CASP Covid-19

Bikash Shrestha

April 2020

1 Introduction

Critical Assessment of Structure Prediction (CASP) is a community-wide experiment to determine and advance the state of the art in modeling protein structure from amino acid sequence. It provides researchers a platform and opportunity to test their structure prediction methods and helps to improve the 3D structure prediction of the protein. They organize a competition every two years where participants from around the world participate in it.

This year the outbreak of novel coronavirus pandemics (COVID-19) set new challenges for the biomedical scientific community. So CASP organizes a competition to predict the structure modeling of SARS-2-CoV protein sequences. By generating and evaluating models for the virus, it can help to understand the spread of the virus and how it functions. The goal is to find the best possible structure hoping it can be useful to others for gaining further insight into the virus structure and function.

The CASP committee provides the target sequence and the groups that are participating in the competition will try to predict the best structure for those targets and submit them. For this competition, CASP Common COVID-19, they released 10 targets for the groups to predict the structure.

2 Our Methodology

The block diagram given below shows the flow of the methods that we have performed for this competition.

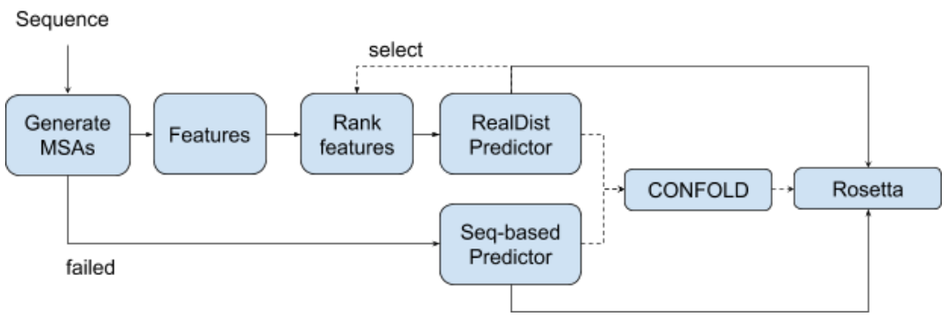

Figure 1: Block diagram showing methodology

2.1 MSA Generation

Multiple sequence alignments were generated for each target. All the alignments are generated using DeepMSA. For each target, three alignments were generated with different threshold values: e0.001, e0.01 and e0.1. The table given below shows the counts of the alignment for each target with their respective threshold values.

As we can see from the table, target C1909 has very fewer alignments that cannot be used for feature extraction using template-based modeling.

Target	L	E-value threshold			E=0.01 & Coverage 35
		0.001	0.01	0.1	
C1901	638	249	253	257	
C1902	500	499	501	501	
C1903	290	357	357	350	
C1904	686	166	341	167	
C1905	275	26	20	20	
C1906	222	160	160	161	
C1907	61	12	12	12	
C1908	121	12	12	1919	
C1909	38	1	1	1	
C1910	43	4	4	2141	

Figure 2: Table showing showing aligments count

2.2 Feature Generation

Using these MSAs features are generated and using these features distance map of the protein was created. However, for target C1909 which does not have MSA, a different approach was used to find the distance map. The distance map was generated directly from the input sequence. For every e-value threshold, the distance map was created. The visualization of the distance map can be seen in the given images below.

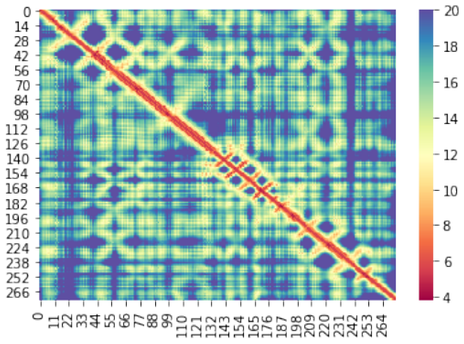

Figure 3: Distance map of target C1905

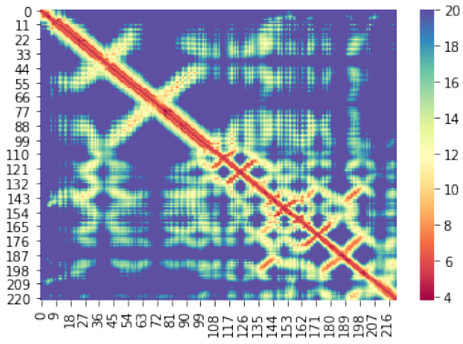

Figure 4: Distance map of target C1908

2.3 Constraints and Fragments Generation

These same distance maps were used to generate the constraint files for the Rosetta which were used in the model building process. The constraint weight used while generating the file ranges from 0.1 to 1.0. Along with the constraint files, another type of file called fragment files were also created which are required for model building in Rosetta.

2.4 Model Building

We feed these files in Rosetta and let it run to build the 200 to 400 models from which one best model was selected using the score value. The table given below shows the number of models built for each target. Because of the limitation in time and resource we were not able to build more models. Those best models were then submitted in the competitions. Before submitting the model various comparison and analyses were done to check the reliability and structure of the 3D model.

Target Name	Number of Models Built
C1901	25
C1902	17
C1903	150
C1904	12
C1905	200
C1906	100
C1907	350
C1908	420
C1909	600
C1910	410

Table 1: Number of models built per target

3 Results and Analysis

The distance map generated from the predicted 3D model was compared against the real predicted distance map from our deep learning model. The given image shows a comparison between those distance maps. The distance maps on the left are the distance maps generated from the deep learning model and the maps on the right are obtained from the predicted 3D model.

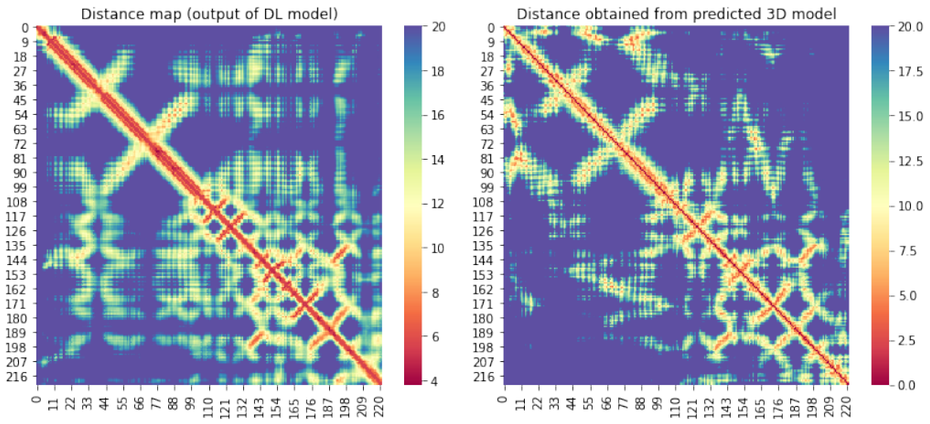

Figure 5: Distance maps comparison of target C1906

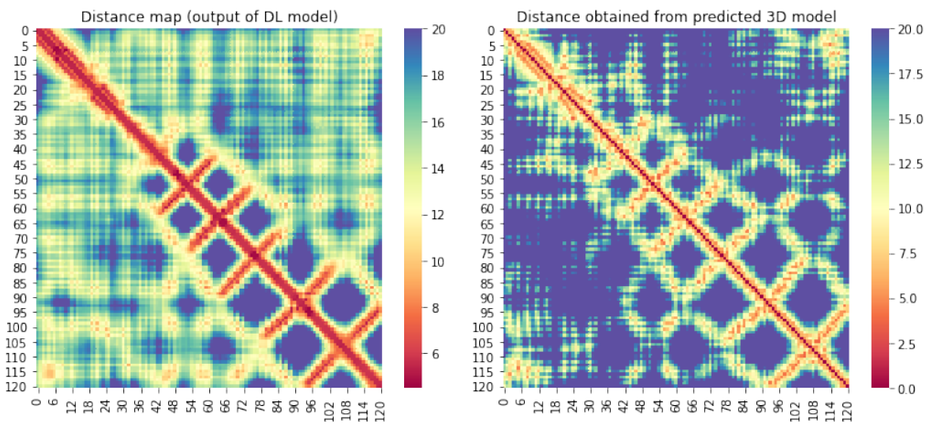

Figure 6: Distance maps comparison of target C1908

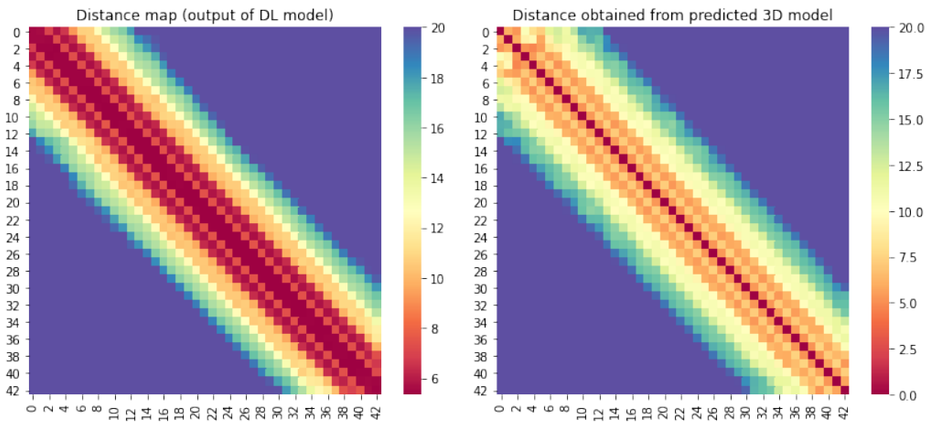

Figure 7: Distance maps comparison of target C1910

Target Name	Compared with	TM-score	RMSD
C1910	Server Group	0.65882	2.05
C1908	Zhang's Labe	0.34965	5.24
C1907	Baker's Lab	0.77837	1.69

Table 2: TM-scores and RMSD values of some models

Along with that various comparisons were made with 3D models of other groups against our models. The software called Chimera was used to compare the 3D structures of the predicted models. 3D models predicted by other groups like Deepmind, Zhang's lab, Baker's Lab as well as models generated by the server groups were compared with our models. The TM-score was calculated for the models by comparing them with other models. Some of the TM-score along with RMSD value are given in the table below.

The images given below are some of the 3D structure comparison of the models. The structure in golden color is our model and the model in blue color is other group's model.

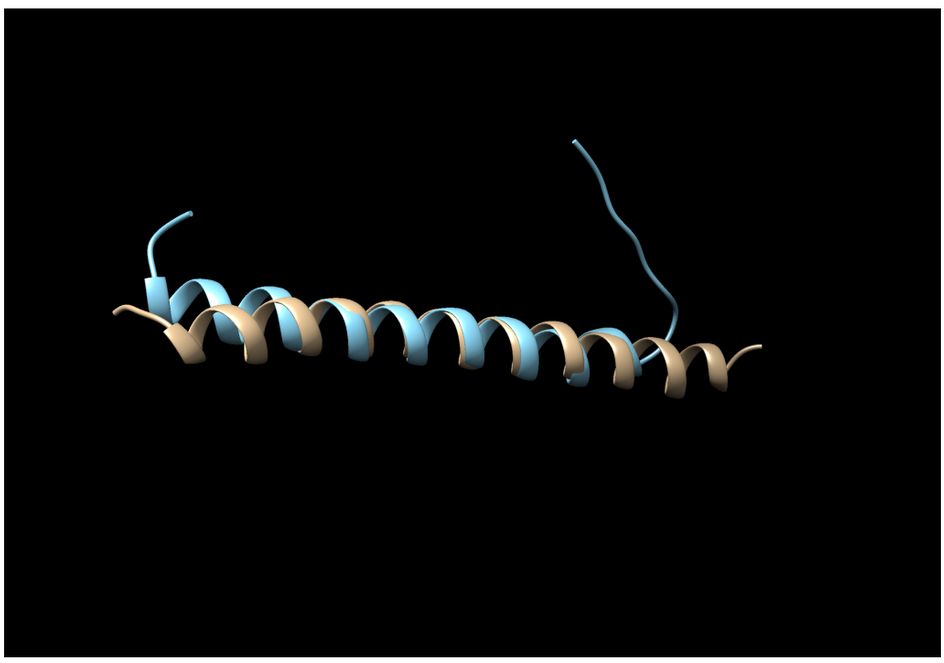

Figure 8: 3D structure comparison of target C1910 with Server Group Model

Figure 9: 3D structure comparison of target C1907 with Baker's Lab model

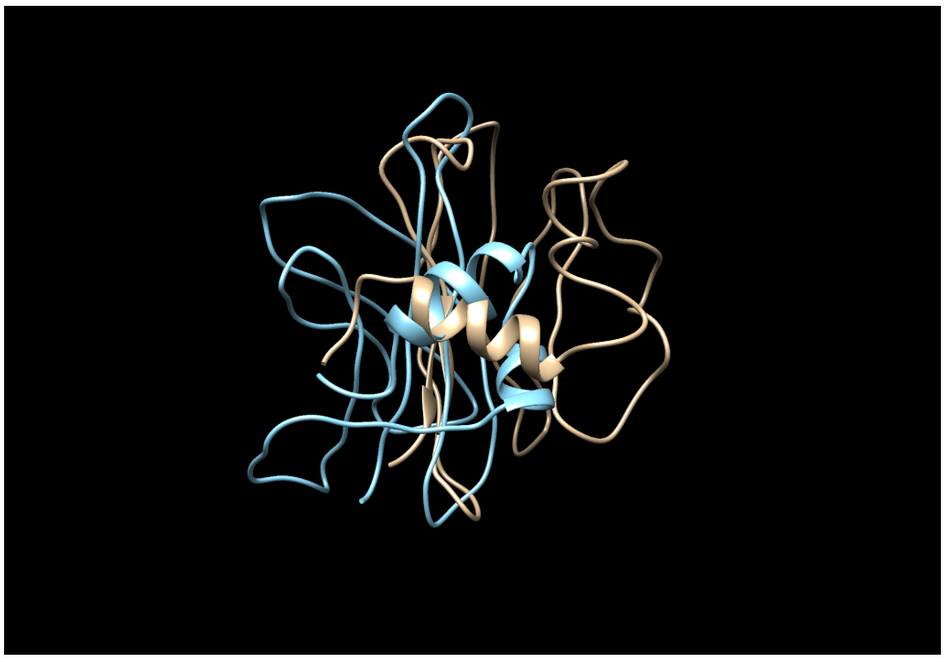

Figure 10: 3D structure comparison of target C1908 with Zhang's Lab Model

4 Conclusion

Since there were time limitations and fewer resources due to we could not build enough 3D models for some of the targets. The 3D models could have been better if there were more time. Moreover, we could not find the quality of the distance maps that were used in the model building because of time limitations. Despite that we were able to submit our prediction for all 10 targets within the time. Some of the models are comparable with the models of groups like Deepmind and Zhang's lab. Nevertheless, it might be useful for the research of SARS-2-COV and this will also help for future research.