Sequence-structure-function relationships in the microbial protein universe

Julia Koehler Leman [1,2,#*], Pawel Szczerbiak [3,#], P. Douglas Renfrew [1,2,#], Vladimir Gligorijevic [1,4], Daniel Berenberg [1,4,5,6], Tommi Vatanen [7,8], Bryn C. Taylor [9, 15], Chris Chandler [1], Stefan Janssen [10,14], Nick Carriero [11], Ian Fisk [11], Ramnik J. Xavier [7], Rob Knight [10, 12, 13], Richard Bonneau [1,2,4,5,6], Tomasz Kosciolek [3,10,#*]

Affiliations

1 Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, NY
2 Department of Biology, New York University
3 Malopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland
4 Prescient design, a Genentech accelerator, NY, NY 10010, USA
5 Center for Data Science, New York University, New York, NY 10011, USA
6 Courant Institute of Mathematical Sciences, Department of Computer Science, New York University, New York, NY, USA
7 Broad Institute, Cambridge, MA, USA
8 Liggins Institute, University of Auckland, New Zealand
9 Department of Pediatrics, University of California San Diego, La Jolla, CA, USA
10 Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA 92093, USA
11 Scientific Computing Core, Flatiron Institute, Simons Foundation, New York, NY, USA
12 Department of Bioengineering, University of California, San Diego
13 Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA.

* corresponding
\# These authors contributed equally

Abstract

For the past half-century, structural biologists relied on the notion that similar protein sequences give rise to similar structures and functions. While this assumption has driven research to explore certain parts of the protein universe, it disregards spaces that don't rely on this assumption. Here we explore areas of the protein universe where similar protein functions can be achieved by different sequences and different structures. We predict ~200,000 structures for diverse protein sequences from 1,003 representative genomes[1] across the microbial tree of life, and annotate them functionally on a per-residue basis. Structure prediction is accomplished using the World Community Grid, a large-scale citizen science initiative. The resulting database of structural models is complementary to the AlphaFold database, with regards to domains of life as well as sequence diversity and sequence length. We identify 161 novel folds and describe examples where we map specific functions to structural motifs. We also show that the structural space is continuous and largely saturated, highlighting the need for shifting the focus from obtaining

structures to putting them into context, to transform all branches of biology, including a shift from sequence-based to sequence-structure-function based meta-omics analyses.

Introduction

Structural biology follows the sequence-structure-function paradigm, which states that the sequence of a protein determines its structure, which in turn, determines its function[2–5]. Experimental structure determination efforts were unable to keep up with the exponential growth of available sequences, yet recent breakthroughs in protein structure prediction and renewed focus on machine learning approaches, through methods like AlphaFold[6], now allow for closing the sequence-structure gap. While disordered sequences, large complexes, multiple chains, and protein-protein interactions remain to be addressed, the large number of available protein structures and models has drastically shifted the perspective in the field. Here, we predict the structures of ~200,000 metagenomic sequences leveraging a citizen-science approach. We annotate these models in terms of protein function[7], specifically providing residue-specific annotations, and analyze the features of the resulting protein structure-function universe, including fold novelty and structure-function relationships. Our work demonstrates how to integrate massive structural datasets into a sequence and function context and motivates a shift in perspective to include functional annotations as the starting point to understand biological questions.

Methods

Here we performed large-scale structure prediction on representative protein domains from the Genomic Encyclopedia of Bacteria and Archaea (GEBA1003) reference genome database across the microbial tree of life[1]. From a non-redundant GEBA1003 gene catalog we extracted protein sequences without matches to any structural databases and which produced HHblits alignments deep enough for robust structure predictions using Rosetta[8,9] or DMPfold[10] (N_eff > 16, see supplement). For each sequence we generated 20,000 Rosetta *de novo* models[8,9] using the World Community Grid via the Microbiome Immunity Project and up to 5 models per sequence using DMPFold[10]. We then curated the initial output dataset (MIP raw) of about 240,000 models to arrive at high-quality models comprising about 75% of the original dataset (MIP curated). All analyses in this paper are either on MIP curated or a subset. Putative new folds were identified by comparing our models against representative domains in CATH[11] and the PDB, using a TM-score[12] cutoff of 0.5. Putative novel folds were also verified by AlphaFold. To contextualize our structural findings, we projected 42-dimensional graphlet vector representations of each model in a representative subset encompassing 10,000 structures from MIP curated onto a 3D space, along with CATH representative structures, using UMAP dimensionality reduction. Functional annotations of the entire dataset were created using structure-based Graph Convolutional Network embeddings from DeepFRI[7].

Results & Discussion

How do protein structures and functions differ across species among the tree of life? How much of the protein structure and fold space is still unexplored? And can we learn anything new about the sequence-structure-function universe from microbial proteins? To gain answers to these questions, we run *de novo* structure prediction for ~240,000 protein sequences from the 1,003 reference genomes in the GEBA1003 dataset, sampled across the tree of life. We predict the structures and functions of these microbial proteins and analyze structure-function relationships in the resulting protein universe - a summary of our workflow is shown in Fig. 1A. Structure prediction was carried out on the World Community Grid, a large-scale citizen science partnership between volunteers donating computational time, the host institution (formerly IBM and now the Krembil Research Institute) and researchers.

The MIP database is orthogonal to existing databases -150

Fig. 1B shows the length distributions of protein structures in various databases. The baseline is the PDB90, which are sequences from the Protein DataBank with a pairwise sequence identity <=90%. CATH superfamilies are a non-redundant subset of the PDB90, covering over 6,000 folds (v4.3.0). The AlphaFold protein structure database contains over 360,000 protein models, vastly increasing the known structure space, and covers a wide range of organisms and sequence lengths, primarily from Eukaryotes[13,14]. Our MIP database is distinct from the other databases because it consists of proteins from Archaea and Bacteria, whose protein sequences are generally shorter than from Eukaryotes[15,16]. MIP models drastically increase the available structure space of smaller proteins and domains from 40 to 200 residues (Fig. 1B). We further split the sequences into domains before structure prediction, unlike structures in the AlphaFold database. Also, only about 3.6% of structures in the AlphaFold database belong to Archaea and Bacteria, indicating that AlphaFold and MIP databases are complementary.

The Rosetta models in our MIP database generally contain fewer coil residues than the DMPFold models (Supplement), yet the quality of the DMPFold models is higher for larger proteins (see Supplement). Further, the model quality assessment score (average TM-score[17] of 10 lowest energy models from Rosetta and the raw score from DMPFold) correlates with the TM-score between the Rosetta and DMPFold models (see Supplement), indicating that models that agree better between Rosetta and DMPFold, are generally of higher quality.

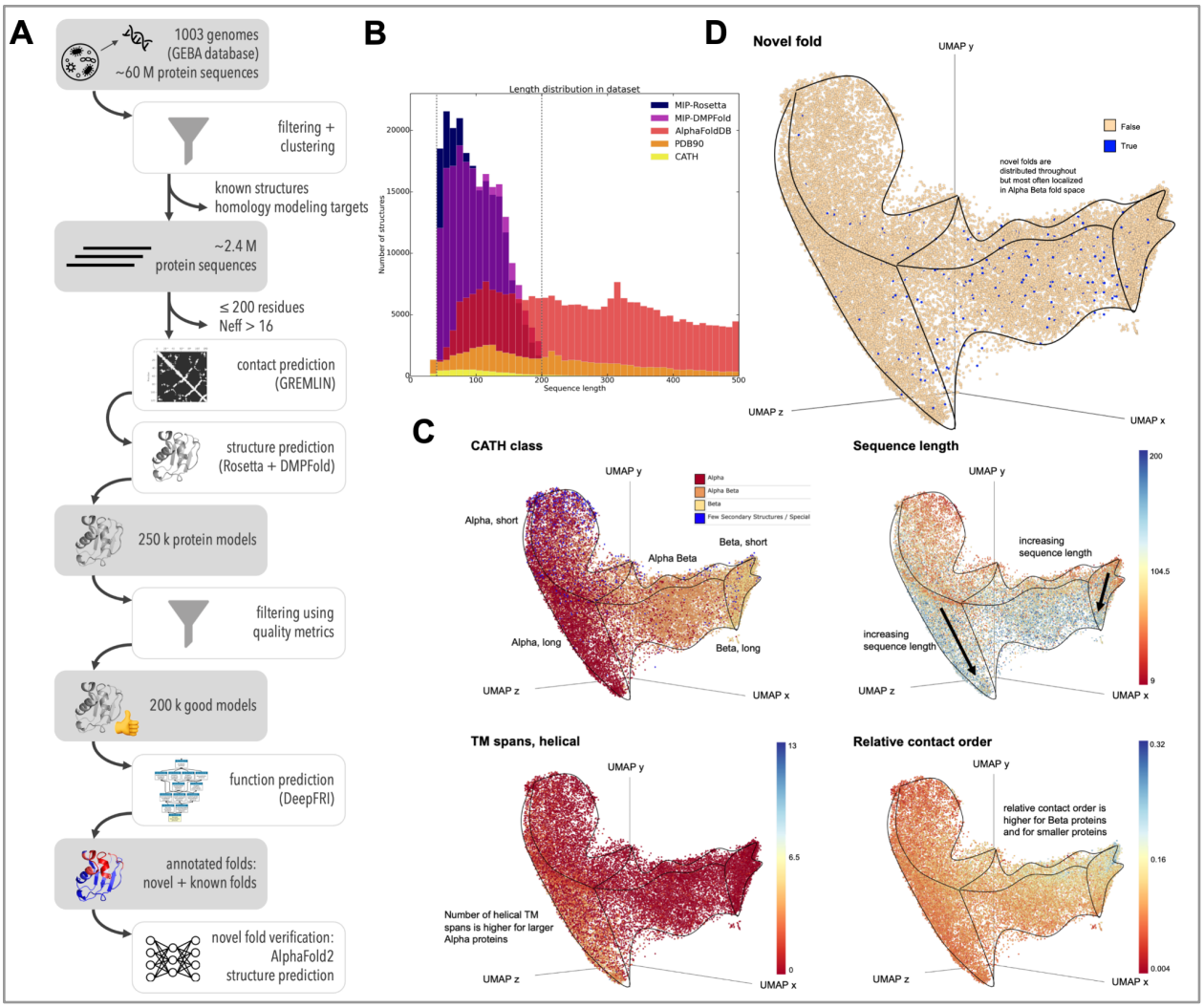

Fig. 1: The fold space covered by the microbial protein structure universe is continuous. (A) Flowchart of our process to arrive at ~200,000 de novo protein models covering a diverse sequence space. (B) The sequence length distribution shows that our sequences are shorter than many of the proteins in the PDB, CATH or AlphaFold databases. Our proteins are between 40 and 200 residues long, which is in agreement with the fact that microbial protein sequences are often shorter than eukaryotic sequences. (C) The protein structure universe in UMAP space is color-coded according to features, such as similarity to CATH classes, sequence length, number of helical transmembrane spans, and relative contact order. (D) Novel folds (blue dots) are spread throughout the fold space with fewer representatives in the purely α-helical and purely β-sheet folds.

The microbial protein universe maps into a continuous fold space - 200

Visualization was created by generating a 42-dimensional graphlet vector representation for each structure in the vizualization dataset and CATH superfamilies and mapping these vectors into 3D space using UMAP dimensionality reduction[18,19] (Fig. 1C and D). The surfaces of the 3D structure cloud are outlined in black. We investigated several features in this mapping, including sequence length, relative contact order, number of transmembrane spans, and mapping to a CATH class.

The 3D mapping of the protein universe allows to distinguish different sequence lengths, the number of helical transmembrane spans and the relative contact order of the protein folds, as different shadings show in Fig. 1C. The visualization further illustrates (see Supplement) that the protein universe space is continuous, indicating that folds may evolve along a trajectory where small changes in the tertiary structure can eventually lead to a different fold, which is in agreement with prior work[20,21]. In contrast, a discrete fold space would display distinct clusters of folds that require larger conformational changes to interconvert between them. We identify 452 previously unseen structures in our MIP dataset that cluster into 161 distinct, novel folds (114 singletons and 47 clusters with multiple structures). Fig. 1D shows that the majority of novel folds are distributed throughout α/β fold space (compare with Fig. 1C) with few novel folds in α or β fold space.

MIP dataset explores the sequence-structure-function universe - 300

We computed pairwise similarities between 3052 random sequences in the curated dataset (*MIP random5000_curated* - see Supplement) in terms of sequence identity, structural similarity (TM-score) and functional similarity (cosine similarity score). This was compared against a PDB baseline of 1000 protein chains, covering pairwise sequence similarities between 0 and 100%.

By design, protein sequences in the MIP dataset are dissimilar (30% sequence identity cutoff). When correlating sequence similarities to structural similarities for pairs of proteins, the vast majority of dissimilar sequences fold into different structures. However, there are a fair number of proteins that have vastly different sequences and still fold into similar structures (Fig. 2B). The PDB baseline that covers sequence similarities across all ranges from 0 - 100% confirms this expected trend, and it also confirms the general notion that similar sequences fold into similar structures (Fig. 2A).

When correlating sequence identity and functional similarity, the majority of sequences have different functions, but still a fair number of dissimilar sequences have similar functions. This originates in the multiplicity of biological systems (Fig. 2D). i.e. achieving the same functional outcome by different pathways. The PDB baseline gives the same trend and has the additional known population where similar sequences achieve similar functions (Fig. 2C).

When correlating structural similarity (of dissimilar sequences!) to functional similarity, we find 4 populations (Fig. 2F): (a) the largest population following expectations of dissimilar structures having different functions - quadrant III, (b) the 2nd largest population of dissimilar structures having similar functions - quadrant I, (c) the third largest population of similar structures having different functions - quadrant IV, and (d) the smallest population following expectations of similar structures having similar functions. Quadrants I and IV are the most interesting ones with examples shown in Fig. 3. The PDB baseline covers all sequence similarities and follows mainly known expectations of quadrants II and III (Fig. 2E).

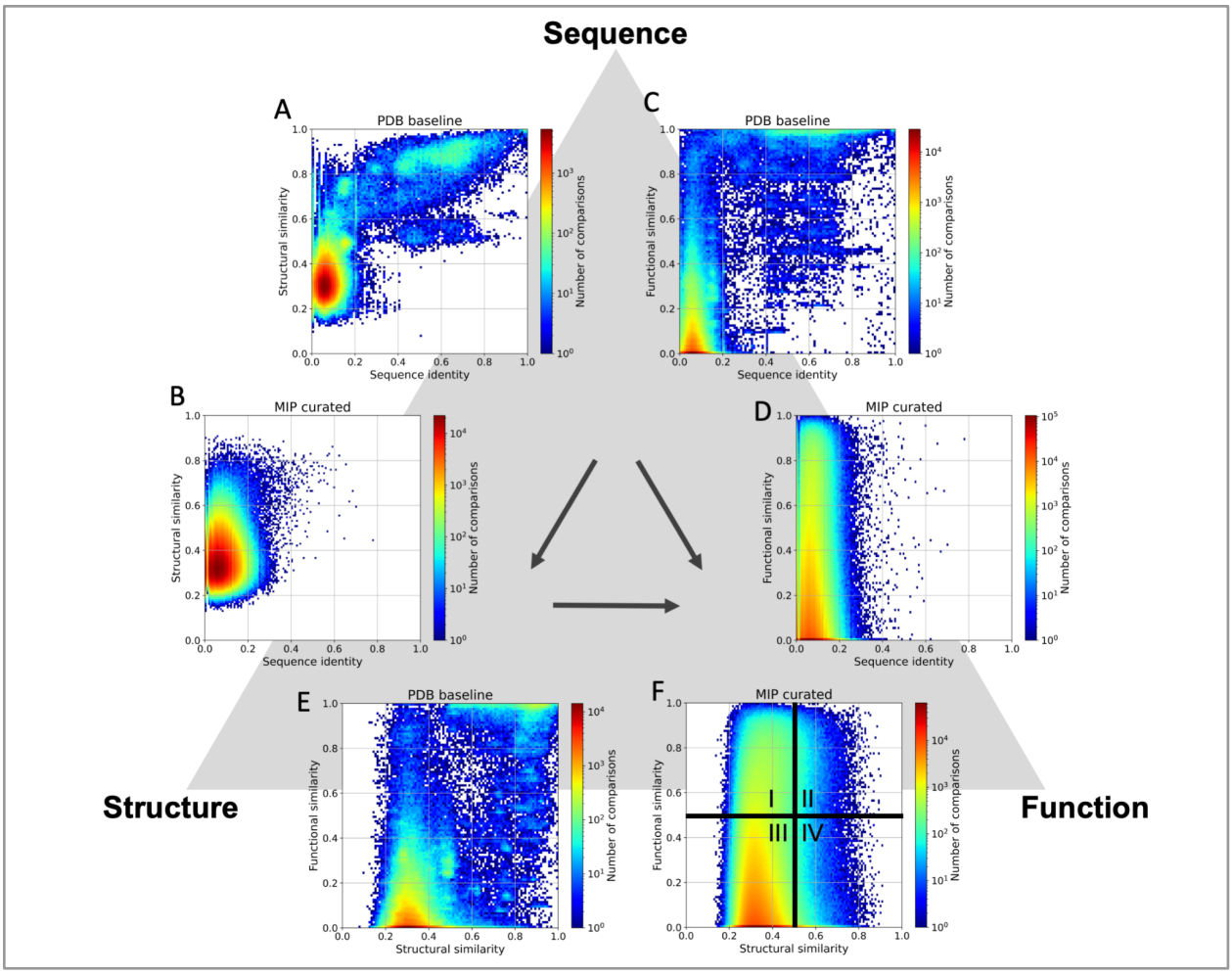

Fig. 2: Sequence-structure-function relationships in both PDB and the MIP dataset. Pairwise comparisons of protein sequences (using sequence identity), structures (TM-score), and functions (cosine similarity) for two datasets: a baseline from the PDB and the MIP random5000_curated dataset, containing 3052 structures (see Supplement). The PDB baseline dataset contains 1000 chains covering pairwise sequence similarities between 0 and 100% while the MIP dataset is a non-redundant set with mostly dissimilar sequences (sequence identity <30% threshold was imposed before sequential domain splitting). Analyses of these two datasets in this way lead us to the following conclusions: sequence identity correlates with structural similarity (A), yet high structural similarity can be achieved by low sequence identity (B). High sequence identity leads to high functional similarity (C), yet high functional similarity can be achieved by proteins with low sequence identity (D). Structural similarity often correlates with functional similarity ((E) and quadrants II and III in (F)). However, there are plenty of examples where low structural similarity can be seen in proteins with high functional similarity (quadrant I in (F)), and highly similar structures can exhibit different functions (quadrant IV in (F)).

Most functions are produced by the same structural motifs, even for dissimilar sequences - 300

For each of the 161 novel fold structural clusters, we compared functional similarities for each protein pair by computing the cosine similarity for the function vectors; this is shown as a heatmap in Fig. 3. We then picked several proteins for each structural cluster and mapped some of the top-scoring functions onto the predicted structures (right panel in Fig. 3). Residues that our function prediction network DeepFRI predicts to have high importance to achieve a particular function are highlighted in red, whereas blue residues are not involved in generating that particular function. Many examples of this analysis are outlined in the supplement, some of which are shown in Fig. 3. We find that the majority of functions in those structural clusters map to the same residues in the structure ("structural motif") as shown for the largest cluster 161 the Supplement. However, we also find more complicated structure-function relationships in these clusters as shown in Fig. 3 and discussed in the next section. Note that the sequences in each structural cluster (and in the MIP dataset) are dissimilar to each other and neither structural nor functional prediction could be inferred by sequence identity for these proteins due to lack of homology.

Per-residue functional annotations reveal a more complex picture of protein structure-function relationships - 300

Some of the structure-function relationships map to quadrants I and IV in Fig. 2F, where similar structures can have different functions or different structures can have similar functions. To a first approximation this is not surprising. Similar structures can achieve different functions due to the fact that the gene ontology database is organized in a hierarchical manner and that parent or child functions are related but still different[22]. Different structures can generate similar functions due to the multiplicity of functional pathways as a back-up plan for organisms to survive. However, a closer look at some of the structure-function disparities reveals some surprises.

Fig. s 3A and B show two proteins that use the same structural motif for different functions. While the overall sequence identity between these proteins is low (~30 and 25% for panels A and B respectively), a short sequence motif underlies the structural motif, which in turn has different functions. Fig. 3A shows two proteins where the terminus of the central helix is involved in phosphatase activitity, where the same motif in a different protein is involved in actin binding. The sequence motif for this region is GGWDXP. In Fig. 3B, the N-terminus of one protein is involved in zinc ion binding and 'catalytic activity, acting on a protein', whereas the N-terminus of another protein of that structural cluster is involved in DNA binding and 'identical protein binding'. The underlying sequence motif for this structural motif is CXCCG.

Fig. s 3C, 3D, and 3E show examples where a different structural motif in the same protein fold achieves the same function. This seems unusual and doesn't seem to rely on a short sequence motif. In the first example (Fig. 3C), transferase activity either maps to a beta-sheet or a C-terminal short helix in two different proteins. In the second example transmembrane transporter activity maps to either two helices or a beta-sheet (Fig. 3D). Fig. 3E shows that ion transmembrane transporter activity maps to different structural motifs for different proteins. This entire structural cluster (cluster 146) has very high similarity across predicted functions, indicated by the heatmap showing mostly yellow hues.

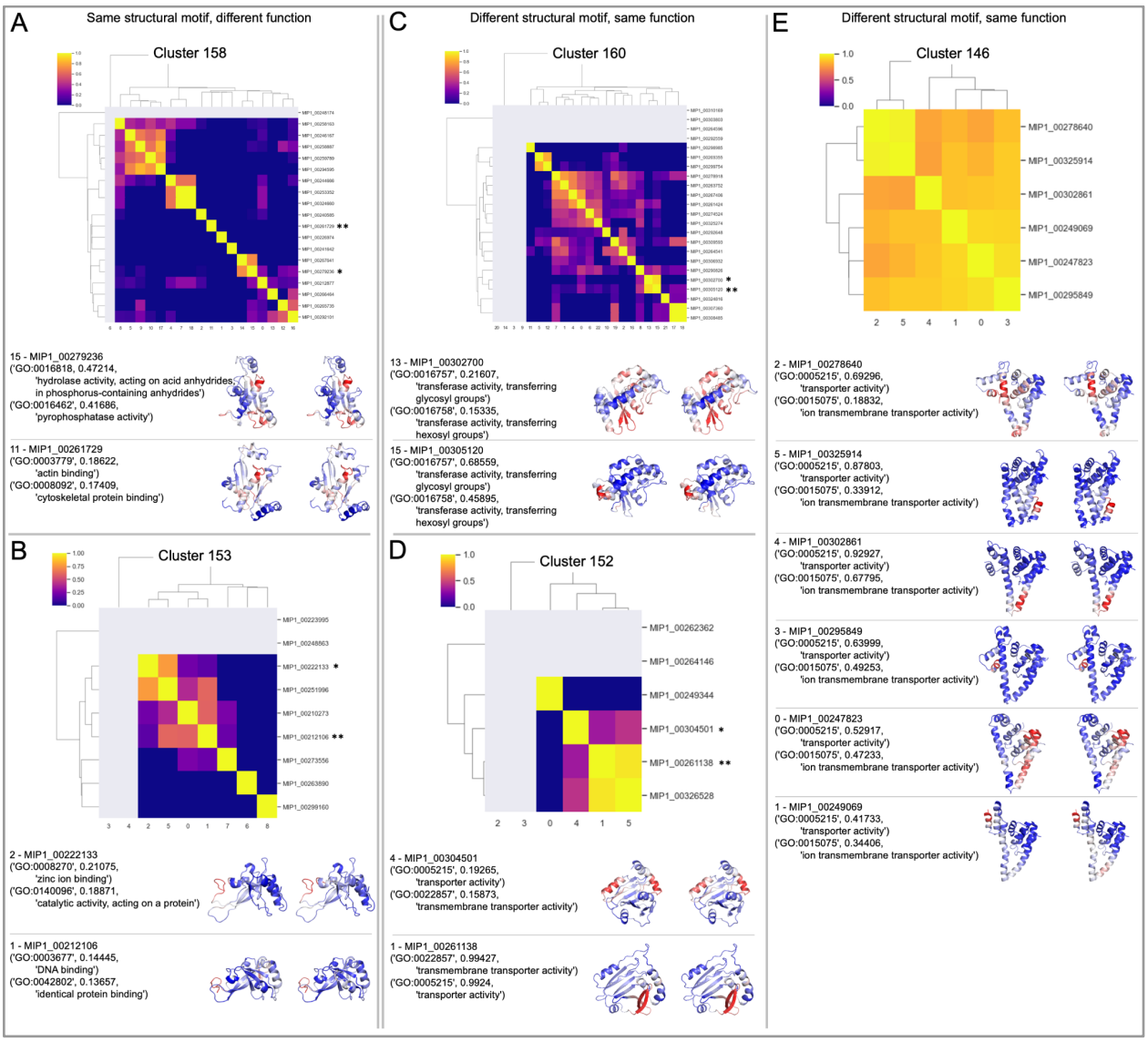

Fig. 3: Functional diversity of proteins with the same structure.
We show examples from several structural clusters that exhibit novel folds. The heatmaps show functional similarity (cosine similarity of the function vectors) of protein pairs within the cluster. Proteins that have predicted functions with scores < 0.1 are shown in gray in the heatmaps. Asterisks highlight the examples shown below. (A) and (B) show cases where the same structural motif in two different proteins produces different, unrelated functions. (C), (D), and (E) show cases where the same function is generated by different structural motifs in different proteins, even though the proteins have the same fold.

Higher functional specificity is carried out by fewer possible folds - 350

We investigated different protein functions and examined the structures with these functions (Fig. 4 and Supplement). Some protein functions are sufficiently general such that they can be achieved by different folds, examples are 'carbohydrate binding' (GO:0030246), 'protein tyrosine kinases' (EC 2.7.10.), and 'mitigation of host immune response by virus' (GO:0030683). More

specific functions are accomplished by fewer folds. Examples of specific functions with a single fold are 'thymidine kinase' (EC 2.7.1.21) and 'sole sub-class for lyases that do not belong in the other subclasses' (EC 4.99.1.).

The functional cluster for carbohydrate binding (Fig. 4A) covers many different folds with high β-sheet propensity, including β-barrels, twisted sheets, and stacked sheets. This class contains a single helical protein, indicated by the single blue line in the heatmap in Fig. 4A with the structure shown in (F). The largest structural cluster in this functional category corresponds to the largest novel-fold cluster (yellow square in the heatmap) and the salient residues in this cluster overlap nicely.

Fig. 4B shows the function 'maintenance of CRISPR repeat elements'. CRISPR repeats are short DNA sequences in bacteria and archaea. They derive from DNA fragments of bacteriophages that previously infected those organisms and allows them to identify recurring invaders. Hence, the CRISPR-Cas system functions like a microbial immune system. Cas1 and Cas2 identify the site in the bacterial genome where viral DNA is inserted and ultimately cleaved by Cas9. The structural cluster in Fig. 4B (A) overlays with part of Cas2 (PDBID 5sd5, chains EF) and the predicted salient residues bind DNA in the structure. Cluster (D) in Fig. 4B is similar to parts of Cas1 (PDBID 5sd5, chains ABCD) but doesn't overlay perfectly.

Fig. 4C shows the function 'sole sub-class for lyases that do not belong in the other subclasses'. Lyases are enzymes that catalyze the breaking of chemical bonds by means other than hydrolysis or oxidation. None of the lyases in the other classes (EC. 4.1 - EC.4.6) has the same fold as our predicted MIP models, even though there are structural similarities. Our models have an (αβ)x3 fold with sequential strand connections – the other lyases have various (αβ)xN folds but their strand connections are non-sequential.

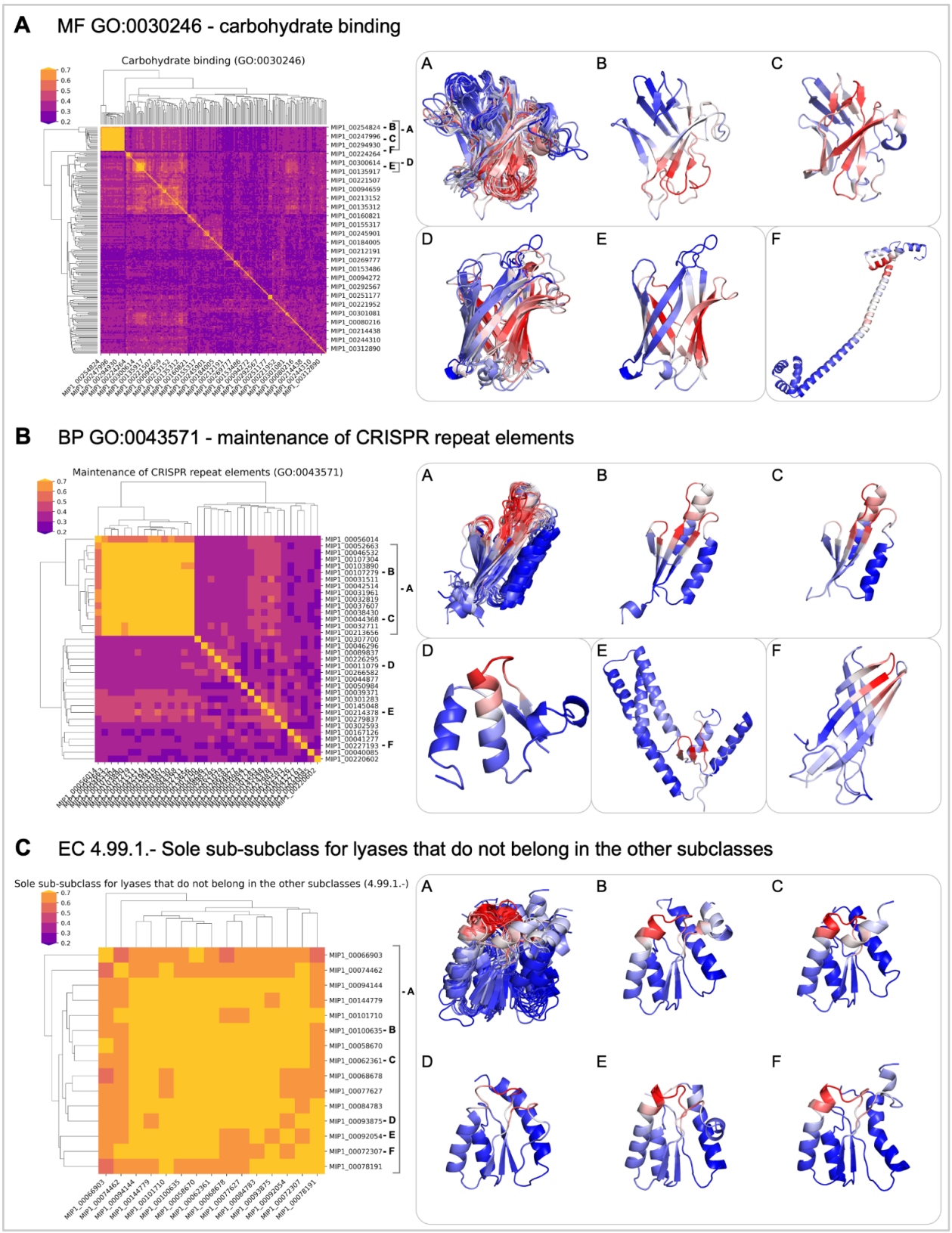

Fig. 4: Structural diversity of proteins with the same function.

We examine proteins that have the same function and plot the TM-score as a measure of structural similarity as a heatmap, with larger numbers (more yellow) representing more similar structures. We also map the residue-specific function predictions onto the structures on the right, where residues in red are responsible for the functions. (A) Gene ontology molecular function carbohydrate binding with GO number GO:0030246. Except for the protein shown in (F) which has high helical propensity, the proteins in this functional cluster have high β-sheet content. The largest cluster in the heatmap in yellow is also the largest novel-fold cluster. The salient residues responsible for this function overlay nicely across the proteins in this cluster. (B) Gene ontology biological process function 'maintenance of CRISPR repeat elements' with GO number GO:0043571. The largest cluster highlighted in yellow superimposes with Cas2 and the salient residues in red interact with DNA. (C) Enzyme commission number EC 4.99.1. with the function 'Sole sub-class for lyases that do not belong in the other subclasses'. All structures in this functional cluster have the same fold and the salient residues responsible for this function overlay onto the same structural motif in the protein. More details in the text.

Conclusion

In summary, we used a citizen-science approach to predict ~200,000 protein structures for non-redundant microbial sequences across the tree of life. Structures were predicted by two state-of-the-art independent methods (Rosetta and DMPFold) and evaluated by quality metrics to indicate model quality. Functional annotations give us a unique look at the microbial protein universe in terms of sequence, structure, and function. Our database is orthogonal to the AlphaFold database in terms of domains of life, sequence diversity and sequence length. We predicted 161 novel folds which were verified by AlphaFold. With functional annotations, we can more closely relate sequence-structure-function relationships in this universe, that go beyond the main homology assumption of similar sequences, structures and functions. In fact, we frequently see that these dissimilar sequences fold into similar structures, indicating that the sequence diversity is much greater than the structural diversity. Moreover, we provide examples that challenge our classic understanding of biological sequence-structure-function relationships. We hope that this research inspires the scientific community to advance our understanding of site-specific protein function by developing experimental and computational tools for high-quality measurements and predictions. Only these new tools can lead to a more complete understanding of the complexities of how proteins fold, function, evolve and interact, to address questions of health and disease.

Contributions

RK, RB, RJX, TK conceived and initiated the project; TK, JKL coordinated the project; PDR, TK prepared and supervised World Community Grid project execution; VG developed the methodology and performed functional annotations; DB, PS developed, implemented and performed low-dimensional representation of the protein space; PS, JKL, TK, PDR analysed the data; JKL, TK, PDR, PS wrote the manuscript with input from all authors; CC, PDR prepared computational framework for data aggregation and analysis; TK, PDR, JKL, BCT, TV performed preliminary data analysis; SJ, TK implemented sequence processing pipeline; IF, NC provided HPC cluster and computational support.

Acknowledgments

We kindly acknowledge the wonderful support and encouragement of the IBM World Community Grid team (Caitlin Larkin, Juan A Hindo, Al Seippel, Erika Tuttle, Jonathan D Armstrong, Kevin Reed, Ray Johnson, and Victors Berstis), and the community of 790K volunteers who donated 140,661 computational years (!) since Aug 2017 to this project. This research was also supported in part by PLGrid Infrastructure.

References

1. Mukherjee, S. *et al.* 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol. 2017 357* **35**, 676–683 (2017).
2. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science (80-.).* **181**, 223–230 (1973).
3. Maynard Smith, J. Natural Selection and the Concept of a Protein Space. *Nat. 1970 2255232* **225**, 563–564 (1970).
4. Aharoni, A. *et al.* The 'evolvability' of promiscuous protein functions. *Nat. Genet. 2004 371* **37**, 73–76 (2004).
5. Redfern, O. C., Dessailly, B. & Orengo, C. A. Exploring the structure and function paradigm. *Curr. Opin. Struct. Biol.* **18**, 394–402 (2008).
6. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nat. 2021 5967873* **596**, 583–589 (2021).
7. Gligorijević, V. *et al.* Structure-based protein function prediction using graph convolutional networks. *Nat. Commun. 2021 121* **12**, 1–14 (2021).
8. Ovchinnikov, S. *et al.* Protein structure determination using metagenome sequence data. *Science (80-.).* **355**, 294–298 (2017).
9. Koehler Leman, J. *et al.* Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature Methods* vol. 17 665–680 (2020).
10. Greener, J. G., Kandathil, S. M. & Jones, D. T. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat. Commun. 2019 101* **10**, 1–13 (2019).
11. Sillitoe, I. *et al.* CATH: increased structural coverage of functional space. *Nucleic Acids Res.* **49**, D266–D273 (2021).
12. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889 (2010).
13. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nat. 2021 5967873* **596**, 590–596 (2021).
14. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
15. Zhang, J. Protein-length distributions for the three domains of life. *Trends Genet.* **16**, 107–109 (2000).
16. Gong, X. *et al.* Comparative analysis of essential genes and nonessential genes in Escherichia coli K12. *Mol. Genet. Genomics* **279**, 87–94 (2008).
17. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
18. Faisal, F. E. *et al.* GRAFENE: Graphlet-based alignment-free network approach integrates 3D structural and sequence (residue order) data to improve protein structural comparison. *Sci. Rep.* **7**, (2017).

19. Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A. & Knight, R. EMPeror: A tool for visualizing high-throughput microbial community data. *Gigascience* **2**, 1–4 (2013).
20. Holm, L. & Sander, C. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.* **26**, 316–319 (1998).
21. Taylor, W. R. Evolutionary transitions in protein fold space. *Curr. Opin. Struct. Biol.* **17**, 354–361 (2007).
22. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet. 2000 251* **25**, 25–29 (2000).