Comparative Protein Modelling

Sai Manikanta S Godavarthi, Deepika Joseph

Computer Science – EECS Department, Wichita State University, Wichita, Kansas

Abstract

Motivation:

Results: With the given 10 proteins; 5 from CASP 11 and 5 from CASP 10 we have obtained better accuracy for 3 proteins comparing with the CASP results. Results for each of the protein are explained separately in the results section. We have also implemented a sample pipeline for automation of the process of modelling using modeller. Implemented few refinement steps on some proteins where the accuracy has been increased compared to the previous results before refinement.

Contact: sxgdavarthi@shockers.wichita.edu

1 Introduction

The term comparative protein modelling or homology modelling or template based modelling are referred as the same where our own main goal is to model a protein 3D structure using the know templates. Here the templates are those with similar sequence to that of our unknown query protein. Based on the known protein i.e. our template we will model the query protein using the properties and alignment of the known template. The template contains sufficient information of spatial arrangement of residues and internal structure which helps in predicting our model. Comparative modelling of protein sequence is more reliable than compared to that of ab initio methods as in the later, the model is entirely built using only the sequence rather depending on the template. Protein are one of most important functional units of our body, they do most of the work in cells and they are required for structure, function, and regulation of body's tissues and organs. The three-dimensional structure of the protein determines the functionality of the protein. The four levels of proteins i.e. the primary structure which is a sequence of amino acid residues determine the peptide chain. In the secondary structure, hydrogen bonds between the amino acids creates alpha helix, which is a spiral or coiled molecule and pleated sheet that looks like ribbon with regular peaks and valleys as a part of the fabric. The tertiary structure is for overall shape of the protein which are either globular or fibrous. Quaternary structures describe the proteins appearance.

Our method of protein modelling starts with taking a query sequence. Query sequence is the one which we want to model as a three-dimensional structure for the protein then later identifying template and build model using the modeller tool and we validate our results using varies validation techniques available online and do structural analysis of the proteins using visualization tools.

2 Methods

The whole process if protein modelling is described as follows:

i. Select the required protein and get the sequence.

ii. Search for template in BLAST, PDB and other protein databases accordingly as required.

iii. Find the better matching sequence for the query sequence, this is our template. We may have single or multiple templates. Align the template with our query.

iv. Prepare our files in PIR format. Download the PDB formats for the templates. Save PIR formatted query file as .ali extension.

v. Start the modeller by giving input the query .ali file and other PDB files, according to the log file generated give inputs to the modeller.

vi. Validate the chosen best model and determine the accuracy.

To align the given sequences, we have used multiple sequence alignment techniques for multiple templates selected, some of them include T-Coffee and Clustal-Omega. At some cases, we have chosen only one template, where we have used Needleman-Wunch algorithm to align the sequences and converted all of them into PIR format. The pipeline that we have implemented asks for sequence and automatically converts them into PIR format and saves query in .ali file and rest template sequences as .pir extension. The PDBs are automatically download once given the template IDs after script 1 execution. The program that we wrote has BioPython packages and uses NCBI pdb API call to download the PDB files given the template IDs. Alternatively, we can use REST API to download the pdb files. The REST API is of XML format.

For validation, we are using different tools available online and for visualization purpose we have used Chimera as our tool. The various online techniques that we have used to measure the accuracy of our model inclue TM-Score, Molprobity scores, and RMSD score.

All the steps are cleared explained clearly below where we have mentioned each of options that we have chosen for protein modelling.

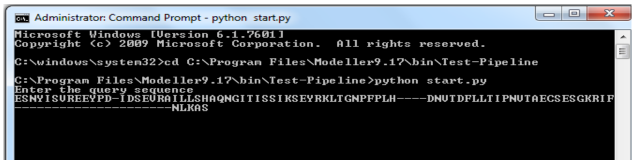
The software distribution that we used for this project is described as below:

a. Test cases i.e. Protein queries are taken from CASP website. CASP10 and CASP 11 are chosen to select 10 proteins which include:

- CASP11 targets T0856, T0843, T0806, T0837, T0792 and
- CASP10 targets T0757, T0666, T0678, T0651, T0694

b. For template identification: BLAST, PDB, and SWISS-Model.
c. Sequence alignment: Needleman-Wunch algorithm, and T-Coffee
d. Software for protein modelling used is Modeller 9.17.
e. Various languages used are python 2.7 and JAVA (Needleman-Wunch algorithm)
f. Protein visualization software's: Chimera, and Rasmol
g. Project Management and version control: Github

Process automation:
Implemented a sample pipeline where the process of using modeller is made lot easier compared to the standard approach. Instead changing inputs in the script files everytime, user needs to just enter the sequences initially and template IDs later for each of the script to run. The script automatically converts the given aligned query sequence into PIR format as saves it as .ali extension file. At each script execution, based on the output generated by the script and after evaluating the results, we given the template ID as input to the next script and the process is the same till the end. Sample screenshot for the pipeline is as below, where the program ask for user input query sequence.



```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\windows\system32>cd C:\Program Files\Modeller9.17\bin\Test-Pipeline

C:\Program Files\Modeller9.17\bin\Test-Pipeline>python start.py
Enter the query sequence
ESNYISVREEYPD-IDSEVRAILLSHAQNGITISSIKSEYRKLTGNPFPLH----DNVTDFLLTIPNVTAECSESGKRIF
---------------------NLKAS
```

Figure 1:start.py script; User entering query sequence

When the user enters the aligned sequence, the script process the sequence and ask for user to confirm by pressing enter, once the user clicks enter the first script of modeller starts running. Now for each script, once the execution is complete the console ask for user input. From script 2 users' needs to enter just the template IDs and model ID for the last one accordingly after evaluating the log results generated by the modeller. The PDB files required for the template can be automatically downloaded when we use BioPython libraries available for python. So, taking these dependencies we can download the PDB files directly. Another option is we can use REST API calls provided by PDB website where are already available in XML format. So, we use them to download the PDB files required for running script2. Similarly, the pipeline process is continued for each of the proteins.

3 Results

With the given ten CASP sequences, we divided the sequences among ourselves in the group and completed the process individually later we validated our results by exchanging the proteins we modelled among ourselves. For template search, we have used BLAST as our primary source and we played around with different option available with BLAST to match the best template that we can obtain from our resources. For most of the results that are obtained we have used PSI-BLAST option in BLAST and searched for template in PDB database. Each of the protein results are explained below separately and all the results obtained are discussed individually at first and then comparison is made with CASP results to show the accuracies of our protein model that we have generated.

3.1 CASP 10 and CASP 11

For the process of modelling the proteins, we have chosen 5 query sequences for CASP 10: T0757, T0666, T0678, T0651, T0694 and CASP 11: T0856, T0843, T0806, T0837, T0792 respectivey we have followed the same procedure as discussed in the pipeline. Initially we have taken the query sequence and searched for template in BLAST with option as pdb database and PSI-BLAST searching techniques. Depending the results obtained and percentage of identity between the chosen templates and query sequence we have used either of pair wise sequence alignment with Needleman-Wunch algorithm and T-Coffee for multiple sequence alignment between query and templates which are explained clearly for each of the protein below with results obtained for each of the protein. For the purpose of pair-wise sequence alignment we have implemented a JAVA program from github and for multiple sequence alignment we have used T-Coffee and Clustal-Omega accordingly when required and for validation and refinement purpose.

3.1.1 Protein modelling using Modeller

We have taken 5 protein query sequences from CASP 10 i.e. T0757, T0666, T0678, T0651, and T0694 respectively and from CASP 11 i.e. T0856, T0843, T0806, T0837, and T0792 as our query sequence individually and the results obtained are as follows:

- **T0757:**

we have taken the query sequence from CASP and used BLAST to obtain the template sequence, although we got couple of matching sequences but as result we took 4gak as out template and performed pair-wise sequence alignment between each of the query and template. Initially we tried with multiple sequence alignment but the TM-Score obtained for the protein sequence alignment is around 0.8, but when we tried pair-wise alignment with selected template we obtained TM-Score as 0.9629. the result of the modeller scores is as follows:

Filename	molpdf	DOPE score	GA341 score
Model1.pdb	1250.17883	-29520.33789	1.00000
Model2.pdb	**1209.11462**	**-29355.93359**	**1.00000**
Model3.pdb	**1170.82837**	**-29699.65625**	**1.00000**
Model4pdb	1374.11084	-29354.61719	1.00000
Model5.pdb	1231.56580	-29481.41602	1.00000
Model6.pdb	1167.89832	-29369.69531	1.00000

We have chosen model3 for our result which has better TM-Score compared to other models. We have also considered Model3 as a part of refinement which has better Molprobity score 2.02 compared to Model2 which has score of 2.43, so as a refinement we have reconsidered Model3 as our solution even though it has high DOPE score but least molpdf score.

- **T0666:**

For T0666 we have started the analysis and obtained multiple templates, after analysis we have chosen 3napA and 3ux4A as our resultant templates and performed multiple sequence alignment with query. The diagonalization matrix obtained as a comparison is as follows:

```
          3napAA@23ux4AA@3
3napAA@2      264       2
3ux4AA@3        1     180
```

Based on the results which are bit confusing to choose the better template, we have performed modelling taking both the templates separately and after obtaining results we have chosen best template and model based on the TM-Score obtained for each of them respectively. Finally, we have chosen 3ux4A as the best matching template which and the results for model structures are as follows:

```
>> Summary of successfully produced models:
Filename                  molpdf     DOPE score    GA341 score
----------------------------------------------------------------------
Model1.pdb          1022.91730   -23121.25391        1.00000
Model2.pdb          1053.24976   -23116.05078        1.00000
Model3.pdb           950.99738   -23461.31836        0.99997
Model4.pdb          1070.12378   -23245.58594        0.99999
Model5.pdb          1003.10345   -23420.93164        0.99994
```

Based on scores, we have choosen model2 as our best model which as average of scores as well comparatively and obtained a TM-Score of 0.9042 and Molprobity score as 2.92.

- **T0678:**

we have performed pair-wise sequence alignment with the query protein taking 4epz as our template. The results obtained are better comparing to multiple sequence alignment that we have performed initially, later with refinement we have chosen pair-wise sequence alignment as best choice for this protein based on our template models. The results obtained for models are follows:

```
>> Summary of successfully produced models:
Filename                  molpdf     DOPE score    GA341 score
----------------------------------------------------------------------
Model1.pdb           643.56921   -18270.39844        1.00000
Model2.pdb           613.73462   -18336.48047        1.00000
Model3.pdb          1147.81177   -17304.24219        1.00000
Model4.pdb           595.51093   -18572.18359        1.00000
Model5.pdb           590.30353   -18337.23242        1.00000
```

We fixed model2 as our best model based on the results of TM-Score of 0.9222 and Molprobity score as 2.19.

- **T0651:**

We have performed multiple sequence alignment for this protein later we have analyzed our results after final decision of template and model selection, here we noticed multiple-sequence alignment performed better by few points compared to pair-wise sequence alignment. The results obtained for models of chosen template are as follows:

```
>> Summary of successfully produced models:
Filename                  molpdf     DOPE score    GA341 score
----------------------------------------------------------------------
Model1.pdb          1590.79980   -30183.49805        1.00000
Model2.pdb          1579.90295   -30485.20117        1.00000
Model3.pdb          1718.90918   -30320.98438        1.00000
Model4.pdb          1441.80750   -30527.74219        1.00000
Model5.pdb          1662.43384   -30278.29883        1.00000
Model6.pdb          1435.20447   -30606.22852        1.00000
```

We have finally decided to take model6 as our result model based on taking average scores of DOPE and molpdf. The TM-Score obtained is 0.97 and Molprobity score of 2.18.

- **T0694:**

Based on the results of BLAST, we have performed multiple sequence alignment of templates and query using T-Coffee and the results obtained for models are as follows:

```
>> Summary of successfully produced models:
Filename                  molpdf     DOPE score    GA341 score
----------------------------------------------------------------------
Model1.pdb          1435.19836   -38772.60547        1.00000
Model2.pdb          1422.11194   -38440.36719        1.00000
Model3.pdb          1332.25159   -38742.54688        1.00000
Model4.pdb          1465.64661   -38483.46484        1.00000
Model5.pdb          1380.60278   -38905.53125        1.00000
```

We have chosen Model3 as our resultant model taking average of scores obtained and we have achieved a TM-Score of 0.9576 and Molprobity score of 2.21 which is comparatively better solution based on analysis of CASP results which we will discuss later in the paper.

- **T0843:**

We have used multiple sequence alignment of different templates with query sequence and obtained better Molprobity score than CASP predictions available on the website. The TM-Score is also appeared to be more reliable and overall, we scored almost equal TM-Score and better Molprobity score compared with top appeared results. The models generated has the scores as below:

```
>> Summary of successfully produced models:
Filename                  molpdf     DOPE score    GA341 score
----------------------------------------------------------------------
Model1.pdb          1987.00964   -45485.44531        1.00000
Model2.pdb          1966.13806   -45068.55859        1.00000
Model3.pdb          1899.45654   -45152.75781        1.00000
Model4.pdb          1810.79456   -45319.01563        1.00000
Model5.pdb          1907.92151   -45049.98828        1.00000
```

based on average of the results we have chosen model3 as our best model and it has better molprobity compared to others.

- **T0806:**

We have used pair-wise sequence alignment for this model and obtained best TM-Score compared with the CASP results. Initially we tried with multiple sequence alignment but comparatively pair-wise sequence alignment performed better and the results obtained are as follows:

```
>> Summary of successfully produced models:
Filename                  molpdf     DOPE score    GA341 score
----------------------------------------------------------------------
Model1.pdb          1111.07874   -31527.54883        1.00000
Model2.pdb          1288.70447   -31395.63867        1.00000
Model3.pdb          1436.59229   -31155.95898        1.00000
Model4.pdb          1199.31531   -31481.81055        1.00000
Model5.pdb          1460.71851   -30915.53516        1.00000
Model6.pdb          1187.11096   -31542.72852        1.00000
```

We have choosen molpdf score as evaluating factor and based on the result as model1 performed better we made further analysis with TM-Score and molprobity as a analysis factors and determined model1 as our best results which outperformed on CASP results obtaining TM-Score as 0.9682.

- **T0837:**

Specifically, to this protein, results can be improved and lot of refinement can be performed. The template that we identified with different means didn't yield any better results for us. We have tried using many techniques, like pair-wise, multiple-sequence alignment with T-Coffee and Clustal-Omega. Although TM-Score improved after identification of better template and pair-wise sequence alignment, still it is in between random and accepted model. Here we show TM-Score and Molprobity score than DOPE scores as they are important for validation and we've obtained the scores as follows:
TM-Score: 0.234
Molprobity: 3.18

- **T0792:**

This protein has the similar case as above, but we have obtained fair results upon refinement. Initially the TM-Score was as low as 0.12. Later with refinement and proper identification of templates the score improved to 0.4172 and for few of the matching sequence i.e. templates didn't show any TM-Scores as there were no matching residues in modeled result and template. The various TM-Scores before refinement and after refinement are as follows:

Before refinement
Templates 54a9, 5cd7, 5a49 has no TM-Scores; template 4obm has TM-Score of 0.12
After refinement
2i6e gave me a score of 0.16 and 3s93 has score of 0.4172.
We have performed different techniques with 3s93 which is best template available, out of all the techniques multiple sequence alignment with templates that are selected after refinement gave better results.

3.2 Analysis

For analyzing the accuracy of our data, we have compared our results with CASP results and based on the TM-Score results that we have obtained, 4 of our proteins outperformed the CASP results and are as shown below in Fig.2

TM-score obtained for the proteins T0806, T0757, T0666, T0694, T0678 and T0651 outperformed compared to that of CASP results and when we consider protein T0843 it almost nearly modelled to that of CASP result and it has better molprobity results compared to that of CASP result so overall, we can say it is well modelled comparing with CASP results. Few of the results are obtained upon refinement of the original modelling which was clearly explained in the paper above and the data table for the above graph is as shown below in Table 1.

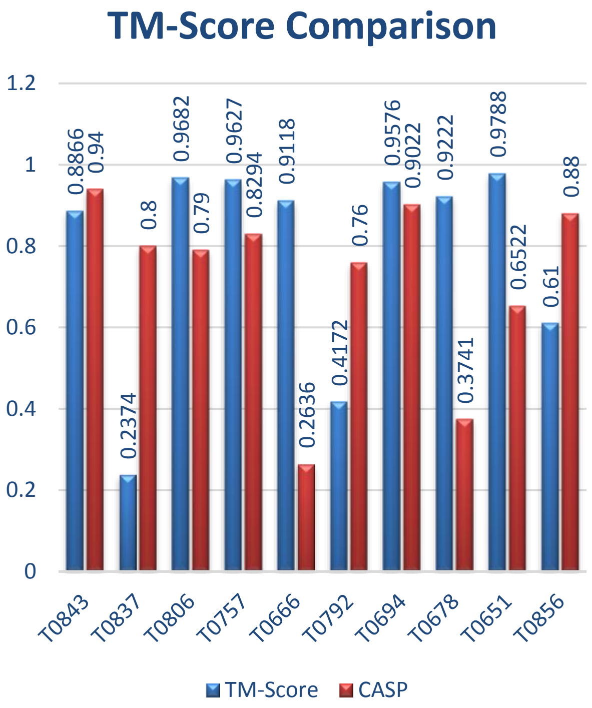


Figure2:TM-Score comparison with best CASP models

Table1: Comparison of TM-scores with CASP best models

Protein	*TM-Score*	*CASP best TM-Score*
T0843	**0.8866**	**0.94**
T0837	**0.2374**	**0.8**
T0806	**0.9682**	**0.79**
T0757	**0.9627**	**0.8294**
T0666	**0.9118**	**0.2636**
T0792	**0.4172**	**0.76**
T0694	**0.9576**	**0.9022**
T0678	**0.9222**	**0.3751**
T0651	**0.9788**	**0.6522**
T0856	**0.61**	**0.88**

and the result comparison with CASP with molporbity scores are viewed in below table along with the type of alignment that we have used for our protein modelling is given in the below table 2.

Table2: Molprobity and sequence alignment use; comparing molprobity scores with best CASP models

Protein	***Alignment Used***	***Molprobity Score***	***CASP-Molprobity Score***
T0843	Multiple	2.42	3.02
T0837	Pair-wise	3.18	1.58
T0806	Pair-wise	2.04	1.37
T0757	Pair-wise	2.02(5th in the CASP list)	0.83
T0666	Multiple	2.92	1.23
T0792	Multiple	3.02	1.18
T0694	Multiple	2.21	0.7
T0651	Multiple sequence	2.18 (13th in CASP List)	1.27
T0678	Pair-wise	2.19	0.74
T0856	Pair-wise	2.82	1.66

3.3 Protein View using Chimera

Fig3: Protein view of T0806 using chimera-Ribbon view

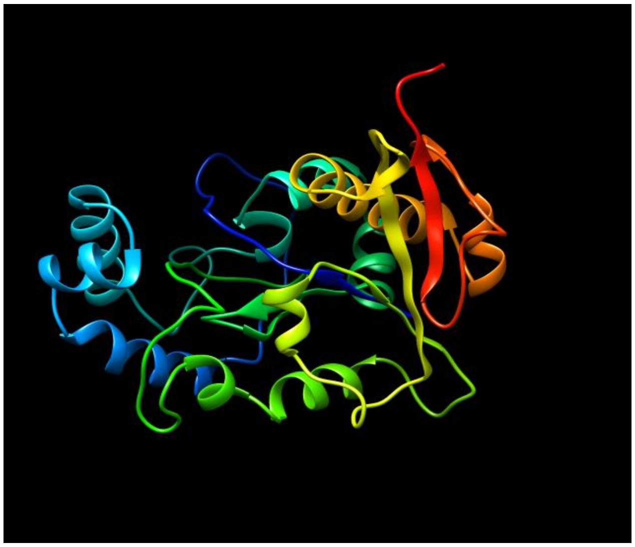
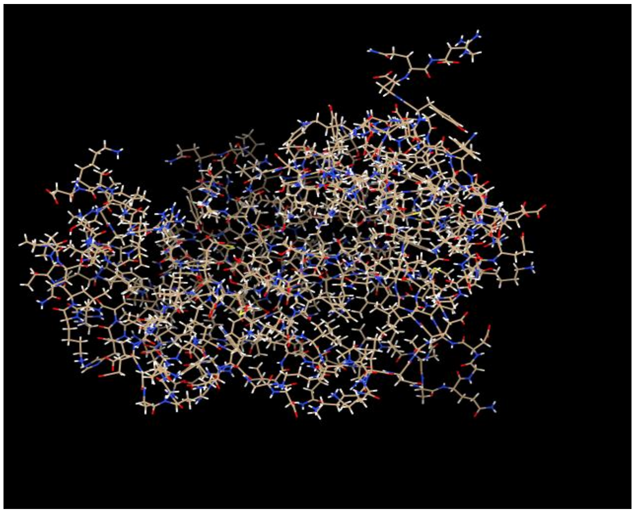


Fig4: Protein view of T0806 Chimera-All atom view



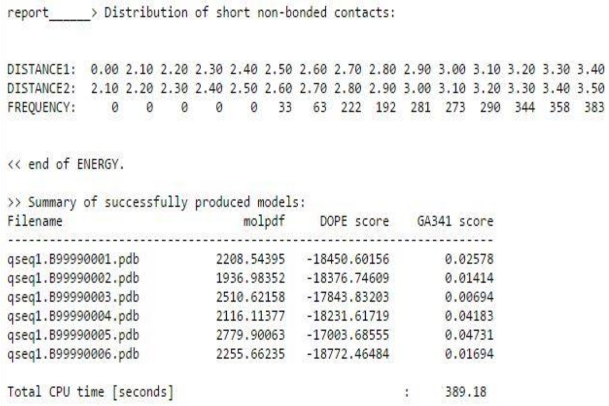
Sample visualization for protein structure modelled is as shown in the above Figure 3 which was represented in Ribbon view and Figure 4 as All atom view using chimera protein visualization software. We can also perform visual analysis using this software where we can see the bonds present and performed many structural and different kinds of visual analysis such as seeing the position of atoms, types of atoms and structural view of the model that was generated etc.

3.4 Case Studies

- We had a problem with protein identification initially as we have chosen pBlast as our option in BLAST search and later we have refined our results and identified better templates using PSI-BLAST option.

- We faced a lot of challenge with choosing the best of modelled templates, mainly because of scores obtained with modeller as they are uneven and hard to select the correct one. For example, the scores obtained for T0651 protein are as follows:

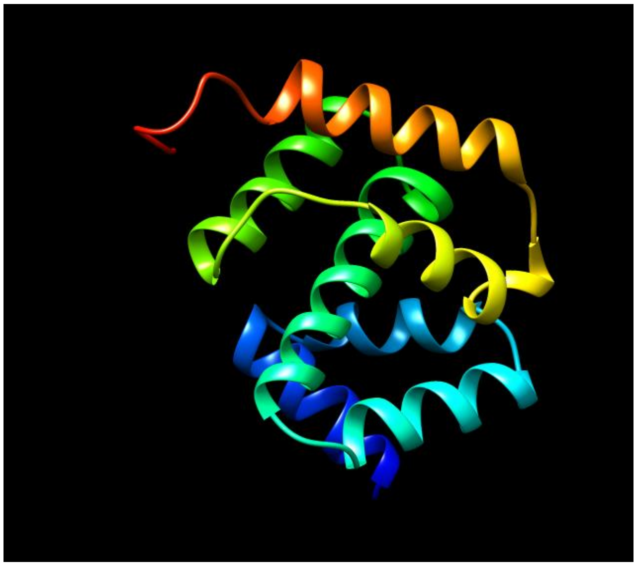
Fig5: T0651 Script 4 results using Modeller



```
report______> Distribution of short non-bonded contacts:


DISTANCE1:  0.00 2.10 2.20 2.30 2.40 2.50 2.60 2.70 2.80 2.90 3.00 3.10 3.20 3.30 3.40
DISTANCE2:  2.10 2.20 2.30 2.40 2.50 2.60 2.70 2.80 2.90 3.00 3.10 3.20 3.30 3.40 3.50
FREQUENCY:     0    0    0    0    0   33   63  222  192  281  273  290  344  358  383


<< end of ENERGY.

>> Summary of successfully produced models:
Filename                          molpdf     DOPE score    GA341 score
----------------------------------------------------------------------
qseq1.B99990001.pdb           2208.54395   -18450.60156        0.02578
qseq1.B99990002.pdb           1936.98352   -18376.74609        0.01414
qseq1.B99990003.pdb           2510.62158   -17843.83203        0.00694
qseq1.B99990004.pdb           2116.11377   -18231.61719        0.04183
qseq1.B99990005.pdb           2779.90063   -17003.68555        0.04731
qseq1.B99990006.pdb           2255.66235   -18772.46484        0.01694

Total CPU time [seconds]                                 :      389.18
```

As we can see that obtained molpdf, DOPE score and GA341 score for protein in the above Fig5 are uneven so for best protein selection we have chosen the model with average scores, if there are bad results, we have chosen the one with best scores for each of category and tried with those models.

- There are some case studies that we identified while carrying on the process such as, the sequence alignment didn't play a major role in obtaining an accurate template, but templates play a major and most of the role in protein modelling and in obtaining best model. The sequence alignment has very minimal role in modelling accurate 3D-structure of a protein. For example, we have considered the sequence T0792 for which we tried the sequence alignment of query with templates i.e. with both pair-wise and multiple sequence alignment along and other option with no sequence alignment for query with any of the templates. For a surprise, still the final obtained scores are the same for both cases. So, we concluded that major part of protein modelling depends on the template that we choose.



- Visual interpretation of protein doesn't always yield better results, and we can't assume that a better structured modelled has better scores. For example, if we visualize the below protein

  It looks like a very well structured protein but when we validated the results, the TM-Score obtained for above protein model is just 0.2287. hence visual interpretation based on structural analysis may completely fail for protein model.

- We also noticed that geometry of the structure plays very important roles in structuring a protein modelling, we can improve the score of protein model by refinement of the best model obtained.

- Another biggest confusion we have is with sequence alignment, at some cases we noticed that sequence alignment doesn't really play a major role. But some other cases the results were changed based on the sequence alignment that we have considered. For some cases with pair-wise sequence alignment we obtained better results, for some other cases with multiple sequence alignment. At time, we didn't find any difference between with and without alignment as both yielded same results for some of the test cases. Hence it is hard to identify the best practice but with results obtained, it is to be noted that pair-wise sequence is bit better compared to multiple but we have aligned only after identifying the template, that's a bad idea here.

Future analysis

We can perform lot more refinement to obtain better proteins. In the future, we want to identify better template by implementing machine learning techniques or by selecting wide range of databases, as we noticed that template plays a major key role in modelling od protein structure. The refinement process can be concentrated more towards the structural analysis of the protein than the sequence alignment, as the bon angles, bond lengths and other properties plays a key role in modelling of protein. Finally, we want to implement a better pipeline with visuals and easy to use integration on the front end for our pipeline there by automating the whole process of modelling and concentrating heavily on the data analysis and data mining part of protein modelling. This helps a lot in time saving and implementation process.

Acknowledgements

We would like to thank all our class mates of Wichita state university, Bio Informatics course and Dr. Debswapna bhattacharya for providing us this wonderful opportunity to learn the process of protein modelling and sharing different ideas on implementation and validation techniques

References

BLAST: Basic Local Alignment Search Tool (https://blast.ncbi.nlm.nih.gov/Blast.cgi)

PDB: Protein Data bank (http://www.rcsb.org/pdb/home/home.do)

Java Implementation of Needleman-Wunsch Algorithm: http://zhanglab.ccmb.med.umich.edu/NW-align/NWalign.java.tar.gz

T-Coffee: http://tcoffee.crg.cat/apps/tcoffee/do:regular

Modeller: https://salilab.org/modeller/

TM-score: http://zhanglab.ccmb.med.umich.edu/TM-score/

MolProbity: http://molprobity.biochem.duke.edu

Chimera: https://www.cgl.ucsf.edu/chimera/

Rasmol: https://www.umass.edu/microbio/rasmol/

CASP11 targets T0856, T0843, T0806, T0837, T0792 http://www.predictioncenter.org/download_area/CASP11/targets/

CASP10 targets T0757, T0666, T0678, T0651, T0694 http://www.predictioncenter.org/download_area/CASP10/targets/

Comparative Protein Modelling: https://en.wikipedia.org/wiki/Homology_modeling

Introduction to Homology Modelling: http://www.proteinstructures.com/Modeling/homology-modeling.html

Results and project management, Github: https://github.com/SaiManikanta23/Comparative-Protein-Modelling