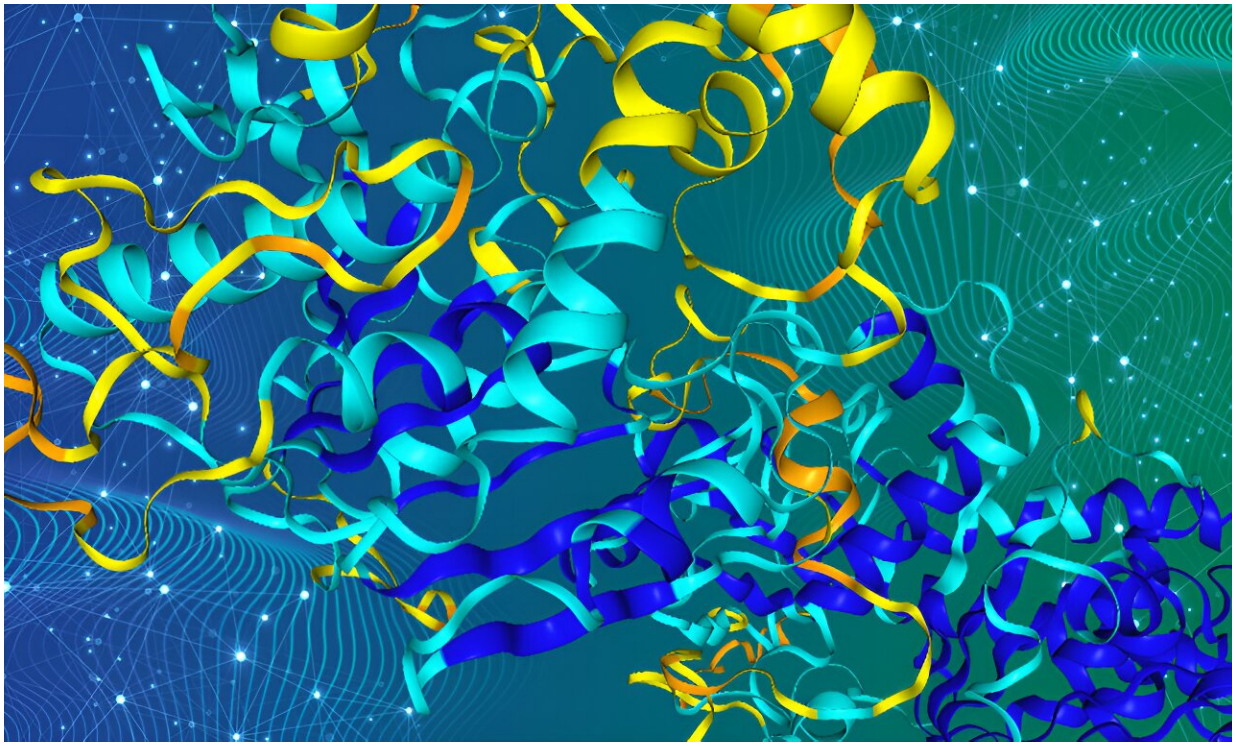
Revealing the secrets of protein evolution using the AlphaFold database

September 13 2023, by Vicky Hatch



Revealing the secrets of protein evolution using the AlphaFold database. Credit: Karen Arnott/EMBL-EBI

By developing an efficient way to compare all predicted protein structures in the AlphaFold database, researchers have revealed similarities between proteins across different species. This work aids our understanding of protein evolution and has uncovered new insights into the origin of human immunity proteins.

The research was conducted by EMBL's European Bioinformatics Institute (EMBL-EBI), the Institute of Molecular Systems Biology ETH Zurich, and the School of Biological Sciences Seoul National University.

The AlphaFold database is a transformative resource in the field of protein research, serving as a comprehensive repository of AI-predicted 3D structures for all known proteins. The database fills a critical gap in understanding protein function and evolution by offering high-quality structural predictions. Although AI predictions are not a substitute for experimentally determined structures, they do provide invaluable insights for the scientific community.

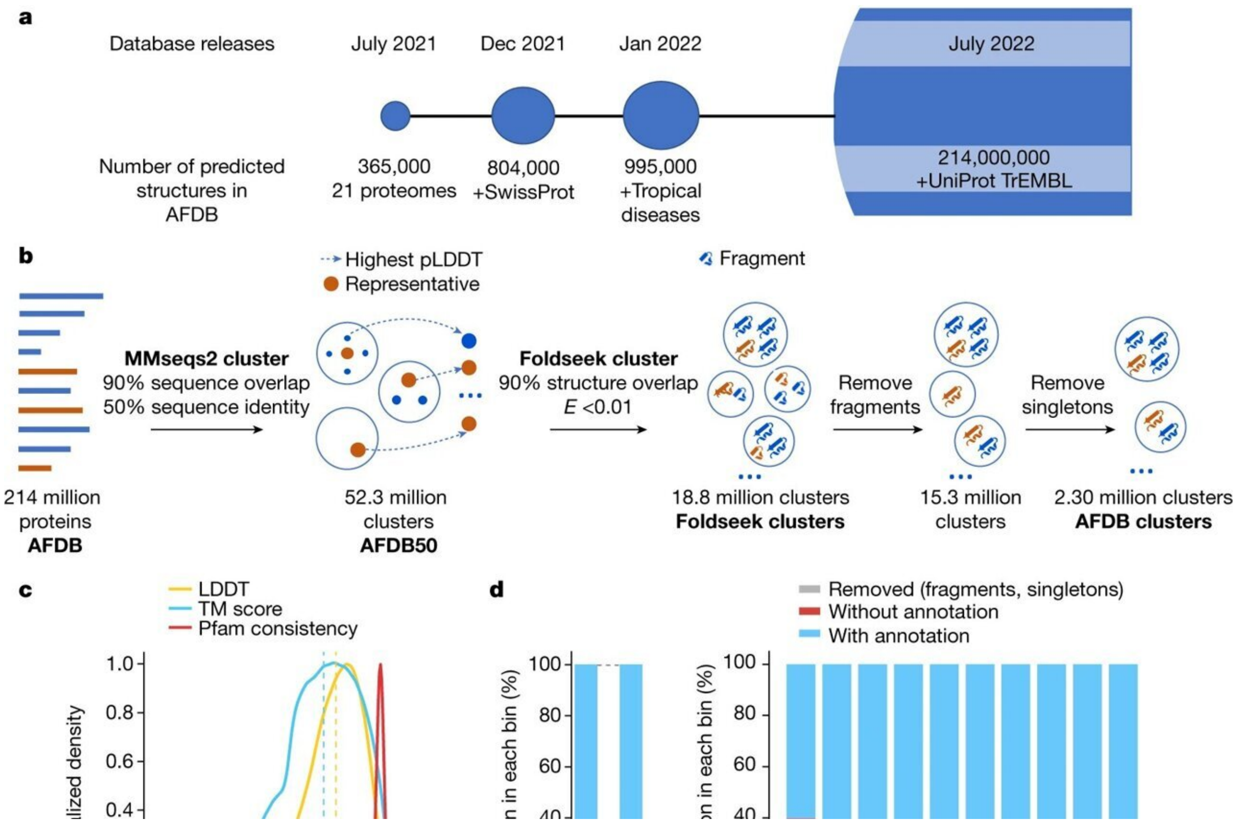
For this study, published in the journal *Nature*, the researchers developed a new algorithm known as Foldseek Cluster that can be used to analyze large sets of protein structures all at once. Foldseek Cluster was applied to the 200 million predicted protein structures in the AlphaFold database, identifying over 2 million unique structural clusters—groups of protein structures that are similar to each other in their three-dimensional shapes. One third of these clusters lack any previous annotations, meaning they had not before been described or categorized.

Bridging the gap in protein science

Proteins are critical to processes that take place in the cell. Understanding protein structure is pivotal for studying their function and evolution. Despite significant advancements in sequence-based predictions of protein structures, computational limitations have made it difficult to study these structures at scale. Foldseek Cluster now enables structural comparisons and clustering at an unprecedented scale, reducing the time for such tasks by several orders of magnitude.

"We've entered a new era in structural biology where computational methods unlock unprecedented access to explore the protein universe," said Martin Steinegger, Assistant Professor at the School of Biological Sciences Seoul National University.

"We estimated that clustering all structures with established methods would have taken a decade when compared to the five days it took using our new method, Foldseek Cluster. Our algorithm can sift through millions of predicted protein structures in the AlphaFold database and cluster them based on their 3D shapes. This acceleration in computational power doesn't just make things faster; it makes things possible."



The AFDB, structural clustering workflow and summary of the clusters. **a**, The AFDB started as a collaborative effort between EMBL-EBI and DeepMind in

2021. The database grew in multiple stages, with the latest version of 2022 containing over 214 million predicted protein structures and their confidence metrics. **b**, A two-step approach was used to cluster proteins in the database. First, MMseqs2 was used to cluster 214 million UniProtKB protein sequences (AFDB) on the basis of 50% sequence identity and 90% sequence overlap, resulting in a reduction of the database size to 52 million clusters (AFDB50). For each cluster, the protein with the highest pLDDT score was selected as the representative. Next, using Foldseek, the representative structures were clustered into 18.8 million clusters (Foldseek clusters) without a sequence identity threshold, but still enforcing a 90% sequence overlap and an *E*-value of less than 0.01 for each structural alignment. As the last step, we removed all sequences labeled as fragments from the clustering, ending up with 2.30 million clusters with at least two structures (AFDB clusters). **c**, AFDB cluster structural and Pfam consistency. Our clusters have a median LDDT of 0.77 and a median TM score of 0.71 across all clusters and 66.5% of clusters with Pfam annotations are 100% consistent. **d**, Summary of sequences and clusters with and without annotation (left) and the relationship of cluster sizes to annotation (right). From left to right, each bin occupies AFDB clusters at rates of 12.24%, 10.59%, 9.20%, 10.07%, 10.46%, 10.05%, 9.04%, 9.20%, 9.19% and 9.96%. Credit: *Nature* (2023). DOI: 10.1038/s41586-023-06510-w

Protein evolution and immunity

The study also delves into the evolutionary implications of these clusters. While most clusters are ancient in origin, around 4% appear to be species-specific. This offers new insights into evolutionary phenomena such as de novo gene birth—when new genes arise from non-coding regions of the genome. The work also illustrates several examples of evolutionary relationships that could enrich our understanding of protein function across different species, including their role in human immunity.

"This work isn't just about making comparisons more efficiently, it's about gaining new insights into the evolutionary history of proteins," said Pedro Beltrao, Associate Professor at the Institute of Molecular Systems Biology, ETH Zurich.

"One of the most interesting findings from this study is our detection of structural similarities between human immune system proteins and those found in bacteria. This suggests that proteins involved in the immune system may have ancient evolutionary origins that we share with bacterial species. If true, this could reshape our understanding of immunity. Our research not only advances current knowledge but also lays out a roadmap for future investigations into the mysteries of protein function and evolution."

Improving the AlphaFold database functionality

As the AlphaFold database and other life science databases continue to grow there is a significant need to help users sift through the vast amount of data while reducing the computational costs of analyzing and managing these data. Approaches such as the Foldseek Cluster algorithm, that is scalable to billions of structures, will be invaluable in helping researchers navigate this wealth of information.

"Foldseek Cluster is more than just a technological advancement; it's an enhancement that elevates the entire AlphaFold database experience for researchers worldwide," said Sameer Velankar, Team Leader at EMBL-EBI.

"With the explosion of predicted protein structures we have in AFDB, managing and navigating these data efficiently has been a significant challenge," he continued. "Foldseek Cluster has revolutionized this process. We are working on integrating FoldSeek clusters into AFDB to streamline the analysis of large sets of protein structures and make it easier for our user community to find exactly what they're looking for."

More information: Martin Steinegger, Clustering-predicted structures at the scale of the known protein universe, *Nature* (2023). DOI: 10.1038/s41586-023-06510-w.
www.nature.com/articles/s41586-023-06510-w

Provided by European Molecular Biology Laboratory

Citation: Revealing the secrets of protein evolution using the AlphaFold database (2023, September 13) retrieved 28 April 2024 from https://phys.org/news/2023-09-revealing-secrets-protein-evolution-alphafold.html