Review for "Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model"

Farzan Farnia, Haitong Li, Ruishan Liu, Wanyi Qian, Fei Xia

I. MOTIVATION

Protein contact and contact-assisted protein folding predictions have long been a critical challenge in computational biology. Two main approaches have been proposed in recent years: unsupervised evolutionary coupling (EC) analysis and supervised methods. The efficiency of the former methods heavily depend on the sequence homologs number, which varies a lot among protein types. In contrast, the supervised methods do not have this intrinsic limitation and are also proved to outperform unsupervised EC analysis.

This paper aims at improving the performance of supervised methods for protein contact prediction. The basic idea of achieving this goal is to increase the neural network depth. Before this paper, existing methods have shallow architectures which limit their performance to some extent. For instance, CoinDCA-NN and MetaPSICOV only have two layers, and CMAPpro is found to saturate with around ten layers. In this paper, two deep residual neural networks are utilized, taking advantage of a great many hidden layers and avoiding the vanishing gradient problem.

II. METHODS

The 6767 training datasets used in the paper was a subset of PDB25 with any two proteins sharing less than 25% sequence identity. Around 6000 proteins from the training set was randomly sampled to train 7 different models and the rest of proteins were used for validation and tuning of hyper-parameter, and the final model was the average of the seven trained models.

This paper involved two residual neural networks to solve for the linear and nonlinear characteristics of sequential features and pairwise features, and each residual block contained two activation layers (ReLU) and two convolution layer. A batch normalization layer is added before each activation layer for a faster training. The batch normalization layer will normalize the output from each convolution layer to have mean 0 and variance 1. The first residual neural network contained 1D convolution layers, and aimed to solve for the linear and nonlinear characteristics of sequential features. For a specific residual block, the output of the residual block (residual function) will be combined with a padded input to add the linearity into the output, then the output will be used as input for the next residual block. The window size used by 1D convolution layer was 17. The neural network can model the long-range relevance between sequential features and contacts, and two different residue pairs by stacking the residual blocks together. According to the paper, the first residual neural networks with depth 6 and 60 neurons for at each position would perform the best.

The second residual neural network contained 2D convolution layers, and aimed to solve for the linear and nonlinear information of pairwise features. The pairwise features was converted from the output of first residual neural network by concatenating the final output of two specific residues.

The negative log-likelihood averaged of all the residue pairs for the proteins was used as loss function. The model aimed to attain the parameters which could maximize the likelihood of occurrence of training labels (native contacts and non-contacts). And L2 regularization was utilized to regulate the parameter space and stochastic gradient descent was used for the optimization.

III. RESULTS

The deep neural network is trained by datasets of proteins with known structures, and is test on public data of CASP and CAMEO test proteins, and membrane proteins. Unsupervised EC analysis including EVfold, PSICOV, and CCMpred, and one supervised methods, MetaPSICOV, are trained and tested by the same data set for comparison.

First, the proposed method is shown to have high accuracy for different proteins including 150 Pfam, 105 CASP11, 76 CAMEO, and 398 membrane proteins. The results are also evaluated for the top L/k (L is the protein sequence lenth and $k = 10, 5, 2, 1$) contacts which are short, medium, and long-range. This method is shown to outperform the other four previous methods for all the situations. A typical result is shown in Fig. 1. Taking L/5 and medium results for instance, the supervised methods (accuracy of 0.59 and 0.77) outperform the three EC analysis (accuracy around 0.3), and the method of this paper (accuracy of 0.77) has an improvement to MetaPSICOV (accuracy of 0.59).

The paper further investigate the dependence of method performance on the amount of homologous information, as shown in Fig. 2. The result shows that EC analysis (ccm pred) is affected by the homologous number dramatically. The supervised methods (metapsicov and the proposed method) also perform better when more non-redundant sequence homologs are present, but not as sensitive as EC analysis does.

To this point, the method proves good accuracy in pro-

Method	Short				Medium				Long			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
EVfold	0.25	0.21	0.15	0.12	0.33	0.27	0.19	0.13	0.37	0.33	0.25	0.19
PSICOV	0.29	0.23	0.15	0.12	0.34	0.27	0.18	0.13	0.38	0.33	0.25	0.19
CCMpred	0.35	0.28	0.17	0.12	0.40	0.32	0.21	0.14	0.43	0.39	0.31	0.23
MetaPSICOV	0.69	0.58	0.39	0.25	0.69	0.59	0.42	0.28	0.60	0.54	0.45	0.35
Our method	0.83	0.71	0.46	0.28	0.86	0.77	0.56	0.36	0.84	0.79	0.70	0.56

FIG. 1: Table: contact prediction accuracy on 105 CASP11 test proteins.

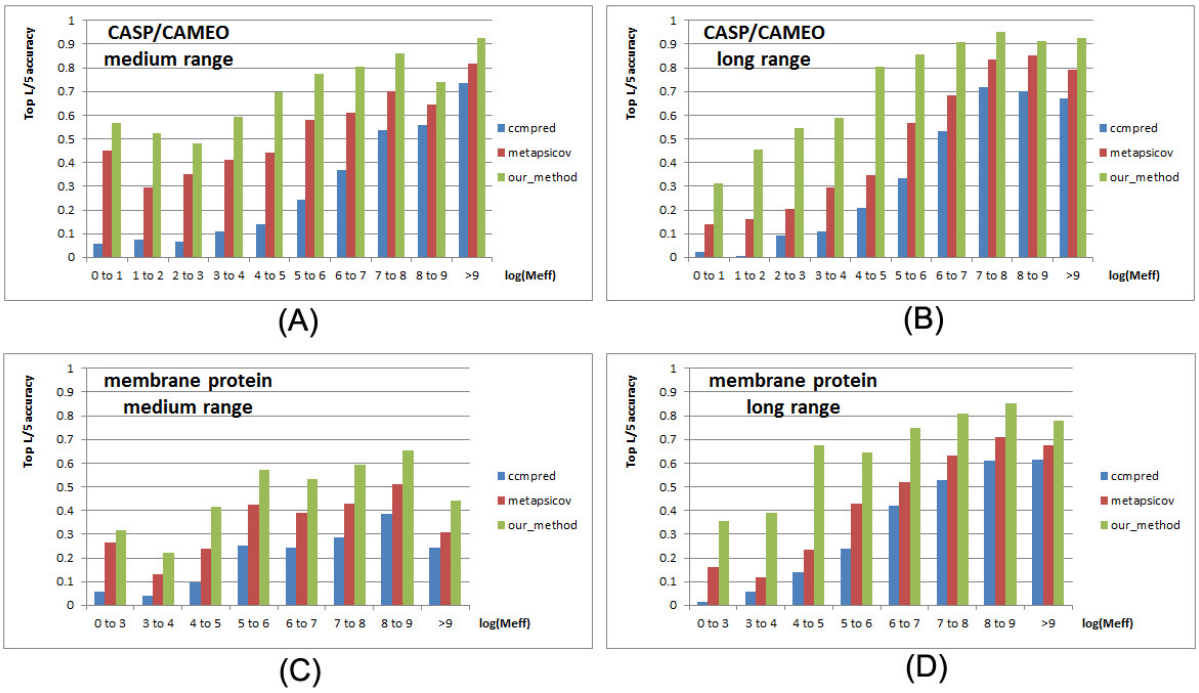

FIG. 2: Top L/5 accuracy of our method (green), CCMpred (blue) and MetaPSICOV (red) with respect to the amount of homologous information measured by $\ln(Meff)$. The accuracy on the combination of CASP and CAMEO is displayed in (A) medium-range and (B) long-range. The accuracy on the membrane protein set is displayed in (C) medium-range and (D) long-range.

tein contact prediction, and the authors further move on to show the use of contact in protein folding prediction. The contact-assisted protein folding prediction is carried out for three types of proteins and compared among CCMpred, metaPSICOV and the proposed method. The method of this paper is shown to outperform the others. When compared with template-based models (TBM), the contact-assisted model shows better performance especially under few or no close templates cases, which is quite straightforward.

IV. DISCUSSIONS

The results indicate that the proposed method performs well in predicting protein contact, and the contact-assisted model is also promising for protein folding prediction. The comparison with unsupervised EC analysis and other supervised methods

One unclear point of this paper is the influence of homologs number on the proposed method. From Fig. 2, we confirm that the CE analysis is affected by homologs number dramatically due to its intrinsic theory. However, the sequence homologs information also affects the proposed method in a large scale. For instance, the proposed method has an accuracy of about 0.3 when $\ln(Meff) < 1$ (equivalently $Meff < 10$), but the accuracy increases above 0.9 when $\ln(Meff) > 6$ (equivalently $Meff > 400$).

The authors do not address the cause of this huge improvement due to homologs number. A doubt immediately arose as to whether the relatively bad perfor-

mance of this method on some kind of protein, such as 76 CAMEO test proteins, may attribute to the less sequence homologs they have. The authors may benefit more from the discussions on why the method performance varies among different protein kinds and different homologs. Corresponding strategies could help make this method be accurate to more universal problems.

Moreover, the batch normalization layers aimed to reduce the training time by rescaling the output data before the activation layer. However, the rescaling can also lead to the saturation of output if the output contained similar values. Even though the features of input proteins sharing less than 25% sequence identity, it could have possibility that the combination of linear and nonlinear information of features will lead to increase of similarity of output. the low accuracy of low $\ln(Meff)$ could also caused by still utilizing normalization of output even though the number of output is small and appropriate for using stochastic gradient descent directly. Also, the reason for fixed depth of first residual neural network was not explained and vary the depth of first residual neural network could also influence the accuracy of prediction.

V. SUMMARY

The paper introduced a new deep learning methods which integrated both evolutionary coupling information and sequence conservation information with linear and nonlinear characteristics by utilizing two residual neural networks. The first residual network processed the 1D linear and nonlinear information of sequential features and converted the output into a 2D pairwise features for the second residual network. The new deep learning methods outperformed other existing contact prediction methods due to much higher accuracy, and performed better than other template-based models due to better quality of contact assisted models. However, the paper didn't clarify the influence of homologs number on the proposed method and the influence of batch normalization layers on small amount of input data with low $\ln(Meff)$. Moreover, the reason for fixed depth of first residual neural network could be elaborated more clearly for an optimal performance of the proposed model.