Application of DNA-Binding Protein Prediction Based on Graph Convolutional Network and Contact Map

Zhiqiang Hui, Nan Zhou

Suzhou University of Science and Technology

Abstract: DNA contains the genetic information for the synthesis of proteins and RNA, and it is an indispensable substance in living organisms. DNA-binding proteins are an enzyme, which can bind with DNA to produce complex proteins, and play an important role in the functions of a variety of biological molecules. With the continuous development of deep learning, the introduction of deep learning into DNA-binding proteins for prediction is conducive to improving the speed and accuracy of DNA-binding protein recognition. In this study, the features and structures of proteins were used to obtain their representations through graph convolutional networks. A protein prediction model based on graph convolutional network and contact map was proposed. The method had some advantages by testing various indexes of PDB14189 and PDB2272 on the benchmark dataset.

Keywords: DNA-Binding Proteins, Graph Convolutional Network, Contact Map, Protein Prediction.

1. Introduction

With the development of gene sequencing, various sequencing studies have left many DNA and proteins, including DNA-binding proteins[1]. In order to improve the accuracy of structure and prediction, combining with the current developing trend of the technology of deep learning, a DNA binding protein prediction[2] model based on GCN[3] and contact map was proposed[4].

The protein graph depends on the sequence of the results of the comparison, so first

introducing the preprocess of the dataset, including sequence comparison and filtering; the part of the output is used to calculate the features, and the other part as the input of Pconsc4 model[5], which is used to predict protein contact map, so the inputs of the model are feature matrix and adjacency matrix[6]. We use them for training and prediction. The research content of this paper is shown in Figure 1.

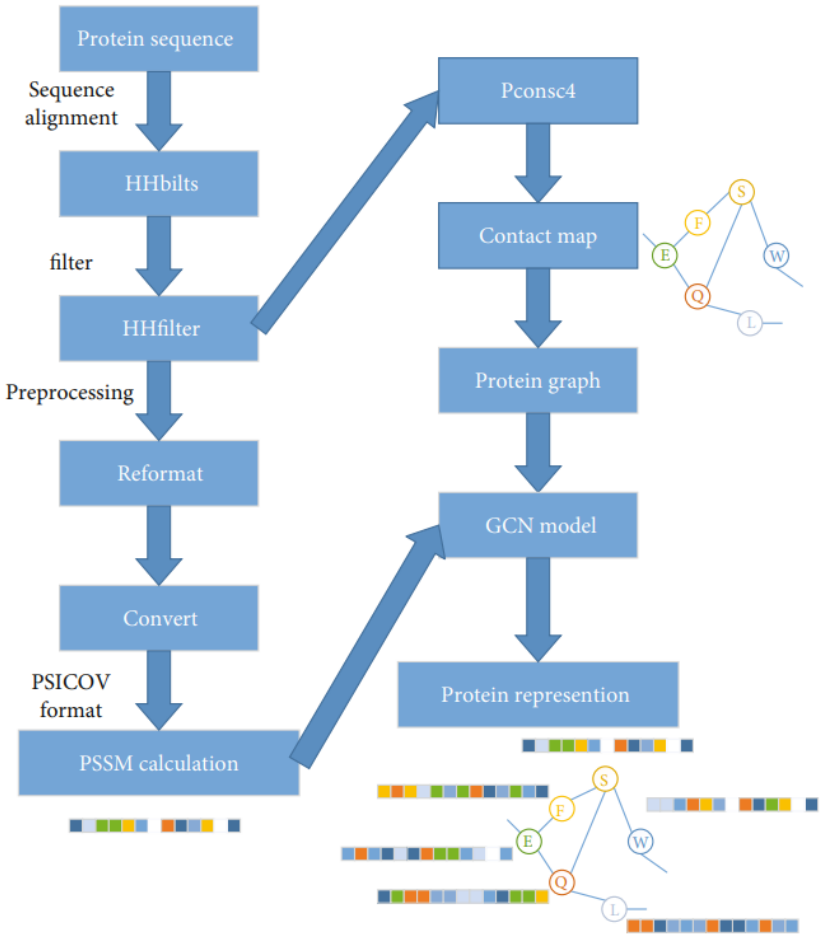

Figure 1: The processing of proteins, including the preprocessing of sequence, the generation of graph structures, and feature extraction, Pconsc4 was used to extract protein structural information. Finally, protein graph was generated higher-level feature graph through GCN.

2. Methods

We have proposed a DNA binding protein prediction model based on graph convolutional network (GCN) and contact graph. This model obtains protein features and structural representations through graph convolutional networks, and extracts protein structural information using contact maps. The specific steps include preprocessing the dataset, using the Pconsc4 model to predict protein structure information[7], extracting protein features, and training and predicting DNA binding protein data. Figure 2 shows the architecture of the model.

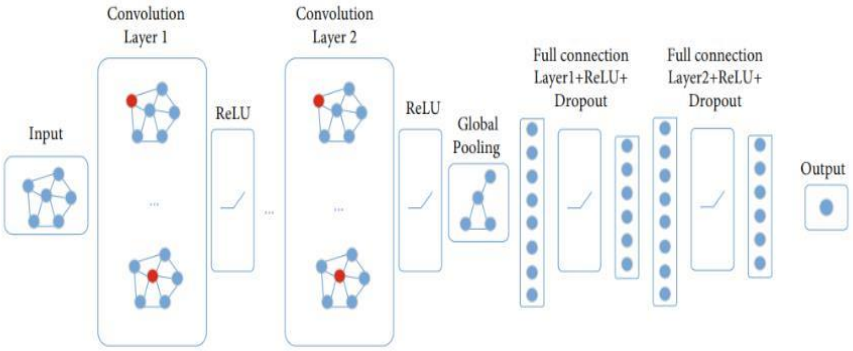

Figure 2: The structure of the GCN network, graphs of DNA-binding proteins through the GCN to get their representation.

Predicting the structure of a protein from its sequence is the purpose of introducing contact map. Specifically, assuming that the length of protein sequence is M, the size of its contact map is M∗M.M(i,j) represents the probability of contact between the ith residue and the jth residue. If the value is less than the threshold value, it can be considered that they are in contact. Pconsc4 is a fast and efficient method to predict contact map. Since its output is a probability value between 0 and 1, the threshold value of 0.5 was set for the obtained contact maps, and the probability value greater than or equal to 0.5 was set as 1.The rest were set as 0, so that the structural information of the protein could be well extracted, corresponding to the adjacency matrix as the input GCN network .[8].

Figure 3 shows a protein contact map.

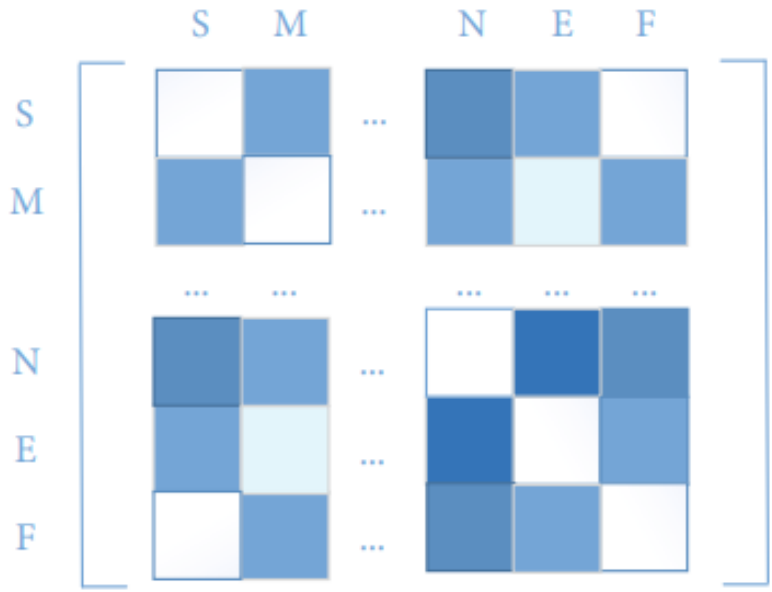

Figure 3: The contact map of protein.

The next step is the extraction of protein features. Since residues are used as nodes, the properties of residues are selected as features. Due to the differences in the R group, different features are displayed, including aromaticity, polarity, and explicit valence [9]. Position-specific scoring matrix (PSSM) is a commonly used representation of protein features, in which the results of each element depend on the results of sequence comparison, and these results represent the feature of proteins [10]. Other features were also used, such as the primary thermal coding of the remaining symbols, whether the residue was aromatic, whether the residue was acidic charged, and whether it was extremely neutral, etc. [11], as shown in Table 1. In summary, the total number of features is 54, so the protein's feature matrix dimension is (M, 54)

For PSSM, the basic position frequency matrix (PFM) [12] is calculated by the number of occurrence of residues at each position in the sequence of sequence alignment results.

Table 1: Node features.

Label	Feature	Size
1	One-hot encoding of the residue symbol	21
2	Position-specific scoring matrix (PSSM)	21
3	Whether the residue is aliphatic	1
4	Whether the residue is aromatic	1
5	Whether the residue is polar neutral	1
6	Whether the residue is acidic charged	1
7	Whether the residue is basic charged	1
8	Residue weight	1
9	The negative of the logarithm of the dissociation constant for the –COOH group	1
10	The negative of the logarithm of the dissociation constant for the –NH3 group	1
11	The negative of the logarithm of the dissociation constant for any other group in the molecule	1
12	The pH at the isoelectric point	1
13	Hydrophobicity of residue (pH = 2)	1
14	Hydrophobicity of residue (pH = 7)	1
	due (pH = 7) 1	54

3. Datasets

The DNA-binding protein dataset selected is the internationally common dataset. PDB14189 and PDB2272 were established by Gomes et al[13]. Among them, the PDB14189 dataset was divided into 7129 DNA-binding protein sequences and 7060 DNA-unbinding protein sequences, and the PDB2272 dataset was divided into 1153 DNA-binding proteins and 1119 nonbinding proteins. PDB14189 was taken as the training set and PDB2272 as the test set. The dataset is detailed in Table 2 below. Among them, positive represents DNA-binding proteins, while negative represents non-DNA-binding proteins.

Table 2: Introduction to the dataset.

Number\dataset	PDB14189	PDB2272
Positive	7129	1153
Negative	7060	1119
Total	14189	2272

4. Results

The experiment was built on PyTorch [14], an open source deep learning framework. The GCN model was based on its PyG implementation [15], PDB14189 was used for testing to find the optimal super parameters, and PDB2272 was used to test model performance.

The Evaluation Index. Accuracy (ACC), Matthews correlation coefficient (MCC), sensitivity (SN), and specificity (SP) were used as the evaluation indexes of the model [16], these indexes were widely used in the studies of biological sequences

In the independent test dataset, PDB14189 was used as the training dataset to train the model, and PDB2272 was used as the test dataset. According to the optimal experimental parameters, the final DNA-binding protein classification model was constructed: the number of GCN[17] layers were three, dropout was 0.2, PSSM was selected as the feature, the input and output dimensions of each layer were (54, 54),

(54,108), and (108,216). Other methods were compared with the method, and the method reached ACC (78.49%), SN (92.59%), SP (64.15%), and MCC (59.27%). Under certain conditions, the method has certain advantages compared with the existing methods， as shown in Table 3.

Table 3: Comparison between the proposed method and existing methods on PDB2272.

Methods	ACC (%)	MCC (%)	SN (%)	SP (%)
Qu et al.[18]	48.33	3.34	48.31	48.35
Local-DPP[19]	50.57	4.56	8.76	93.66
Pse-DNA-Pro[20]	61.88	24.30	75.28	48.08
DPP-Pse-AAC[21]	58.10	16.25	56.63	59.61
Ms-DBP[22]	66.99	33.97	70.69	63.18
GCN-method	78.49	59.27	92.59	64.15

To evaluate the impact of different dropout values, Figure 4 shows the performance of the model according to different dropout values. When the dropout is 0.2, the model has the highest performance compared to other parameters.

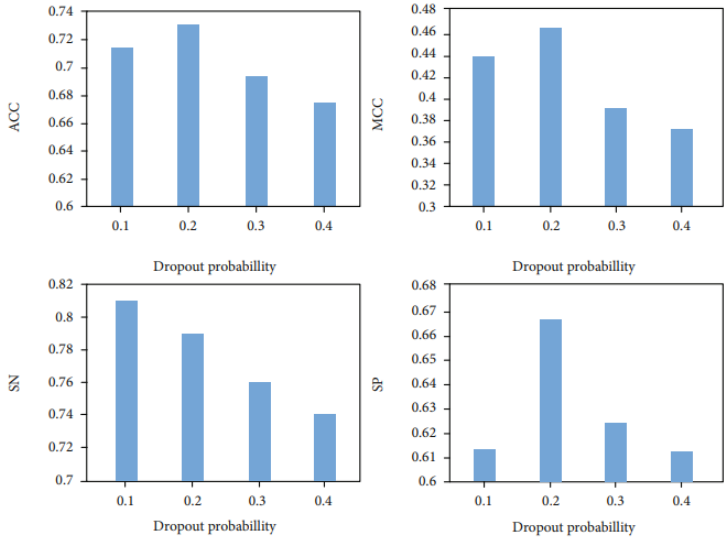

Figure 4: Comparison of prediction performance of different dropout probabilities.

5. Conclusions

DNA-binding proteins are enzymes, which can bind with DNA to produce complex proteins and play important roles in the functions of a variety of biological molecules. In order to improve the accuracy of prediction of DNA-binding protein, a DNA-binding protein prediction model based on GCN and contact map was proposed. In this model, the dataset was preprocessed by sequence alignment; then, the structural information is extracted by Pconsc4 model; PSSM and some biological characteristics are used as features. Finally, the GCN model was constructed to train and predict DNA-binding protein data. The protein graph contained information about the interactions and positions of each residue pair, which was important for feature learning and predicting binding proteins. The protein graph was input into the GCN to extract the features, and the prediction included two full connection layers. Using GCN to map proteins to the representation of rich features has also become a method of protein feature extraction. Through training and parameter tuning, the performance of GCN model was better than some existing methods. It also provides some thoughts for other fields of biological information.

In the future, we plan to carry out a research on feature extraction and network model to improve the accuracy of DNA-binding proteins and related prediction. Different biological features can be combined, and methods such as attention mechanism can be considered to improve the model, in order to achieve the goal of improving the prediction effect and other indicators.

References

1. M. S. Nogueira and O. Koch, “The development of targetspecific machine learning models as scoring functions for docking-based target prediction,” Journal of Chemical Information and Modeling, 2019.
2. Y. Wang, Y. Ding, F. Guo, L. Wei, and J. Tang, “Improved detection of DNA-binding proteins via compression technology on PSSM information,” PLoS One, vol. 12, no. 9, 2017.
3. J. Hanson, T. Litfin, K. Paliwal, and Y. Zhou, “Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning,” Bioinformatics, vol. 36, no. 4, 2019.
4. A. S. Rifaioglu, H. Atas, M. J. Martin, R. Cetin-Atalay, V. Atalay, and T. Doğan, “Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases,” Briefings in Bioinformatics, vol. 20, no. 5, pp. 1878–1912, 2019.
5. L. Jiang, S. Wang, B. Zhang et al., “"A more probable explanation" is still impossible to explain GN-z11-flash: in response to Steinhardt et al. (arXiv:2101.12738),” 2021, https://arxiv.org/ abs/2102.01239.
6. K. Liu, X. Sun, L. Jia et al., “Chemi-net: a Molecular graph convolutional network for accurate drug property prediction,” International Journal of Molecular Sciences, vol. 20, no. 14, p. 3389, 2019.
7. S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, “Accurate de novo prediction of protein contact map by ultra-deep learning model,” PLoS Computational Biology, vol. 13, no. 1, article e1005324, 2017.
8. V. Le, T. P. Quinn, T. Tran, and S. Venkatesh, “Deep in the bowel: highly interpretable neural encoder-decoder networks predict gut metabolites from gut microbiome,” BMC Genomics, vol. 21, no. S4, 2020.
9. Z. Hakime, Z. Arzucan, and O. Elif, “DeepDTA: deep drugtarget binding affinity prediction,” Bioinformatics, vol. 17, p. 17, 2018.
10. M. Sun, S. Zhao, C. Gilvary, O. Elemento, J. Zhou, and F. Wang, “Graph convolutional networks for computational drug development and discovery,” Briefings in Bioinformatics, vol. 21, no. 3, pp. 919–935, 2020.
11. T. Wen and R. B. Altman, “Graph convolutional neural networks for predicting drug-target interactions,” Journal of Chemical Information and Modeling, vol. 59, no. 10, pp. 4131–4149, 2019.

12. T. Nguyen, H. Le, and S. Venkatesh, “GraphDTA: prediction of drug-target binding affinity using graph convolutional networks,” BioRxiv, vol. 2019, p. 684662, 2019.
13. J. Gomes, B. Ramsundar, E. N. Feinberg, and V. S. Pande, “Atomic convolutional networks for predicting proteinligand binding affinity,” https://arxiv.org/abs/1703.10603, 2017.
14. A. Paszke, S. Gross, S. Chintala et al., Automatic differentiation in PyTorch, 2017.
15. S. Akbar, S. Khan, F. Ali, M. Hayat, M. Qasim, and S. Gul, “iHBP-DeepPSSM: identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach,” Chemometrics and Intelligent Laboratory Systems, vol. 204, article 104103, 2020.
16. T. Song, S. Wang, D. Liu et al., “SE-OnionNet: a convolution neural network for protein–ligand binding affinity prediction,” Frontiers in Genetics, vol. 11, article 607824, 2021.
17. K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How powerful are graph neural networks?,” 2018, https://arxiv.org/abs/1810 .00826.
18. Y. Qu, J. A. Fitzgerald, H. Rauter, and N. Farrell, “Approaches to selective DNA binding in polyfunctional dinuclear platinum chemistry. The synthesis of a trifunctional compound and its interaction with the mononucleotide 5'-guanosine monophosphate,” Inorganic Chemistry, vol. 40, no. 24, pp. 6324–6327, 2001.
19. L. Wei, J. Tang, and Q. Zou, “Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information,” Information Sciences, vol. 384, pp. 135–144, 2017.
20. B. Liu, J. Xu, S. Fan, R. Xu, J. Zhou, and X. Wang, “PseDNApro: DNA-binding protein identification by combining Chou's PseAAC and physicochemical distance transformation,” Molecular Informatics, vol. 34, no. 1, pp. 8–17, 2015.
21. Y. D. Khan, M. Jamil, W. Hussain, N. Rasool, S. A. Khan, and K. C. Chou, “pSSbond-PseAAC: prediction of disulfide bonding sites by integration of PseAAC and statistical moments,” Journal of Theoretical Biology, vol. 463, pp. 47–55, 2019.
22. X. du, Y. Diao, H. Liu, and S. Li, “MsDBP: exploring DNAbinding proteins by integrating Multiscale sequence information via Chou's Five-Step rule,” Journal of Proteome Research, vol. 18, no. 8, pp. 3119–3132, 2019.