List of CBSB3 miniprojects

Yong Wang

Project 1. Protein structure prediction using AlphaFold2

Protein structure prediction, the inference of the three-dimensional structure (secondary and tertiary structure) from the amino acid sequence (primary structure) is a longstanding problem in computational biology. It is one of the most important goals pursued by computational biologists. Protein structure prediction also has a big impact on drug design and biotechnology (for example, in the design of novel biomaterials and enzymes). Therefore protein structure prediction has been evolving as one of the standard tools of the biology community.

In this project, I will provide two small proteins, protein #1 and protein #2 to predict.

The sequence of protein#1:
GSSHHHHHHSSGMDALNSKEQQEFQKVVEQKQMKDFMRLYSNLVERCFTDCVNDF
TTSKLTNKEQTCIMKCSEKFLKHSERVGQRFQEQNAALGQGLGR

The sequence of protein#2:
GSSHHHHHHSSGGSFLGFGGGQPQLSSQQKIQAAEAELDLVTDMFNKLVNNCYKKC
INTSYSEGELNKNESSCLDRCVAKYFETNVQVGENMQKMGQSFNAAGKF

Tasks:

1. Figure out which proteins they are by searching the sequences.
2. Do protein structure prediction online using cloud-based Alphafold2 Colab notebook (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb).
3. Visualize the protein structure using the most popular protein visualization software PyMol/Chimera and compare the predicted structures with available NMR/crystal structures in the Protein Data Bank if there are.
4. Understand the per-residue confidence metric and the predicted aligned error.
5. Compare the structural difference between protein#1 and protein#2, as well as the difference between your predicted structures with the available AF2 structures if there are. Report the RMS deviation.

Some useful links:

- https://alphafold.ebi.ac.uk/
- https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb
- https://blast.ncbi.nlm.nih.gov/Blast.cgi
- https://pymol.org/
- https://www.ks.uiuc.edu/Research/vmd/
- https://www.cgl.ucsf.edu/chimera/

Project 2. Protein-protein interaction and complex structure prediction

Besides a well-defined structure, protein functions by forming specific interactions with other molecules, such as proteins, small molecules, DNA, etc. While the structure of a single protein alone can now be predicted to high accuracy thanks to the advances of AI-based protein structure prediction methods, such as AlphaFold2 and RosettaFold, the prediction of protein-protein interactions and their complexes with other proteins remains a challenge.

In this project, I provide three proteins that can interact with each other. It is known that they could form multimer with multiple copies and with variable stoichiometry. Their sequences are given below:

The sequence of protein#1:
MDALNSKEQQEFQKVVEQKQMKDFMRLYSNLVERCFTDCVNDFTTSKLTNKEQTCI
MKCSEKFLKHSERVGQRFQEQNAALGQGLGR

The sequence of protein#2:
GSFLGFGGGQPQLSSQQKIQAAEAELDLVTDMFNKLVNNCYKKCINTSYSEGELNK
NESSCLDRCVAKYFETNVQVGENMQKMGQSFNAAGKF

The sequence of protein#3:
MSFFLNSLRGNQEVSQEKLDVAGVQFDAMCSTFNNILSTCLEKCIPHEGFGEPDLTK
GEQCCIDRCVAKMHYSNRLIGGFVQTRGFGPENQLRHYSRFVAKEIADDSKK

Tasks:

1. Figure out which proteins they are by searching the sequences.
2. Predict the multimer structures of the complex between protein#1 and protein#2 using the AlphaFold Multimer colab code (https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb).
3. Based on the predicted structures, tell how many proteins to form the complex and what is the optimal copy number of each protein in the complex.
4. Now taking protein#3 in, and redo the structure prediction of the complex, and play with different stoichiometries (e.g. three copie of #1, two copies of #2 and one #3).
5. Visualize the protein structure using the most popular protein visualization software PyMol/Chimera.

This project is a bit more challenging than project#1. If it is too difficult for you to finish all tasks, it is also ok to do tasks 1), 2) and 5).

Some useful materials and links:

- https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb
- https://wenmr.science.uu.nl/
- https://pymol.org/
- https://www.cgl.ucsf.edu/chimera/
- Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S and Steinegger M. ColabFold - Making protein folding accessible to all. Nature Methods (2022) doi: 10.1038/s41592-022-01488-1

- Jumper et al. "Highly accurate protein structure prediction with AlphaFold." Nature (2021) doi: 10.1038/s41586-021-03819-2
- Evans et al. "Protein complex prediction with AlphaFold-Multimer." BioRxiv (2022) doi: 10.1101/2021.10.04.463034v2

Project 3. Modelling mutational effects on protein structure

The stability of protein 3D structure (either as its own or in a complex) is one of the most important features to ensure the protein is functional. You will explore mutational effects based on a protein, namely T4 lysozyme, on its folding stability.

Tasks:
1) Use PDB database to find right protein structure(s) to work with (https://www.rcsb.org) Check “206L”, “3DMV” and “2LC9”, what differences do you see? Are there more of the same protein structure? Choose one structure to work on and tell why you chose that.
2) Explore the built-in structure viewer.
3) Download PDB files and explore the information.
4) By carefully inspecting the structure, choose one candidate residue to mutate and tell why.
5) Mutate it to other residues (virtual mutational scan), remodel the protein structure and obtain the structure stability data.
6) Check the $\Delta\Delta G_{folding}$ (compared to the wild type) and see which mutations destabilize the protein most.
7) If you are interested, keep moving on and do ligand (could be benzene) docking on the mutants.

Some useful links:

- https://pymol.org/
- https://www.cgl.ucsf.edu/chimera/
- ‘FoldX’ https://foldxsuite.crg.eu/products#foldx
- https://elifesciences.org/articles/17505
- https://www.pnas.org/doi/epdf/10.1073/pnas.2106195118