A bio-informatics approach to identify new drug targets in multidrug-resistant bacteria

Imogen Ramsden[1]*, Dominique de Jong-Hoogland[1], Aryeh Chiam[1], Martin B. Ulmschneider[1]

[1]Department of Chemistry, King's College London, 7 Trinity Street, SE1 1DB, London, UK

*Correspondence: imogen.ramsden@kcl.ac.uk

Abstract

Antibiotic resistance poses a global health crisis. In order to develop new antibiotic agents, it is crucial to identify drug targets in multidrug-resistant bacteria. Criteria for such a target are an α-helical, essential membrane protein, that is non-homologues with the human membrane proteome, and present across multiple bacterial species. Using a stepwise subtractive genomics approach, the membrane protein F0F1 ATP synthase subunit C was identified as a non-human analogues drug target that is present in 11 bacterial species.

Introduction

Antibiotic resistance is the root cause of a rapidly worsening global public health crisis. In 2019, Antimicrobial Resistance Collaborators found that drug-resistance infections killed at least 1.27 million people worldwide, and nearly 5 million deaths associated with bacterial antimicrobial resistance (AMR).[1] With predictions raising the number of deaths due to bacterial AMR to 10 million deaths annually by 2050,[2] the UN Environment Programme, the US Centers for Disease Control and Prevention (CDC), World Health Organisation (WHO) and European Commission have recognised the need for urgent global action.[3-7]

The cost of antibiotic research and low rate of novel antibiotic translation chiefly due to the extremely unfavourable cost-benefit ratio, has discouraged big pharmaceutical companies from developing new antibiotics.[8-10] Consequently, only 17 new antibiotics have been approved between 2010 and 2021,[1, 11] five of which are not in circulation due to the companies' finances.[8, 9] Only four of these new antibiotics represent new classes of antibiotics, targeting bacteria through novel mechanisms.[12, 13] Though there have been three new FDA-approved antibiotics last year (2024), experts worry that this could be a peak due to their associated scientific, regulatory and financial challenges.[14] It remains that many of the broad-spectrum antibiotics currently in use are not effective against multidrug-resistant bacteria, thus it is

critical to look for novel drug targets specific to these species. In 2017, the WHO published a priority pathogens list for research and development of new antibiotics[5], updated in 2024,[15] and dominated by Gram-negative bacteria. This list was composed of multidrug-resistant bacteria – against some of which even the last-resort antibiotic colistin is ineffective.[16, 17] This list was used as the starting point of this investigation which endeavours to identify new drug targets to combat multidrug-resistant bacterial infections.

Membrane proteins constitute the largest class of drug targets as they are responsible for vital physiological processes such as nutrient and waste transportation, signal transduction, and enzymatic catalysis.[18, 19] Due to their placement in the membrane, membrane proteins fold in a very specific way so that the hydrophobic residues are buried in the hydrophobic membrane core, while the more polar residues tend to reside either side of the membrane where they are exposed to aqueous environments.[20] Membrane proteins divide overwhelmingly into two major structural classes, which reflects the two secondary structural motifs that are stable in lipid bilayer membranes: the alpha (α)-helix and the beta (β)-barrel.[21, 22] α-helical proteins are found in the inner membranes of Gram-negative bacteria, Gram-positive bacteria, and eukaryotes while β-barrel proteins are almost exclusively found in the outer membranes of Gram-negative bacteria.[23-25] In an effort to identify a new target for broad-spectrum antibiotics this investigation will focus on α-helical transmembrane proteins.

Methods

A stepwise subtractive genomics approach was used for the identification of unique essential membrane proteins in the 21 bacterial species selected based on the 2017 WHO priority pathogens list, including 8 species of the Enterobacteriaceae family and subspecies of Shigella and Providencia.[5] The scientific names and accession numbers are listed in **Table 1**. Proteins were clustered according to sequence similarity across species to identify potential drug targets in multiple bacterial species, and the protein sequences were aligned and analysed. The overall workflow is illustrated in **Figure 1**.

Retrieving bacterial genomic data and human proteome

The most significant complete genomes (e.g. reference or representative genomes) of the chosen 21 bacterial species were retrieved as translated coding sequences in FASTA files from the NCBI Assembly database at https://www.ncbi.nlm.nih.gov/.[26] The human reference genome GRCh38 was retrieved in the same format.[27]

66 *Identification of membrane protein sequences*

67 The translated coding sequences of each bacterial species and the *Homo sapiens* GRCh38 reference
68 genome were analysed using Phobius algorithm at https://phobius.sbc.su.se/, to locate and predict
69 the topology of trans-membrane (TM) segments, thus identifying membrane proteins[28]. In assembly
70 entries with chromosomal and plasmid DNA, the plasmid sequences were ignored due to their
71 transiency and inconsistency between strains. The proteins predicted to contain TM segments were
72 taken forward for further analysis, though sequences with predicted signal peptides were not taken
73 further due to possible inaccurate predictions.

74 A Python script was written to create a dictionary of the translated coding sequences and their
75 genomic annotations for the human and bacteria proteomes, which created a new FASTA file
76 containing only the membrane protein sequences. These FASTA files were used in the next stages of
77 comparison genome analysis using local BLASTp[29, 30].

78 *Identification of non-homologous proteins*

79 A local NCBI BLASTp searching for non-homologous bacterial membrane proteins was conducted with
80 E-value cut-off of 0.001 and 35% identity[31-36] against the human membrane proteome.[29, 30, 37] The
81 homologous proteins were excluded, and the non-homologs merged into a list, by parsing the outputs
82 in formats 0 and 6.

83 *Identification of essential non-homologous proteins*

84 The human non-homologous membrane proteins were subjected to a local BLASTp similarity search –
85 with E-value cut-off of 0.001, 35% identity and bit-score cut-off of 100 – against the Database of
86 Essential Genes (DEG) to focus on the proteins required for the survival of the bacteria.[38] The
87 parameters used were similar to filtering out the human homologous proteins, with the additional bit-
88 score cut-off achieving closer matches to identify essential proteins. This greater similarity threshold
89 has been used by previous computational studies identifying potential drug targets in bacteria.[31-36, 39-
90 42]

91 *Clustering of membrane proteins*

92 The essential non-homologous membrane proteins were subjected to the CD-HIT suite with a cut-off
93 score 0.9 (90% sequence identity) to cluster the proteins across bacterial species according to
94 sequence similarity.[43-46] Clusters containing ≥5 proteins were selected for sequence analysis; the
95 protein ID headers were paired with the protein sequences and genome annotations used to record
96 the protein names for each cluster.

97 *Sequence alignment and analysis of TM segments*

98 The consistency of the genome annotations of proteins in clusters of ≥5 proteins indicates function
99 similarity, so the proteins used for sequence alignment and analysis were those with consistent
100 genome annotations across species. The largest protein clusters were prioritised with each protein's
101 sequence in the clusters being aligned using MUSCLE[47]. The multiple sequence alignments, saved as
102 CLUSTALW (.aln) files, were viewed using UGENE[48] and predicted topologies from Phobius were used
103 as a guide to compare sequences and infer conserved structure.

104 *Metabolic pathway analysis*

105 The representative protein sequences of each of the 109 protein clusters were concatenated into one
106 FASTA file and run through the BLASTKOALA server[49]. The genes were functionally characterised by
107 BLASTp and KEGG orthology mapping[50-52] to identify the pathways for the proteins showing their most
108 likely role in the organism.

109 *3D protein structures (PyMOL)*

110 The structures containing proteins SecY and Sec61 were downloaded as PDB files from the RCSB PDB
111 website (rcsb.org)[53, 54] with the PDB IDs 3DIN[55, 56] and 6R7Q[57, 58]. The structures were viewed with
112 PyMOL[59] and the chains of SecY and Sec61α were extracted to separate objects from the complexes
113 in 3DIN and 6R7Q, respectively: Chains B/C and F for SecY in 3DIN and Chain HC/XX for Sec61α in
114 6R7Q. The two identical chains of SecY were superimposed to show the similarity, and SecY and
115 Sec61α aligned in PyMOL to compare experimental protein structure.

116

117 **Results and Discussion**

118 All 75,535 translated coding sequences from the chosen bacterial genomes were filtered down after
119 running through Phobius to leave just 13,671 membrane proteins' sequences that did not include
120 signal peptides. The percentage of the genomes encoding membrane proteins was calculated for each
121 bacterium and averaged across all species, resulting in 18% which closely matches the 19% of
122 transmembrane proteins in the human proteome. Though TM proteins containing signal peptides
123 (SPs) were included in this statistic, they were not taken forward for further analysis as SPs may be
124 cleaved or form part of an excreted protein; the number of TM proteins with a signal-peptide
125 compared to the total number of TM proteins in each species of bacteria is shown in **Table S1**.

126 The set of TM proteins was analysed to determine the number and distribution of TM helices. The
127 results of this quantification were summarised in **Figure 2**. Proteins found to be single-spanning (i.e.,
128 only one α-helical segment traversing the membrane) were not taken forward as the simpler structure
129 tends to provide fewer options for a drug binding site.[60, 61]

130 Analysis of the distribution of the number of TM helices among membrane proteins was repeated on
131 the human proteome after extracting them from Phobius results. Like the distributions found in
132 bacteria, single-spanning proteins were the most dominant by far, rising to 42% of the whole
133 membrane proteome (see **Figure 3**).

134 The local BLASTp searches conducted with each bacterial membrane protein FASTA file against the
135 human membrane proteome reduced the number of sequences by around 5%. The thresholds used
136 here were informed from previous work reporting subtractive genomic analysis of other bacteria.[31, 35,
137 41] The sets of non-homologous proteins included proteins labelled as "No hits" as well as those which
138 satisfied the thresholds. The BLASTp tool was used for a selective approach with the 21,967 membrane
139 protein sequences being compared like for like.

140 The membrane proteins found to be non-homologous with the human membrane proteome, were
141 subjected to a local BLASTp search against the Database of Essential Genes (DEG) which utilises the
142 minimal-gene-set concept for bacteria[38, 62]; this process revealed about 40% of them to be genes
143 required to sustain life. The proportion of essential to non-essential membrane proteins is shown in
144 **Figure S1**. These remaining proteins are both non-homologous with the host proteome and essential
145 for survival of the bacteria – 2 key aspects which make for species-specific potential drug targets.

146 Having gathered a set of human non-homologous essential proteins in each of the chosen 21 bacterial
147 species, the aim was to find a drug target that is present across species. To narrow down the number
148 of proteins for this purpose, they were clustered according to sequence similarity using CD-HIT suite.[43,
149 44] A threshold of 80% sequence similarity was initially used since membrane proteins are known to
150 have highly conserved structures due to their restricted folding in the membrane, despite having
151 distinct functions in the cell; therefore, this higher than typical threshold ensured proteins with both
152 similar structure and function were grouped together, while reducing noise.[21, 63, 64] However, latterly
153 the need for sequence alignment required a threshold of 90% similarity. To extract the full protein ID
154 of every protein in the clusters and for retrieval of the protein sequences for sequence analysis later,
155 additional settings were used – the details can be found on Github.[43] Protein sequences containing
156 non-letter characters (i.e. sequencing gaps) were removed, leaving 5183 proteins for clustering.

157 There were 109 clusters which contained 5 or more proteins, the clusters were split into their own
158 files of sequences and accession numbers for sequence analysis and comparisons. The output file from
159 the CD-HIT clustering can be found as a CSV file in the linked data repository (**Dataset D1**). Collating
160 the data showed the distribution of bacterial species in these clusters (see **Figure 3**) where 6 Gram-
161 negative bacteria appeared the most frequently: *E. cloacae, E. coli, K. pneumoniae, S. typhimurium, S.*

162 *flexneri* and *S. dysenteriae*. There were 13 clusters which contained these 6 bacteria in addition to
163 *Serratia marcescens*; *M. tuberculosis* did not share similar membrane proteins with other bacteria.
164 There were 2 protein clusters containing 11 bacterial species, suggested from genome annotations to
165 be preprotein translocase SecY and F0F1 ATP synthase subunit C, which were investigated latterly as
166 potential drug targets.

167 The clustered protein sequences aligned by MUSCLE were visualised in UGENE, to infer the structural
168 similarity among each cluster – the alignment files can be found in the linked data repository (**Dataset**
169 **D2**). This high conservation among many of the TM sequences further supports the homology of the
170 proteins in each cluster. As sequence leads to structure, which encodes function, these unique
171 essential proteins are potential novel antibiotic targets, a hypothesis that is supported by the functions
172 listed in the genome annotation.

173

174 After completing the analysis, Kyoto Encyclopaedia of Genes and Genomes (KEGG) was used to look
175 into the metabolic pathways of the 109 sets of proteins found.[32, 33, 41] The KEGG Pathway Analysis
176 results overview broken down by category and subcategories are summarised in **Table S3**. The 109
177 representative sequences were input as the query and 8 were eliminated due to the absence of
178 recognised metabolic pathways (see **Table S2**), leaving 101 possible proteins across 5 or more high-
179 priority antibiotic-resistant bacterial species. A summary of the results of using BLASTKOALA[49] are
180 reported in **Table 2**, showing the distribution of metabolic pathways. The most highly populated
181 functional categories were signalling and cellular processes and environmental information processing
182 with 29 and 28 proteins, respectively. Proteins were involved with biochemical pathways such as
183 oxidative phosphorylation and membrane transport, with 26 ABC transporters reported.

184 The protein clusters were evaluated for conservation among bacterial species as ideally a drug could
185 be used to target more than one species of antibiotic-resistant bacteria. Starting with proteins
186 conserved in 5 species, namely *E. cloacae, E. coli, S. typhimurium, S. flexneri* and *S. dysenteriae*, there
187 were 36 proteins identified as potential membrane protein drug targets (see **Table 3**). Looking at non-
188 homologous membrane proteins essential and common to 6 species – these same 5 species and *K.*
189 *pneumoniae* – there were 47 potential antibiotic targets identified (see **Table 3**). While these 6 species
190 appeared most often in the protein clusters, there were 10 proteins found in these species that were
191 also present in *S. marcescens* (see **Table 3**). These are dominated by permease and other transport
192 proteins which would be a significant target for antibiotics as inhibition would prevent vital
193 biomolecules and ions from reaching the cell.

194 Although there were a potential 109 proteins to work with as an output from the protein clustering,
195 some candidates were eliminated where annotated gene functions differed across species or
196 metabolic pathways could not be obtained, justification for elimination shown in **Table S2**.

197 There were only 2 membrane proteins found to be common to 11 species, following the workflow of
198 this study: SecY and F0F1 ATP synthase subunit C. SecY has been cited in literature as a homolog of
199 the human translocase subunit Sec61α,[65] so the translated coding sequences for Sec61α were
200 extracted from the GRCh38 genome and aligned with the SecY protein sequences. The overall
201 percentage sequence similarity between S. flexneri protein NP_709088.1 and the 4 Sec61α isoforms
202 was calculated in UGENE, with protein NP_709088.1 chosen as the reference sequence from being set
203 as the representative sequence of the cluster by the CD-HIT suite. Across the 4 isoforms of human
204 protein Sec61 subunit α, the similarity was between 14-16% excluding gaps, supporting the result of
205 the initial BLASTp search against the human proteome. However, characterised structures of SecY and
206 Sec61α were obtained from the RCSB PDB database.[53-58] The SecY pdb file came as a dimer so one
207 chain was extracted for structural comparison; the SecY and Sec61α structures were superimposed
208 (see **Figure 4**). Although not identical, it is clear there are many general structural similarities: though
209 the cytoplasmic and extracellular loops differ – e.g. the intracellular loop at residues 231 to 271 in
210 SecY compared to Sec61α loop (G260 to P290), the TM helices share a very similar structure. This
211 visual comparison could point to the extracellular loops at the top of the proteins shown in **Figure 4**
212 being a possible unique binding site for a drug due to their differences. Further comparative analysis
213 of these 2 proteins should be carried out to determine if SecY remains a potential non-homologous
214 drug target for 11 of the priority pathogens studied here. This inconsistency of sequence similarity and
215 structural similarities may highlight the limitations of purely sequence-based analyses.

216 F0F1 ATP synthase subunit C, though only 79 amino acids long, has a highly conserved structure,
217 possibly a dimeric structure with similar patterns of glycine and alanine residues in both TM segments
218 (see **Figure S2**). According to the KEGG orthology analysis, this protein is classified under
219 "Metabolism", associated with essential pathways such as oxidative phosphorylation (energy
220 generation) and photosynthesis (light to chemical energy conversion). Since there is already a class of
221 clinically used antibiotics, diarylquinolines, that targets the ATP synthase subunit C in *M. tuberculosis*,
222 the F0F1 ATP synthase subunit C found across the 11 species is a potential drug target.

223

224 **Conclusion**

225 In conclusion, being human non-homologous and present in 11 of the 21 investigated priority
226 bacteria, F0F1 ATP synthase subunit C may be a target for potential therapeutic drug development in
227 future. In addition, this study yielded 99 potential membrane protein targets for antibiotics against
228 5-7 species of multidrug-resistant bacteria. However, homology modelling, binding site analysis and
229 *in vivo* experimental work needs to be performed to evaluate these findings.

References

1. Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet* **399**, 629-655 (2022).

2. O'Neil, J. Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations. 2014. *The Review on Antimicrobial Resistance* (2014).

3. CDC. Antibiotic Resistance Threats in the United States, 2019. (2019).

4. World Health Organization. Global Priority List of Antibiotic-Resistant Bacteria to Guide Research, Discovery, and Development of New Antibiotics. , 1-7 (2017).

5. Shrivastava, S., Shrivastava, P. & Ramasamy, J. World health organization releases global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics. *Journal of Medical Society* **32**, 76-77 (2018).

6. European Commission. A European One Health Action Plan against Antimicrobial Resistance (AMR). *European Commission* (2017).

7. United Nations Environment Programme. Bracing for Superbugs: Strengthening environmental action in the One Health response to antimicrobial resistance. (2023).

8. Plackett, B. Why big pharma has abandoned antibiotics. *Nature (London)* **586**, S50-S52 (2020).

9. McKenna, M. The antibiotic paradox: why companies can't afford to create life-saving drugs. *Nature (London)* **584**, 338-341 (2020).

10. Ploy, M. *et al*. Antibiotic resistance: Tools for effective translational research. *Therapies* **75**, 7-12 (2020).

11. Chahine, E. B., Dougherty, J. A., Thornby, K. & Guirguis, E. H. Antibiotic Approvals in the Last Decade: Are We Keeping Up With Resistance? *Annals of Pharmacotherapy* **56**, 441-462 (2022).

12. Hutchings, M. I., Truman, A. W. & Wilkinson, B. Antibiotics: past, present and future. *Current Opinion in Microbiology* **51**, 72-80 (2019).

13. https://sitn.hms.harvard.edu/flash/2018/less-rebooting-antibiotic-pipeline/.

14. Walrath, R. New antibiotics are hard to come by. Red tape is making the problem worse. *Chemical & Engineering News* **102** (2024).

15. World Health Organization. WHO bacterial priority pathogens list, 2024: bacterial pathogens of public health importance, to guide research, development and strategies to prevent and control antimicrobial resistance. (2024).

16. Aghapour, Z. *et al*. Molecular mechanisms related to colistin resistance in Enterobacteriaceae. *Infection and Drug Resistance* **12**, 965-975 (2019).

17. Andrade, F. F., Silva, D., Rodrigues, A. & Pina-Vaz, C. Colistin Update on Its Mechanism of Action and Resistance, Present and Future Challenges. *Microorganisms (Basel)* **8**, 1-12 (2020).

18. Jianting Gong *et al*. Understanding Membrane Protein Drug Targets in Computational Perspective. *Current Drug Targets* **20**, 551-564 (2019).

19. Almeida, J. G., Preto, A. J., Koukos, P. I., Bonvin, Alexandre M. J. J & Moreira, I. S. Membrane proteins structures: A review on computational modeling tools. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1859**, 2021-2039 (2017).

20. Ulmschneider, M. B. & Sansom, M. S. P. Amino acid distributions in integral membrane protein structures. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1512**, 1-14 (2001).

21. Vinothkumar, K. R. & Henderson, R. Structures of membrane proteins. *Quarterly Reviews of Biophysics* **43**, 65-158 (2010).

22. Luckey, M. in *Computational Biophysics of Membrane Proteins* 1-18 (The Royal Society of Chemistry, 2017).

23. Fairman, J. W., Noinaj, N. & Buchanan, S. K. The structural biology of β-barrel membrane proteins: a summary of recent reports. *Current Opinion in Structural Biology* **21**, 523-531 (2011).

24. Schulz, G. E. The structure of bacterial outer membrane proteins. *BBA - Biomembranes* **1565**, 308-317 (2002).

25. Silhavy, T. J., Kahne, D. & Walker, S. The Bacterial Cell Envelope. *Cold Spring Harbor Perspectives in Biology* **2**, a000414 (2010).

26. Kitts, P. A. *et al*. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Research* **44**, D73-D80 (2016).

27. O'Leary, N. A. *et al*. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Scientific Data* **11**, 732 (2024).

28. Käll, L., Krogh, A. & Sonnhammer, E. L. L. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of Molecular Biology* **338**, 1027-1036 (2004).

29. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410 (1990).

30. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

31. Rafiq, H. *et al*. A computational subtractive genome analysis for the characterization of novel drug targets in Klebsiella pneumonia strain PittNDM01. *Microbial Pathogenesis* **146**, 104245 (2020).

32. Uddin, R. & Saeed, K. Identification and characterization of potential drug targets by subtractive genome analyses of methicillin resistant Staphylococcus aureus. *Computational Biology and Chemistry* **48**, 55-63 (2014).

33. Rahman, M. A. *et al*. Identification of potential drug targets by subtractive genome analysis of Bacillus anthracis A0248: An in silico approach. *Computational Biology and Chemistry* **52**, 66-72 (2014).

34. Sudha, R., Katiyar, A., Katiyar, P., Singh, H. & Prasad, P. Identification of potential drug targets and vaccine candidates in Clostridium botulinum using subtractive genomics approach. *Bioinformation* **15**, 18-25 (2019).

35. Wadood, A. *et al*. Subtractive genome analysis for in silico identification and characterization of novel drug targets in Streptococcus pneumonia strain JJA. *Microbial Pathogenesis* **115**, 194-198 (2018).

36. Khan, M. T., Mahmud, A., Iqbal, M. A. & Hasan, M. Subtractive Genomic Analysis for Identification of Novel Drug Targets and Vaccine Candidates against Bartonella bacilliformis subsp. Ver097. *bioRxiv* (2019).

37. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A Greedy Algorithm for Aligning DNA Sequences. *Journal of Computational Biology* **7**, 203-214 (2000).

38. Luo, H. *et al*. DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. *Nucleic Acids Research* **49**, D677-D686 (2021).

39. Sharma, O. P. & Kumar, M. S. Essential proteins and possible therapeutic targets of Wolbachia endosymbiont and development of FiloBase-a comprehensive drug target database for Lymphatic filariasis. *Scientific Reports* **6** (2016).

40. Mondal, S. I. *et al*. Identification of potential drug targets by subtractive genome analysis of Escherichia coli O157: H7: an in silico approach. *Advances and Applications in Bioinformatics and Chemistry* **8**, 49-63 (2015).

41. Damte, D. *et al*. Putative drug and vaccine target protein identification using comparative genomic analysis of KEGG annotated metabolic pathways of Mycoplasma hyopneumoniae. *Genomics (San Diego, Calif.)* **102**, 47-56 (2013).

42. Khan, M. T., Mahmud, A., Iqbal, A., Hoque, S. F. & Hasan, M. Subtractive genomics approach towards the identification of novel therapeutic targets against human Bartonella bacilliformis. *Informatics in Medicine Unlocked* **20** (2020).

43. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

44. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)* **28**, 3150-3152 (2012).

45. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282-283 (2001).

46. Li, W., Jaroszewski, L. & Godzik, A. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* **18**, 77-82 (2002).

47. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797 (2004).

48. Okonechnikov, K., Golosova, O. & Fursov, M. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics (Oxford, England)* **28**, 1166-1167 (2012).

49. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology* **428**, 726-731 (2016).

50. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27-30 (2000).

51. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Science* **28**, 1947-1951 (2019).

52. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research* **49**, D545-D551 (2021).

53. Berman, H. M. *et al*. The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).

54. Burley, S. K. *et al*. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research* **49**, D437-D451 (2021).

55. Zimmer, J., Nam, Y. & Rapoport, T. A. Structure of a complex of the ATPase SecA and the protein-translocation channel. *Nature* **455**, 936-943 (2008).

56. Zimmer, J., Nam, Y. & Rapoport, T. A. Crystal structure of the protein-translocation complex formed by the SecY channel and the SecA ATPase. (2008).

57. Shanmuganathan, V. *et al*. Structural and mutational analysis of the ribosome-arresting human XBP1u. *eLife* **8** (2019).

58. Shanmuganathan, V. *et al*. Structure of XBP1u-paused ribosome nascent chain complex with Sec61. (2019).

59. Schrödinger. The PyMOL Molecular Graphics System. **2.3.3** (2017).

60. Simon C. Bull & Andrew J. Doig. Properties of Protein Drug Target Classes. *PLOS ONE* **10**, e0 (2015).

61. Bakheet, T. M. & Doig, A. J. Properties and identification of human protein drug targets. *Bioinformatics* **25**, 451-457 (2009).

62. Koonin, E. V. How Many Genes Can Make a Cell: The Minimal-Gene-Set Concept. *Annual Review of Genomics and Human Genetics* **1**, 99-116 (2000).

63. Hopf, T. A. *et al*. 3D structures of membrane proteins from genomic sequencing. *Cell (Cambridge)* **149**, 1607-1621 (2012).

64. Flock, T., Venkatakrishnan, A. J., Vinothkumar, K. R. & Babu, M. M. Deciphering membrane protein structures from protein sequences. *Genome Biology* **13**, 160 (2012).

65. Osborne, A. R., Rapoport, T. A. & van den Berg, B. PROTEIN TRANSLOCATION BY THE SEC61/SECY CHANNEL. *Annual Review of Cell and Developmental Biology* **21**, 529-550 (2005).

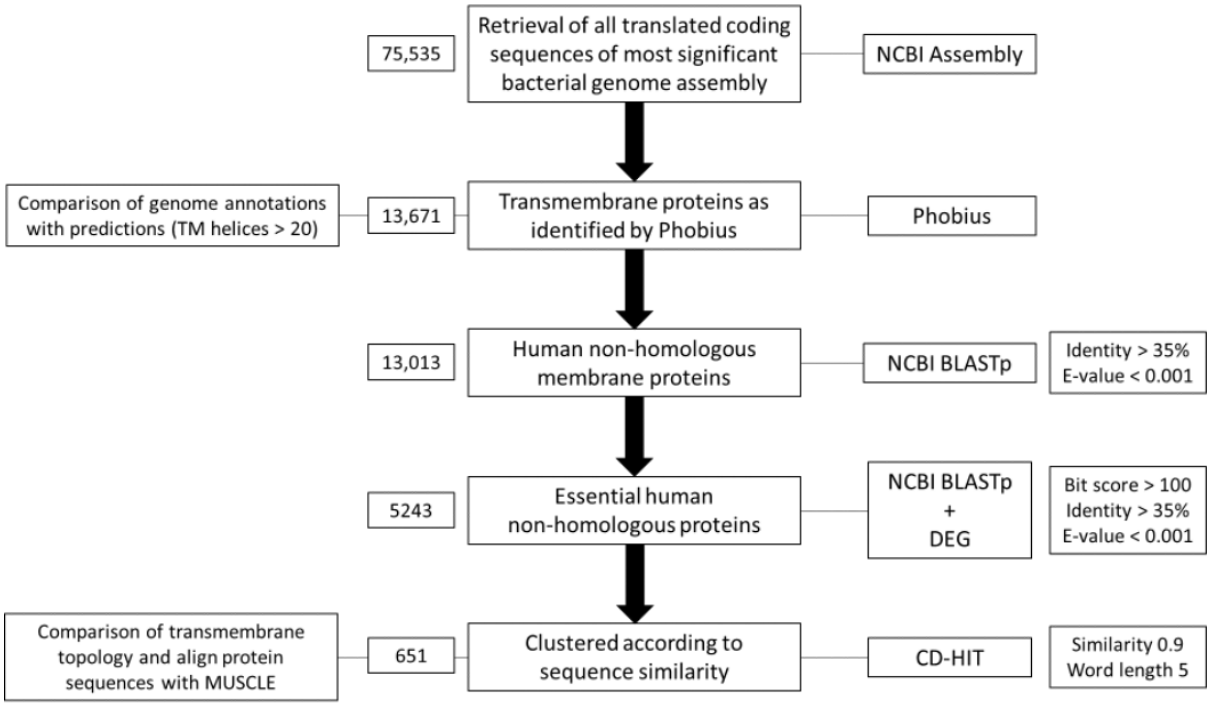

Figure 1. Overall workflow chart. Bit score is a measure of sequence similarity independent of query and database size (the higher the better for this study); E-value is the number of expected hits whih would be found randomly (the lower the better for this study).

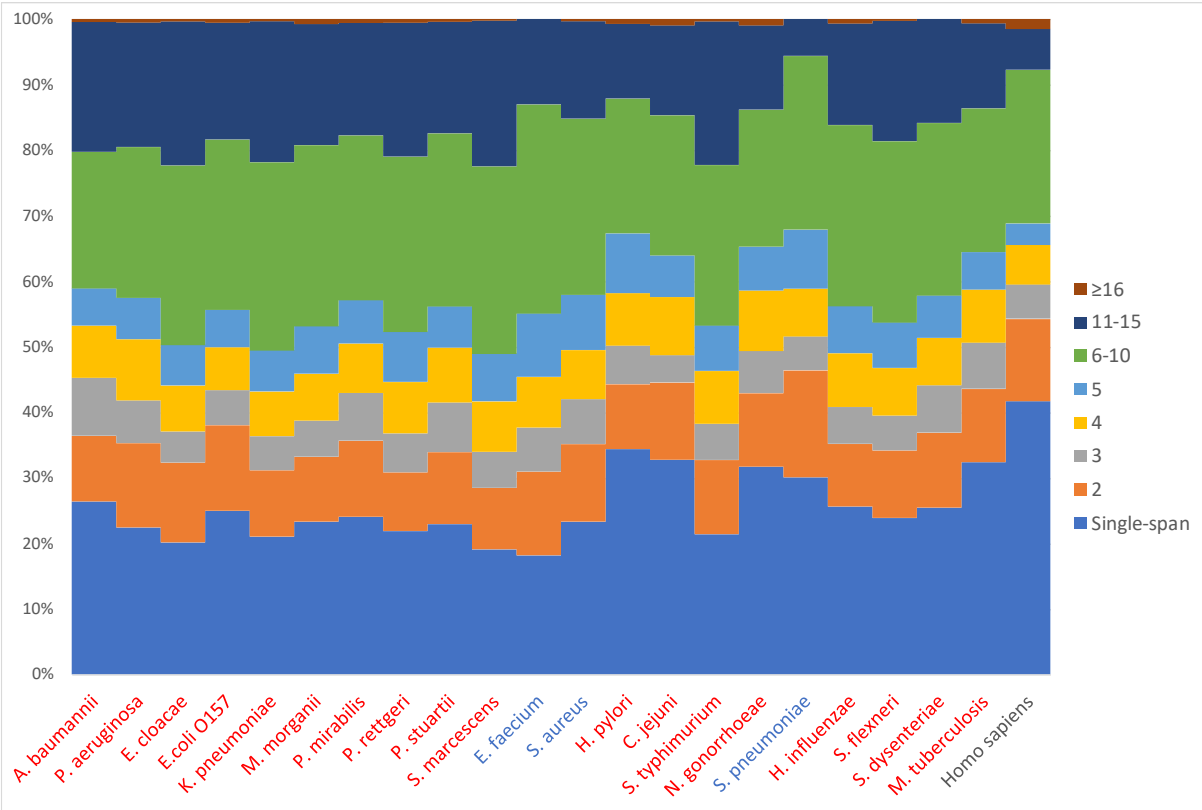

Figure 2. Transmembrane helices distribution in bacterial and human membrane proteins. Bar charts showing the distributions of helices among membrane proteomes in each species of bacteria (gram-negative labelled in red, gram-positive in blue) and in humans (*Homo sapiens* – labelled in black).

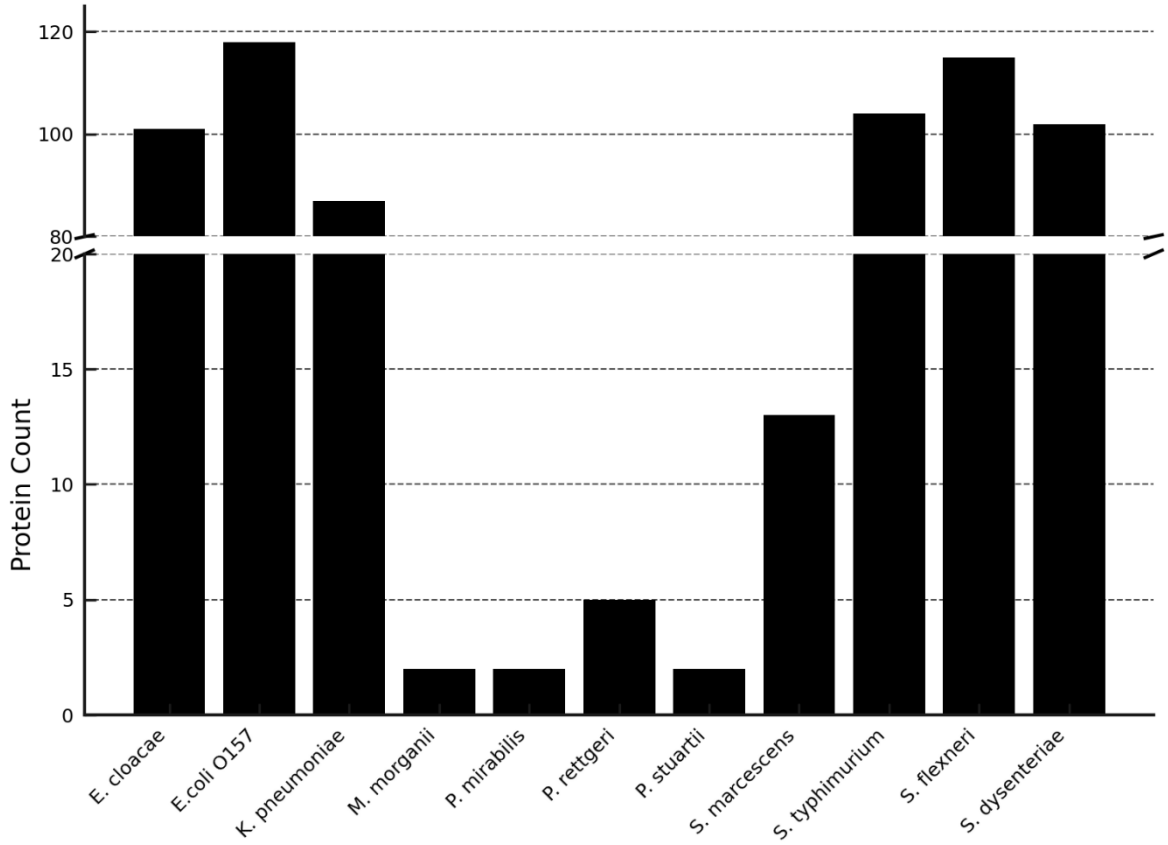

Figure 3. Bar chart showing the populations of clusters containing ≥5 proteins with ≥ 90% similarity across different species of bacteria, bacteria absent from the clusters are omitted.

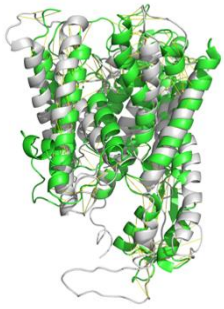

Figure 4. Tertiary structure overlay of SecY in gray (pdb: 3DIN) and Sec61α in green (pdb: 6R7Q).

Table 1. The chosen 21 species of bacteria with accession numbers and (-) or (+) indicating Gram-negative or Gram-positive, respectively. Based on 2017 WHO priority pathogens list and ordered accordingly.

Species	Accession Number	Group
Acinetobacter baumannii	NZ_CP015121.1	Bacteria(-)
Pseudomonas aeruginosa	NC_002516.2	Bacteria(-)
Enterobacter cloacae	NZ_CP009756.1	Bacteria(-)
Escherichia coli O157	NC_002695.2	Bacteria(-)
Klebsiella pneumoniae	NC_016845.1	Bacteria(-)
Morganella morganii	NZ_CP034944.1	Bacteria(-)
Proteus mirabilis	NC_010554.1	Bacteria(-)
Providencia rettgeri	NZ_CP029736.1	Bacteria(-)
Providencia stuartii	NC_017731.1	Bacteria(-)
Serratia marcescens	NZ_CP063354.1	Bacteria(-)
Enterococcus faecium	NZ_CP039729.1	Bacteria(+)
Staphylococcus aureus	NC_007795.1	Bacteria(+)
Helicobacter pylori	NC_017379.1	Bacteria(-)
Campylobacter jejuni	NC_002163.1	Bacteria(-)
Salmonella typhimurium	NC_003197.2	Bacteria(-)
Neisseria gonorrhoeae	NZ_CP012028.1	Bacteria(-)
Streptococcus pneumoniae	NZ_CP007593.1	Bacteria(+)
Haemophilus influenzae	NZ_CP009610.1	Bacteria(-)
Shigella flexneri	NC_004337.2	Bacteria(-)
Shigella dysenteriae	NZ_CP061527.1	Bacteria(-)
Mycobacterium tuberculosis	NC_000962.3	Bacteria(-)

Table 2. Summary of the results of the metabolic pathway analysis through BLASTKOALA showing the number of proteins in each functional category.

BLASTKOALA Category	*Number of Proteins*
Protein families: signalling and cellular processes	29
Environmental Information Processing	28
Energy metabolism	10
Carbohydrate metabolism	8
Protein families: genetic information processing	5
Genetic Information Processing	5
Glycan biosynthesis and metabolism	4
Unclassified: signalling and cellular processes	3
Protein families: metabolism	3
Unclassified	2
Unclassified: metabolism	2
Human Diseases	1
Lipid metabolism	1

Table 3. Protein names, K number and gene from KEGG orthology annotation for proteins found in 5-7 species.

KO Entry	Protein Name	Gene	No. of Species found in
K00427	L-lactate permease	lldP, lctP	7
K03498	trk/ktr system potassium uptake protein	trkH, trkG, ktrB, ktrD	7
K02779	glucose PTS system EIICB or EIICBA component	ptsG	7
K02445	MFS transporter, OPA family, glycerol-3-phosphate transporter	glpT	7
K01000	phospho-N-acetylmuramoyl-pentapeptide-transferase	mraY	7
K15581	oligopeptide transport system permease protein	oppB	7
K15582	oligopeptide transport system permease protein	oppC	7
K11070	spermidine/putrescine transport system permease protein	potC	7
K03562	biopolymer transport protein TolQ	tolQ	7
K10037	glutamine transport system permease protein	glnP	7
K02109	F-type H+-transporting ATPase subunit b	ATPF0B, atpF	7
K18138	multidrug efflux pump	acrB, mexB, adeJ, smeE, mtrD, cmeB	6
K01873	valyl-tRNA synthetase	VARS, valS	6
K03669	membrane glycosyltransferase	mdoH	6
K23676	pyruvate/proton symporter	btsT	6
K05515	penicillin-binding protein 2	mrdA	6
K03495	tRNA uridine 5-carboxymethylaminomethyl modification enzyme	gidA, mnmG, MTO1	6
K03980	putative peptidoglycan lipid II flippase	murJ, mviN	6
K00425	cytochrome bd ubiquinol oxidase subunit I	cydA	6
K00342	NADH-quinone oxidoreductase subunit M	nuoM	6
K11733	lysine-specific permease	lysP	6
K06901	adenine/guanine/hypoxanthine permease	pbuG, azgA, ghxP, ghxQ, adeQ	6
K00343	NADH-quinone oxidoreductase subunit N	nuoN	6
K03475	ascorbate PTS system EIIC component	ulaA, sgaT	6
K11737	D-serine/D-alanine/glycine transporter	cycA	6
K11734	aromatic amino acid transport protein AroP	aroP	6
K11736	proline-specific permease ProY	proY	6
K07638	two-component system, OmpR family, osmolarity sensor histidine kinase EnvZ	envZ	6
K06901	adenine/guanine/hypoxanthine permease	pbuG, azgA, ghxP, ghxQ, adeQ	6
K06156	Gnt-I system low-affinity gluconate transporter	gntU	6
K09808	lipoprotein-releasing system permease protein	lolC_E	6
K03837	serine transporter	sdaC	6
K02824	uracil permease	uraA, pyrP	6

K11103	aerobic C4-dicarboxylate transport protein	dctA	6
K05837	rod shape determining protein RodA	rodA, mrdB	6
K11720	lipopolysaccharide export system permease protein	lptG	6
K12369	dipeptide transport system permease protein	dppB	6
K02037	phosphate transport system permease protein	pstC	6
K10440	ribose transport system permease protein	rbsC	6
K03284	magnesium transporter	corA	6
K01997	branched-chain amino acid transport system permease protein	livH	6
K12370	dipeptide transport system permease protein	dppC	6
K02038	phosphate transport system permease protein	pstA	6
K02796	mannose PTS system EIID component	manZ	6
K06212	formate transporter	focA	6
K11074	putrescine transport system permease protein	potI	6
K02108	F-type H+-transporting ATPase subunit a	ATPF0A, atpB	6
K02066	phospholipid/cholesterol/gamma-HCH transport system permease protein	mlaE, linK	6
K10015	histidine transport system permease protein	hisM	6
K10002	glutamate/aspartate transport system permease protein	gltK, aatM	6
K10009	L-cystine transport system permease protein	tcyB, yecS	6
K02072	D-methionine transport system permease protein	metI	6
K03617	Na+-translocating ferredoxin:NAD+ oxidoreductase subunit A	rnfA	6
K00338	NADH-quinone oxidoreductase subunit I	nuoI	6
K06190	intracellular septation protein	ispZ	6
K03101	signal peptidase II	lspA	6
K03571	rod shape-determining protein MreD	mreD	6
K03210	preprotein translocase subunit YajC	yajC	6
K05365	penicillin-binding protein 1B	mrcB	5
K06200	carbon starvation protein	cstA	5
K02800	mannitol PTS system EIICBA or EIICB component	mtlA, cmtA	5
K03587	cell division protein FtsI (penicillin-binding protein 3)	ftsI	5
K02770	fructose PTS system EIIBC or EIIC component	fruA	5
K03688	ubiquinone biosynthesis protein	ubiB, aarF	5
K03446	MFS transporter, DHA2 family, multidrug resistance protein	emrB	5
K11928	sodium/proline symporter	putP	5
K03762	MFS transporter, MHS family, proline/betaine transporter	proP	5
K14392	sodium/pantothenate symporter	panF	5
K11732	phenylalanine-specific permease	pheP	5
K03293	amino acid transporter, AAT family	TC.AAT	5
K03535	MFS transporter, ACS family, glucarate transporter	gudP	5
K03838	threonine transporter	tdcC	5
K09808	lipoprotein-releasing system permease protein	lolC_E	5
K07862	serine/threonine transporter	sstT	5
K03835	tryptophan-specific transport protein	mtr	5
K02001	glycine betaine/proline transport system permease protein	proW	5
K09811	cell division transport system permease protein	ftsX	5

K10541	methyl-galactoside transport system permease protein	mglC	5
K04087	modulator of FtsH protease HflC	hflC	5
K19228	cationic peptide transport system permease protein	sapC	5
K02047	sulfate/thiosulfate transport system permease protein	cysW	5
K11071	spermidine/putrescine transport system permease protein	potB	5
K02795	mannose PTS system EIIC component	manY	5
K09816	zinc transport system permease protein	znuB	5
K00655	1-acyl-sn-glycerol-3-phosphate acyltransferase	plsC	5
K02018	molybdate transport system permease protein	modB	5
K09125	queuosine precursor transporter	yhhQ	5
K19416	modulator of FtsH protease	yccA	5
K02299	cytochrome o ubiquinol oxidase subunit III	cyoC	5
K03558	membrane protein required for colicin V production	cvpA	5
K03282	large conductance mechanosensitive channel	mscL	5
K00241	succinate dehydrogenase / fumarate reductase, cytochrome b subunit	sdhC, frdC	5
K03586	cell division protein FtsL	ftsL	5
K00247	fumarate reductase subunit D	frdD	5