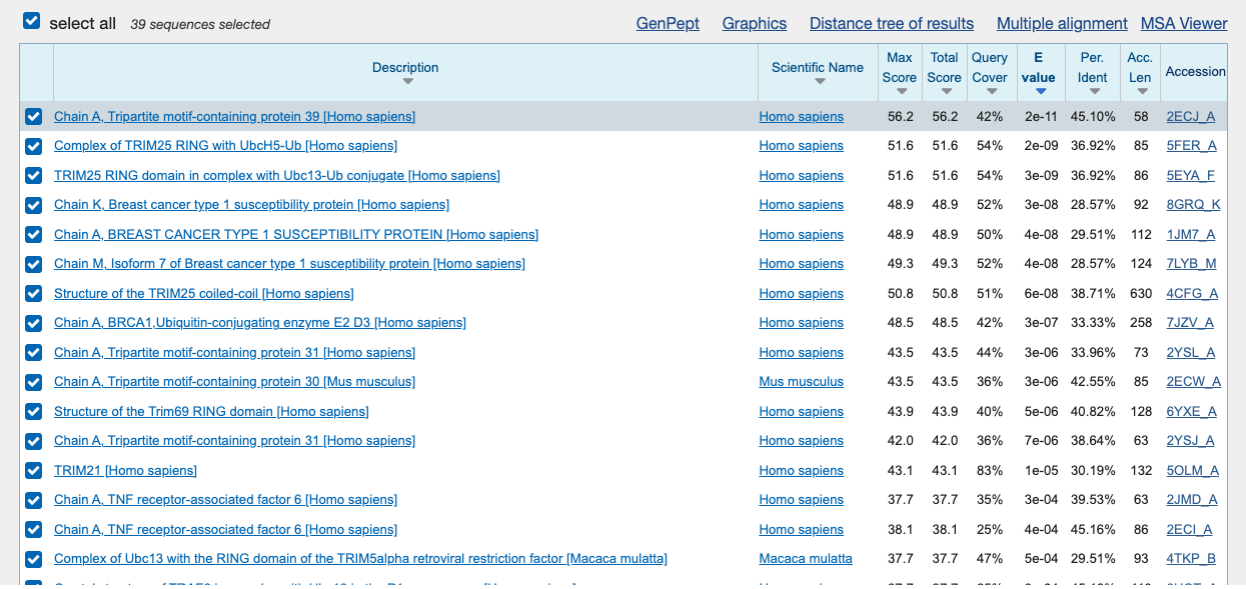
IQB Assignment-3

> Given protein sequence
MALPAGPAEAACALCQRAPREPVRADCGHRFCRACVVRFWAEEDGPFPCPECADDCWQRA
VEPGRPPLSRRLLALEEAAAAPARDGPASEAALQLLCRADAGPLCAACRMAAGPEPPEWE

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

Ans.



select all *39 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	56.2	56.2	42%	2e-11	45.10%	58	2ECJ_A
Complex of TRIM25 RING with UbcH5-Ub [Homo sapiens]	Homo sapiens	51.6	51.6	54%	2e-09	36.92%	85	5FER_A
TRIM25 RING domain in complex with Ubc13-Ub conjugate [Homo sapiens]	Homo sapiens	51.6	51.6	54%	3e-09	36.92%	86	5EYA_F
Chain K, Breast cancer type 1 susceptibility protein [Homo sapiens]	Homo sapiens	48.9	48.9	52%	3e-08	28.57%	92	8GRQ_K
Chain A, BREAST CANCER TYPE 1 SUSCEPTIBILITY PROTEIN [Homo sapiens]	Homo sapiens	48.9	48.9	50%	4e-08	29.51%	112	1JM7_A
Chain M, Isoform 7 of Breast cancer type 1 susceptibility protein [Homo sapiens]	Homo sapiens	49.3	49.3	52%	4e-08	28.57%	124	7LYB_M
Structure of the TRIM25 coiled-coil [Homo sapiens]	Homo sapiens	50.8	50.8	51%	6e-08	38.71%	630	4CFG_A
Chain A, BRCA1,Ubiquitin-conjugating enzyme E2 D3 [Homo sapiens]	Homo sapiens	48.5	48.5	42%	3e-07	33.33%	258	7JZV_A
Chain A, Tripartite motif-containing protein 31 [Homo sapiens]	Homo sapiens	43.5	43.5	44%	3e-06	33.96%	73	2YSL_A
Chain A, Tripartite motif-containing protein 30 [Mus musculus]	Mus musculus	43.5	43.5	36%	3e-06	42.55%	85	2ECW_A
Structure of the Trim69 RING domain [Homo sapiens]	Homo sapiens	43.9	43.9	40%	5e-06	40.82%	128	6YXE_A
Chain A, Tripartite motif-containing protein 31 [Homo sapiens]	Homo sapiens	42.0	42.0	36%	7e-06	38.64%	63	2YSJ_A
TRIM21 [Homo sapiens]	Homo sapiens	43.1	43.1	83%	1e-05	30.19%	132	5OLM_A
Chain A, TNF receptor-associated factor 6 [Homo sapiens]	Homo sapiens	37.7	37.7	35%	3e-04	39.53%	63	2JMD_A
Chain A, TNF receptor-associated factor 6 [Homo sapiens]	Homo sapiens	38.1	38.1	25%	4e-04	45.16%	86	2ECI_A
Complex of Ubc13 with the RING domain of the TRIM5alpha retroviral restriction factor [Macaca mulatta]	Macaca mulatta	37.7	37.7	47%	5e-04	29.51%	93	4TKP_B

This template (i.e. Chain A, Tripartite motif-containing protein 39) covers 42% of query protein (>30%), E-value is the least, and the Percentage Identity is 45.10%.
E-value - 2e-11
Score - 56.2
Query Coverage - 42%

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.

Ans.

Chain A, Tripartite motif-containing protein 39 [Homo sapiens]

Sequence ID: 2ECJ_A **Length:** 58 **Number of Matches:** 1

Range 1: 8 to 58 GenPept Graphics ▼ Next Match ▲ Previous Match

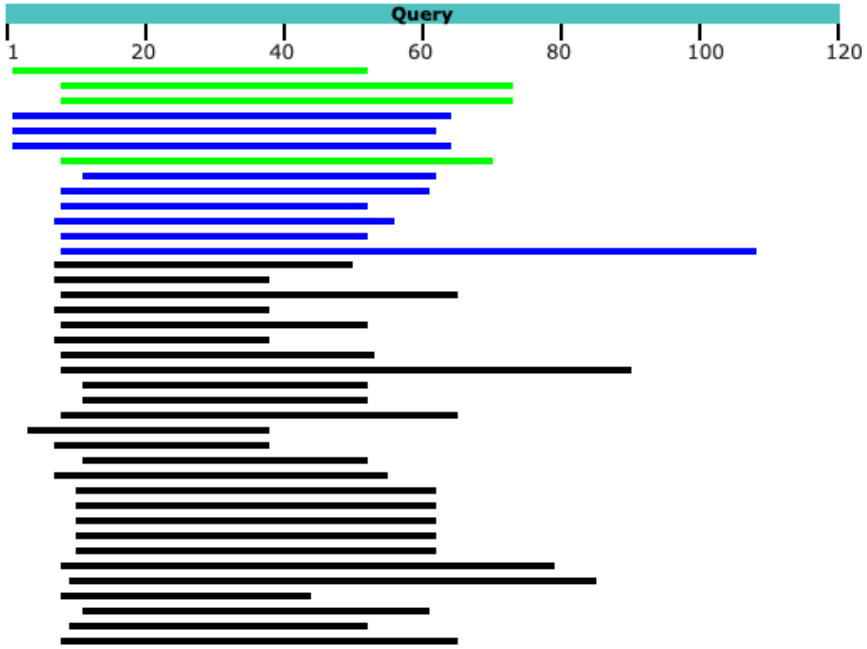
Score	Expect	Method	Identities	Positives	Gaps
56.2 bits(134)	2e-11	Compositional matrix adjust.	23/51(45%)	33/51(64%)	0/51(0%)

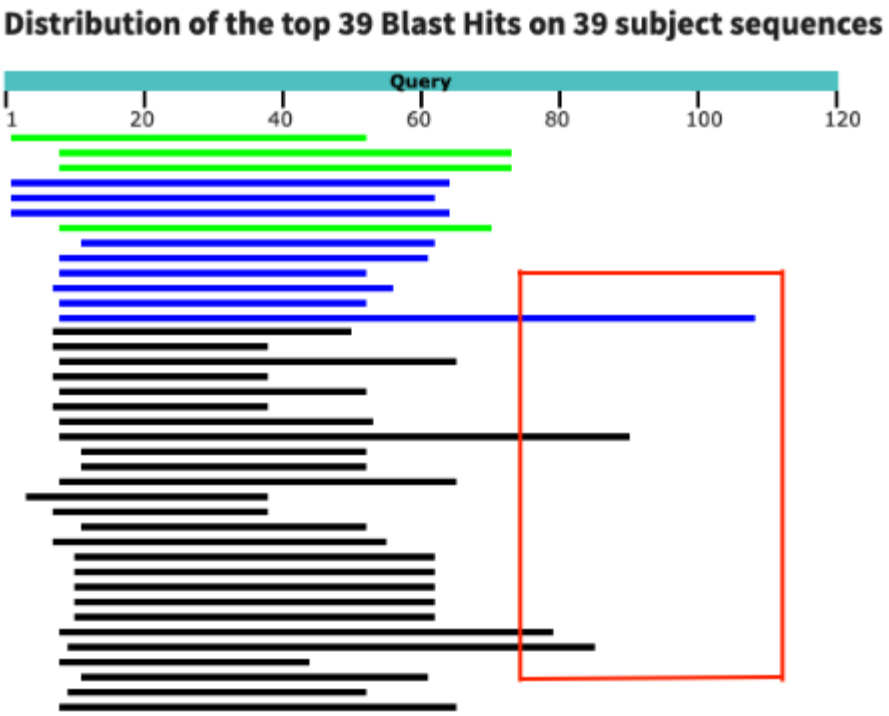
```
Query  2   ALPAGPAEAACALCQRAPREPVRADCGHRFCRACVVRFWAEEDGPFPCPEC  52
           AL     EA+C++C    +EPV  +CGH FC+AC+ R+W + +  FPCP C
Sbjct  8   ALENLQVEASCSVCLEYLKEPVIIECGHNFCKACITRWWEDLERDFPCPVC  58
```

The region not covered by the template is from 1 and 53-120 amino acids.

Other templates like TRIM21 can be used to cover the uncovered regions from 53-108.
It is not used here because of the low percentage identity value.

Distribution of the top 39 Blast Hits on 39 subject sequences





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?
Ans.

Method: Solution NMR (Nuclear Magnetic Resonance)
There is only one chain in the asymmetric unit of the crystal structure.
Only a unique chain is observed.

Question 4. Which were the two default parameters or objective functions on which you chose the best model here? Give the significance of both.
Ans.

GA341 and DOPE score is used to evaluate the best models.

Generated models:

```
>> Summary of successfully produced models:
Filename                          molpdf     DOPE score    GA341 score
----------------------------------------------------------------------
protein.B99990001.pdb          878.04004    -5300.28320        0.55137
protein.B99990002.pdb          738.08984    -5434.99170        0.99407
protein.B99990003.pdb          783.82684    -5385.80811        0.99436
protein.B99990004.pdb          789.58148    -5492.69385        0.85361
protein.B99990005.pdb          769.27081    -5416.01465        0.66993
```

SA341 score is near to 1, and the lower DOPE value is considered best. Hence I used the 3rd protein model for evaluation.

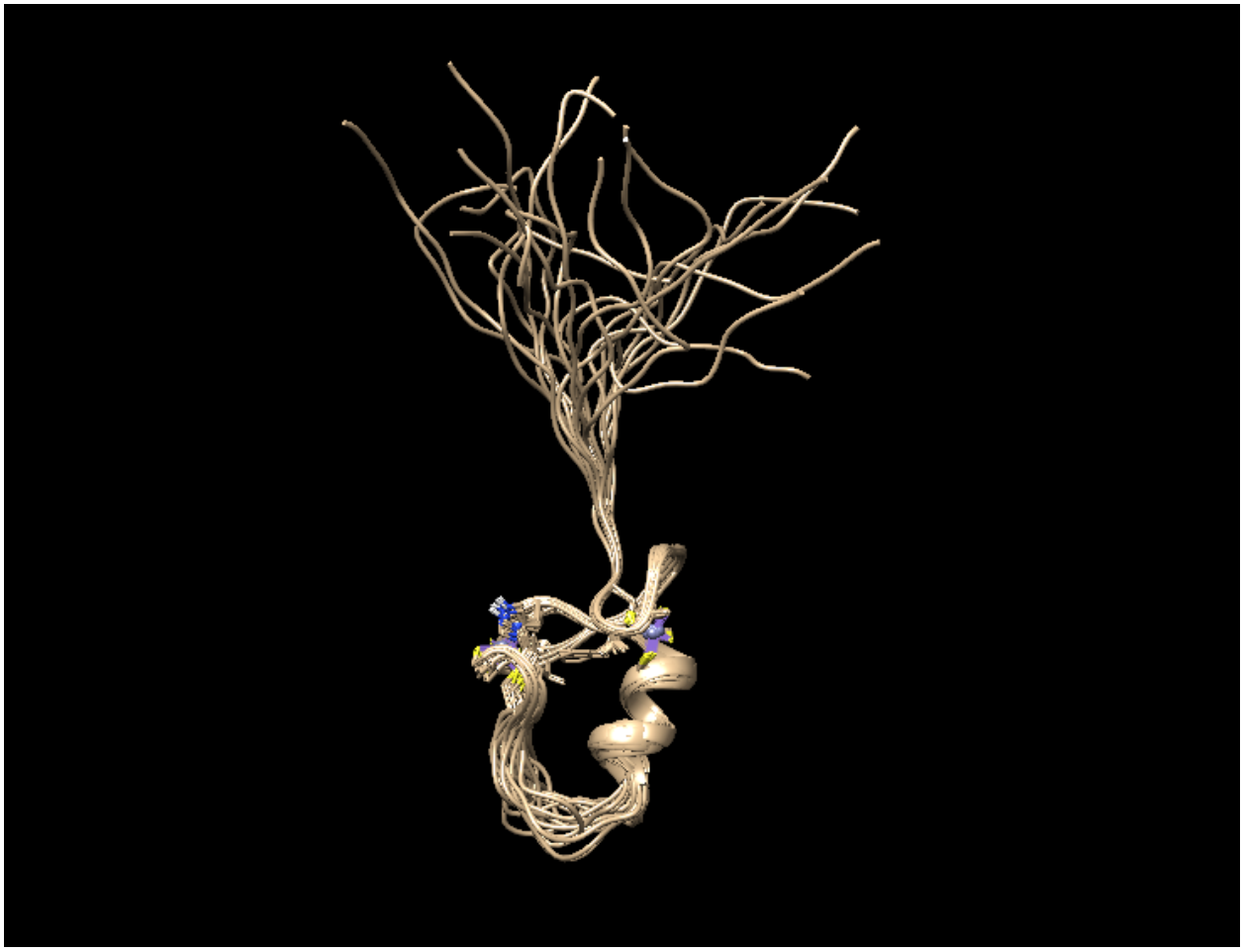
GA341 is a scoring function based on a machine-learning algorithm that was trained on a set of experimentally determined protein structures. It uses a variety of features to predict the quality of a protein model, including the degree of sequence conservation, the number of long-range interactions, and the compactness of the structure. GA341 has been shown to be effective in selecting the best models from a pool of predicted structures and has been used in various protein structure prediction competitions.

DOPE (Discrete Optimized Protein Energy) is another widely used scoring function that measures the energy of a protein model in terms of its compatibility with known physical and chemical properties of protein structures. DOPE calculates the energy of a protein structure based on statistical potentials derived from a non-redundant set of experimentally determined protein structures. The energy score is used to compare the quality of different protein models and select the most accurate one. DOPE has been shown to be effective in discriminating native-like models from incorrect ones and is commonly used in protein structure refinement and model selection.

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.
Ans.

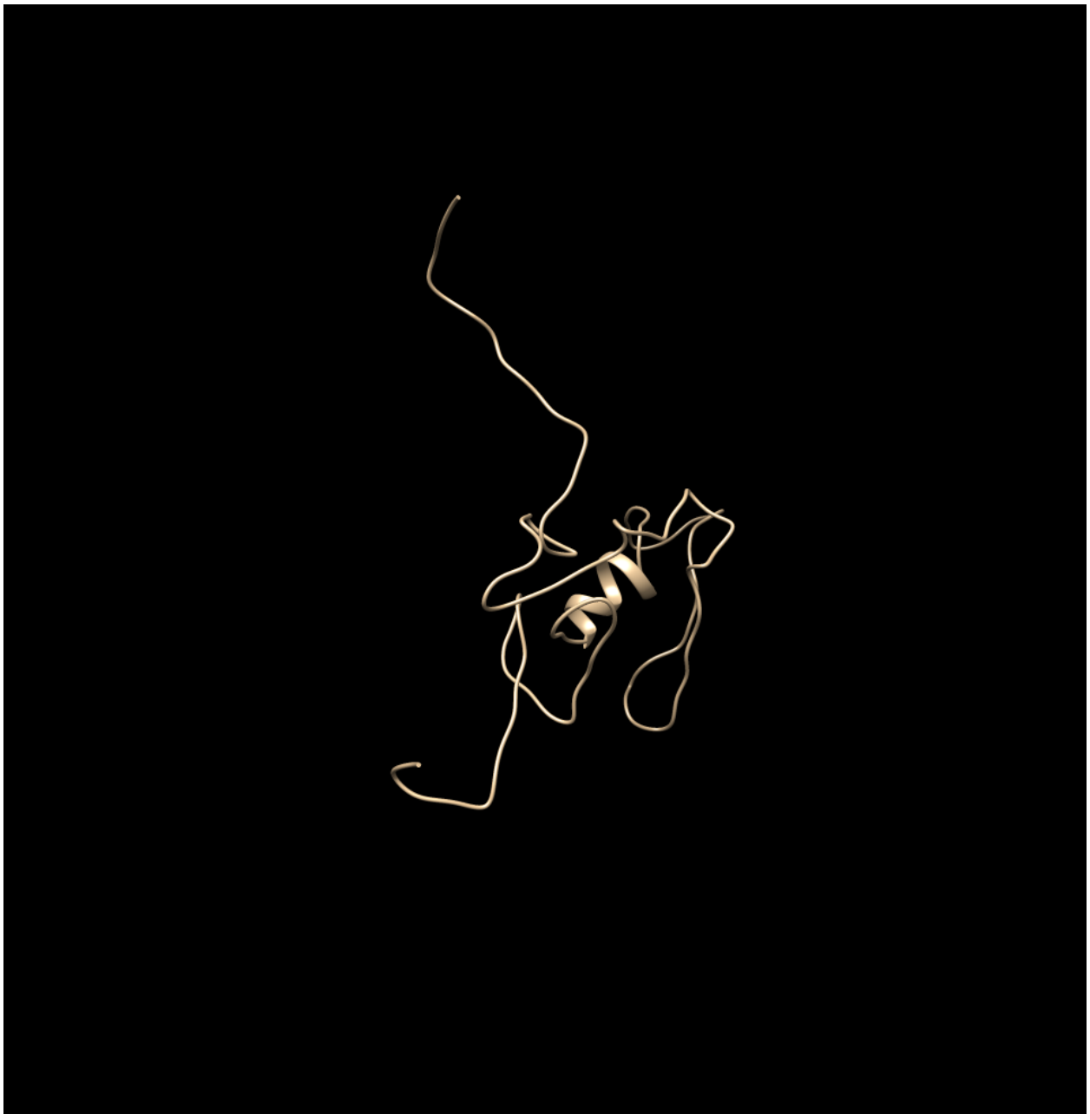
Yes, The modelled structure is similar to the original model of 2ecj.pdb.

PDB Structure of 2ecj.pdb:



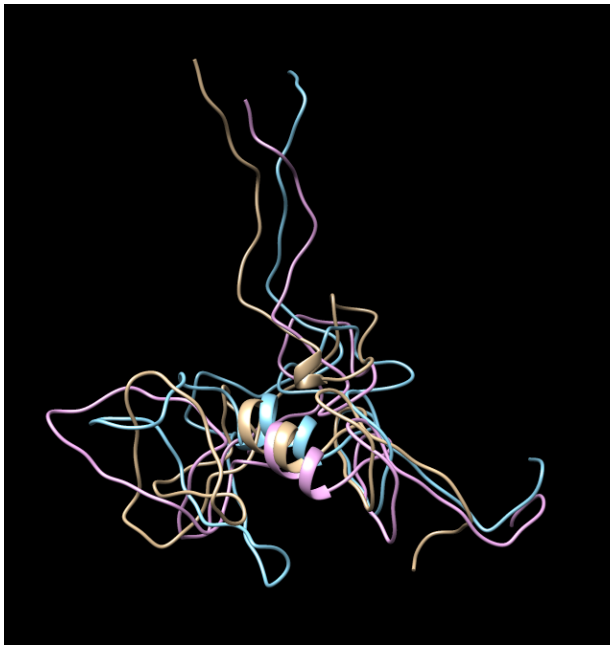
My Modelled Structure:

Single model:

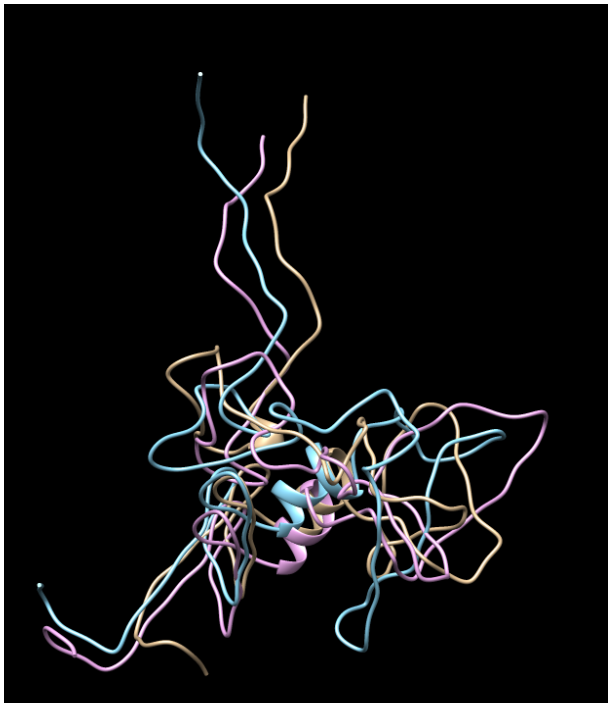


A combined model of 3 PDBs.

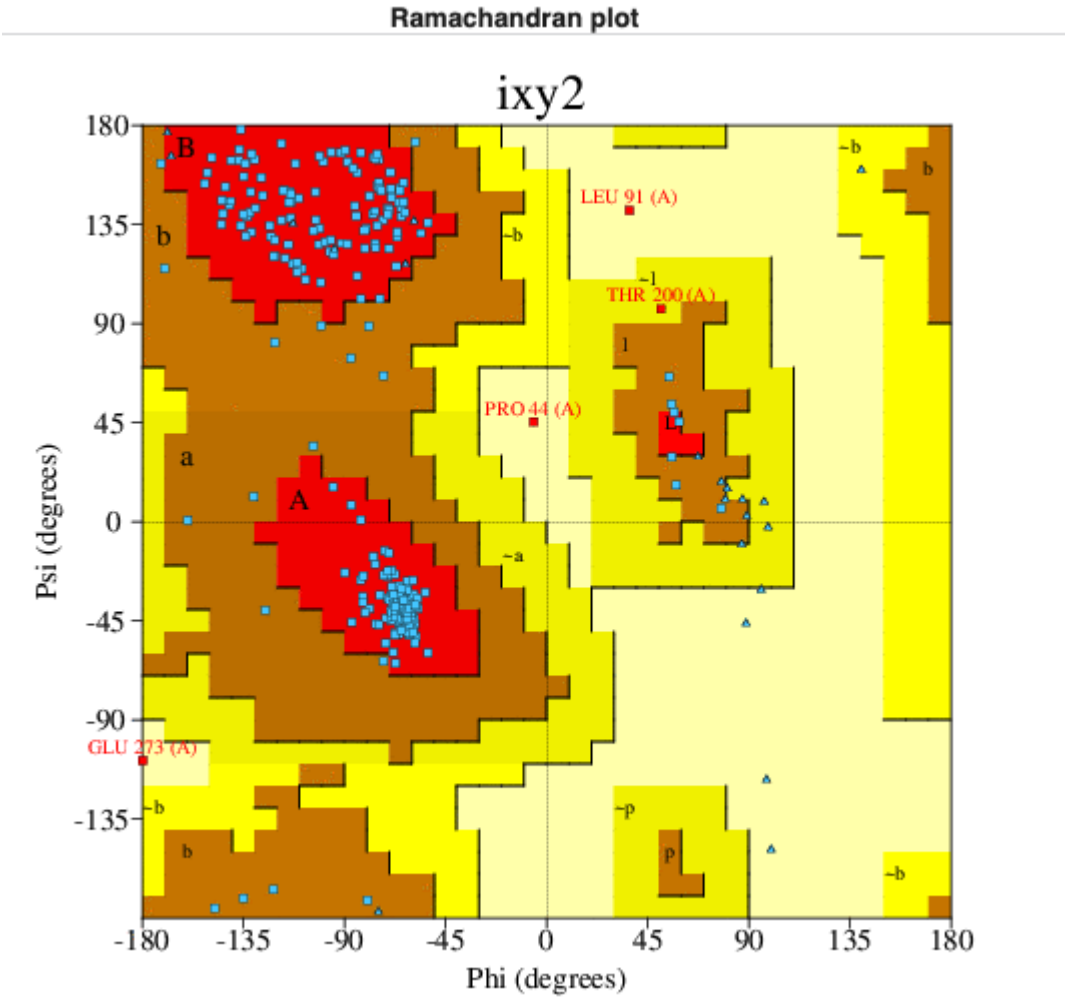
Normal View



Horizontally Flipped View



Question 6: 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?
Ans.



PROCHECK statistics

1. Ramachandran Plot statistics

		No. of residues	%-tage
Most favoured regions	[A,B,L]	269	91.8%
Additional allowed regions	[a,b,l,p]	21	7.2%
Generously allowed regions	[~a,~b,~l,~p]	1	0.3%
Disallowed regions	[XX]	2	0.7%*
Non-glycine and non-proline residues		293	100.0%
End-residues (excl. Gly and Pro)		1	
Glycine residues		25	
Proline residues		16	
Total number of residues		335	

Based on an analysis of **118** structures of resolution of at least **2.0** Angstroms and *R*-factor no greater than **20.0** a good quality model would be expected to have over **90%** in the most favoured regions [A,B,L].

2. G-Factors

Parameter	Score	Average Score
Dihedral angles:-		
Phi-psi distribution	0.07	
Chi1-chi2 distribution	-0.10	
Chi1 only	0.25	
Chi3 & chi4	0.40	
Omega	-0.23	
		-0.02
Main-chain covalent forces:-		
Main-chain bond lengths	-0.13	
Main-chain bond angles	-0.30	
		-0.23
OVERALL AVERAGE		-0.10

G-factors provide a measure of how **unusual**, or out-of-the-ordinary, a property is.

Values below -0.5* - unusual
Values below **-1.0**** - highly unusual

Important note: The main-chain bond-lengths and bond angles are compared with the Engh & Huber (1991) ideal values derived from small-molecule data. Therefore, structures refined using different restraints may show apparently large deviations from normality.

Results show that the model has **two** residues in the disallowed regions. Although the model is not so accurate, it can be used for further studies and refined. Also, the G-factor is also around -0.10, which is unusual but better than the highly unusual, which is -1.0.

Question 7. Do you think using multiple templates (or multi-template homology modelling) could have resulted in a better structure? Justify your answer.
Ans.

Yes, multi-template modelling could have been a better option. We used 2ecj template because it covers 42% of the query and hence is the best template. The multi-template model provides a better way of choosing templates corresponding to different regions of the query proteins. It can help to account for the diversity in the protein family, which can be difficult to capture with a single template structure.

However, the effectiveness of multi-template modelling can depend on the quality and relevance of the templates used, as well as the algorithm and parameters used for modelling.