International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 4, ISSN No. 2455-2143, Pages 173-176
Published Online August 2022 in IJEAST (http://www.ijeast.com)

AN ANALYSIS OF PROTEIN STRUCTURE PREDICTION WITH HELP OF ARTIFICIAL INTELLIGENCE

‘

Aditi Bhadoriya
High School Science Department
The Pacific International Public School
Gwalior, Madhya Pradesh, India

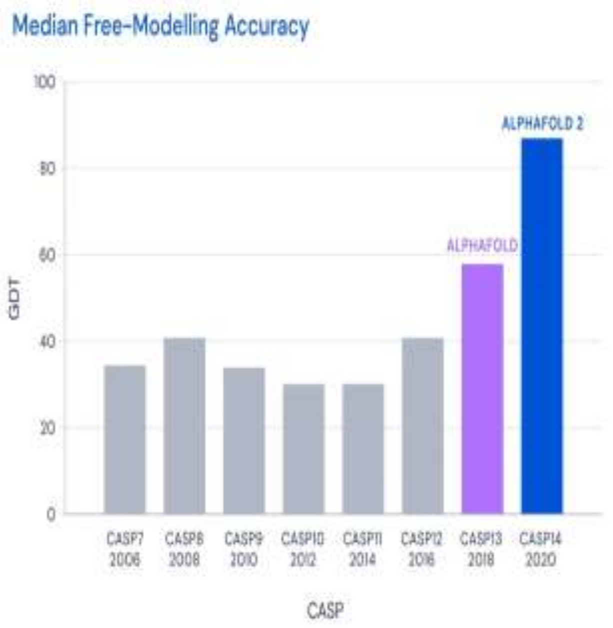
Abstract— **Proteins are essential to life, supporting practically all its functions. Figuring out what shapes proteins fold into is known as "protein folding problem" and stood as a grand challenge in biology for the past fifty years. In a major scientific advance, the latest version of the Artificial Intelligence system Alpha Fold has been recognized as a solution to this grand challenge by the organizers of the biennial Critical Assessment of Protein Structure Prediction (CASP). This breakthrough demonstrates the impact of AI on scientific discovery and its potential to dramatically accelerate progress in some of the most fundamental fields and shape our world.**

Keywords— **Alpha Fold, Proteins, Artificial Intelligence.**

I. INTRODUCTION

Proteins are large, complex molecules that play many critical roles in the body. Proteins are essential in order to give structure, function, and regulation of the body's tissues and organs. Proteins are nitrogen containing substances that are formed by long chains of amino acid. These sequences of amino acid is responsible for its unique three dimensional structure and its particular functions. The simple sequences of amino acid in a polypeptide chain is the primary structure. The next level of protein structure is the secondary structure refers to folded structures that form within a polypeptide due to interaction between atoms of the backbone. The tertiary structure is mainly due to interactions between the R groups of the amino acids that make up the protein.

The research on predicting the protein structures has been going on from more than the last 50 years. Finally, an evolutionary algorithm has been showed by Google Artificial Intelligence team named DeepMind Technologies (a British artificial intelligence subsidiary of Alphabet Inc. and research laboratory) in presenting its project called AlphaFold1 at CASP 13 competition in 2018 and a more improved AlphaFold2 in the CASP 14 competition of 2020. This review paper is going to give an analysis of both the AlphaFold1 and AlphaFold2 and explain how they have created this algorithm. The specialty of AlphaFold model that makes its unique from other protein detecting technologies is its accuracy and efficiency. Interestingly, the AI system Alpha Fold proved the accuracy of their algorithm, by scoring a median of 92.5 out of 100 and earning an overwhelming first place.



In order to uncover a protein's 3D shape, prior to computational approach X-ray Crystallography was a one method to solve this problem. In comparison with Alpha Fold, it is arduous, expensive and time-consuming method. Now Alpha Fold is able to produce highly accurate prediction of proteins structure, many of which do not share any part of its primary structure with other known proteins. Thus, although Alpha Fold cannot achieve 100 percent accurate predictions, some experts consider the protein folding problem to be solved sufficiently for experimental purpose. One of the older method for protein structure prediction was XRAY crystallography which is used to obtain a macromolecular structure. The basis of XRAY Crystallography is to obtain the distribution of the electron density which is related to the atomic positions in the unit cell, starting from the diffraction data. Moreover, one other method called Molecular Replacement method consists of fitting a “probe structure” into the experimental unit cell. The probe structure is starting atomic model, from which estimates of the phases can be computed. The previously used technologies contain some disadvantages that becomes the reason for its less popularity as compared to AlphaFold. In case of XRAY crystallography, there are a fixed amount of protein samples that can be analyzed. Other disadvantage is that it requires a crystallized sample. In particular, membrane proteins and large molecules are difficult to crystallize, due to their large molecular weight and relatively poor solubility.

II. ALPHAFOLLD SYSTEM

A. Working of AlphaFold1

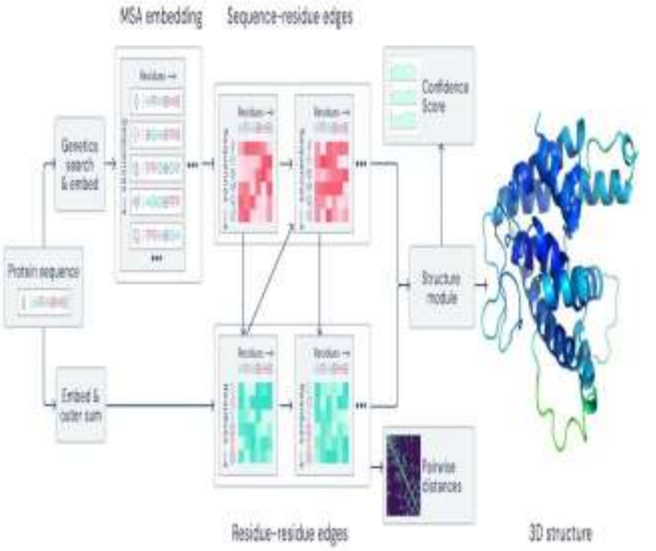
AlphaFold1 of 2018 consists of two stages. The first stage talks about CNN (Convolutional Neural Network). In the first step, it doesn’t simply track contact points moreover it also predicts probability distribution of distances between amino acid. For instance, there are amino acids A,B,C and D. From this, we can create a table of distances. It’s obvious that distance between an amino acid and itself is zero. Further, distance between A and B will be same as that of between B and A. Similarly, by calibrating the distance between each amino acid pair like this, the entire table can be filled in. AlphaFold1 even took a step further, and predicted probability distribution of these distances. By looking at the distance map of AlphaFold1 paper, each pixel of its represents a probability distribution. DeepMind named it as ‘distogram’. To determine the final folding structure of protein mathematically, the easiest way will be through 3D coordinates between amino acids central molecules. AlphaFold1 called this as torsion angles. Torsion angles are basically angles between amino acids. It is worth noting that, when proteins are in a folded state, the amino acid structure remains unchanged. Only the torsion angle between one amino acid and other changes. In relation to this, alphafold1 even determines the probability distribution of these torsion angles. In the first step, CNN receives MSA (Multiple Sequence Alignment) data and amino acid sequence.

In second stage, using distogram and probability distribution of torsion angles obtained from step 1, it determines optimal solution of the protein folding structure. For this, they set up a potential function that evaluates protein structures stability. It is the protein folding structure with the minimal function value, that becomes the final protein folding structure that alphafold1 estimates. The final step is filling in the estimation structure, a tentative answer will be saved temporarily. From iteration, an optimal solution is stored as one of the expected answers. Lastly, among the expected answers, the answer with the least potential structure is selected as the final answer.

B. Working of AlphaFold2 –

AlphaFold2 relies mainly on pattern recognition and it’s an attention based neural network architecture combined with a deep learning framework. It’s similar to jigsaw puzzle, by connecting smaller sections of puzzle. In this case, amino acids sections are joined together to form the picture of three dimensional protein structure. AlphaFold2 can be divided into three stages for the evaluation of protein structure. The stages are embedding, trunk and heads. For the first stage, it uses MSA data and protein template information for embedding. You might be noticed that MSA data was also one of the fundamental background information for predicting the entire structure in AlphaFold1. Coming to trunk stage, it shows two tracks. It’s about updating the sequence –residue edge in one part and residue-residue edge in the other. In the last step, probably the outputs are for finding the initial stage of the graph network. Further, sequence-residue edge updates attention between several different sequences and residue-residue edge is similar to finding the distance map in Alpha Fold1. In order to explain self-attention, just like we can find connection between words in a sentence. If we take this in context of protein structure, we link the relation between amino acids within protein. Therefore, this could be substituted by distogram.

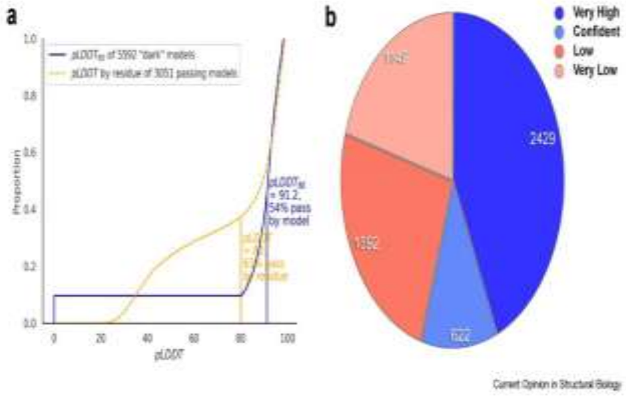
There are two avenues of communication in the Evoformer: pair representations communicated to the MSA via biases, and MSA representations communicated to the pair representation via Outer Product Mean. Each of these two forms of communication occur once in each block, implying a total of 2×48 = 96 communications per cycle. The continuous conversation between MSA and pair representation allows for the iterative refinement of an initial structural hypothesis. For example, similar genetic sequences may reveal amino acids A and B have closely grouped residues.



This information is then passed to the pair representation through the Outer Product Mean layer. Upon update, the pair representation may present a hypothesis that, since A and B are close, C and D must also be close. This is passed back to the MSA through pair bias, where similar genetic sequences are consulted as to the validity of this hypothesis.

III. RESULT OF ALPHAFOLD

One of the research paper called "AlphaFold illuminates half of the dark human proteins". According to it, known protein conformational changes can help us assess the effects of model accuracy on AlphaFold2 model readiness for target-based virtual screening (TBVS). It evaluated which AlphaFold2 models of the human understudied proteins, Tdark which currently lack an experimental PDB structure might be TBVS-ready. Of the set of 5592 "dark" proteins with AF2 models, 3051(54.6%) meet their criteria for possibly being accurate enough for TBVS studies. Talking into account the estimated false- positive rate (~6% of total), this implies that AF2 provides TBVS-ready models for about half of the understudied human proteins.



Current Opinion in Structural Biology

Fig a) Of the set of 5592 unique "dark" proteins with AF2 models, 3051 (54%) pass the proposed selection criteria of pLDDT80 greater to or equal to 91.2 while having at least 20 residues with pLDDT≥80.

Fig **b)** Pie chart illustrating AF2 model quality according to pLDDT80-derived criteria (54%) proteins associated with "very high" or "confident" AF2 models are likely to be TBVS-ready, whereas 2541 proteins are not.

In AlphaFold1, the final folding structure of a protein is expressed as the pair of torsion angles. In AlphaFold2, it seems to be expressed as position vector of the central carbon atom. For the final module, it employs a transformer in the system that updates the graph neural network with attention.

IV.CONCLUSION

AlphaFold1 proved that neural networks posses the complexity required in order to be capable of modeling the protein folding mechanism. AlphaFold2 further improves accuracy by using a more representative internal presentation and embedding equivariance knowledge in the model. This allows the network to concentrate on the underlying folding mechanism.

In the history of protein structure prediction, definitely the development of Alpha Fold is watershed moment. It achieved near angstrom accuracy for single apo domain prediction given sufficiently deep MSAs. Nonetheless, there are numerous cases exist in which latest PSP system remain uneven and untested. It include (i) MSA free prediction from individual protein sequences. (ii) ultra-high accuracy prediction important for drug discovery and enzymology, and (iii) predictions sensitive to minor sequence changes that lead to major structure changes, important for understanding the molecular bases of genetic diseases. We expect rapid developments in the areas of multidomain protein, quaternary complexes, and protein ligand complexes.

V. REFERENCES

[1]. Evans J Jumper., Pritzel A.(2021).Highly accurate protein structure prediction with AlphaFold, (pp. 583-589). https://doi.org/10.1038/s41586-021-03819-2

[2]. Hoffman, J. R., & Falvo, M. J. (2004). Protein- Which is best?,(pp. 118-130)

[3]. M Akdel., DEV Pires., Pardo Porta. (2021). A structural biology community assessment of AlphaFold 2 applications. https://doi.org/10.1101/2021.09.26.461876

[4]. QuraishiAl.,(2021). Machine learning in protein structure prediction.(pp. 1-8) https://doi.org/10.1016/j.cbpa.2021.04.005

[5]. Rao RM, Liu J, Verkuil R, Meier J, Canny J, Abbeel P, Sercu T, Rives A. Msa transformer. (2021). doi: 10.1101/2021.02.12.430858

[6]. Binder, J. L., Berendzen, J., Stevens, A. O., He, Y., Wang, J., Dokholyan, N. V., & Oprea, T. I. (2022). AlphaFold illuminates half of the dark human proteins.]. https://doi.org/10.1016/j.sbi.2022.102372

[7]. Evans J Jumper., Pritzel A, 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). Applying and improving AlphaFold at CASP14. Pg no. 89

[8]. KoJ, LeeJ. (2021). Can AlphaFold2 predict protein-peptide complex structures

[9]. accurately?.doi:10.1101/2021.07.27.453972

[10]. Stefan-Bogdan Marcu, Sabin Tabirca, and Mark Tangney. (2002). An overview of AlphaFold’s breakthrough. https://doi.org/10.3389/frai.2022.875587

[11]. Ilari, Andrea & Savino, Carmelinda. (2008). Protein Structure Determination by X-Ray Crystallography. (pp. 63-87)

[12]. Jaskolski, M., Dauter, Z., & Wlodawer, A. (2014). A brief history of macromolecular crystallography, 281(18), 3985-4009.

[13]. Dill, K. A., Ozkan, S. B., Shell, M. S., & Weikl, T. R. (2008). The protein folding problem.pg no 289.

[14]. Steinegger, M., Mirdita, M., & Söding, J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. (pp 603-606).

[15]. Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., ... & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning.(pp 706-710).

[16]. Qian, N., & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. (pp 865-884).