**OX-Pro: Classification of Oxygen Binding Proteins based on Deep Learning**

**Abstract**

How to annotate the function of protein is important to understand life. The development of sequencing technologies generates lots of protein sequences. There is an urgent need for computational methods to annotate the function of protein sequences. Oxygen-binding protein is an ancient and important protein, widely distributed in various species. Classification of oxygen-binding proteins helps to understand these proteins. In this paper, we propose a deep learning based method to achieve classification of oxygen-binding proteins. We use a convolutional neural network (CNN) and bidirectional gate recurrent unit (GRU) hybrid architecture to learn oxygen-binding proteins’ feature information from their amino acid composition and PSSM feature. We evaluate our model by acc, loss, precision, recall and Fscore, and the hybrid model achieve great performance.

Key Words：oxygen binding protein classification; feature extraction; deep learning

**Introduction**

Oxygen-binding protein, as an extremely important protein, can reversely bind oxygen and store it or transport it from the lungs and support cell respiration through the body's cell deposition, providing important biological energy through photosynthesis. Now, oxygen binding proteins have been found in all creatures [1] [2]. These proteins are mainly found in vertebrate red blood cells and specific non-vertebrate tissues, in mammals, and in many prokaryotes and protozoa. The occurrence of oxygen binding proteins in all kingdoms of organisms shows the importance of these proteins. According to UniProt[3], oxygen binding proteins are divided into 9 different types called cytoglobin, erythrocruorin, flavohemoprotein, hemerythrin, hemocyanin, hemoglobin, leghemoglobin, myoglobin, neuroglobin, each has its own functional characteristics and structure with unique oxygen binding capacity.

With the development of sequencing technology, biological genomic data and transcriptomic data are increasing rapidly. However, only using genomic and transcriptomic information is not fully to understand complex organism[4]. Proteomics research is the logical next step after genomics in understanding the working of the organism and therefore is highly demanded by biomedical research and pharmaceutical applications[5]. Proteome studies protein structure, function, interaction, etc. to understand life on molecular level. It is important to discover function of proteins for proteome research. The size of protein sequence databases is growing at an exponential rata because the advancement in sequencing technology. But accurate annotation of protein function is complex and only a few protein sequences in UniProt correspond directly to experimental annotation. Thus, it is much needed to develop bioinformation methods for proteins annotation including identification of oxygen binding proteins[6].

Some bioinformation methods have been developed to identify category of oxygen binding proteins. Oxypred classes oxygen binding proteins based on support vector machine (SVM) using amino acid and dipeptide composition, achieved 85.5% and 87.8% of accuracy respectively[7]. And a Random Forest machine learning tool which was proposed by Matthew Curcio achieved 89.22% of accuracy using amino acid composition[8]. Deep learning technology is developing rapidly recently years and demonstrated high performance for a wide range of applications. A fully connected deep learning model was used to achieve classification of oxygen binding proteins using amino acid composition and obtained 95.04% accuracy[9].

In fully connected deep learning model, each neuron form connections with each neuron in lower layer. The potential problem is the expansion of the number of parameters, which is not only easy to overfit, but also easy to fall into the local optimum. In this context, we propose a deep learning-based model composed of convolutional neural network (CNN) and bidirectional gate recurrent unit (GRU). The features of the model consist of amino acid composition and position-specific scoring matrix (PSSM) [10], more information can improve classification accuracy. And we evaluated the model by accuracy rate, precision rate, recall rate, and Fscore to make the experiment more meaningful. Compared with DNN and CNN, our complex model achieved better performance.

**Methods**

***Dataset***

We obtained the oxygen binding proteins from UniProt and downloaded it in FASTA format. In order to obtain a reliable dataset, we used the keyword “oxygen binding proteins“ and removed all those proteins annotated as “potentials”, “isoform”, “probable fragment”, “subunit” and “like”. And collected the oxygen binding proteins from three credible levels: the evidence of protein existence as evidence at protein level, evidence at transcript level and inferred from homology.

We used Cd-hit [11] to create a dataset which has redundancy below 0.9, reducing the error caused by too much similarity. According to the statistics on the length of the obtained oxygen binding protein, several sequences with a length greater than 800 are deleted for the convenience and normalization of subsequent processing. The remaining sequences are padding 0 to 800 during the subsequent processing (as shown in Figure 1).

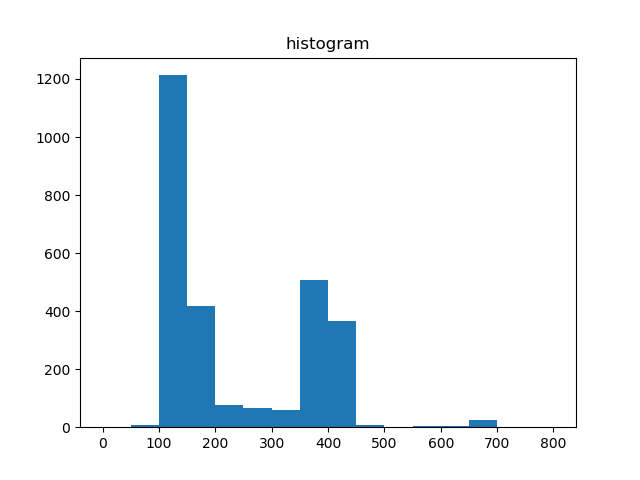


Figure 1 sequence length distribution of Oxygen binding protein

Finally, a total of 2751 common oxygen-binding proteins were obtained: 13 Erythrocruorin、120 Myoglobin、35 Hemerythrin、31 Hemocyanin、1374 Hemoglobin、28 Leghemoglobin、943 Flavohemoprotein、100 Neuroglobin and 107 Cytoglobin. Oxygen binding protein was labeled 9 types with 1-9 (Erythrocruorin、 Myoglobin、 Hemerythrin、 Hemocyanin、Hemoglobin、Leghemoglobin、 Flavohemoprotein、 Neuroglobin and Cytoglobin).

The sequence of oxygen-binding protein was encoded with one-hot, so oxygen-binding protein sequence can be represent by a 800×20 matrix. And PSI-blast was used to obtain the PSSM of each sequence. We used a 800×40 matrix which was consisted of sequence matrix and PSSM matrix.

***Model***

There are 9 types of oxygen-binding protein : Erythrocruorin、 Myoglobin、 Hemerythrin、 Hemocyanin、Hemoglobin、Leghemoglobin、 Flavohemoprotein、 Neuroglobin and Cytoglobin. Therefore, this experiment requires the use of a model to predict the type of a new given oxygen-binding protein:

M(Dnew)=L

L ∈ { Erythrocruorin、 Myoglobin、 Hemerythrin、 Hemocyanin、Hemoglobin、Leghemoglobin、 Flavohemoprotein、 Neuroglobin and Cytoglobin }.

We proposed a model composed of convolutional neural network and bidirectional GRU. CNN achieved good performance in feature extraction, so we assumed that CNN can identify the characteristics of different types of oxygen-binding proteins. Since the convolutional neural network can only learn local features, the improvement measure is to use GRU to learn the aggregate features of the entire sequence.

As shown in the architecture diagram(Figure 2): Convolution neural network with rectified linear unit (ReLU) activation functions was used to learn oxygen-binding proteins’ patterns. After the 2 layers convolution neural network, a max pooling layer is used to reduce the dimension of the data. Pooling layer can reduce information redundancy, improve model invariance and rotation invariance and prevent overfitting. Convolutional neural networks recognize local features generally. So bidirectional GRU with Tanh activation functions was used to learn long-range features. Dropout layer is used in every layers to prevent overfitting. Finally, it is still regularized and output using the softmax function.

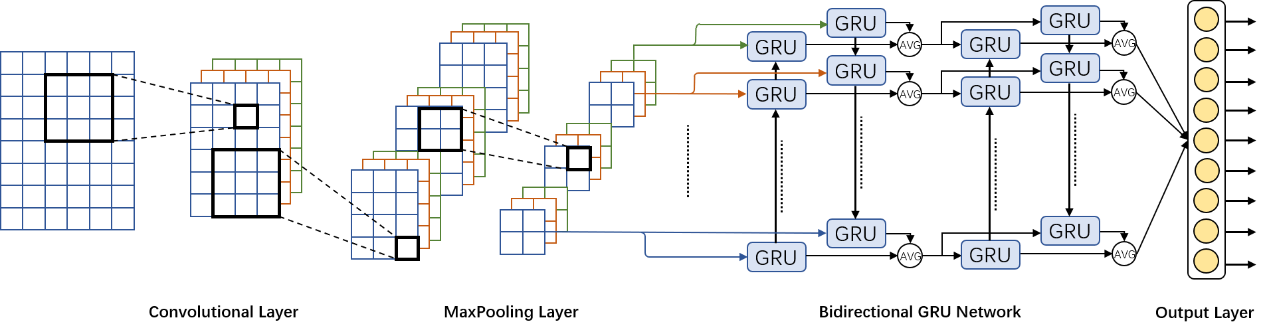


Figure 2 Architecture of the network used to identify oxygen-binding proteins. The input is one-hot encode protein sequence matrix and PSSM

***Train***

The model was constructed with Keras and Tensorflow [12], and hyperopt was employed. Here, the function categorical loss (multi-class cross entropy loss value) [13] was used as the loss function for the charge prediction task.

We split our datasets in a way to get 80% for training and 20% for testing purpose [14]. We initialized the random number a constant random seed to ensure that the results obtained are achieved again accurately. 5 folds cross-validation was used for training and verification. Hyperas which is in package hyperopt was used to select hyperparameters [15] which includes filter size of the convolutional neural network, the size of the convolution kernel, the stride, the dropout [16], the size of the pooling layer, the output dimension of the recurrent neural network, the regularizer weight, the learning rate, the epoch size, the batch size to get best performance.

**Result and Discussion**

In order to compare with the results of others, a three-layer CNN neural network model which was proposed by Soumiya Hamena et al. [9] was first established. Using the sequence + PSSM feature in this experiment, the obtained classification index was compared with the results of 2CNN + 2BIGRU.

The results obtained (shown in Table 1) show that the effect of this experiment is better than the other two methods.

Table 1 Comparison of evaluation indicators

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | loss | precision | recall | f1score | acc |
| 2CNN+2BIGRU | 0.2967 | 0.8244 | 0.7572 | 0.77 | 0.9824 |
| 3CNN | 0.509 | 0.5898 | 0.4878 | 0.5032 | 0.9662 |
| 3-layer linear model | 0.3884 | 0.5541 | 0.5488 | 0.5463 | 0.9017 |

3 models can achieve high acc, but acc is not enough to evaluate a model which is used to solve a multi-classification problem. We also calculated the precision, recall, f1socre of different model. And our model got 0.8244 precision, 0.7572 recall and 0.77 f1score, higher than other 2 models significantly. It means that our method has an excellent overall performance.

CNN generally learns the short-range information and ignore long-range and the entire information. Compared with model that only use CNN, BiGRU made our model have the ability to learn long-range information, making the results significantly better than the other two methods.

PSSM matrix and the sequence encoded by one-hot as the feature includes more biological information. More effective architectures and ricer information input feature make our model performance better than 3-layer linear model.

It can be seen from the loss\_acc diagrams (Figures 3) that the results of this experiment are not overfitting. Compared with other methods, the accuracy is indeed greatly improved.

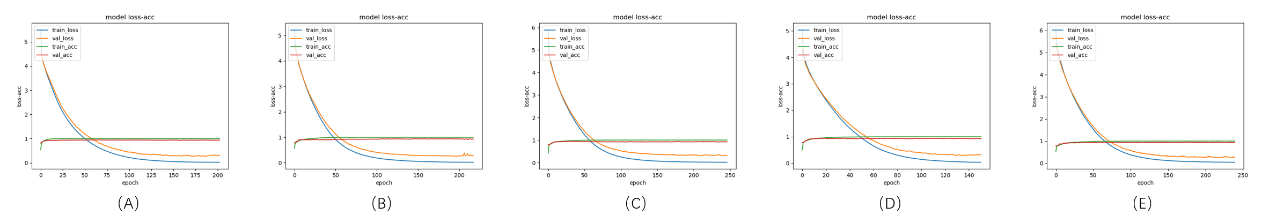


Figure 3 Loss-acc diagrams. (A) Loss-acc diagrams of first fold. (B) Loss-acc diagrams of second fold. (C) Loss-acc diagrams of third fold. (C) Loss-acc diagrams of fourth fold. (E) Loss-acc diagrams of fifth fold.

In this case, the multi-classified problem can be seen as a binary classification problem for each class of oxygen binding proteins. We evaluated the model’s classification capabilities for each classification of oxygen binding protein by precision, recall and fscore (shown in Table 2). As the table show, the amount of data for each type of oxygen binding protein would influence the classification performance. Some types of oxygen binding protein only include a small amount of sequence data, which can only provide little information to the model. On the contrary, large amount of sequence data can provide rich information to the model to achieve a great performance.

Table 2 Various types of oxygen binding protein indicators

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1score | Number of sequence |
| Cytoglobin | 0.701 | 0.5886 | 0.6238 | 107 |
| Erythrocruorin | 0.4 | 0.1666 | 0.2334 | 13 |
| Flavohemoprotein | 0.9682 | 0.9554 | 0.9616 | 943 |
| Hemerythrin | 1 | 1 | 1 | 35 |
| Hemocyanin | 0.9428 | 0.8524 | 0.874 | 31 |
| Hemoglobin | 0.9186 | 0.9568 | 0.937 | 1374 |
| Leghemoglobin | 0.8564 | 0.76 | 0.7584 | 28 |
| Myoglobin | 0.8896 | 0.8832 | 0.8786 | 120 |
| Neuroglobin | 0.742 | 0.65 | 0.6628 | 100 |

By improving the model architecture and using features which include more information, our method got better performance than other methods. CNN combine bidirectional GRU make the model can learn local information and long-range information efficiently. Using hyperas to adjust hyper-parameters can also cover a wider range and obtain better results. Although our method has achieved good performance in general, its performance is unstable when the amount of data sets is small. It is difficult to learn feature information from small train sets. With the development of sequencing technology, the number of oxygen binding proteins sequences will increase rapidly, which will provide more train data to improve model’s performance.

**Conclusion**

We propose a deep learning based method to achieve classification of oxygen-binding proteins. We use a convolutional neural network (CNN) and bidirectional gate recurrent unit (GRU) hybrid architecture to learn oxygen-binding proteins’ feature information from their amino acid composition and PSSM feature. 3 models can achieve high acc, but acc is not enough to evaluate a model which is used to solve a multi-classification problem. The precision, recall, f1socre of the method are 0.8244 , 0.7572 and 0.77 ,respectively, and they are higher than the other 2 benchmark methods significantly.

We note that the deep layers generally increase the learning information and greatly improve the accuracy, but the problem is also very obvious, that is, it is common to see the phenomenon of overfitting and time-consuming. Furthure research should be done in the future.

**References**

[1] L. Zhang, Y. Li, Z. Wang, Y. Xia, W. Chen, and K. Tang, “Recent developments and future prospects of Vitreoscilla hemoglobin application in metabolic engineering,” *Biotechnology advances,* vol. 25, no. 2, pp. 123-136, 2007.

[2] G. Wu, L. M. Wainwright, and R. K. Poole, “Microbial globins,” *Advances in microbial physiology,* vol. 47, pp. 257-310, 2003.

[3] U. Consortium, “UniProt: a worldwide hub of protein knowledge,” *Nucleic acids research,* vol. 47, no. D1, pp. D506-D515, 2019.

[4] P. Wright, J. Noirel, S.-Y. Ow, and A. Fazeli, “A review of current proteomics technologies with a survey on their widespread use in reproductive biology investigations,” *Theriogenology,* vol. 77, no. 4, pp. 738-765. e52, 2012.

[5] G. L. Kenyon, D. M. DeMarini, E. Fuchs, D. J. Galas, J. F. Kirsch, T. S. Leyh, W. H. Moos, G. A. Petsko, D. Ringe, and G. M. Rubin, “Defining the Mandate of Proteomics in the Post-Genomics Era: Workshop Report:© 2002 National Academy of Sciences, Washington, DC, USA. Reprinted with permission from the National Academies Press for the National Academy of Sciences. All rights reserved. The original report may be viewed online at <http://www>. nap. edu/catalog/10209. html,” *Molecular & Cellular Proteomics,* vol. 1, no. 10, pp. 763-780, 2002.

[6] M. Debnath, G. B. Prasad, and P. S. Bisen, *Molecular diagnostics: promises and possibilities*: Springer Science & Business Media, 2010.

[7] S. Muthukrishnan, A. Garg, and G. Raghava, “Oxypred: prediction and classification of oxygen-binding proteins,” *Genomics, proteomics & bioinformatics,* vol. 5, no. 3-4, pp. 250-252, 2007.

[8] M. Curcio. "Classification of oxygen binding proteins using Random Forest Machine Learning," <http://rpubs.com/oaxacamatt/Random_Forest_Oxygen_Binde>

rs.

[9] S. Hamena, and S. Meshoul, "Deep neural network for classification and prediction of oxygen binding proteins." pp. 72-77.

[10] G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht, “Use of the ‘Perceptron’algorithm to distinguish translational initiation sites in E. coli,” *Nucleic acids research,* vol. 10, no. 9, pp. 2997-3011, 1982.

[11] W. Li, and A. Godzik, “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences,” *Bioinformatics,* vol. 22, no. 13, pp. 1658-1659, 2006.

[12] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, and M. Devin, “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” *arXiv preprint arXiv:1603.04467*, 2016.

[13] E. Lee, and D. Kim, “Accurate traffic light detection using deep neural network with focal regression loss,” *Image and Vision Computing,* vol. 87, pp. 24-36, 2019.

[14] J. Z. Huang, "An Introduction to Statistical Learning: With Applications in R By Gareth James, Trevor Hastie, Robert Tibshirani, Daniela Witten," Springer, 2014.

[15] T. Reviewer-Heiman, “Pattern Recognition and Neural Networks is written by Brian D. Ripley, and published by Cambridge University Press, 2007, Paperback, ISBN 978-0521-71770-0, pp., 0--403,” *ACM SIGSOFT Software Engineering Notes,* vol. 34, no. 6, pp. 28-28, 2009.

[16] C. Johnco, J. F. McGuire, T. Roper, and E. A. Storch, “A meta‐analysis of dropout rates from exposure with response prevention and pharmacological treatment for youth with obsessive compulsive disorder,” *Depression and anxiety,* vol. 37, no. 5, pp. 407-417, 2020.