Computation Biology Project Two Protein Structure Prediction
Name: Peng-Jhen,Lee; Tsai-Yi, Chen; Jiaying TU

Introduction

In the previous project, we rebuilt the sequence, and in this project, we try to build protein structure. However, building the protein structure in the traditional way (X-ray crystallography) is time-consuming. Therefore, in this project, we would predict the protein structure of Salmonella that is high resistant to tetracycline.

Data

According to the provided sequence of Tet Repressors protein, we first use this sequence to find multiple sequence alignmnent (MSA) by Pfam. The result of MSA can identify the evolutionary relationships that may descend of common ancestors. From the mutation of deleting or inserting of the single amino acids may produce the difference between the sequence, and this could help us to understand the conserved sequence, secondary structure, or protein tertiary structure.

Methods and Model

Mutual Information

To estimate the coevolutionary relationship between two positions of the amino acids, we could use mutual information theory (MI). MI could provide information of the amino acid on position how it interacted with the other amino acid on the other amino acid.
The method of the MI is as follows

$$\boldsymbol{MI(i,j) = \sum_{a,b} P(a_i, b_j) log(\frac{P(a_i, b_j)}{P(a_i)P(b_j)})}$$

Before computing the MI, we first need to trim the original MSA, since there were numerous gaps compared to the reference sequence (we selected the first sequence as the reference sequence). If all positions contained more than 50% gaps, we would delete the certain position in all sequences. After the computation, the threshold would be 0.55 to build the contact map, which could observe whether the amino acid connected with each other or not.

Protein Structure

According to the contact map, we would build the protein structure by the Fault Tolerant Contact Map Reconstruction (FT-COMAR). It is a program that is easily be used. After predicting the protein structure, we would use PyMol to calculate the Root Mean Square Deviation (RMSD), which could measure the similarity between the predicted and the original structures. RMSD is calculated the by comparing the coordinates between predicted

and known protein structures.

Result

When we compared two protein structures in Pymol, if we calculated the RMSD directly, we would get a large RMSD value. Hence, it is important to align two protein structures first by translating and rotating the predicted protein structure to get the minimum RMSD. We use the align function in Pymol and by testing different parameters, we set the gap value as 30.
The command is: Pymol> align predicted_protein, tetR, gap=30

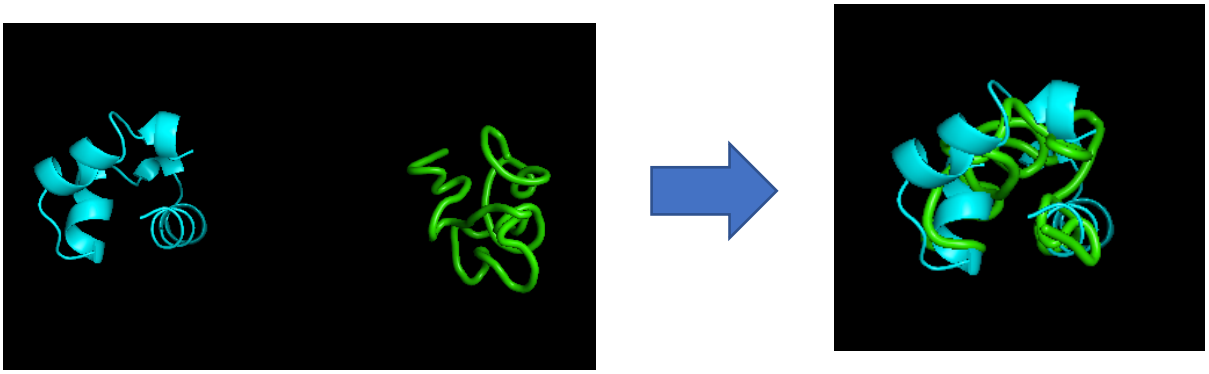


Figure 1 Comparison two protein structure without alignment (left)

RMSD = 2.451 (5 to 5 atoms)

Figure 2 Comparison two protein structure with alignment(right)

RMSD = 5.637 (29 to 29 atoms)

We observed that although the RMSD between the two protein structures before alignment was relatively small, only five atoms were aligned, with an average error of 0.49 to each atom, while the RMSD of the two protein structures after alignment was 5.637, with an average error of 0.194 between each atom.
In practice, it is very difficult to find the predicted protein structure that most closely resembles the expected protein structure. Because we need to consider the following parameters at the same time:

1. The length of the original MSA after cutting

We tested MSA contact maps with lengths of 44, 48, 49, 50, 51, and 54. Comparing hundreds of protein structures with the RMSD of the original structure by Pymol and we found that the MSA length of 49 was the most appropriate, possessing a relatively small RMSD.

2. The threshold used to build the contact map

The smaller our threshold is, the curlier the structure will be, which was shown below. Thus, we needed to choose a suitable threshold

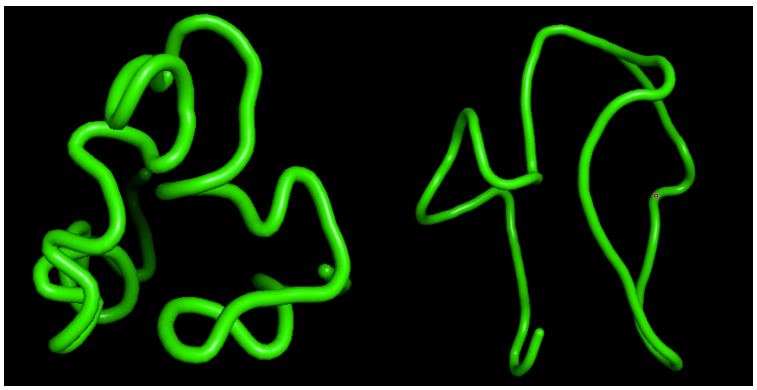


Figure 3: Predicted protein when threshold is 0.03 (left)

Figure 4: Predicted protein when threshold is 0.07 (right)

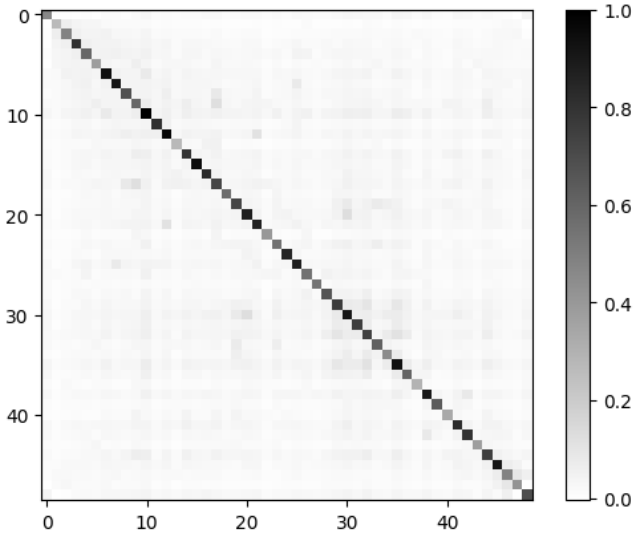


Figure 6: Heat Map of MI scores

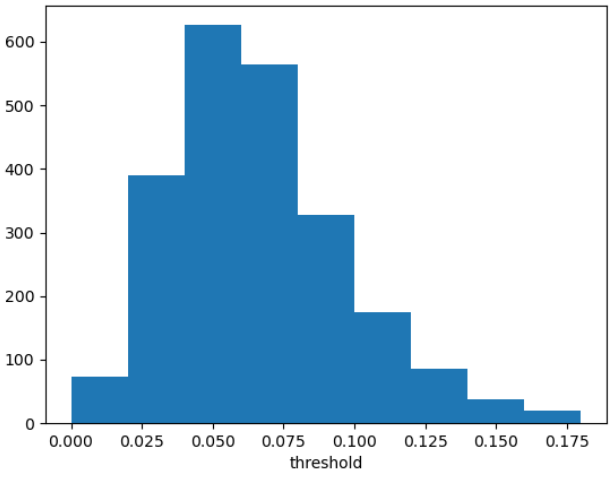


Figure 5: Histogram of the distribution of MI scores

We could find that MI values are mostly distributed around 0.05. After several tests of different thresholds, as the threshold value was 0.055, the RMSD was the smallest, so we set the threshold value for building the map to 0.055.

3. The parameter threshold used by FT-COMAR

When FT-COMAR uses a smaller parameter threshold, the more curled the protein structure is, but we do not know what it should be set to that would be closer to the original structure, so for each contact map, we choose to measure all the threshold values from 7 to 18.

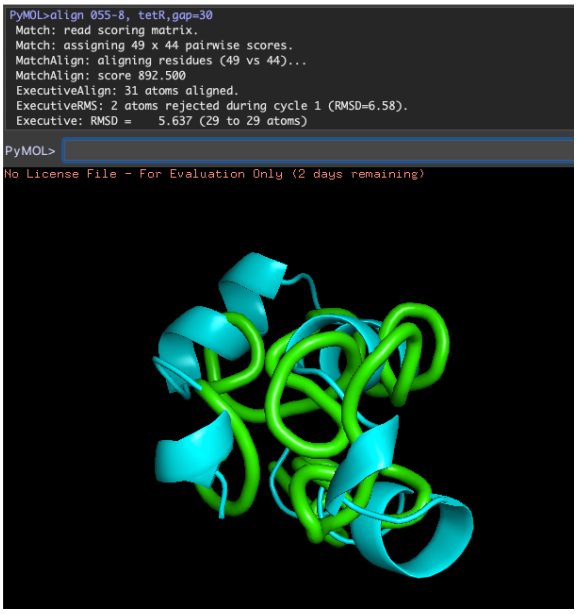
4. whether FT-COMAR uses common neighbors filter.

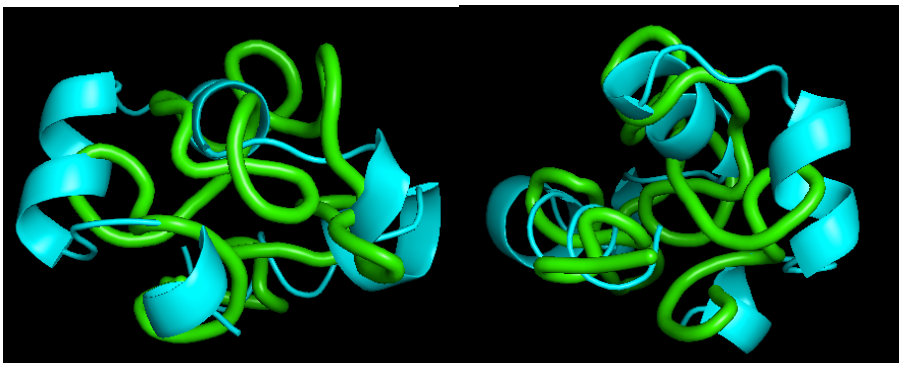
We found that the difference between the structure when set or not is not big, so we uniformly do not use filter

The best result

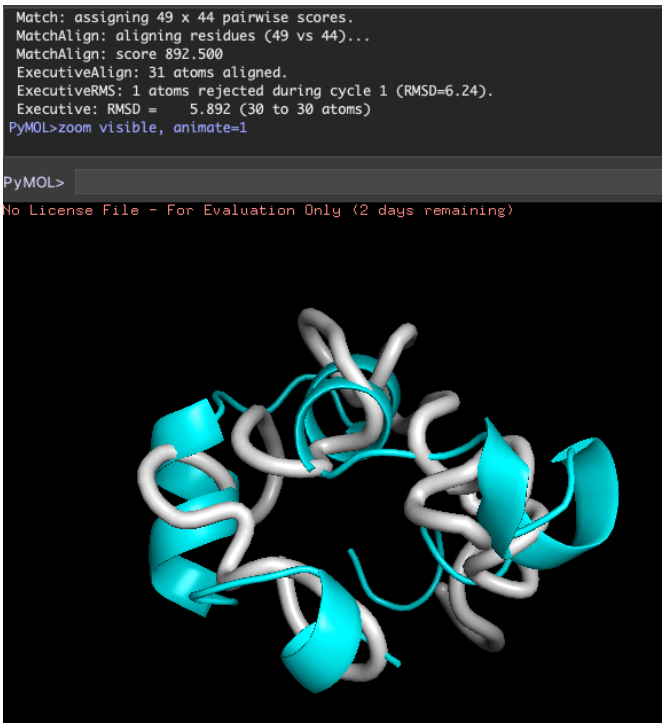
We finally selected two models that are more similar. One is when the MSA threshold is 0.55 and the FT-COMAR threshold is 8, there are 30 atoms aligned and the RSMD value is 5.892. The other is that when the MSA threshold is 0.5 and the FT-COMAR threshold is 14, there are 29 atoms aligned and the RMSD value is 5.637.

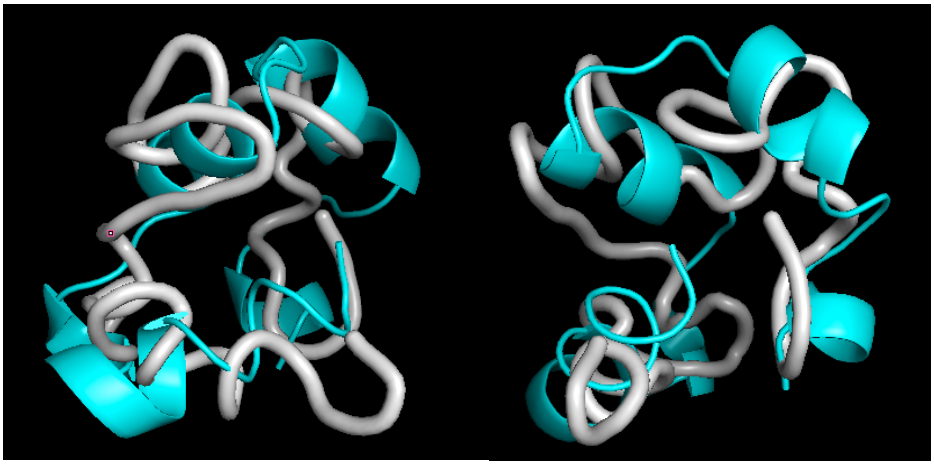
Comparison of several different angles for the predicted protein structure with threshold 0.55 and 8(Green) and Original protein structure(blue)





Comparison of several different angles for the predicted protein structure with threshold 0.5 and 14(White) and Original protein structure(blue)





We think that the first model might be a little better because the average RMSD value between each atom is 0.1943, which is slightly smaller than the value of 0.1964 of the second model. Visually, maybe the second one is a little more similar, and our predicted protein has a similar structure to the original protein structure. Although the error is still relatively large, it is the more similar result we could get.