GENOME WATCH

Machine learning sheds light on microbial dark proteins

Aeron Tynes Hammack and Crysten E. Blaby-Haas

This month's Genome Watch highlights the recent use of machine learning to uncover functional 'dark matter' in the microbial protein universe.

Metagenomics projects have revealed more than 8 billion non-redundant microbial protein sequences from across the Earth's biosphere[1]. Of these, 1.17 billion proteins do not have recognizable homologues in any of the more than 100,000 reference genomes available[1]. Understanding the function of these microbial proteins is a daunting task. Fortunately, machine learning (ML), has recently achieved unprecedented accuracy in modelling complex biological data and making predictions. At the forefront of these advancements are ML-based approaches that can confidently predict atomic-level protein structures for many (but not all) amino acid sequences. . A recent study used the ESMFold predictor[2] that takes advantage of a large language model (LLM) to quickly generate 617 million structures from the European Bioinformatics Institute (EBI)'s MGnify metagenome database. Of the total predictions, around 36% of them were considered to have high confidence. The computed structures were made available to the community through the Evolutionary Scale Modelling (ESM) Metagenomic Atlas database. Many of the predicted proteins are from unculturable and/or genetically intractable microorganisms; therefore, such structures could aid microbiologists in devising hypotheses about the molecular function of specific proteins. However, the size of the resulting dataset limits the extent to which these structures can be analysed en masse. To overcome this challenge, three recent studies have used protein clustering to narrow down the data that needs to be examined to find novel proteins and structures.

Two of these studies analysed 215 million precomputed structures in the AlphaFold Database (AFDB)[3,4]. One of the works[3] developed a method, Foldseek cluster, which uses a combination of ultra-fast sequence and structural aligners to cluster sequences and then cluster representative structures. After quality filtering, this approach reduced the AFDB protein space to 2.3 million structures. Of these, a little over 700,000 protein clusters (~30%) do not have matches to experimentally determined structures and could not be functionally annotated with Pfam or TIGRFAM annotations. However, in several cases, structural similarity to annotated clusters, including leveraging human proteins to inform on bacterial proteins, enabled functional predictions for several bacterialproteins in 'dark' —that is, poorly annotated— clusters.

The other study[4] used precomputed clusters from the UniProt database to define a set of 6 million representative structures. These sequences were then used to build an interactive sequence similarity network, where nodes were given an estimated 'brightness' score based on the ability to assign a given cluster to an experimentally characterized protein family. A deeper analysis into 'dark' areas of the network led to the identification and subsequent experimental verification of a new family of toxin proteins that function in toxin-antitoxin systems in bacteria.

A third study[1] analysed the 8 billion sequences encoded by metagenomes and metatranscriptomes stored in the Integrated Microbial Genomes and Microbiomes (IMG/M) database. This large dataset was first reduced by removing proteins with similarity to Pfam or sequences encoded by reference genomes. The resulting sequences were clustered using a graph-based approach. , Nearly 100,000 protein families, named novel metagenome protein families (NMPFs), were idnetified. The use of AlphaFold and the clustering of NMPFs based on structure resulted in ~4,000 unique predicted structures. Although not apparent at the sequence level, structural similarity placed 62% of protein structures in a known family.

In sum, these studies identified new protein families and demonstrated the value of structural similarity in identifying family association, especially for highly divergent sequences. Although defining the structure of an uncharacterized protein does not necessarily reveal its function, structural similarity to characterized proteins can provide an invaluable inference when seeking to decode the vast functional information contained within microbial genomes.

Aeron Tynes Hammack[1] and Crysten E. Blaby-Haas[1,2]*
[1]Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA; [2]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
**e-mail: cblaby@lbl.gov*

https://doi.org/10.1038/s41579-XXX-XXXX-X

1. Pavlopoulos, G. A. et al. Unraveling the functional dark matter through global metagenomics. *Nature* **622,** 594–602 (2023).
2. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
3. Barrio-Hernandez, I. et al. Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637–645 (2023).
4. Durairaj, J. et al. Uncovering new families and folds in the natural protein universe. *Nature* **622**, 646–653 (2023).

Credit: Philip Patenall/Springer Nature Limited