Lab Session 6: Prediction of protein structure

Date: Feb 2023

Objective: Derive the structure of proteins using comparative modelling methods.

Note: 3D models obtained by comparative modelling are approximations.

Step1: Select any sequence to model.

Model 1: Major Cold shock protein from *Staphylococcus aureus*

> The Major Cold Shock protein (CspA) is present in many bacteria and is induced by cold shock. This protein binds to single stranded nucleic acids and it is believed to be involved in the stability of mRNA upon these conditions (as a RNA "chaperone"). Our objective here is to structurally characterise the CspA from *Staphylococcus aureus*, a human pathogen that has its protein sequence determined, but for which no experimental structural information exists.
>
> The sequence can be found in:
> http://www.uniprot.org/uniprot/Q6G9F9
> Staphylococcus aureus (strain MSSA476) [TaxID:282459]
> MKQGTVKWFN AEKGFGFIEV EGENDVFVHF SAINQDGYKS LEEGQAVEFE VVEGDRGPQA
> ANVVKL

Model 2: Dihydrofolate reductase from *Vibrio cholerae*

> Dehydrofolate reductase is an enzyme that catalyses the conversion of folic acid (folate) into tetrahydrofolic acid, a fundamental molecule in the synthesis of DNA precursors. The homology with proteins in the PDB database is much lower than the one found for CpsA from *S. aureus*, presenting a more challenging problem for comparative modelling procedures. The sequence of this protein is 157 residues long and can be found at http://www.uniprot.org/uniprot/Q7BN39 :
>
> MKLSLMAAIS KNGVIGNGPD IPWSAKGEQL LFKAITYNQW LLVGRKTFES
> MGALPNRKYA VVTRSSFTSS DENVLVFPSI DEALNHLKTI TDHVIVSGGG
> EIYKSLIDKV DTLHISTIDI EPEGDVYFPE IPSSFRPVFS QDFVSNINYS
> YQIWQKG

Assuming that they are unknown sequences, can you predict the structure with Alpha fold.

Predict the structure with Alphafold
https://alphafold.ebi.ac.uk/

Step by step procedure:

1. Obtain the sequence of the protein of interest, e.g. at UniProt. Click on the FASTA button above the sequence

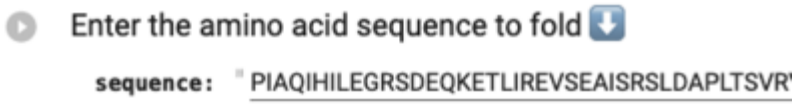
in UniProt. Copy only the sequence, excluding the FASTA header line that begins with ">".

2. Login with a google account at AlphaFold2_advanced. You can register for a free gmail account to use for login.

Link:

https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/beta/AlphaFold2_advanced.ipynb

3. Paste in your sequence, making sure to completely replace the default sequence:



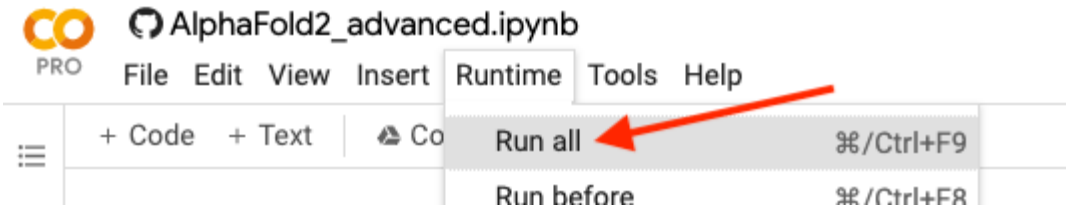
This input slot can accept sequences >1,000 amino acids, even though it is only one line. Sequence lengths of ~1,000 amino acids, or longer, may cause the Colab to fail, but can be predicted by submitting in two halves.[4]

4. Enter a jobname in the slot below the sequence slot. The results.zip filename will begin with this jobname (but none of its contents include the jobname).

5. Scroll down to the section titled *run alphafold*, subsection *Sampling options*:

- **num_models**, the number of models to be predicted, is 5 by default. You could reduce this to 3 if you are in a hurry.
- **max_recycles**: Set this to 48 (or at least 12). The actual number of "recycles" performed will stop when the model has converged to the specified tolerance. The default of 3 recycles is often not enough for an optimal result.
- **tol** (tolerance): Set this to 0.5 Å (or 1.0 to get a faster result). When a prediction differs from the previous "recycle" prediction by less than this value (RMSD in Å between alpha carbons), the recycles will stop.
- **num_samples** (random seeds): Leave this at 1. Beware that if you increase this above 1, you will generate a number of models equal to the product of this value times num_models. This will proportionally increase the time to complete a result.

6. Open the Runtime menu at the very top of the page, and select **Run all**.



Don't worry about the "Warning". It is just Google's disclaimer that they did not write the code you are about to execute. **Click *Run anyway***.

NOTES:

1. NO GAPS BETWEEN AMINOACIDS IN THE SEQUENCE
2. GIVE THE JOBNAME as your-first-name1. The results are archived as compressed (RAR) file. Open the directory

WORK ELEMENTS

1. Choose one predicted model from the directory
2. Go to PDB and download the solved structure; for example **1MJC.pdb** for Model 1. **4DFR.pdb** for Model 2.
3. Compare the structures. Calculate RMSD between the predicted structure and the PDB structure
4. Comment on the efficiency of AlphaFold.

Please follow the link for step by step procedure

https://proteopedia.org/wiki/index.php/How_to_predict_structures_with_AlphaFold#:~:text=the%20following%20instructions.-,Submitting%20A%20Sequence,skip%20to%20Interpreting%20Results%20below.
