Molecular Microbiology 2024

Bioinformatics Protocol 10

Beyond BLAST and HMMer: Sequence-based 3D structural prediction

This protocol covers the use of ESMFold to predict protein folding from sequences, the use of Uniprot to find sequences previously folded using AlphaFold2.

1. Open the fasta amino acid file (`.faa` file) from your Prokka annotation and search for a protein sequence annotated as `hypothetical protein` (a protein with no known function). Looking in Artemis for a gene encoding a hypothetical protein that is in the same operon as some genes with a known function might help to narrow down a hypothetical protein that is more interesting for you, but it really doesn't matter what protein sequence you choose.

2. Visit [https://esmatlas.com/resources?action=fold] and paste the protein sequence in the input box then click on the picture of the magnifying glass. Ensure the drop-down menu on the left remains at "Fold sequence". ESMfold will now attempt to fold this sequence to a 3D protein structure *in silico*.

3. Investigate the 3D sequence by rotating and zooming. How much of the input protein sequence was folded? Also note the colours indicating prediction confidence.

4. Click on the blue **PDB file** button to download this 3D structure as a file.

5. Now let's use this PDB file as a query for searching databases of 3D structures. Visit https://search.foldseek.com/search. Click on **UPLOAD PDB** and then on **SEARCH**.

6. You will now see lists of hits based on the database searched. To visualise a hit click on the icon of three horizontal lines to the right of a hit - here you will see a 3d visualisation of how the structures are aligned and a linear sequence alignment.

Some of these databases are based on automated predictions of 3D structure (such as AFDB_PROTEOME and MGNIFY_ESM30) - such larger databases are more likely to have good hits, but here we're comparing a computationally predicted structure to another computationally predicted structure so it's more likely to be incorrect. Experimentally-derived databases (like PDB100) are more reliable, but smaller (so there's less chance of getting a good hit).

7. Now let's see if there's an Alphafold-predicted structure for the protein you have chosen. Visit https://www.uniprot.org/blast and paste your protein sequence in the box labeled `Protein or nucleotide sequence(s) in FASTA format`. Then click **Run BLAST** in the bottom right-hand corner. Your sequence will be BLASTed against the UniProt database.

8. Check the organism name, sequence identity, and e-values for your results. Hopefully you should a sequence out there that is very similar to your

sequence. Click on its acession number in the **Entry** column. On the hit's Uniprot page, scroll down to the **Structure** section to see a structure for this protein predicted by AlphaFold2. How does this compare to the structure produced by ESMfold?

9. Click on the accession number next to **AlphaFoldDB** in the structure section, you will now be taken to the **AlphaFoldDB** website. Click **PDB File** to download the PDB file for this structure.

10. Now submit this structure predicted by AlphaFold to FoldSeek. Does this reveal the protein's function any better than the ESMfold-predicted structure?