How AlphaFold 1 and 2 outperformed the CASP competition of 2020

Marion Pavaux

February 2021

1 Introduction

Proteins are macro molecules that have a lot of functionalities in our body. For instance, they take part in the renewing of our skin, muscular tissues, they regulate our hormones, and antibodies help to eradicate virus and bacteria.

We know nowadays 200 millions of them in the living world, but all of them are made of only twenty one amino acids. An amino acid is a molecule which has an amino group as well as a carboxyl group, and finally a radical, which is different for every amino acid (Figure 1).

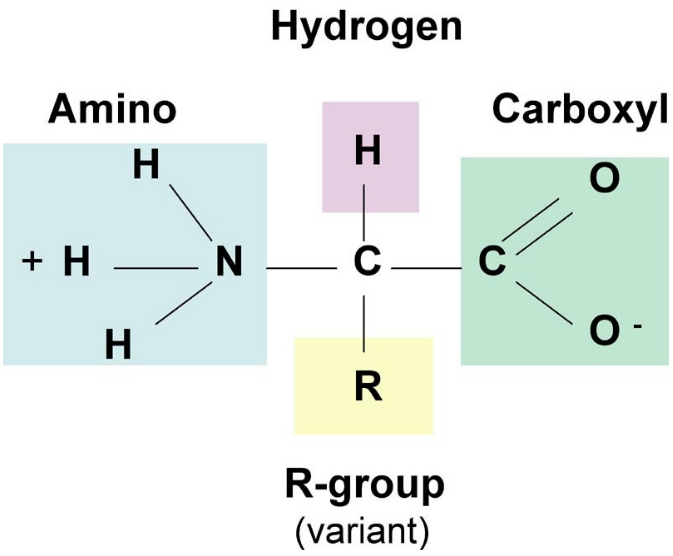

Figure 1: Structure of an Amino Acid
biochemanics.wordpress.com

What we can easily know of a protein is how many amino acids it has and from which type. However the shape is the most important property of a protein and for instance, it allows it to trap the antigens for the antibodies. It is often the case that the protein has the complementary shape of an other macro molecule, to fix to it and influence its behaviour [1]. However, this structure can only be experimentally determined, with a lot of efforts and difficulties (Figure 2). To compare with the 200 millions known proteins, we only know the shape of 70 000 of them. This problem is known as protein folding.

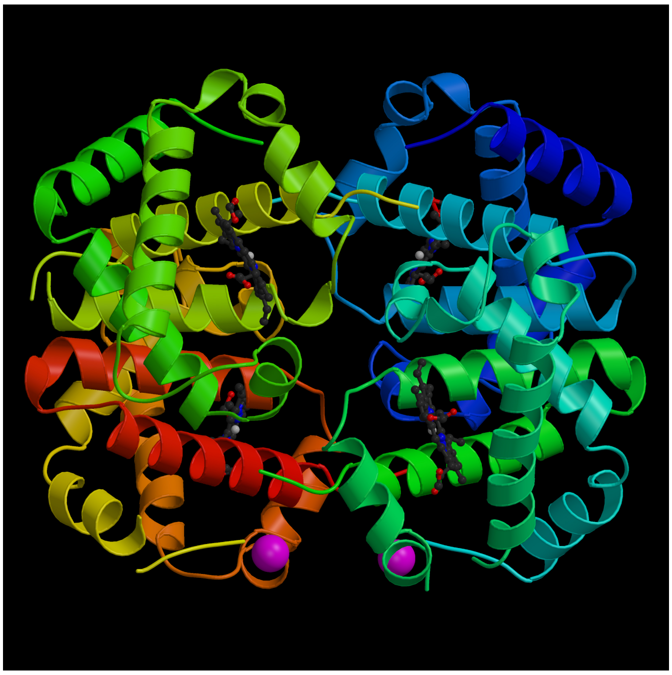

Figure 2: Shape of the Hemoglobin Protein
Protein Data Bank UCSD

To predict with accuracy the shape of proteins will be a major step in the medical field, because it will make it much more easier to find new cures to diseases. In that purpose was the Critical Assessment of Protein Structure Prediction (CASP) competition created. It takes place every two year since 1994, and gather a hundred of teams. For a hundred of protein is the shape experimentally determined, and then compared with the results of the competitors. In 2018 and 2020, the Team of AlphaFold 1 and 2 respectively from Deepmind outperformed the competition. In the second part, we try to understand how this algorithm works.

2 What is AlphaFold ?

AlphaFold [2] is a very powerful Deep Learning Algorithm that uses new tools of Neural Networks like attention networks. It was trained on over 170 000 proteins with known sequence or structure.
The two versions have two steps:

1. From the amino acids sequence determine the euclidean distance matrix of the residues
2. From the Euclidean distance matrix guess the shape of the protein

2.1 Finding the Euclidean distance matrix

In a two dimensional space with n points, the coefficient $d_{i,j}^2$ of the D distance matrix represents the square of the euclidean distance from point i to j.

$$D = \begin{pmatrix} 0 & d_{1,2}^2 & \dots & d_{1,n}^2 \\ d_{2,1}^2 & 0 & \dots & d_{2,n}^2 \\ . & & & \\ . & & & \\ . & & & \\ d_{n,1}^2 & d_{n,2}^2 & \dots & 0 \end{pmatrix}$$

To find this matrix but this time in a 3 dimensional space, AlphaFold 1 (Figure 3) used entries $L * L$ pairwise amino acids sequence with physicochemical features of the proteins for an hundred of channels. In fact, they took proteins from a data base with a sequence similar to the target protein, that is more likely to have a similar shape, and generated a multiple sequence alignment (MSA). From the MSA we can infer which residues are more likely to be in contact.
They then used 220 residual blocks of $64 * 64 * 128$ of 3 convolutional layers to take 64 amino-acids at a time. This allowed to make correlations between the amino acid chain, and the structure of the molecule. In AlphaFold the distance matrix is related to a probability density that the two residues are close to each other.

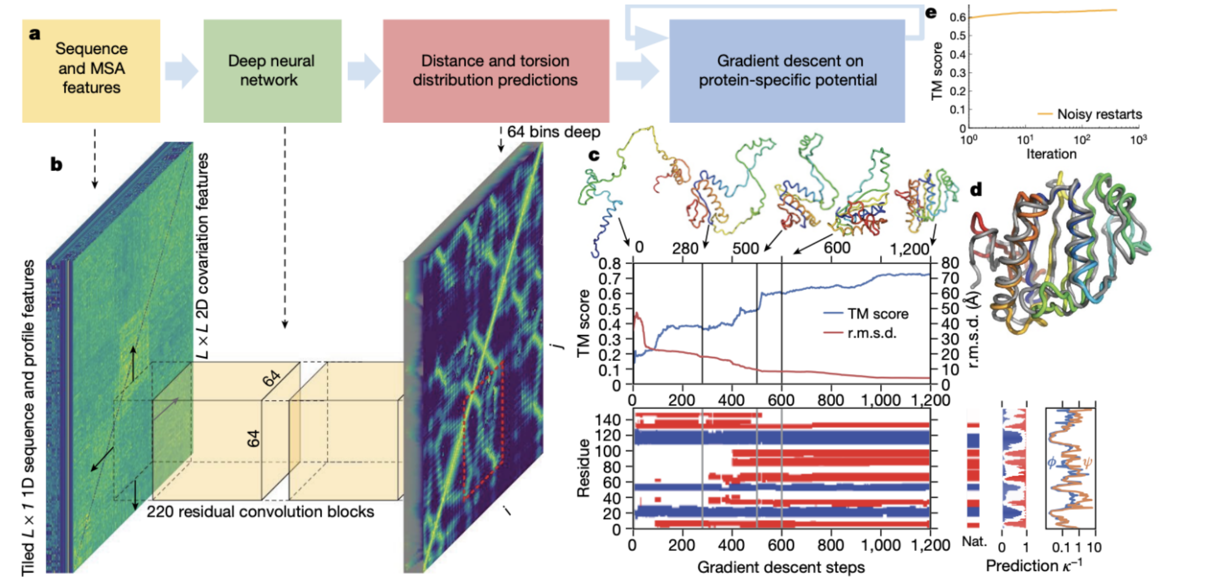
a
Sequence and MSA features
Deep neural network
Distance and torsion distribution predictions
Gradient descent on protein-specific potential
b
Tiled L × 1 1D sequence and profile features
L × L 2D covariation features
64
64
220 residual convolution blocks
64 bins deep
i
j
c
0
280
500
600
1,200
TM score
r.m.s.d. (Å)
TM score
r.m.s.d.
Residue
Gradient descent steps
Nat.
Prediction κ⁻¹
φ
ψ
d
e
TM score
Noisy restarts
Iteration

Fig. 2 | The folding process illustrated for CASP13 target T0986s2. CASP target T0986s2, $L = 155$, PDB: 6N9V. **a**, Steps of structure prediction. **b**, The neural network predicts the entire $L \times L$ distogram based on MSA features, accumulating separate predictions for 64 × 64-residue regions. **c**, One iteration of gradient descent (1,200 steps) is shown, with the TM score and root mean square deviation (r.m.s.d.) plotted against step number with five snapshots of the structure. The secondary structure (from SST[33]) is also shown (helix in blue, strand in red) along with the native secondary structure (Nat.), the secondary structure prediction probabilities of the network and the uncertainty in torsion angle predictions (as κ^{-1} of the von Mises distributions fitted to the predictions for φ and ψ). While each step of gradient descent greedily lowers the potential, large global conformation changes are effected, resulting in a well-packed chain. **d**, The final first submission overlaid on the native structure (in grey). **e**, The average (across the test set, $n = 377$) TM score of the lowest-potential structure against the number of repeats of gradient descent per target (log scale).

Figure 3: Block design of AlphaFold 1
Nature 577.7792 (2020): 706-710

We can note here that the order of the amino acids are taken into account using the convolutional layers. In fact, if there is a correlation between how close two residues are in chain, and how close they are in space. The residual blocks make also sense because the last amino acid could fold on the first one and we do not want to forget this possibility.

In AlphaFold 2 (Figure 4), MSA features were also used. Here two networks were coupled together in an attention-based neural network model: one block try to predict how residues are related to the sequence, and the other how residues are related to residues. There is an iteration with an attention block for each block, that are also coupled at each step as shown in the picture below. From the residue-residue edges we can guess the distance matrix as before with the probability distribution.

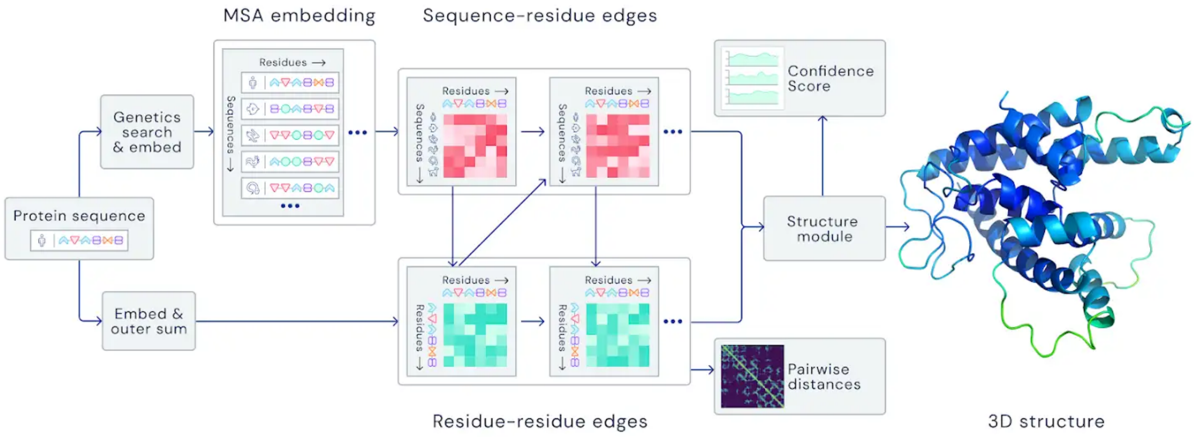

Figure 4: Block design of AlphaFold 2
Deepmind.com

2.2 Guessing the shape of the protein

With the distance matrix in a given space, you can approximately predict what is the absolute position of the points. Moreover each point has two angles Φ and Ψ defined according to chemical aspects.
One important property in physics, is that every physical system wants to minimize its energy, and that is also the case for the molecules. The components of a molecule are subjected to interactions with each other, like Coulomb forces for example. The goal is to bring the molecule in a position, so that those interactions are minimized, and so its energy.
Like we want to minimize the cost function in a neural network we want here to minimize the energy of the molecule and find the angles Φ and Ψ, so that the distance matrix is respected. In AlphaFold 1 the gradient descent were used with about 1200 steps. In AlphaFold 2, end-to-end folding was used with coordinate restrained gradient descent according to the CASP presentation of AlphaFold 2 by Deepmind. This would mean that the training will not only be on the first part of the algorithm like in Alpha Fold 1 were we then apply a minimization problem, but all the way to the obtention of the 3D structure.

3 Conclusion

According to the authors, the network took few weeks to train, to finally take a matter of days to converge for each structure. It would be equivalent to more than a hundred of GPUs.
AlphaFold 1 and 2 got respectively 60 percents and 87 percents of global distance test (GDT) at the competition. The accuracy of experimental techniques are around 90 percents. It seems that AlphaFold got the accuracy of experimental techniques. However 87 percents is a mean and there is still molecules were AlphaFold is totally wrong.
The problem of finding medicines is still not solved with this technique, the structure of molecules can dramatically change with the environment of the body and the pH. Nevertheless, it is a huge progress in the field and we can keep thinking, that the algorithm will become more accurate as it did with AlphaFold and take into account the features of its environment.

4 Sources

1. DeepMind's protein-folding AI has solved a 50-year-old grand challenge of biology. MIT technology review (2020).

2. AlphaFold: Using AI for scientific discovery.Deepmind.com

3. https://www.predictioncenter.org/casp14/doc/presentations/2020_12_01_TS_predictor_AlphaFold2.pdf