INFO8010: Reading assignment
Jumper et al, "Highly accurate protein structure prediction with AlphaFold"

Victor Mangeleer,[1] **Axelle Schyns**,[2] and **Lucie Navez**[3]

[1] *victor.mangeleer@student.uliege.be (s181670)*
[2] *axelle.schyns@student.uliege.be (s180598)*
[3] *lucie.navez@student.uliege.be (s180703)*

I. A BRIEF HISTORY OF PROTEINS

The *proteins* are fundamental to life and they can be seen as building blocks to every living things that surrounds us. Therefore, figuring out their structure is key to understand the natural world.

For the time being it is considered in biology to be one of the biggest problems of our time and huge means are deployed to solve it. However, despite all these efforts, the research for the determination of protein's structure encounters a big challenge. Indeed, determining the 3D structure of even a single protein based on its amino acid sequence with existing methods is excessively time-consuming and is thus not scalable. Moreover, these methods are inconclusive, since the accuracy of the obtained results is often too low.

The paper presents a new computational method that proposes to speed up this process while obtaining a really good accuracy and, hopefully, to allow improvements in that field by using deep learning with a new neural network architecture called *AlphaFold*. Actually, the problem tackled by the paper is not recent and has been covered many times before but the presented methodology is quite innovative since it involves attention prediction, which was not used in most previous work.

II. SIGNIFICANCE AND CHALLENGES

Taking up the challenge of solving the matter of 3D structure prediction for proteins would be a *huge breakthrough* for the scientific community since it would imply that the matter would be, almost, completely solved. Indeed, some predictions might still not be perfect. Moreover, the solution could be deployed and *made available for the scientific community*. Therefore, expensive and time consuming experiments to infer the structure of proteins will no longer be as needed.

III. TECHNICAL SUMMARY

The operating principle of *AlphaFold* is relatively simple. Indeed, it first finds similar sequences to the input and extracts the combined information of these sequences through a first block. Then, the result is passed to a second block that outputs the final predicted 3D structure of the protein. In more details, the network introduced in the paper has the following structure :

1. *Input segment of a protein* : Actually, it is combined to other structures (such as MSA) obtained from databases and paired to create a complete input that enters the network in two different representations which are:
   - Multiple Sequence Alignment (s, r, c)
   - Pair (r, r, c).

   It is important to notice that these representations are further described in the Glossary (Sec. VI).
2. *EvoFormer* : They enters the first building block of the network which is called the EvoFormer (composed of 48 blocks). Actually, it is just a fancy name for a transformer and its outputs takes the following representations:
   - Refined Multiple Sequence Alignment (r, c)
   - Pair (r, r, c)
3. *Structure model* : This corresponds to the next and final block of the architecture which is itself composed of 8 blocks.

Finally, the output of this block is the final output and gives a prediction of the 3D representation of the input protein. It is also important to notice that the single representation (r, c) and the pair representation (r, r, c) outputted by the EvoFormer block and the final 3D structure prediction are fed back into the network. Surely, this procedure is called *recycling* and is performed three times in total.

IV. MAIN CONTRIBUTIONS TO THE FIELD AND RESULTS

A. Contributions

The methodology presented in the paper adds a few innovations to the current state of the research field of 3D structure predictions using deep learning, such as:

- *New Architecture* : The new architecture introduced in the paper allows the model to deal with multiple sequences alignment.

Indeed, this occurs in the Evoformer block, whose objective is to extract the most information out of the input (combination of templates and MSA). As a matter of fact, the novelty is that the update applied to this framework are applied in each block instead of only once on the entire structure. This leads to **better communication** between the MSA representation and the pair representation which allows to update the hypothesis made on the structure multiple times. Another point of interest in this architecture is the **triangle update** (triangle composed of 3 nodes of the graph) which combines a multiplicative phase and attention in order to update the "missing" edge of the triangle based on the 2 others.

- *Iterative refinement* : It consists in applying the loss to the outputs and then feeding them back into the network, thus the entire is used once again network (**recycling mechanism**).

- *New output representation* and *associated loss* : The representation used consists in a 3D backbone structure based on the pair representations and the original sequence are represented by independent rotations and translations. This representation causes some constraint on the peptide bound geometry to be crossed frequently which actually simplify computation of the rest of the chain and allows **local refinement** of all parts of the chain. Then this constraint is satisfied by adding a **violation term** to the loss used such that the orientation of the residue is improved.

- *New attention architecture* : It is specifically made to deal with 3D structure and it is named IPA (= invariant point attention). The main property of IPA is that it benefits from invariance to translations and rotations. Therefore, it helps the network deal with unrepresented side-chain atoms by making the updates give results that are symmetrical to residue gas.

- *Unlabeled Learning* : It uses **self distillation** and **self estimates** to learn unlabeled proteins sequences. This method allows AlphaFold to reach even better accuracy result by using as data a set of highest confidence predictions from a previous training.

B. Results

This network was tested in the 14^{th} *Critical Assessment of protein Structure Prediction* challenge which is known to be a reference for protein structure prediction and it achieved the best results by far. As it was said before, one of the major need was scalability which seems to be reached by AlphaFold as it works on very long proteins while still giving really accurate results.

Furthermore, the model can estimate its results which makes it quite robust and is capable to also accurately predict new folds and chains that have a lot of hetero-contacts. However, the results are less satisfying when confronted to proteins with few intra-chain or homotypic contact.

V. LIMITATIONS AND DISCUSSION

According to what is written in the discussion section of the paper, the AlphaFold network would have several advantages over other existing methods, such as : it learns far more efficiently from the limited data in the *Protein Data Bank* (PDB) and is still able to cope with the complexity and variety of new structural data. Actually, this is a good point since the model is more efficient on the PDB data while also being able to manage new data. Another advantage of AlphaFold is that it is able to *handle the lack of physical* context and can still produce accurate models, especially in the case of intertwined homomers or proteins that only fold in the presence of an unknown haem group. In the case of the trustworthiness of the article, it was released in the wake 14th Critical Assessment of Structural Prediction competition[1], led in November 2020. The team presenting using the technology described in this paper is a team of AI scientists from Google *DeepMind*. They furnished lots of details, figures and ways of evaluating their results and their references are varied and relevant. In conclusion, it is probably fair to assume that this article can indeed be trusted.

VI. GLOSSARY

- *Multiple Sequence Alignment* (MSA) is generally the alignment of three or more biological sequences, protein or nucleic acid, of similar length. From the output, homology can be inferred and the evolutionary relationships between the sequences studied.[2]

- *Pair representation* represents a kind of initial representation of the structure. Roughly speaking, this can be viewed as a table that puts in relation each amino acid of the protein with the others and gives a likelihood estimation for these amino acids to be in contact with each others.

REFERENCES

[1] Protein Structure Prediction Center. 14th community wide experiment on the critical assessment of techniques for protein structure prediction, 2020.

[2] Wikipedia. Multiple sequence alignment — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Multiple%20sequence%20alignment&oldid=1048784914`, 2022.