Health Informatics Vision: From Data via Information to Knowledge
J. Mantas et al. (Eds.)
IOS Press, 2019

doi:10.3233/SHTI190061

DeepCNPP: Deep Learning Architecture to Distinguish the Promoter of Human Long Non-Coding RNA Genes and Protein-Coding Genes

Tanvir ALAM[a,1], Mohammad Tariqul ISLAM[b], Mowafa HOUSEH[a], Samir Brahim BELHAOUARI[a] and Ferdaus Ahmed KAWSAR[c]
[a] *Information and Computing Technology Division, College of Science and Engineering, Hamad Bin Khalifa University (HBKU), Doha, Qatar*
[b] *Computer Science Department, Southern Connecticut State University, USA*
[c] *Department of Computing, East Tennessee State University, USA*

Abstract. Promoter region of protein-coding genes are gradually being well understood, yet no comparable studies exist for the promoter of long non-coding RNA (lncRNA) genes which has emerged as a global potential regulator in multiple cellular process and different diseases for human. To understand the difference in the transcriptional regulation pattern of these genes, previously, we proposed a machine learning based model to classify the promoter of protein-coding genes and lncRNA genes. In this study, we are presenting DeepCNPP (deep coding non-coding promoter predictor), an improved model based on deep learning (DL) framework to classify the promoter of lncRNA genes and protein-coding genes. We used convolution neural network (CNN) based deep network to classify the promoter of these two broad categories of human genes. Our computational model, built upon the sequence information only, was able to classify these two groups of promoters from human at a rate of 83.34% accuracy and outperformed the existing model. Further analysis and interpretation of the output from DeepCNPP architecture will enable us to understand the difference in transcription regulatory pattern for these two groups of genes.

Keywords. deep learning, convolution neural network, long non-coding RNA, promoter

1. Introduction

Although the human genome project [1] primarily focused on the protein-coding genes exclusively, long non-coding RNA (lncRNA) genes [2], which do not encode a protein, later, emerged as a potential global regulator for different cellular processes and have shown to be involved in different diseases including cancer [3, 4]. Given the diversity in biogenesis for lncRNAs, their low-level expression and conservation make them more cryptic than protein-coding genes. Therefore, it becomes more challenging to understand their regulation and functional relevance in different pathways and diseases [4].

[1] Corresponding Author, Tanvir Alam, Information and Computing Technology Division, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar; E-mail: talam@hbku.edu.qa.

To better understand the differences between these two groups of genes, many computational methods have been developed to distinguish the non-coding regions from the coding regions of the genome across multiple species [5]. But very few studies focused on comparing the difference in their promoter regions, which is considered as the key regulatory region for genes. In order to understand the difference in global transcriptional regulatory pattern of protein-coding genes and lncRNA genes, we need to investigate their promoter regions thoroughly. To achieve this goal, previously, we developed a traditional machine-learning models to classify these two groups of promoters using genomics, epigenomics and regulatory information in the promoter regions [6]. We used different sequence-related features (k-mer, palindromes, skewness etc.), chromatic states [7] and the putative transcription factor binding sites (TFBSs) [8] that could bind in the promoter region of these two groups of genes to elucidate their differences in transcription regulation. Though gene expression, transcription factors (TFs) controlling the genes and epigenetic information are crucial to understand the transcription regulatory network of genes, yet there is a scarcity of such information. For example, we do not have gene expression data for all cells/tissues in human; we do not have ChIP-seq data for all known TFs across all cell types in human. So, it is always advantageous to build a computational model using sequence information only so that it can work independent of other available factors. Here we interrogated, at genome-wide scale, to test the hypothesis that only DNA sequence information of the promoter region is well enough to elucidate the underlying pattern of promoter of lncRNA genes from protein-coding genes. To reach in conclusion, we set up this problem as a machine learning classification framework for classifying two broad groups of gene promoters using sequence information only.

Recently convolutional neural network (CNN) has shown to achieve ground-breaking results in the classification of images [9]. The examples of using CNN in biological problems are increasing in recent time as well [10]. Though we are the first to introduce the machine learning model to distinguish the promoter of lncRNA genes and protein-coding genes, no deep learning (DL) based system has been developed for the classification of promoters from human lncRNA genes and protein-coding genes. Therefore, it is unknown whether DL based architecture can achieve reasonable accuracy in classifying the promoter of lncRNA genes and protein-coding genes. So, the objective of this study is to build a DL based architecture to check the effectiveness of such a network structure in this particular problem. Hence, we introduce DeepCNPP (deep coding noncoding promoter predictor), the first DL based architecture to classify the promoter of these two broad groups of genes using the sequence information of the promoter region only. DeepCNPP outperformed the existing model [6] considering all evaluation metrics.

2. Methods

We downloaded the publicly available promoter dataset from our previous study [6]. The dataset contains promoter information of 18,787 protein-coding genes and 18,487 lncRNA genes. We considered the [-1000, +1000] region of transcription start sites (TSS) as the putative promoter region of genes as prescribed in [6]. The promoter sequence was fetched from the human genome (hg19 version). Each nucleotide of promoter sequence was encoded using one-hot encoding approach of four length vector, A:(1,0,0,0),

C:(0,1,0,0), G:(0,0,1,0) and T:(0,0,0,1). The two-dimensional encoded promoter sequences were used as input to build DeepCNPP architecture (Figure 1).

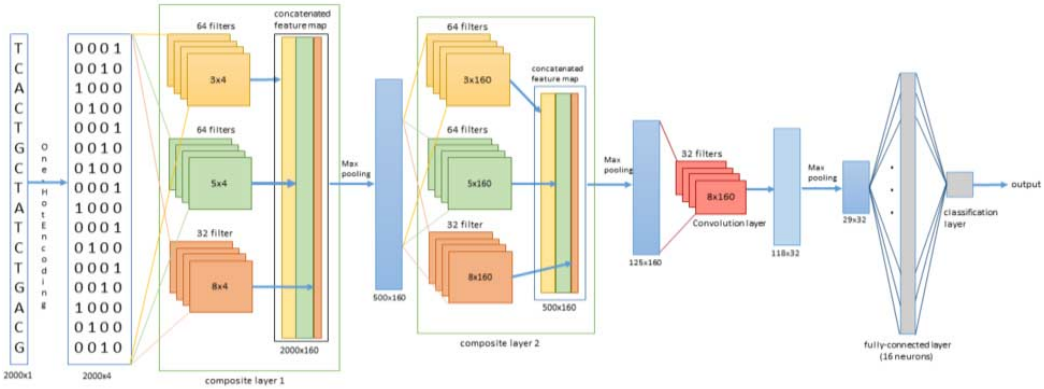

Figure 1. Proposed architecture for DeepCNPP.

We used one-dimensional convolution operations, commonly used for sequence data analysis, in our model. Keeping consistency with the convolution layers the max-pooling operation used were also one-dimensional, having size of 4. DeepCNPP consists of two inception-like [11] layers (hereafter referred to as composite layers) followed by one conventional convolutional layer, and finally, two fully connected layers. Each composite layer is a cascade of three convolutional layers with filter size 3, 5, and 8, respectively. We used 64 kernels of size 3 and 5, and 32 of size 8. The outputs of these convolution layers are then concatenated along the feature axis to produce a feature map that acts as the input to the next layer. Using convolutions of multiple sizes in one layer helps the network analyze the layer's input at different scales, and produces a feature map for the next layer incorporating information from different visual levels. DeepCNPP has two of these composite layers one after another with dropout and max pooling layers following each. After these two composite layers, the network has a regular convolution layer containing 32 kernels of size 8 with dropout and max-pooling layers following. The last two layers of the network are fully connected layers of size 16 and 1 (the final classification layer for coding promoter (0) and non-coding promoter (1) prediction), respectively. The first fully-connected layer has a dropout layer after it for regularization. We used ReLU as the activation function for the inner layers, and sigmoid for the classification layer. The dropout rate for the composite and the convolution layers was 0.4, and for the fully connected layer, 0.5.

We trained our model using the Adam [12] optimization algorithm using a minibatch size of 256. We used the default values for the β1 and β2 parameters for Adam, and used the stochastic gradient descent with warm restart [13] as the learning rate scheduler with a minimum and maximum learning rate of 3x10-5 and 10-3, respectively. We used Keras for implementing the DeepCNPP. The model was trained on GeForce GTX TitanX (Pascal) on single GPU machine for 400 epochs. Each epoch took around 40 seconds to complete. We used 10-fold cross validation to evaluate the performance of our model.

3. Results & Discussions

In the training process of DeepCNPP, we considered the promoter regions of lncRNA genes and protein-coding genes as positive and negative class respectively. We used the following metrics to evaluate the performance of the model: Sensitivity= (TP)/(TP+FN), Specificity = (TN)/(FP+TN), Accuracy = (TP+TN)/(TP+FN+FP+TN), where TP, FN, FP, TN stand for true positive, false negative, false positive and true negative respectively. Using 10 fold cross-validation, we achieved 83.88% sensitivity, 82.81% specificity, and 83.34% accuracy (see Table 1). From Table 1 we can notice that DeepCNPP outperformed the existing model in all evaluation metrics.

Table 1. Summary of model performance for DeepCNPP and the previous model.

Model	Sensitivity (%)	Specificity (%)	Accuracy (%)
Previous model [6]	82.77	80.60	81.69
DeepCNPP	83.88	82.81	83.34

4. Conclusion

We developed DeepCNPP, the first deep CNN based architecture to classify the promoter of lncRNA genes and protein-coding genes from human and it outperformed the existing model in all evaluation metrics. In future, we will investigate the output from different filters at different layers of CNN to interpret the model. The interpretation of the model will help us to better understand the transcription regulatory pattern of these two groups of genes. In addition to improving the proposed methods, we will extend this for other model organisms.

References

[1] E.S. Lander et al., Initial sequencing and analysis of the human genome, *Nature* **409 (6822)** (2001), 860-921.
[2] T. Derrien et al., The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression, *Genome Research* **22 (9)** (2012), 1775-1789.
[3] T. Alam et al., FARNA: knowledgebase of inferred functions of non-coding RNA transcripts, *Nucleic Acids Res* **45 (5)** (2017), 2838-2848.
[4] C.C. Hon et al., An atlas of human long non-coding RNAs with accurate 5' ends, *Nature* **543 (7644)** (2017), 199-204.
[5] G. Wang et al., Characterization and identification of long non-coding RNAs based on feature relationship, *Bioinformatics*, 2019.
[6] T. Alam et al., Promoter analysis reveals globally differential regulation of human long non-coding RNA and protein-coding genes, *PLoS One* **9 (10)** (2014), e109443.
[7] J. Ernst et al., Mapping and analysis of chromatin state dynamics in nine human cell types, *Nature* **473 (7345)** (2011), 43-49.
[8] I.V. Kulakovskiy et al., HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models, *Nucleic Acids Res* **44 (D1)** (2015), D116-125.
[9] A. Krizhevsky, I. Sutskever and G.E. Hinton, Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, 2012.
[10] T. Ching et al., Opportunities and obstacles for deep learning in biology and medicine, *J R Soc Interface* **15 (141)** (2018).
[11] C. Szegedy et al., Going deeper with convolutions, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.
[12] D.P. Kingma and J.L. Ba, *ADAM: A method for stochastic optimization*.
[13] I. Loshchilov and F. Hutter, SGDR: Stochastic gradient descent with warm restarts, *Learning* **10**, 3.