Protein Report

January 24, 2020

1 Protein Theory

DeepMind's AlphaFold and AlQuraishi's End to End prediction model were both entries in the most recent Critical Assessment of Protein Structure Prediction (CASP13). AlphaFold was able to predict highly accurate structures for 24 out of 43 free modelling domains[4] while End to End prediction fared much worse[1]. This report will describe both models.

1.1 AlphaFold

DeepMind's AlphaFold is a system of neural networks that is capable of making protein structure predictions and their CASP13 submissions utilize 3 different methods. The inputs for Alphafold are a protein sequence and a Multiple Sequence Alignment (MSA) for that protein sequence. The general idea of AlphaFold is to generate an MSA dependant knowledge based statistical potential which can be minimized to obtained parameters that produce an accurate structure of the protein. Methods 1 and 3 use a potential based on the pairwise distances of the AA residues, while Method 2 uses a potential based on the GDT_TS. Methods 1 and 2 minimize the potential using simulated annealing while Method 3 uses gradient descent on the distance potential. Methods 1 and 3 produced the best results [3]. This report focuses on Method 3.

1.1.1 Distance Based Potential

Residues in the MSA that co-evolve (i.e. if one mutates, the other has a complementary mutation), do so because they interact with each other and therefore are in close physical proximity. This information is generally used to produce contact pairs but AlphaFold uses this information to predict probability distributions of residue pair distances. However, this distribution is biased to predict distances that are too large and requires a reference state to correct this bias [4]. The reference distribution is produced by predicting the distances of the same set of training proteins but without using the sequence or MSA data and is then subtracted from the distance prediction. The equation for this calculation is given below:

$$V_{distance}(x) = \sum_{i,j,i\neq j} -\log P(d_{ij}|S, MSA(S) - \log P(d_{ij}|length)$$

This distance based potential is used in Method 1 and 3 and is substituted with a GDT potential in Method 2.

1.1.2 Gradient Descent

Next in order to predict protein structure, the distance based potential must be minimized. However, the problem with this is the distance based potential provides information about non local backbone structures, so along with the distanced based potential, torsion angles are also predicted and a torsion angle based potential is created. This provides information regarding local backbone structures. Finally, Rosetta, a physics based energy function, is used to incorporate a van der Waals term to account for the positioning of side chains. This combined function is then minimized using gradient descent to find the parameters for predicting the optimal structure.

1.2 End to End Prediction

AlQuraishi's End to End differentiable protein prediction model is a single neural network that takes a protein sequence and Position Specific Scoring Matrix (PSSM) and outputs a 3D structure [2].It consists of recurrent geometric networks (RGNs) which comprises three stages known as computation, geometry and assessment.

The computational units each represent residues in the sequence and outputs the three torsional angles for the residue. This layer of units is constructed in a recurrent and bidirectional manner so that this layer carries all the information of the protein from the N terminus to the C terminus. Each geometric unit takes in the bond angles of its corresponding computational unit and the partially completed protein backbone from the previous unit and predicts the Cartesian coordinate of the amino acid structure, thus adding one new amino acid to the backbone. The final geometric unit completes the protein 3D structure. The assessment layers takes the predicted structure and compares it to the training structure by calculating the distance-based root mean squared deviation (dRMSD) between the two. RGN parameters are then optimized by trying to reduce the dRMSD, which ensures the global structure is accurate while each individual RGN ensures the local structure is accurate.

Unlike AlphaFold's technique, no distance sampling is being done in RGNs, therefore no reference state is needed to make corrections. Additionally, the end to end model does not rely on any physics based energy function and is a purely knowledge based machine learning model.

1.3 Conclusion

While AlphaFold performed significantly better than End to End prediction at CASP13 [1], it can be argued that AlQuraishi's method is significantly better. While both are template free methods, End to end does not require any coevolutionary data which means it can make predictions for sequences that do not have any homologs unlike AlphaFold. End to end can also make predictions within seconds, unlike all other protein folding methods which take hours or days, making it the only practical solution to real world applications. But the most compelling argument is the fact that end to end differentiable models have completely outclassed human engineered methods in speech recognition and computer vision. The same will hold true for protein structure prediction and the wisest course of action to take would be to improve AlQuraishi's methods as this is the way forward.

2 Protein Practical

2.1 Introduction

The objective of this assignment is to identify the degree of variability of the side chains of all the amino acids excluding GLY and ALA. These are excluded because they lack degrees of freedom due to their small side chains. The side chains of amino acid are capable of rotation, therefore they can assume different conformations in different protein structures. This difference in conformations can be determined by calculating the RMSD of the amino acid side chains. Here, side chains are defined as Cα, Cβ and all atoms beyond this point excluding hydrogen atoms.

For each of these 18 amino acids, 1000 pairs are randomly sampled from the protein data set with replacement. The amino acid in each pair comes from different proteins. The RMSD is then calculated for each pair, and the variability of RMSD scores is observed. All proteins were obtained from the top500 collection of high quality protein structures from the PDB database to ensure accurate results.

2.2 Materials and Methods

The Root Mean Squared Deviation (RMSD) of two sets of vectors gives the similarity between the two sets. The smaller the RMSD value, the more similar the sets. The optimal RMSD algorithm works by finding the rotation matrix that minimizes the RMSD between two sets, i.e. one set of vectors is rotated until it is superimposed on the other. This method is only valid for sets that have a one to one correspondence. By using Singular Value Decomposition (SVD) the following formula can be derived which can be used to calculate the optimal RMSD without rotations:

$$RMSD(x,y) = \sqrt{\frac{1}{n}(E_0 - 2(\sigma_1 + \sigma_2 + \sigma_3))}$$

Using the optimal RMSD algorithm requires the center of mass of both objects to be at the origin. The center of mass is calculate by finding the mean of all x, y and z values in each set of vectors. The center of mass of each object is then moved by subtracting the vector of each point in the object by the vector of the center of mass. This translates the entire object towards the origin by a distance corresponding to the distance between the center of mass and the origin. This is done using the following code:

```
def get_center(coord_matrix):
    #calcualte center of mass of side chain
    center_of_mass = coord_matrix.sum(1)/coord_matrix.shape[1]
    #center the matrix
    centered_matrix = coord_matrix - center_of_mass
    return centered_matrix
```

In order to compare two side chains, the correct side chain atoms have to be selected. This is done by checking the name of each atom in the residue. If the name matches one of the backbone atoms (N, C, O), or the atom is a hydrogen atom, it is not selected. The following code performs the selection:

```
def get_sidechain(res):
    '''Get a list of side chain atoms from a residue'''
    #make sure residue is actually an amino acid
    assert PDB.is_aa(res)
    sidechain = []
    #exclude nitrogen, carbonyl carbon, oxygen and special case
    exclude = ["N", "C", "O", "OXT"]
    for atom in res.get_atoms():
        #ignore exclusion list and hydrogen atoms
        if atom.get_id() in exclude or atom.element == "H":
            continue
        else:
            sidechain.append(atom)
    #either returns list of side chain atoms or an empty list
    return sidechain
```

2.3 Results

After the optimal RMSD was calculated, a histogram was plotted for the RMSDs for every amino acid. The mean and standard devaiation for all amino acid RMSDs is given below:

```
CYS: Mean = 0.0290012400204499 SD = 0.018423218843645076
ASP: Mean = 0.28550837764823234 SD = 0.21980163139635842
GLU: Mean = 0.5489323198811482 SD = 0.2526307245847303
PHE: Mean = 0.4130261493645994 SD = 0.27942344720219364
HIS: Mean = 0.44117259744400544 SD = 0.26113310799930906
ILE: Mean = 0.2891414519800054 SD = 0.2847186980965948
LYS: Mean = 0.6450245092432924 SD = 0.2423340455860364
LEU: Mean = 0.37057393017195306 SD = 0.29498213121914063
MET: Mean = 0.6499130294174245 SD = 0.25259217627999436
ASN: Mean = 0.3709698151023477 SD = 0.23146920053414138
PRO: Mean = 0.19898407631737405 SD = 0.14643510546192648
GLN: Mean = 0.6016080780742812 SD = 0.275879633817692
ARG: Mean = 0.8183618297261899 SD = 0.2835738569212089
SER: Mean = 0.02306300424120141 SD = 0.01771666208658028
THR: Mean = 0.03335703546681525 SD = 0.022820963481960277
VAL: Mean = 0.03126374399163392 SD = 0.02892954625327702
TRP: Mean = 0.39535871992819 SD = 0.22777274152197471
TYR: Mean = 0.3856644063349895 SD = 0.28083230247811086
```

Four main groups were observed: Amino acids with low variability in RMSD, Amino acids with high variability in RMSD, amino acids which had two peaks in their RMSD distribution and PRO which also had two peaks but these were much more compressed.

The low variability group included: CYS, VAL and SER. The histogram of CYS is plotted below:

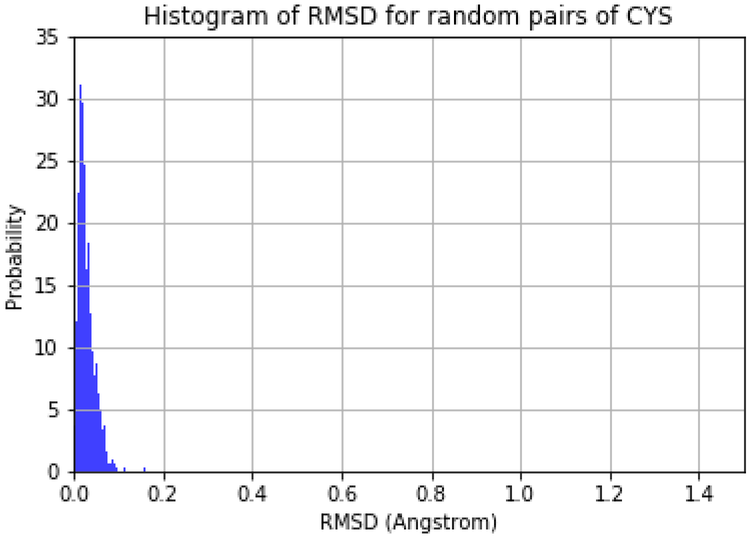


Figure 1: Histogram of RMSD for 1000 random pairs of Cystine

The high variability group incuded: ARG, ASN, ASP, GLN, GLU and MET. The histogram of ARG is plotted below:

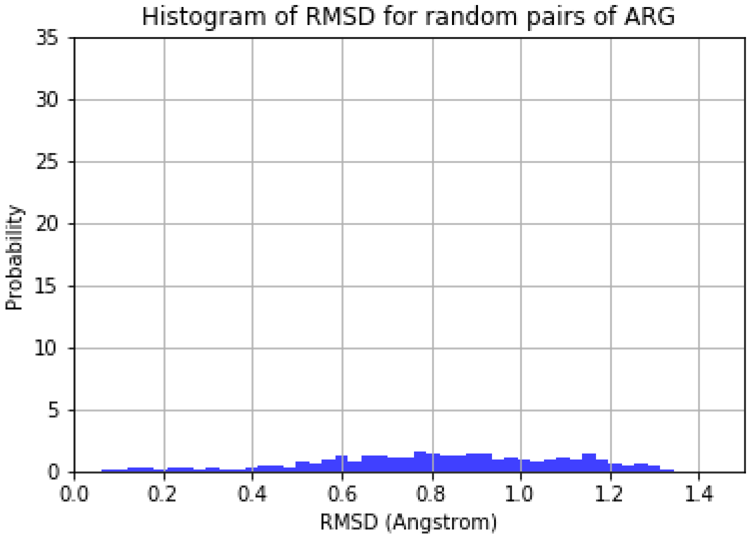


Figure 2: Histogram of RMSD for 1000 random pairs of Arginine

The twin peaks group included: HIS, ILE, LEU, PHE, TRP, TYR. The histogram of LEU is plotted below:

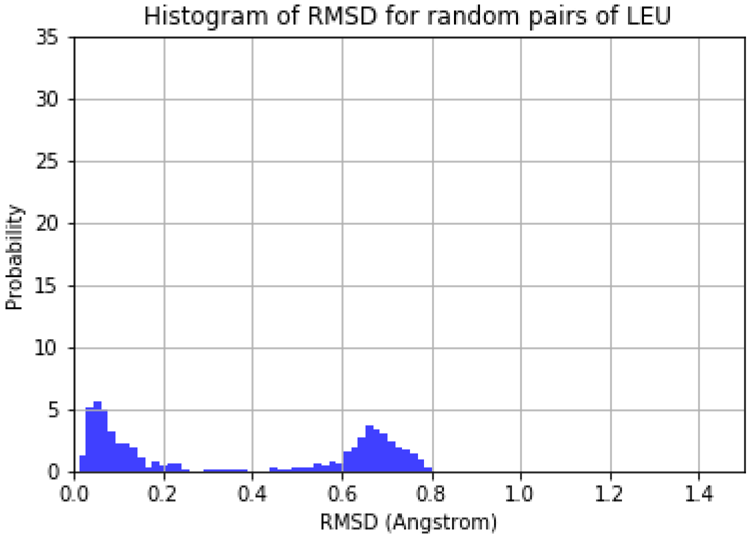


Figure 3: Histogram of RMSD for 1000 random pairs of Leucine

The histogram of PRO is plotted below:

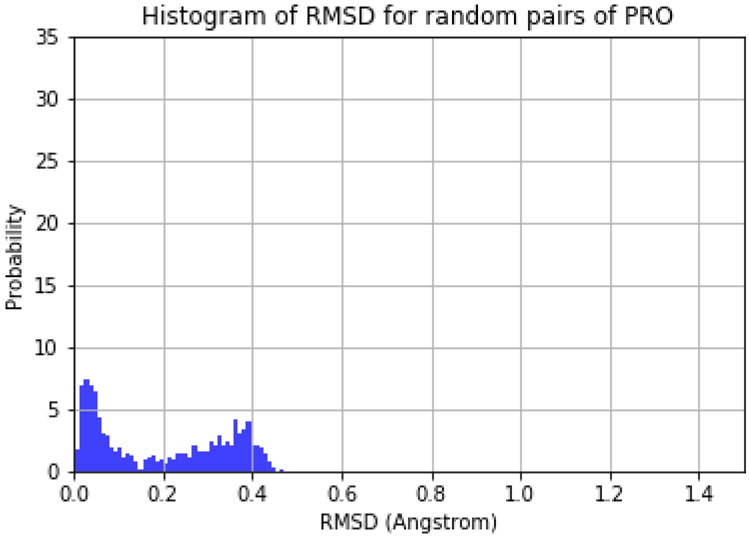


Figure 4: Histogram of RMSD for 1000 random pairs of Proline

Additional figures are plotted in the Additional figures section.

2.4 Conclusion

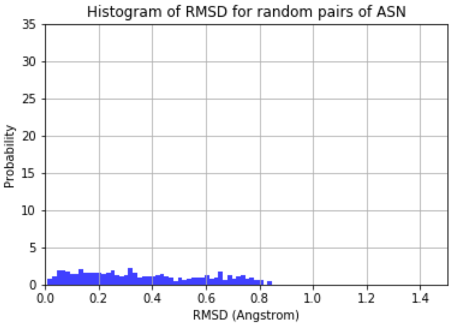
The RMSD value represents how different the various conformations are from each other, so a peak at a low RMSD value means most of the conformations vary only slightly from each other. This pattern is observed for the low variability group, which all have short side chains, so it can be reasoned that they have small side chains, with reduced degrees of freedom so they do not have much variation in their conformation.

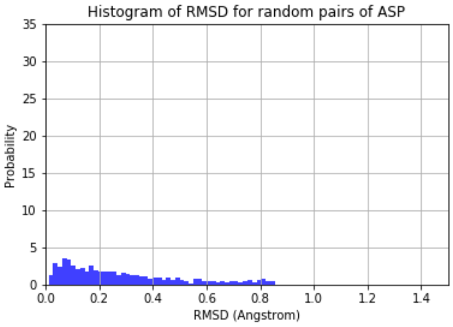
Therefore, the opposite should be true for amino acids with long side chains. Their RMSD values should be spread over a large range which would indicate they have many possible conformations since their long chain allows many degrees of freedom. Indeed, this is what is observed for the high variability group that consists of amino acids with long side chains.

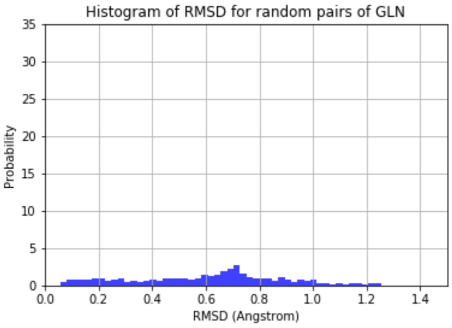
The two peaks observed in the third group mean that most of the amino acids in this group are found in one of those two conformations. This could be due to having different conformations in alpha helices and beta sheets. All the amino acids in this group are all hydrophobic in nature so another possibility is they have a different conformation deep in the hydrophobic core compared to other locations in a protein.

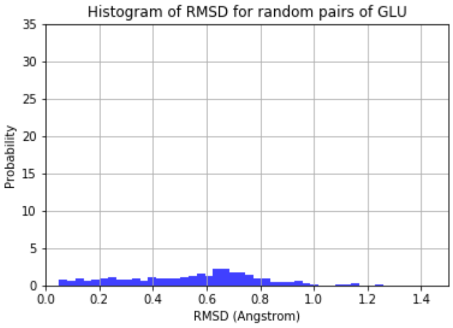
Proline has a unique structure hence it having a unique RMSD distribution seems logical. The smaller distance between peaks means the two conformations Proline can take are quite similar to each other, which makes sense given its constrained structure.

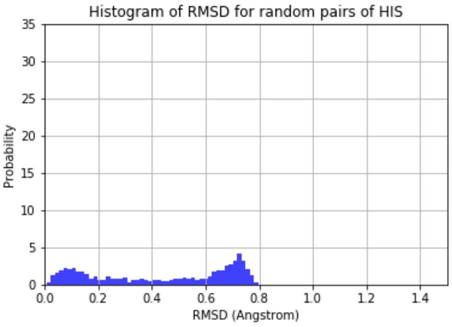
2.5 Additional Figures



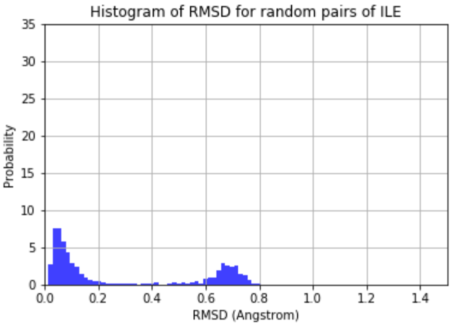




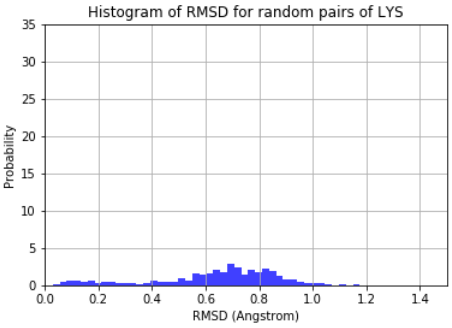




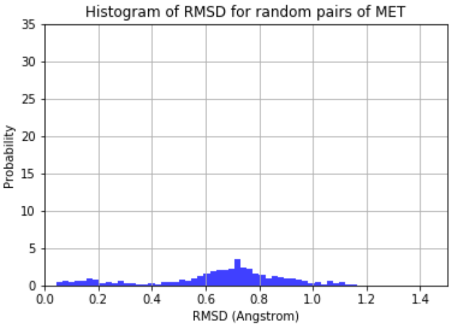
Histogram of RMSD for random pairs of HIS



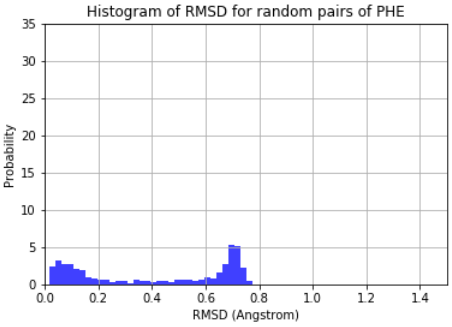
Histogram of RMSD for random pairs of ILE



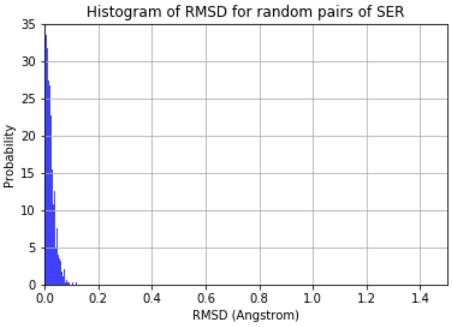
Histogram of RMSD for random pairs of LYS



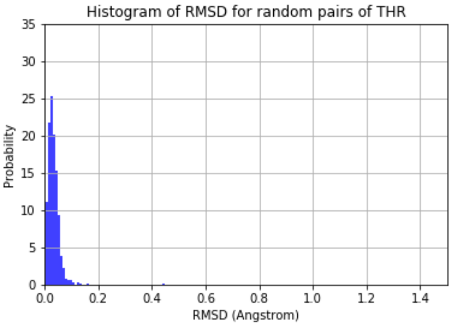
Histogram of RMSD for random pairs of MET



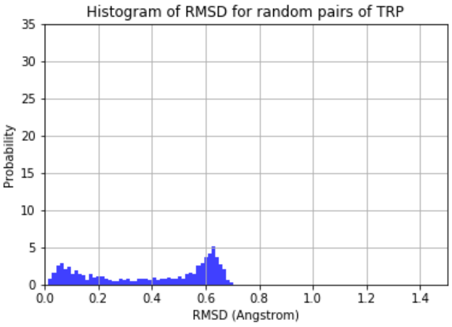
Histogram of RMSD for random pairs of PHE



Histogram of RMSD for random pairs of SER



Histogram of RMSD for random pairs of THR



Histogram of RMSD for random pairs of TRP

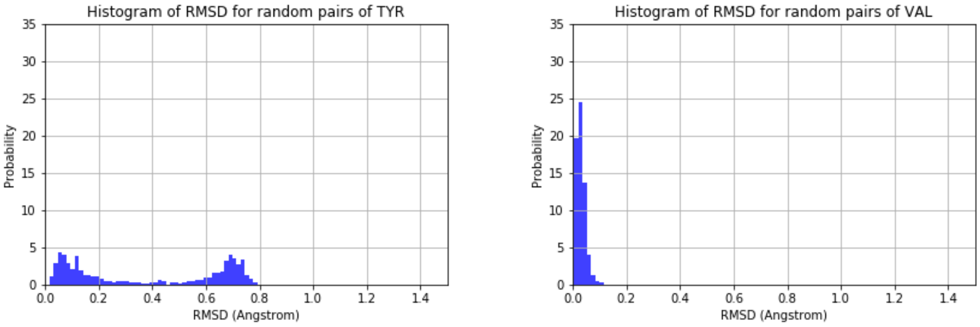


Figure 5: Histogram of RMSD for the 14 remaining amino acids

References

[1] Mohammed AlQuraishi. Alphafold @ casp13: "what just happened?". *Some Thoughts on a Mysterious Universe*, 2018.

[2] Mohammed AlQuraishi. End-to-end differentiable learning of protein structure. *Cell Systems*, 2019.

[3] Andrew Senior. Protein structure prediction using multiple deep neural networks in the 13th critical assessment of protein structure prediction (casp13). *Proteins: Structure, Function, and Bioinformatics*, 2019.

[4] Andrew Senior. Improved protein structure prediction using potentials from deep learning. *Nature*, 2020.