|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | |
|  | | Project 2: Template-based Protein Structure Modeling | | | | | |  |
|  |  | | | | | |  | |
|  | | | |  |  | | | |
|  | | | | Jian Liu, Yanli Wang, Weinan Zhang |  | | | |
|  | | | | Feb,25,2022CMP\_SC 8170 Computational Modeling of Molecular Structures Supervised by Dr. Jianlin Cheng |  | | | |
|  | | |  | | |  | | |

**Project Content**

Analyze COVID-19 Spike Protein Sequence and 5 “all group” CASP14 Single-chain Protein Targets by building, analyzing, and visualizing multiple sequence alignments of these proteins

**Project Tasks**

1. Data Collection
2. Database Collection
3. MAS Generation
4. Quality Comparison and Analysis

**Project Procedure**

1. **Targets Download**

Covid-19 spike protein sequence and 5 “all group” targets sequences were found online. We chose 5 FM targets with different PDB codes

|  |  |  |
| --- | --- | --- |
| Target | PDB | FASTA Sequence Source |
| P0DTC2 |  | https://www.uniprot.org/uniprot/P0DTC2#sequences |
| T1027 | 7D2o | https://www.rcsb.org/sequence/7D2O |
| T1049 | 6y4f | https://www.rcsb.org/structure/6Y4F |
| T1064 | 7jtl | https://www.rcsb.org/structure/7JTL |
| T1074 | 7oc9 | https://www.rcsb.org/structure/7oc9 |
| T1090 | 7k7w | https://www.rcsb.org/structure/7k7w |

1. **MSA Generation**

5 databases were used to generated MSAs for 6 protein sequences. As different databases work with different tools, 4 apps were used in the experiment (Hhblits, Jackmmer,DeepMSA, and alignment generation pipeline in Alphafold).

* UniRef Databases
* UniProt Databases
* Big Fantastic Databases (BFD)
* Metagenomics Databases
* ColabFold Databases

1. **MSA Qualitys**

* **Neff**

One way to measure the quality of sequency alignment is by the looking at the normalized number of effective sequences (Neff) in an MSA.

A picture containing text, watch

Description automatically generatedDiagram, schematic

Description automatically generated

Where N is the sequences in the MSA, and is the sequence identity between the m-th and n-th sequences. is equal to 1 if , or zero otherwise. is set to 0.8 in the example showing in Fig 1.

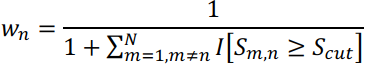
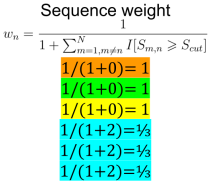
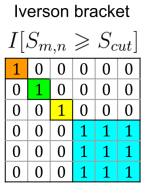
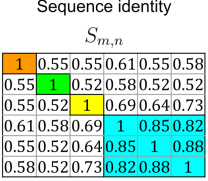
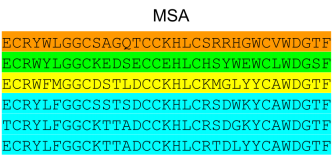


Fig 1. calculation of sequence weights and the number of effective sequences.

Table below show the quality comparison of the MSAs Neff result

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Target | Alpha Fold | Colab Fold | Deep MSA | UniRef 30 | UniProt | Bfd | Mgnify |
| T1027 | 36.9 | 30.5 | 4.8 | 4.7 | 5.7 | 36.6 | 1 |
| T1049 | 80.4 | 56.1 | 151.9 | 125.6 | 31.7 | 29.7 | 16.1 |
| T1064 | 6.4 | 7.4 | 5.1 | 6 | 4.6 | 4 | 1 |
| T1074 | 10.3 | 12.3 | 8 | 6 | 5 | 5 | 1 |
| T1090 | 54.4 | 77.9 | 94.7 | 50.4 | 32.6 | 45.9 | 1 |
| PoDTC2 | 15.8 | 45.5 | 32.4 | 21.6 | 12 | 13.5 | 1 |
| Average | 34.03333 | 38.28333 | 49.48333 | 35.71667 | 15.26667 | 22.45 | 3.51667 |

According to the result, Colab Fold and Deep MSA have the top two best qualities of Neff

* **Global Distance Test Total Score (GDT-TS)**

GDT-TS = , where Pi is the percentage of Cα atoms whose distance errors between predicted and native structure are within i =(1, 2, 4, 8) Å. Table below shows the quality result of GDT-TS

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Target | Alpha Fold | Colab Fold | Deep MSA | UniRef 30 | UniProt | Bfd | Mgnify |
| T1027 | 0.38306 | 0.38542 | 0.35356 | 0.34972 | 0.35626 | 0.38362 | 0.29762 |
| T1049 | 0.92686 | 0.94106 | 0.93318 | 0.92052 | 0.93693 | 0.91306 | 0.91868 |
| T1064 | 0.30148 | 0.32402 | 0.3446 | 0.22352 | 0.2711 | 0.29508 | 0.23578 |
| T1074 | 0.90798 | 0.90302 | 0.92346 | 0.91176 | 0.89432 | 0.8553 | 0.36516 |
| T1090 | 0.89894 | 0.90106 | 0.88148 | 0.87352 | 0.85266 | 0.8746 | 0.14868 |
| Average | 0.683664 | 0.690916 | 0.687256 | 0.655808 | 0.66225 | 0.664332 | 0.39318 |
| Correl | 0.516723 | 0.570174 | 0.651119 | 0.584762 | 0.65832 | 0.275856 | 0.96507 |

According to the table: from average Neff we can find that the two databases Colab Fold and Deep MSA can get more effect sequences for the six query sequences that aligned against than the rest of other five databases, at the same time we can get the top two average GDT-TS for the databases Colab Fold and DeepMSA corresponding to the top two Neff databases based on the alignment. However, when calculating the correlation of Neff and GDT-TS, we find that both the highest Neff and GDT-TS does not mean they have the strongest relation between them; the lowest Neff combined with lowest GDT-TS for Mgnify has the strongest correlation.

Fig 2. T1090 3D structure

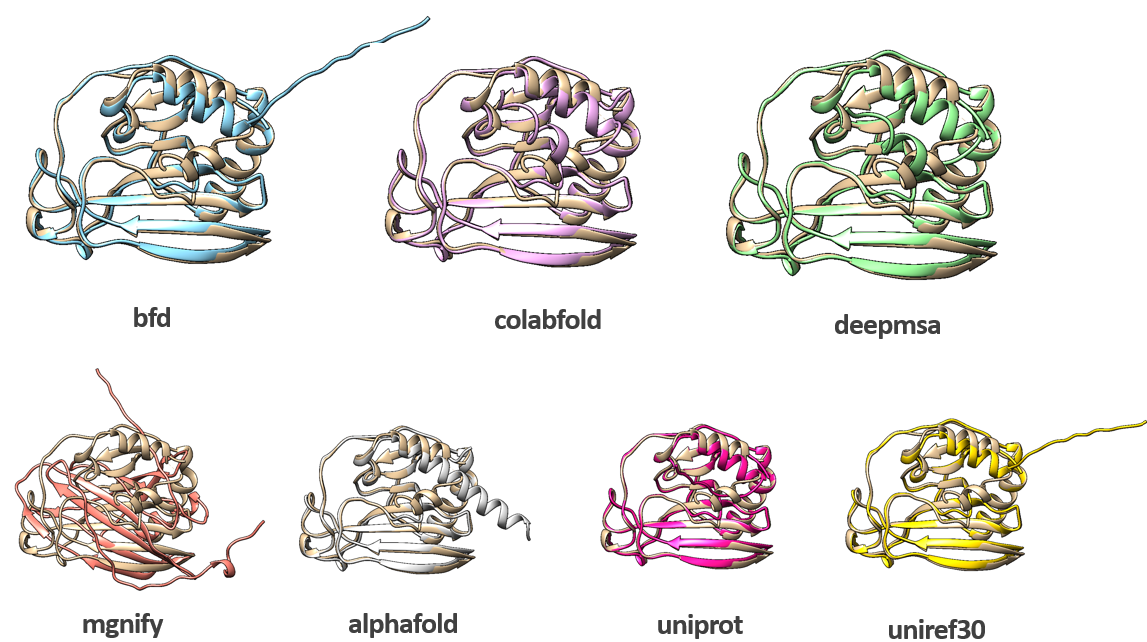


Fig 2 shows the 3D structure of T1090 sequences draw by chimera

1. **MSA Generation**

Finally, the MSAs visualization was achieved by using a website NCBI Multiple Sequence Alignment Viewer.

Fig 3 is a picture of the P0DTC2\_BFD database visualization demo. For more information, please check the link below:

[https://www.ncbi.nlm.nih.gov/projects/msaviewer/?key=12RNvctmFL4SFq4m1lsRjzwrjLxQ\_9G80DbVs9S3XxMcjFk5F6vA\_kUzI7EdrsL6hO3B6kj8ia1PKExpzerLJkFpAuIN6I,TP\_WJlD9jyQj08EjAML33adn6\_m0iLqNtouenYqZmLcJuXSvoZU\_atVW4GkxcScMdhQrADUkbiEpOz02OzA3KwUCOAwUMD4](https://www.ncbi.nlm.nih.gov/projects/msaviewer/?key=12RNvctmFL-4SFq4m1lsRjzwrjLxQ_9G80DbVs9S3XxMcjFk5F6vA_kUzI7EdrsL6hO3B6kj8ia1PKExpzerLJkFpAuIN6I,TP_WJlD9jyQj08EjAML33adn6_m0iLqNtouenYqZmLcJuXSvoZU_atVW4GkxcScMdhQrADUkbiEpOz02OzA3KwUCOAwUMD4)

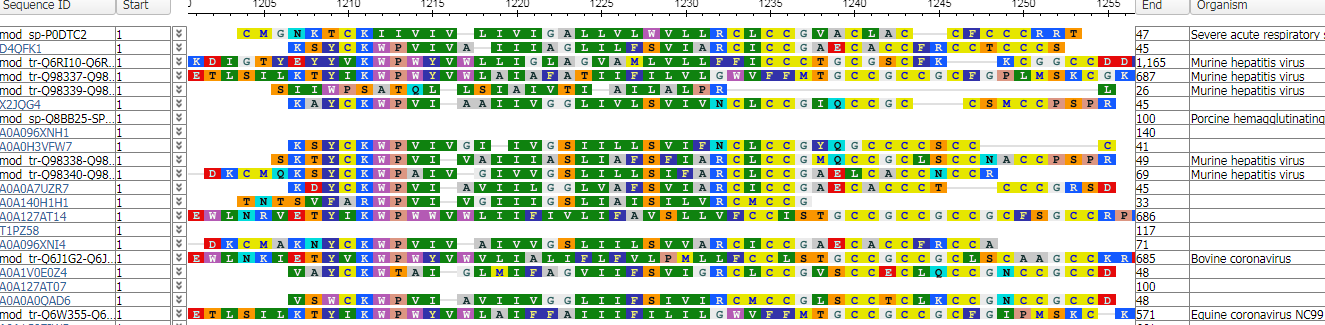


Fig 3. P0DTC2 MSA Visualization

**Conclusion:**

In this project, COVID-19 spike protein sequence and 5 single-chain protein targets were searched, aligned and their MSAs were visualized. With the help of different tools, we now have a better understanding of the sequence alignment and how their structures look like in 3D.