Computational Biology and Bioinformatics
2018; 6(2): 31-35
http://www.sciencepublishinggroup.com/j/cbb
doi: 10.11648/j.cbb.20180602.11
ISSN: 2330-8265 (Print); ISSN: 2330-8281 (Online)

Prediction of Escherichia Coli K-12 Promoters Using Convolutional Neural Network

Lu Wang, Ping Wan*

College of Life Sciences, Capital Normal University, Beijing, China

Email address:
wanping@cnu.edu.cn (Ping Wan)

*Corresponding author

To cite this article:
Lu Wang, Ping Wan. Prediction of Escherichia Coli K-12 Promoters Using Convolutional Neural Network. *Computational Biology and Bioinformatics*. Vol. 6, No. 2, 2018, pp. 31-35. doi: 10.11648/j.cbb.20180602.11

Received: October 11, 2018; **Accepted**: October 31, 2018; **Published**: November 30, 2018

Abstract: Promoters are significant cis-acting elements in genomes and play important roles in gene regulation. Each gene is regulated by a specific type of promoter, so determining the type of promoter for regulation of a gene is crucial to explore the gene function. Although some computational methods to predict promoters have been proposed, their performances are not satisfying. Convolutional neural network (CNN) is a powerful model in deep learning, it has been applied in bioinformatics in recent years. To improve the performance of promoter prediction, in this study, six types of Escherichia coli K-12 promoter DNA sequences were collected from the RegulonDB database, and constructed a CNN model to predict promoters using the Keras platform. The CNN model is composed of two convolutional layers, three dropout layers, four batch normalization layers and one hidden layer. To evaluate the performances of the CNN model, the 10-fold cross-validation and the receiver operating characteristic (ROC) curve plotting were performed. The results show, the accuracies of predictions for promoters sigma 24, sigma 28, sigma 32, sigma 38, sigma 54 and sigma 70 are 94%, 97%, 95%, 95%, 97% and 83%, respectively. The convolutional neural network model achieves the highest accuracy in promoter prediction up to now. In conclusion, CNN is the best model in promoter prediction, and it will be a promising model both in DNA and protein sequence analysis.

Keywords: Convolutional Neural Network (CNN), Escherichia Coli, Promoter, Prediction

1. Introduction

Promoters are vital cis-acting element in the regulation of gene expression. It is recognized by RNA polymerase and related sigma factors [1, 2]. Till now, seven types of promoters are found in Escherichia coli K-12 genome, including sigma19, sigma24, sigma28, sigma32, sigma38, sigma54 and sigma70.

Since experimental approaches for promoter detection are expensive and time-consuming, several computational methods have been developed for predicting promoters, they are: PSSM (Position-specific scoring matrix), IDQD (Increment of diversity with quadratic discriminant analysis), PCSM (Position-correlation scoring matrix) and PWM (position weight matrices) [3-6]. However, none of these methods could reach good performance for all seven types of promoters.

Deep learning is an extremely crucial branch of machine learning, it has developed rapidly in recent years and has been widely applied in the fields of protein sequence prediction, enhancer recognition, and medical data modeling [7-10]. Deep learning exhibits excellent generalization ability in extracting and discerning various features in dataset [11, 12]. With the increasing of computer performance, deep learning has been feasible in dealing with more complex machine learning models [13].

Convolutional neural network (CNN) is one of the most important model in deep learning [14, 15]. In recent years, CNN has been widely used to solve biological problems. However, no study has been reported on prediction of bacterial promoters using CNN. In this study, with the Keras platform, a convolutional neural network was constructed to predict Escherichia coli K-12 promoters. The CNN model achieves the best performance in promoter prediction up to now.

2. Data & Methods

2.1. Collection of Escherichia Coli K-12 Promoter Nucleic Acid Sequences

Escherichia coli K-12 promoter nucleic acid sequences were collected from the Regulon DB database (http://regulondb.ccg.unam.mx/). Due to only one sigma19 promoter sequence is stored in the RegulonDB database, the sigma19 promoter was not considered in this study. The numbers of sequences of sigma24, sigma28, sigma32, sigma38, sigma54 and sigma70 are 517, 141, 307, 169, 94 and 1894, respectively.

2.2. Encoding of Promoters Sequences

Each promoter sequence consists of 81 bases. Since the CNN receives a two-dimensional matrix, DNA sequences were first transformed into an 4×81 two-dimensional matrix by encoding each base with four digital list: A = [1,0,0,0], T = [0,1,0,0], C = [0,0,1,0], G = [0,0,0,1].

2.3. Construction of Convolutional Neural Network

A typical structure of CNN consists of input layer, convolutional layer, pooling layer, full layer and output layer [16]. Figure 1 shows the CNN structure in this study. The input layer receives the 4×81 two-dimensional matrix transformed from the promoter DNA sequence. Since the input data is not a RGB image, the number of channel in the input layer is one. To obtain more features from the promoter sequences, three convolutional layers were set, the numbers of kernels for each layer is 16, 8 and 32, respectively. The kernel size for each convolutional layer is 2×2, 1×2 and 2×2, respectively. Each convolutional layer follows a batch normalization layer and a dropout layer, the dropout value is 50%. After the third dropout layer is the flatten layer and hidden layer. Flatten layer transforms the 2-dimentional data generated in the fore layer into 1-dimentional data. The hidden layer is a normal full connected neural network with 50 neurons. The output of the hidden layer is transmitted to the output layer. Since the prediction in this study is a binary classification question, the sigmoid activation function was used in the output layer, while the rectifier (relu) activation function is used in other layers.

The pooling layer is usually following the convolutional layer. Two pooling methods, maxpooling and meanpooling, are available in CNN [17]. Considering the small size of the input matrix in this study, the pooling layer was not set. In order to improve the generalization ability of the network, the batch normalization method was applied [18]. Besides, the dropout regularization method was adopted to avoid overfitting [19].

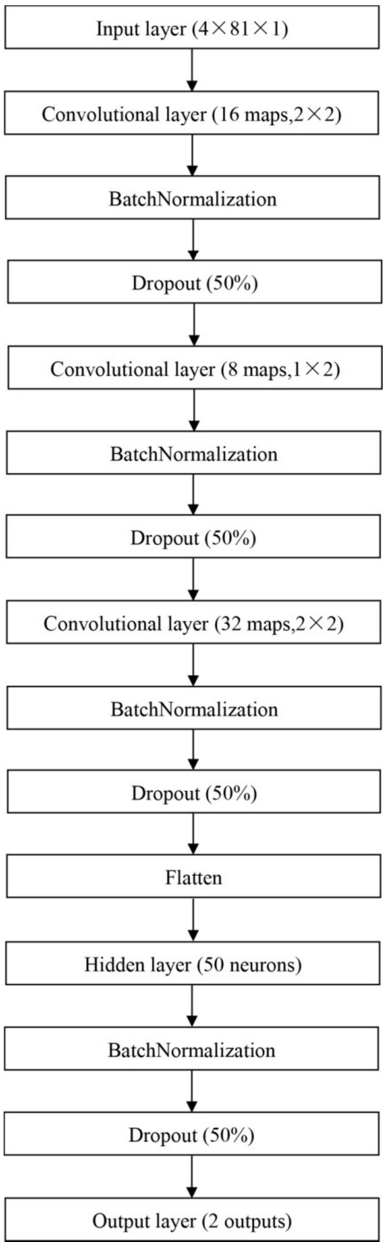


Figure 1. *Structure of the CNN model.*

2.4. Evaluation of the Prediction Performance

To evaluate the performances of the predictions for each types of promoters, the 10-fold cross-validation was performed for the CNN model.

2.5. ROC Curve

ROC curve is a tool to evaluate the performance of a prediction model. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The TRP is also known as sensitivity, the FPR is also known as the probability of false alarm [20]. An ROC space is defined by FPR and TPR as x and y axes, respectively. The diagonal divides the ROC space. Points above the diagonal represent good classification results (better than random); points below the line represent bad

results (worse than random). The area under the curve is called AUC, short for Area Under Curve. The more the AUC nears 1, the more a model performs better.

Specificity (Spec) and sensitivity (Sens) are also two important indicators to evaluate a prediction performance [20]. When calculating the Spec and Sens, four items are included: TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative). Concretely:

$$Sens = TP / TP + FN \quad (1)$$

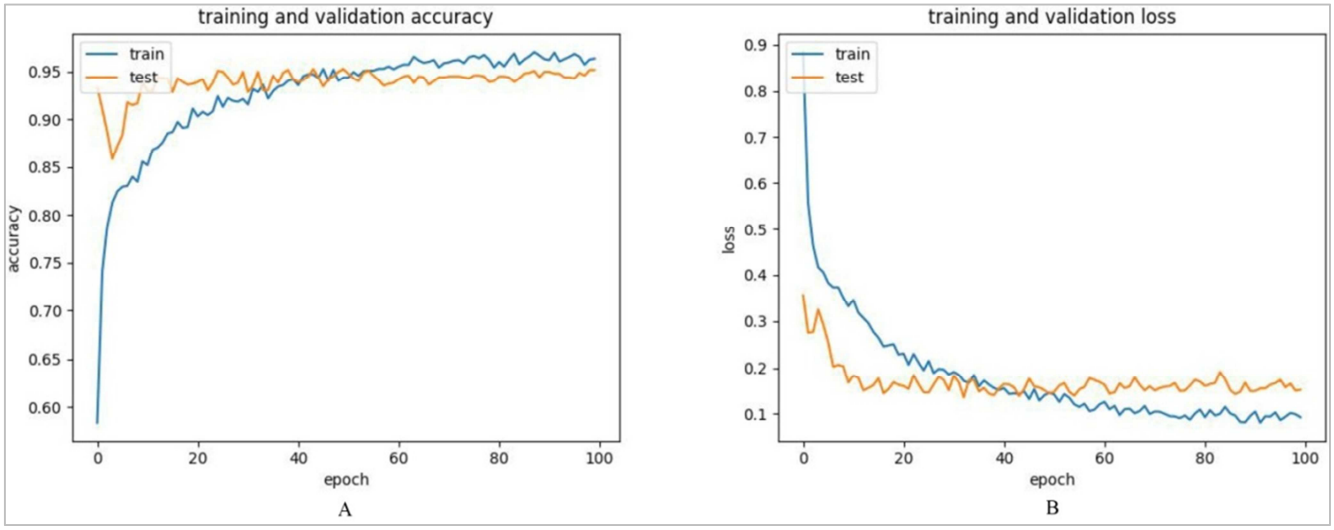
$$Spec = TN / FP + TN \quad (2)$$

In this study, the ROC curve was applied to evaluate the CNN model. For a specific type of promoter, the true positive (TP) dataset is itself, and the true negative (TN) dataset is composed of the other five types of promoters.

3. Result and Discussions

3.1. Determine the Number of Training Epoch

The number of training epoch is an important parameter in a CNN model. When the number of training epoch is inadequate, the model could not reach its best performance; if the number is too large, the model will waste too much time and computing resources. The changing curve of the training and validation accuracy (Figure 2A) and training and validation loss (Figure 2B) were used to optimize the parameter.



(Taking sigma 24 for an example)

Figure 2. *CNN model accuracy and loss on train and validation dataset*

Figure 2 shows, the accuracy curve of training datasets and validation datasets is not improved significantly after 40 epochs (Figure 2A), and keeps stable after 100 epochs (not shown). The loss curve of training datasets and validation datasets also exhibits no significant fluctuation after 40 epochs (Figure 2B), and keeps stable after 100 epochs (not shown). Combining the Figure 2A and 2B, the number of training epoch was set to 100 in the final CNN model.

3.2. Accuracy of the Convolutional Neural Network Prediction

Accuracy is an important measure to evaluate a prediction model. Table 1 compares the prediction accuracy of the CNN model with two previous models. Clearly, the CNN model surpasses PSSM and BacPP models in all types of promoters [21, 22]. Except for promoter σ38, PSSM model is 1% better than CNN (96% vs 95%). And for promoter σ70, BacPP model is 1% better than CNN (84% vs 83%).

Table 1. *Comparison of Accuracy (%) among CNN, PSSM and BacPP models in prediction of E. coli promoters.*

Methods	σ24	σ28	σ32	σ38	σ54	σ70
CNN	94	97	95	95	97	83
PSSM	86	96	93	96	97	74
BacPP	87	93	92	89	97	84

Obviously, the convolutional neural network is a powerful model in promoter prediction.

3.3. ROC Curves

ROC curve is a standard measurement to evaluate a prediction model. Figure 3 shows the ROC curves for the six types of promoters. For promoters sigma24, sigma28, sigma32, sigma70, their AUC are greater than 0.9, indicating the predictions for these promoters are ideal. For sigma38 and sigma54, their AUC is less than 0.9, the possible reasons may be the features within these two type of promoter are too weak to extract with the CNN model. The previous study also has

observed that the conservative around the -30 region in sigma38 is very weak [21]. The conservatives around -30 region and -10 region in sigma 54 are ambiguous as well.

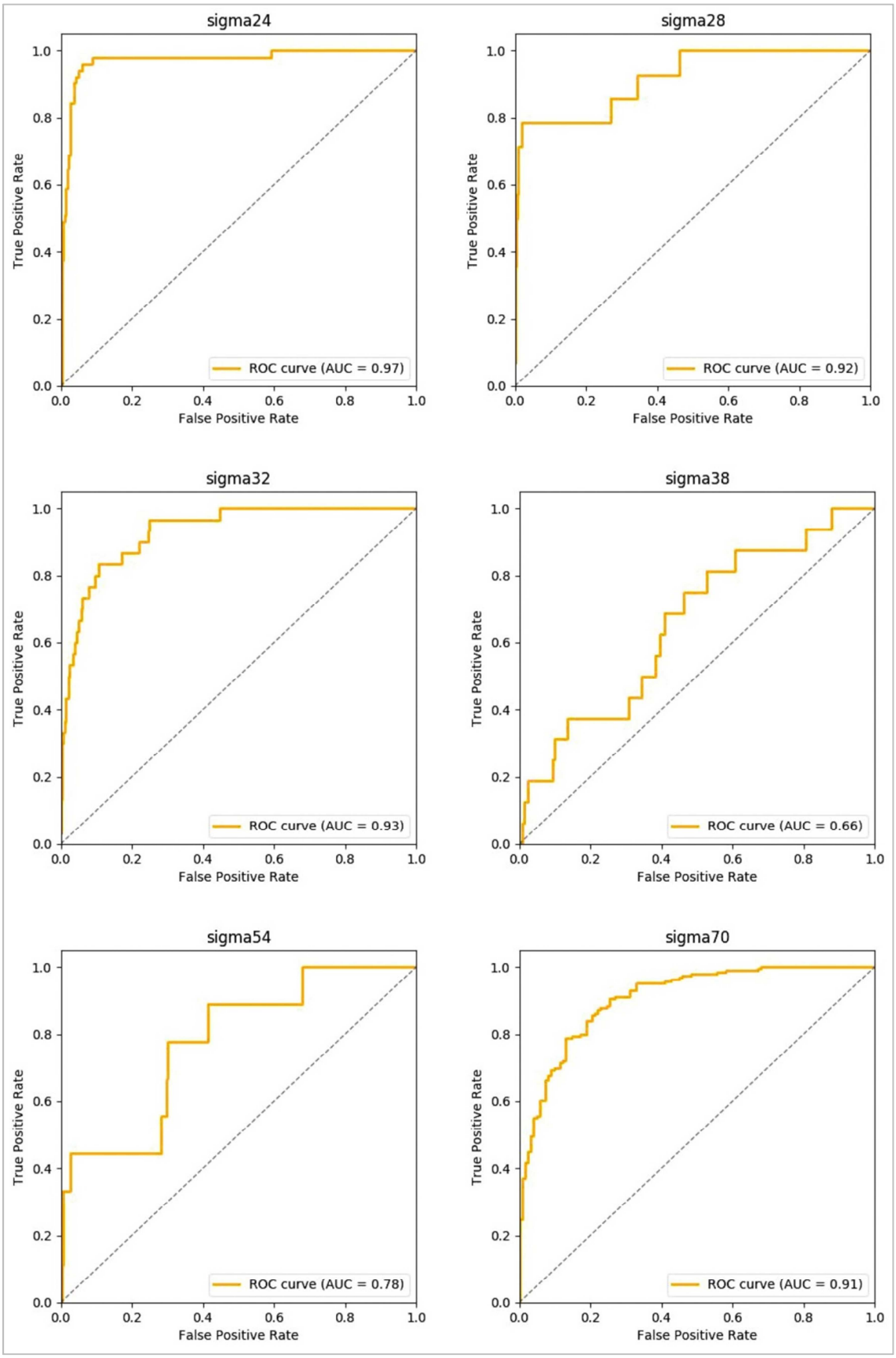


Figure 3. *The ROC curves of E. coli promoter predictions using CNN.*

The sensitivities and specificities provide more details of the CNN model. Table2 shows the sensitivities and specificities of the CNN model in predicting the six types of promoters. Consistently, the sensitivities of sigma38 and sigma54 are both less than 0.6, and their specificities are also not high. The results suggested that, in order to obtain better performance, more specific models should be designed for these two types of promoters.

Table 2. *Sensitivities and specificities of CNN model in predicting E. coli promoters.*

Evaluate	σ24	σ28	σ32	σ38	σ54	σ70
Sensitivity	0.77	0.70	0.74	0.51	0.58	0.75
Specificity	0.88	0.82	0.85	0.69	0.77	0.74

On the whole, the ROC curves also indicate that CNN model is a powerful approach in predicting promoters.

4. Conclusions

In this study, with the deep learning approach, a CNN model was constructed to predict the six types of Escherichia coli K-12 promoters. The CNN model achieves the best performance in prediction of E. coli promoters up to now. The study suggests that the CNN model is a powerful model in DNA sequence analysis, and it could be applied in more fields of DNA and protein sequence analysis.

Acknowledgments

This study was funded by the scientific research project of Beijing Municipal Commission of education, KM201610028010.

References

[1] He W, Jia C, Duan Y, et al. 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. [J] BMC Systems Biology, 2018, 12(4):44.

[2] Barrios H, Valderrama B, Morett E. Compilation and analysis of sigma(54)-dependent promoter sequences. [J] Nucleic Acids Research, 1999, 27(22):4305-4313.

[3] Gershenzon N I, Stormo G D, Ioshikhes I P. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. [J] Nucleic Acids Research, 2005, 33(7):2290-301.

[4] Zhang L, Luo L. Splice site prediction with quadratic discriminant analysis using diversity measure. [J] Nucleic Acids Research, 2003, 31(21):6214-6220.

[5] Drioli S, Felluga F, Forzato C, et al. The recognition and prediction of σ 70, promoters in Escherichia coli K-12. [J] Journal of Theoretical Biology, 2006, 242(1):135.

[6] Gordon J J, Towsey M W, Hogan J M, et al. Improved prediction of bacterial transcription start sites. [J] Bioinformatics, 2006, 22(2):142-148.

[7] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. [J] Neural Computation, 2006, 18(7):1527-1554.

[8] Tran N H, Zhang X, Xin L, et al. De novo peptide sequencing by deep learning. [J] Proceedings of the National Academy of Sciences of the United States of America, 2017:201705691.

[9] Yang B, Liu F, Ren C, et al. BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone. [J] Bioinformatics, 2017, 33(13).

[10] Hong H, Xiao-Chen B O, Fei L I. Application of Deep Learning in Biomedical Data. [J] Journal of Medical Informatics, 2018, 39(03):2-9.

[11] Bengio Y. Learning Deep Architectures for AI. [J] Foundations & Trends® in Machine Learning, 2009, 2(1):1-127.

[12] Serre T, Kreiman G, Kouh M, et al. A quantitative theory of immediate visual recognition. [J] Progress in Brain Research, 2007, 165(6):33-56.

[13] Zhou F Y, Jin L P, Dong J. Review of Convolutional Neural Network. [J] Chinese Journal of Computers, 2017, 40(06):1229-1251.

[14] Lecun Y L, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE. [J] Proceedings of the IEEE, 1998, 86(11):2278-2324.

[15] Lecun Y, Boser B, Denker J, et al. Backpropagation Applied to Handwritten Zip Code Recognition. [J] Neural Computation, 2014, 1(4):541-551.

[16] Gao L, Chen P Y, Yu S. Demonstration of Convolution Kernel Operation on Resistive Cross-Point Array. [J] IEEE Electron Device Letters, 2016, 37(7):870-873.

[17] Boureau Y L, Ponce J, Lecun Y. A Theoretical Analysis of Feature Pooling in Visual Recognition. International Conference on Machine Learning. DBLP, 2010:111-118.

[18] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. [J] 2015:448-456.

[19] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. [J] Journal of Machine Learning Research, 2014, 15(1):1929-1958.

[20] Zhou X, Li Z, Dai Z, et al. Predicting promoters by pseudo-trinucleotide compositions based on discrete wavelets transform. [J] Journal of Theoretical Biology, 2013, 319(5):1-7.

[21] Yan Y, Wan P. Prediction of Escherichia coli K-12 promoters using position-specific scoring matrix (PSSM) method. [J] Chinese Journal of Bioinformatics, 2015, 13(02):125-130.

[22] De A E S S, Echeverrigaray S, Gerhardt G J. BacPP: bacterial promoter prediction--a tool for accurate sigma-factor specific assignment in enterobacteria. [J] Journal of Theoretical Biology, 2011, 287(1):92.