Indraprastha Institute of Information Technology Delhi (IIITD)
Department of Computational Biotechnology

BIO213 – Introduction to Quantitative Biology

ASSIGNMENT-3 (April 25, 2022)

Objective of this assignment is to get you acquainted with Modeller (https://salilab.org/modeller/), one of the most popularly used homology modelling tool.

The sequence given below is a part of human E3 ubiquitin-protein ligase for which structure has not been solved and therefore you will be developing a structural model for the same using the following instructions. During the modelling process answer the questions given below.

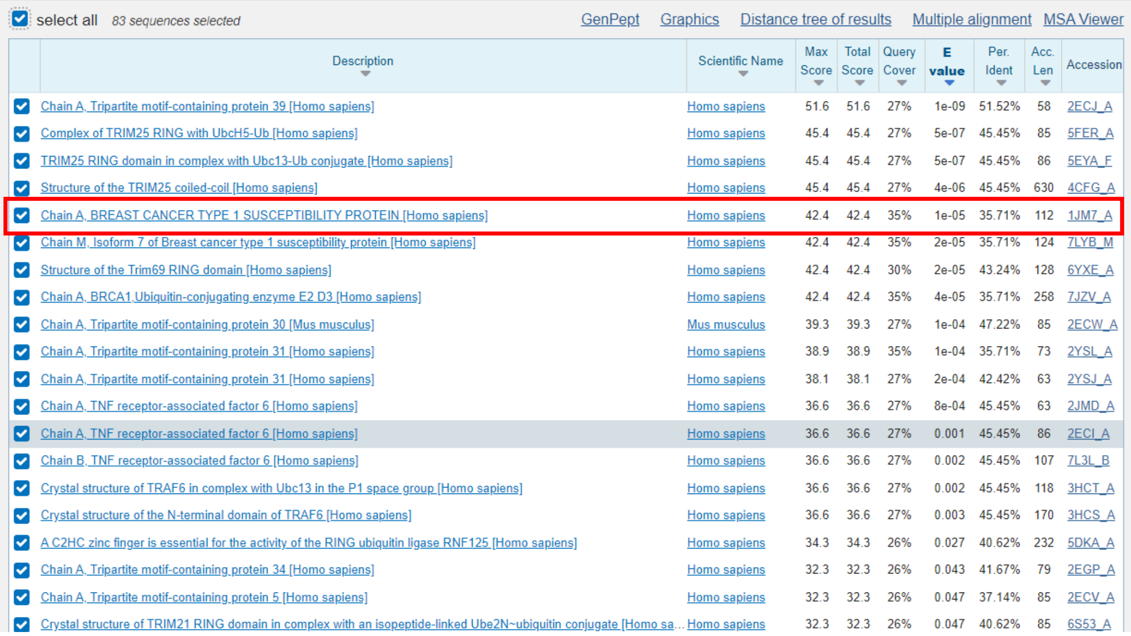
```
>protein
MALPAGPAEAACALCQRAPREPVRADCGHRFCRACVVRFWAEEDGPFPCPECADDCWQRA
VEPGRPPLSRRLLALEEAAAAPARDGPASEAALQLLCRADAGPLCAACRMAAGPEPPEWE
```

STEP 1. *Search for homologous proteins with solved structure to be used as template*

Firstly, use the NCBI blastp program available at the link https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp. Input the given FASTA sequence, choose 'Protein Data Bank proteins (pdb)' as the database and blastp as the algorithm. Now submit the job, wait for the results and then analyze it carefully.

Question 1. Which sequence can serve as the best template for modelling the E3 ubiquitin-protein ligase structure? Give reason for the same. Use the parameters like score, identity, similarity, query coverage, E-value, etc. to make the choice. **(10 marks)**

Answer:



☑ select all *83 sequences selected* GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

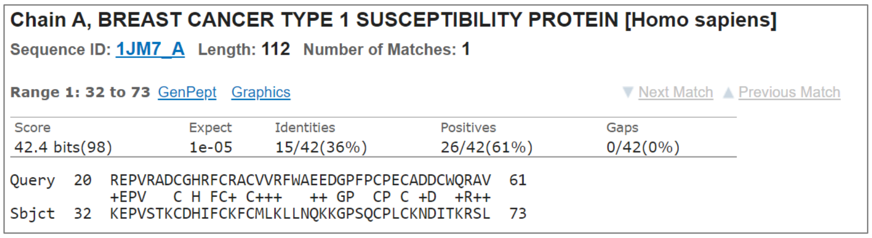
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Chain A, Tripartite motif-containing protein 39 [Homo sapiens]	Homo sapiens	51.6	51.6	27%	1e-09	51.52%	58	2ECJ_A
☑	Complex of TRIM25 RING with UbcH5-Ub [Homo sapiens]	Homo sapiens	45.4	45.4	27%	5e-07	45.45%	85	5FER_A
☑	TRIM25 RING domain in complex with Ubc13-Ub conjugate [Homo sapiens]	Homo sapiens	45.4	45.4	27%	5e-07	45.45%	86	5EYA_F
☑	Structure of the TRIM25 coiled-coil [Homo sapiens]	Homo sapiens	45.4	45.4	27%	4e-06	45.45%	630	4CFG_A
☑	Chain A, BREAST CANCER TYPE 1 SUSCEPTIBILITY PROTEIN [Homo sapiens]	Homo sapiens	42.4	42.4	35%	1e-05	35.71%	112	1JM7_A
☑	Chain M, Isoform 7 of Breast cancer type 1 susceptibility protein [Homo sapiens]	Homo sapiens	42.4	42.4	35%	2e-05	35.71%	124	7LYB_M
☑	Structure of the Trim69 RING domain [Homo sapiens]	Homo sapiens	42.4	42.4	30%	2e-05	43.24%	128	6YXE_A
☑	Chain A, BRCA1 Ubiquitin-conjugating enzyme E2 D3 [Homo sapiens]	Homo sapiens	42.4	42.4	35%	4e-05	35.71%	258	7JZV_A
☑	Chain A, Tripartite motif-containing protein 30 [Mus musculus]	Mus musculus	39.3	39.3	27%	1e-04	47.22%	85	2ECW_A
☑	Chain A, Tripartite motif-containing protein 31 [Homo sapiens]	Homo sapiens	38.9	38.9	35%	1e-04	35.71%	73	2YSL_A
☑	Chain A, Tripartite motif-containing protein 31 [Homo sapiens]	Homo sapiens	38.1	38.1	27%	2e-04	42.42%	63	2YSJ_A
☑	Chain A, TNF receptor-associated factor 6 [Homo sapiens]	Homo sapiens	36.6	36.6	27%	8e-04	45.45%	63	2JMD_A
☑	Chain A, TNF receptor-associated factor 6 [Homo sapiens]	Homo sapiens	36.6	36.6	27%	0.001	45.45%	86	2ECI_A
☑	Chain B, TNF receptor-associated factor 6 [Homo sapiens]	Homo sapiens	36.6	36.6	27%	0.002	45.45%	107	7L3L_B
☑	Crystal structure of TRAF6 in complex with Ubc13 in the P1 space group [Homo sapiens]	Homo sapiens	36.6	36.6	27%	0.002	45.45%	118	3HCT_A
☑	Crystal structure of the N-terminal domain of TRAF6 [Homo sapiens]	Homo sapiens	36.6	36.6	27%	0.003	45.45%	170	3HCS_A
☑	A C2HC zinc finger is essential for the activity of the RING ubiquitin ligase RNF125 [Homo sapiens]	Homo sapiens	34.3	34.3	26%	0.027	40.62%	232	5DKA_A
☑	Chain A, Tripartite motif-containing protein 34 [Homo sapiens]	Homo sapiens	32.3	32.3	26%	0.043	41.67%	79	2EGP_A
☑	Chain A, Tripartite motif-containing protein 5 [Homo sapiens]	Homo sapiens	32.3	32.3	26%	0.047	37.14%	85	2ECV_A
☑	Crystal structure of TRIM21 RING domain in complex with an isopeptide-linked Ube2N~ubiquitin conjugate [Homo sa...	Homo sapiens	32.3	32.3	26%	0.047	40.62%	85	6S53_A

This template covers 35% of the query protein (better than the first 4 hits, which cover only 27% of the query protein), E-value is much less than 0, and the percentage identity is 35.71%.

(If the same template is used give 10 marks, if not 6 marks)

Answers to most of the following questions will vary depending upon the template used. So, evaluate accordingly.

Question 2. Show the alignment of the chosen template with your query protein. Is there any region of the query that is not being covered by the template? If yes, mention the residue numbers. Use the graphical summary on the results page to check if any other sequence can serve as a template for the uncovered region or not. **(5 marks)**



Chain A, BREAST CANCER TYPE 1 SUSCEPTIBILITY PROTEIN [Homo sapiens]

Sequence ID: 1JM7_A Length: 112 Number of Matches: 1

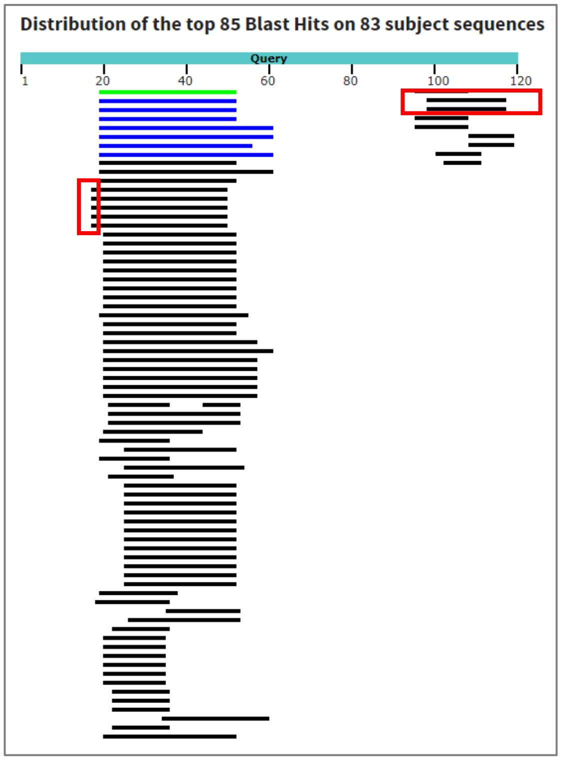
Range 1: 32 to 73 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps
42.4 bits(98)	1e-05	15/42(36%)	26/42(61%)	0/42(0%)

```
Query  20   REPVRADCGHRFCRACVVRFWAEEDGPFPCPECADDCWQRAV  61
            +EPV   C H FC+ C+++   ++ GP  CP C +D  +R++
Sbjct  32   KEPVSTKCDHIFCKFCMLKLLNQKKGPSQCPLCKNDITKRSL  73
```

(1.5 marks)

Region not covered by the template: 1-19 and 62-120 amino acids (1 mark)



(2.5 marks)

Multiple templates (marked with red box on the left) are available to cover a few more amino acids to the left of the aligned region, and two more templates (marked with red box on the right) can cover region from 96-117 amino acids.

STEP 2. ***Retrieve the PDB structure of the chosen template***

Note the Accession of the chosen template from the blastp results page. The first four characters correspond to the PDB ID, and the character after the underscore represents the chain of PDB aligned to your query sequence. Copy the PDB ID (only the first four characters) and search for the structure in the Protein Data Bank at https://www.rcsb.org/. Open and explore more about the template structure. Download the structure in PDB format.

Question 3. Which experimental method was used to solve this structure? How many total chains are there in the structure? Are the other chains different from the chain of your interest? **(5 marks)**
Answer: Method used: NMR (1.5 marks)
No. of chains: 2 (1.5 marks)
Yes, the two the chains are different (2 marks)

STEP 3. ***Prepare the files in format supported by Modeller to carry out further steps***

You will be using the basic modelling approach here. Follow the tutorial available at https://salilab.org/modeller/tutorial/basic.html. Download the example input and output files, which can be modified and used for the modelling of your protein ligase. As described in the tutorial prepare the protein sequence file in PIR (.ali) format.

Note: build_profile.py and compare.py help in selecting the template as you did in the previous steps. Skip these for now as you have already chosen the template protein.

STEP 4. ***Align the query and template using align2d.py***

This help in the alignment of the query to the template sequence, taking into consideration its structural information as well. Follow the instructions provided in the tutorial for alignment of the two sequences.

STEP 5. ***Model building***

As described in the tutorial, use model-single.py to build the model. By default, 5 models will be generated. Check the summary in the log file to get more information related to the generated models. Choose the best model.

Question 4. Which were the two default parameters or objective functions on which you chose the best model here? Give the significance of both. **(6 marks)**
Answer: The GA341 and DOPE score is being to evaluate or choose the best models. The best GA341 score is near to 1 or 1 as it ranges from 0 to 1, while the model with the lowest DOPE value is considered to be the best.

GA341 score assesses a model by combining a Z-score calculated with a statistical potential function, target-template sequence identity and a measure of structural compactness. The GA341 score ranges from 0 for models that tend to have an incorrect fold to 1 for models that tend to be comparable to at least low-resolution X-ray.

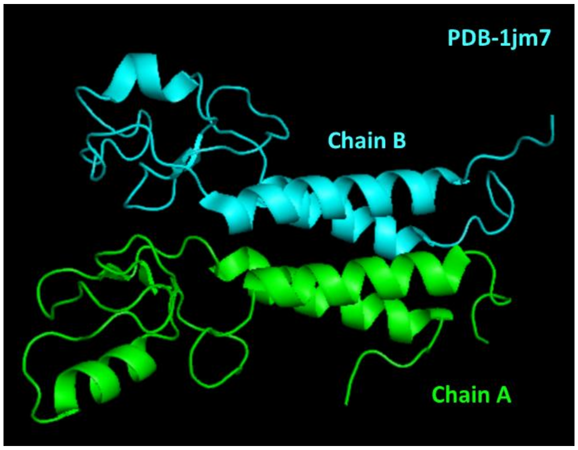
DOPE, or Discrete Optimized Protein Energy, is a statistical potential used to assess homology models in protein structure prediction. DOPE is based on an improved reference state that corresponds to non-interacting atoms in a homogeneous sphere with the radius dependent on a sample native structure; it thus accounts for the finite and spherical shape of the native structures. (If no explanation, give 2 marks)

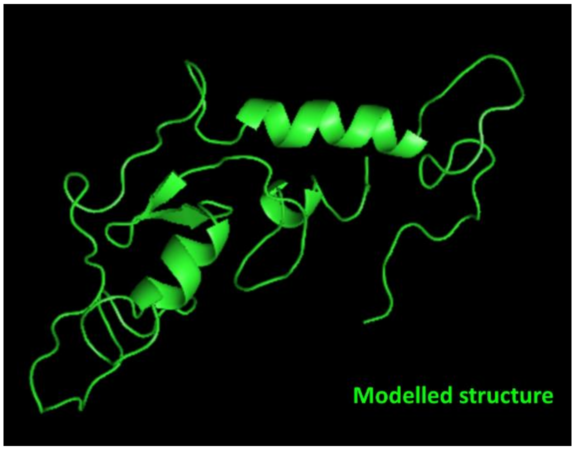
STEP 6. ***Visualization of the developed model***

Though any viewing platform can be used, you will be using Chimera that is freely available at https://www.cgl.ucsf.edu/chimera/download.html. Open the best model in Chimera.

Question 5. Compare the structure of your model to the PDB structure (3D view is also available with each entry). Does it carry similar structural folds? Provide a screenshot of the modelled structure. **(10 marks)**

Answer:





Yes, similar structural folds can be seen in the modelled and PDB structure (respective chains shown in green color).

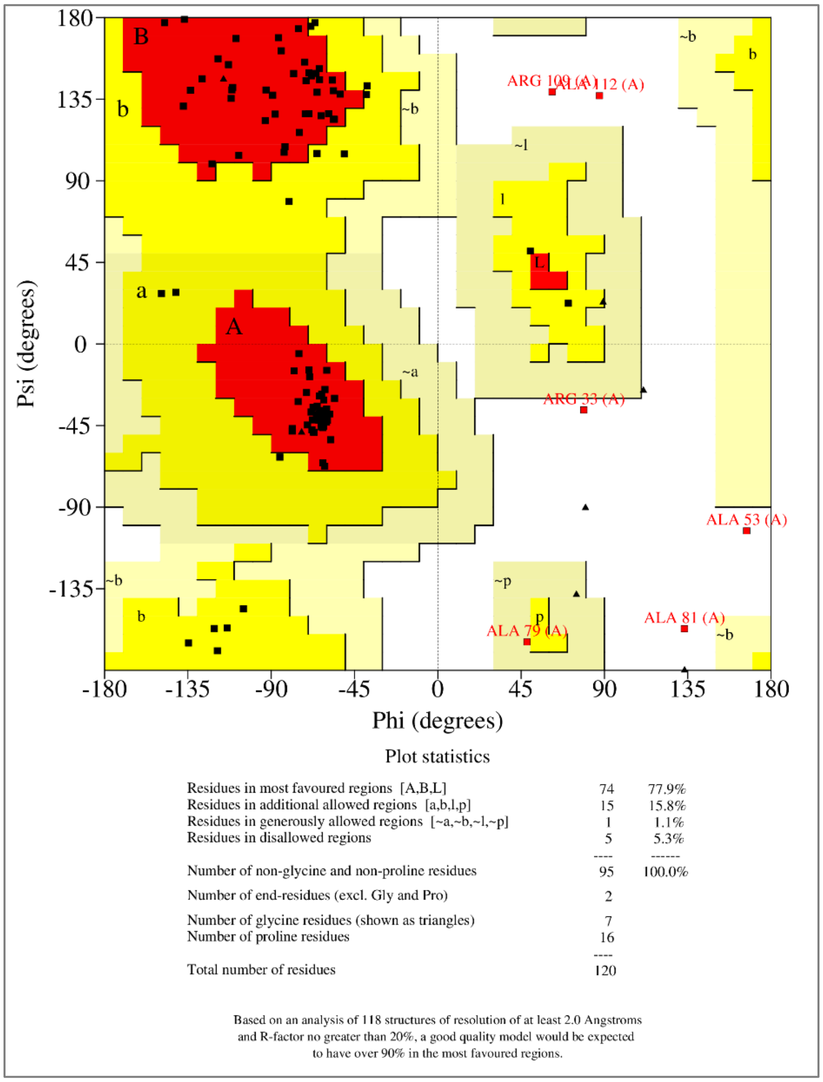
STEP 7. ***Evaluation of the developed model***

We discussed different methods for evaluating the correctness of the structural models. Here you will be generating only the Ramachandran Plot to assess if any amino acids fall in the disallowed region.

Use the PDBsum service (http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/Generate.html) to generate the Ramachandran plot for your model. Upload the .pdb file of the best model generated using Modeller and provide your e-mail ID. The link to the results page will be shared with you once ready. Click on the image depicting the Ramachandran plot to get further information.

<u>Question 6.</u> Provide the plot obtained and briefly discuss the results. Do you think it can be a reliable structure that can be used for other studies? **(10 marks)**

Answer:



Plot statistics

Residues in most favoured regions [A,B,L]	74	77.9%
Residues in additional allowed regions [a,b,l,p]	15	15.8%
Residues in generously allowed regions [~a,~b,~l,~p]	1	1.1%
Residues in disallowed regions	5	5.3%
	----	------
Number of non-glycine and non-proline residues	95	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	7	
Number of proline residues	16	

Total number of residues	120	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

As shown in the statistics, only 5 residues were found in the disallowed region. Though not a very good quality model, it can still be used for further studies after refinement. Also, we have to make sure that the region of the protein to be studied should fall within 21-60 amino acids.

Advanced modelling

<u>Question 7.</u> Do you think using multiple templates (or multi-template homology modelling) could have resulted in a better structure? Justify your answer. **(4 marks)**

Answer: Yes, multi-template modeling could have been a better option. 1jm7 was used as a template because it alone covered 35% part of the query and hence was the best option when single template was to be used. Multi-template modeling provides us a better way of choosing the templates corresponding to different regions of the query proteins, which are then used in parts to generate the complete protein model.