Prediction of Protein Contact Map by Ultra-Deep Learning Model: A Review for CS273B

Amr Mohamed [1], Wisam Reid [2], and Irán Román [2]

I. Overview

In the paper, the authors developed new methods for ”de novo protein structure prediction”, which consists of predicting a protein’s 3D structure from a sequence of amino-acids. Recently, approaches to this problem have benefited from advances in protein contact prediction. However, these approaches rely on data containing proteins information with a large number of homologues sequences. Since these datasets are limited, the results of many state-of-the-art methods tend to yield results that do not generalize. Recent work in computer vision using residual neural networks has made significant progress at the level of pixel-level image labeling [1]. The authors drew inspiration from this problem in computer vision by thinking about a single amino-acid in a protein sequence as an analogy to a single pixel in a picture. Additionally, they thought of a protein’s contacts as parallel to the predicted label of an image in the computer vision problem.

Although the authors motivated their problem and approach in the bio-sciences by drawing connections to a problem in computer science, their analogy between pixels and amino-acids seems to be underdeveloped. They did not present a natural way to translate the methods from computer vision to a problem focused on sequences of amino acids. The authors do mention limitations to their analogy between fields, including the limited literature in single-pixel labeling tasks, the fact that pictures in image classification tasks are pre-processed to be of a single size while functional sequences of proteins are of different sizes, and the fact that protein contacts are characterized by more features than the ones used to train their model. Nevertheless, they proceeded to train a neural network inspired by methods in computer vision after making several assumptions about the significance of features that are relevant for protein contact prediction. In summary, they showed that the deepness of their model will help with the identification of long-range dependencies in the sequences of proteins. However, such dependencies might also be captured by other models such as recurrent neural networks or recursive neural networks. The authors did not provide sufficient justification or motivation for their particular choice of model.

II. Dataset

The dataset of choice is a subset of proteins from the widely used CASP and CAMEO databases. Their selection of proteins is somewhat conservative, rejecting proteins that do not have a large-enough number of elements in its sequence. They also exclude a number of proteins that could have contact information but the resolution of the data is limited. Thus, their data selection is inherently biased for larger proteins, and also for proteins that have been sequenced multiple times, and with state-of-the-art methods for peptide sequencing.

III. Methods

The authors utilized a cutting edge deep learning method that predicted contacts by integrating both evolutionary coupling and sequence conservation information through an ultra-deep neural network formed by two deep residual neural networks. One of the two networks took the sequence as input while the other one took the embedded pairs. Even though the authors explored other methods, including template-based models, the most important tool exploited in this paper is the residual neural network. In the following subsection, we briefly

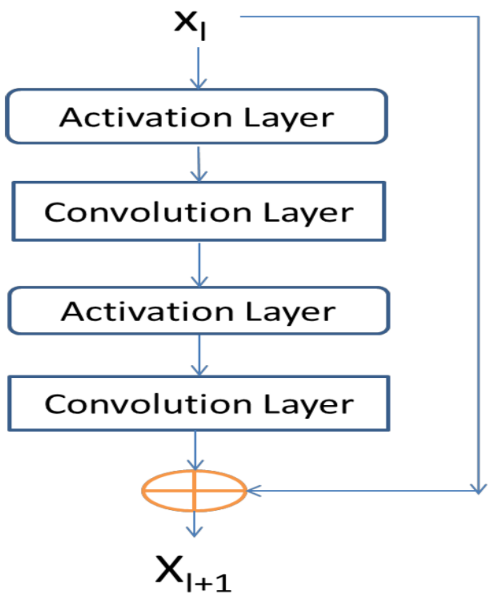
[1] Department of Computer Science

[2] Center for Computer Research in Music and Acoustics

summarize the mechanics of the basic building blocks of this architecture.

A. *Residual Network Blocks*

Each of the two residual networks that the authors used consisted of residual blocks concatenated together. Here is a figure of one such block:



Let the input to the above block be denoted by X_l. Each block applied two activation layers and two convolution layers, alternatingly. The activation layers used a rectified linear unit (ReLU) non-linearity. Let $f(X_l)$ denote the output of the last layer. Then the output of the block, X_{l+1}, is equal to $X_l + f(X_l)$. The function f is called the residual function because $f(X_l)$ is equal to the difference $X_{l+1} - X_l$. A network with such residual blocks is called a residual network.

A drawback of the method used by the authors is the small size of the filters used. Nevertheless, by stacking many residual blocks together, the authors were able to predict long-range interdependencies between input features and contacts as well as the long-range interdependency between two different residual pairs. In the results section, we describe some of the most significant results.

B. *Criticism*

The authors built a model that achieves fine accuracy when using around 60-70 convolutional layers. Similar models in computer vision had been shown to have better accuracy with a 1001-layer residual neural network than with a 100-layer network [2]. The authors mentioned that they were limited in terms of GPU memory. One way to fix this would be to use multiple GPUs in parallel.

IV. Results

Since one of the important goals of contact prediction is contact-assisted protein folding (the physical process by which protein chains acquire their native 3-dimensional structure), the authors rightly tested their methods performance on 3D structure modeling. The authors measured the quality of their 3D model against competitor (CCMpred and MetaPSICOV) built structure models using a template modeling score (TM-score), which ranges from 0 to 1, (0 indicating the worst and 1 the best respectively).

The TM-score was a good choice, it is a more accurate measure of the quality of full-length protein structures than the often used RMSD and GDT measures. The TM-score overcomes several problems afflicting these other measures by exploiting a variation of LevittGerstein (LG) weight factoring (weighting the residue pairs at smaller distances relatively stronger than those at larger distances) [3]. In other words, the TM-score is more sensitive to the global topology than to the local structural variations. Additionally, the value of the TM-score is normalized in a way that the score magnitude relative to random structures is not dependent on the protein's size. This means that TM-score can measure the similarity between two protein structures with different tertiary structures [4].

This is a good strategy since the evolutionary coupling (EC) requires multiple sequence alignments for a large group of homologs to work well and performs poorly for smaller homology groups. The author´s method here still performs better when there are more effective sequence homologs for the target protein, but its performance does not drop off as badly for proteins with less homologous information.

The performance gains seem quite impressive; some of their headlines are: They can correctly fold 224 of 579 proteins when previous methods, including the CASP11 winner, can only do 79 or

62. When evaluating accuracy of the top L long range contacts, where L is the sequence length, they are substantially better than the competitors. They also outperform a template-based method (TBM) in almost all cases, which is especially useful for proteins like membrane proteins that don't have many relevant templates available.

These results indicate that when a query protein has no close templates, their contact-assisted model may have much better quality than TBM. These results imply that their Ultra-Deep model does not predict contacts by simply copying contacts from the training proteins. It also implies that contact-assisted modeling might be very useful for membrane proteins since many of them have no close templates in the database.

A. Criticism

We agree with the decision to use TM-score. It is effective when comparing two models based on their given and known residue equivalency. However, TM-score is usually not applied to compare two proteins of different sequences.

The authors should consider using TM-align. TM-align is a structural alignment program for comparing two proteins whose sequences can be different, allowing for sequence-order independent protein structure comparisons. For two protein structures of unknown equivalence. The TM-align will first find the best equivalent residues of two proteins based on the structure similarity and then output a TM-score [5].

Note: The TM-score values in both programs have the same definition.

V. CONCLUSIONS

Overall, the authors successfully applied a bleeding edge deep learning technique to the protein contact map prediction problem. They were able to show that the new method was promising, surpassing performance of existing methods. However, the authors did not carry out an exhaustive search for the optimal hyper-parameters and did not experiment with bigger architectures that were known to perform better on image classification tasks. These drawbacks raised a natural question: could the model have been improved or could a different architecture achieve better prediction accuracy? Investigating this question would be a needed first-order extension of the paper.

REFERENCES

[1] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *arXiv preprint arXiv:1603.05027*, 2016.

[3] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

[4] Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.

[5] Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.