutionary contact map analysis.

C-terminal signal by protein type

The C-termini of TMBBs are known to have a sequence motif (41) that assists membrane insertion by interacting with the outer membrane translocation protein BAM (42-44) called a β-signal. (45) We analyzed the predicted TMBB sequences for the presence of the β-signal in the eleven types using the MEME (46) sequence-motif detector (Fig. 3D). MEME searched the final quarter of each sequence so that the signal could be detected even if it is not in the terminal strand. (47) We find that all TMBB types have sequence motifs within the last quarter of their sequence (precise location in Fig. S3). The prototypical barrels have the classic signal sequence with a hydrophobic, glycine, hydrophobic, X, tyrosine, positive, phenylalanine motif. However, although all protein types have a sequence motif none of the non-prototypical protein types are similar to the known β-signal nor to our knowledge any other previously described β-signals. (44) Both PorT-like and Hypotheical Protein Type 1 have an aromatic-X-aromatic within their motif. Fim/Usher and TorF-like have a glycine in the -6 position with respect to an aromatic and LptD-like has a glycine in the -4 position with respect to an aromatic. Though these similarities do exist they also underscore how many differences there are among the sequence motifs of different TMBB types. Regardless, most of these sequence motifs do still show typical TMBB hydrophobicity/hydrophilicity alternation typical of β-barrels. (6)

Distribution by phyla

The protein types are distributed differently across phyla. We find TMBBs in 38 phyla and 99 classes (Fig. 4A). Prototypicals are found in all phyla and LptD-like proteins in most phyla. All TMBB types are found in at least some members of the Proteobacteria class and Firmicutes phylum and most are found in the Bacteroidetes phylum. Ordering the phyla by taxonomy reveals that more similar phyla have dissimilar TMBB types possibly indicating that TMBBs are more frequently confered by horizonal gene transfer than by inheritance.

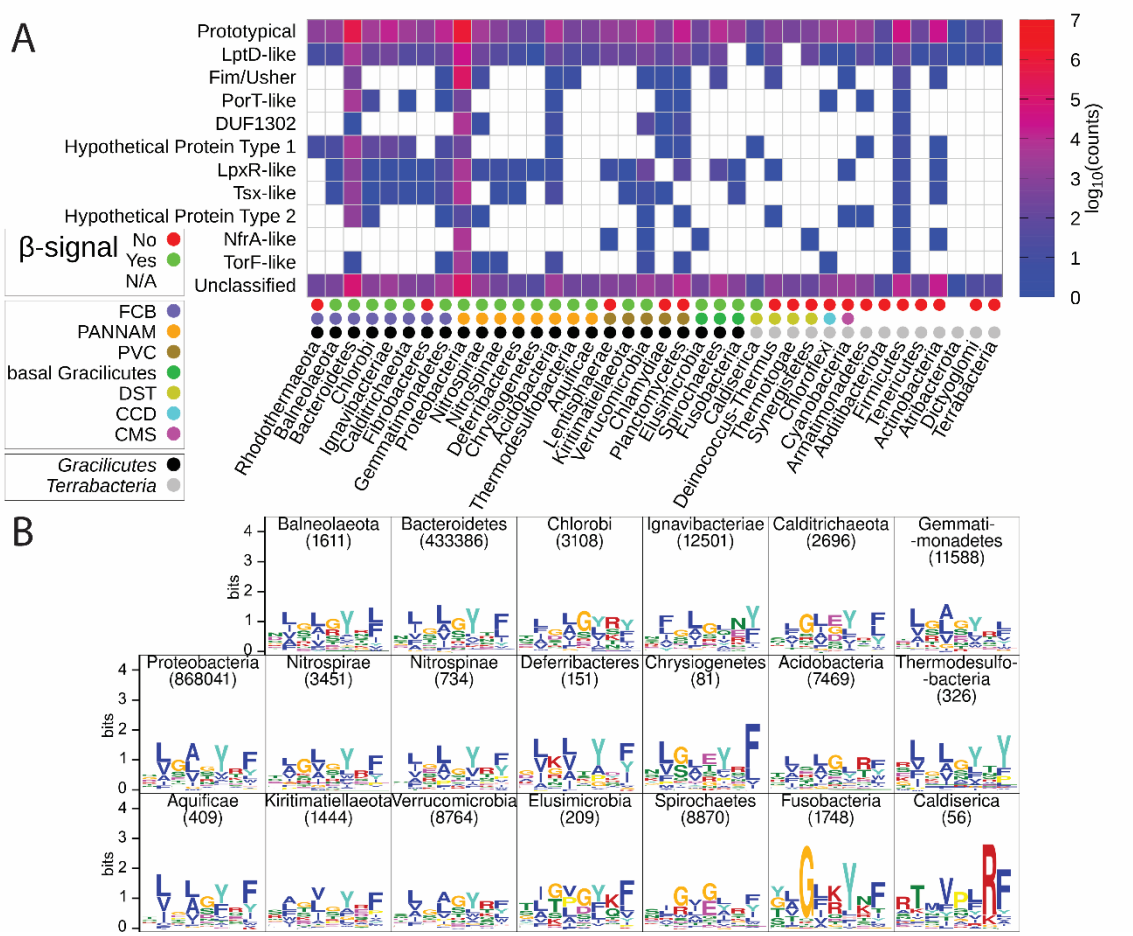

Figure 4. Analysis of IsItABarrelDB by Phylum. (A) The distribution of the eleven protein types and unclassified IsItABarrelDB TMBB sequences among 38 phyla. A red to blue scale indicates greater to fewer numbers of that protein type or no proteins of that type (white). Phyla are organized by Terrabacteria and Gracilicutes (gray and black dots, lowest row), and colored dots that show clades within those categories (middle row), clades are as previously described (48, 49) FCB – Fibrobacteres/Chlorobi/Bacteroidetes, PANNAM - Proteobacteria/Acidobacteria/Nitrospinae/Nitrospirae/Aminicenantes/Methylomirabilis PVC - Planctomycetes/Verrucomicrobia/Chlamydiae, DST - Deinococcus-Thermus/ Synergistetes/Thermotogae, CCD – CPR/Chloroflexi/Dormibacteraeota, CMS – Cyanobacteria/Melainabacteria/Margulisbacteria-Saganbacteria. Dots are also used to indicate the presence of a classic β-signal (top row, red and green). The Unclassified type across all phyla includes 274,245 sequences **(B)** Sequence logos of the prototypical β-signal of each phyla that has a clear traditional β-signal. Logos as described in Fig. 2D. No logos plot was created for Atribacterota because it only contained two proteins.

β-signal by Organism

Searching for sequence motifs by phyla within the prototypical protein type we find that at least 20 of the 38 phyla have a recognizable traditional β-signal (Fig. 4B, estimated location of β-signal Fig. S4). Another 13 of the phyla have at least some features of a traditional β-signal such as a hydrophobic-G-hydrophobic or an aromatic-R-aromatic (Fig. S5, estimated location of β-signal Fig. S6), though this may simply be indicative of typical hydrophobic alternation known for beta strands. (6) We find that although the motif for each phylum has subtle variations, the consensus sequence is LGLGYRF. Each amino acid feature is found to dominate the motif (in N

to C order) in 15, 14, 16, 15, 13, 18, 16 times out of the 20 phyla respectively. We find no similarity between the non-prototypical sequence motifs and the non-traditional prototypical motifs found here.

Discussion

Beyond a useful database, our analysis provides signposts for future discovery of previously unknown protein functions in bacterial outer membranes. Here we find types of proteins that had never been annotated and identify the organisms that use them. Moreover, the extent of bacterial outer membrane diversity and the particulars of the diversity discovered here will be important for developing new and more specific vaccines and antibiotics. Finally, the completeness of the database illustrates patterns of evolution from a bird's-eye-view, highlighting the frequency of duplication events in this protein fold within protein types.

Because of the extensiveness of the database and the lack of relationships found between protein types described herein, it is most likely the case that the eleven types of proteins convergently evolved to a beta barrel fold. The convergence onto this fold by so many proteins provides further evidence that the outer membrane itself must strongly favor this fold. Furthermore, the wide utilization of disparate barrel types in the most distantly related phyla indicates tremendous usage of horizontal gene transfer for membrane barrels.

Algorithm Benefits and Limitations

One of the difficulties of generalizing about the family of outer membrane β-barrels is the small number of solved crystal structures for proteins with low sequence similarity (6, 50, 51) With 1,938,936 sequences, our high-accuracy IsItABarrelDB database contributes a large set of sequences for analyzing TMBBs. The IsItABarrel algorithm can also be used to identify more TMBBs within genomes sequenced in the future.

In addition to IsItABarrelDB being at least two times larger than any previous TMBB database, the IsItABarrel algorithm selects between three and seven times fewer false positives (Table 1). Thus, the database is both larger and more reliable. However, the fact that IsItABarrel is more selective also results in a higher false negative rate. IsItABarrel incorrectly predicts 8% of TMBBs as non-TMBBs. The algorithms that have fewer false negatives have more false positives.

To improve TMBB prediction, IsItABarrel applies knowledge of the β-barrel fold to contact maps generated by homology. RaptorX generates contact maps using a machine learning method that models the probabilities of secondary and tertiary structures by comparison with a large database of homologous natural proteins. The results of this statistical model provide IsItABarrel with the location of the structural features that can be used as heuristics (hairpins and closing contact) as well as with a level of confidence that these structures exist (strength of contact map pixels). By using knowledge of how these features are represented in the contact map—diagonal lines of a certain length, the location relative to the main diagonal, the spacing between consecutive hairpins— IsItABarrel is able to use the statistical distribution as a scaffold for extracting concrete evidence that a map was generated from a TMBB. Using both knowledge of TMBBs and homology information improves upon methods that only use one or the other such as sequence-based methods that rely on HMMs or single sequence properties such as presence of motifs and amino acid composition. The three predictors we compared IsItABarrel with use independent discriminatory features such as hydrophobicity alternation in strand residues, average length of strands, detection of β-signal motif, differences in amino acid composition across the outer membrane and topology modeling with hidden Markov models. We speculate

that the statistical modeling of contacts given by RaptorX contact maps combined with the judicial use of heuristics based on the knowledge of the beta-barrel fold to extract visual features directly from the map is what allows IsItABarrel to improve on prediction accuracy over previous methods.

The two sources of error for our algorithm are the algorithm itself which has a balanced accuracy of 95.88% and the labeling assignments by sequence similarity > 40% which has accuracy of approximately 98% (Fig. 1F).

Though the contact maps from RaptorX are a key element of our algorithm, the main drivers of prediction accuracy are our features that describe strand-strand interactions and closing contacts. Therefore other contact map generation methods could have been employed, such as those from GREMLIN (35) or AlphaFold (36).

To assist with appropriate assessment of the impact of this inaccuracy in downstream tasks, we provide annotation of sequences that were predicted directly using IsItABarrel from those that had their predictions assigned by similarity (available online). Future analysis of the cases that receive different prediction labels when predicted by IsItABarrel and when being assigned by sequence similarity may reveal avenues for further improvement.

TMBB by Phylogenetic Phylum

Using reference genomes, we find that previous reports of TMBB content being 2 – 3% of diderm bacterial genomes (52) are accurate for the most frequently studied of γ-proteobacteria genomes. However, over a larger range of organisms there is higher variance, from 7% to less than 1% with an overall mean of 2.14% and two modes at 1.5% and 2.2%. The category with the largest median genomic percentage of TMBBs is Bacteroidetes at a median of 4.19% per genome. It is possible that the greater number of TMBBs is due to Bacteroidetes localization in the gut where they need to utilize a variety of nutrients. For example, Bacteroidetes use several polysaccharide utilization loci encoding outer membrane proteins that allow these organisms to adapt quickly to differences in the set of nutrient available. (53)

One notable feature is that some of the classes with the fewest numbers of TMBBs have the widest variety of TMBBs. Firmicutes, the second largest phylum found in the human microbiome after Bacteroidetes, has only a small number of organisms with TMBBs. This is likely due to the organisms in this phylum being low-GC Gram-positives with only some organisms being diderms with outer membranes. However, though Firmicutes tend to have few TMBBs as a percentage of their genome, they have the greatest variety of TMBBs along with Proteobacteria with both containing all eleven types of TMBB. We similarly find that several organisms in the Actinomycetia class have only a small number of TMBBs, yet there is a large variety of TMBB types, with eight of the ten types of TMBBs in the phylum Actinobacteria. The existence of these phyla with few TMBBs in number, but wide in TMBB diversity may be linked to their lineage. Firmicutes and Actinobacteria are closely related and are the only phyla to display mixed monoderm diderm lineages. As such these have been hypothesized to be closer to the root leading to the last bacterial common ancestor than other diderm bacteria. (54) Perhaps this ancestor had all possible TMBB types and some were lost from disuse over evolution.

Though we would have expected that the more related bacteria would have more similar varieties of TMBBs, we do not find such a correlation. Rather, when ordering bacteria by their phylogenetic relationships (48, 49), we find that different TMBB types are scattered all over the phylogenetic tree. This may indicate that horizontal gene transfer is more common for TMBBs than previously anticipated. The magnitude of the variety of bacteria using different types of barrels may suggest that some barrel architectures are appropriated and repurposed for alternate needs by extremely different bacteria from the ones in which the barrels originated.

Domains of Unknown Function

We also analyzed the presence and distribution of DUFs in predicted TMBB sequences. In our reference genome analysis, we found 6,972 proteins that included domains labeled as DUF, which represents 4.31% of the set of new TMBB proteins found in these reference organisms. A total of 99 different folds categorized as DUFs are found among those 6,972 proteins (Table S4). These folds are distributed differently among the 16 largest organismal classes of reference genomes (Table S5). For the whole IsItABarrelDB analysis we found 74,124 predicted TMBBs with DUFs (this represents 3.82% of our final dataset that contains a total of 1,938,936 sequences). The distribution of sequences for each of the 260 DUFs follows a power law (Table S6), where only a few groups are large while most other groups have a much smaller number of sequences. In particular, we note that the two largest groups contain sequences that are mostly in the Prototypical type. The third group, DUF1302, formed a type of its own in our analysis (Fig. 3). In the reference genome analysis, we see that some of the domains we find have similarity to other barrels (DUF481, DUF3187) while others are less similar to whole barrels and may be a globular domain attached to a barrel protein and exposed on the cellular surface such as domains for cell adhesion (DUF1566).

TMBB Sequence Length

The distribution of barrel lengths in the IsItABarrelDB is different across the eleven protein types we identified (Fig. 3A). The largest type, prototypical TMBBs, has a trimodal distribution where the peaks roughly indicate the possibility of doubling events from which outer membrane β-barrels may have evolved and increased in size by domain duplication (8, 16 and 24-stranded barrels). (3, 4) Similar ~160 residue jumps are also apparent in the barrel sizes of the Fim/Usher proteins and an ~320 residue jump (2 x 160) in the Hypothetical Type 1 proteins. This supports the notion of the 8-stranded barrel being a fundamental evolutionary starting point and building block. (3) Evidence of smaller steps of barrel growth are present in LptD-like and DUF 1302 types. This may indicate smaller domain accretion, possibly the duplication of double-hairpins. (2)

β-signal Motif

We find that SignalP predicts that usage of a signal peptide varies by organism, however it is unclear if this is truly due to the lack of signal peptide or if SignalP is simply not as well trained on the signal peptides used by those organisms. Regardless, we find no such variation by protein type (Fig. S7) which consistently has a high percentage of signal peptide found.

With respect to C-terminal sequence motifs, we find that although sequence motifs can be found in all TMBB types, only the prototypical type has a previously documented β-signal motif. The Fim/Usher motif we found did not conform to that found in previous reports. (44, 55) Overall, there is no consensus among the β-signals when they are broken out into protein type. Since the traditional β-signal is in the last strand of the barrel, we evaluated the locational distribution of these other motifs by estimating the location of the final residue of the barrel within the sequence (Fig. S3). We find that the motifs of prototypical, LptD-like, DUF1302 and NfrA-like are closest to the end of the barrel while the other motifs are further from the final strand

Our computational approach to detect consensus motifs may be influenced by sample size and sequence redundancy, which may have impacted detection of the true motif in some of the phyla with more variation. More studies are needed to understand the role of these motifs. However, because we see hydrophobicity alternation in all motifs, these motifs are likely localized within the barrel inserted in the membrane.

We take different lessons from the subtle variation in β-signal among the bacterial phyla and the extreme differences in sequence motifs among the different types of TMBBs.

The subtle variations among most phyla within the known β-signal motif likely indicate conserved use of subtly different Barrel Assembly Machineries (BAMs) among all these bacterial classes with a conserved mechanism of prototypical membrane insertion.

The subtle variations among most phyla within the known β-signal motif likely indicate conserved use of subtly different barrel assembly machinery (BAM) among all these bacterial classes with a conserved mechanism of prototypical membrane insertion.

However, though we see C-terminal sequence motifs in all varieties of TMBBs, there is extreme variation among these motifs. The frequent use of a C-terminal sequence may indicate that this region is consistently used for membrane insertion though the β-signal also has a role in activating proteolysis machinery when unfolded barrels accumulate in the periplasm (56), and these differences in motifs may also be related to this additional function. If some of these alternate sequence motifs are used for membrane insertion, the lack of similarities between most types of C-terminal signal sequences and the traditional β-signal give rise to two possible alternate mechanisms of TMBB membrane insertion: 1) some alternate sequence motifs are used to target BAM though with a different membrane insertion mechanism and 2) some alternate sequence motifs target a different membrane insertion machinery—possibly even TAM which has been shown to play a role in Fim/Usher membrane insertion (57). These mechanisms are not mutually exclusive and future studies may show that one, the other, or both are found in nature.

Future studies may also allow for IsItABarrel to be extended to identify other membrane β-barrel proteins such as those in the mitochondria or chloroplasts. Because the parameters for IsItABarrel were optimized for bacterial barrels, we find that IsItABarrel does not predict other membrane β-barrels with high accuracy. Difficulties arising from differences in membrane depth and known differences in barrel structures—odd-stranded mitochondrial barrels, unclosed chloroplastic barrels—would need to be overcome for such an extension to be successful.

Conclusion

We have developed an algorithm for TMBB prediction and have created a new database of curated TMBB sequences that is 40% more accurate and at least two times larger than those created with previous algorithms. TMBB content of bacterial genomes varies with some having 7% of the genome devoted to encoding TMBBs and some with less than 1%. For the most part we find that the β-signal varies in small ways by organism, and that the β-signal consensus is LGLGYRF. A study of eleven types of TMBB sequences indicates that there are many types of TMBBs without this consensus signal. We offer our TMBB identifying algorithm and database as a useful tool for researchers working with bacterial TMBBs who need a curated set of these proteins at https://isitabarrel.ku.edu andhttps://github.com/SluskyLab/isitabarrel. Awareness of the types and quantities of outer membrane proteins in different bacterial classes will be a foundation for better understanding of bacterial phylogeny/evolution and clearer understanding of outer membrane protein function.

Materials and Methods

IsItABarrel Algorithm

To classify sequences as TMBBs or non-TMBBs we combined RaptorX-Contact co-evolutionary contact map generation (58) with the development of a rule-based algorithm for feature extraction and classification. Map creation is more fully described in Supplemental Information, *Contact Map Generation*).

Our IsItABarrel algorithm was based on the following principles: Because the β-strands in TMBBs are ordered contiguously (strand 1 is next to strand 2, strand 2 is next to strand 3, etc.) TMBBs have consistent and distinctive contact map features. These features include map lines that indicate strand-strand interactions, and a line in the corner indicates the closing contacts between the first strand and the last strand. Non-TMBB folds rarely have both these features. The strand-strand contact feature is extracted by scanning the contact map along the main diagonal and counting how many strand-strand contact pairs are above a length threshold (further described in Supplemental Information, *Heuristic Features and IsItABarrel Algorithm*). The closing contacts feature is extracted by scanning the corner regions of the map and ensuring the detected closing contacts align with previously found first and last strand. Details of the search for closing contacts is provided in Supplemental Information, *Heuristic Features and IsItABarrel Algorithm*. The two features are combined to classify the map as a TMBB or not-TMBB (Supplemental Information, *Heuristic Features and the IsItABarrel Algorithm*). We observed that contact maps of sequences sharing more than 40% similarity receive the same IsItABarrel prediction 98% of the time. This made it possible to use IsItABarrel to predict large datasets in three steps: (1) clustering the set, (2) making a prediction for the smaller number of cluster representatives and (3) broadcasting this prediction to all other members of the cluster (further described in Supplemental information).

Datasets

To train and validate the IsItABarrel algorithm we created two datasets of proteins with solved structures in the PDB. We call these datasets TMBB121 (Dataset S1) and NONTMBB1K (Dataset S2). The TMBB121 dataset includes TMBB examples (TMBBs), while the sets NONTMBB1K include non-TMBB examples. We use only one small set of TMBB examples because there is a limited number of non-redundant TMBBs with solved structures. The list of TMBB proteins were taken from a previously published dataset of outer membrane proteins (4) and the non-TMBB examples were randomly sampled from PISCES (59) and PDBSelect. (60) The negative examples were manually verified to ensure they did not include TMBBs. Since the dataset NONTMBB1K is much larger than the set of TMBB examples, we worked with unbalanced data throughout our training, validation and testing. (40) A summary of the datasets and how they were used is given in Table S3. The parameters of the IsItABarrel algorithm were optimized by grid search using balanced accuracy as a scoring metric on our validation set and details are provided in section *Lowering the False Negative Rate of IsItABarrel* in Supplemental Information and in Fig. S1).

To perform large-scale comparisons between different TMBB predictors (Fig. S2) we created two non-redundant datasets of protein sequences. The TMBB29183 (Dataset S4) contains the top 29,183 putative TMBB sequences from the TMBB-DB. (31) The OMPDB-NR (Dataset S5) contains 651,874 non-redundant representative sequences extracted from the OMPdb. Finally, using our IsItABarrel algorithm, we predicted TMBBs in the large non-redundant set of protein sequences from NCBI restricted to only bacterial sequences (taxonomic ID 2). This search resulted in the creation of our final IsItABarrelDB dataset with 1,938,936 TMBB sequences which is available for search and download at https://isitabarrel.ku.edu. Further details of how the database was created and features of our webapp are provided in section *Creation of the IsItABarrelDB* of the Supplemental Information.

To understand how TMBBs are distributed across taxonomic lineages, the proteomes of 2,959 prokaryotic organisms were analyzed with IsItABarrel. We created contact maps for all sequences of the largest proteome for each of the 16 organismal categories and these maps were predicted with IsItABarrel. All sequences from the 16 reference proteomes were used in this analysis. The sequences predicted to be TMBBs were counted by class/phylum (16 categories), producing updated estimates of TMBB content in prokaryotic organisms. We also used these

sequences to observe differences in the distribution of lengths and C-terminal β-signal motif across the 16 taxonomic categories (Supplemental Information, *Distribution of TMBBs in Representative Organisms*).

To detect new types of TMBBs we performed a thorough sequence search of our final IsItABarrelDB dataset. Sequences of TMBBs with solved structures from all known TMBB types were used as seed queries against the IsItABarrelDB. Sequences that were not related to any of the known classes were clustered at 25%, and the largest clusters (designated 'types') were selected for analysis.

To understand how sequence length and β-signal motif are distributed across these new types of TMBBs we computed kernel density estimates (KDE) for the distribution of lengths of sequences and used MEME to compute motifs on the C-terminal quarter of the sequence. Details of this analysis are given in Supplemental Information, A*nalysis of Sequences in IsItABarrelDB*.

The IsItABarrel algorithm was originally designed only for classification of protein sequences as TMBBs or not TMBBs. However, we also used the closing contact to identify the first and last amino acids and thereby estimate barrel size. Details of the availability of our data and code are given in the Supplemental Information, *Code availability* and *Programming*.

Acknowledgments

We thank Rachel Kolodny for important discussions about the algorithm, Harris Bernstein for helpful feedback on the manuscript, and Robert Unckless from the University of Kansas for help with our analysis of distribution of transmembrane β-barrels in prokaryotes. Financial support for this work came from National Institute of Health award DP2 GM128201 and P20GM103418, National Science Foundation (NSF) award 2226804, an American Scandinavian Foundation fellowship, and Kansas University (KU) Startup funding to Joanna Slusky. Daniel Montezano acknowledges support from the National Institute of Health award P20GM103418.

References

1. G. E. Schulz, The structure of bacterial outer membrane proteins. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1565**, 308-317 (2002).
2. M. Remmert, A. Biegert, D. Linke, A. N. Lupas, J. Soding, Evolution of outer membrane beta-barrels from an ancestral beta beta hairpin. *Molecular biology and evolution* **27**, 1348-1358 (2010).
3. M. W. Franklin *et al.*, Evolutionary pathways of repeat protein topology in bacterial outer membrane proteins. *Elife* **7** (2018).
4. M. W. Franklin *et al.*, Efflux Pumps Represent Possible Evolutionary Convergence onto the beta-Barrel Fold. *Structure* **26**, 1266-1274 e1262 (2018).
5. S. E. Rollauer, M. A. Sooreshjani, N. Noinaj, S. K. Buchanan, Outer membrane protein biogenesis in Gram-negative bacteria. *Philos Trans R Soc Lond B Biol Sci* **370** (2015).
6. R. Dhar, R. Feehan, J. S. G. Slusky, Membrane Barrels Are Taller, Fatter, Inside-Out Soluble Barrels. *J Phys Chem B* **125**, 3622-3628 (2021).
7. W. C. Wimley, Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. *Protein science : a publication of the Protein Society* **11**, 301-312 (2002).
8. D. Hsieh, A. Davis, V. Nanda, A knowledge-based potential highlights unique features of membrane alpha-helical and beta-barrel protein insertion and folding. *Protein science : a publication of the Protein Society* **21**, 50-62 (2012).
9. R. Jackups, Jr., J. Liang, Interstrand pairing patterns in beta-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction. *J Mol Biol* **354**, 979-993 (2005).
10. M. M. Gromiha, M. Suwa, Discrimination of outer membrane proteins using machine learning algorithms. *Proteins* **63**, 1031-1037 (2006).
11. Q. Liu, Y. Zhu, B. Wang, Y. Li, Identification of beta-barrel membrane proteins based on amino acid composition properties and predicted secondary structure. *Computational biology and chemistry* **27**, 355-361 (2003).
12. Y. Zhai, M. H. Saier, Jr., The beta-barrel finder (BBF) program, allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes. *Protein science : a publication of the Protein Society* **11**, 2196-2207 (2002).
13. T. C. Freeman, Jr., W. C. Wimley, A highly accurate statistical approach for the prediction of transmembrane beta-barrels. *Bioinformatics* **26**, 1965-1974 (2010).
14. S. Hayat, C. Peters, N. Shu, K. D. Tsirigos, A. Elofsson, Inclusion of dyad-repeat pattern improves topology prediction of transmembrane beta-barrel proteins. *Bioinformatics* **32**, 1571-1573 (2016).
15. F. S. Berven, K. Flikka, H. B. Jensen, I. Eidhammer, BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. *Nucleic acids research* **32**, W394-399 (2004).
16. M. M. Gromiha, Motifs in outer membrane protein sequences: applications for discrimination. *Biophysical chemistry* **117**, 65-71 (2005).
17. S. Hayat, A. Elofsson, Ranking models of transmembrane beta-barrel proteins using Z-coordinate predictions. *Bioinformatics* **28**, i90-96 (2012).
18. J. Waldispuhl, B. Berger, P. Clote, J. M. Steyaert, transFold: a web server for predicting the structure and residue contacts of transmembrane beta-barrels. *Nucleic acids research* **34**, W189-193 (2006).
19. J. Waldispuhl, C. W. O'Donnell, S. Devadas, P. Clote, B. Berger, Modeling ensembles of transmembrane beta-barrel proteins. *Proteins* **71**, 1097-1112 (2008).
20. K. D. Tsirigos, A. Elofsson, P. G. Bagos, PRED-TMBB2: improved topology prediction and detection of beta-barrel outer membrane proteins. *Bioinformatics* **32**, i665-i671 (2016).

21. W. Tian, M. Lin, K. Tang, J. Liang, H. Naveed, High-resolution structure prediction of beta-barrel membrane proteins. *Proc Natl Acad Sci U S A* **115**, 1511-1516 (2018).
22. H. Bigelow, B. Rost, PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins. *Nucleic acids research* **34**, W186-188 (2006).
23. A. G. Garrow, A. Agnew, D. R. Westhead, TMB-Hunt: an amino acid composition based method to screen proteomes for beta-barrel transmembrane proteins. *BMC bioinformatics* **6**, 56 (2005).
24. M. Remmert, D. Linke, A. N. Lupas, J. Soding, HHomp--prediction and classification of outer membrane proteins. *Nucleic acids research* **37**, W446-451 (2009).
25. N. K. Natt, H. Kaur, G. P. Raghava, Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. *Proteins* **56**, 11-18 (2004).
26. A. Randall, J. Cheng, M. Sweredoski, P. Baldi, TMBpro: secondary structure, beta-contact and tertiary structure prediction of transmembrane beta-barrel proteins. *Bioinformatics* **24**, 513-520 (2008).
27. P. L. Martelli, P. Fariselli, A. Krogh, R. Casadio, A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics* **18 Suppl 1**, S46-53 (2002).
28. P. G. Bagos, T. D. Liakopoulos, I. C. Spyropoulos, S. J. Hamodrakas, A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins. *BMC bioinformatics* **5**, 29 (2004).
29. C. Savojardo, P. Fariselli, R. Casadio, Improving the detection of transmembrane beta-barrel chains with N-to-1 extreme learning machines. *Bioinformatics* **27**, 3123-3128 (2011).
30. H. B. Kazemian, C. M. Grimaldi, Cascading classifier application for topology prediction of transmembrane beta-barrel proteins. *Journal of bioinformatics and computational biology* **18**, 2050034 (2020).
31. T. C. Freeman, Jr., W. C. Wimley, TMBB-DB: a transmembrane beta-barrel proteome database. *Bioinformatics* **28**, 2425-2430 (2012).
32. K. D. Tsirigos, P. G. Bagos, S. J. Hamodrakas, OMPdb: a database of {beta}-barrel outer membrane proteins from Gram-negative bacteria. *Nucleic acids research* **39**, D324-331 (2011).
33. D. S. Marks *et al.*, Protein 3D structure computed from evolutionary sequence variation. *PloS one* **6**, e28766 (2011).
34. S. Seemayer, M. Gruber, J. Soding, CCMpred--fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* **30**, 3128-3130 (2014).
35. H. Kamisetty, S. Ovchinnikov, D. Baker, Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci U S A* **110**, 15674-15679 (2013).
36. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
37. I. B. Hernandez *et al.*, Clustering predicted structures at the scale of the known protein universe. *bioRxiv* 10.1101/2023.03.09.531927, 2023.2003.2009.531927 (2023).
38. B. W. Matthews, Comparison of Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim Biophys Acta* **405**, 442-451 (1975).
39. D. Megrian, N. Taib, J. Witwinowski, C. Beloin, S. Gribaldo, One or two membranes? Diderm Firmicutes challenge the Gram-positive/Gram-negative divide. *Molecular microbiology* **113**, 659-671 (2020).
40. J. J. Almagro Armenteros *et al.*, SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature biotechnology* **37**, 420-423 (2019).
41. M. Struyvé, M. Moons, J. Tommassen, Carboxy-terminal phenylalanine is essential for the correct assembly of a bacterial outer membrane protein. *Journal of Molecular Biology* **218**, 141-148 (1991).
42. M. T. Doyle, H. D. Bernstein, Bacterial outer membrane proteins assemble via asymmetric interactions with the BamA beta-barrel. *Nat Commun* **10**, 3358 (2019).

43. D. Gessmann *et al.*, Outer membrane beta-barrel protein folding is physically controlled by periplasmic lipid head groups and BamA. *Proc Natl Acad Sci U S A* **111**, 5878-5883 (2014).
44. X. Wang, J. H. Peterson, H. D. Bernstein, Bacterial Outer Membrane Proteins Are Targeted to the Bam Complex by Two Parallel Mechanisms. *mBio* **12** (2021).
45. V. Robert *et al.*, Assembly factor Omp85 recognizes its outer membrane protein substrates by a species-specific C-terminal motif. *PLoS Biol* **4**, e377 (2006).
46. T. L. Bailey, J. Johnson, C. E. Grant, W. S. Noble, The MEME Suite. *Nucleic acids research* **43**, W39-49 (2015).
47. C. L. Hagan, J. S. Wzorek, D. Kahne, Inhibition of the beta-barrel assembly machine by a peptide that binds BamD. *Proc Natl Acad Sci U S A* **112**, 2011-2016 (2015).
48. D. Megrian, N. Taib, A. L. Jaffe, J. F. Banfield, S. Gribaldo, Ancient origin and constrained evolution of the division and cell wall gene cluster in Bacteria. *Nat Microbiol* **7**, 2114-2127 (2022).
49. J. Witwinowski *et al.*, An ancient divide in outer membrane tethering systems in bacteria suggests a mechanism for the diderm-to-monoderm transition. *Nat Microbiol* **7**, 411-422 (2022).
50. J. S. Slusky, R. L. Dunbrack, Jr., Charge asymmetry in the proteins of the outer membrane. *Bioinformatics* **29**, 2122-2128 (2013).
51. H. Naveed, Y. Xu, R. Jackups, Jr., J. Liang, Predicting three-dimensional structures of transmembrane domains of beta-barrel membrane proteins. *J Am Chem Soc* **134**, 1775-1781 (2012).
52. W. C. Wimley, The versatile β-barrel membrane protein. *Current Opinion in Structural Biology* **13**, 404-411 (2003).
53. S. M. Lee *et al.*, Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature* **501**, 426-429 (2013).
54. N. Taib *et al.*, Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nat Ecol Evol* **4**, 1661-1672 (2020).
55. N. Paramasivam, M. Habeck, D. Linke, Is the C-terminal insertional signal in Gram-negative bacterial outer membrane proteins species-specific or not? *BMC Genomics* **13**, 510 (2012).
56. N. P. Walsh, B. M. Alba, B. Bose, C. A. Gross, R. T. Sauer, OMP peptide signals initiate the envelope-stress response by activating DegS protease via relief of inhibition mediated by its PDZ domain. *Cell* **113**, 61-71 (2003).
57. C. Stubenrauch *et al.*, Effective assembly of fimbriae in Escherichia coli depends on the translocation assembly module nanomachine. *Nat Microbiol* **1**, 16064 (2016).
58. J. Ma, S. Wang, Z. Wang, J. Xu, Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics* **31**, 3506-3513 (2015).
59. G. Wang, R. L. Dunbrack, Jr., PISCES: recent improvements to a PDB sequence culling server. *Nucleic acids research* **33**, W94-98 (2005).
60. S. Griep, U. Hobohm, PDBselect 1992-2009 and PDBfilter-select. *Nucleic acids research* **38**, D318-319 (2010).