Gram-negative outer-membrane proteins with multiple β-barrel domains

Ron Solan[a,1], Joana Pereira[b,1,2], Andrei N. Lupas[b,3], Rachel Kolodny[c,3], and Nir Ben-Tal[a,3]

[a]Department of Biochemistry and Molecular Biology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel; [b]Department of Protein Evolution, Max Planck Institute for Developmental Biology, Tübingen 72076, Germany; and [c]Department of Computer Science, University of Haifa, Haifa 3498838, Israel

Edited by Barry Honig, Columbia University, New York, NY, and approved June 28, 2021 (received for review March 1, 2021)

Outer-membrane beta barrels (OMBBs) are found in the outer membrane of gram-negative bacteria and eukaryotic organelles. OMBBs fold as antiparallel β-sheets that close onto themselves, forming pores that traverse the membrane. Currently known structures include only one barrel, of 8 to 36 strands, per chain. The lack of multi-OMBB chains is surprising, as most OMBBs form oligomers, and some function only in this state. Using a combination of sensitive sequence comparison methods and coevolutionary analysis tools, we identify many proteins combining multiple beta barrels within a single chain; combinations that include eight-stranded barrels prevail. These multibarrels seem to be the result of independent, lineage-specific fusion and amplification events. The absence of multibarrels that are universally conserved in bacteria with an outer membrane, coupled with their frequent de novo genesis, suggests that their functions are not essential but rather beneficial in specific environments. Adjacent barrels of complementary function within the same chain may allow for functions beyond those of the individual barrels.

gram-negative bacteria | outer-membrane beta barrels | sequence analysis | coevolution analysis of bacterial sequences | evolutionary analysis

Outer-membrane beta barrels (OMBBs) are an important class of membrane proteins in gram-negative bacteria, mitochondria, and chloroplasts (1, 2). In bacteria, OMBBs are the most common family of outer-membrane proteins, and their functions are very diverse (e.g., adhesion, pilus formation, specific and nonspecific forms of import and efflux, proteolysis, and even outer-membrane protein assembly). Structurally, OMBBs are closed β-sheets of antiparallel β-strands, forming pores that traverse the membrane. In the majority of the cases, the sheet is formed from a single protein chain, but in some cases, it can result from the assembly of smaller β-sheets, as in trimeric autotransporters (3). With the exception of the mitochondrial 19-stranded porins (4), OMBBs have an even number of strands, and OMBBs with between eight and 36 strands have been described so far (5, 6). For their biological activity, many OMBBs form complexes with each other, as for example, the outer-membrane phospholipase A (OMPLA) homodimer (7, 8), the trimeric porins (9, 10), and the type-9 translocon heterodimer (6).

Previous studies highlighted the common evolutionary origins of OMBB proteins (5, 11, 12). Remmert et al. (11) argued that all OMBBs in gram-negative bacteria evolved from a single ancestral subunit of two β-strands, arranged as a hairpin. Their study and a later work by Franklin et al. (5) showed that the diverse structures of OMBBs evolved through amplification and recombination of these hairpins and the accretion of mutations. These analyses point to the important contributions of duplication and fusion events to the evolution of OMBBs. In an experimental study, Arnold et al. (13) showed that the fusion of two β-barrels connected by a short linker can yield a single barrel of twice the size, demonstrating that concatenating hairpins may result in single, larger barrels. Given the importance of amplification for the genesis of new β-barrels, it is notable that despite the propensity of OMBBs to form oligomers and the fact that some OMBBs function solely in complex with other OMBBs, naturally occurring proteins with multiple barrel domains (referred to herein as "multibarrel" proteins) have not yet been identified [though the possibility of their existence has been acknowledged, for example, by Reddy and Saier (12)].

We sought to expand the current repertoire of known OMBB architectures and provide insights to their evolution by searching for proteins with multiple OMBB domains. To identify yet unknown protein architectures, one must search beyond the Protein Data Bank (14) in the structurally uncharacterized space curated in sequence databases (15). Using a combination of state-of-the-art sensitive sequence comparison methods and coevolutionary analysis tools, we searched the UniRef100 and the National Center for Biotechnology Information (NCBI) nonredundant (nr_bac) databases for sequences with more than one nonoverlapping match to any OMBB family of known structure. After classification and annotation of the identified sequences, we were able to predict many OMBB families with previously unknown strand topologies but most importantly with a wide variety of multibarrel domain combinations. Grouping these multibarrel chains into families and superfamilies, we characterize them phylogenetically. Based on the function of their closest single-barrel homologs, we discuss putative biological roles for some of the multibarrel families. Our findings highlight that OMBBs have a richer repertoire of architectures than previously known, expanding our current knowledge of these proteins, providing hints about their

Significance

All currently known architectures of outer-membrane beta barrels (OMBBs) have only one barrel. While the vast majority function as oligomers, with barrels from different chains packing against each other in the membrane, it was assumed that these multiple chains are needed to form multibarrel structures. And yet, here we show that multibarrel chains exist. Using state-of-the-art sequence and structure analysis tools, we report the discovery of more than 30 multibarrel architectures from gram-negative bacteria. The discovery of these architectures reveals another interesting chapter in OMBB evolution and has implications for protein engineering. The evolutionary advantages of multibarrels are yet to be discovered.

Author contributions: R.S., A.N.L., R.K., and N.B.-T. designed research; R.S., J.P., and R.K. performed research; R.S. contributed new reagents/analytic tools; R.S., J.P., A.N.L., R.K., and N.B.-T. analyzed data; and R.S., J.P., A.N.L., R.K., and N.B.-T. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]R.S. and J.P. contributed equally to this work.

[2]Present address: Biozentrum, University of Basel, 4056 Basel, Switzerland.

[3]To whom correspondence may be addressed. Email: bental@tauex.tau.ac.il, trachel@cs.haifa.ac.il, or andrei.lupas@tuebingen.mpg.de.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2104059118/-/DCSupplemental.

Published July 30, 2021.

evolution, and revealing 34 previously unknown architectures of gram-negative surface proteins.

Results and Discussion

To identify putative multibarrel proteins and single barrels with yet unknown architectures, we followed two parallel and complementary approaches. In one approach, we used HMMER (16) to align Hidden Markov models (HMMs) of all single-barrel proteins of known structure to the sequences in UniRef100 (15) and searched for cases with multiple matches along their sequence. In a second approach, we used PsiBLAST (17) to search the bacterial protein sequences in the nonredundant sequence database at NCBI (nr_bac), starting from representative sequences of all OMBB families of known structure. Both approaches resulted in a similar set of proteins, which were combined, classified, and annotated using state-of-the-art sequence analysis tools.

We grouped the protein sequences that we found based on their global sequence similarity and denote these multibarrel families (MB-families) and single-barrel families. In some cases, we further grouped MB-families that share some sequence similarity, but not necessarily over their entire chains, and denoted these multibarrel superfamilies (MB-superfamilies). An alternative grouping of the MB-families is based on their predicted structure, denoted as multibarrel architecture (MB-architecture). The MB-architecture annotates the sizes of the barrel domains, ordered from the N-terminal domain to the C-terminal domain. For example, the MB-architecture of a protein composed of a 16-stranded N-terminal barrel and a 12-stranded C-terminal barrel is 16–12. The search and annotation procedures are described in detail in *Methods*. The seeds and MB-architectures found are available online at https://trachel-srv.cs.haifa.ac.il/rachel/MOMBB/.

34 Different MB-Architectures. We identified 12,643 unique proteins with previously unknown architecture and provide an overview of this large set using CLANS's (CLuster ANalysis of Sequences) coarse clustering based on local sequence similarity (*SI Appendix*, Fig. S1). A finer clustering based on global similarity and manual inspection allowed separating these coarse clusters into 186 MB-families, each with a single MB-architecture. Of these MB-families, available at https://trachel-srv.cs.haifa.ac.il/rachel/MOMBB/, 79 have four or more proteins (*SI Appendix*, Table S1). Size distribution shows many small MB-families and a few large MB-families of ∼1,000 proteins or more (*SI Appendix*, Fig. S2).

Annotation of representative proteins from each cluster suggests that 34 MB-architectures are represented in our set: 17 double-barrels and 17 of three or more barrels (Fig. 1 and *SI Appendix*, Fig. S3). These mostly comprise proteins with multiple nonoverlapping full-length matches to known OMBBs, indicating concatenation of known OMBBs. The linkers connecting these matches in representative sequences from each MB-family provide a clear delimitation between the putative barrels, with a median length of 22 ± 13 residues (*SI Appendix*, Fig. S4). These findings suggest that a large variety of multibarrel domain architectures exist in nature, far beyond the previously documented repertoire of OMBBs.

Of the 34 putative MB-architectures, 27 include an eight-stranded barrel (i.e., are either two [or more] concatenated eight-stranded barrels or an eight-stranded barrel concatenated with a barrel of different size) (Fig. 1 and *SI Appendix*, Fig. S3). All but one of the architectures with at least three barrels are composed of multiple eight-stranded OMBBs (Fig. 1 and *SI Appendix*, Fig. S3) or a combination of these and exactly one barrel of another size (of 10, 12, 16, or 22 strands). Some architectures are two barrels with the same number of strands: these are either a repeat, as in the PLA1–PLA1 (12, 12) architecture, or two barrels from different OMBB families, as in the FhaC–Porin (16, 16) architecture. Not all known single-barrel topologies appear in multibarrels: we did not find any MB-architecture with a 24-stranded barrel or with a 36-stranded barrel.

Grouping by the taxonomy of their respective bacteria shows that the MB-architectures are present mostly in Proteobacteria and Bacteroidetes, although a few are in Cyanobacteria, Firmicutes, Fusobacteria, and other phyla (Fig. 1).

We inspected the distribution of species within each MB-family. Some MB-families include only a few homologous proteins in a single species. All MB-families save one are endemic to a single phylum (with up to only 2% proteins in other phyla). Fig. 2 shows an example of two different distributions of two clusters: the 18–8-8 MB-family (marked with yellow circles) has 48 proteins but is spread among more Proteobacterial clades than the 22–8 MB-family (marked with red circles), which has 99 proteins.

Additional Single-Barrel Families. Other instances beyond those that clearly include multibarrels show only partial and/or overlapping matches to known OMBBs, representing putative single-barreled families. Within these, we identified candidates for families with architectures of 10, 12, 16, and 34 strands. More interestingly, we also identified a superfamily of large OMBBs,

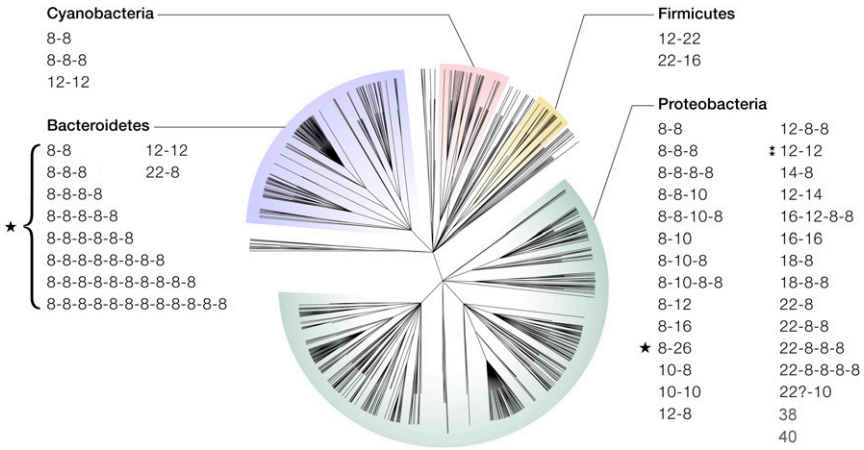

Fig. 1. Taxonomic classification of the multi-barrel and large barrel architectures. All the species with MB-architectures were collected, and their taxonomic tree of 483 families was composed from the NCBI taxonomy database. The architectures that are manifested in each of the major clades are listed, with asterisks marking the ones that are discussed in the main text. The tree was rendered using Dendroscope (45).

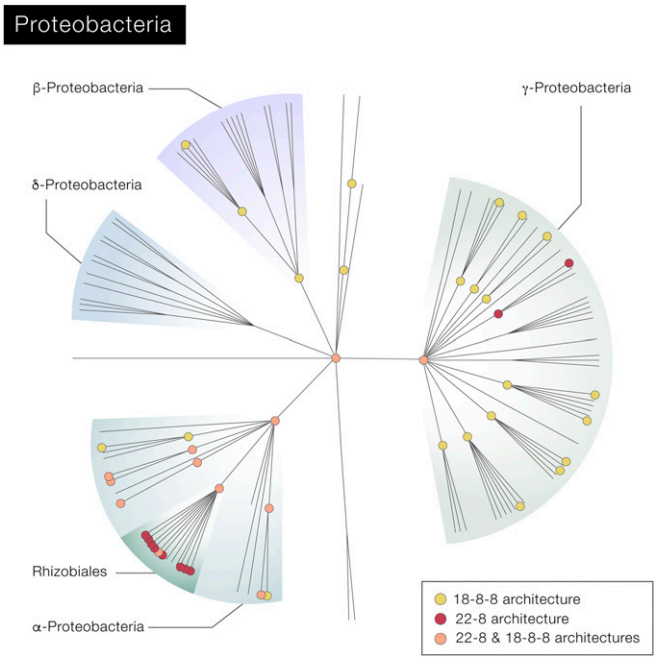

Fig. 2. The taxonomic tree of the MB-families in Proteobacteria with multibarrel proteins. Clades with an architecture of 22–8 are marked with a red circle, clades with an architecture of 18–8-8 are marked with a yellow circle, and clades with both architectures are marked with an orange circle. Even though the 22–8 MB family with its 99 proteins (marked in red and orange) has twice as many representatives than the 48 proteins in the 18–8-8 MB-family (marked in yellow and orange), it appears mostly in Rhizobiales, which are α-Proteobacteria, and four other orders, while the proteins of the 18–8-8 MB-family are present in many (14) orders in α, β, and γ Proteobacteria. The tree was rendered using Dendroscope (45).

which contains barrel families predicted to have 38 and 40 strands, larger than any OMBB reported to date. Some families in our set include other domain families that are not integral to the membrane, some well-characterized (such as POTRA domains) and others without any clear homology to a known family (*SI Appendix*, Fig. S1).

Contact Map Predictions Support the Multibarrel Nature of the Protein Chains. The sequence matches for the architectures described above are generally full-length matches to known OMBBs, separated by extended linkers of 22 ± 13 residues (*SI Appendix*, Fig. S4). Nevertheless, the connected barrel domains might still fold into larger fused barrels (13). We use contact map predictions to examine this.

When there are many homologous and variable protein sequences that fold into the same structure, computational tools can predict the two-dimensional residue–residue contact map of their structure (18, 19). Because the prediction of the contacts exploits the correlated mutation signals in the multiple sequence alignment, a homologous protein with a known structure that will serve as a template is not required. In theory, correlations between mutations are a consequence of compensatory changes between residues in close physical contact (20). In practice, however, correlated mutations are also observed between residues that are not in contact. To extract only true contacts from correlated mutations, state-of-the-art contact map predictors employ deep networks (21). Herein, we used four contact maps predictors, which were identified as accurate (21–24): RaptorX (18), TripletRes (25), trRosetta (26), and DeepMetaPSICOV (27).

Accurate prediction of contact maps requires multiple sequence alignments of many diverse and homologous sequences (4, 18, 24); when only a few homologs are available, predictions tend to be of low quality. For example, 148 effective homologs were needed by RaptorX to reach an accuracy of 0.55 (in a 0-through-1 scale) in the top L/5 medium-range contacts in membrane proteins (L being sequence length) (18). Only 13 MB-families or families with single barrels of outstandingly large size have more than 148 homologous proteins, and even a smaller number have over 148 effective homologs (i.e., after redundancy cleaning). Here, we predicted contact maps for all families with at least 50 proteins to find cases in which despite the small number of homologs, the signal is strong enough for accurate predictions. The anticipation was that in most of them there would be no signal.

To filter the many expected contact maps with no strong signal, we defined strict criteria for contact maps that we consider accurate. In proteins with one or more barrel domains, we expected that (e.g., Fig. 3) 1) within each of the barrels, there would be contact signals near the diagonal between every two consecutive strands, allowing to infer the number of strands in the barrel even if the predictions are noisy; 2) there would be a barrel-closing signal far from the diagonal, indicative of contact between the N- and C-terminal strands, which are adjacent in the folded barrel; and 3) there would be no significant contacts between strands in different barrels or contacts between nonadjacent strands in the same barrel. With these criteria in mind, we used RaptorX, TripletRes, trRosetta, and DeepMetaPSICOV to predict contact maps for the 21 families containing more than 50 protein sequences (*SI Appendix*, Table S2).

Fig. 3 shows an example of a contact map predicted by RaptorX using the 1,959 proteins of MB-family 001, which supports the hypothesis that MB-family 001 has two separate eight-stranded barrel domains. The contact maps predicted using TripletRes,

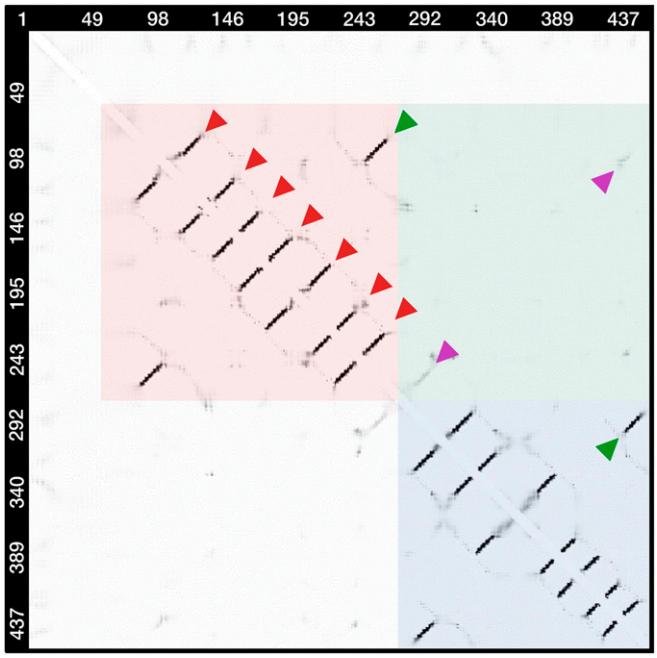

Fig. 3. RaptorX prediction of the contact map of MB-family 001, comprising 1,959 homologous proteins and manifesting an 8–8 architecture (a pair of 8-stranded barrels). This predicted double-barrel architecture is strongly supported by the clear (dark) signals for contacts between the spatially adjacent beta strands of the N- and C-terminal barrels within the red and blue frames, respectively. That there is, in essence, no signal within the green frame is a clear indication that the alternative 16-stranded barrel architecture is not supported.

trRosetta, and DeepMetaPSICOV also support an 8–8 architecture (*SI Appendix*, Fig. S5).

Contact maps are, by definition, symmetric; thus, we focus our discussion on the triangle above and to the right of the diagonal. Looking at the first part of the predicted contact map (~200 residues; red background), we see seven lines perpendicular to the main diagonal (red arrows). Each of these diagonal lines describes contacts between two consecutive strands, with residues of increasing number in the first strand in contact with residues of decreasing number in the second strand. Thus, the seven lines describe contacts between strands 1–2, 2–3, 3–4, . . ., and 7–8. Finally, contacts between strands 8 and 1, in the upper right corner and marked by a green arrow, manifest a barrel-closing signal. The second barrel domain (blue background) has a similar pattern, with contacts between strands 8 and 1 (i.e., barrel-closing signal, marked by a green arrow). Beyond these clear patterns, there are only a few weak correlations in the predicted contact map, which we consider to be noise. In particular, the predicted signals of contacts between the two barrels (pale green background, marked by the purple arrows) are very weak. The proteins in this MB-family have two segments homologous to eight-stranded barrels. Theoretically, rather than forming two separate barrels, they could fuse to form a single large barrel. Since there are sequence matches to all the strands, the most likely single barrel would comprise 16-strands. In this case, we would expect to see a contact between strands 8 and 9 and between strands 1 and 16 (marked by purple arrows), which are only weakly observed in the predicted contact map. Namely, the contacts supporting the 8–8 architecture hypothesis are much stronger than the contacts supporting the 16-strand hypothesis, and by visual inspection, we conclude that the contact map supports the prediction of two consecutive eight-stranded barrel domains.

Although the contact prediction programs output contact probabilities, these have dependencies among them that makes combining them in a mathematically meaningful way challenging. However, as the example in Fig. 3 demonstrates, the signal is sometimes clear enough to identify the MB-architectures. To determine whether the contact prediction supports an MB-architecture, we compare the predictions for closing signals of the single barrels and those for closing a single joined barrel (Fig. 4). In an MB-architecture, we expect the predicted probabilities of the former to be large and of the latter to be small. We consider these probabilities in the context of a null distribution. For the null distribution, we consider the predicted probabilities of intrabarrel contacts between residues that are not in contact with each other in both the single-barrel and MB-architectures and so are likely false positives. We considered a contact significant when it was outside the null distribution. The probabilities assigned to the contacts between strands 1 and 8 and between strands 9 and 16, which support the 8–8 architecture, are both 1, much higher than any contact in the null distribution. In comparison, the probability assigned to the contact between strands 1 and 16, which supports a single-barrel architecture, is 0.0088, and it is in the 75th percentile. The probability assigned to the contact between strands 8 and 9, which would be an intrabarrel contact in a single-barrel architecture and a weaker interbarrel contact in an 8–8 architecture, is 0.15, and it is in the 92nd percentile. Overall, the contact probabilities clearly support an 8–8 architecture. Notice however, that since the true null distribution is unknown and the sample size is small, this is not a true statistical analysis but rather a demonstration of the clarity of the prediction signal.

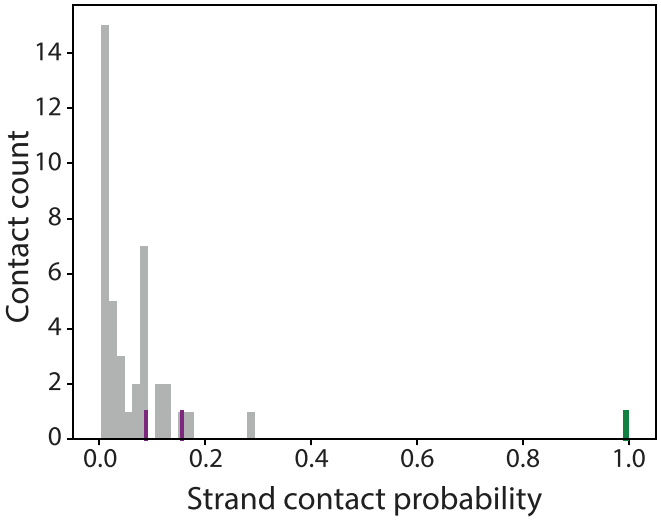

Fig. 4. RaptorX's contact probabilities between the beta strands in the predicted contact map of MB-family 001 shown in Fig. 3 clearly support an 8–8 architecture. The probabilities for contacts supporting an 8–8 architecture (Fig. 3, green arrows) are marked in green, and the probabilities for contacts supporting a 16-stranded single-barrel architecture (Fig. 3, purple arrows) are marked in purple. The histogram of probabilities that are expected to be intrabarrel contacts are shown in gray and used as a null distribution.

SI Appendix, Figs. S6 and S7 show the predicted contact maps of two more MB-families that clearly support their predicted MB-architectures: three contact maps of MB-family 007 with 8–8-8 architecture (*SI Appendix*, Fig. S6; although a barrel-closing signal for the C-terminal barrel is missing) and two contact maps for MB-family 012 with 12–12 architecture (*SI Appendix*, Fig. S7). Similarly, *SI Appendix*, Fig. S8 shows two predicted contact maps of family 002, and *SI Appendix*, Fig. S9 shows the predicted contact map of family 003. The contact maps confirm the presence of a barrel with 40 strands in family 002 and a barrel with 38 strands in family 003, both larger than any previously observed barrel domain. Altogether, predicted contact maps support five previously undocumented architectures (Fig. 3 and *SI Appendix*, Figs. S5–S9). However, contact maps do not always yield clearly interpretable signals, even when there are many homologs. For example, in MB-family 004, predicted to form an 8–8 architecture, there are 813 sequences, but the contact map does not unequivocally support a single architecture (*SI Appendix*, Fig. S10). Another case is MB-family 000, with 4,187 homologs, predicted to form a 12–12 architecture. The corresponding contact map clearly supports the N- but not the C-terminal barrel domain (*SI Appendix*, Fig. S11). This could indicate that the correlated mutation signal for the C-terminal barrel is particularly weak or that it is not a barrel domain. The contact predictions results are summarized in *SI Appendix*, Table S2.

Overall, the predicted contact maps provide further support for five previously undescribed barrel architectures (*SI Appendix*, Table S2). Three of these are MB-architectures, 8–8 (Fig. 3 and *SI Appendix*, Fig. S5), 8–8-8 (*SI Appendix*, Fig. S6), and 12–12 (*SI Appendix*, Fig. S7), and two feature single barrels of 40 strands (*SI Appendix*, Fig. S8) and 38 strands (*SI Appendix*, Fig. S9), larger than previously observed. Only in one MB-family, the contact map contained clear signals that did not match our predicted architecture (*SI Appendix*, Fig. S11). That the predicted contact maps agree with the multibarrel annotation predicted by the sequence homology methods and that only one of the five clear contact maps conflicts with the homology-based predictions suggests that most of the other architectures predicted only by sequence homology are also correct.

Functional and Taxonomic Analysis of the Multibarrel MB-families. To better understand the significance of multibarrel proteins and the evolutionary pathways that may have led to their emergence,

we study their predicted functions and the taxonomic distribution within each MB-family. We predict the functions by homology transfer, as identified by HHpred, and predict disordered regions using Quick2D (28). In cases in which multiple MB-architectures are evolutionarily linked, we used these relationships to trace their evolution. From the 186 MB-families, we elaborate on two MB-families in which the functions of the individual barrel domains are complementary and an MB-family with representatives in *Escherichia coli*.

The PLA1-PLA1 (12–12) MB-family. MB-family 052 includes six proteins from γ-Proteobacteria with an architecture of 12–12 (Fig. 1, *Right*; 12–12 marked with two stars). The proteins in this MB-family have two repeated 12-stranded OMBBs homologous to OMPLA (annotated as PLA1 in Pfam), connected by a linker of 40 to 50 residues. OMBBs of the OMPLA MB-family are known to natively form homodimers, suggesting that in this case, interchain interactions may have been replaced with interdomain interactions. Possibly, the interdomain interactions facilitate regulation.

OMPLAs are widespread in gram-negative bacteria and were implicated in the virulence of some pathogenic species (29), acting as hydrolases that recognize and act on a broad spectrum of phospholipids. Their monomeric form is inactive and only becomes active upon calcium-dependent homodimerization (7, 8). Dimerization forms two binding pockets at the interface between the barrels, each harboring two calcium-binding and two active sites, which we label I and II. While residues from both monomers are involved in forming each calcium-binding site, all active residues in one active site belong to the same barrel. These residues form a catalytic triad in each active site, which has phospholipase A1 and A2, lysophospholipase A1 and A2, and mono- and diacyl glyceride lipase activities, hydrolyzing phospholipids to fatty acids and lysophospholipids (29).

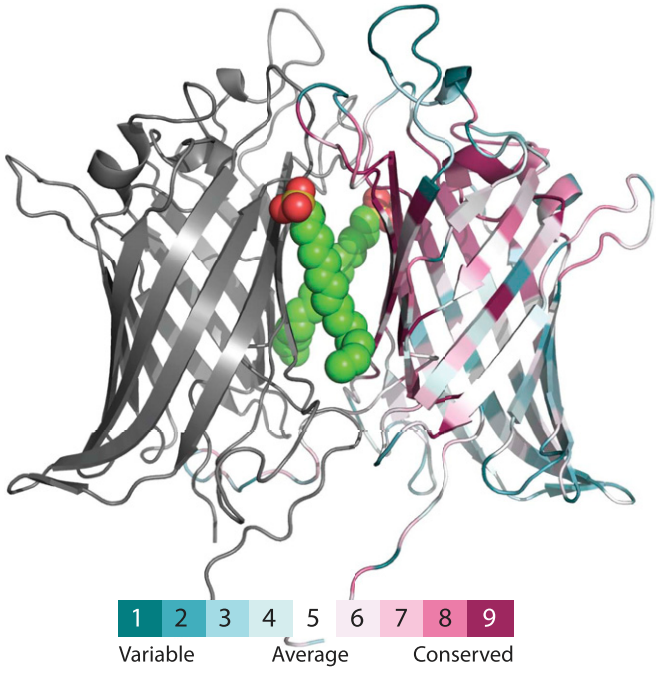

Fig. 5. Homology model of a 12–12 double barrel based on the homodimeric structure of *E. coli* OMPLA (1QD6) (8) as template. One barrel is colored by the ConSurf conservation grades of the *E. coli* homolog and the other in gray. Based on the template structure, two inhibitor molecules are tentatively docked in the interface between the barrels. The core of the barrel and the interface between the barrels are evolutionarily conserved, and the residues facing the membrane are variable.

The 052 MB-family includes six proteins, all belonging to γ-Proteobacteria. Four are from three *Oceanospirilalles* species (*Zymobacter palmae*, *Halotalea alkalienta*, and *Carnimonas nigrificans*) and the remaining two belong to the *Agarivorans* genus (*Agarivorans albus* and *Agarivorans gilvus*; *SI Appendix*, Fig. S12). Proteins in this MB-family are within a conserved genomic environment only in *Oceanospirillales*, composed of genes encoding proteins that depend or act on cations. In *Agarivorans*, such an environment is absent (*SI Appendix*, Fig. S13).

When the individual barrel domains of the six multibarrel proteins are clustered together with the barrel domains of single-barrel OMPLA proteins (*SI Appendix*, Fig. S14), the barrel domains from the multibarrel proteins in *Oceanospirilalles* cluster closely with Proteobacterial single-barrel proteins, while those from *Agarivorans* are far separated, substantiating that although small, this MB-family is the result of two independent duplication events from two different single-barrel ancestors (*SI Appendix*, Fig. S12).

Fig. 5 shows a homology model of a double barrel based on the homodimeric structure of the *E. coli* single-barrel OMPLA as template. Sequence comparison suggests that all catalytic and calcium-binding residues are conserved in both barrels from *Oceanospirillales* but only in the second barrel from *Agarivorans* (*SI Appendix*, Fig. S15). In the first barrel from *Agarivorans*, two out of the three catalytic residues and three out of four calcium-binding residues involved in the formation of active site I are mutated to residues with biochemical properties incompatible with the function of their counterparts in *E. coli,* indicating that this active site was either lost or evolved into a different function. This latter observation suggests that one benefit of connecting the barrels within the same polypeptide chain may be the ability to diverge and explore functional adaptations not possible in a homodimer.

The PagP-LptD (8–26) MB-family. MB-family 093 is composed of 19 proteins from the Rhodocyclaceae branch of β-Proteobacteria with an 8–26 architecture (Fig. 1, *Right*; 8–26 marked with a star), annotated as a combination of a Lipid A modifying barrel of eight strands (PagP) and the 26-stranded LptD barrel of the lipopolysaccharide transport (Lpt) complex, connected by a loop of more than 100 residues predicted to be disordered. Here, the two barrels share the same lipid substrate, and the connection between them might increase the efficiency of the process.

Lipid A is the lipid component of lipopolysaccharide (LPS), the major component of the outer membrane of gram-negative bacteria (30), which is synthesized in the inner membrane and transferred to the outer membrane by the LPS transport system. The last protein in the transport system is LptD, which transports LPS to the outer leaflet of the outer membrane. The link to the Lipid A modifying barrel might shorten the distance an LPS molecule has to travel between the transport system and the modifying enzyme, thereby increasing the efficiency of the modification. This MB-family comprises proteins from β-proteobacteria, specifically species of the *Propionivibrio* and *Rhodocyclus* genera, whose conserved genomic environment is the same as that of LptD but not PagP in *E. coli* (*SI Appendix*, Fig. S16).

The YjbH-GfcD (12–12) MB-family. This MB-family is the largest in our set, comprising 4,187 proteins, mostly from Proteobacteria (Fig. 1, top-right; 12–12 marked with two stars), with representatives also from other bacterial phyla. As noted above, our attempt to provide further support for the proposed 12–12 architecture based on contact prediction was unsuccessful (*SI Appendix*, Fig. S11). Regardless, however, this MB-family is of particular interest as it is the only one with a representative in *E. coli.* Proteins from this MB-family can be divided into several subfamilies using CLANS with an E-value of 10^{-5}. The central one encompasses 3,440 proteins (82%) from almost exclusively γ-Proteobacteria. The second largest is composed of 537 proteins (13%) almost exclusively from α-Proteobacteria. The third is composed of 82 proteins

(2%) and has representatives mostly from α- and γ-Proteobacteria. Two smaller subfamilies are originated from Chlamydiae (30 proteins) and from δ-Proteobacteria (26 proteins). The rest of the subfamilies are small and contain mostly Proteobacteria. Five sequences in this MB-family are attributed to Euryarchaeota, a phylum in archaea. It is very unlikely that the sequences fold into functional barrels. Possible explanations are that the sequences are attributed to archaea due to sequencing error or that they were horizontally transferred to archaea but are nonfunctional.

Two paralogs of these multibarrels, YjbH and GfcD/YmcA, are found in related operons in *E. coli* (*SI Appendix*, Fig. S16) and are clearly connected to the production of exopolysaccharide capsules (31–33). Of these, the *gfcABCDE-etp-etk* operon has been studied extensively for its role in the formation of the group 4 capsule polysaccharide, and the structures and functions of most proteins encoded are understood (*SI Appendix*, Fig. S17) (34–37), except for GfcA (which is a short protein predicted to be natively unstructured) and GfcD. The yjbEFGH operon, which is a paralog of *gfcABCD*, has also been implicated in the production of an extracellular polysaccharide but is much less studied. Both YjbH and GfcD are predicted to carry an N-terminal OMBB of the FapF family, a middle globular peptidoglycan-binding domain, and an unknown C-terminal OMBB without full-length homology to any known OMBB family but predicted to be composed of 12 strands. The molecular advantage in coupling these two OMBB domains into one protein is unclear, as neither domain has close homologs among single barrels.

Poly-8 MB-Families. Finally, tandemly repeated eight-stranded OMBBs are the most common architectures in our set, with many different MB-families from diverse taxonomic groups emerging in parallel from different eight-stranded single barrels (Fig. 1). Some of the MB-families can be grouped to two MB-superfamilies, containing evolutionarily related MB-families. The first poly-8 MB-superfamily includes 289 proteins from 28 MB-families, with between two and 11 OMBB domains, found only in Bacteroidetes (Fig. 1, multiple architectures near the bottom marked with a star), especially in species from the *Prevotella* (105 proteins) and *Bacteroides* (157 proteins) genera, which represent the dominant bacteria of the human microbiome (38). The second poly-8 MB-superfamily includes 118 proteins, with two or three barrel domains and an α-helical linker domain, found in various Bacteroidetes genera. The evolution of the two MB-superfamilies includes both amplification events that formed the $(8)_{11}$ protein from an 8–8-8 ancestral protein and deletion events that formed 8–8 proteins from 8 to 8-8 proteins and $(8)_6$ from $(8)_7$ and $(8)_9$ proteins. Fully tracing the evolutionary pathways that formed the poly-8 proteins can reveal insights to the evolutionary process itself and will be described in a separate manuscript.

Conclusions

Given their internal symmetry and their distinction from soluble forms, OMBBs are an attractive class of folds for tracing evolutionary processes (5, 11, 39). Scientists have long been familiar with proteins forming single OMBB domains (5, 11, 12), as well as with complexes of multiple barrels such as the OMPLA homodimer (7, 8), the porin trimer (9, 10), and the type-9 translocon heterodimer (6). Our study allowed us to identify several previously unknown OMBB forms, including the largest currently reported, and many proteins with multiple OMBB domains in the same chain, often in tandem arrangement. The existence of such multibarrel proteins has not been established to date, although it has been a matter of conjecture (12).

In soluble proteins, the concatenation of two domains is expected to yield a two-domain protein. It is not obvious that this also applies to the membrane-embedded beta barrels, since, owing to their nature, their concatenation in tandem may also result in a single barrel of greater diameter, as shown with the concatenation of two eight-stranded OmpX barrels by Arnold et al. (13). Indeed, Remmert et al. (11) and Franklin et al. (5) suggested that some large barrels evolved through the addition of β-hairpins to smaller barrels. In the cases we describe here, the multibarrels almost invariably combine full-length matches to existing single-barrel proteins, connected by substantial linkers with a median length exceeding 20 residues. This is in contrast to the extremely short linker of two residues used by Arnold et al. (13) to successfully connect the two eight-stranded barrels into a larger barrel. Additionally, in the cases for which we could predict contact maps, these clearly support the presence of multiple independent barrels along the polypeptide chain.

The large number of 34 previously unknown MB-architectures that we identified is notable in light of the conservative search procedures that we used. In particular, to minimize the number of falsely reported architectures, we 1) inspected only proteins with at least two full matches to the seed sequences and 2) inspected the proteins we found with HHpred and further sequence annotation tools. We believe that with a larger set of seed sequences or using more sensitive homology search tools, it is likely that we would have found even more previously unknown MB-architectures. Even with these conservative choices, we found many multibarrels from a multitude of bacterial phyla, highlighting how frequent these proteins are in nature.

Our example analyses show that multiple evolutionary processes lead to the diversity of multibarrel proteins. One is descent with modification from an ancestor that already contained multiple barrel domains, as illustrated by the YjbH/GfcD family, which is ancient and conserved in many bacterial lineages that have an outer membrane. Another is the amplification of a beta barrel that normally forms oligomers for functionality into a single chain with multiple barrel domains, as observed in two independent lineages of OMPLA. Also, we saw the fusion of different barrels acting within the same biological process, as described for the PagP-LptD MB-family involved in LPS biogenesis. The poly-8 MB-superfamilies allow one to trace the evolutionary events forming them. A preliminary examination revealed lineage-specific deletion of individual domains from a multibarrel ancestor and multiple independent amplification of barrel domains. A detailed study tracing the evolutionary relationships among these multibarrels is underway. The diversity of these evolutionary mechanisms underscores the frequency with which multibarrels occur in many lineages.

We find that multibarrels evolve frequently, but only some of these appear to be retained over long evolutionary periods. Most appear in only one or a few bacterial genera and some even only in individual species, suggesting a large turnover through de novo evolution and subsequent gene loss. This is in agreement with the results of genetic processes in general, which typically remain lineage-specific even when they become fixated in the genome and disappear when the lineage dies out. Nevertheless, the large number of multibarrels that have become fixated in genomes even only with a very narrow phylogenetic distribution begs the question of the biological advantages they may confer over their single-barrel homologs. Several can be envisaged: 1) increasing the efficiency of a given pathway, as we propose for the PagP-LptD fusion in LPS biogenesis; 2) opening avenues for the divergence and separate functional adaptation of domains previously engaged in homo-oligomerization, as in the OMPLA; and 3) increasing the avidity of proteins engaged in binding multiple adjacent epitopes. This last consideration may offer a hypothesis for the prevalence of poly-8 proteins in the gut genera *Bacteroides* and *Prevotella*, since one of the functions reported for OmpA is adhesion (40) and the long polysaccharides in the human gut offer many adjacent epitopes of the same or similar nature.

An important open question in understanding multibarrels is their mechanism of insertion in the outer membrane. In single-barrel outer-membrane proteins, this process is mediated by the β-barrel assembly machinery (BAM) complex (41), which recognizes a sequence signal in the C-terminal strand of the barrel to be inserted. The process has been studied mainly in

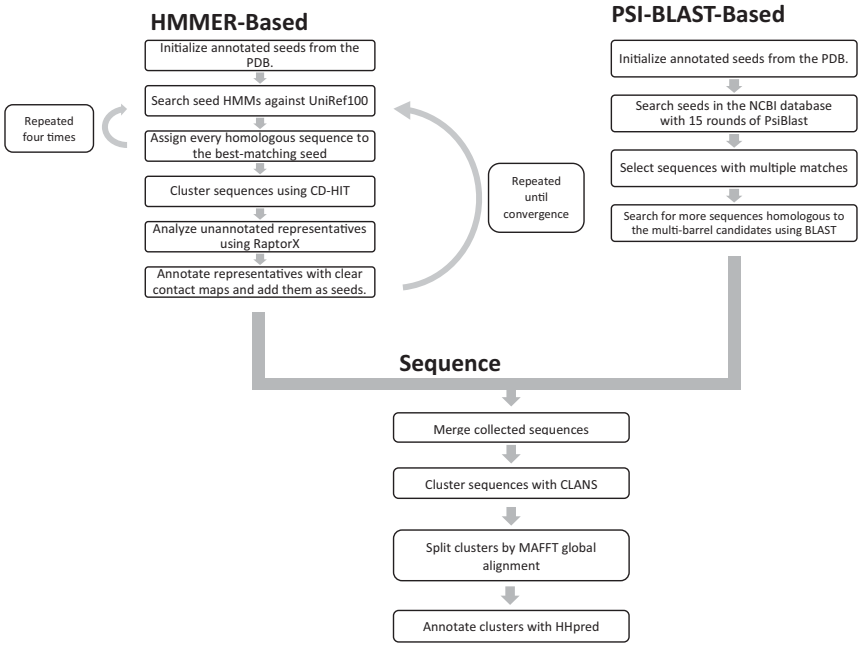

Fig. 6. A flowchart of our protocol to find multibarrel proteins.

Proteobacteria and mitochondria, so the nature of the sorting signal is still understood only in general terms and may be partly lineage specific (42–44). In some of our multibarrels, mainly in those that have clearly arisen recently, we can readily recognize the presence of the sorting signal in each constituent barrel, but this may be the result of recent fusion, as opposed to the need for two independent signals within the same protein. So far, we have not attempted to build a prediction tool for sorting signals and are not aware that such a tool is available. Hence, it is unknown whether one signal at the C terminus is sufficient for the insertion of the entire protein into the outer membrane or each barrel is inserted independently by the BAM machinery.

The extent of previously unknown OMBBs identified herein highlights their functional importance. Furthermore, OMBBs may have even more undescribed functions: a preliminary search that we carried out revealed many proteins combining beta barrels and nonbarrel domains; the latter include DNA binders, histidine kinase sensors, and ABC transporters. Overall, the architectures identified herein suggest that OMBBs have an even greater functional role than was previously thought.

Materials and Methods

Two independent pipelines were implemented and used to discover previously unknown OMBB architectures based on existing OMBB structures (Fig. 6). One is based on the HMMER (16) search engine and the other based on PsiBLAST (17). In both cases, the goal was to detect protein sequences comprising two or more nonoverlapping sequence segments which are similar to OMBBs of known structure. These hits are putative multi-OMBBs or OMBBs with outstandingly large barrels. To corroborate this homology-based inference, where possible, we predicted contact maps using four contact prediction programs: RaptorX (18), TripletRes (25), trRosetta (26), and DeepMetaPSICOV (27). A detailed description of both pipelines as well as the off-the-shelf methodology that was used to further support the predictions based on residue–residue contact predictions for phylogenetic analysis and for producing the homology model of Fig. 5 is included in *SI Appendix, Supplementary Methods*.

Data Availability. All study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. This research has been supported by Grant 94747 by the Volkswagen Foundation. N.B.-T.'s research is supported in part by the Abraham E. Kazan Chair in Structural Biology, Tel Aviv University.

1. S. A. Paschen *et al.*, Evolutionary conservation of biogenesis of β-barrel membrane proteins. *Nature* **426**, 862–866 (2003).
2. D. Duy, J. Soll, K. Philippar, Solute channels of the outer membrane: From bacteria to chloroplasts. *Biol. Chem.* **388**, 879–889 (2007).
3. J. Bassler, B. Hernandez Alvarez, M. D. Hartmann, A. N. Lupas, A domain dictionary of trimeric autotransporter adhesins. *Int. J. Med. Microbiol.* **305**, 265–275 (2015).
4. J. Pereira, A. N. Lupas, The origin of mitochondria-specific outer membrane β-barrels from an ancestral bacterial fragment. *Genome Biol. Evol.* **10**, 2759–2765 (2018).
5. M. W. Franklin *et al.*, Evolutionary pathways of repeat protein topology in bacterial outer membrane proteins. *eLife* **7**, e40308 (2018).
6. F. Lauber, J. C. Deme, S. M. Lea, B. C. Berks, Type 9 secretion system structures reveal a new protein transport mechanism. *Nature* **564**, 77–82 (2018).
7. N. Dekker, J. Tommassen, A. Lustig, J. P. Rosenbusch, H. M. Verheij, Dimerization regulates the enzymatic activity of *Escherichia coli* outer membrane phospholipase A. *J. Biol. Chem.* **272**, 3179–3184 (1997).
8. H. J. Snijder *et al.*, Structural evidence for dimerization-regulated activation of an integral membrane phospholipase. *Nature* **401**, 717–721 (1999).
9. M. S. Weiss *et al.*, Molecular architecture and electrostatic properties of a bacterial porin. *Science* **254**, 1627–1630 (1991).
10. M. S. Weiss, T. Wacker, J. Weckesser, W. Welte, G. E. Schulz, The three-dimensional structure of porin from *Rhodobacter capsulatus* at 3 A resolution. *FEBS Lett.* **267**, 268–272 (1990).
11. M. Remmert, A. Biegert, D. Linke, A. N. Lupas, J. Söding, Evolution of outer membrane β-barrels from an ancestral β β hairpin. *Mol. Biol. Evol.* **27**, 1348–1358 (2010).
12. B. L. Reddy, M. H. Saier Jr., Properties and phylogeny of 76 families of bacterial and eukaryotic organellar outer membrane pore-forming proteins. *PLoS One* **11**, e0152733 (2016).
13. T. Arnold, M. Poynor, S. Nussberger, A. N. Lupas, D. Linke, Gene duplication of the eight-stranded β-barrel OmpX produces a functional pore: A scenario for the evolution of transmembrane β-barrels. *J. Mol. Biol.* **366**, 1174–1184 (2007).
14. H. M. Berman *et al.*, The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
15. B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, C. H. Wu, UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).
16. S. R. Eddy, Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
17. S. F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
18. S. Wang, S. Sun, Z. Li, R. Zhang, J. Xu, Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.* **13**, e1005324 (2017).
19. S. Wang, S. Sun, J. Xu, Analysis of deep learning methods for blind protein contact prediction in CASP12. *Proteins* **86** (suppl. 1), 67–77 (2018).
20. D. de Juan, F. Pazos, A. Valencia, Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **14**, 249–261 (2013).

21. J. Schaarschmidt, B. Monastyrskyy, A. Kryshtafovych, A. M. J. J. Bonvin, Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins* **86** (suppl. 1), 51–66 (2018).
22. J. Xu, S. Wang, Analysis of distance-based protein structure prediction by deep learning in CASP13. *Proteins* **87**, 1069–1081 (2019).
23. L. A. Abriata, G. E. Tamò, B. Monastyrskyy, A. Kryshtafovych, M. Dal Peraro, Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins* **86** (suppl. 1), 97–112 (2018).
24. C. Bassot, A. Elofsson, Accurate contact-based modelling of repeat proteins predicts the structure of new repeats protein families. *PLoS Comput. Biol.* **17**, e1008798 (2021).
25. Y. Li *et al*., Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. *PLoS Comput. Biol.* **17**, e1008865 (2021).
26. J. Yang *et al*., Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 1496–1503 (2020).
27. S. M. Kandathil, J. G. Greener, D. T. Jones, Prediction of inter-residue contacts with DeepMetaPSICOV in CASP13. *Proteins* **87**, 1092–1099 (2019).
28. L. Zimmermann *et al*., A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.* **430**, 2237–2243 (2018).
29. H. J. Snijder, B. W. Dijkstra, Bacterial phospholipase A: Structure and function of an integral membrane phospholipase. *Biochim. Biophys. Acta* **1488**, 91–101 (2000).
30. I. Botos *et al*., Structural and functional characterization of the LPS transporter LptDE from Gram-negative pathogens. *Structure* **24**, 965–976 (2016).
31. L. Ferrières, S. N. Aslam, R. M. Cooper, D. J. Clarke, The yjbEFGH locus in *Escherichia coli* K-12 is an operon encoding proteins involved in exopolysaccharide production. *Microbiology (Reading)* **153**, 1070–1080 (2007).
32. A. Peleg *et al*., Identification of an *Escherichia coli* operon required for formation of the O-antigen capsule. *J. Bacteriol.* **187**, 5259–5266 (2005).
33. C. Whitfield, Biosynthesis and assembly of capsular polysaccharides in *Escherichia coli*. *Annu. Rev. Biochem.* **75**, 39–68 (2006).
34. K. Sathiyamoorthy, E. Mills, T. M. Franzmann, I. Rosenshine, M. A. Saper, The crystal structure of *Escherichia coli* group 4 capsule protein GfcC reveals a domain organization resembling that of Wza. *Biochemistry* **50**, 5465–5476 (2011).
35. C. Dong *et al*., Wza the translocon for *E. coli* capsular polysaccharides defines a new class of membrane protein. *Nature* **444**, 226–229 (2006).
36. M. Salomone-Stagni, F. Musiani, S. Benini, Characterization and 1.57 Å resolution structure of the key fire blight phosphatase Amsl from *Erwinia amylovora*. *Acta Crystallogr. F Struct. Biol. Commun.* **72**, 903–910 (2016).
37. D. C. Lee, J. Zheng, Y. M. She, Z. Jia, Structure of *Escherichia coli* tyrosine kinase Etk reveals a novel activation mechanism. *EMBO J.* **27**, 1758–1766 (2008).
38. E. L. Johnson, S. L. Heaver, W. A. Walters, R. E. Ley, Microbiome and metabolic disease: Revisiting the bacterial phylum Bacteroidetes. *J. Mol. Med. (Berl.)* **95**, 1–8 (2017).
39. V. Alva, A. N. Lupas, From ancestral peptides to designed proteins. *Curr. Opin. Struct. Biol.* **48**, 103–109 (2018).
40. A. W. Confer, S. Ayalew, The OmpA family of proteins: Roles in bacterial pathogenesis and immunity. *Vet. Microbiol.* **163**, 207–222 (2013).
41. N. Noinaj, J. C. Gumbart, S. K. Buchanan, The β-barrel assembly machinery in motion. *Nat. Rev. Microbiol.* **15**, 197–204 (2017).
42. S. Kutik *et al*., Dissecting membrane insertion of mitochondrial β-barrel proteins. *Cell* **132**, 1011–1024 (2008).
43. T. J. Knowles, A. Scott-Tucker, M. Overduin, I. R. Henderson, Membrane protein architects: The role of the BAM complex in outer membrane protein assembly. *Nat. Rev. Microbiol.* **7**, 206–214 (2009).
44. K. A. Diederichs *et al*., Structural insight into mitochondrial β-barrel outer membrane protein biogenesis. *Nat Commun* **11**, 3290 (2020).
45. D. H. Huson *et al*., Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinf.* **8**, 460 (2007).