ENSEMBLE-CNN: Predicting DNA Binding Sites in Protein Sequences by an Ensemble Deep Learning Method

Yongqing Zhang[1,2], Shaojie Qiao[3(✉)], Shengjie Ji[1], and Jiliu Zhou[1]

[1] School of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China
[2] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
[3] School of Cybersecurity, Chengdu University of Information Technology, Chengdu 610225, China
sjqiao@cuit.edu.cn

Abstract. Detection of DNA binding sites in proteins plays an essential role in gene regulation processing. However, the difficult problem in developing machine learning predictors of DNA binding sites in protein is that: the number of DNA binding sites is significantly fewer than that of non-binding sites. Aiming to handle this issue, we propose a new predictor, named ENSEMBLE-CNN, which integrates instance selection and bootstrapping techniques for predicting imbalanced DNA-binding sites from protein primary sequences. ENSEMBLE-CNN uses a protein's evolutionary information and sequence feature as two basic features and employs sampling strategy to deal with the class imbalance problem. Multiple initial predictors with CNNs as classifiers are trained by applying SMOTE and a random under-sampling technique to the original negative dataset. The final ensemble predictor is obtained by majority voting strategy. The results demonstrate that the proposed ENSEMBLE-CNN achieves high prediction accuracy and outperforms the existing sequence-based protein-DNA binding sites predictors.

Keywords: Protein-DNA binding sites · Deep learning · Ensemble method Imbalance learning

1 Introduction

DNA-binding proteins are the proteins composed of DNA-binding domains. The interactions between these proteins and DNA play a crucial role in vital biological process [1, 2]. A number of high throughput experimental techniques have been developed to confirm the interactions between DNA and proteins, such as protein binding microarray(PBM) [3], ChIP-seq [4] and protein microarray assays [5]. However, the existing approaches are costly and time-consuming. Thus, there is urgent need to propose computational methods for predicting protein-DNA binding sites from sequences in an efficient and effective fashion.

D.-S. Huang et al. (Eds.): ICIC 2018, LNCS 10955, pp. 301–306, 2018.
https://doi.org/10.1007/978-3-319-95933-7_37

Currently, a series of computational methods have been proposed to predict DNA-binding sites in protein. Based on the discovered features for prediction, these methods can be partitioned into three groups: evolutionary features based methods, sequence features based methods, and the ones based on structure features. During the past decade, a number of machine learning algorithms have been used to predict DNA-binding sites from protein sequences, including BindN [6], BindN+ [7], ProteDNA [8], DP-Bind [9], MetaDBSite [10], DNABind [11] and TargetDNA [12]. These sequence-based predictors only utilize protein sequence information and recognize DNA-binding sites with one or more machine learning algorithms. Despite the promising results of these methods, there remains room for further improvements in accurately predicting DNA-binding sites form protein sequence.

Another important issue in machine learning predictors of protein-DNA binding sites is the severe intrinsic class imbalance problem: the number of DNA binding sites (minority class) is apparently fewer than that of non-binding sites (majority class). Re-sampling is the most straightforward strategy for dealing with the issue of class imbalance [13]. Based on the aforementioned problem, we proposed a sequence-based predictor, named ENSEMBLE-CNN, for the computational identification of DNA binding sites. First, we employ the protein evolutionary information and sequence features, which are determined solely from protein sequences. Next, SMOTE [14] is used to over-sample positive data. Then, we train multiple DNA binding site predictors with CNNs as a basic classifier by applying a bootstrap technique on the original imbalanced data. Lastly, we obtain the ensembled predictor by using the majority voting strategy.

2 Methods

2.1 Feature Descriptors

From the point of view of machine learning, prediction of DNA binding sites in proteins is actually a traditional binary classification problem. Various effective sequence-based feature, such as position specific scoring matrix (PSSM) [9], predicted secondary structure [15] and physicochemical properties [16], have been explored for predicting protein DNA binding residues. In this study, we employ PSSM feature and sequence feature for predicting DNA binding sites in proteins.

PSSM, being a very important type of evolutionary features, is obtain by running the PSI-BLAST [17] program to search the SwissProt database [18] via three iteration, with 10–3 as the E-value cutoff for multiple sequence alignment. In PSSM, there are 40 scores for each sequence position and each score implies the conservation degree of a specific residue type on that position. For each data instance, all the scaled scores in the PSSM are used as its evolution features. In this study, we use the window size $w = 15$.

Sequence features include local amino acid composition, predicted second structure and predicted solvent accessible area. Each probe sequence is converted into a $4 \times L$ one-hot coded binary matrix (L is the probe length) and the intensity values are normalized.

2.2 SMOTE: Synthetic Minority Over-Sampling Technique

SMOTE [14] is an over-sampling approach in which the minority class is oversampled by creating "synthetic" examples rather than over-sampling with replacement. The minority class is over-sampled by taking each minority class sample and generating synthetic examples by the line segments joining approached on k nearest neighbors corresponding to each minority class. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly selected. In our model, we use 5 nearest neighbors.

2.3 Convolutional Neural Networks (CNN)

Most modern deep learning models are based on an artificial neural network [19–21]. Recently, deep learning techniques have demonstrated their capability of improving the discriminative power compared with other machine learning methods [22], and have been widely applied to the field of bioinformatics. The basic components of a CNN include convolutional, pooling and fully connected layers.

2.4 ENSEMBLE-CNN

We trained SMOTE and multiple different classifiers on balanced datasets obtained by applying random under-sampling method. Then, these trained classifiers are ensembled by the majority voting strategy. In this study, ENSEMBLE-CNN is trained by using the standard back-propagation algorithm and mini-batch gradient descent with the Adagrad variation. Dropout and early stopping strategies are used for regularization and model selection.

3 Experimental Results

3.1 The Datasets

In this study, three datasets, datasets PDNA-543 [12] and PDNA-TEST [12], were used to evaluate the performance of our method. PDNA-543 consists of 543 protein sequences; there are 9,549 DNA-binding residues as positive samples and 134,995 non-binding residues as negative samples. PDNA-TEST has 41 protein chains, which includes 734 positive samples and 14,021 negative samples.

Six famous evaluation measurements are used, which is the same as in [12]. Sensitivity (Sen), Specificity (Spe), Accuracy (Acc), Precision (Pre), the Mathew's Correlation Coefficient (MCC) and AUC are utilized to evaluate prediction performance.

3.2 Comparison with Different Features

This set of experiments examines the contributions of the three different kinds of features in ENSEMBLE-CNN for the DNA binding sites in proteins prediction on the training dataset. The detail results are given in Table 1. As mentioned above, Sen, Spe,

Acc, Pre, MCC and AUC are the main metrics. It can be observed that the PSSM2 + One-hot coding features outperforming the PSSM2 features by 5.04% for Sen, 15.79% for Spe, 14.49% for ACC, 29.81% for Pre, 0.276 for MCC and 0.114 for AUC. When the three kinds of features are combined, ENSEMBLECNN achieves 0.632 for MCC and 0.933 for AUC, which indicates that these features are complementary for each other and the one-hot coding feature is the important method for effectively predicting protein-DNA binding.

Table 1. The performance on PDNA-543 for various features by ten-fold cross-validation.

Feature	Sen (%)	Spe (%)	Acc (%)	Pre (%)	MCC	AUC
PSSM2	71.40	77.06	76.38	29.98	0.349	0.812
PSSM2 + One-hot coding	76.44	**92.85**	**90.87**	**59.79**	0.625	0.926
PSSM1 + PSMM2 + One-hot coding	**79.48**	92.23	90.69	58.70	**0.632**	**0.933**

3.3 Predicted Results on Independent Test

In this section, we demonstrate the effectiveness of the proposed method, ENSEMBLECNN, by comparing it with other commonly-used predictors of DNA binding sites in proteins, including BindN [6], BindN+ [7], ProteDNA [8], DP-Bind [9], MetaDBSite [10], DNABind [11] and TargetDNA [12], by performing independent validation tests on PDNA-TEST, the results of which are shown in Table 2.

Table 2. The predicting performance compared with other predictors on PDNATEST (the value of other predictors are from [12])

Predictor	Sen (%)	Sep (%)	Acc (%)	Pre (%)	MCC
Bind	45.64	80.90	79.15	11.12	0.143
ProteDNA	4.77	99.84	95.11	60.30	0.160
BindN + (FRP $\approx$ 5%)	24.11	95.11	91.58	20.51	0.178
BindN + (Sep $\approx$ 85%)	50.81	85.41	83.69	15.42	0.213
MetaDBSite	34.20	93.35	90.41	21.22	0.221
DP-Bind	61.72	82.43	81.40	15.53	0.241
DNABind	70.16	80.28	79.78	15.70	0.264
TargetDNA(Sen $\approx$ Spe)	60.22	85.79	84.52	18.16	0.269
TargetDNA(FPR $\approx$ 5%)	45.50	93.27	90.89	26.13	0.300
ENSEMBLE-CNN	48.10	91.20	89.08	21.99	0.274

Table 2 shows that ENSEMBLE-CNN achieves satisfactory results with the second-best MCC value of 0.274. When compared with TargetDNA(Sen $\approx$ Spe), ENSEMBLE-CNN achieves an high Sen value. As for BindN+, which is an improved version of BindN, ENSEMBLE-CNN achieves an improvement of 6.1% on MCC. By comparing with MetaDBSite, the proposed ENSEMBLE-CNN also achieves

improvements of 13.9 and 5.3% on Sen and MCC, respectively. DNABind achieves the best performance on Sen (70.16%), but a much lower Spe value, implying too many false positive are incurred during prediction.

4 Conclusions

In this study, we proposed a new sequence-based predictor of protein-DNA binding sites, called ENSEMBLE-CNN. It is trained on the DNA-binding protein dataset collected from the most recently released PDB with a SMOTE, CNN, and the bootstrap classifier ensemble strategy. Experimental results with a training dataset and an independent validation dataset have demonstrated the effectiveness of the proposed ENSEMBLE-CNN. In terms of our future work, we will further investigate the applicability of our model to other types of molecules binding sites prediction problems.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grants (No. 61702058, 61772091), the China Postdoctoral Science Foundation funded project (No. 2017M612948), the Scientific Research Foundation for Advanced Talents of Chengdu University of Information Technology under Grant (No. KYTZ201717, KYTZ201715, KYTZ201750), the Scientific Research Foundation for Young Academic Leaders of Chengdu University of Information Technology under Grant (No. J201701, J201706), the Planning Foundation for Humanities and Social Sciences of Ministry of Education of China under Grant (No. 15YJAZH058), and the Innovative Research Team Construction Plan in Universities of Sichuan Province under Grant (No. 18TD0027).

References

1. Si, J., Zhao, R., Wu, R.: An overview of the prediction of protein DNA-binding sites. Int. J. Mol. Sci. **16**(3), 5194–5215 (2015)
2. Wong, K.C., Li, Y., Peng, C., Wong, H.S.: A comparison study for DNA motif modeling on protein binding microarray. IEEE/ACM Trans. Comput. Biol. Bioinform. **13**(2), 261–271 (2016)
3. Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., Bulyk, M.L.: Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol. **24**(11), 1429–1435 (2006)
4. Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M., Sidow, A.: Genomewide analysis of transcription factor binding sites based on chip-seq data. Nat. Methods **5**(9), 829–834 (2008)
5. Ho, S.W., Jona, G., Chen, C.T., Johnston, M., Snyder, M.: Linking DNA-binding proteins to their recognition sequences by using protein microarrays. Proc. Nat. Acad. Sci. U.S.A. **103** (26), 9940–9945 (2006)
6. Wang, L., Brown, S.J.: BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res. **34**(Web Server issue), W243 (2006)
7. Wang, L., Huang, C., Yang, M.Q., Yang, J.Y.: BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst. Biol. **4**(S1), S3 (2010)

8. Chu, W.Y., Huang, Y.F., Huang, C.C., Cheng, Y.S., Huang, C.K., Oyang, Y.J.: ProteDNA: a sequence-based predictor of sequence-specific DNA-binding residues in transcription factors. Nucleic Acids Res. **37**(Web Server issue), W396 (2009)
9. Hwang, S., Gou, Z., Kuznetsov, I.B.: DP-bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics **23**(5), 634–636 (2007)
10. Si, J., Zhang, Z., Lin, B., Schroeder, M., Huang, B.: MetaDBSite: a meta approach to improve protein DNA-binding sites prediction. BMC Syst. Biol. **5**(S1), S7 (2011)
11. Li, B.Q., Feng, K.Y., Ding, J., Cai, Y.D.: Predicting DNA-binding sites of proteins based on sequential and 3D structural information. Mol. Genet. Genomics **289**(3), 489–499 (2014)
12. Hu, J., Li, Y., Zhang, M., Yang, X., Shen, H.B., Yu, D.J.: Predicting protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs. IEEE/ACM Trans. Comput. Biol. Bioinform. **PP**(99), 1389–1398 (2016)
13. Hu, J., Li, Y., Yan, W.X., Yang, J.Y., Shen, H.B., Yu, D.J.: KNN-based dynamic query-driven sample rescaling strategy for class imbalance learning. Neurocomputing **191**, 363–373 (2016)
14. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**(1), 321–357 (2011)
15. Ahmad, S., Gromiha, M.M., Sarai, A.: Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics **20**(4), 477–486 (2004)
16. Wong, K.C., Li, Y., Peng, C., Moses, A.M., Zhang, Z.: Computational learning on specificity-determining residue-nucleotide interactions. Nucleic Acids Res. **43**(21), 10180–10189 (2015)
17. Schffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E. V., Altschul, S.F.: Improving the accuracy of psi-blast protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. **29**(14), 2994–3005 (2001)
18. Bairoch, A., Apweiler, R.: The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. **28**(1), 45–48 (2000)
19. Huang, D.-S.: Radial basis probabilistic neural networks: model and application. Int. J. Pattern Recogn. Artif. Intell. **13**(07), 1083–1101 (1999)
20. Huang, D.S., Du, J.X.: A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks. IEEE Trans. Neural Netw. **19**(12), 2099–2115 (2008)
21. Zhang, J.-R., Zhang, J., Lok, T.-M., Lyu, M.R.: A hybrid particle swarm optimization–back-propagation algorithm for feedforward neural network training. Appl. Math. Comput. **185**(2), 1026–1037 (2007)
22. Huang, D.-S.: A constructive approach for finding arbitrary roots of polynomials by neural networks. IEEE Trans. Neural Netw. **15**(2), 477–491 (2004)