Protein Structure Prediction

Aditi Goyal, Alan Zhang, Nivedita Attada

GitHub: https://github.com/aditi-goyal-523/ecs129

Introduction:

Protein structure determines protein function and is made up of smaller building blocks: amino acids. A protein's three dimensional structure is directly determined by its primary structure, or its amino acids sequence. The interactions between a protein's amino acids fold the amino acids into the complex 3D structure, which is critical in understanding a protein and therefore, gene's function. Given the complexity of the chemical bonds occurring between different amino acids, predicting protein structure is a difficult task. Over the course of nearly sixty five years, researchers have sought out various methods of predicting protein structures (Yang 2018) . There have been three major generations of predicting protein structure starting from using amino acid residues to statistical information and neural networks to support vector machines (Yang 2018). Today, protein structure predictions can be calculated on the average household laptop in a matter of minutes.

AlphaFold, a newly developed artificial intelligence system by Google, has been lauded as one of the most accurate tools for protein structure prediction. Given an amino acid sequence, AlphaFold predicts the 3D structure that the amino acid sequence produces with complementary files and figures. AlphaFold uses neural networks and other machine learning techniques to produce models of protein structures as pdb files which can then be used for deeper analysis (Jumper 2021). This can then be used in the comparison of protein structures and in the case of our project, to help calculate root-mean-square deviation (RMSD). According to an article on the accuracy of AlphaFold, it was found that AlphaFold structures show the most accuracy through

the comparison of central atoms provided by AlphaFold and other methods of protein structure prediction (Jumper 2021). To this end, this report also utilizes the central atoms in its calculations.

We are also curious to see how different amino acid substitutions can affect the predicted protein structure. As different amino acids share more or less characteristics from each other, the level of disruption of a protein's structure due to a single amino acid substitution should depend on how similar the original amino acid and the substitute amino acid are to each other. Therefore, we hypothesized that the RMSD calculated between the predicted structure from the original sequence and the predicted structure from the substituted sequence is inversely related to the BLOSUM80 score between the two amino acids.

Methods:

This project first takes two amino acid sequences, fimbrial adhesin of *Proteus mirabilis* (Fig. 1) and CST complex subunit CTC1 of *Homo sapiens* (Fig. 2), and uses AlphaFold to predict their three dimensional structures. AlphaFold predicted the protein structures as the following:

Figure 1:

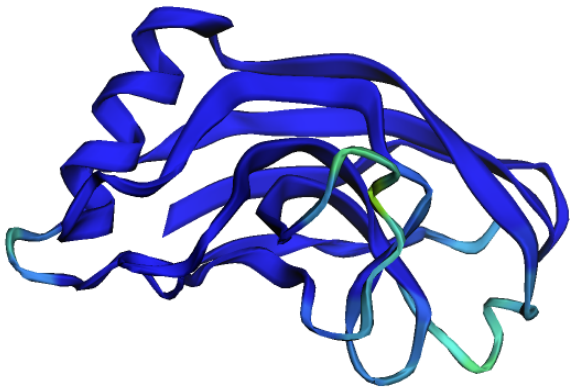


Figure 2:

The predicted structures are then compared to the gold standard structures determined using x-ray crystallography, which can be found on the Protein Data Bank. The root-mean-square deviation (RMSD) is calculated using Quaternions between the central atoms of the predicted structure and the central atoms of the gold standard structure.

In the context of this report, the RMSD is a measurement of how close the corresponding central atoms of the predicted structure and gold standard structure are to each other. The formula for RMSD:

$$RMSD = \sqrt{\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{n}} \quad (1)$$

where x_i and y_i are the central atoms.

The least RMSD is calculated using Quaternions and our implementation follows the paper by Coutsias, Seok, and Dill, "Using quaternions to compute RMSD." Following are the main formulas used for the least RMSD algorithm and can be found in the Coutsias, Seok, and Dill's paper:

$$e_q = \sqrt{\frac{\sum_{k=1}^{N} (|x_k|^2 + |y_k|^2 - 2\lambda_{max})}{N}} \quad (2)$$

where e_q is the least RMSD and λ_{max} is the maximum eigenvalue of $\mathcal{F}$:

$$\mathcal{F} := -\sum_{k=1}^{N} \mathcal{A}_L(y_k)\mathcal{A}_R(x_k). \quad (3)$$

where A_L and A_R are:

$$\mathcal{A}_R(p) = \begin{pmatrix} p_0 & -p_1 & -p_2 & -p_3 \\ p_1 & p_0 & p_3 & -p_2 \\ p_2 & -p_3 & p_0 & p_1 \\ p_3 & p_2 & -p_1 & p_0 \end{pmatrix}, \quad \mathcal{A}_L(p) = \begin{pmatrix} p_0 & -p_1 & -p_2 & -p_3 \\ p_1 & p_0 & -p_3 & p_2 \\ p_2 & p_3 & p_0 & -p_1 \\ p_3 & -p_2 & p_1 & p_0 \end{pmatrix}. \quad (4)$$

In our algorithm, we created two functions to represent A_L and A_R. These return arrays which represent the matrix equations outlined by the paper.

Our algorithm can be broken into three sections as summarized below:

- Initialization: pdb files are read and the central atoms are parsed out; necessary functions are defined.
- Linear translation: the barycenters of the model and target are obtained and the vectors linearly translated
- Get least RMSD: the least RMSD is calculated using equations 2, 3, and 4 with linearly translated vectors as the parameters x_k and y_k.

These sections are identified accordingly in the accompanying python script `rmsd.py`

We also wanted to test the effect of individual amino acids on the overall structure of a protein. Specifically, we wanted to know that if we mutate a single amino acid, how much of an impact would it have on the overall RMSD of the protein.

Results:

In order to test our question, we first generated pdb files for each of the mutations. We then used these as the ‘model’ file. The ‘target’ file was the alphafold pdb file generated using the original sequence. The results of the post-transformation RMSD values are shown in the table below.

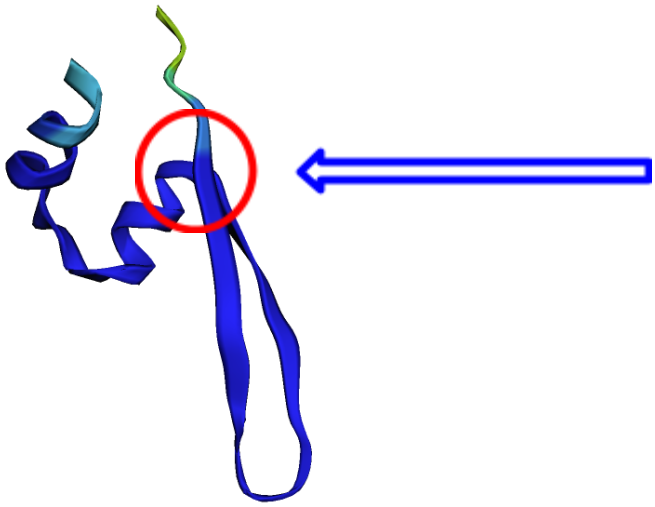
Original Protein: AISQAIIRLLVEDGTA**E**AVVTCRNHHVAAALGLCPREWASLLD

Amino Acid	Blosum80 Substitution Score	RMSD
E (original)	NA	12.173
C	-5	12.183
I	-4	12.249
G	-3	12.183
M	-2	12.200
A	-1	12.174
H	0	12.195
D	1	12.227
Q	2	12.120

Our results did not support our hypothesis that the RMSD would be inversely related to the BLOSUM 80 score. When we calculated the RMSD for the amino acid sequences after mutating one amino acid, we still got values very similar to our original RMSD.

Discussion:

We were not able to draw a conclusion from our results since it did not support our original hypothesis. Here we propose variables that can be investigated in further iterations of this experiment. One factor that could be tested in the future is how the RMSD would change if we changed more than one amino acid. Also, upon analysis of the protein shape, we believe that for future trials, we should mutate amino acids close to the ‘overlap’ portion of this protein structure.



We believe that given the long loop structure, it is possible that the mutation we were making was not in a section of the protein that was critical to its function. If we make a mutation along different areas of the protein, then we might be able to identify which amino acids contribute to this connection.

It should also be noted that while researching the topic of RMSD, we came across an article presenting various methods of protein structure comparison and listed the drawbacks of calculating RMSD. RMSD makes it difficult to find differences between protein structures that may have a similar backbone since it is a very general calculation (Kufareva 2011). Therefore, we believe that the RMSD score may not be an appropriate indicator of the similarity between each protein sequence and its corresponding structure.

Finally, we propose that the same experiment be run using a PAM scoring matrix instead of the BLOSUM scoring system. PAM, which stands for percent accepted mutations, developed by Margaret Dayhoff (Mount 2008), is designed to score alignments of highly similar protein sequences. PAM is built off of a Markov chain. By definition, it assumes that the state of each amino acid is independent, and that a mutation in one position is independent of the positions surrounding it. Unlike BLOSUM, it uses a global alignment algorithm as opposed to local

alignment. As a result, a mutation is weighed more heavily in a global alignment than compared to local alignment.

Furthermore, there exist an infinite number of PAM matrices, unlike BLOSUM. BLOSUM matrices are defined using percent thresholds that describe the percent relatedness between the two target species. BLOSUM62 is for sequences at least 62% related, etcetera. PAM matrices have a base case: PAM1. This represents the situation where there is a 1% chance that an amino acid mutates into a different amino acid. For each amino acid, there exists a list of probabilities that it will mutate into 19 other mutations. To that end, the probability that amino acid stays the same is the complement of the sum of these probabilities. This matrix can be multiplied onto itself a variable number of times to represent the increase in variability between two sequences. For example, PAM250, a commonly used matrix for analyzing distantly related proteins, is calculated by multiplying the PAM1 matrix by itself 250 times. As the n value increases, the percent similarity between the two matrices decreases. For example, the PAM250 matrix is intended to be used for sequences that are only about 20% similar.

For this experiment, we would probably have to use the PAM1 matrix, given that our amino acid sequences are highly similar (only one mutation). The PAM1 matrix is given here:

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

This matrix provides the probability of an amino acid from the top sequence mutating into a nucleic acid on the left column. For readability, and to demonstrate the rarity of a mutation occurring, these probabilities have been scaled up by a factor of 10,000. One can note that the sum of each column (or row) adds to 10,000, which is a probability of 1 overall.

To put this into the context of our experiment, Glutamine ("E") only has a .0053 percent chance of mutating into Aspartic Acid ("D"). This helps put each relative mutation into perspective. We also see that there are certain mutations that are predicted to never happen, as they have a probability of zero. Realistically, this zero implies that the probability this occurs is so rare, it is essentially negligible. Using this type of scoring matrix may provide a higher resolution as to how each amino acid change will impact the overall protein. However, in order to get visible results, we may still need to use a longer protein sequence, and mutate more than one amino acid at a time.

References:

Coutsias, E. A., Seok, C., & Dill, K. A. (2004). Using quaternions to calculate RMSD. *Journal of computational chemistry*, *25*(15), 1849–1857. https://doi.org/10.1002/jcc.20110

Yuedong Yang, Jianzhao Gao, Jihua Wang, Rhys Heffernan, Jack Hanson, Kuldip Paliwal, Yaoqi Zhou, Sixty-five years of the long march in protein secondary structure prediction: the final stretch?, *Briefings in Bioinformatics*, Volume 19, Issue 3, May 2018, Pages 482–494, https://doi.org/10.1093/bib/bbw129

Kufareva I., Abagyan R. (2011) Methods of Protein Structure Comparison. In: Orry A., Abagyan R. (eds) Homology Modeling. Methods in Molecular Biology (Methods and Protocols), vol 857. Humana Press. https://doi.org/10.1007/978-1-61779-588-6_10

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Mount D. W. (2008). Comparison of the PAM and BLOSUM Amino Acid Substitution Matrices. *CSH protocols*, *2008*, pdb.ip59. https://doi.org/10.1101/pdb.ip59