TDA507 - Computational Methods in Bioinformatics
Assignment 4: Review of a topic in bioinformatics
Highly accurate protein structure prediction with AlphaFold

Stefano Ribes

December 16, 2021

1 Introduction

Given the sequence of the twenty amino acids defining a given protein, what would be the 3D structure of such protein? This apparently simple question has always been considered one of the most challenging problems in structural biology: the protein-folding problem. In fact, knowing the protein 3D structure it folds into, allows to understand the protein biological mechanisms and function at the molecular level.

However, the number of possible conformations a protein can take is extremely large and impossible to fully explore computationally. Because of that, many computational methods, algorithms, and physical and statistical models have been developed to *predict* the native structure of a protein given its amino acid sequence. Unfortunately, the accuracy of such predictions has never overcome the need of time-consuming and expensive experimental protein structure determination, *e.g.* via X-ray crystallography.

This inconvenient methodology might soon been replaced by the use of AlphaFold, a new groundbreaking method that stood out for its remarkable accuracy, advancing the protein folding problem many steps closer to a possible solution.

AlphaFold is a deep learning algorithm developed by DeepMind which achieved, among others, all-atom accuracy of 1.5Å r.m.s.d. (root mean square deviation) (95% confidence interval=1.2–1.6Å) with respect to the 3.5Å r.m.s.d. (95% confidence interval=3.1–4.2Å) of the best alternative method in CASP14 in 2020.

The rest of this report includes a brief background on previous computational methods tackling the protein folding problem. It then describes the design of AlphaFold, together with its performance and limitations. Finally, a short summary is given in the conclusions section.

2 Background

Many physical forces drive the folding of proteins: hydrogen bonds, van der Walls interactions, backbone angle preferences, hydrophobic interactions and chain entropy, to name a

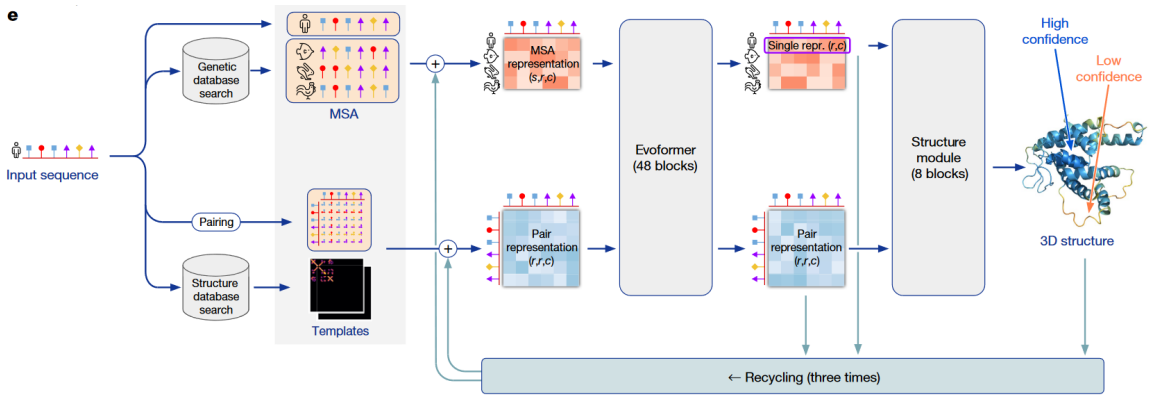

Figure 1: The AlphaFold architecture. The arrows represent information flow.

few [1]. Computational modelling of protein folding focuses on extensive molecular dynamics simulations, where the aforementioned forces are modeled as "forcefields", or models of potential energies, in computer simulations. The idea is to simulate the protein conformation changes starting from a random configuration. The approach aims at simulating the protein conformation states reaching a lower and stable energy state. This methodology has shown to be working, in practice, for short sequences of amino acids.

On the other end, modern, more successful, prediction algorithms assume that similar sequences lead to similar structures. Template-based modeling for instance relies on sequences already present in the protein's PDB. When those sequences are not present, the predictions are referred as "free modeling" and in general make the task much harder.

AlphaFold belongs to this latter category: it features a neural network pipeline which is able to predict the 3D coordinates of all heavy atoms of a protein, given its sequence of amino acids. It exploits recent advances and successes in attention-based networks for interpreting protein sequences. The idea behind AlphaFold's algorithm is to treat the protein structure prediction " as converting an 'image' of evolutionary couplings to an 'image' of the protein distance matrix and then integrating the distance predictions into a heuristic system that produces the final 3D coordinate prediction." [2]. Attention-mechanisms [3], as the name suggests, enable neural networks to focus more on certain regions of the input data compared to others. By exploiting a set of values, keys and queries weights (all trainable), a network can learn where salient features are and in turn can lead to better predictions.

3 AlphaFold Architecture

This Section gives a high level overview of the AlphaFold neural network and its components. The architecture is illustrated in Figure 1 and it's divided in two main segments: the first is a novel block defined as Evoformer, while the subsequent block is referred as Structure module.

The Evoformer is in charge of solving a graph inference problem in 3D given a Multi

Sequence Alignment (MSA) input. The MSA representation allows the network to identify and exploit relation between individual residues and the sequences in which those residues appear.

The Structure module finally takes the Evoformer output and generate the torsion angle of the peptide chains.

The network is run both in training and inference in a "recycling" technique, where its outputs are fed back into the input space. In this way, the algorithm is better performing both in terms of computation and memory utilization. On top of that, having recycling and additional training losses allow AlphaFold to be free to learn without constraints, thus generating a high quality structure.

4 Limitations

AlphaFold surely demonstrates a great potential in determining the protein structure from an amido acid sequence, but it currently requires extensive computational resources to run. Because of that, despite being open source, it's still in its early development stage to be used by the rest of the academic world and eventually other companies.

5 Conclusions

In this report I've given an overview of the research work behind the AlphaFold algorithm developed by DeepMind and presented in CASP14 in 2020. The algorithm is based on a large and complex deep neural network which heavily exploits attention mechanisms. Thanks to this architecture, AlphaFold is able to achieve impressive GDT accuracy and pLDDT scores, such that many experts consider it as a valid solution to the protein folding problem. The authors of AlphaFold believe there's still much work to do for AI to make to real-world impact,

References

[1] Ken A Dill and Justin L MacCallum. The protein-folding problem, 50 years on. *science*, 338(6110):1042–1046, 2012.

[2] 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.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.