Gram-negative outer membrane proteins with multiple β-barrel domains – supplementary material

Ron Solan[1#], Joana Pereira[2#†], Andrei N. Lupas[2*], Rachel Kolodny[3*], Nir Ben-Tal[1*]

[1] Department of Biochemistry and Molecular Biology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel.

[2] Department of Protein Evolution, Max-Planck-Institute for Developmental Biology, Tübingen 72076, Germany.

[3] Department of Computer Science, University of Haifa, Mount Carmel, Haifa, 3498838, Israel.

[#] Co-first authorship

*Correspondence: Andrei N. Lupas, Email: andrei.lupas@tuebingen.mpg.de; Rachel Kolodny, Email: trachel@cs.haifa.ac.il; Nir Ben-Tal, Web: http://bental.tau.ac.il, Email: bental@tauex.tau.ac.il;

[†]Current address: Biozentrum, University of Basel, 4056 Basel, Switzerland

Supplementary Methods

We used two methods to gather protein sequences likely to have multiple barrel domains: A HMMER-based method and a PsiBLAST-based method. The sequences found by either method were clustered by similarity, and the clusters were manually inspected to ensure every cluster had a single architecture (Figure 6).

HMMER-based sequence searches over UniRef100

An overview of the search procedure is depicted in Figure 6, and the steps are described below.

1. Constructing the set of seeds (Single-barrel HMMs annotated with barrel size)

We started with all 275 PDB sequences of OMBB proteins as identified by MemProtMD (46); the most recently solved structure in the set we used is 6FSU (released 11/2018). We trimmed the N- and C-terminal ends that were not part of the barrels. Then, we selected from this set a less-redundant subset of 98 proteins using CD-HIT v4.7 (47) and a 90% identity threshold. Using the corresponding structures, we annotated each protein sequence with the number of strands in its barrel.

2. Enriching the seed HMMs

For each OMBB in our set, we identified homologs iteratively, building our seed HMMs in four rounds. Initially, the set of homologs for each OMBB was the trimmed sequence itself. In every iteration, we used HMMER 3.1b2 (17) to list all the sequence homologs in UniRef100 of any of the seed HMMs (using an E-value threshold of 10^{-5}). Then, we traversed the list in ascending order of E-value: for every sub-sequence that matched a seed, provided it did not intersect with already-considered subsequences on the same sequence, we added it (only) to its best-matched seed. This way, a sequence with high similarity to one barrel and low similarity to another would be added only to the set of homologs of the former. We further filtered the set of homologs, enforcing high coverage: the sequences must match 90% of the seed HMM. Finally, we added the identified homologous sections to the seed alignment and rebuilt the MSAs and HMMs of the seeds. This procedure constructed seed HMMs that included remote homologs from the same family, while not being corrupted by close homologs from a different barrel family.

3. Adding seeds for new predicted single OMBBs

In the process of building the HMMs for the seeds, we found UniRef100 sequences that could not be added to the set of homologs of any seed, because the coverage of their matches was too low. These sequences may be members of OMBB families that are not represented in the PDB, or remote homologs of families that are represented. Given that we wanted to find as many OMBB families as possible, we inspected representatives of these sequences, and added those which we found to be OMBBs to the seed set. We performed four rounds of new seed selection, and after each round we reconstructed the HMMs using the protocol described above.

To limit the number of representatives checked, we selected only the subset that is more likely to contain barrel domains. For that, in the first three rounds of new seed selection, we collected sequences with at least two partial matches. We clustered them with CD-HIT and used RaptorX to construct an MSA and to predict a contact map for a representative from each cluster. When the predicted contact map of a sequence had a clear β-barrel pattern from which we could infer the strand number via a manual inspection, we concluded that this is likely an OMBB and added the barrel domain as a seed. In the fourth and last round, we collected all sequences with only a single partial match, clustered them using CD-HIT, and picked representatives from the largest clusters. We predicted contact maps for the representatives using RaptorX, and if the predicted contact map manifested a clear β-barrel-like pattern we added them as seeds as well. Our process converged after the third and fourth rounds of new seed selection.

4. Using the seed HMMs to identify multi-barrel proteins

Our goal was to find with high sensitivity multi-barrel proteins, and to accurately annotate their individual domains. To achieve this, we started with a two-step procedure to collect a set of multi-barrels, aimed at minimizing the false-negatives: (1) we searched UniRef100 for homologs of the proteins in our seed set, using HMMER and the very lax E-value of 1000. This was the initial database of potential homologs, which were possibly false. (2) Within this initial database, we searched again using a stricter (maximal) E-value of 0.1.

The challenge in annotating the barrel domains of the sequences found in UniRef100 is that each sequence has many alignments to our seeds, sometimes matching the same residues in that sequence. To address this challenge, we devised the following procedure: First, we identified and corrected false local matches. HMMER searches for local alignments between the seed HMM and the UniRef100 sequences. Because of this, large insertions split full-length matches between the seed HMM and the target sequence into multiple non-overlapping short matches. To correct this, our program merged these fragments into a larger single match. Then, we used a greedy approach to iteratively collect a list of the trusted matches for each target UniRef100 sequence. Initially, the trusted list was empty. We traversed the list of all the target sequence matches to the seed HMMs in ascending order of E-value. For every match, we checked if it had long intersections with matches that were already in the trusted list. If there were intersections longer than 10% of the shorter match, we discarded the currently considered match. Otherwise, we added the match to the list of trusted matches. Finally, relying on this list, we used the barrel sizes of the matching seeds to annotate the sequence.

Having collected and annotated the set of candidate sequences, we removed from our set sequences with fewer than two full matches, or with fewer than three partial matches; we then clustered the sequences using CLANS (48) (and an E-value of 10^{-20}). In cases where the cluster had sequences annotated with different sizes (e.g., a cluster with sequences annotated with two consecutive 8-stranded barrels, and sequences annotated with an 8-stranded barrel followed by a 10-stranded barrel), we further split the cluster, so that all sequences in each cluster would have the same barrel annotation. We used MAFFT v7.409 (49) to build an MSA for every cluster, and HMMER to expand these MSAs four times.

The search used 239 seed HMMs. In the initial search, 1,387,337 proteins with at least one partial match to the seed HMMs were found. After removing proteins with fewer than two full matches or three partial matches, 4,194 proteins which were likely multi-barrels were found.

The 4,194 proteins were divided to clusters, and the clusters were divided by architectures. For every architecture in every cluster we built an HMM and performed another HMMER search against UniRef100. After this final search, we had 9,062 annotated multi-barrel proteins.

PsiBlast-based sequence searches over NCBI's non-redundant protein sequences database (*nr*)

Independently, we used another approach to find proteins with multi-barrel matches that complements the method described below (Figure 6, upper-right).

1. Constructing the set of initial sequences

To collect a set of reliable initial OMBB protein sequences, global and local structural homologs of 6 OMBB families were collected using HHpred. Input sequences corresponded to *E. coli* BamA (UniprotKB P0A942), *E. coli* FepA (UniproKB P05825), *E. coli* OmpX (UniprotKB P0A917), *E. coli* Phospholipase A (UniprotKB P0A921), *K. pneumoniae* LptD (UniproKB C4T9I0), and mouse VDAC1 (UniprotKB Q60932). Searches were carried out over the PDB database filtered to a sequence identity of 70% as of January 2019, using default parameters. The resulting PDB entries matched were collected, the barrel identified, its sequence saved along with the number of strands, the length of internal and external loops collected and any N- and C-terminal decorations trimmed. In total, 157 unique OMBB sequences were collected, with barrel topologies ranging between 8 and 36 strands.

2. PsiBlast searches for bacterial proteins with multi-barrel matches

These OMBB sequences were then used as seeds for PsiBlast searches over the bacterial sequences in the NCBI's non-redundant sequence database, filtered to a sequence identity of 70% (nr_bac70). For all seeds, a maximum of 15 PsiBlast rounds were carried out, using those matches with an E-value better than 10^{-4} to generate the PSSM for the next round, and saving the hits with an E-value better than 10 after each round. The searches were carried out with PsiBlast 2.6.0+, allowing for the filtering out of low complexity matches with SEG (-seg option on). All matches with a minimum query coverage of 90% were then processed together, to find sequences with at least 2 non-overlapping matches to any of the queries.

Given that domain boundaries are always difficult to define with certainty at the sequence level, two matches were considered non-overlapping even if they overlapped by up to 10 residues. This search resulted in 482 unique EntrezIDs comprising multiple OMBB matches in a single chain. As we observed that matches to the large 36-stranded OMBB were generating ambiguous matches to families with a well-defined barrel topology (e.g., 22-stranded barrels of the FhuA family), these matches were excluded from further consideration.

3. Enriching the set of bacterial multi-barrel proteins

To enrich this set of proteins containing multi-barrel matches, we used the 482 unique, full-length, sequences as seeds for a BLAST search over the complete set of bacterial NCBI's non-redundant sequences (nr_bac). Searches were carried out as described above, collecting matches with an E-value better than 1 and a sequence coverage of the multi-barrel region higher than 90%. The resulting set of full-length sequences was then filtered to a sequence identity of 100%, eliminating any redundant match. A total of 6992 sequences were collected. Here, no preliminary topology annotation of putative multi-barrel sequences was carried out based on the queries as the deep searches preformed in step 2 may lead to the identification of larger barrel topologies when starting from a smaller number of strands (e.g., due to indels).

Classification and domain annotation

To perform a rough clustering of the sequences and visually inspect the clusters, we used CLANS (48). CLANS takes a list of protein sequences as an input, performs all-against-all BLAST comparisons between the sequences, and stores the p-value of the similarity of every similar sequence pair. To allow visual inspection of the sequence clusters, CLANS represents every sequence as a point in a two dimensional space and places the points in a way that minimizes the distance between points representing homologous sequences, while keeping points representing non-homologous sequences away from each other.

To allow a finer clustering of similar sequences, CLANS can use a p-value cutoff, and consider only distances between sequences whose homology has a p-value smaller than this cutoff.

The proteins collected using both the HMMER-based and a PsiBLAST-based approaches were combined, redundant sequences excluded with CD-HIT at a sequence identity of 100% and clustered with CLANS (48). Clusters were identified at a p-value of 1×10^{-25} and only those comprising at least 3 sequences were considered. The p-value was picked manually so that each MB-family will be clustered to a single cluster, but different MB-families may also be clustered together.

To assign to every protein its architecture, we split the clusters found by CLANS further. For every CLANS cluster, we calculated an MSA using MAFFT. Then, we split the proteins to sub-clusters using the alignment. Two proteins were determined to be in the same sub-cluster if they shared more than 70% of their positions in the MSA. This made sure that proteins with a completely different architecture would be in different sub-clusters, since the unmatched barrels would create a large number of unmatched positions in the MSA. This method ensures that the proteins in every cluster are aligned to each other well for further analysis.

After splitting to sub-clusters, we separated close architectures. Two architectures could be sent to the same sub-cluster if one was contained in the other – a protein with an 8-8 architecture can share most of its positions with an 8-8-8 architecture protein. To separate those, we manually inspected the distribution of the lengths of sequences in the clusters. When there were sequences whose lengths were at least 100 amino acids longer or shorter than the most common length, we found their architecture using HHpred and checked if it differed from the architecture of the other sequences in the cluster. This method identified one cluster which contained false positives, and five small groups of sequences which were in the same cluster with sequences with different architectures.

This method also over-splits clusters when there are large insertions, or non-barrel domains. This caused cluster 15, which contains a large single barrel and has both long insertions and domains other than the barrel domain, to be split to many small clusters. To separate the proteins in cluster 15 to subclusters properly, we used two steps: We first identified the barrel domains, and then we applied the previously described splitting method using only the barrel domain in every sequence.

To detect the barrel domains automatically in the sequence, we used HMMs. We manually identified the barrel domains in six sequences using HHpred. We used them to create six initial HMMs, and for each HMM, we searched for homologs among the sequences of cluster 15, which we then added to the HMM. We repeated this process four times, until the barrel domains were detected in all proteins of cluster 15. We then split the proteins in cluster 15 using the similarity of the barrel domains. The cluster was split to four sub-clusters this way.

From every sub-cluster, a representative protein was analyzed with HHpred, and the closest homologs were used to determine its architecture. When there were only partial overlapping matches to homologous proteins, we assumed the section was a single barrel, and determined its number of strands by HHpred's secondary structure prediction and by the matches to strands of known structures.

At the end of the procedure, we had 186 sub-clusters, to which we refer in the paper as MB-families.

Homologous structures and contact prediction

Our MSAs include multi-barrel proteins and other barrel architectures that extend the currently documented repertoire. We used sequence analysis tools to obtain further support for the correctness of our predicted architectures. First, we used HHpred (50) as an alternative homology-based tool to remove likely erroneous cases. To this end, we constructed and searched for HMMs (one round of PSI-BLAST, no secondary structure scoring) in the PDB, ECOD, and Pfam. This procedure identified five cases that were predicted by HHpred to have only a single barrel domain, and we removed them from our datasets.

To predict contact maps, we used four contact prediction programs: RaptorX (19), TripletRes (26), trRosetta (27), and DeepMetaPSICOV (28). We used the RaptorX webserver, and ran local versions of TripletRes, trRosetta, and DeepMetaPSICOV (downloaded from their respective GitHub repositories). Because contact map prediction relies on the coevolutionary signal, it requires many homologous proteins. Thus, we could use the server to predict the maps only for the 21 families containing more than 50 homologues. RaptorX limits the length of the MSAs to 1300, and the MSAs of our families were often longer. To overcome the length limitation, we arbitrarily picked a single protein as query, and removed all positions in the MSA in which the query protein had a gap. Finally, we inspected each of the predicted contact maps to deduce a structural characterization of the multi-barrels. In many cases the β-barrel signal in the contact-map was missing, blurry, or riddled with too many false positives to use. We obtained a clear multi-barrel signal for four cases out of 21. Due to technical limitations, we were able to run

TripletRes only for MSAs with length no longer than 600. To predict contact maps for longer sequences, we divided the contact map to squares, representing the contacts between groups of 300 amino acids, and predicted each square individually. This produced visible artifacts in the predicted contact maps. In seven of twenty-one cases, DeepMetaPSICOV failed to return any output.

To compare the strengths of the contact signals of beta strands in the predicted contact maps, we manually annotated the strand positions in each contact map and measured for each pair of strands the probability assigned to the contact between them. We divided the contacts between strands to three groups: contacts supporting the multi-barrel hypothesis, contacts supporting the single-barrel hypothesis, and contacts between strands in the same barrel disagree with both hypotheses. We then plotted the histogram of the strengths of contacts in the third group as a null distribution, and plotted the contacts supporting one of the hypotheses on the histogram.

Structural modelling the PLA1-PLA1 (12-12) MB-family

To model the structure of the PLA1-PLA1 (12-12) double-barrel, we used HHpred to find structural templates for both OMPLA domains, and picked 1QD6, which is a structure of an OMPLA homo-dimer bound to a hexadecanesulfonyl fluoride inhibitor. We cut the protein to two parts according to HHpred and removed the linker section, which was unaligned to the template.

To align the double-barrel sequences to 1QD6, we built an HMM for the two domains of one of the double-barrel proteins using HMMER. We searched for homologs in UniRef100 using an E-value of 10^{-10}. We used MAFFT to create an MSA which contained the sequences of the two domains of the double-barrel, the sequence of the template, and sequences of other homologs of the 12-stranded barrel. We deleted all sequences in the MSA beyond the sequences of query and template and removed empty columns in the alignment. The result was an alignment of the domains of the double barrel to the template sequence, which used the information in other homologs of the OMPLA barrel to improve the accuracy.

We used this alignment to create a PIR file for MODELLER 9.23, and generated five models. We aligned the models we generated to the model of 1QD6, and copied the inhibitor from 1QD6 to the models we generated. We picked the model manually, attempting to minimize the steric clashes with the inhibitor.

We calculated conservation scores for the fused protein and for the template using ConSurf with the default settings. The N-terminal barrel domain of the 12-12 protein had 1,830 homologs, 150 of which were used. Since the PLA1-PLA1 (12-12) query has only six homologs, ConSurf calculated the conservation scores using homologs of the monomer and provides only an illustration of the conservation scores of the monomer and a sanity check for the predicted model.

Classification of individual OMBBs from selected families and analysis of the genomic context

To assess the likelihood of the OMBBs in the PLA1-PLA1 MB-family to have emerged by the tandem repeat of an ancestral OMBB, we further classified the individual barrel domains with their single-barrel homologs in close species. For that, the individual barrel domains in the reference sequences selected for the respective family were used as seeds for a BLAST search over the nr_bac database, and only hits covering a minimum of 90% of the query collected. For the PLA1 proteins, given their high sequence identity to the single-barrel PLA1, only hits with an E-value better than 1×10^{-40} were collected in order to reduce the number of sequences to cluster. The resulting sets of individual barrel proteins were clustered with CLANS (48) with a p-value threshold of 1×10^{-80}.

To assess the conservation of the genomic context of the PagP-LptD, PLA1-PLA1, YjbH-GfcD MB-families, the genomic context of each protein in the MB-family (or a set of representative cases) was carried out by identifying the products of the *n* flanking genes in their full genomic assembly and their further comparison using BLAST. For each case, clusters of homologous flanking genes were identified by all-against-all BLAST searches of their protein products at a maximum E-value of 1×10^{-3}. Protein products from genes conserved throughout the genomic contexts analyzed were further annotated for their domain composition, secondary structure and sequence features as described above.

In the BLAST search of the PagP-LptD family, 17 additional proteins belonging to the MB-family were found.

Calculating the taxonomic tree of the species containing multi-barrels

To display the taxonomic classification of the bacteria containing the multi-barrel proteins, we constructed a tree using NCBI's taxonomy database (51), and rendered it using Dendroscope (1).

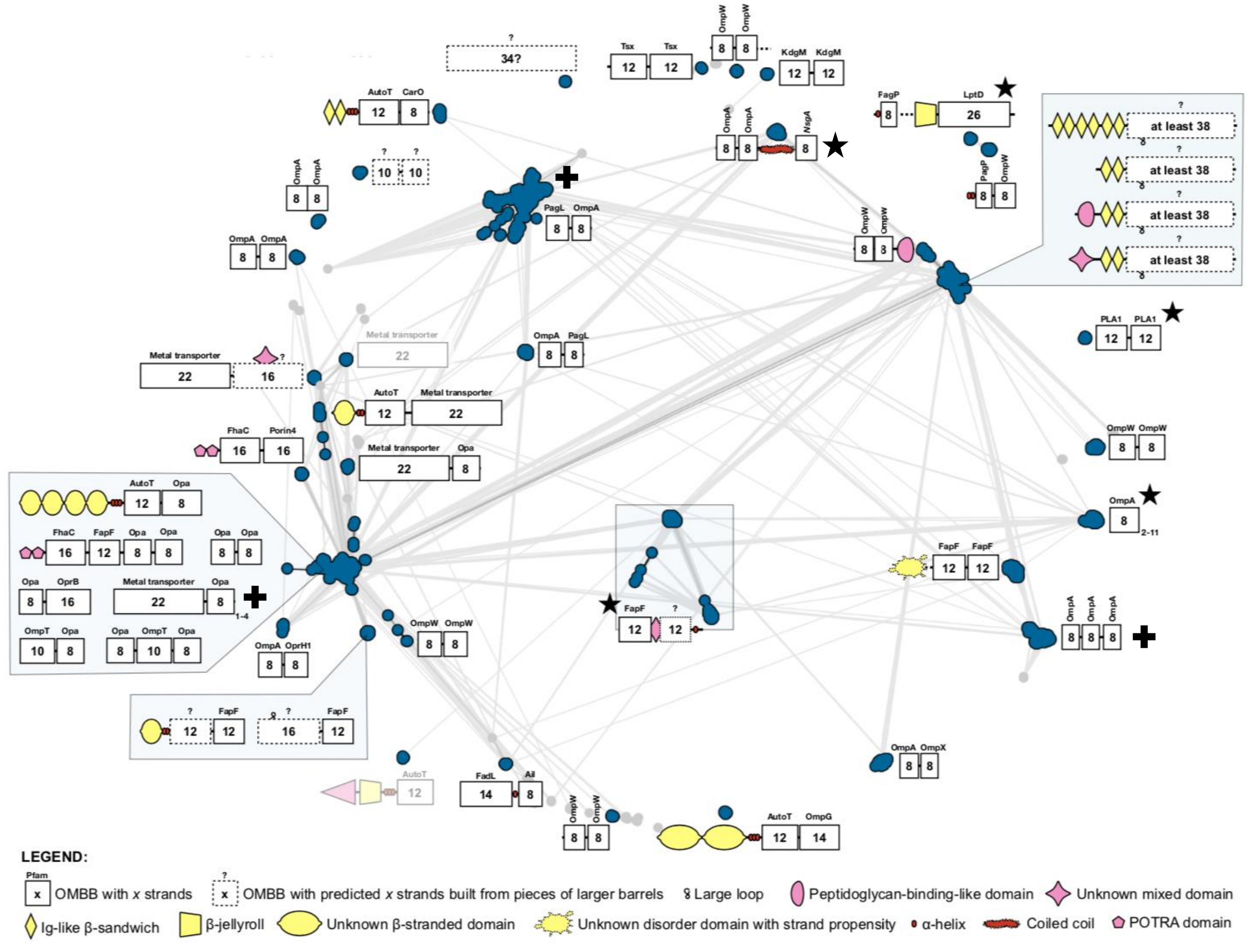


Figure S1. Classification and domain annotation of 12,643 OMBB proteins with new architectures. Clustering was carried out with CLANS (1) at a p-value of 1×10^{-25} and shown at 1×10^{-15}. For each cluster, sequences were binned by their size, with a step of 100 residues, and a representative from each bin collected. Each of these representatives was manually annotated for their predicted domain composition with HHpred (2) over Pfam (3), ECOD (4, 5) and PDB70 (6), for their secondary structure content with Quick2D (2), and for their transmembrane topology with BOCTOPUS (7). Coiled coil domains were identified with PCOILS (8) and signal peptides with SignalP 5.0 (9). Shaded architectures represent clusters that are composed of false-positive sequences carrying only one barrel belonging to a known OMBB family. Clusters with 3 or less sequences were not considered and are colored light grey. Black crosses mark four architectures with clear contact maps. Black stars mark the five families/MB-superfamilies selected for detailed discussion: The 12-12 OMPLA family (right edge), the 8-26 PagP-LptD family (top-right), the poly-8 MB-superfamily (bottom-right, marked with a single box), the 8-8-Linker-8 MB-superfamily (center-top), and the 12-12 YjbHGfcD family (center).

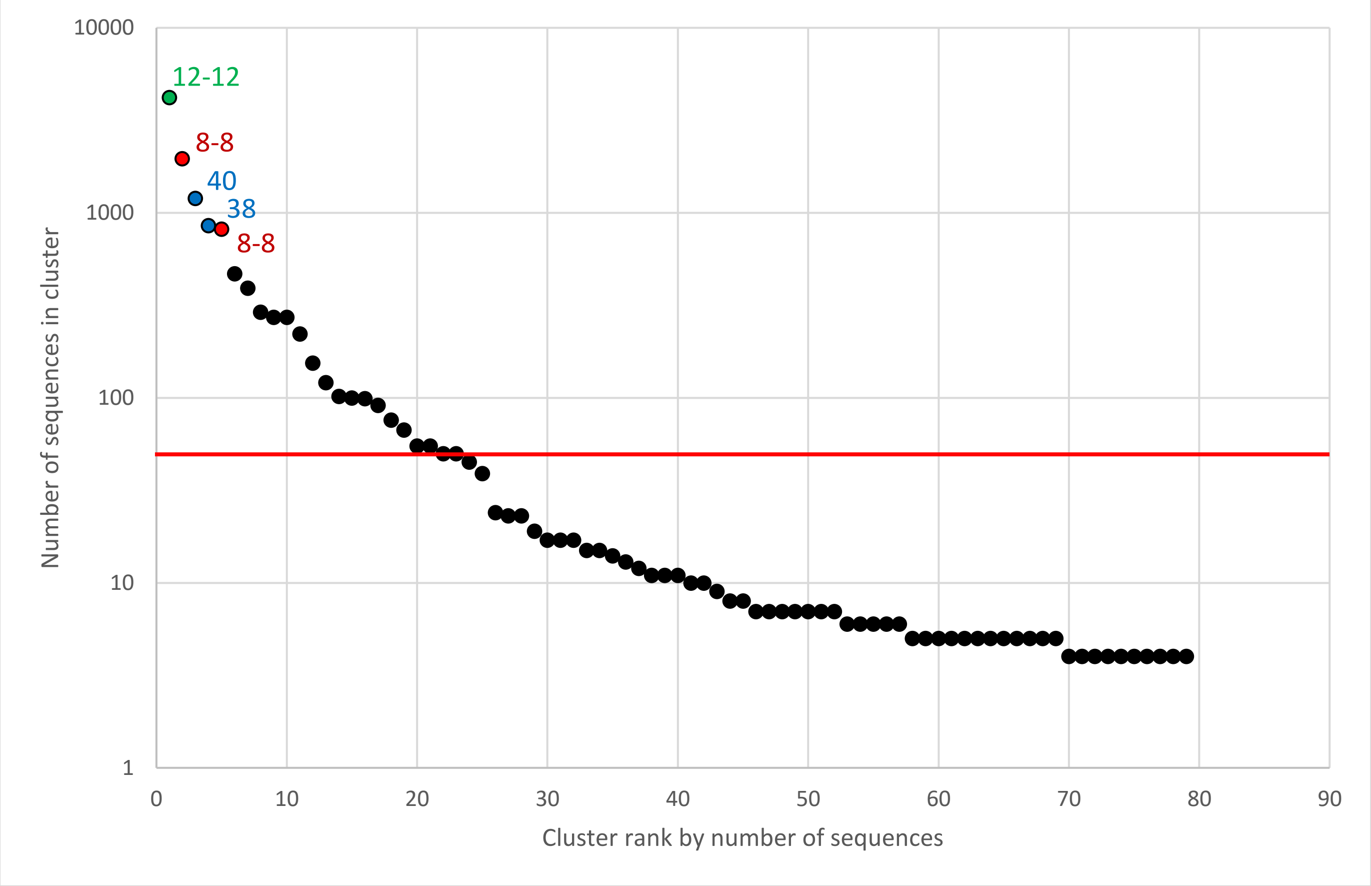


Figure S2. Families of barrels with new architectures (multi-barrels and large barrels) with at least four proteins, sorted by decreasing size, i.e., number of proteins per family, displayed in log scale. The points representing the five largest families are labeled and colored according to their architecture – the 12-12 architecture in green, the 8-8 architecture in red, and the 40 and 38 architectures in blue. We see that quite a few MB-architectures are populated by many homologous proteins. For the 21 families with more than 50 homologous proteins that fall above the marked red line, we used RaptorX, TripletRes, trRosetta, and DeepMetaPSICOV to predict their contact maps.

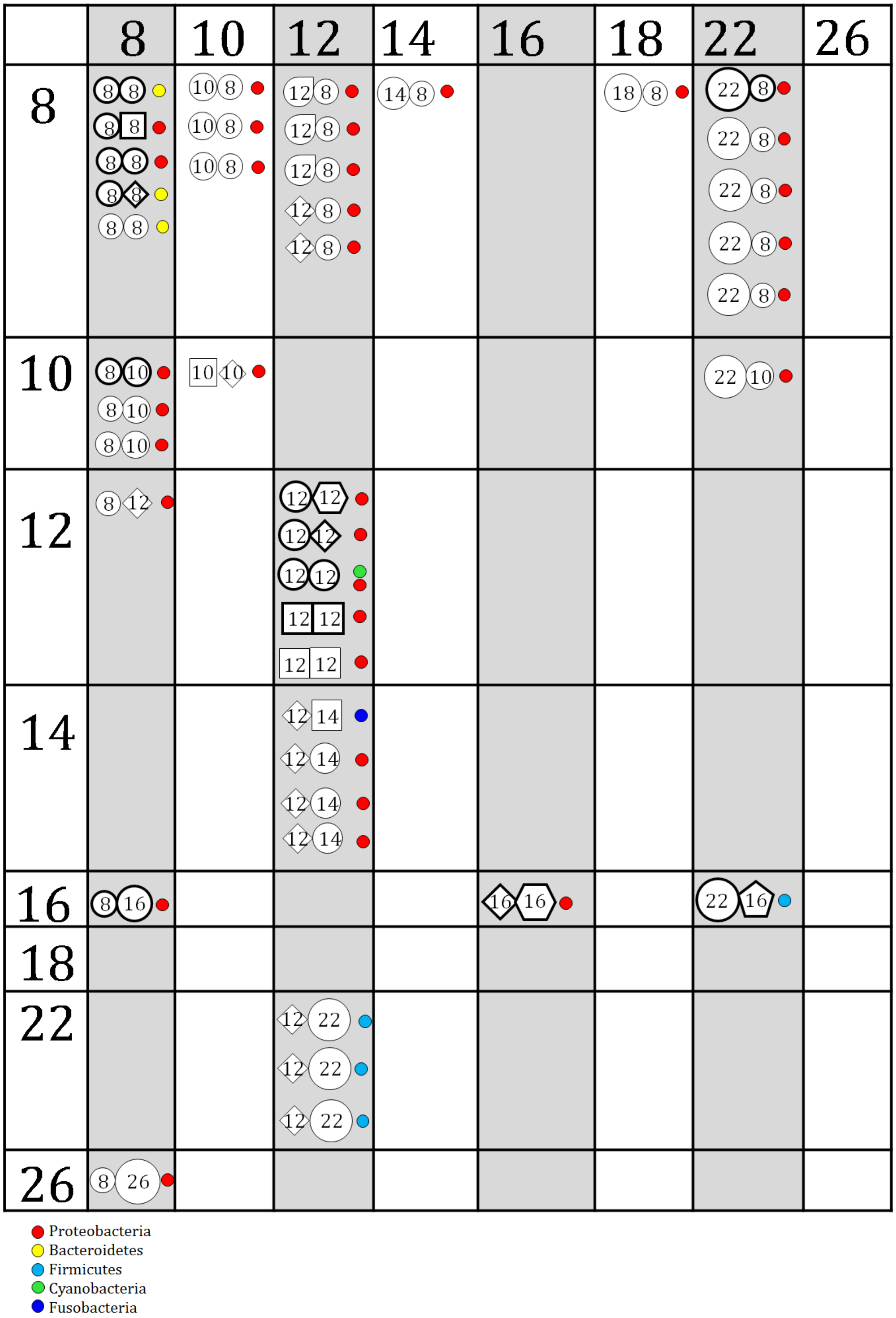


Figure S3. All double-barrel architectures. The largest MB-families (five at most) within each architecture are shown. MB-families with 50 homologous proteins or more are represented with a bold outline. The colored dot(s) show the taxonomy of the bacteria with the architecture. The topmost 12-12 architecture and the 8-26 architecture are discussed in the main text.

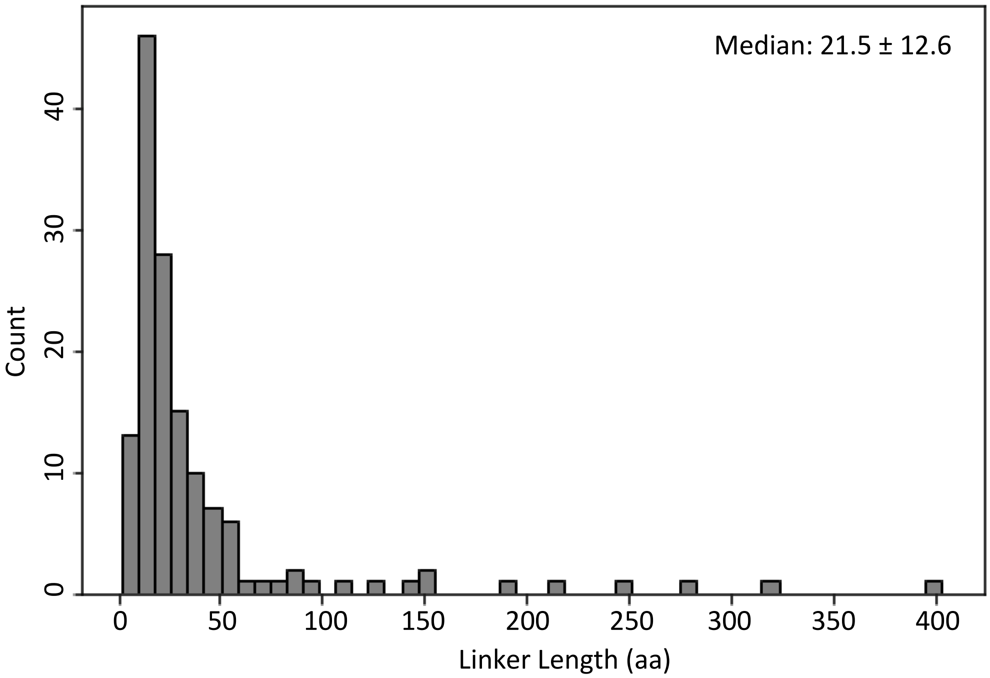


Figure S4. Histogram of lengths for the linkers between the barrels in the MB-families of Figure 1.

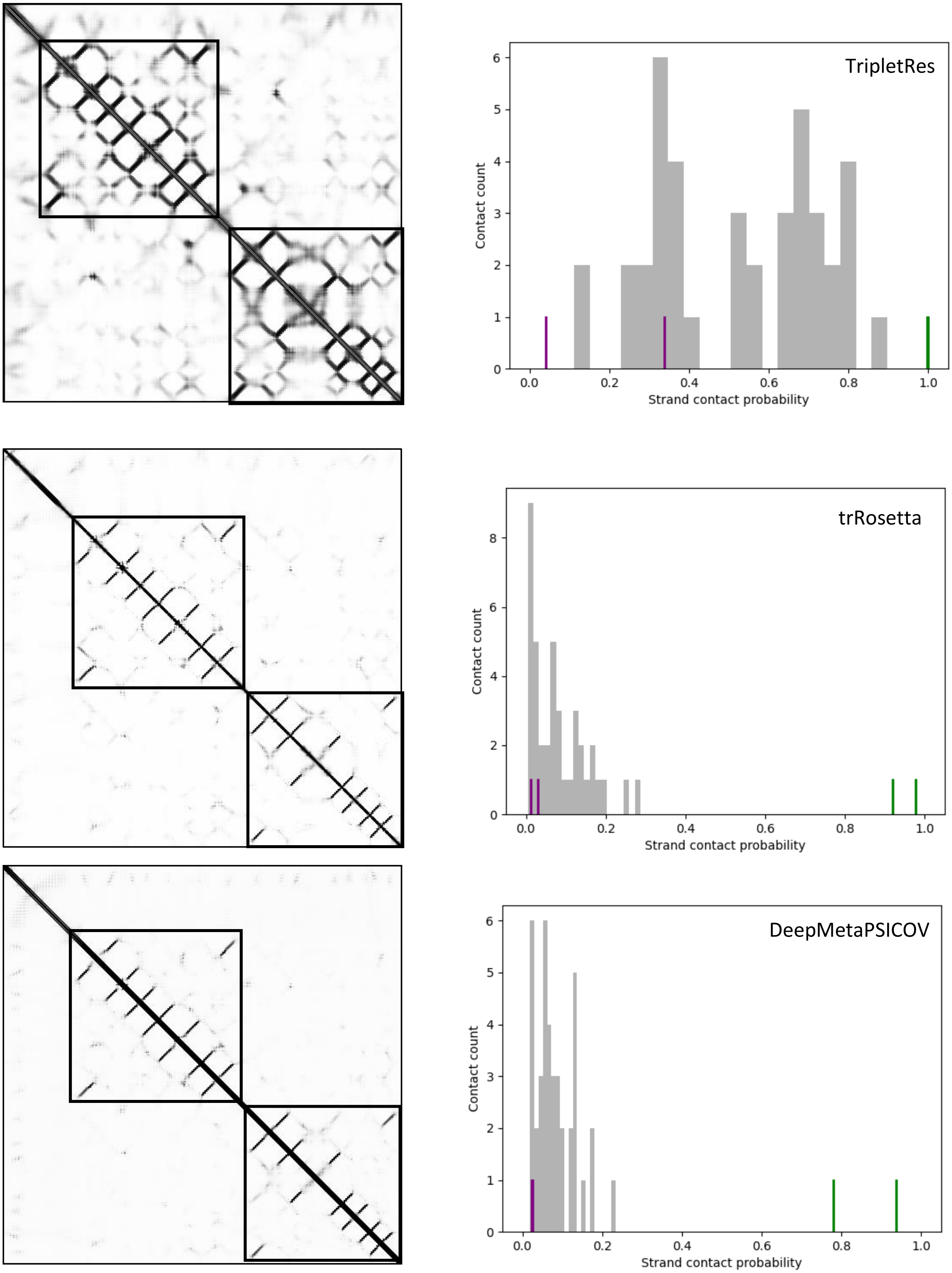


Figure S5. The contact maps and the distribution of relevant contact strengths for family 001 (of 1,591 homologs)

predicted to form an 8-8 architecture, by TripletRes, trRosetta, and DeepMetaPSICOV. Clear contact signals are observed for a total of 16 beta-strands in all three contact maps. Reassuringly, a clear ‘barrel-closing signal’, i.e., the contact between the first and last strands of the barrel, is observed for both predicted barrel domains. The two putative barrel domains are marked with black squares. The probabilities for the closing contacts supporting an 8-8 architecture are marked in the histograms in green, and the probabilities for the closing contacts supporting an 16 stranded-barrel architecture are marked in purple. Probabilities that are expected to be intra-barrel contacts are marked in grey and used as a null distribution. The probabilities supporting an 8-8 architecture are higher than the null distribution in all three contact maps (and in the RaptorX prediction shown in the main text), and the probabilities describing the contacts among residues that would support an 16 stranded-barrel architecture are in the low values of the null distribution (i.e., describing residues that are not in contact). In summary, contact prediction by all 4 predictors strongly support an 8-8 architecture.

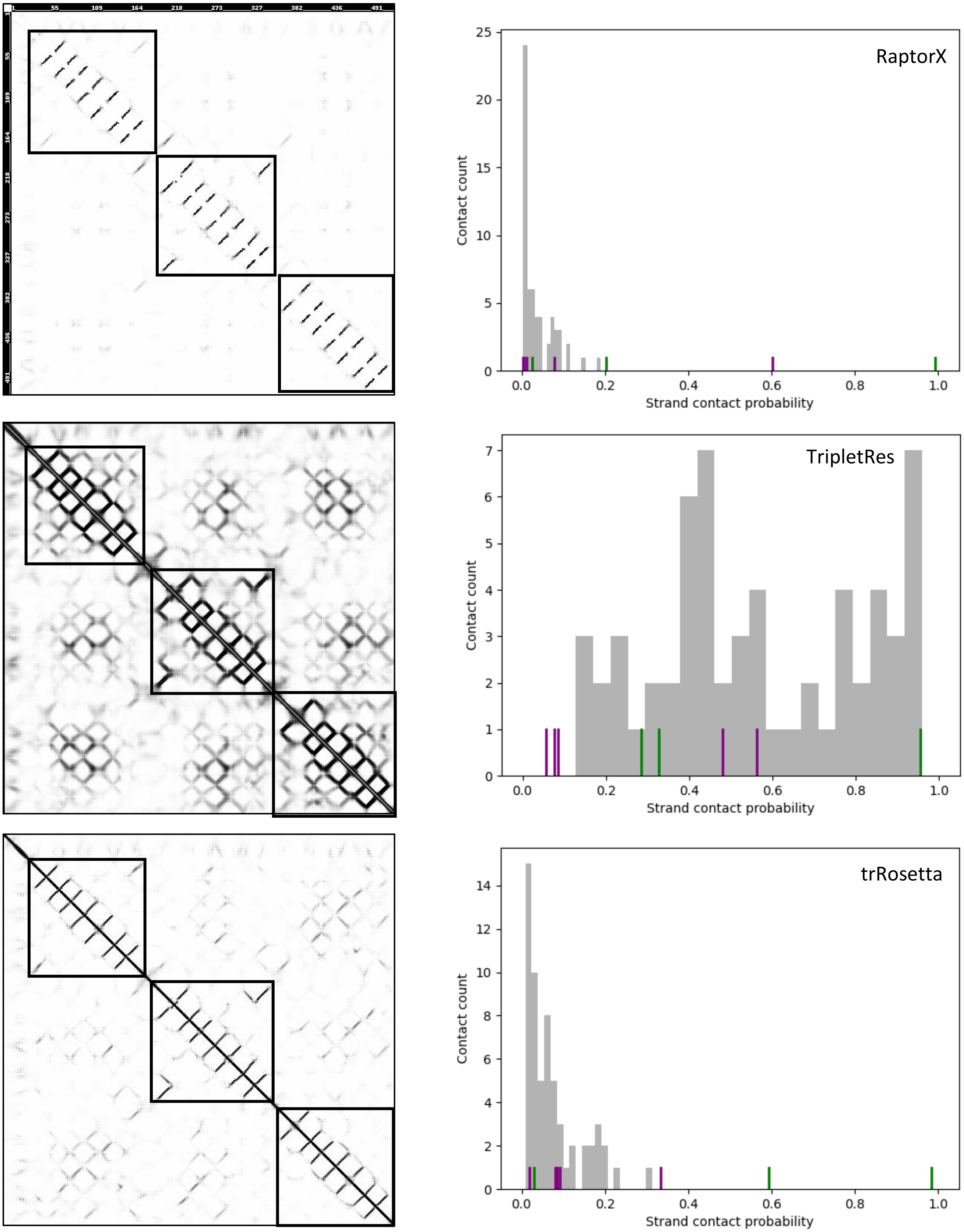

1
55
109
164
218
273
327
382
436
491
RaptorX
Contact count
Strand contact probability
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
TripletRes
Contact count
Strand contact probability
0
1
2
3
4
5
6
7
trRosetta
Contact count
Strand contact probability
0
2
4
6
8
10
12
14

Figure S6. The contact maps and the distribution of relevant contact strengths for family 007 (of 291 homologs), predicted to form an 8-8-8 architecture, by RaptorX, TripletRes, and trRosetta. Clear contact signals are observed for a total of 24 beta-strands in all three contact maps. Reassuringly, a clear 'barrel-closing signal', i.e., the contact between the first and last strands of the barrel, is observed for the middle barrel in all three contact maps, and a weak barrel-closing signal for the first barrel in the contact maps predicted by RaptorX and trRosetta. There is no barrel-closing signal for the C-terminal barrel in all three contact maps. The three putative barrel domains are marked with black squares. The probabilities for the closing contacts supporting an 8-8-8 architecture are marked in the histograms in green, and the probabilities for the closing contacts supporting a 24 stranded-barrel architecture are marked in purple. Probabilities that are expected to be intra-barrel contacts are marked in grey and used as a null distribution. In all three figures, the probabilities of contacts between residues that close the central barrel are the highest. In the contact map predicted by RaptorX, the contact between strands 8 and 9 is also significant. In the contact map predicted by TripletRes, the predicted probabilities for contacts between residues in strands 1 and 16, 9 and 24, and 1 and 24, are significantly below the null distribution. In the contact map predicted by trRosetta, the probabilities for contact closing the N-terminal barrel are also significant. Taken together, the three contact maps support an 8-8-8 architecture, although the support is not as strong as the support for MB-families 001 and 012.

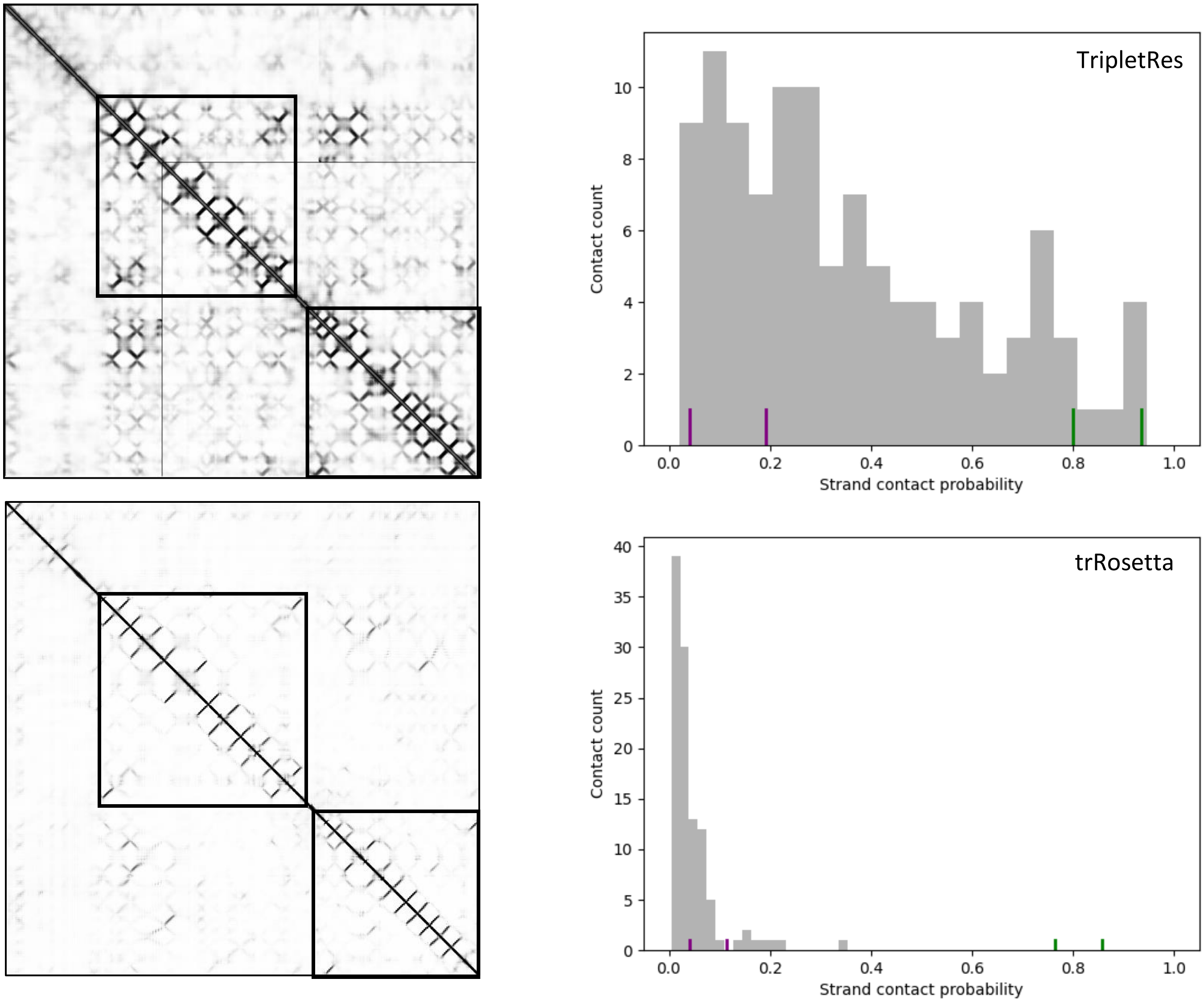
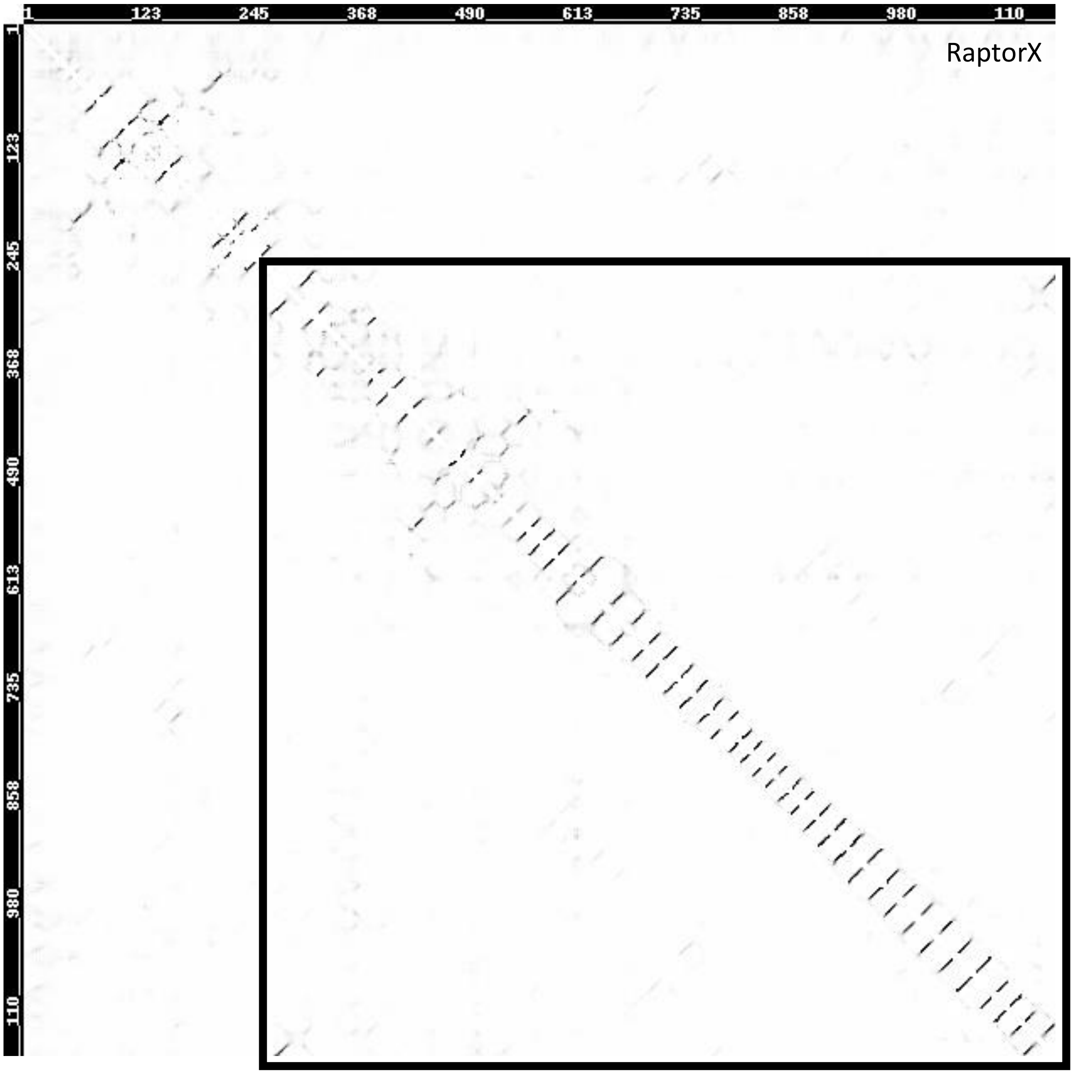


Figure S7. The contact maps and the distribution of relevant contact strengths for family 012 (of 121 homologs), predicted to form a 12-12 architecture, by TripletRes and trRosetta. (Left panels) Clear contact signals are observed for a total of 24 beta-strands in both contact maps. Also, a clear 'barrel-closing signals', i.e., the contact between the first and last strands of the barrel, are observed for both barrels. Furthermore, a barrel-closing signal between the stands that would indicate a single 24-stranded barrel is missing, making this alternative architecture much less likely. The two putative barrel domains are marked with black squares. (Right panels) The probabilities assigned for contacts between beta strands in MB-family 012, which is predicted to form a 12-12 architecture, by the two contact map predictors. The probabilities for the closing contacts supporting a 12-12 architecture are marked in the histograms in green, and the probabilities for the closing contacts supporting a 24 stranded-barrel architecture are marked in purple. Probabilities that are expected to be intra-barrel contacts are marked in grey and used as a null distribution. The closing signals matching the 12-12 architecture are higher than the null distribution or at a high percentile, while the signals matching a 24 stranded-barrel architecture are explained by the null distribution.


1
123
245
368
490
613
735
858
980
110
RaptorX

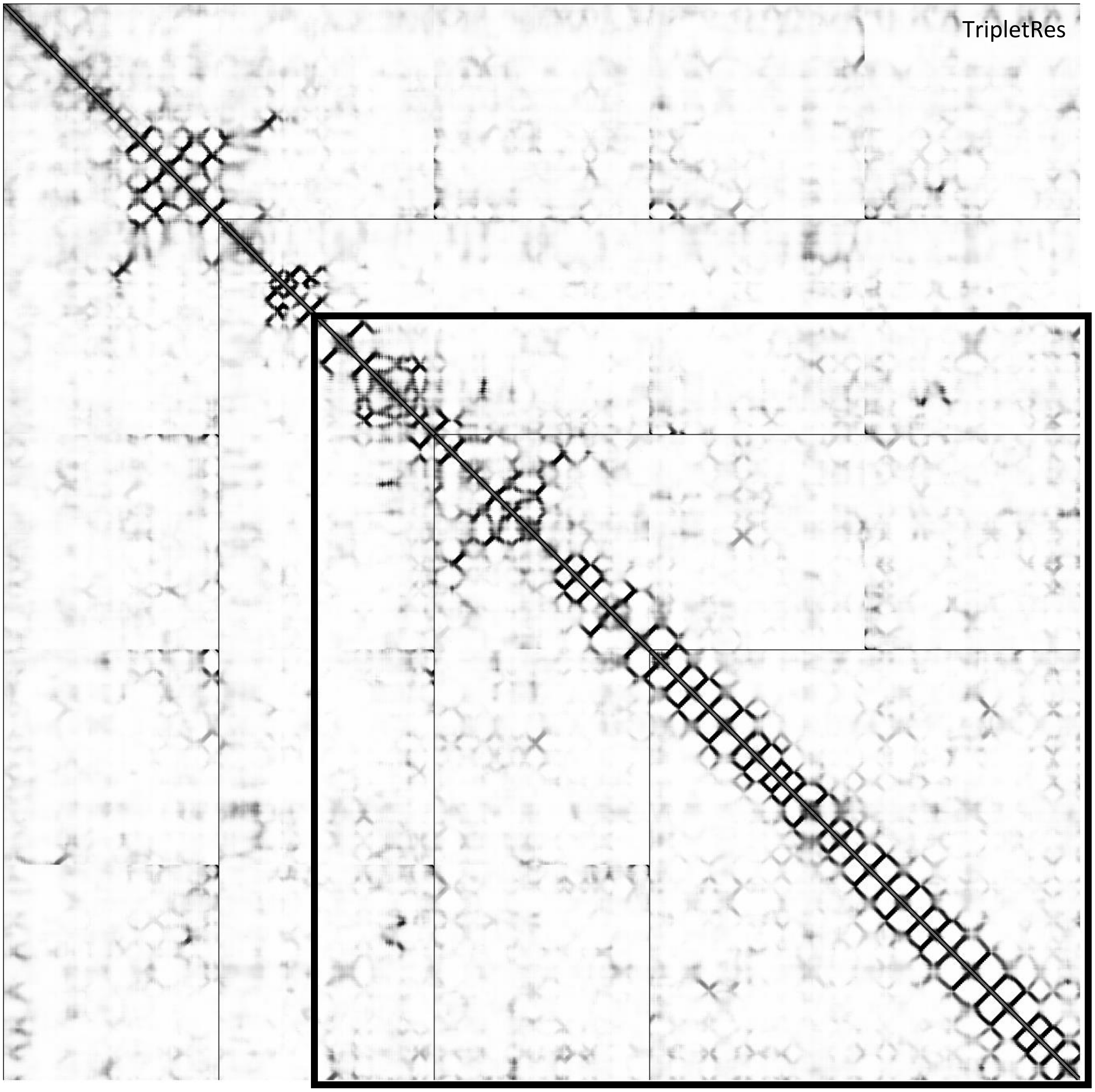


Figure S8. The contact maps predicted by RaptorX (Top) and TripletRes (Bottom) for family 002, with 1,194 homologous proteins, predicted to form a single large barrel. Clear contact signals between adjacent strands are observed for a total of 40 beta-strands in both contact maps. Furthermore, clear 'barrel-closing signals' between the first and last strands of the barrel are also visible. The putative barrel is marked with a black square.

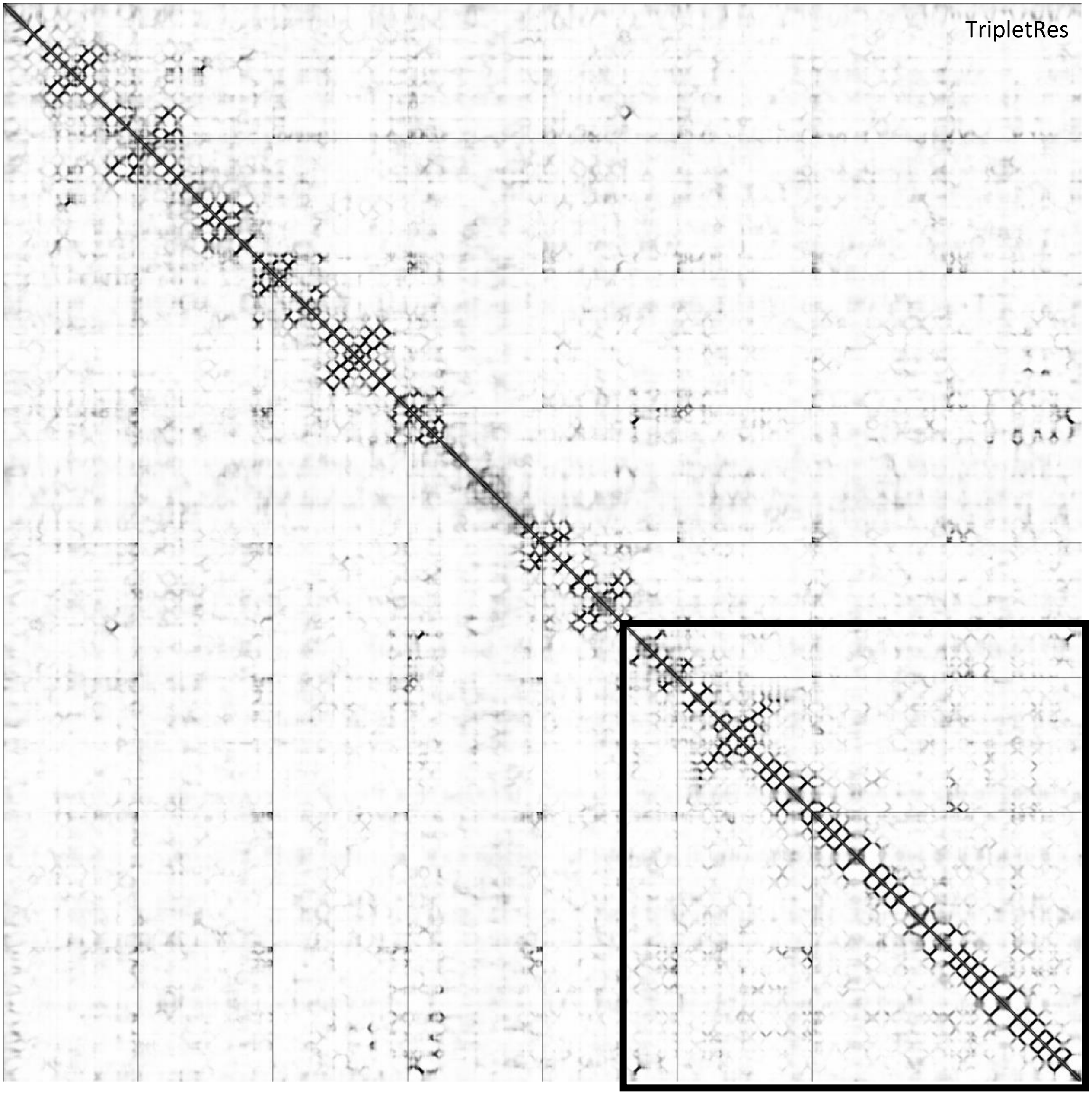


Figure S9. The contact map predicted by TripletRes for family 003, with 849 homologous proteins, predicted to form a single large barrel. Clear contact signals between adjacent strands are observed for a total of 38 beta-strands, although the contact between strands 22 and 23 is missing. Furthermore, a clear 'barrel-closing signal' between the first and last strands of the barrel is also visible. The putative barrel is marked with a black square.

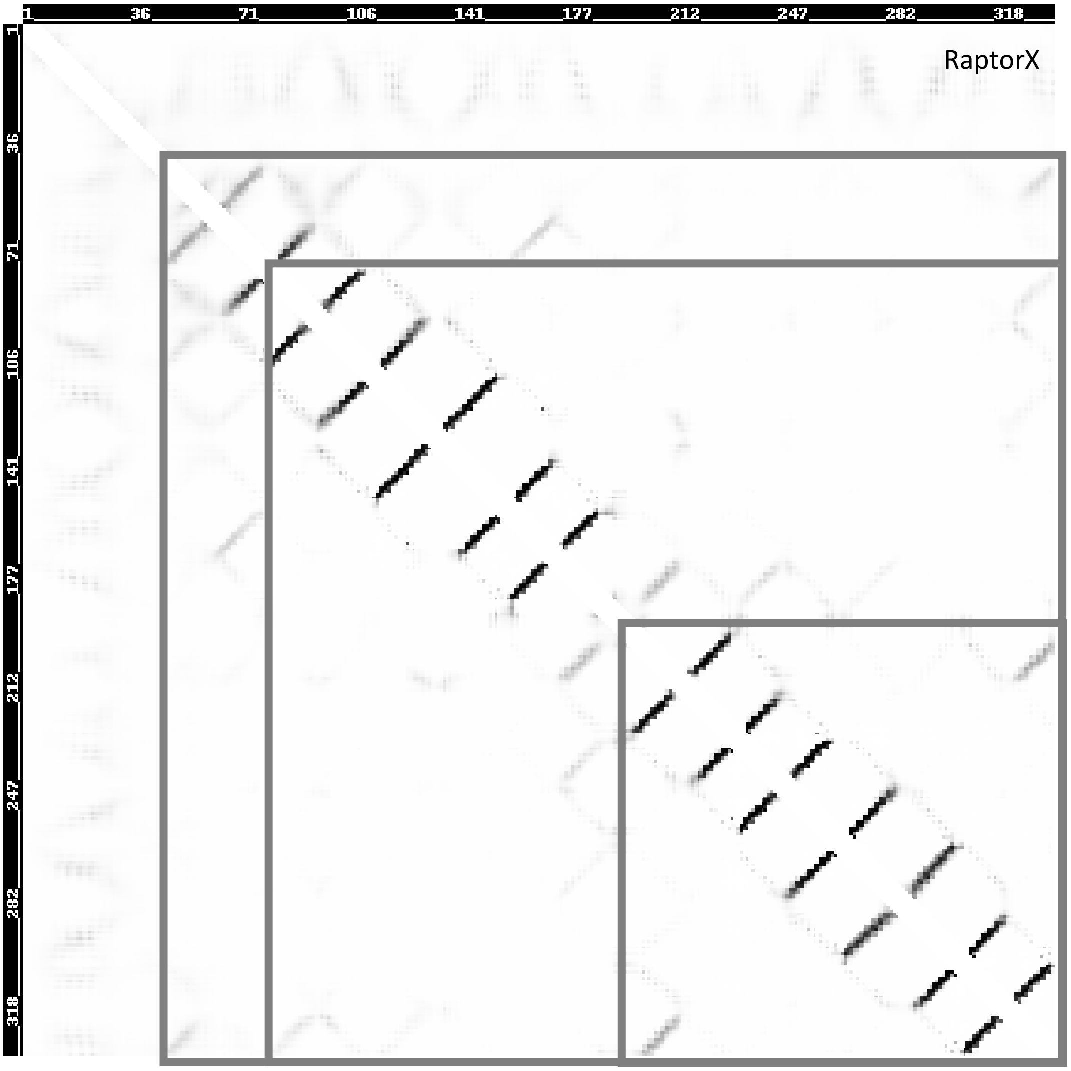


Figure S10. Example of an uninterpretable contact map predicted by RaptorX: family 004, with 813 homologous proteins, predicted to form an 8-8 architecture. The signals of contacts between adjacent strands within both putative barrels are clear. There is also a barrel closing signal for the second domain, in supports of the predicted 8-8 architecture. However, there is also a weak contact signal between the last strand of the first putative barrel domain and the first strand of the second putative barrel domain, suggesting a single barrel architecture. Furthermore, there are weak closing signals for hypothetical single barrel architectures with 14 and 16- strands, as well as a weak closing signal between the second strand of the first barrel and the seventh strand of the first barrel, which is unlikely. Overall, the contact map does not provide unequivocal evidence in support of a single architecture.

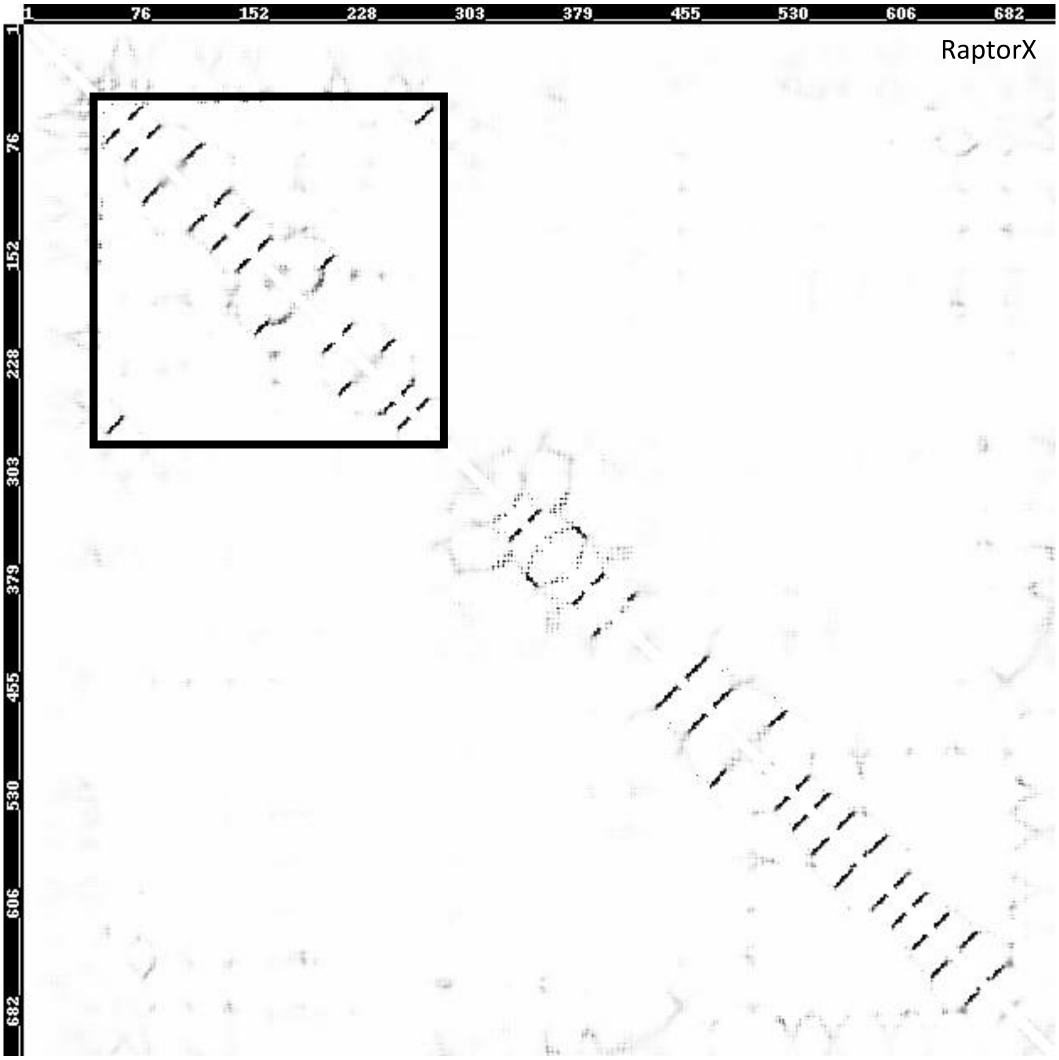


Figure S11. Example of a contact map that does not support a multi-barrel architecture. The homologies for family 000, with 4,187 homologous proteins, suggest these form a 12-12 architecture. Reassuringly, the contact map predictions (here, RaptorX is shown) has footprints of a 12-stranded barrel, including the barrel closing signal are clearly available in the N-terminal. However, the probabilities for contact closing the C-terminal barrel domain are very low. Rather, high probability contacts are predicted between 12 adjacent strands, suggesting a 13 stranded barrel.

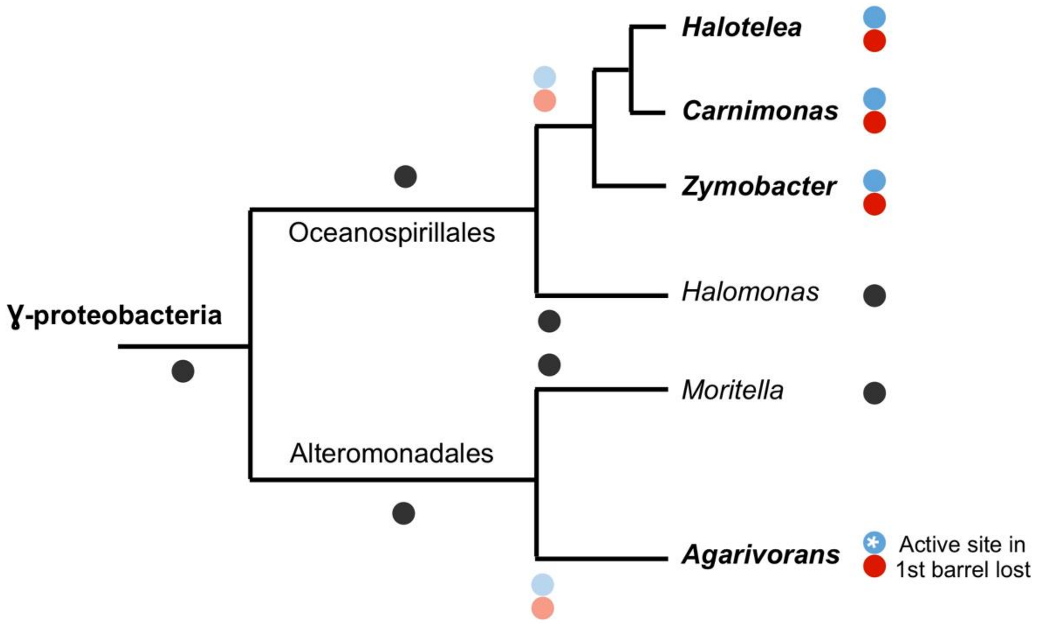


Figure S12. An evolutionary scenario for the independent, parallel emergence of PLA1-PLA1 proteins in the two *Ocenospirillales* and *Agarivorans* lineages. The dots mark the barrels domains – a single dot marks a protein with a single beta barrel, and two dots mark a double-barrel protein, with blue for the N-terminal barrel and red for the C-terminal barrel.

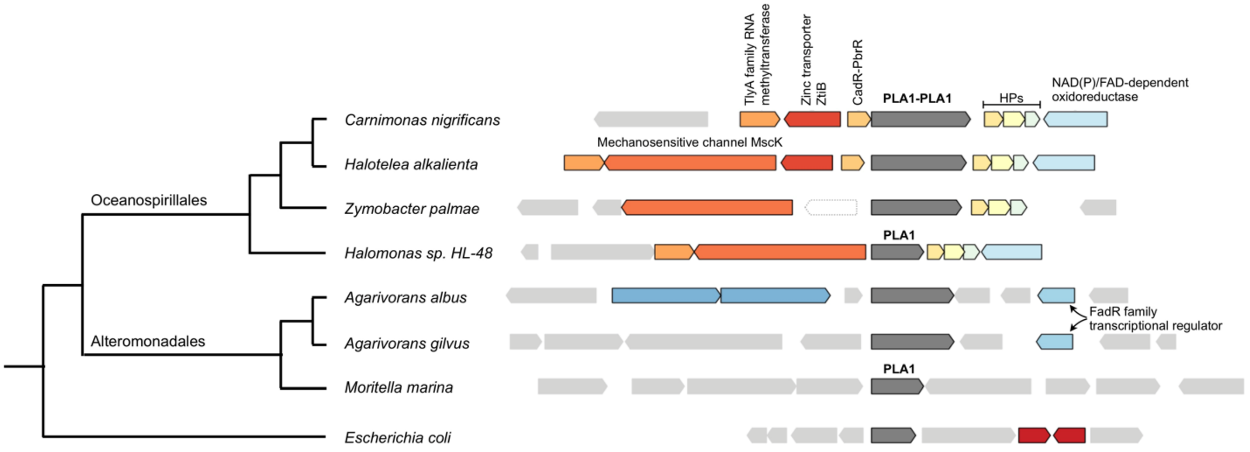


Figure S13. The genomic context of PLA1-PLA1 proteins in comparison to that of their close homologs in two related species (Halomonas and, Mortella marina) and *E. coli.* Homologous genes are colored the same, light grey boxes represent the non-conserved genes, and white boxes the pseudogenes. HP stands for ‘hypothetical protein’.

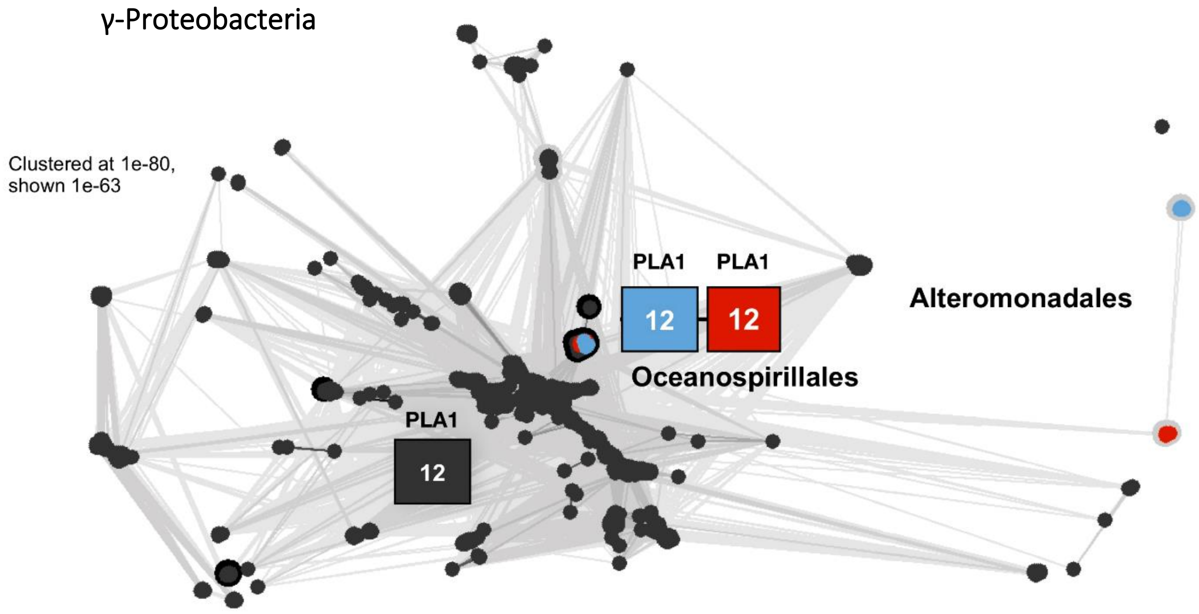
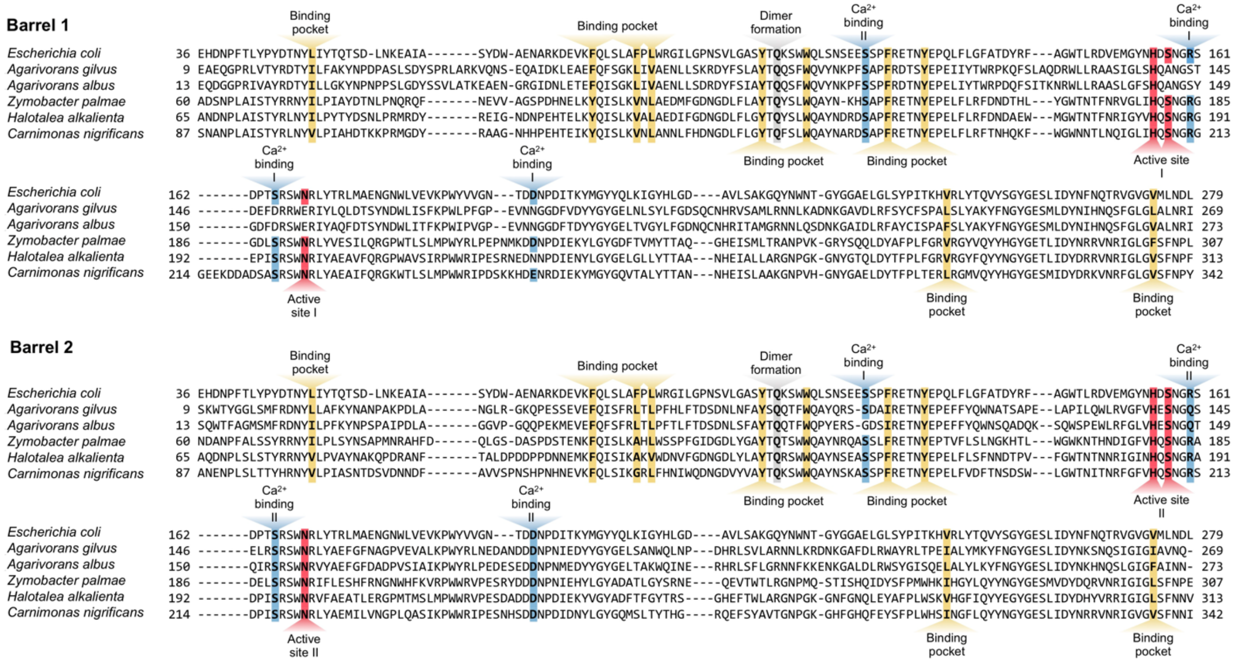


Figure S14. Classification of the PLA1 domains in the six PLA1-PLA1 proteins with their close single barrel homologs in γ-proteobacteria. The dots represent PLA1 barrel domains, and the boxes show the architecture. Black dots represent barrel domains from single-barrel proteins with an architecture of 12, shown in the black box. The blue dots represent the N-terminal barrel domains from the PLA1-PLA1 proteins, and the red dots represent the C-terminal barrel domains from the PLA1-PLA1 proteins. The connected blue and red boxes in the center of the figure represent the 12-12 architecture, whose barrel domains are the red and blue dots near it. The two barrel domains are very similar to each other, and are clustered very close to each other. The barrel domains in *Alteromonadales* are less similar to each other, and are far from each other. Edges connect sequences whose similarity has an E-value of at least 1×10^{-63}. Clustering was carried out with CLANS (1) at a p-value of 1×10^{-80} and is shown at 1×10^{-63}.



```
Barrel 1
Escherichia coli         36 EHDNPFTLYPYDTNYLIYTQTSD-LNKEAIA-------SYDW-AENARKDEVKFQLSLAFPLWRGILGPNSVLGASYTQKSWWQLSNSEESSPFRETNYEPQLFLGFATDYRF---AGWTLRDVEMGYNHDSNGRS 161
Agarivorans gilvus        9 EAEQGPRLVTYRDTYILFAKYNPDPASLSDYSPRLARKVQNS-EQAIDKLEAEFQFSGKLIVAENLLSKRDYFSLAYTQQSFWQVYNKPFSAPFRDTSYEPEIIYTWRPKQFSLAQDRWLLRAASIGLSHQANGST 145
Agarivorans albus        13 EQDGGPRIVAYRDTYILLGKYNPNPPSLGDYSSVLATKEAEN-GRGIDNLETEFQISGKLIVAENLLSDRDYFSIAYTQQSFWQVYNKPFSSPFRDTNYEPELIYTWRPDQFSITKNRWLLRAASLGFSHQANGSY 149
Zymobacter palmae        60 ADSNPLAISTYRRNYILPIAYDTNLPNQRQF-------NEVV-AGSPDHNELKYQISLKVNLAEDMFGDNGDLFLAYTQYSLWQAYN-KHSAPFRETNYEPELFLRFDNDTHL---YGWTNTFNRVGLIHQSNGRG 185
Halotalea alkalienta     65 ANDNPLAISTYRLNYILPYTYDSNLPRMRDY-------REIG-NDNPEHTELKYQISLKVALAEDIFGDNGDLFLGYTQYSLWQAYNDRDSAPFRETNYEPELFLRFDNDAEW---MGWTNTFNRIGYVHQSNGRG 191
Carnimonas nigrificans   87 SNANPLAISTYRLNYVLPIAHDTKKPRMGDY-------RAAG-NHHPEHTEIKYQISLKVNLANNLFHDNGDLFLGYTQFSLWQAYNARDSAPFRETNYEPELFLRFTNHQKF---WGWNNTLNQIGLIHQSNGRG 213

Escherichia coli        162 -------DPTSRSWNRLYTRLMAENGNWLVEVKPWYVVGN---TDDNPDITKYMGYYQLKIGYHLGD----AVLSAKGQYNWNT-GYGGAELGLSYPITKHVRLYTQVYSGYGESLIDYNFNQTRVGVGVMLNDL 279
Agarivorans gilvus      146 -------DEFDRRWERIYLQLDTSYNDWLISFKPWLPFGP--EVNNGGDFVDYYGYGELNLSYLFGDSQCNHRVSAMLRNNLKADNKGAVDLRFSYCFSPALSLYAKYFNGYGESMLDYNIHNQSFGLGLALNRI 269
Agarivorans albus       150 -------GDFDRSWERIYAQFDTSYNDWLITFKPWIPVGP--EVNNGGDFTDYYGYGELTVGYLFGDNQCNHRITAMGRNNLQSDNKGAIDLRFAYCISPAFSLYAKYFNGYGESMLDYNIHNQSFGLGVALNRI 273
Zymobacter palmae       186 -------GDLSRSWNRLYVESILQRGPWTLSLMPWYRLPEPNMKDDNPDIEKYLGYGDFTVMYTTAQ---GHEISMLTRANPVK-GRYSQQLDYAFPLFGRVRGYVQYYHGYGETLIDYNRRVNRIGLGFSFNPL 307
Halotalea alkalienta    192 -------EPISRSWNRIYAEAVFQRGPWAVSIRPWWRIPESRNEDNNPDIENYLGYGELGLLYTTAA---NHEIALLARGNPGK-GNYGTQLDYTFPLFGRVRGYFQYYNGYGETLIDYDRRVNRIGLGVSFNPF 313
Carnimonas nigrificans  214 GEEKDDADSASRSWNRLYAEAIFQRGKWTLSLMPWWRIPDSKKHDENRDIEKYMGYGQVTALYTTAN---NHEISLAAKGNPVH-GNYGAELDYTFPLTERLRGMVQYYHGYGESMIDYDRKVNRFGLGVSFNPY 342

Barrel 2
Escherichia coli         36 EHDNPFTLYPYDTNYLIYTQTSD-LNKEAIA-------SYDW-AENARKDEVKFQLSLAFPLWRGILGPNSVLGASYTQKSWWQLSNSEESSPFRETNYEPQLFLGFATDYRF---AGWTLRDVEMGYNHDSNGRS 161
Agarivorans gilvus        9 SKWTYGGLSMFRDNYLLAFKYNANPAKPDLA-------NGLR-GKQPESSEVEFQISFRLTLPFHLFTDSDNLNFAYSQQTFWQAYQRS-SDAIRETNYEPEFFYQWNATSAPE--LAPILQWLRVGFVHESNGQS 145
Agarivorans albus        13 SQWTFAGMSMFRDNYILPFKYNPSPAIPDLA-------GGVP-GQQPEKMEVEFQFSFRLTLPFTLFTDSDNLSFAYTQQTFWQPYERS-GDSIRETNYEPEFFYQWNSQADQK--SQWSPEWLRFGLVHESNGQT 149
Zymobacter palmae        60 NDANPFALSSYRRNYILPLSYNSAPMNRAHFD------QLGS-DASPDSTENKFQISLKAHLWSSPFGIDGDLYGAYTQTSWWQAYNRQASSLFRETNYEPTVFLSLNGKHTL---WGWKNTHNDIGFVHQSNGRA 185
Halotalea alkalienta     65 AQDNPLSLSTYRRNYVLPVAYNAKQPDRANF-------TALDPDDPPDNNEMKFQISIKAKVWDNVFGDNGDLYLAYTQRSWWQAYNSEASSPFRETNYEPELFLSFNNDTPV---FGWTNTNNRIGINHQSNGRA 191
Carnimonas nigrificans   87 ANENPLSLTTYHRNYVLPIASNTDSVDNNDF-------AVVSPNSHPNHNEVKFQLSIKGRLFHNIWQDNGDVYVAYTQKSWWQAYNSKASSPFRETNYEPELFVDFTNSDSW---LGWTNITNRFGFVHQSNGRS 213

Escherichia coli        162 -------DPTSRSWNRLYTRLMAENGNWLVEVKPWYVVGN---TDDNPDITKYMGYYQLKIGYHLGD----AVLSAKGQYNWNT-GYGGAELGLSYPITKHVRLYTQVYSGYGESLIDYNFNQTRVGVGVMLNDL 279
Agarivorans gilvus      146 -------ELRSRSWNRLYAEFGFNAGPVEVALKPWYRLNEDANDDDNPNIEDYYGYGELSANWQLNP---DHRLSVLARNNLKRDNKGAFDLRWAYRLTPEIALYMKYFNGYGESLIDYNKSNQSIGIGIAVNQ- 269
Agarivorans albus       150 -------QIRSRSWNRVYAEFGFDADPVSIAIKPWYRLPEDESEDDNPNMEDYYGYGELTAKWQINE---RHRLSFLGRNNFKKENKGALDLRWSYGISQELALYLKYFNGYGESLIDYNKHNQSLGIGFAINN- 273
Zymobacter palmae       186 -------DELSRSWNRIFLESHFRNGNWHFKVRPWWRVPESRYDDDNPNIEHYLGYADATLGYSRNE----QEVTWTLRGNPMQ-STISHQIDYSFPMWHKIHGYLQYYNGYGESMVDYDQRVNRIGIGLSFNPE 307
Halotalea alkalienta    192 -------DPISRSWNRVFAEATLERGPMTMSLMPWWRVPESDADDDNPDIEKYVGYADFTFGYTRS----GHEFTWLARGNPGK-GNFGNQLEYAFPLWSKVHGFIQYYEGYGESLIDYDHYVRRIGIGLSFNNV 313
Carnimonas nigrificans  214 -------DPISRSWNRLYAEMILVNGPLQASIKPWWRIPESNHSDDNPDIDNYLGYGQMSLTYTHG----RQEFSYAVTGNPGK-GHFGHQFEYSFPLWHSINGFLQYYNGYGESLIDYNRRVNRIGVGVSFNNI 342
```

Figure S15. Sequence alignment of the 1st and 2nd PLA1 barrel domains in PLA1-PLA1 proteins to the *E. coli* homolog. Sequence alignment was carried out with PROMALS3D (10), and functionally relevant positions highlighted based on (11).

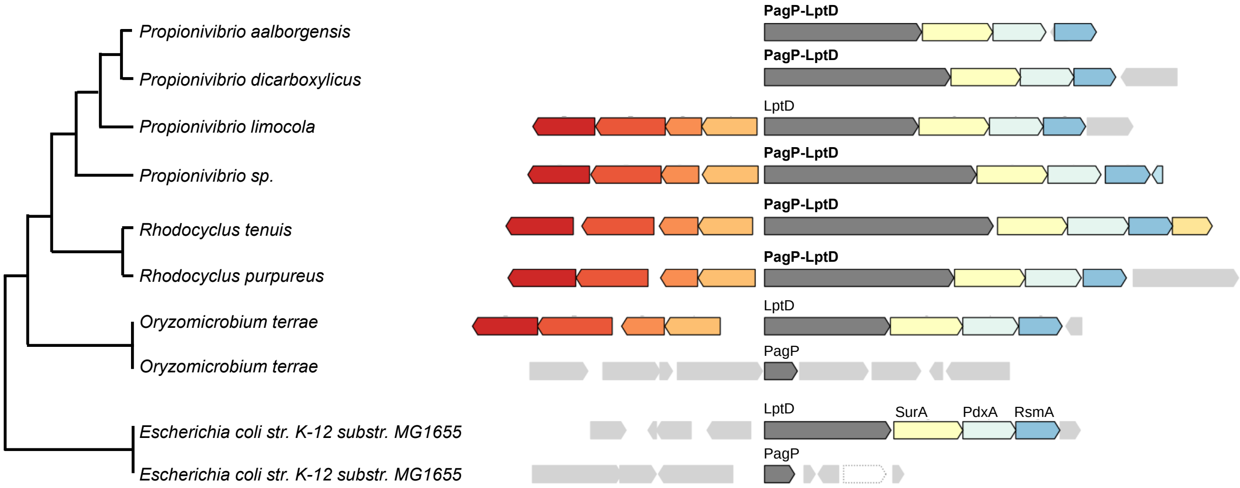


Figure S16. The genomic context of the PagP-LptD MB-family, in comparison to that of the single barrels in *Escherichia coli* (*E. coli*) *K-12* and close species of Propionivibrio. Paralogous genes are colored with the same color, light grey boxes represent the non-conserved genes. Arrow lengths are proportional to the number of the amino acids in the domains.

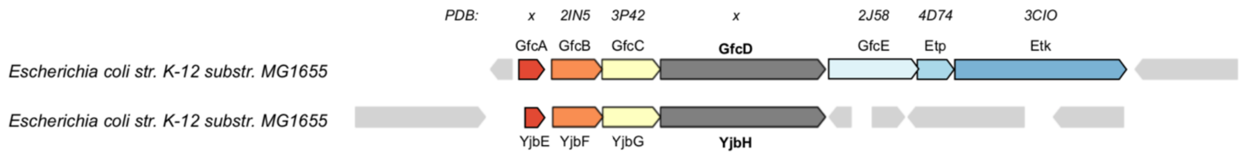

PDB:
x
2IN5
3P42
x
2J58
4D74
3CIO
GfcA
GfcB
GfcC
GfcD
GfcE
Etp
Etk
Escherichia coli str. K-12 substr. MG1655
Escherichia coli str. K-12 substr. MG1655
YjbE
YjbF
YjbG
YjbH

Figure S17. The *gfcABCDE-etp-etk* and *yjbEFGH* operons in *Escherichia coli* (*E. coli*) *K-12*. Homologous genes are colored with the same color and those in the neighborhoods that are not part of the operons are colored grey. The PDB codes for those products for which a full-length or partial structure is known for at least one homolog are shown above the corresponding gene.

Architecture	**#Proteins**	**PDB Homologs**
12-12	4187	5O65_B-4Y25_A fragment
8-8	1959	2K0L_A-2K0L_A
large_single_barrel	1194	6H3I_A fragment-6R2Q_B
large_single_barrel	849	6H3I_A fragment-6R2Q_B
8-8	813	3NB3_A-3QRA_A
8-16	469	2MLH_A-4GEY_A
8-8	392	2MLH_A-2MLH_A
8-8-8	291	3NB3_A-3QRA_A-3QRA_A
22-16	273	3FHH_A-4AFK fragment-6EHB_B fragment
large_single_barrel	273	6H3I_A fragment-6R2Q_B
12-12	222	5O65_B-very weak 3FIP_B
8-10	154	2MLH_A-1I78_B
12-12	121	5O65_B-5O65_B
12-12	102	6R2Q_B partial-5O65_B
large_single_barrel	100	6H3I_A fragment-6R2Q_B
22-8	99	3CSL_A-2MLH_A
8-8-helical-8	91	2K0L_A-4RLC_A-5IJN_Z-2X27_X
8-8-8	76	3NB3_A-6QAM_A-6QAM_A
8-8-8-8-8	67	3NB3_A-6QAM_A-2K0L_A-6QAM_A-2K0L_A
16-16	55	4QL0_A-5MDR_F
8-8	55	2K0L_A-2ERV_A
18-8-8	50	5ONU_B-2MLH_A-2MLH_A
large_single_barrel	50	6H3I_A fragment-6R2Q_B
8-8-8-8	45	3NB3_A-6QAM_A-2K0L_A-6QAM_A
8-8	39	3NB3_A-4RLC_A
8-8-helical	24	3NB3_A-6QAM_A-5IJN_Y
12-8	23	3AEH_B-2MLH_A
8-8	23	3NB3_A-6QAM_A
8-8	19	3GP6_A-6QAM_A
12-12	17	6R2Q_B partial-5O65_B
12-22	17	3QQ2_A-2FFH_A
22-8	17	4RDR_A-2MLH_A
12-12	15	4FQE_A-4FQE_A
8-8-10-8	15	2MLH_A-2MLH_A-1I78_B-2MLH_A
8-8-8-8	14	3NB3_A-2K0L_A-2K0L_A-1QJ8_A
8-8-8-8	13	2MLH_A-2MLH_A-2MLH_A-2MLH_A
14-8	12	3PGU_A-2K0L_A
18-8	11	4AFK_A-2MLH_A
8-8	11	2MLH_A-2F1V_D
8-8	11	3NB3_A-2K0L_A
8-10-8	10	2MLH_A-1I78_B-2MLH_A
8-8	10	3NB3_A-3NB3_A
12-14	9	1UYN_X-2X9K_A
12-8	8	1UYN_X-4FUV_A
8-8-8	8	3NB3_A-2K0L_A-2MLH_A
10-8	7	2X55_A-2MLH_A
10-8	7	1I78-B-2MLH_A
12-12	7	5O65_B-very weak 6GIE_A
12-8	7	3QQ2_A-2MLH_A
8-8	7	2K0L_A-1QJ8_A
8-8	7	2K0L_A-3QRA_A
8-8-helical	7	3NB3_A-6QAM_A-4PX7_A
12-12	6	1QD5_A-1QD5_A
8-8	6	2K0L_A-6QAM_A
8-8	6	2MLH_A-2MLH_A

8-8	6	1QJ8_A fragment-2MLH_A-2MLH_A
8-8-8-8-8-8	6	3NB3_A-2K0L_A-6QAM_A-3QRA_A-2K0L_A-2K0L_A
10-8	5	1I78-B-2MLH_A
12-12	5	1TLY_A-1TLY_A
12-12	5	5O65_B-4Y25_A fragment
12-22	5	3AEH_B-2QLB_A
22-8	5	6HCP_B-2MLH_A
8-10	5	2MLH_A-2X55_A
8-10-8-8	5	2K0L_A-1I78_B-2MLH_A-2MLH_A
8-12	5	3NB3_A-5O65_B
8-8	5	2ERV_A-2K0L_A
8-8	5	2MLH_A-2MLH_A
8-8	5	2MLH_A-2MLH_A
8-8-8	5	2K0L_A-3QRA_A-2MLH_A
22-10	4	4AFK_A-1I78_B
22-8	4	6HCP_B-2MLH_A
22-8	4	6I97_A-2MLH_A
8-8	4	3NB3_A-6QAM_A
8-8	4	2MLH_A-2MLH_A
8-8	4	2F1V_D-2F1V_D
8-8	4	3NB3_A-1QJ8_A
8-8	4	2K0L_A-2X27_X
8-8-8-8	4	3NB3_A-6QAM_A-1QJP_A-1P4T_A
large_single_barrel	4	

Table S1. The number of proteins in each -family with at least 4 similar proteins. **Left column**: barrel architecture. **Central column**: number of similar proteins. **Right column**: The respective PDB homologs. Matches to only part of the target protein are labelled as “fragment”. The central column numbers are plotted in Figure S2.

Cluster Name	Homolog Count	Homology Prediction	Architecture predicted by RaptorX	Architecture predicted by TripletRes	Architecture predicted by trRosetta	Architecture predicted by DeepMetaPSICOV
000	4,187	12-12	12	12	12	12
001	1,959	8-8	8-8	8-8	8-8	8-8
002	1,194	Large Single Barrel	40	40		No Contact Map
003	849	Large Single Barrel		38		
004	813	8-8				
005	469	8-16				
006	392	8-8				
007	291	8-8-8	8-8-8	8-8-8	8-8-8	
008	273	22-16				No Contact Map
009	273	Large Single Barrel				No Contact Map
010	22	12-12				
011	154	8-10				
012	121	12-12		12-12	12-12	
013	102	12-12				
014	100	Large Single Barrel				No Contact Map
015	99	22-8				No Contact Map
016	91	8-8-Helical-8				
017	76	8-8-8				
018	67	8-8-8-8-8				No Contact Map
019	55	16-16				No Contact Map
020	55	8-8				

Table S2: The architectures predicted by RaptorX (12), TripletRes (13), trRosetta (14), and DeepMetaPSICOV (15) for the 21 clusters with more than 50 sequences. Empty cells indicate that the contact maps do not clearly support one architecture. In seven of the twenty-one clusters, DeepMetaPSICOV did not produce any contact map. There are no conflicts among the predicted contact maps, and only in one cluster, the predicted contact map conflicts with the architecture derived from the homologies along the chain. There are four clusters with contact map predictions supporting that they have multiple barrels.

References

1. Frickey T & Lupas A (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* 20(18):3702-3704.
2. Zimmermann L*, et al.* (2018) A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology* 430(15):2237-2243.
3. El-Gebali S*, et al.* (2019) The Pfam protein families database in 2019. 47(D1):D427-D432.
4. Cheng H*, et al.* (2014) ECOD: An Evolutionary Classification of Protein Domains. *PLoS Comput Biol* 10(12):e1003926.
5. Cheng H, Liao Y, Schaeffer RD, Grishin NVJPS, Function,, & Bioinformatics (2015) Manual classification strategies in the ECOD database. 83(7):1238-1251.
6. Berman HM*, et al.* (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235-242.
7. Hayat S, Peters C, Shu N, Tsirigos KD, & Elofsson A (2016) Inclusion of dyad-repeat pattern improves topology prediction of transmembrane β-barrel proteins. *Bioinformatics* 32(10):1571-1573.
8. Gruber M, Söding J, & Lupas ANJJosb (2006) Comparative analysis of coiled-coil prediction methods. 155(2):140-145.
9. Nielsen H (2017) Predicting secretory proteins with SignalP. *Protein function prediction*, (Springer), pp 59-73.
10. Pei J, Kim B-H, & Grishin NV (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Research* 36(7):2295-2300.
11. Snijder H & Dijkstra B (2000) Bacterial phospholipase A: structure and function of an integral membrane phospholipase. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids* 1488(1-2):91-101.
12. Wang S, Sun S, Li Z, Zhang R, & Xu J (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Computational Biology* 13(1):e1005324.
13. Li Y*, et al.* (2021) Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. *PLoS computational biology* 17(3):e1008865.
14. Yang J*, et al.* (2020) Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences* 117(3):1496-1503.
15. Kandathil SM, Greener JG, & Jones DT (2019) Prediction of inter-residue contacts with DeepMetaPSICOV in CASP13. *BioRxiv*:586800.