Protein theory part

January 24, 2020

1. How do these methods deal differently with local versus non-local structure?

AlphaFold. In order to model the non-local structure, AlphaFold uses a convolutional neural network that is trained on PDB structures to predict the distances between the C_β atoms of pairs, ij, of residues of a protein. Given an amino acid sequence, S, it derives features from an $\mathrm{MSA}(S)$ of that sequence and the networks predict the probability of residues being separated by different distances. The network is trained to predict distances between two 64-residue fragments of a chain, then they combine these discrete probability distributions to produce distance predictions for the entire protein (distogram) [1].

In order to model the local structure, they use a convolutional auto-regressive latent neural network based on the work of Karol Gregor on recurrent neural network for image generation [2]. Basically a separate output of the contact prediction network is trained to predict discrete probability distributions of backbone torsion angles $P\left(\varphi_i, \psi_i | S, \mathrm{MSA}(S)\right)$, and after fitting a von Mises distribution, this is used to add a smooth torsion modelling term to the potential. Finally, a Rosetta2 score is added to the potential to prevent steric clashes [1].

Then the combined potential (torsion potential, Rosetta2 score and the log probability of the distance predictions) is optimized by gradient descent to achieve accurate structure predictions [3].

RGN. Recurrent Geometric Networks (RGN) couples local and global protein structure via geometric units that optimize global geometry without violating local covalent chemistry. Protein sequences are fed one residue at a time to the computational units of an RGN, which compute an internal state that is integrated with the states of adjacent units. Based on these computations, torsional angles are predicted and fed to geometric units, which sequentially translate them into Cartesian coordinates to generate the predicted structure. Finally, RGN parameters are optimized to minimize the dRMSD between predicted and experimental structures using backpropagation [4].

2.1. How is a "reference state" used by AlphaFold and why?

AlphaFold use a reference state to obtains the correct distance-based potential. The distance potential is created from the negative log likelihood of the distances, summed over all pairs of residues i, j.

$$V_{\text{distance}}\left(\mathbf{x}\right) = -\sum_{i,j,i\neq j} \log P\left(d_{ij} | S, \mathrm{MSA}(S)\right) \tag{1}$$

Given a protein with N residues, this potential accumulates L^2 terms from marginal distribution predictions, so the reference state is used to correct the overrepresentation of the prior by subtracting the reference distribution from the distance potential in the log domain. The distance potential with a reference state become:

$$V_{\text{distance}}\left(\mathrm{x}\right) = -\sum_{i,j,i\neq j} \log P\left(d_{ij}|\mathcal{S},\mathrm{MSA}(\mathcal{S})\right) - \log P\left(d_{ij}|\,\mathrm{length},\delta_{\alpha\beta}\right) \tag{2}$$

The reference distribution used by AlphaFold models the distance distributions independent of the protein sequence and it is computed by training a small version of the distance prediction neural network on the same structures, without sequence or MSA input features [1].

2.2. What is the purpose of the reference state?

The prediction of protein structure requires conformational-energy-based score functions that can correctly pick the native conformation out of many incorrect folds. In order to properly evaluate the nativeness of the interactions in a given conformation, its conformational energy is measured relative to a so-called reference state, a hypothetical “random” state where those interactions are absent [5].

2.3. Why does the end-to-end prediction not use a reference state?

RGN doesn’t need a reference state because it doesn’t use any sampling. His neural network makes a local prediction of angles and then “remember” it to makes a global prediction so there is no need for bias correction [4].

3. To what extend are they knowledge-based and to what extend are they physics-based?

AlphaFold. AlphaFold uses a combined potential of the mean force that can accurately describe the shape of a protein, and then it optimizes it by a simple gradient descent algorithm. Its protein-specific potential combines the distance potential, the torsional potential and the Rosetta2 score:

$$V_{\text{total}}\left(\phi,\psi\right) = V_{\text{distance}}\left(G(\phi,\psi)\right) + V_{\text{torsion}}\left(\phi,\psi\right) + V_{\text{score2 smooth}}\left(G(\phi,\psi)\right) \tag{3}$$

The distance potential and the torsional potential represent the knowledge-based potentials, they are used to model the non-local and local structure respectively. Finally, Rosetta2 score, which include a van der Waals term, is the physics-based potential used to prevent steric clashes [1].

RGN. RGN energy potential is entirely knowledge based because it doesn’t use any physics-based energy function. That is the main reason why RGN-predicted structures having non-physical torsion angles can often have a poor local structure.[6]

4. What are their major similarities and major differences?

They both are ab initio modeling methods that use neural networks for protein structure prediction [3]. One of the main differences is that AlphaFold is a co-evolutionary method that convert its predictions into geometric constraints to guide a conformation sampling process [4]. It can predict distances with sufficient accuracy to outperform state-of-the-art search methods, but because of the complexity of this method the prediction can still be slow taking tens to hundreds of hours [7]. In contrast a trained RGN model, which doesn’t use evolutionary data, is a single mathematical function evaluated once per prediction, and while training RGNs can take weeks to months, once trained, they are able to make predictions in few seconds [4]. Another difference, as already mentioned, is in how they deal with local versus non-local structure and in the energy potential they use. Because of these differences, AlphaFold model local structure much better than long-range interaction [3], while RGN often predict global fold correctly but do less well with secondary structure [4].

5. In your opinion, what are their strengths and limitations compared to each other?

One of the greatest strengths of AlphaFold is probably the level of accuracy that its predictions can achieve, while one of the greatest RGN strengths is probably the speed of its predictions. One limitation of AlphaFold approach is that it is heavily dependent on coevolutionary data [3]. In fact, coevolutionary methods are not able to predict structures for which no sequence homologs exist. Also, because they use features derived from MSA, they are not able to predict effects of a small sequence changes, like mutations, on the structure of the individual sequence. One limitation of AlQuraishi end-to-end prediction is that it relies on PSSMs. This can be an advantage in term of computational complexity but PSSMs are much weaker than MSA, as they are based on single residue mutation frequencies and ignore coevolutionary interactions [4].

References

[1] Andrew W. Senior, Richard Evans et al. (2020) Improved protein structure prediction using potentials from deep learning. Nature.

[2] Karol Gregor et al. (2015) DRAW: A Recurrent Neural Network For Image Generation

[3] Andrew W. Senior, Richard Evans et al. (2019) Protein structure prediction using multiple deep neuralnetworks in the 13th Critical Assessment of ProteinStructure Prediction (CASP13). Proteins.

[4] Mohammed AlQuraishi. (2019) End-to-End differentiable learning of protein structure. Cell Systems 8, 292–301.

[5] Armando D Solis and Shalom R Rackovsky. (2011) Information-Theoretic Analysis of the Reference State in Contact Potentials used for Protein Structure Prediction. Proteins.

[6] AlQuraishi, M. (2020). AlphaFold @ CASP13: “What just happened?”. Some Thoughts on a Mysterious Universe.

[7] Mohammed AlQuraishi. (2020) Protein-structure prediction gets real. Nature.