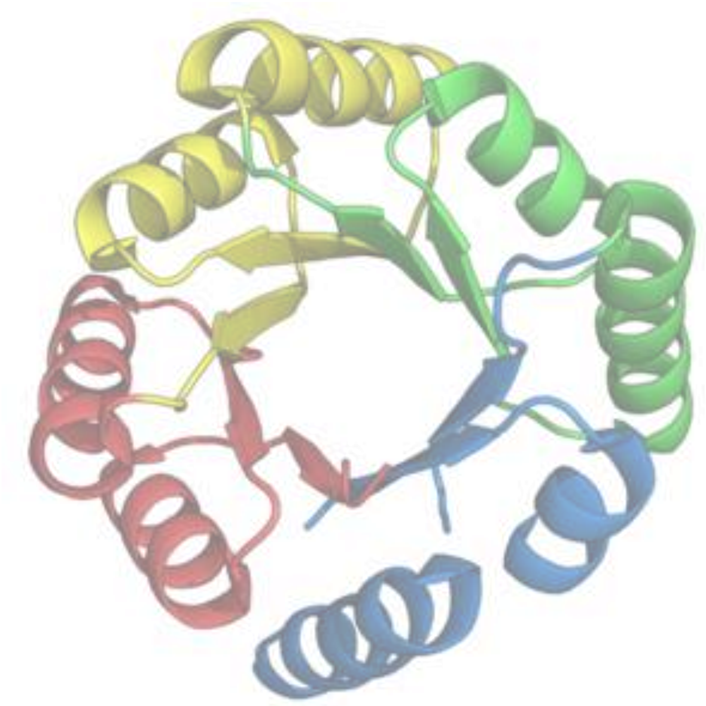
MSc in Data Science and Information Technologies

National and Kapodistrian University of Athens

Algorithms in Structural Bioinformatics

Course Project: TIM Barrels



Evangelia Giannaki (7115152100025)
Katerina Mastoraki (7115152100031)
Sophia Nteli (7115152100032)

Athens, 2022

Contents List

Introduction

The TIM-barrel fold, also known as alpha/beta barrel, is a conserved protein fold (its sequence is maintained through time and natural selection, despite all the different forces of recombination) and it was named after triose phosphate isomerase, which was the first enzyme that was crystalized ("TIM barrel", Wikipedia, para. 1). The most probable theory about their creation is that they were evolved from a common ancestor through gene duplication and domain fusion. The theory has been accepted by the scientific community as many TIM-barrel proteins possess 2-fold, 4-fold, or 8-fold internal symmetry (Sergio Romero-Romero et al. , 2021).

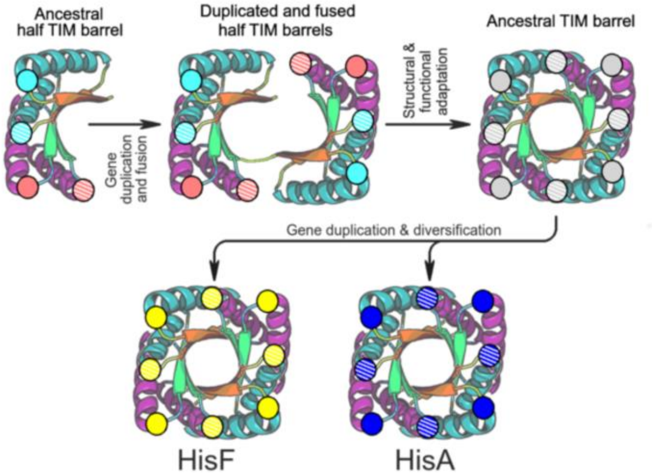

Ancestral half TIM barrel
Duplicated and fused half TIM barrels
Ancestral TIM barrel
Gene duplication and fusion
Structural & functional adaptation
Gene duplication & diversification
HisF
HisA

Figure 1: Model for the evolution of TIM barrels through gene duplication and domain fusion

This fold is adopted by approximately 10% of the enzymes and most of them consist of 8 alpha helices and 8 parallel beta strands that alternate along the peptide backbone, connecting between them through loops (Po-Ssu Huang et al., 2015). However, some TIM-barrels contain 7 or even 9 strands and helices. According to CATH classification, each protein contains from 150 to 500 residues, with an average of 298 amino acid residues per structure. Despite the low sequence homology, TIM-barrel structure is well conserved (Nagano et al., 2002). Approximately 160 residues are structurally equivalent among the different proteins while, the remaining residues are located into loop regions. The residues on the loop regions are mainly responsible for the structural maintenance and the enzymatic catalysis of the protein. Thus, each enzyme contains different subsets of residues according to its action and its exact structure and the loop regions are if needed, to contain all these distinct protein domains.

There are a lot of studies about the function and the comparison of the different TIM-barrel families, however none of them focuses on identifying the regions of the proteins which are conserved among them. In the absence of illuminating studies, we were motivated to study TIM-barrels and construct a pipeline for calculating an average structure using numerous TIM barrel PDB files.

Scope of the project

The aim of this project was the construction of a pipeline to extract a mean TIM-barrel structure.

Steps of the analysis

Data collection

The first step of the analysis was the collection of PDB IDs of TIM-barrels. For this purpose, we searched for "TIM-barrels" in the Protein Data Bank (PDB), which returned 8656 PDB entries. To ensure that our dataset contains only proteins with the TIM-barrel fold, we did the same search in CATH database, which returned 1716 results. However, some of CATH's results were not included in PDB's. We decided to continue our analysis only with the common results of both databases, which were 1144 proteins. We used PDB's Advanced Search to select the entries with refinement resolution < 2.5 Å. Then, we sorted them by the number of modeled residues of chain A and we used the first 30 results to proceed with our analysis. The sorting step was crucial to ensure that the proteins used for the analysis will contain only the TIM-barrel fold and not any other structures. The following analysis was also performed for another set of 30 PDB IDs. For this set, before sorting our results, we applied one more filter, which was the "Isomerase" classification of the proteins. We will refer to the two datasets as "IC" (Isomerase Classification) and "NIC" (No Isomerase Classification).

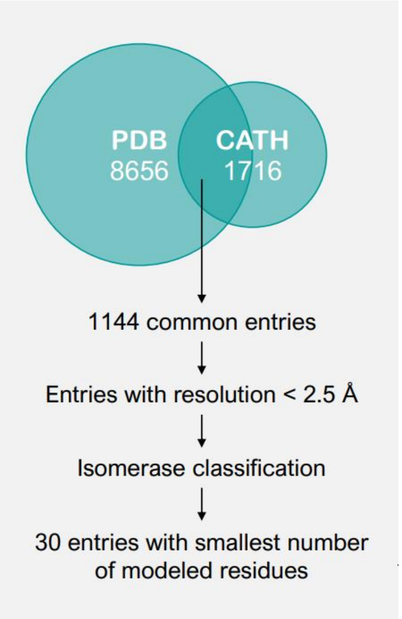


Figure 2: Data collection procedure

Multiple Structure Alignment

The next step of our analysis was the performance of Multiple Structure Alignment on the two datasets. For this purpose, we used PDBeFold.

PDBeFold

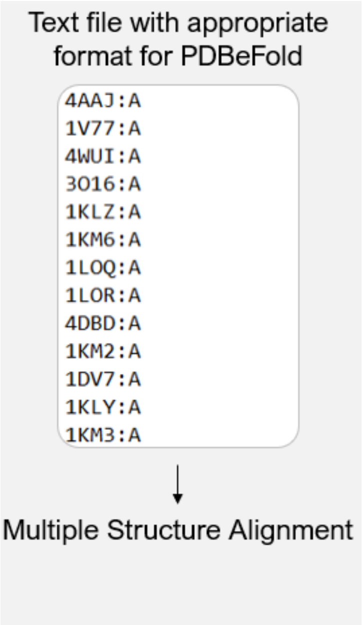
PDBeFold is a very powerful structure alignment tool which can perform both pairwise and multiple three-dimensional alignment. The structure alignment of the proteins is based on a graph-theoretical approach where the objects in question are represented in graphs, then the graphs are being matched and finally the common subgraphs found are evaluated to form conclusions about similarity.

The algorithm of PDBeFold consists of the following steps (E. Krissinel et al., 2004):

1. The secondary elements (SSEs) of each protein (helices or strands) are represented in a three-dimensional graph
2. The three-dimensional graphs representing the proteins are being matched
3. The individual residues of matched and non-matched SSEs are being mapped by their **Ca** atoms
4. Quality filters to improve the alignment (ex. unmapping Ca atoms of less similar parts) are applied.
5. The quality of the alignment using the Q-score is measured. This function offers a compromise between contradicting requirements of achieving a lower RMSD and a higher number of aligned residues and, therefore, Q is expected to be a more objective indicator of quality of alignment than RMSD and Nalign alone.

$$Q = \frac{N_{align}^2}{\left(1 + \left[\frac{RMSD}{R_0}\right]^2\right) N_1 N_2}$$

To perform the Multiple Structure Alignment, we decided to use only chain A of each protein. Until now, we have the PDB ID’s saved in a csv file. However, PDBeFold requires a specific input file format. We have created a Python script that takes as input a csv file with PDB ID’s and outputs a text file of the appropriate PDBeFold format.



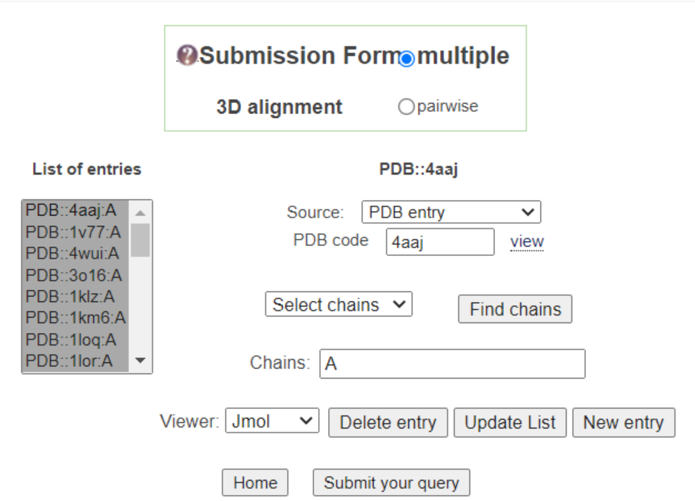
Text file with appropriate format for PDBeFold

```
4AAJ:A
1V77:A
4WUI:A
3O16:A
1KLZ:A
1KM6:A
1LOQ:A
1LOR:A
4DBD:A
1KM2:A
1DV7:A
1KLY:A
1KM3:A
```

Multiple Structure Alignment

Figure 3: Format of PDBeFold's input

To perform Multiple Structure Alignment, we launch PDBeFold, we select the "multiple" submission form and "List of PDB codes" as source. We then select the file that we generated and submit our query.



Submission Form ● multiple

3D alignment ○ pairwise

List of entries

```
PDB::4aaj:A
PDB::1v77:A
PDB::4wui:A
PDB::3o16:A
PDB::1klz:A
PDB::1km6:A
PDB::1loq:A
PDB::1lor:A
```

PDB::4aaj

Source: PDB entry

PDB code: 4aaj

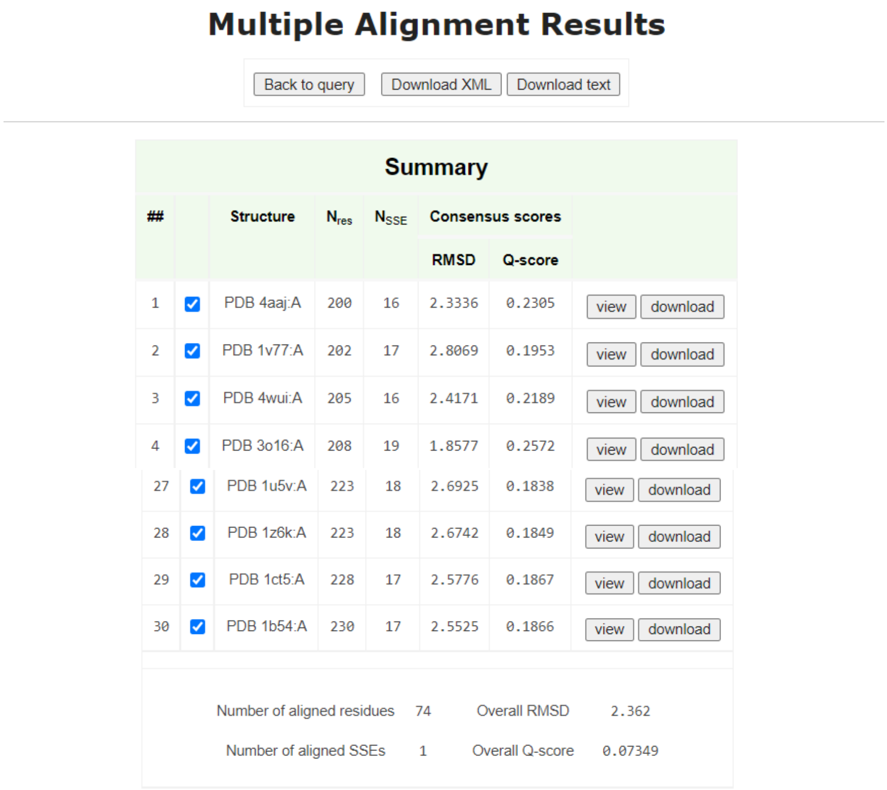
Chains: A

Viewer: Jmol

Figure 4: Multiple Structure Alignment using PDBeFold

Figures 5 and 6 show the results of the Multiple Structure Alignment for the two datasets. The Q-scores for both our experiments are too low. This means that the quality of the alignment is low, either because the RMSDs are high, or because the number of aligned residues is low. However, the IC's dataset Q-score is extremely low because the filters that we applied were too many and the proteins that we ended up with did not have similar sequence lengths.

For our analysis, we needed more proteins with small sequence length to include only the TIM-barrels fold. In the NIC dataset, we included more proteins with small sequence length, so the final alignment was better because we aligned almost only TIM-barrels.



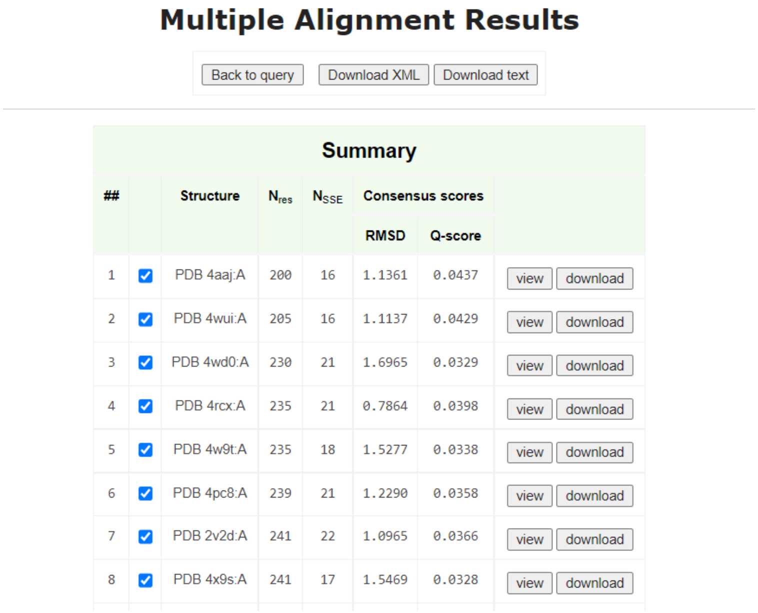
Multiple Alignment Results

Summary

##		Structure	N_{res}	N_{SSE}	Consensus scores		
					RMSD	Q-score	
1	☑	PDB 4aaj:A	200	16	2.3336	0.2305	view download
2	☑	PDB 1v77:A	202	17	2.8069	0.1953	view download
3	☑	PDB 4wui:A	205	16	2.4171	0.2189	view download
4	☑	PDB 3o16:A	208	19	1.8577	0.2572	view download
27	☑	PDB 1u5v:A	223	18	2.6925	0.1838	view download
28	☑	PDB 1z6k:A	223	18	2.6742	0.1849	view download
29	☑	PDB 1ct5:A	228	17	2.5776	0.1867	view download
30	☑	PDB 1b54:A	230	17	2.5525	0.1866	view download

Number of aligned residues	74	Overall RMSD	2.362
Number of aligned SSEs	1	Overall Q-score	0.07349

Figure 5: Multiple Structure Alignment of the NIC dataset



Multiple Alignment Results

Summary

##		Structure	N_{res}	N_{SSE}	Consensus scores		
					RMSD	Q-score	
1	☑	PDB 4aaj:A	200	16	1.1361	0.0437	view download
2	☑	PDB 4wui:A	205	16	1.1137	0.0429	view download
3	☑	PDB 4wd0:A	230	21	1.6965	0.0329	view download
4	☑	PDB 4rcx:A	235	21	0.7864	0.0398	view download
5	☑	PDB 4w9t:A	235	18	1.5277	0.0338	view download
6	☑	PDB 4pc8:A	239	21	1.2290	0.0358	view download
7	☑	PDB 2v2d:A	241	22	1.0965	0.0366	view download
8	☑	PDB 4x9s:A	241	17	1.5469	0.0328	view download

Figure 6: Multiple Structure Alignment of the IC dataset

Protein Structural Superposition

The next step of the analysis consists of the batch download of all files that correspond to the PDB IDs that we collected in the previous step by applying several assumptions(NIC or IC protein dataset) to perform structural superposition of the coordinates of each entry.

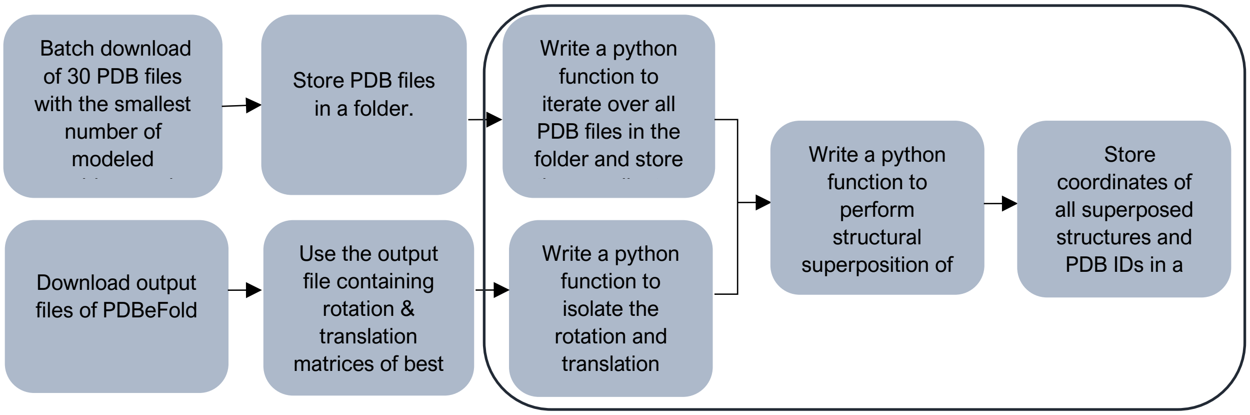

Batch download of 30 PDB files with the smallest number of modeled
Store PDB files in a folder.
Write a python function to iterate over all PDB files in the folder and store
Download output files of PDBeFold
Use the output file containing rotation & translation matrices of best
Write a python function to isolate the rotation and translation
Write a python function to perform structural superposition of
Store coordinates of all superposed structures and PDB IDs in a

Figure 7: Workflow diagram of protein structural superposition

The batch download was executed in the command line using a shell script that can download multiple PDB archive files by providing a file containing a comma-separated list of PDB IDs. The PDB files were stored in a folder. Then using a Python script, that iterates through each PDB file the coordinates of the Ca atoms of chain A of each entry were isolated. The following step was the usage of the output of PDBeFold that gives information about the rotation-translation matrices of best superposition for each protein of the dataset. The rotation and translation matrices for each protein were isolated using a Python function. To perform structural superposition of the coordinates of each PDB entry we created a Python function that transforms each coordinate by exploiting the information given by the matrices. To be more specific, each coordinate was multiplied by the rotation matrix and then the translation matrix was added to the product.

$$\begin{array}{lclcc} & & \textit{Coordinates} & & \\ \textit{Rotation matrix} & & [-3.897, 40.259,-5.661]' & & \textit{Translation matrix} \\ [-0.27468\ \ -0.06498\ \ 0.95934] & & & & [-13.26081 \\ [0.75720\ \ -0.62954\ \ 0.17416] & * & [-1.514, 38.959, -2.998]' & + & 36.28885 \\ [0.59262\ \ 0.77425\ \ 0.22212\,] & & \ldots & & -53.75269] \\ & & [1.734, 40.687, -2.044]' & & \end{array}$$

Figure 8: Transformation of the coordinates

The final step of the protein structural superposition was to store the coordinates of all superposed structures in a text file and label them with the corresponding PDB ID, respectively.

Mean structure calculation

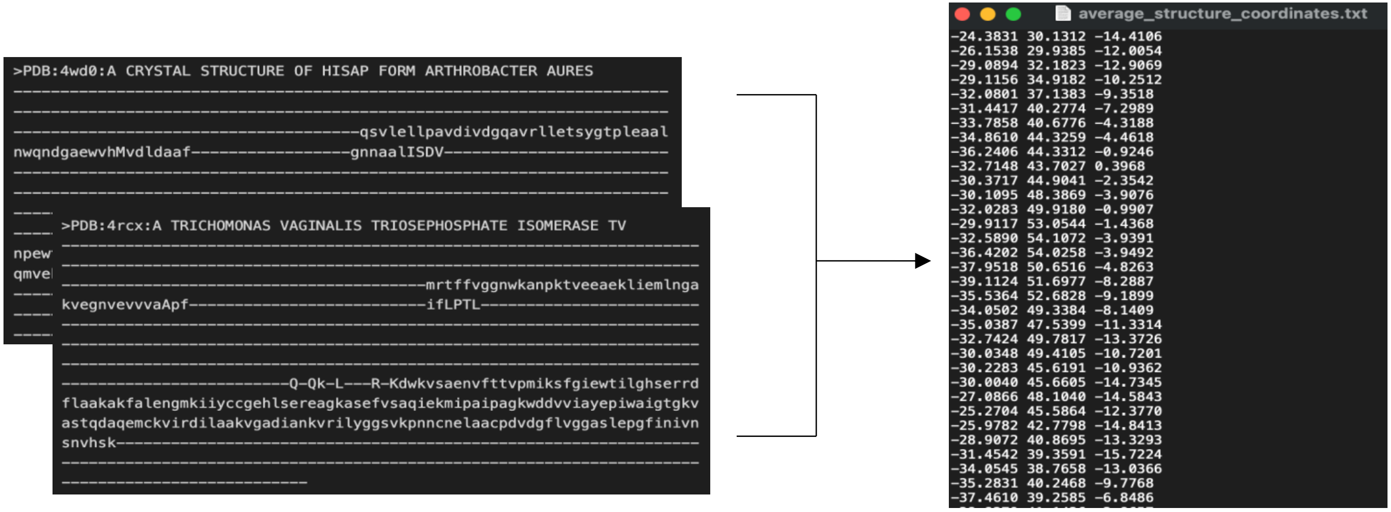
The next step in the pipeline is to calculate the coordinates of the Ca atoms of the mean structure consist of all the TIM-barrel entries of PDB we have used so far. To do that we need:

- The Fasta file of the PDBeFold results, that contains the alignment of every TIM-barrel entry
- The text file with the superposed coordinates for every TIM-barrel entry

The resulted Fasta file of the PDBeFold contains one entry for every PDB file we have used so far in the pipeline. Each entry consists of multiple lines. The first one contains the ID of the PDB file in Protein Data Bank and a small description of the TIM-barrel structure. In the next lines the actual alignment of the structure is depicted. Every position in the alignment may contain:

- A *dash*, a gap in the alignment for the specific residue (position) of the structure, compared with the overall alignment.
- A *lowercase letter*, this residue is aligned with another residue of at least another TIM-barrel structure.
- An *uppercase letter*, this residue is matched with one residue for every TIM-barrel structure.

The text file with the superposed coordinates contains one entry for every TIM-barrel structure. Each entry contains multiple lines. The first line consists of the ID of the PDB file in Protein Data Bank and each of the next lines contains the 3D superposed Ca atoms' coordinates (rotated and translated according to the protein structural superposition mentioned in the previous section) of the structure. Also, before we move on to the calculation of the mean coordinates, we transform the text file with the superposed coordinates, so every number inside a line is separated by the next number with only one space character.



```
>PDB:4wd0:A CRYSTAL STRUCTURE OF HISAP FORM ARTHROBACTER AURES
----------------------------------------------------------------------
----------------------------------------------------------------------
-------------------------------------qsvlellpavdivdgqavrlletsygtpleaal
nwqndgaewvhMvdldaaf----------------gnnaalISDV-------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----
----
npew
qmve
----
----
----
```

```
>PDB:4rcx:A TRICHOMONAS VAGINALIS TRIOSEPHOSPHATE ISOMERASE TV
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------mrtffvggnwkanpktveeaekliemlnga
kvegnvevvvaApf--------------------------ifLPTL------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
-------------------------Q-Qk-L---R-Kdwkvsaenvfttvpmiksfgiewtilghserrd
flaakakfalengmkiiyccgehlsereagkasefvsaqiekmipaipagkwddvviayepiwaigtgkv
astqdaqemckvirdilaakvgadiankvrilyggsvkpnncnelaacpdvdgflvggaslepgfinivn
snvhsk----------------------------------------------------------------
----------------------------------------------------------------------
---------------------------
```

average_structure_coordinates.txt

```
-24.3831 30.1312 -14.4106
-26.1538 29.9385 -12.0054
-29.0894 32.1823 -12.9069
-29.1156 34.9182 -10.2512
-32.0801 37.1383 -9.3518
-31.4417 40.2774 -7.2989
-33.7858 40.6776 -4.3188
-34.8610 44.3259 -4.4618
-36.2406 44.3312 -0.9246
-32.7148 43.7027 0.3968
-30.3717 44.9041 -2.3542
-30.1095 48.3869 -3.9076
-32.0283 49.9180 -0.9907
-29.9117 53.0544 -1.4368
-32.5890 54.1072 -3.9391
-36.4202 54.0258 -3.9492
-37.9518 50.6516 -4.8263
-39.1124 51.6977 -8.2887
-35.5364 52.6828 -9.1899
-34.0502 49.3384 -8.1409
-35.0387 47.5399 -11.3314
-32.7424 49.7817 -13.3726
-30.0348 49.4105 -10.7201
-30.2283 45.6191 -10.9362
-30.0040 45.6605 -14.7345
-27.0866 48.1040 -14.5843
-25.2704 45.5864 -12.3770
-25.9782 42.7798 -14.8413
-28.9072 40.8695 -13.3293
-31.4542 39.3591 -15.7224
-34.0545 38.7658 -13.0366
-35.2831 40.2468 -9.7768
-37.4610 39.2585 -6.8486
```

Figure 9: Calculation of the coordinates of the mean structure

To calculate the coordinates of the mean structure we follow two main steps:

- For every entry of the Fasta file we look at a specific position (we start with the first and we move along until no more residues exist). For this specific position we add the superposed coordinates of all the TIM-barrel structures that contains either an uppercase or a lowercase letter.
- When we've checked all the entries for that specific position, we divide with the total number of the PDB entries that contributed to its coordinates' calculation.

Once the previous process is completed, then we save the coordinates into a new text file. This file contains the coordinates for every Ca atom and form the mean structure of all the TIM-barrels we have used in the pipeline. As we have already discussed, we executed the pipeline for two different groups of 30 files each. The text file produced from this step contains:

- 867 3D coordinates – group of isomerase classification (IC protein dataset).
- 446 3D coordinates – group with the smallest residues' number (NIC protein dataset).

Visualization of the mean structure

The final step of the pipeline is to use the mean coordinates we have calculated to visualize the mean structure and get a 3D representation of the 30 PDB files for the IC and NIC datasets.

For this purpose:

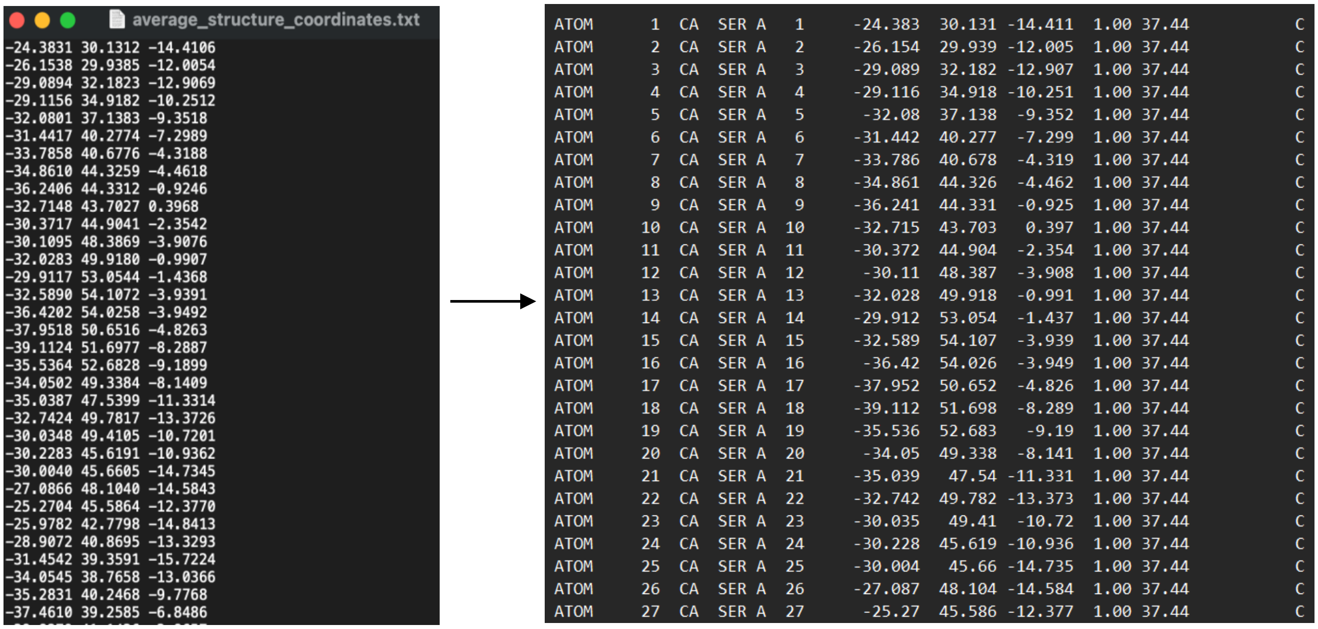
- ✓ UCSF ChimeraX v1.3, a molecular visualization program, was used
- ✓ A PDB file containing the coordinates of the mean structure was created, for the tool to be easier to interpret the results to a 3D structure

To create the PDB file to use it as input to the ChimeraX, we created a Python script that gets as input the text file containing the coordinates of the mean structure. For each line of the text file that corresponds to a triplet of (X, Y, Z) coordinates, we concatenate it with strings that correspond to:

1. The number of the atom, which increases along with the
2. The atom type, which was set to Ca since we have used only carbon atoms of the backbone of the proteins to calculate the mean structure
3. The amino acid type, which randomly was set to SER to all atoms
4. The chain ID, which corresponds to A based on the initial assumptions
5. The residue numbers
6. The coordinates (X, Y, Z) of the mean structure
7. The atom occupancy: most of the time the number is 1.0 (100%) because atoms occupy only one position

8. The atomic displacement parameter, which describes the displacement of atoms from an equilibrium position, and it was randomly set to 37,44
9. The element of the atom, which was set to C

The new lines are written in a text file which simulates a PDB file format, as seen below.



```
-24.3831 30.1312 -14.4106
-26.1538 29.9385 -12.0054
-29.0894 32.1823 -12.9069
-29.1156 34.9182 -10.2512
-32.0801 37.1383 -9.3518
-31.4417 40.2774 -7.2989
-33.7858 40.6776 -4.3188
-34.8610 44.3259 -4.4618
-36.2406 44.3312 -0.9246
-32.7148 43.7027 0.3968
-30.3717 44.9041 -2.3542
-30.1095 48.3869 -3.9076
-32.0283 49.9180 -0.9907
-29.9117 53.0544 -1.4368
-32.5890 54.1072 -3.9391
-36.4202 54.0258 -3.9492
-37.9518 50.6516 -4.8263
-39.1124 51.6977 -8.2887
-35.5364 52.6828 -9.1899
-34.0502 49.3384 -8.1409
-35.0387 47.5399 -11.3314
-32.7424 49.7817 -13.3726
-30.0348 49.4105 -10.7201
-30.2283 45.6191 -10.9362
-30.0040 45.6605 -14.7345
-27.0866 48.1040 -14.5843
-25.2704 45.5864 -12.3770
-25.9782 42.7798 -14.8413
-28.9072 40.8695 -13.3293
-31.4542 39.3591 -15.7224
-34.0545 38.7658 -13.0366
-35.2831 40.2468 -9.7768
-37.4610 39.2585 -6.8486
```

```
ATOM      1  CA  SER A   1     -24.383  30.131 -14.411  1.00 37.44           C
ATOM      2  CA  SER A   2     -26.154  29.939 -12.005  1.00 37.44           C
ATOM      3  CA  SER A   3     -29.089  32.182 -12.907  1.00 37.44           C
ATOM      4  CA  SER A   4     -29.116  34.918 -10.251  1.00 37.44           C
ATOM      5  CA  SER A   5      -32.08  37.138  -9.352  1.00 37.44           C
ATOM      6  CA  SER A   6     -31.442  40.277  -7.299  1.00 37.44           C
ATOM      7  CA  SER A   7     -33.786  40.678  -4.319  1.00 37.44           C
ATOM      8  CA  SER A   8     -34.861  44.326  -4.462  1.00 37.44           C
ATOM      9  CA  SER A   9     -36.241  44.331  -0.925  1.00 37.44           C
ATOM     10  CA  SER A  10     -32.715  43.703   0.397  1.00 37.44           C
ATOM     11  CA  SER A  11     -30.372  44.904  -2.354  1.00 37.44           C
ATOM     12  CA  SER A  12      -30.11  48.387  -3.908  1.00 37.44           C
ATOM     13  CA  SER A  13     -32.028  49.918  -0.991  1.00 37.44           C
ATOM     14  CA  SER A  14     -29.912  53.054  -1.437  1.00 37.44           C
ATOM     15  CA  SER A  15     -32.589  54.107  -3.939  1.00 37.44           C
ATOM     16  CA  SER A  16      -36.42  54.026  -3.949  1.00 37.44           C
ATOM     17  CA  SER A  17     -37.952  50.652  -4.826  1.00 37.44           C
ATOM     18  CA  SER A  18     -39.112  51.698  -8.289  1.00 37.44           C
ATOM     19  CA  SER A  19     -35.536  52.683   -9.19  1.00 37.44           C
ATOM     20  CA  SER A  20      -34.05  49.338  -8.141  1.00 37.44           C
ATOM     21  CA  SER A  21     -35.039   47.54 -11.331  1.00 37.44           C
ATOM     22  CA  SER A  22     -32.742  49.782 -13.373  1.00 37.44           C
ATOM     23  CA  SER A  23     -30.035   49.41  -10.72  1.00 37.44           C
ATOM     24  CA  SER A  24     -30.228  45.619 -10.936  1.00 37.44           C
ATOM     25  CA  SER A  25     -30.004   45.66 -14.735  1.00 37.44           C
ATOM     26  CA  SER A  26     -27.087  48.104 -14.584  1.00 37.44           C
ATOM     27  CA  SER A  27      -25.27  45.586 -12.377  1.00 37.44           C
```

Figure 10: Creation of the PDB file to be used as input to ChimeraX.

Finally, the PDB file of the mean structure is loaded in ChimeraX and as a result we get the 3D visualization of it. Below we observe the representation results of the mean structure extracted from the IC and NIC dataset.

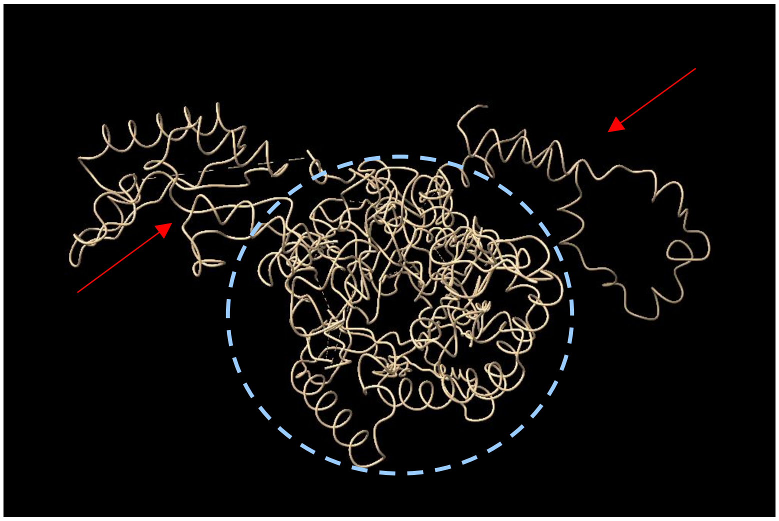


Figure 11: 3D visualization of the mean structure of the IC dataset

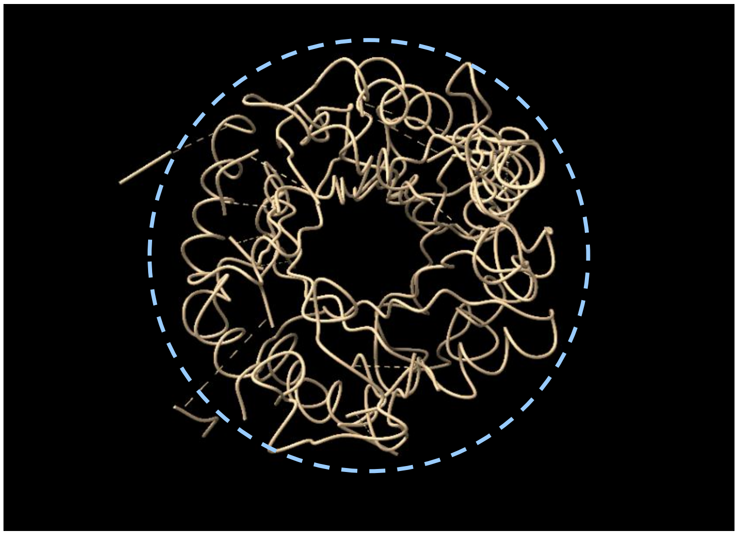


Figure 12: 3D visualization of the mean structure of the NIC dataset

Conclusions

Based on the results of our experiments as presented in Figure 11 and 12 we can conclude that the NIC dataset gives a better representation of a mean TIM-Barrel structure than the IC dataset, because it includes almost only structures that belong to the TIM-barrel fold.

To get a better understanding of our results, we launched PDBeFold of the 30 proteins with no isomerase classification along with the mean structure that we have calculated. The goal was to find a minimum and a maximum value of the RMSD score of the structural alignment.

Unfortunately, PDBeFold did not produce any results. Our hypothesis is that either there was an external web - server error in running PDBeFold, or the PDB format that we have created is not compatible with the rest of PDB files.

Future work

To expand the analysis, we can create a better PDB format of the mean structure and try to re-run the structural alignment using PDBeFold and observe the minimum and the maximum values of the RMSD scores. Furthermore, we can find a distribution of the RMSD values of all pairs of the 30 proteins of the NIC dataset and then check if the median or the mean value of the RMSD values of the mean structure along with the 30 proteins is a good representative.

Bibliography

1. Nozomi Nagano, Christine A Orengo, Janet M Thornton. (2002). One fold with Many Functions: The Evolutionary Relationships between TIM Barrel Families Based on their Sequences, Structures and Functions. Journal of Molecular Biology, Volume 321, Issue 5.
2. Sergio Romero-Romero, Sina Kordes, Florian Michel, Birte Höcker. (2021) Evolution, folding, and design of TIM barrels and related proteins. Current Opinion in Structural Biology, Volume 68.
3. Kevin T Halloran , Yanming Wang, Karunesh Arora, Srinivas Chakravarthy , Thomas C Irving, Osman Bissel, Charles L Brooks 3rd, C Robert Matthews. (2019). · Frustration and folding of a TIM barrel protein. PNAS, Volume 116.
4. 'TIM barrel' (2020). Wikipedia. Available at: https://en.wikipedia.org/wiki/TIM_barrel (Accessed: 22 February 2022).
5. Po-Ssu Huang , Kaspar Feldmeier , Fabio Parmeggiani , D Alejandro Fernandez Velasco, Birte Höcker, David Baker. (2015). De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. Natural Chemical Biology.
6. E Krissinel, K Henrick. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallographica Section D.
7. 'Special relativity' (2018). Wikipedia. Available at: https://en.wikipedia.org/wiki/Special_relativity (Accessed: 5 August 2019).