The mechanism and impact of AlphaFold

Junyang Deng (120090791)

December 11, 2022

In 2021, DeepMind proposed a model called *AlphaFold* to predict protein structure based on sequence [1]. This model became the winner of the 14th Critical Assessment of protein Structure Prediction (CASP14). Also, this is the first time in the history when people became able to predict protein structure at an atomic resolution. This essay introduces AlphaFold from the viewpoint of machine learning and its implication for biological research.

1 Understand AlphaFold from a machine learning prospective

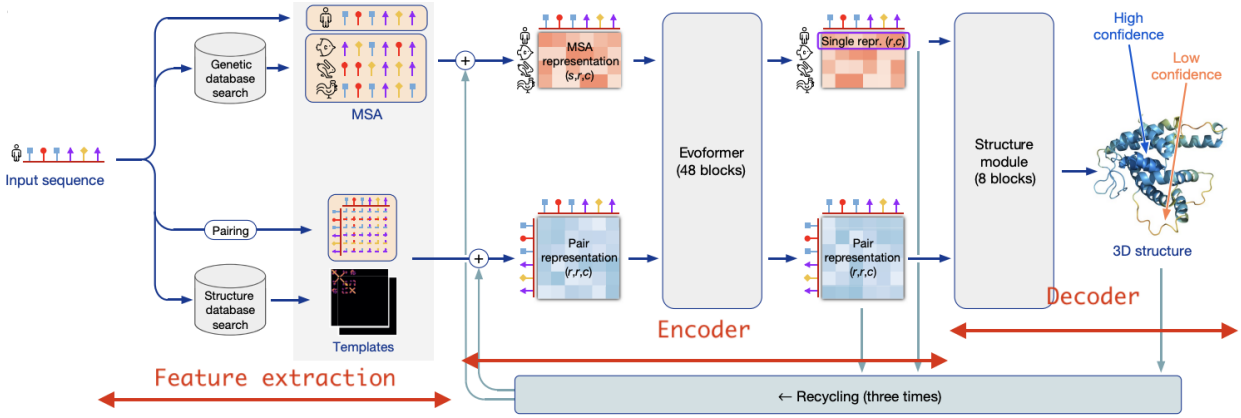

Figure 1: Model architecture of AlphaFold

The complete network can be divided into three parts (as illustrated in Figure 1), which are responsible for feature extraction, encoding and decoding, respectively. *Feature extraction* means transforming the original data into numerical features that can be processed while the information of the data is preserved. For most large input data to be processed, feature selection usually finds a new set of k dimensions that are combinations of the original d dimensions to represent data. But in sequence prediction task, feature extraction often increases dimensionality. Sequences that are originally written in letters can be converted to numerical vectors through functions. Common feature extraction methods for protein sequences include calculating amino acid compositions (AAC), using physicochemical property indexes, or using substitution matrix like BLOSUM62 to extract evolutionary information. For structure prediction, sequence and the structures of similar proteins are essential. Therefore, when a new sequence is inputted, AlphaFold will search against the database for similar proteins by multiple sequence alignment (MSA). Meanwhile, AlphaFold also tries to identify proteins that may have a similar structure as the query sequence. These potentially similar structures are called templates, which is represented through contact map – the relative distance from the i-th amino acid (AA_i) to AA_j is the (i, j) entry of the matrix. Till now, the feature extraction step is finished. The input sequence, accompanied by the MSA result and the searched template (as a contact map), was sent to the encoder block.

Before dealing with the next two blocks, let's first look at what is an encoder-decoder structure in machine learning. Encoder-decoder is a network structure commonly used in machine translation. In this structure, en-

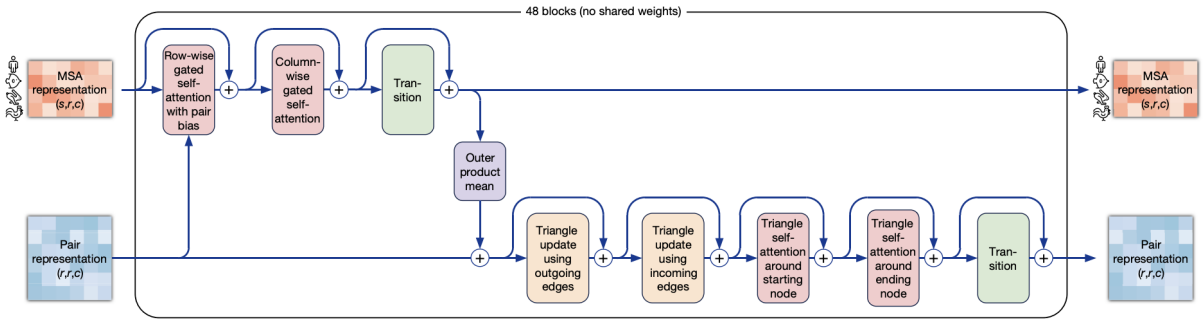

Figure 2: The architecture of the Evoformer block

coder converts the words (in language A) to some vector, and decoder converts the vector back to words (but in language B). What we need in structure prediction is similar – we want to convert sequences (in amino acid world) to structure (relative locations). To make the prediction accurate, the encoder should find the best vector the represent the information.

For this purpose, several transformer blocks were used inside the Evoformer. Transformer utilizes a self-attention mechanism. This mechanism allows people to find out what words are closely related. Figure 2 shows the architectural detail of Evoformer. Four red blocks appear in the first, second, sixth, and seventh places are multi-headed attention modules. The first block conduct gated self-attention between rows, extracting the information between similar proteins. The second block did the same thing between columns, extracting the amino acids' representation in different context. The green block "transition" is a multi-layer perceptron (MLP) module. The orange blocks (triangle multiplicative update) updates the pair representation by combining information within each triangle of graph edges. These update patterns are inspired by the necessity of consistency of pair-wise representation. Physical constraints, like the triangle inequality on distances, are used to post restriction on the predicted structure.

After the encoder module, two matrices are produced to encode a sequence. The first one is the MSA representation with dimension $s \times r \times c$ (s for sequences, r for residues and c for channels). The second one is the pair representation with dimension $r \times r \times c$. Then, these matrices will be sent to the structure module (also called decoder in this essay).

The structure module will turn the matrix representation into the form of relative coordinates.The essential part of the structure module is the invariant point attention (IPA) module. "Invariance" here means any possible rotation or translation of data will lead to the same answer, which is very important in structure prediction. The IPA operates in 3D space, calculating the affinity between residues and between atoms within residues. Finally, this module updates an N_{res} set of neural activations without changing the 3D positions.

The loss function of the whole model is called FAPE (Frame Aligned Point Error). It compares the predicted atom positions to the true positions under many different alignments. Same as other neural network models, AlphaFold uses gradient descent to update parameters. But in order to save storage space, weights are not shared between blocks, which means backpropagation will stop within blocks.

2 Implications

2.1 Revolutionize structural biology research

After introducing the architecture of AlphaFold and how it work, it's time to look at how does it transform biological research. On seeing the astonishing accuracy of AlphaFold2, I realized that the structural biology research

would be revolutionized. Before the AlphaFold2, although there existed many structure prediction softwares, those structures could only be used as reference, but not real-life applications, like the early stage of drug design.

Before AlphaFold, the most convenient and accurate method to get the structure of protein is cryo-EM. The amount of time needed to solve a protein structure through Cryo-EM can vary greatly depending on the complexity of the protein, the quality of the data, and the methods used. Generally, the process can take anywhere from several days to several months. Although other structural prediction softwares exist, they can't provide reliable structures, or the resolution is too low.

AlphaFold is the first software that can predict protein structure from sequences at an atomic resolution. Even though plenty of computational resources are needed for AlphaFold to predict a structure, it's less troublesome. After decades of effect by structural biologists, 17% of total human proteins have experimentally determined structures. Now, a new protein database that is totally calculated by AlphaFold has been proposed [2]. This database covers 98.5% of human proteins. The huge coverage of new database provide people with many new ways to formulate hypotheses. From the high-resolution protein structures, people can make reasonable inference on the biochemical functions of the protein. Sometimes, structures with slightly lower confidence scores can be used, as long as the predict of the binding pocket is accurate [3].

2.2 Current short-comings

AlphaFold is impressive, but not perfect. It also has many shortcomings. First, the accuracy of its prediction result is highly variable. The algorithm used predicted local-distance difference test (pLDDT) as the metric for confidence. For example, when predicting the structure of the 14-kDa phosphohistidine phosphatase, AlphaFold gives a highly reliable result compared to the model given by NMR. The superposition of two structures is almost perfect, despite for some disordered loop region. When it comes to the prediction of human insulin, the result bears no resemblance to real structure. Therefore, if the structure needed cannot be predicted well by AlphaFold, experiments should still be done for further research. This limits its application, since 9.8% of predicted results have low confidence ($50 < \text{pLDDT} < 70$), and 28% of results have very low confidence ($\text{pLDDT} < 50$) [4].

Second, the structure produced by AlphaFold is static, while proteins are flexible and constantly moving inside cells. Many proteins also have different conformations when they interact with different ligands or bear modifications like phosphorylation. Conformational changes define the "active" and "inactive" stages for proteins. If people want to make drugs targeting some proteins, they must know the change of protein conformation when the potential drug binds to it. Currently, AlphaFold only provides "the most probable" protein structure given its sequence, it cannot distinguish states or conformations. This makes it less useful in some applications like drug design. The inability for AlphaFold to understand conformations and the lack of training data on ligand-protein interaction also bring up another problem: AlphaFold cannot predict ligand-protein interaction [5].

Finally, AlphaFold is insensitive to mutation. This algorithm is not totally based on physical laws to make predictions, although triangular inequality is considered. While missense point mutation can sometimes cause dramatic change in protein structure, AlphaFold cannot predict that, and even gives a wrong structure labeled with high confidence.

2.3 Conclusion

Overall, AlphaFold is a powerful algorithm that provides accurate 3D structure of proteins. The new database provided by AlphaFold have profound effect on biological research. However, the application of AlphaFold is limited by its variance in accuracy, inability to predict different conformation, the lack of ligand-protein interaction and its insensitive to mutation. Further research can focus on improving these situations to make AlphaFold more meaningful to biological research.

References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[2] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, 2021.

[3] Jeffrey Skolnick, Mu Gao, Hongyi Zhou, and Suresh Singh. Alphafold 2: why it works and its implications for understanding the relationships of protein sequence, structure, and function. *Journal of chemical information and modeling*, 61(10):4827–4831, 2021.

[4] Janet M Thornton, Roman A Laskowski, and Neera Borkakoti. Alphafold heralds a data-driven revolution in biology and medicine. *Nature Medicine*, 27(10):1666–1669, 2021.

[5] Asher Mullard. What does alphafold mean for drug discovery? *Nature reviews. Drug Discovery*, 2021.