EPiC Series in Computing
Volume 60, 2019, Pages 113–119
Proceedings of 11th International Conference
on Bioinformatics and Computational Biology
EPiC Computing

Prediction of MoRFs Based on n-gram Convolutional Neural Network

Chun Fang[1, 2], Yoshitaka Moriwaki [2], Caihong Li [1] and Kentaro Shimizu[2]
[1] Shandong University of Technology, Shandong, China
[2] The University of Tokyo, Tokyo, Japan
fangchun0409@gmail.com, shimizu@bi.a.u-tokyo.ac.jp

Abstract

MoRFs usually play as "hub" site in interaction networks of intrinsically disordered proteins. With more and more serious diseases being found to be associated with disordered proteins, identifying MoRFs has become increasingly important. In this study, we introduce a multichannel convolutional neural network (CNN) model for MoRFs prediction. This model is generated by expanding the standard one-dimensional CNN model using multiple parallel CNNs that read the sequence with different n-gram sizes (groups of residues). In addition, we add an averaging step to refine the output result of machine learning model. When compared with other methods on the same dataset, our approach achieved a balanced accuracy of 0.682 and an AUC of 0.723, which is the best performance among the single model-based approaches.

1 Introduction

Molecular recognition features (MoRFs) are important short binding regions in intrinsically disordered proteins (IDPs) [1]. They easily undergo a state change from unfolding to folding state when interacting with their molecular chaperone [2]. MoRF is very "sticky" and easily forms a "hub" in the protein interaction network [3]. With more and more serious diseases found to be associated with IDPs, identifying MoRFs has become increasingly important for understanding the functional aspects of IDPs, for studying the folding mechanism of proteins, and for finding drug targets [4].

To date, MoRF's prediction has attracted interest of many researchers, and many sequenced based methods have been developed for MoRF's prediction, such as MoRFpred [5], ANCHOR [6], Retro-MoRFs [7] and MoRFPred-plus [8]. These approaches adopted a large number of predicted feature obtained from other tools as input, such as predicted residue disorder probability, predicted secondary structure features, predicted accessible surface area of residues, and predicted dihedral angles. Retro-MoRFs [7], MoRFPred-plus [8] and MoRFCHiBi [9] are ensemble methods that combined the outcomes of several models to obtain a better performance. The design of these methods is complicated and their accuracy still needs to be improved.

O. Eulenstein, H. Al-Mubaid and Q. Ding (eds.), BiCOB 2019 (EPiC Series in Computing, vol. 60), pp. 113–119

The problem of protein functional sites prediction using sequence information is essentially a sequence classification problem. Since protein sequences can be regarded as a kind of natural language, techniques used in natural language processing can also be used for biological sequence classification.

Existing research [10] suggests that convolutional network is useful in extracting information from raw signals, they can be directly applied to distributed or discrete embedding of characters, do not require the knowledge about the syntactic or semantic structure of a language, and simple statistics of some ordered characters combinations (such as n-grams) usually bring better performance in text classification. Therefore, the n-gram probabilistic language model, which has been successfully used in text classification, can also be used for protein functional sites prediction.

In this study, we propose a method based on n-gram multichannel convolutional neural networks for MoRF's prediction. In our approach, we treat protein sequence as a kind of raw signal at character level, and quantize each character using position-specific scoring matrix generated from sequence and 13 amino acid indices. Then one-dimensional Convolutional Neural Network (CNN) with simple statistics of ordered character combinations (n-grams) was adopted to build the prediction model. Finally, an average strategy was applied to the output prediction scores to further improve the accuracy.

2 Methods

2.1 Dataset

For a fair comparison, we adopted the same datasets with the research of MoRFpred [5]. The training dataset comes from 421 MoRF-containing protein chains, all the 5,396 MoRF residues of the 421 chains and an equal number of randomly selected non-MoRF residues were used to build a balanced training dataset. The validation dataset includes 419 MoRF-containing chains, and the test dataset includes 45 MoRF-containing chains. The validation dataset and test dataset shared up to 30% sequence identity with training dataset.

2.2 Feature representation

We adopt position specific scoring matrix (PSSM) generated by PSI-BLAST searching against the NR database, and 13 amino acid indices selected from the AA index database [11] also be adopted to code protein sequences. Each protein chain of length l will be encoded as an $l \times (20+13)$, i.e., $l \times 33$ matrix (*20* values from PSSM and *13* values from AA indices). We set a sliding window size of *51* to incorporate all possible MoRF residues into the window area for a central residue (*25* residues on each side). By such encoding, the input feature of each residue is transformed to a 51×33 matrix for the model.

2.3 Optimized the prediction results

Since MoRFs are short regions, we adopted an averaging step to filter the points that were predicted to be isolated MoRF or non-MoRF residue. The output probabilities of the CNN model were averaged by a sliding window. ZEROs were padded to the terminals of a sequence. The average probability within the sliding window is the final result of the center residue. Detailed averaging step was applied according to Equation (1).

$$average\_C_i = \frac{1}{2m+1}\sum_{i-m}^{i+m} C_i, (i = 1,...N) \quad (1)$$

where N is the sequence length, *2m+1* is the averaging window size, C_i is the output probability of residue *i* from the CNN model.

2.4 Classification model

Overview of the proposed method is shown in Figure 1. Firstly, each sequence is mapped to an l × 33 feature matrix; secondly, input vectors constructed by n-gram model are imported into a convolutional neural network to make a prediction; finally, the output probabilities of CNN model were further processed by an average step to generate the final prediction result. In this study, the combination of a 2-gram channel and a 3-gram channel were adopted to build the multi-channel model. The detailed structure of model is shown in Figure 2.

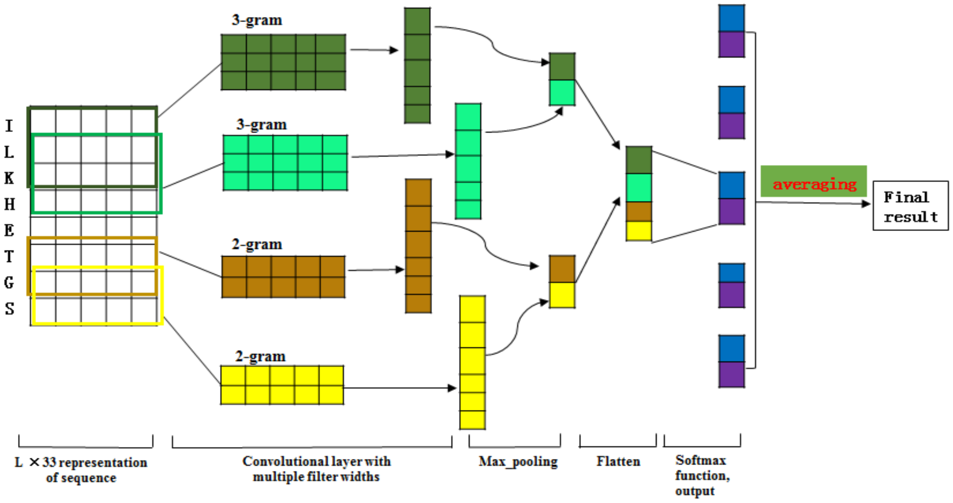

Figure 1: Overview of the proposed method

2.5 Evaluation criteria

Performance of the classification methods in this study are evaluated by the following measures: (1) Area under the ROC curves (AUC); (2) Balanced accuracy (ACC); (3) False positive rate (FPR) which measures the fraction of negative examples that are misclassified as positive; (4) True positive rate (TPR) which measures the fraction of positive examples that are correctly labeled. These measures are defined as follows:

$$Specifity = \frac{TN}{TN + FP} \qquad (2)$$

$$TPR = Sensitivity = \frac{TP}{TP + FN} \qquad (3)$$

$$FPR = 1 - Specificity = \frac{FP}{TN + FP} \qquad (4)$$

$$ACC = \frac{1}{2}(Sensitivity + Specifity) \qquad (5)$$

where true positives (TP) refers to examples correctly labeled as positives, false positives (FP) refers to negative examples incorrectly labeled as positive, true negatives (TN) refers to negatives correctly labeled as negative, and false negatives (FN) refers to positive examples incorrectly labeled as negative.

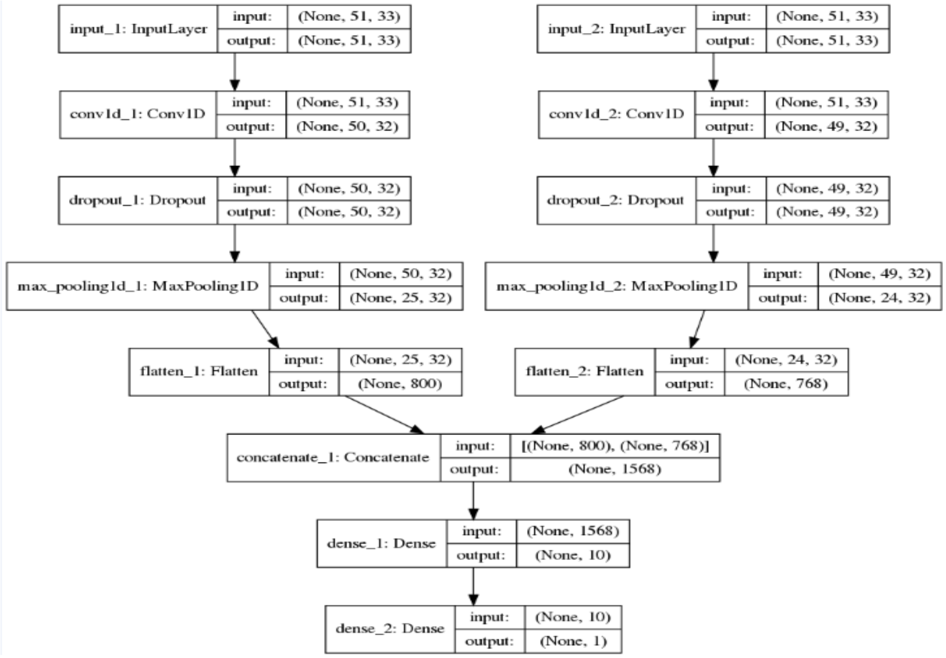

Figure 2: The structure of multi-channel CNN model

3 Results and discussion

3.1 Performance comparison of methods with different integration of n-gram channels

In order to optimize the structure of the prediction model, we have compared four methods with different integration of n-gram channels: (1) the “2_gram” method based on the single 2-gram CNN channel; (2) “2_3_gram” method based on the combination of 2-gram and 3-gram CNN channels; (3) the “2_3_4_gram” method based on the combination of 2-gram, 3-gram and 4-gram CNN channels; (4) the “2_3_4_5_gram” method based on the combination of 2-gram, 3-gram, 4-gram and 5-gram CNN channels. The corresponding ROC plots of the four methods tested on the validation dataset are shown in Figure 3. Figure 3 demonstrates that the “2_3_gram” method outperforms the other methods. Therefore, the combination of a 2-gram channel and a 3-gram channel were adopted to build the multi-channel model of our model in this study.

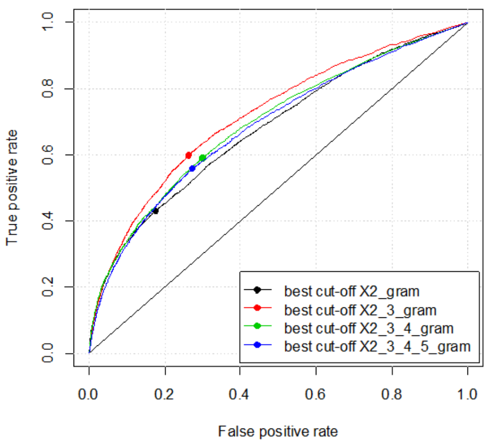

Figure 3: ROC plots of the methods with different integration of n-gram channels tested on the validation dataset

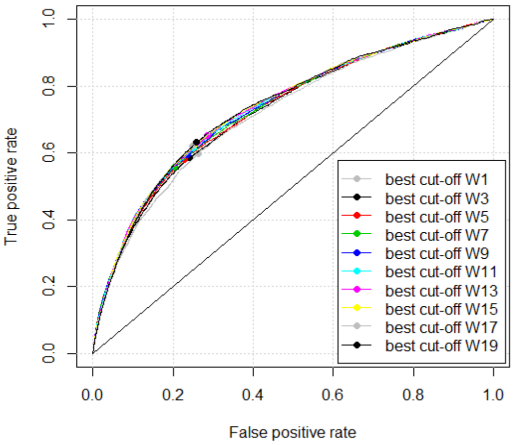

Figure 4: ROC plots of "2_3_gram" method with different averaging window size tested on the validation dataset

3.2 Optimizing the averaging window size

In order to obtain an appropriate sliding window size, we tested our method with different average window sizes. ROC plots of the "2_3_gram" method with different average window size tested on the validation dataset are shown in Figure 4. The "2_3_gram" method with average window size of 17 achieves the best performance. Thus, the average sliding window size was set to 17 for our approach. To illustrate the effectiveness of the averaging step, ROC plots of "2_3_gram" method with and without the average step tested on the validation dataset are shown in Figure 5. The performance has been significantly improved by the average step, achieving an AUC of 0.737 (Figure 6).

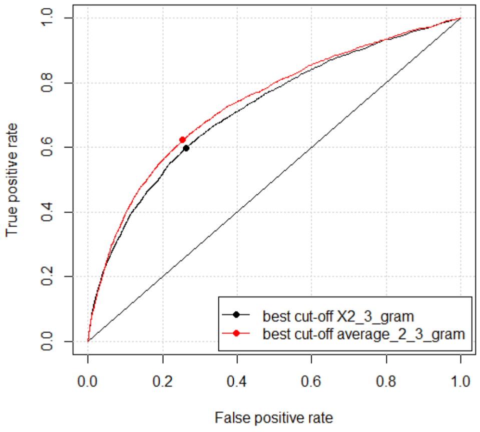

Figure 5: ROC plots of "2_3_gram" method with and without the average step on the validation dataset

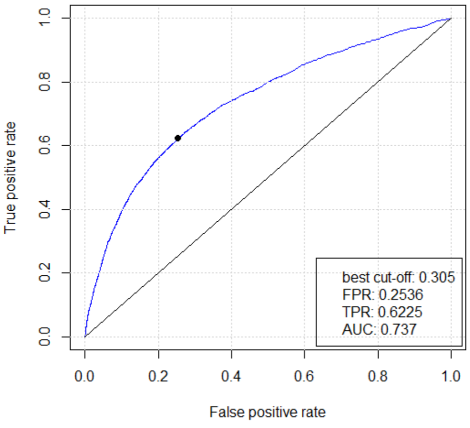

Figure 6: ROC plot of our proposed method tested on the validation dataset

3.3 Performance Comparison with the existing predictors

Some existing methods for predicting MoRFs mainly rely on combining results of multiple classifiers. Their accuracy largely depends on the strategy of “win by more”. Our method in this study is based on a single machine-learning model with multiple channels to extract features with different group of residues (n-gram), so we only compare our method with three representative single model based methods: the MoRFpred [5], ANCHOR [6] and MFSSPSMpred [12]. ROC plots of all the methods tested on the test dataset are shown in Figure 7 (ROC plot of the MoRFpred cannot be drawn because their web tool has limitation on length of input sequences). The detailed performance comparisons with respect to ACC, TPR, FPR and AUC are shown in Table 2. Figure 7 and Table 2 show that, our proposed method performs the best compared to the other methods, achieving an ACC of 0.682 and an AUC of 0.723.

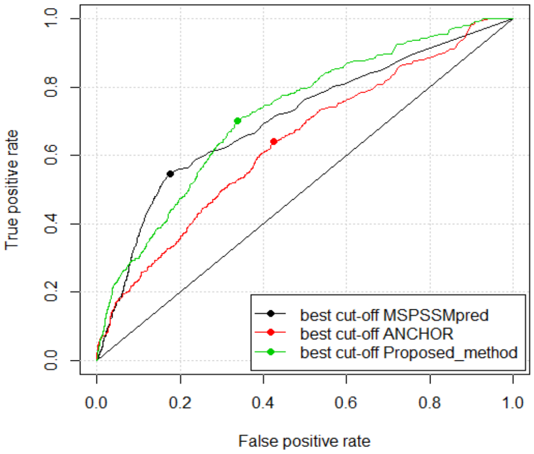

Figure 7: Performance comparisons with existing methods on the test dataset.

Table1: Performance comparisons with the existing predictors on the test dataset.

Method	ACC	TPR	FPR	AUC
MoRFpred [5]*	0.596	0.236	0.045	0.697
ANCHOR [6]	0.608	0.641	0.425	0.638
MFSPSSMPred [12]	0.682	0.540	0.176	0.706
Proposed method	0.682	0.700	0.337	0.723

*Note: Result of MoRFpred was quoted from paper [5]; the other results come from their web servers.

4 Conclusions

In this study, we proposed a deep learning approach, which leverages the local contexts of protein sequence with 2-gram and 3-gram characters via one-dimensional CNN for MoRFs prediction. First, sequences were mapped to feature matrixes by calculating PSSM files of sequences and 13 amino acid indices; second, the 2-gram and 3-gram models were applied to construct the input vectors, and these vectors were imported into a one-dimensional CNN to make predictions; finally, the output probabilities of the CNN model were further refined by the averaging step. Experiments conducted on the same datasets demonstrated that, our proposed method outperformed the other single model-based approaches, achieving a balanced accuracy of 0.682 and an AUC of 0.723, which provided significant performance improvement for the MoRFs prediction based on single machine learning model.

Acknowledgment

This research is supported by National Natural Science Foundation of China (61602280, 61473179), the Natural Science Foundation of Shandong province (ZR2014FQ028), the Japan Society for the Promotion of Science (17F17050), and Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS) from the Japan Agency for Medical Research and Development (AMED).

References

[1] Uversky VN.(2014). Introduction to intrinsically disordered proteins (IDPs). *Chem Rev.* 114(13):6557-60,

[2] Mohan A, Christopher JO, Radivojac P, et al. (2006) Analysis of Molecular Recognition Features (MoRFs). *J. Mol. Biol,* 362: 1043–1059.

[3] Gang H, Zhonghua W, Vladimir N.U, et al. (2017) Functional Analysis of Human Hub Proteins and Their Interactors Involved in the Intrinsic Disorder-Enriched Interactions. *Int. J. Mol. Sci.* 18(12), 2761.

[4] Uversky VN, Oldfield CJ, Dunker AK. (2008).Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys.* 37:215-46.

[5] Fatemeh MD, Wei-Lun H**,** Marcin JM, et al. (2012). MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics.* 28(12): i75-i83.

[6] Dosztanyi Z, Mészáros SI. (2009). ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics. 25(20): 2745-2746.

[7] Xue B, Dunker AK, Uversky VN. (2010). Retro-MoRFs: identifying protein binding sites by normal and reverse alignment and intrinsic disorder prediction. *Int J Mol Sci.* 11(10): 3725-47.

[8] Sharma R, Bayarjargal M, Tsunoda T, et al. (2018). MoRFPred-plus: Computational Identification of MoRFs in Protein Sequences using Physicochemical Properties and HMM profiles. *J Theor Biol.* 437:9-16

[9] Malhis N and Gsponerr J. (2015) Computational identification of MoRFs in protein sequences, *Bioinformatics,* 31(11): 1738–1744.

[10] Ye Zhang, Byron C. Wallace. (2015). A Sensitivity Analysis of Convolutional Neural Networks for Sentence Classication, In *Proceedings of the The 8th International Joint Conference on Natural Language Processing,* (pp. 253-2634)

[11] Kawashima S, Pokarowski P, Pokarowska M, et al. (2008) AA index: amino acid index database. *Nucleic Acids Research,* 36:D202-D205.

[12] Fang C, Noguchi T, Tominaga D, et al. (2013). MFSPSSMpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation. *BMC Bioinformatics,* 14:300.