Report - Module 1: Advanced Homology Modeling

N Narotam
narotam21477@iiitd.ac.in

Lakshya Kumar Raikwar
lakshya21470@iiitd.ac.in

Lakshya Goel
lakshya21469@iiitd.ac.in

Manav Saini
manav20518@iiitd.ac.in

Navidha Jain
navidha20223@iiitd.ac.in

Paras Dhiman
paras21482@iiitd.ac.in

Udit
udit21213@iiitd.ac.in

Abstract

This report details modelling the gene P01130, identified as LDLR_HUMAN, assigned to the group with IDK. We systematically explored gene modules specified in an Excel sheet, documenting each step and its corresponding results. This report compiles the findings along with tables and images.

1. Retrieval and Format Conversion

1.1. Downloading the FASTA Sequence (P01130)

The UniProt ID, P01130, was queried within the Uniprot database. The search led to the specific entry at LDLR_HUMAN. The canonical FASTA sequence was downloaded from this entry.

The Fasta file is given by $sequence.fasta$

1.2. Conversion to PIR Format

The downloaded FASTA sequence was converted to the PIR format using the web tool available at bugaco. This online tool facilitates conversion from Fasta to PIR.

The PIR file is given by $seq.pir$

2. Template Selection through BLASTp

2.1. Running BLASTp

The BLASTp analysis was conducted using the web tool available at NCBI-PDB Blast, with the uploaded FASTA file as the query.

This tool, hosted on the National Center for Biotechnology Information (NCBI) website, allowed for the analysing of the uploaded sequence against the Protein Data Bank (PDB) database.

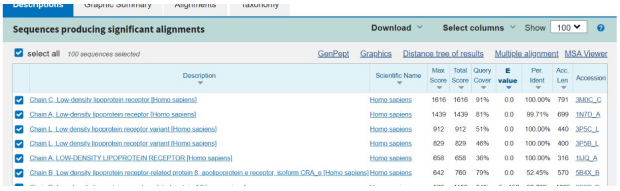

Figure 1. Top 5 BLASTp Results

2.2. Template Selection

Templates were selected based on criteria including a lower E-value for greater significance, a higher score indicating improved alignment quality, and substantial query coverage to ensure a comprehensive representation of the query sequence. Among the top five templates—3M0C_C, 1N7D_A, 3P5C_L, 3P5B_L, and 11JQ_A—3M0C_C emerged as the most promising. Notably, 3M0C_C exhibited 91% % query coverage, 100% % identity, and a total score of 1616 with an E-value of 0, indicating its high significance for further analysis.

The PDB file for 3M0C_C is given by $3m0c.pdb$

3. Model Generation and Evaluation

3.1. Residue Analysis

Since all of the top five templates had sufficient crystallised residues with negligible missing residues, group members ran models on all of them. This was done to compare and contrast Models from other templates to the flagship model from the 3M0C_C template.

3.2. Template-Target Alignment

Using the 'align2D.py' script, the query sequence (P01130) is aligned with the BLASTp resultant template, represented by the template. The resulting alignment is saved in PIR ('Target-template.ali') and PAP ('Target-template. pap') formats. The PIR format is employed for

subsequent model building in MODELLER, while the PAP format is used for visual inspectability.

3.3. Model Generation

The 'build.py' script was modified to produce 10 models using the 'Target-template.ali' file. The resultant 10 files are stored in "LDLR.B9999000XX.pdb" name format files.

3.4. Model Evaluation

The 'evaluate_model.py' evaluates the 10 generated models and prints the dope scores for all the models and the best dope score.

```
>> Summary of successfully produced models:
Filename                          molpdf     DOPE score    GA341 score
----------------------------------------------------------------------
LDLR.B99990001.pdb            9111.24316   -38366.13281        0.31232
LDLR.B99990002.pdb            7454.44434   -40051.30469        0.64691
LDLR.B99990003.pdb            7149.82422   -40059.98047        0.63011
LDLR.B99990004.pdb            7542.05322   -40448.85938        0.58810
LDLR.B99990005.pdb            8763.91211   -38478.26953        0.79046
LDLR.B99990006.pdb            7808.96191   -39707.19531        0.23484
LDLR.B99990007.pdb            7463.12891   -39565.96094        0.50015
LDLR.B99990008.pdb            7266.69043   -40791.96484        0.69118
LDLR.B99990009.pdb            7646.58447   -40729.93359        0.90722
LDLR.B99990010.pdb            7441.27734   -40674.18750        0.84995
```

Figure 2. Summary for 10 models (Model with lowest dope score highlighted)

4. Results and Visualisation

The dope score profile graphs are generated using the 'plot_graphs.py' script. The RMSD score for the best model is calculated using the 'rmsd_calculation.py' script, and the overlapping was visualised using the PyMol software. The Resultant Dope score and specific dope score of the Model and Template, along with the Rmsd value, have been attached as images.

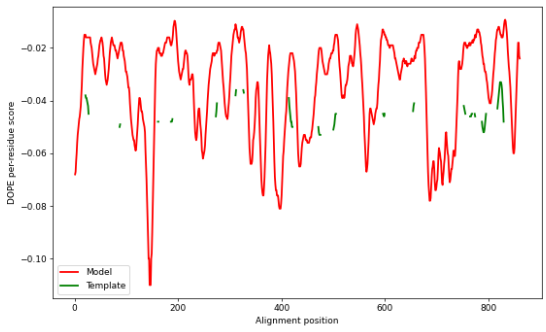

Figure 3. profile graph

```
DOPE score               : -9789.397461
Dope score of Model:  -40792.4140625
Dope score of Template:  -9789.3974609375
Rmsd Value 20.242899
```

Figure 4. Model scores

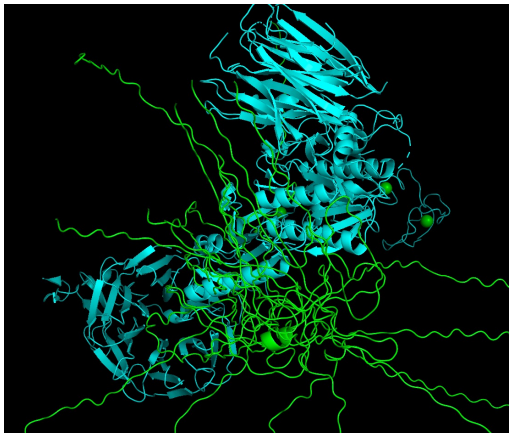

Figure 5. alignment visualisation [Green: Model, Blue: Template]

4.1. Discussion

Alignment to our model to 3M0C_C, despite the 100% identity with 3M0C_C, was driven by objective metrics.

However, While aligning our model to 3M0C_C, we can observe poor results based on multiple criteria, including DOPE scores, RMSD values, and visual inspection of the overlap. These findings suggest that, despite the identical sequence, the structural arrangement and conformation in 3M0C_C may not be the most suitable for our model.

We, therefore, aligned our model with other templates from the top 5, amongst which 1N7D_A showed the most promising results. The alignment with 1N7D_A, which shares a slightly lower identity of 99.71%, yielded significantly better results across the evaluated metrics. The decision to choose 1N7D_A as an alternative template was motivated by the improved alignment quality and structural compatibility demonstrated by this template.

The selection of 1N7D_A with a slightly lower identity underscores the importance of considering structural features and overall alignment quality beyond sequence identity alone. This approach ensures that our model aligns with a template that shares a high sequence similarity and provides a more accurate representation of the structural characteristics relevant to our study.

We followed exactly the same steps and scripts as with the 3M0C_C sequence as a template, and the images show a much better target modelling.

5. Advance Modelling

5.1. Template Alignment

The top five templates—3M0C_C, 1N7D_A, 3P5C_L, 3P5B_L, and 11JQ_A—3M0C_C, all have above 99% identity with the query sequence(1N7D has 99.71%, while all the other have 100%), were used to model missing regions using the 'salign.py' script. The file 'template.ali' was generated, which will be used for modelling, and the 'template.

```
>> Summary of successfully produced models:
Filename                          molpdf     DOPE score    GA341 score
------------------------------------------------------------------------
LDLR.B99990001.pdb             7012.02051   -58772.65234        0.10639
LDLR.B99990002.pdb             7580.27539   -58617.90234        0.12704
LDLR.B99990003.pdb             6997.23877   -58566.95703        0.26364
LDLR.B99990004.pdb             6655.41016   -57477.19141        0.73856
LDLR.B99990005.pdb             6668.26709   -61348.18359        0.79095
LDLR.B99990006.pdb             7411.10693   -56486.53125        0.31859
LDLR.B99990007.pdb             6794.46191   -59907.94922        0.37483
LDLR.B99990008.pdb             7607.08301   -60138.19141        0.50798
LDLR.B99990009.pdb             7323.66943   -59472.04688        0.11404
LDLR.B99990010.pdb             6863.21289   -60217.92969        0.10831
```

Figure 6. 1n7d: Summary for 10 models (Model with lowest dope score highlighted)

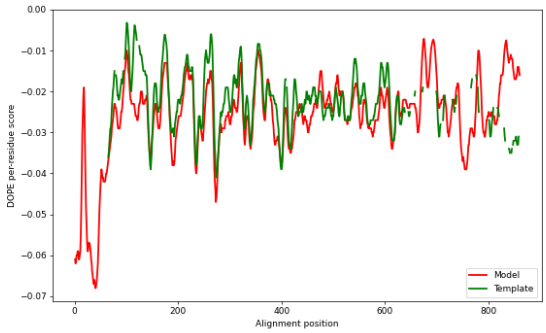

Figure 7. 1n7d: profile graph

```
DOPE score               : -36969.753906
Dope score of Model:  -54111.93359375
Dope score of Template:  -36969.75390625
Rmsd Value 42.56299
```

Figure 8. 1n7d: Model scores

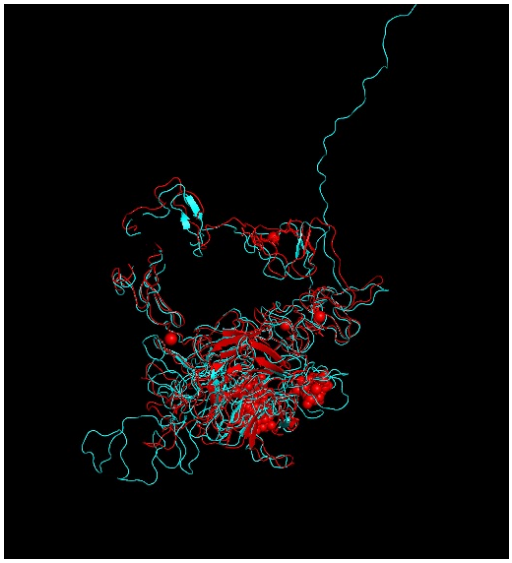

Figure 9. 1n7d: alignment visualisation [Red: Template, Blue: Model]

pap' was generated that will be used for visualisation.

5.2. Structure and Sequence Alignment

We used the 'align2d_mult.py' script to perform a pairwise alignment of the query sequence with multiple template structures. The 'templates.ali,' PIR file generated during template alignment was used as the input.

Two output files were generated: the 'LDLR-mult.ali' file in PIR format for modelling and the 'LDLR-mult.pap' file in PAP format for visualisation.

5.3. Model Building

For Model Building, we used the 'model_mult.py' script to generate 10 models for the query sequence based on a selection of template structures. The known structures are -'3m0cC', '1n7dA', '3p5cL', '3p5bL', and '1ijqA'. The query sequence is 'LDLR'.

The resultant output comprises 10 model files, each in the format 'LDLR.B999900XX.pdb', where 'XX' represents the specific model number.

```
>> Summary of successfully produced models:
Filename                          molpdf     DOPE score    GA341 score
------------------------------------------------------------------------
LDLR.B99990001.pdb             38102.22266   -56122.45703        1.00000
LDLR.B99990002.pdb             39068.32031   -55492.73438        1.00000
LDLR.B99990003.pdb             38233.09375   -56250.44922        1.00000
LDLR.B99990004.pdb             37906.61328   -56526.69531        0.99918
LDLR.B99990005.pdb             38417.97656   -54608.55859        1.00000
LDLR.B99990006.pdb             39103.89453   -54649.35156        0.99947
LDLR.B99990007.pdb             37887.75391   -55985.68359        0.99389
LDLR.B99990008.pdb             38032.98828   -55467.35938        1.00000
LDLR.B99990009.pdb             38323.54297   -54840.72266        1.00000
LDLR.B99990010.pdb             37898.65625   -55745.62500        1.00000
```

Figure 10. Summary for 10 models from model_mult.py (Model with lowest dope score highlighted)

'LDLR.B99990004.pdb', has the lowest DOPE score among the 10 models created.

5.4. Model Evaluation

We used the 'evaluate_model.py' script to assess and select the best model based on the DOPE and RMSD scores. The PyMOL software was used to visualise and analyse the structural alignment of the model. After selecting the optimal model, further evaluation was conducted to generate a profile file.

The 'plot_profile.py' script was used to plot the dope score profile.

The profile plotting process assessed all atoms with DOPE, generating an energy profile file ('LDLR-model10.profile'). The script used the 'LDLR-mult.ali' file, plotting the model alongside the template.

6. Comparing Basic and Advanced Models

Attached is the dope score comparison from the best basic modelled structure from basic and advanced models and overlap visualisation using the PyMol software.

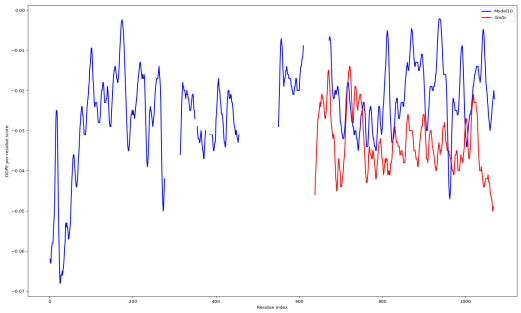

Figure 11. Comparison of DOPE per-residue scores between Model10 and template 3m0c.

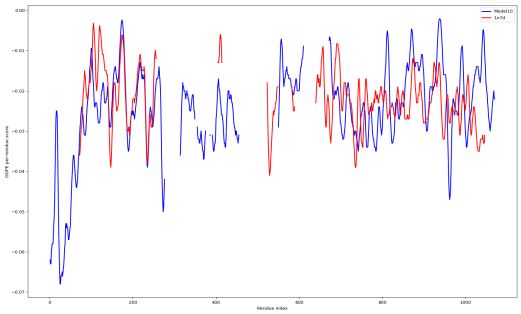

Figure 12. Comparison of DOPE per-residue scores between Model10 and template 1n7d.

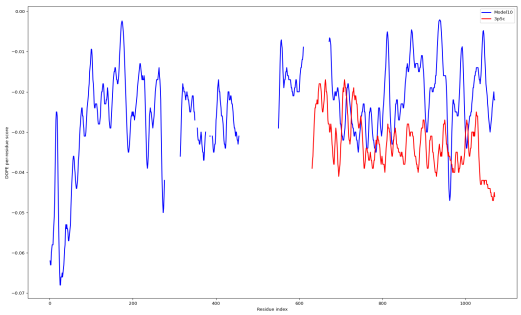

Figure 13. Comparison of DOPE per-residue scores between Model10 and template 3p5c.

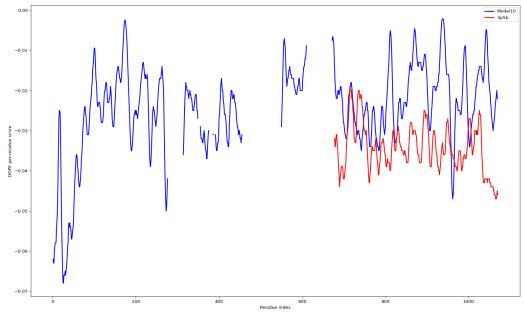

Figure 14. Comparison of DOPE per-residue scores between Model10 and template 3p5b.

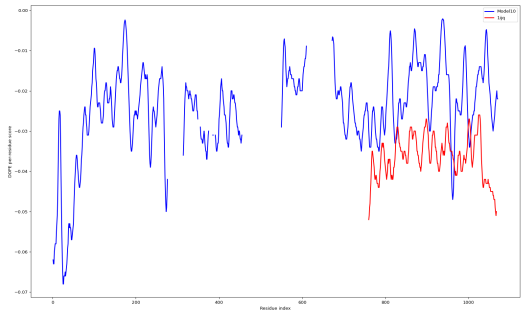

Figure 15. Comparison of DOPE per-residue scores between Model10 and template 1ijq.

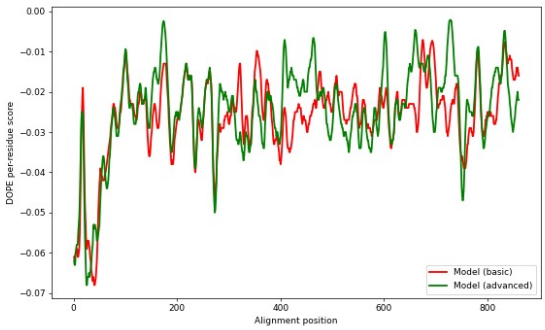

Figure 16. Comparison of DOPE per-residue scores between the best basic and advanced model.

The best basic and advanced models were alignment. The 'align2D.py' aligns the two best models with their equivalent profiles in the 'LDLR-ModelX.profile' files. The 'LDLR.ali' file is the alignment file in PIR format, 'the plot_graph.py' plots the dope score comparison graph. The 5 template pdb files, along with the best basic and advanced models in pdb format, are present to be used by the scripts. RMSD scores are calculated using the 'rmsd_calculations.py' script.

6.1. Structural Differences

Both models show fluctuations in the DOPE score across the alignment positions, with the advanced model generally having a less negative score than the basic model. This could imply that the advanced model predicts a more stable protein structure than the basic model across most of the se-

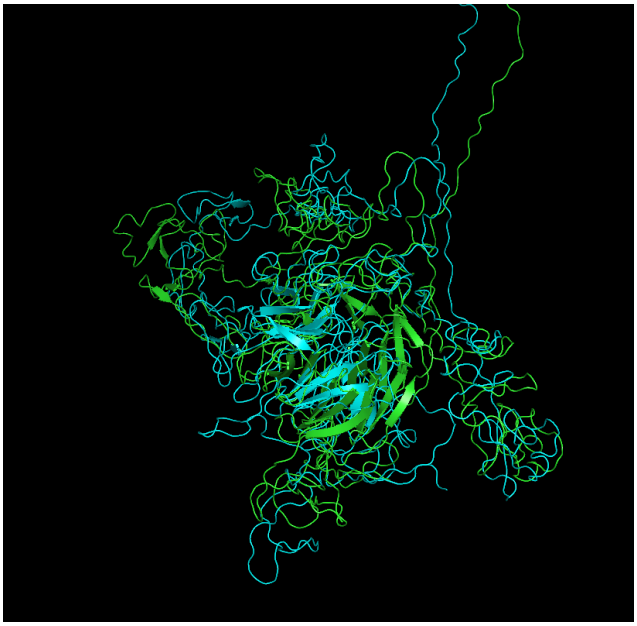

Figure 17. alignment visualisation of basic and advanced models.[Green: Basic, Blue: Advanced]

```
Basic modelling -

RMSD Value for 1n7d:  42.79813
RMSD Value for 1ijq:  42.85779
RMSD Value for 3p5c:  60.46339
RMSD Value for 3p5b:  60.450447
RMSD Value for 3m0c:  60.42523

Advanced modelling -

RMSD Value for 1n7d:  42.511456
RMSD Value for 1ijq:  42.728638
RMSD Value for 3p5c:  60.908653
RMSD Value for 3p5b:  60.89618
RMSD Value for 3m0c:  60.85931
```

Figure 18. alignment visualisation of basic and advanced models.

quence, except in certain regions where the lines cross, indicating positions at which the basic model predicts a more stable structure.

The regions where the lines are closer or cross might be points of interest, as they suggest positions where the difference in prediction between the basic and advanced models is minimal.

The structural differences implied by this graph suggest that the advanced model could better predict protein structure in some regions but not uniformly across the entire sequence. Interpretation of such a graph would typically be context-dependent, requiring knowledge of the specific proteins being modelled and the methods used by the basic and advanced models.

6.2. Position Details

The graph depicts two trends across the range of alignment positions, represented by the red and green lines for the basic and advanced models, respectively:

Both lines exhibit a sinusoidal-like pattern with multiple peaks and valleys. This suggests a recurring pattern in the stability of the protein structure at different points in the sequence. Initial Segment

6.2.1 Positions 0 to 200:

There is a steep decline in the DOPE score for both models at the beginning, with the basic model (red) starting at a higher score and then intersecting with the advanced model (green). This indicates a region where, initially, the basic model predicts a more stable structure but then converges with the advanced model's predictions. Middle Segment.

6.2.2 Positions 200 to 600:

Both models in this broad segment fluctuate and follow a similar trend. However, the basic model consistently predicts a slightly more stable structure with a lower DOPE score than the advanced model for most of this range. There are multiple points where the trends intersect, indicating specific alignment positions where the models' predictions are similar. Later Segment.

6.2.3 Positions 600 to 800:

Towards the end, both models' lines continue to fluctuate and largely parallel each other, with the basic model still often predicting a slightly more stable structure than the advanced model. There is less crossing over in this segment compared to the middle segment, suggesting that the predictions of the two models are more divergent in this protein region.

6.2.4 Variability:

The variability (the magnitude of fluctuation in the scores) appears consistent across the sequence for both models, indicating a similar confidence level in their predictions at each position. The trends suggest that while the basic model often predicts a more stable structure (lower DOPE score) across most positions, there are several specific points where the advanced model predicts equal or greater stability. These trends could be influenced by the models' differing treatment of local versus global protein

structure features, including different computational parameters, or using different training datasets. The alignment positions where the two models converge or diverge significantly could be particularly interesting for further analysis or experimental validation.