Improved Predicting of The Sequence Specificities of RNA Binding Proteins by Deep Learning

Hilal Tayara and Kil To Chong

Abstract—RNA-binding proteins (RBPs) have a significant role in various regulatory tasks. However, the mechanism by which RBPs identify the subsequence target RNAs is still not clear. In recent years, several machine and deep learning-based computational models have been proposed for understanding the binding preferences of RBPs. These methods required integrating multiple features with raw RNA sequences such as secondary structure and their performances can be further improved. In this paper, we propose an efficient and simple convolution neural network, RBPCNN, that relies on the combination of the raw RNA sequence and evolutionary information. We show that conservation scores (evolutionary information) for the RNA sequences can significantly improve the overall performance of the proposed predictor. In addition, the automatic extraction of the binding sequence motifs can enhance our understanding of the binding specificities of RBPs. The experimental results show that RBPCNN outperforms significantly the current state-of-the-art methods. More specifically, the average area under the receiver operator curve was improved by 2.67 percent and the mean average precision was improved by 8.03 percent. The datasets and results can be downloaded from https://home.jbnu.ac.kr/NSCL/RBPCNN.htm

Index Terms—Convolution neural network, deep learning, evolutionary information, RNA-binding protein, sequence motifs

1 Introduction

RNA binding site or binding motif is a subsequence of RNA where the binding between the RBP and its RNA subsequence targets take place. Thus, identifying these binding sites helps for a better understanding of the processes of the post-transcriptional modification. RNA-binding proteins (RBPs) are extremely engaged in different regulatory mechanisms, such as gene splicing and localization, and providing significant functional data for patient healthcare [1]. It has been observed that RBPs have a key function in several important biological processes [1], [2], [3], [4], [5], [6], [7], [8]. Therefore, the search for binding RBP sites is a vital study objective.

Different high-throughput RBPs detection technologies have been introduced such as CLIP-Seq [9], PAR-CLIP [10], RIP-Seq [11], but they still require long processing time and high cost. However, these technologies provide the key bases data for developing powerful computational models [12], [13], [14], [15], [16], [17]. Thus, researchers have put a huge effort into the development of accurate, low-cost, and fast computational models for identifying RBPs sites. For instance, Livi and Blanzieri [18] have proposed a method called Oli by which they extracted tetranucleotide features and used a support vector machine (SVM) as a classifier. Maticzka *et al.* proposed GraphProt model [15] by which they learned the features from RNA structure and sequence and fed them to SVM classifier. On the other hand, deep learning has shown an unprecedented performance in different domains such as image processing [19], [20], [21], [22], [23], text understanding [24], speech recognition [25], [26], [27], and genome analysis [28], [29], [30], [31], [32], [33]. For instance, the DeepBind model [13] was proposed to study the DNA and RNA specificities from large datasets. DeepSEA and DanQ models were proposed to study to effects of non-coding variants [34], [35]. The DeepCpG model was proposed to study CpG sites [36]. All of these successful examples have proven that deep learning can effectively extract the features automatically from raw genomic sequences and provide better outcomes in terms of prediction and analysis. Moreover, deep learning-based models can deal with large scale datasets better than conventional methods. Also, a moderate noise level and misleading training data can be tolerated efficiently using deep learning. Therefore, different deep learning-based models have been proposed for RNA protein binding sites prediction. For instance, recently, Pan *et al.* [37] proposed iDeepS model by which they integrated RNA secondary structure with raw RNA sequences. This model combined a convolution neural network with bidirectional long short-term memory (BLSTM). Shen *et al.* [38] proposed MSCGRU model by which they integrated a multi-scale convolution neural network with gated recurrent network GRU.

- *H. Tayara is with the Department of Electronics and Information Engineering, Chonbuk National University, Jeonju 54896, South Korea. E-mail: hilaltayara@jbnu.ac.kr.*
- *K. T. Chong is with the Advanced Electronics and Information Research Center, Chonbuk National University, Jeonju 54896, South Korea. E-mail: kitchong@jbnu.ac.kr.*

Manuscript received 23 Nov. 2019; revised 17 Feb. 2020; accepted 8 Mar. 2020. Date of publication 18 Mar. 2020; date of current version 8 Dec. 2021. (Corresponding author: Kil To Chong.)
Digital Object Identifier no. 10.1109/TCBB.2020.2981335

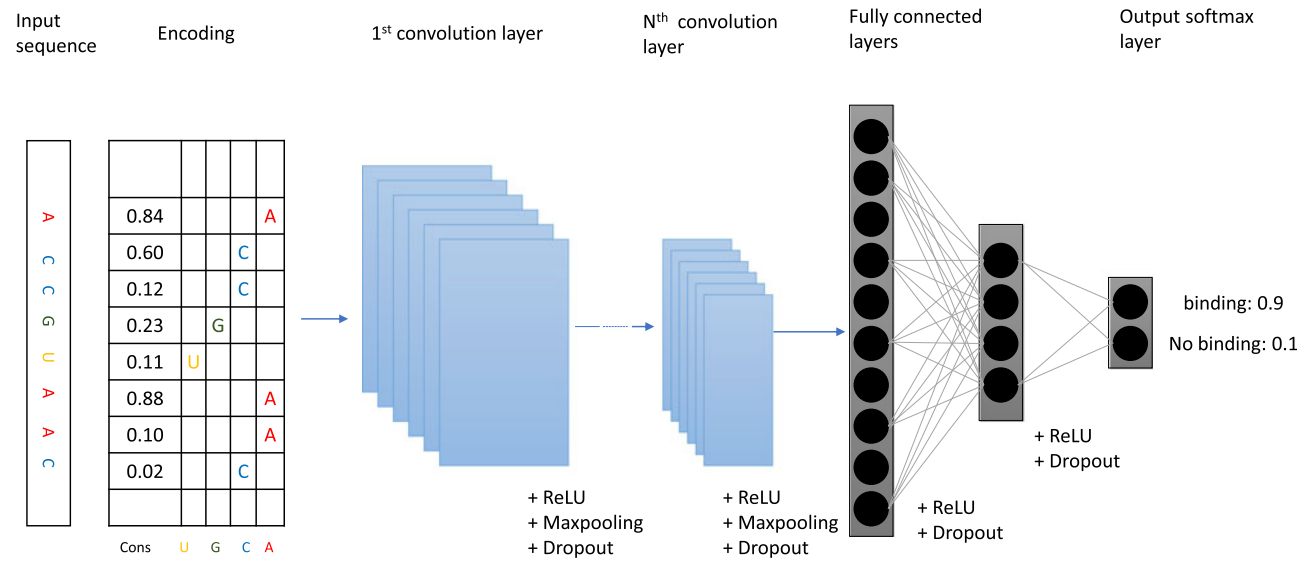

Fig. 1. An overview of the RBPCNN model. A raw RNA sequence is first encoded into a one-hot matrix and conservation scores. The first convolution layer searches the motifs in the encoded input sequence. The subsequent convolution layers discover the interactions between the learned motifs of the first convolution layer. The learned features from the convolution layers go through fully connected layers with a softmax layer at the output for prediction.

High order nucleotide encoding was used as input to the CNN model in [39] for predicting binding sites in lncRNA chains. RNA binding sites in circRNAs has been studied in [40] using word embedding technique as input to CNN and RNN model followed by the conditional random field. Pan and Shen introduced the combination of local and global CNN models for the identification of RNA binding sites [41]. A Multi-scale CNN model was proposed for identifying cancer-specific circRNA Binding sites [42] and RNA binding sites [43]. A hybrid deep learning model was proposed in [44] for RNA binding sites prediction using a codon encoding method. Different representations such as motifs and RNA structures were used in deep belief network and CNN models for generating shared representation. This representation was used in the classifier for prediction RNA binding sites [45]. The positions and the intensities of the CLIP-seq peaks were used for predicting binding sites of functional mRNA targets [46]. Ghanbari and Ohler proposed a deep neural network in which they integrated region type of binding sites and raw RNA sequence [47]. Capsule network using hybrid features was also proposed for finding RNA binding sites by [48]. Furthermore, various deep learning models have been proposed for studying DNA bindings using different features such as [49], [50], [51].

Various studies showed that transcription factor binding sites are conserved among species [52], [53], [54], [55], [56]. For example, Rosanova *et al.* showed that transcription factor binding sites in higher eukaryotes are conserved over the last 600 million years [57]. Another study showed that gene expression can be predicted from the conservation of the transcription factor binding sites [58]. Therefore, in this paper, we study the importance of conservation information to improve the performance of the RNA transcription factor binding site predictors. We propose a simple and efficient convolutional neural network for RNA protein binding sites prediction Fig. 1. The input combines the raw RNA sequences with evolutionary information. We show that this evolutionary information helps in achieving outstanding results compared with the state-of-the-art models. We call our model RBPCNN. The proposed architecture is simpler than the state-of-the-art models that required using BLSTM or GRU. In addition, RBPCNN is able to learn the binding motifs and visualize the learned conservation scores.

The rest of the paper is organized as follows. Section 2 introduces the benchmark datasets, the RNA sequence encoding strategy, and the proposed model. Section 3 presents the performance of the proposed model and the comparison results with competing methods. Section 4 concludes the paper.

2 Material and Methods

2.1 Materials

In this paper, we utilized 31 RBP datasets obtained from iDeepS paper [37]. These datasets were originally obtained from DoRiNA [59] and iCount website (http://icount.biolab.si/). Each nucleotide in the cluster of the interaction sites was treated as binding sites and thus, from which the positive datasets were constructed. On the other hand, the negative datasets were constructed by sampling from RNA sequences that were not identified as binding sites. Each experiment has 30,000 samples for training and hyper-parameter optimization and 10,000 samples for testing the performance of the trained model. The training datasets contain 6,000 positive sequences and 24,000 negative ones while the testing datasets contain 2,000 positive sequences and 8,000 negative ones. For a fair comparison with other models, we have used the same configurations of datasets preparations.

Three experimental protocols namely iCLIP, PAR-CLIP, and CLIP-SEQ/HITS-CLIP were used to prepare the datasets as shown in Table 1. iCLIP stands for individual-nucleotide resolution Cross-Linking and ImmunoPrecipitation and it is used for detecting the protein-RNA interactions by utilizing the ultraviolet light to help in binding the RNA molecules

TABLE 1
An Overview of RBP Datasets Used in This Study

Experimental protocol	Protein
iCLIP	TIAL1, TIA1 [65] TDP-43 [66] Nsun2 [67] hnRNPL, hnRNPL-like [68] hnRNPC [69] U2AF2, hnRNPC [70]
PAR-CLIP	MOV10 [71] FUS, ESWR1,TAF15 [72]; ELAVL1, ELAVL1A, ELAVL1-MNase, Ago2MNase [73]; IGF2BP1-3, Ago/EIF2C1-4, PUM2 [3];
CLIP-SEQ/HITS-CLIP	Ago2 [74] SRSF1 [75]; eIF4AIII [76]; Ago2, ELAVL1 [73];

and proteins [60]. PAR-CLIP stands for photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation and it works by integrating nascent RNA with photoactivatable nucleoside [3]. This technique has been used to study transcriptome-wide binding sites of different RNA-binding proteins RBPs [61], [62], [63]. HITS-CLIP stands for High-throughput sequencing of RNA combined with crosslinking immunoprecipitation and it uses ultraviolet crosslinking with next-generation sequencing [9], [64].

2.2 Encoding Sequence and Evolutionary Information

Each input RNA sequence $S = (s_1, s_2, , s_n)$ was one-hot encoded. Thus, A, C, G, U, and N were encoded as (1000), (0100), (0010), (0001), and (0000) respectively. The length of the input sequence is $n = 101$ nt. In addition to one-hot encoding, we added conservation (evolutionary) information of each nucleotide of the input sequence. The evolutionary information was obtained from (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phylo P100way/) where we used the conservation scores of multiple alignments of 99 vertebrate genomes to the human genome. These scores were obtained from the PHAST package (http://compgen.bscb.cornell.edu/phast/). The values of these scores were scaled to [0-1]. Thus, each input sequence S with n nucleotides is encoded as $n \times 5$ such as four channels for one-hot encoding and the last channel for conservation scores.

2.3 The Proposed Model

We propose a simple and efficient deep learning model based on the convolution neural network (CNN) [77] for the identification of RNA protein binding sites as shown in Fig. 1. It is called RBPCNN. Unlike other models that require BLSTM [78] or GRU [79] we only use a well-optimized convolution neural network that outperforms the state-of-the-art models. The most important hyper-parameters were selected using the grid search algorithm. The ranges of the tuned hyper-parameters are shown in Table 2.

The length of the transcription factor binding sites in eukaryotes ranges from 5nt to 30nt as reported by Stewart *et al.* [80]. Therefore, the input length of the proposed models is set to 101nt. Each sequence is centered on the transcription factor binding site and the additional nucleotides were used for providing contextual information. We followed the same configurations for the input length of other studies we compare with for the fair comparison. The RBPCNN is composed of convolution layers and fully connected layers. Each convolution layer is followed by a ReLU activation layer, a max-pooling layer with a pool size of 2 and a stride of 2, and a drop out layer with a dropout rate of 0.3 [81]. The fully connected layers are followed by a ReLU activation function and a dropout layer with a dropout rate of 0.5. The last layer is a softmax layer that outputs the probability results of the classification task.

TABLE 2
The Ranges of the Tuned-Hyper Parameters

The hyper-parameters	The range
The number of the convolution layers	[1,2,3]
The number of the kernels	[32,64,128]
The kernel length	[5, 7, 9,11]
The number of the fully connected layers	[1,2]
The number of nodes in the dense layer	[16, 32]

The convolution layer is a one-dimensional convolution expressed in Eq(1) where I is the input, o and k are the indices of the output position and the kernels, respectively, and W^f is the weight matrix of $S \times N$ shape with S filters and N input channels.

$$\mathrm{Conv}(I)_{ok} = \mathrm{ReLU}\left(\sum_{s=0}^{S-1}\sum_{n=0}^{N-1} W_{sn}^{f} I_{o+s,n}\right). \tag{1}$$

The dense layer is expressed mathematically in Eq(2) where z_k is a one-dimension feature vector, The weights of the z_k from the previous layer is w_k, and the additive bias term is w_{d+1}.

$$f = w_{d+1} + \sum_{k=1}^{d} w_k z_k. \tag{2}$$

The dropout layer is added to switch off certain neurons at training time in order to reduce overfitting. Adding dropout after dense layer results in Eq. (3) where m_k is sampled form Bernoulli distribution.

$$f = w_{d+1} + \sum_{k=1}^{d} m_k w_k z_k. \tag{3}$$

The rectified linear unit activation function was used in this design and it is given in Eq(4) [82].

$$\mathrm{ReLU}(z) = max(0, z). \tag{4}$$

The final layer is the Softmax layer that normalizes its input vector z into a probability distribution having C probabilities proportional to the exponential of the input numbers.

$$\mathrm{Softmax}(z)_j = \frac{\exp_j^z}{\sum_{i=1}^{C} \exp^{z_i}}. \tag{5}$$

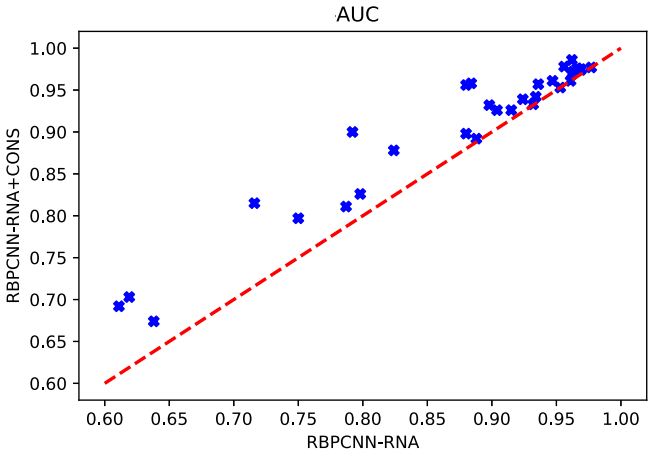

Fig. 2. A scatter plot comparing the achieved AUC of the proposed model RBPCNN using raw RNA sequences only and by integrating conservation scores to raw RNA sequences.

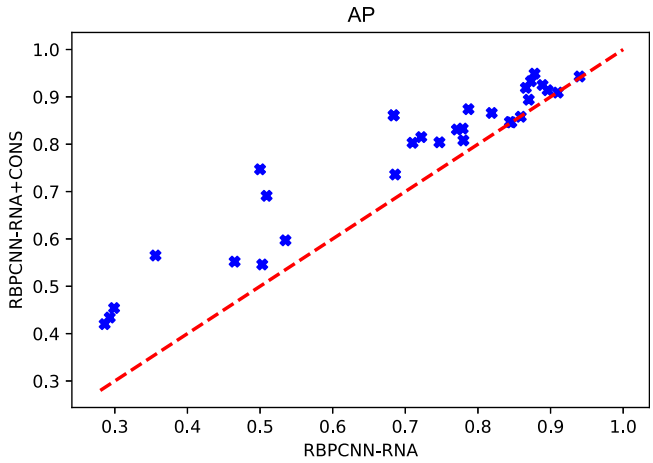

Fig. 3. A scatter plot comparing the achieved AP of the proposed model RBPCNN using raw RNA sequences only and by integrating conservation scores to raw RNA sequences.

The proposed models were constructed using Keras (https://keras.io/). The optimizer was Adam [83] with a learning rate of 5e-4. The number of epochs was set to 500 with early stopping based on validation loss. Kernels were initialized using a random uniform in the range [-0.05, 0.05]. Max norm weight constraint was applied with a value of 3 [81]. Categorical cross-entropy function was used for parameters update. The motifs and the conservation scores were visualized using TOMTOM [84] and pysster framework [85].

3 Results

In this section, we study the performance of the proposed model RBPCNN and compare it with the state-of-the-art-models. In addition, we visualize the learned motifs and conservation scores by the RBPCNN model.

3.1 Evaluation Metrics

In this paper, we used the area under the receiver operating characteristic curve (AUC) and the average precision score (AP). Since the datasets are imbalanced the AP is a more important metric for reflecting the real performance of the proposed model [86]. Scikit-learn package (https://scikit-learn.org) was used to compute these metrics.

3.2 The Importance of Evolution Information

In order to study the importance of adding evolutionary information, we trained the proposed model using raw RNA sequences only. For a fair comparison, we have searched the best hyper-parameters again in the case of using raw RNA sequences only using similar grid search parameters as shown in Table 2. The average AUC of using raw RNA sequence only was 87.40 percent while it was 90.44 percent after integrating the conservation scores. On the other hand, the mean AP of using raw RNA sequences only was 69.64 percent while it was 77.54 percent after integrating the conservation scores. Thus, adding conservation scores to the raw RNA sequences improved the performance by 3.04 and 7.90 percent in terms of AUC and AP, respectively. The Figs. 2 and 3 show that AUC and AP scores of all 31 experiments were improved by integrating the conservation scores with raw RNA sequences.

3.3 Competing Methods

We compare the proposed model RBPCNN with the following methods:

3.3.1 *Oli [18]*

They designed their features based on tetranucleotide frequency of the RNA sequence. The SVM was used for classification.

3.3.2 *GraphProt [15]*

They designed their features from the RNA sequence and the secondary structure and then passed them to the SVM classifier.

3.3.3 *DeepBind*

This is a deep learning-based model in which the authors predicted the binding sites using the raw RNA sequences only.

3.3.4 *iDeepS [37]*

This is also a deep learning-based model in which the authors predicted the binding sites using the integration of the raw RNA sequences and the secondary structure.

3.3.5 *MSCGRU [38]*

This is another deep learning model in which authors designed a multi-scale convolution neural network with the gated recurrent neural network.

The comparison results show that the proposed model outperformed the aforementioned methods significantly in terms of AP and AUC as shown in Table 3. The average AUC was improved by 2.67 percent and the average AP was improved by 8.03 percent. The comparison results are shown graphically in Figs. 4 and 5. In more detail, the average AUC of RBPCNN outperformed Oli, GraphPort, DeepBind, iDeepS, MSCGRU by 13.03, 9.69, 7.34, 4.10, and 2.67 percent, respectively. On the other hand, the average AP of RBPCNN outperformed Oli, GraphPort, DeepBind,

TABLE 3
The Comparison of the Performance of RBPCNN With Other State-of-the-Art Methods

Method	AUC	AP
Oli	0.7741	0.5174
GraphPort	0.8075	0.5295
DeepBind	0.8310	0.5893
iDeepS	0.8634	0.6725
MSCGRU	0.8777	0.6951
RBPCNN	**0.9044**	**0.7754**

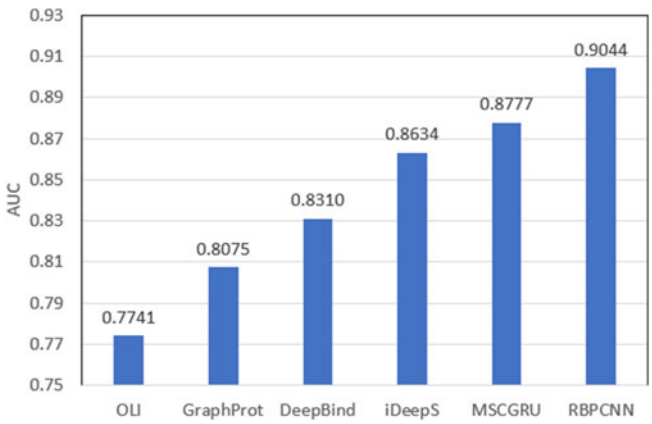

Fig. 4. The comparison of the performance of RBPCNN with other state-of-the-art methods in term of average AUC.

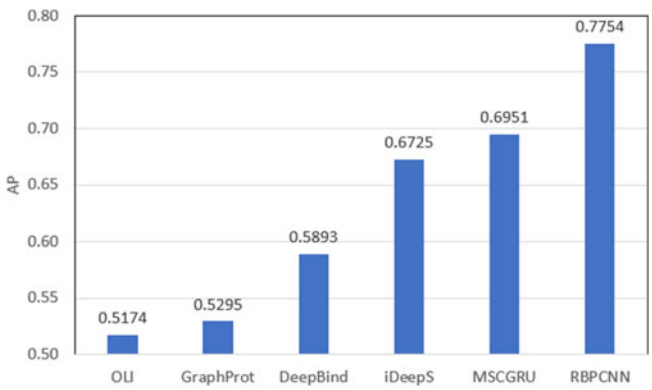

Fig. 5. The comparison of the performance of RBPCNN with other state-of-the-art methods in term of average AP.

iDeepS, MSCGRU by 25.80, 24.59, 18.61, 10.29, 8.03 percent, respectively. Since the dataset is imbalanced the AP is a more important metric to consider. Thus, it is clear that the performance of the proposed model is better than the other state-of-the-art models with a big margin.

The box plots for the 31 experiments for the proposed model RBPCNN and the competing methods are shown in Fig. 6 for AUC and Fig. 7 for AP. These results indicate that the proposed model RBPCNN outperforms the competing methods in almost all 31 experiments in terms of AUC and AP.

The detailed results of the proposed RBPCNN model and other competing methods for the 31 experiments are shown in (Table I supplementary file, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2020.2981333.) for AUC and (Table II supplementary file, available online) for AP. From these tables, we

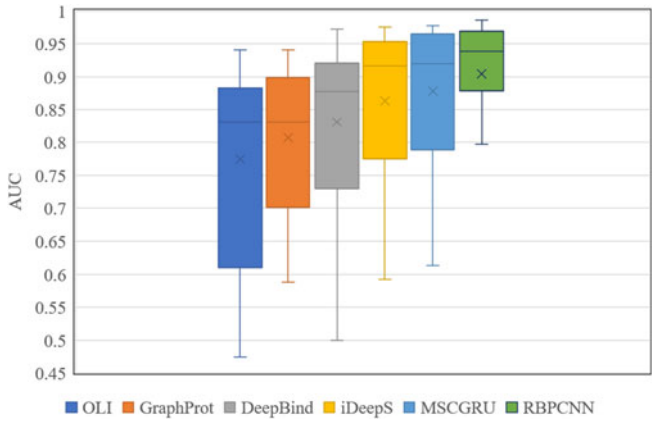

Fig. 6. The box plot of the performance of RBPCNN and other state-of-the-art methods in term of AUC.

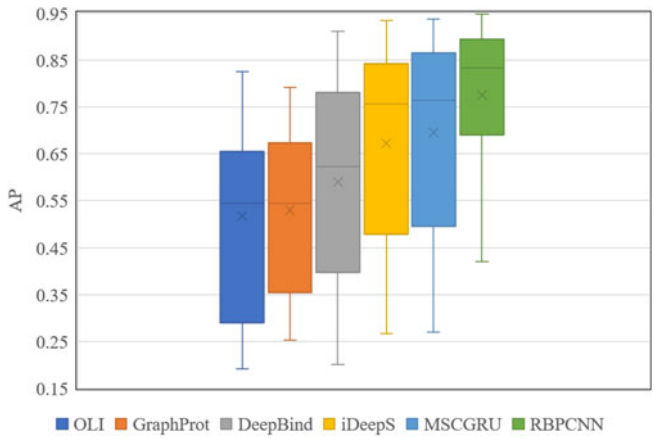

Fig. 7. The box plot of the performance of RBPCNN and other state-of-the-art methods in term of AP.

can see that adding conservation scores improved the performance significantly for almost all experiments. In more detail, and by looking at achieved AP results in (Table II supplementary file, available online) , we can see that the Ago proteins family were improved remarkably by 12.75 $\sim$ 28.36 percent by using conservation scores. The eIF4AIII-1 and eIF4AIII-2 were improved by 4.31 and 7.27 percent, respectively. All ELAVL1 protein family were improved and more remarkably ELAVL1-MNase which was improved by 15.39 percent. In general, the AP of all proteins in this study were improved except hnRNPC-1 and hnRNPC-2 did not show improvement by using the conservation scores.

3.4 The Learned Motifs

The most important advantage of the proposed model over other machine learning-based ones -such as Oli and GraphPort- is its ability to visualize the learned motifs easily. The machine learning-based methods require complex post-processing steps however, the learned convolution filters of the proposed model can be easily converted to position weight matrices (PWMs) (the motifs and conservation scores visualization are explained in the supplementary file, available online). Motifs visualization help in obtaining biological insights into the results. Fig. 8 shows examples of the detected sequence motifs. The full list of the detected motifs can be downloaded from https://home.jbnu.ac.kr/NSCL/

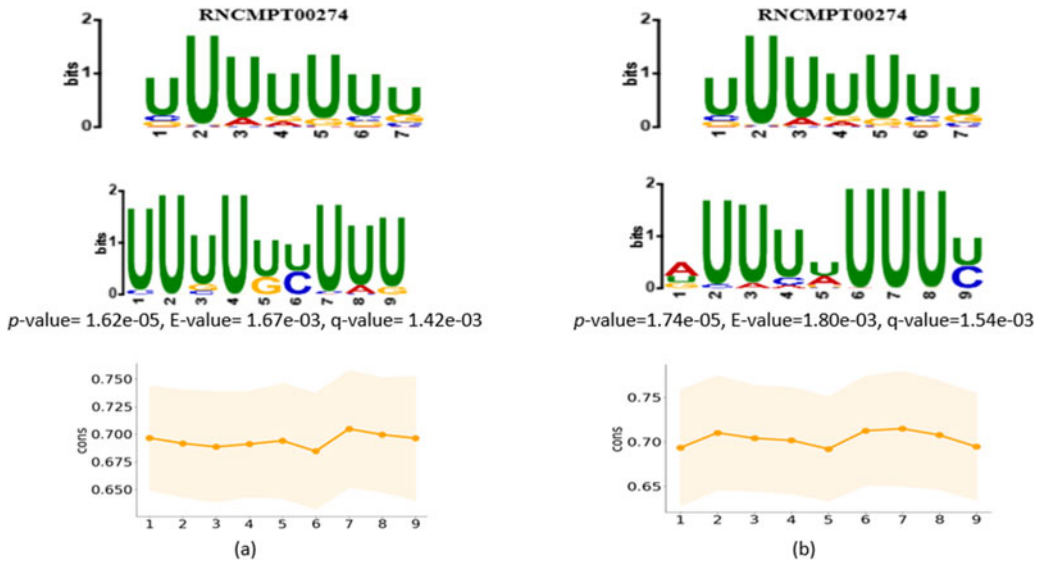

Fig. 8. Examples of the learned motifs by RBPCNN (a) ELAVL1-2 protein and (b) ELAVL1-1 protein. The first row represents the reported motifs by the CISBP-RNA database. The second row represents the learned motifs by our model. The third row shows the average conservation scores.

RBPCNN.htm. The proposed model was able to detect already reported motifs in CISBP-RNA database [87]. Thus, the RBPCNN has the potential to discover new sequence motifs. The predicted motifs have been compared with the known ones reported by the CISBP-RNA database using the TOMTOM tool [84] with significant E-value cutoff 0.10.

When we inspected the visualized conservation scores for all the kernels in all 31 experiments we found that when average maximum activation of the binding class is greater than the average maximum activation of the no binding class, the average conservation scores of the kernel is usually more than 70 percent. Fig. 9 shows examples of some of the learned kernels in the Ago2-1 protein experiment. Figs. 9a and 9b show the learned kernel 33 and kernel 58 where the average maximum activation of the binding class is higher than the average maximum activation of the no binding class. Thus, the average conservation scores for these two kernels is higher than 70 percent. On the other hand, Figs. 9c and 9d show the learned kernel 4 and kernel 66 where the average maximum activation of no binding class is higher than the average max activation of the binding class. Thus, the average conservation scores for these two kernels is less than 70 percent. These patterns were observed in almost all kernels of all experiments. These results show that the nucleotides in the binding sites are highly conserved which agrees with the biological studies that show the transcription factor binding sites are highly conserved [52], [53], [54], [55], [56]. The full list of the learned conservation scores by the proposed model can be downloaded from https://home.jbnu.ac.kr/NSCL/RBPCNN.htm.

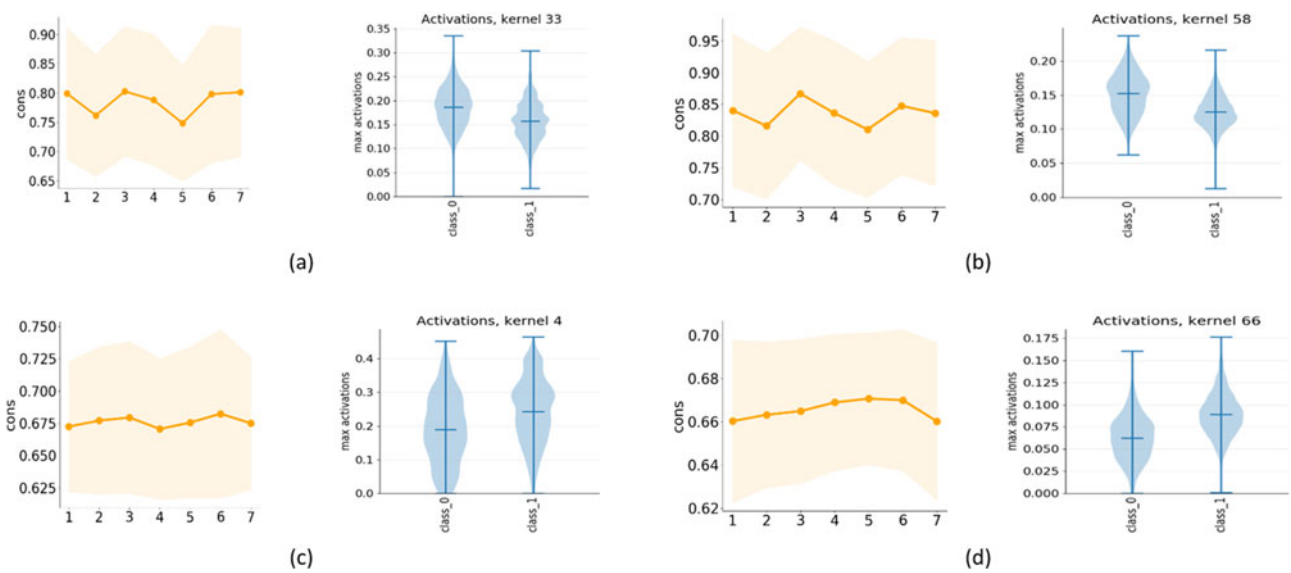

Fig. 9. Examples of the learned conservation patterns in the Ago2-1 protein experiment. The (a) and (b) show examples of the kernels that have the average maximum activation of binding class-class_0- greater than the average maximum activation of no-binding class -class_1- and the corresponding conservation scores patterns. The (c) and (d) show examples of the kernels that have the average maximum activation of no-binding class-class_1- greater than the average maximum activation of binding class -class_0- and the corresponding conservation scores patterns.

4 Conclusion

Accurate identification of RNA transcription factor binding sites is a very important step for a better understanding of different biological tasks. In this paper, we have introduced a simple and efficient deep learning model that integrates the evolutionary information with raw RNA sequences. This integration helped in achieving outstanding results compared with state-of-the-art methods. In addition, we have visualized the learned motifs by the RBPCNN model and matched them with already reported ones in the CISBP-RNA database. Moreover, we have visualized the average conservation scores learned by the deep learning kernels. The datasets and results can be downloaded from https://home.jbnu.ac.kr/NSCL/RBPCNN.htm

Acknowledgments

This research was supported by the Brain Research Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. NRF-2017M3C7A1044815).

References

[1] F. Ferre, A. Colantoni, and M. Helmer-Citterich, "Revealing protein–lncrna interaction," *Brief. Bioinformatics*, vol. 17, no. 1, pp. 106–116, 2015.
[2] R. Bandziulis, M. Swanson, and G. Dreyfuss, "RNA-binding proteins as developmental regulators," *Genes Dev.*, vol. 3, no. 4, pp. 431–437, 1989.
[3] M. Hafner *et al.*, "Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip," *Cell*, vol. 141, no. 1, pp. 129–141, 2010.
[4] M. Haneklaus, J. D. O'Neil, A. R. Clark, S. L. Masters, and L. A. O'Neill, "The RNA-binding protein tristetraprolin (TTP) is a critical negative regulator of the NLRP3 inflammasome," *J. Biol. Chem.*, vol. 292, no. 17, pp. 6869–6881, 2017.
[5] E. L. Van Nostrand *et al.*, "Robust transcriptome-wide discovery of rna-binding protein binding sites with enhanced clip (eCLIP)," *Nature Methods*, vol. 13, no. 6, 2016, Art. no. 508.
[6] S.-P. Deng and D.-S. Huang, "SFAPS: An R package for structure/function analysis of protein sequences based on informational spectrum method," *Methods*, vol. 69, no. 3, pp. 207–212, 2014.
[7] S.-P. Deng, L. Zhu, and D.-S. Huang, "Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks," *BMC Genomics*, vol. 16, no. 3, 2015, Art. no. S4.
[8] C.-H. Zheng, D.-S. Huang, L. Zhang, and X.-Z. Kong, "Tumor clustering using nonnegative matrix factorization with gene selection," *IEEE Trans. Inf. Technol. Biomedicine*, vol. 13, no. 4, pp. 599–607, Jul. 2009.
[9] R. B. Darnell, "HITS-CLIP: Panoramic views of protein–RNA regulation in living cells," *Wiley Interdisciplinary Reviews: RNA*, vol. 1, no. 2, pp. 266–286, 2010.
[10] M. Hafner *et al.*, "PAR-clip-a method to identify transcriptome-wide the binding sites of RNA binding proteins," *J. Vis. Exp.*, vol. 7, no. 41, 2010, Art. no. e2034.
[11] J. Zhao *et al.*, "Genome-wide identification of polycomb-associated RNAs by RIP-seq," *Mol. Cell*, vol. 40, no. 6, pp. 939–953, 2010.
[12] X. Pan, L. Zhu, Y.-X. Fan, and J. Yan, "Predicting protein–RNA interaction amino acids using random forest based on submodularity subset selection," *Comput. Biol. Chem.*, vol. 53, pp. 324–330, 2014.
[13] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of dna-and rna-binding proteins by deep learning," *Nature Biotechnol.*, vol. 33, no. 8, 2015, Art. no. 831.
[14] J. Yan, S. Friedrich, and L. Kurgan, "A comprehensive comparative review of sequence-based predictors of DNA-and RNA-binding residues," *Brief. Bioinformatics*, vol. 17, no. 1, pp. 88–105, 2015.
[15] D. Maticzka, S. J. Lange, F. Costa, and R. Backofen, "Graphprot: Modeling binding preferences of RNA-binding proteins," *Genome Biol.*, vol. 15, no. 1, 2014, Art. no. R17.
[16] M. Stražar, M. Žitnik, B. Zupan, J. Ule, and T. Curk, "Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins," *Bioinformatics*, vol. 32, no. 10, pp. 1527–1535, 2016.
[17] H. Tayara and K. T. Chong, "Improving the quantification of DNA sequences using evolutionary information based on deep learning," *Cells*, vol. 8, no. 12, 2019, Art. no. 1635.
[18] C. M. Livi and E. Blanzieri, "Protein-specific prediction of mRNA binding using rna sequences, binding motifs and predicted secondary structures," *BMC Bioinformatics*, vol. 15, no. 1, 2014, Art. no. 123.
[19] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.
[20] D.-S. Huang and J.-X. Du, "A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks," *IEEE Trans. Neural Netw.*, vol. 19, no. 12, pp. 2099–2115, Dec. 2008.
[21] D.-S. Huang, "Systematic theory of neural networks for pattern recognition," *Publishing House of Electronic Industry of China*, Beijing, 1996.
[22] H. Tayara and K. Chong, "Object detection in very high-resolution aerial images using one-stage densely connected feature pyramid network," *Sensors*, vol. 18, no. 10, 2018, Art. no. 3341.
[23] H. Tayara, K. G. Soo, and K. T. Chong, "Vehicle detection and counting in high-resolution aerial images using convolutional regression neural network," *IEEE Access*, vol. 6, pp. 2220–2230, 2018.
[24] D. Yogatama, C. Dyer, W. Ling, and P. Blunsom, "Generative and discriminative text classification with recurrent neural networks," 2017, *arXiv: 1703.01898*[stat.ML]. [Online]. Available: https://arxiv.org/abs/1703.01898
[25] D.-S. Huang and W. Jiang, "A general CPL-ads methodology for fixing dynamic parameters in dual environments," *IEEE Trans. Syst. Man Cybern. B, Cybern.*, vol. 42, no. 5, pp. 1489–1500, Oct. 2012.
[26] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Netw.*, vol. 92, pp. 60–68, 2017.
[27] D.-S. Huang, "Radial basis probabilistic neural networks: Model and application," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 13, no. 07, pp. 1083–1101, 1999.
[28] I. Nazari, H. Tayara, and K. T. Chong, "Branch point selection in RNA splicing using deep learning," *IEEE Access*, vol. 7, pp. 1800–1807, 2019.
[29] M. Oubounyt, Z. Louadi, H. Tayara, and K. T. Chong, "Deep learning models based on distributed feature representations for alternative splicing prediction," *IEEE Access*, vol. 6, pp. 58 826–58 834, 2018.
[30] M. Tahir, H. Tayara, and K. T. Chong, "iRNA-pseknc(2methyl): Identify RNA 2′-O-methylation sites by convolution neural network and chou's pseudo components," *J. Theor. Biol.*, vol. 465, pp. 1–6, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022519318306349
[31] H. Tayara, M. Tahir, and K. T. Chong, "iSS-CNN: Identifying splicing sites using convolution neural network," *Chemometrics Intell. Lab. Syst.*, vol. 188, pp. 63–69, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169743919300218
[32] Z. Louadi, M. Oubounyt, H. Tayara, and K. T. Chong, "Deep splicing code: Classifying alternative splicing events using deep learning," *Genes*, vol. 10, no. 8, 2019, Art. no. 587.
[33] M. Oubounyt, Z. Louadi, H. Tayara, and K. T. Chong, "Deepromoter: Robust promoter predictor using deep learning," *Frontiers in genetics*, vol. 10, 2019, Art. no. 286.
[34] D. Quang and X. Xie, "DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences," *Nucleic Acids Res.*, vol. 44, no. 11, pp. e107–e107, 2016.
[35] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning–based sequence model," *Nature Methods*, vol. 12, no. 10, 2015, Art. no. 931.
[36] C. Angermueller, H. J. Lee, W. Reik, and O. Stegle, "DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning," *Genome Biol.*, vol. 18, no. 1, 2017, Art. no. 67.
[37] X. Pan, P. Rijnbeek, J. Yan, and H.-B. Shen, "Prediction of rna-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks," *BMC Genomics*, vol. 19, no. 1, 2018, Art. no. 511.
[38] Z. Shen, S.-P. Deng, and D.-S. Huang, "RNA-protein binding sites prediction via multi scale convolutional gated recurrent unit networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, 10.1109/TCBB.2019.2910513.

[39] S.-W. Zhang, Y. Wang, X.-X. Zhang, and J.-Q. Wang, "Prediction of the RBP binding sites on lncRNAs using the high-order nucleotide encoding convolutional neural network," *Anal. Biochem.*, vol. 583, 2019, Art. no. 113364.
[40] Y. Ju, L. Yuan, Y. Yang, and H. Zhao, "CircSLNN: Identifying RBP-binding sites on circRNAs via sequence labeling neural networks," *Frontiers Genetics*, vol. 10, 2019, Art. no. 1184.
[41] X. Pan and H.-B. Shen, "Predicting RNA–protein binding sites and motifs through combining local and global deep convolutional neural networks," *Bioinformatics*, vol. 34, no. 20, pp. 3427–3436, 2018.
[42] Z. Wang, X. Lei, and F.-X. Wu, "Identifying cancer-specific circRNA–RBP binding sites based on deep learning," *Molecules*, vol. 24, no. 22, 2019, Art. no. 4035.
[43] T. Chung and D. Kim, "Prediction of binding property of RNA-binding proteins using multi-sized filters and multi-modal deep convolutional neural network," *PloS One*, vol. 14, no. 4, 2019, Art. no. e0216257.
[44] K. Zhang, X. Pan, Y. Yang, and H.-B. Shen, "CRIP: Predicting circRNA–RBP-binding sites using a codon-based encoding and hybrid deep neural networks," *Rna*, vol. 25, no. 12, pp. 1604–1615, 2019.
[45] X. Pan and H.-B. Shen, "RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach," *BMC Bioinformatics*, vol. 18, no. 1, 2017, Art. no. 136.
[46] J. Lin, Y. Zhang, W. N. Frankel, and Z. Ouyang, "Pras: Predicting functional targets of rna binding proteins based on clip-seq peaks," *PLoS Comput. Biol.*, vol. 15, no. 8, 2019, Art. no. e1007227.
[47] M. Ghanbari and U. Ohler, "Deep neural networks for interpreting rna-binding protein target preferences," *Genome Res.*, vol. 30, 2020, Art. no. gr–247 494.
[48] Z. Shen, S.-P. Deng, and D.-S. Huang, "Capsule network for predicting RNA-protein binding preferences using hybrid feature," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/TCBB.2019.2943465.
[49] Q. Zhang, L. Zhu, W. Bao, and D.-S. Huang, "High-order convolutional neural network architecture for predicting dna-protein binding sites," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 4, pp. 1184–1192, Jul./Aug. 2019.
[50] Q. Zhang, L. Zhu, W. Bao, and D.-S. Huang, "Weakly-supervised convolutional neural network architecture for predicting protein-DNA binding," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/TCBB.2018.2864203.
[51] Q. Zhang, Z. Shen, and D.-S. Huang, "Predicting in-vitro transcription factor binding sites using DNA sequence+ shape," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/TCBB.2019.2947461.
[52] D. Boffelli *et al.*, "Phylogenetic shadowing of primate sequences to find functional regions of the human genome," *Sci.*, vol. 299, no. 5611, pp. 1391–1394, 2003.
[53] D. Boffelli, M. A. Nobrega, and E. M. Rubin, "Comparative genomics at the vertebrate extremes," *Nature Rev. Genetics*, vol. 5, no. 6, pp. 456–465, 2004.
[54] A. M. McGuire, J. D. Hughes, and G. M. Church, "Conservation of dna regulatory motifs and discovery of new motifs in microbial genomes," *Genome Res.*, vol. 10, no. 6, pp. 744–757, 2000.
[55] H. Li, V. Rhodius, C. Gross, and E. D. Siggia, "Identification of the binding sites of regulatory proteins in bacterial genomes," *Proc. Nat. Acad. Sci.*, vol. 99, no. 18, pp. 11 772–11 777, 2002.
[56] A. Woolfe *et al.*, "Highly conserved non-coding sequences are associated with vertebrate development," *PLoS Biol.*, vol. 3, no. 1, 2004, Art. no. e7.
[57] A. Rosanova, A. Colliva, M. Osella, and M. Caselle, "Modelling the evolution of transcription factor binding preferences in complex eukaryotes," *Sci. Rep.*, vol. 7, no. 1, pp. 1–12, 2017.
[58] M. Hemberg and G. Kreiman, "Conservation of transcription factor binding events predicts gene expression across species," *Nucleic Acids Res.*, vol. 39, no. 16, pp. 7092–7102, 2011.
[59] K. Blin, C. Dieterich, R. Wurmus, N. Rajewsky, M. Landthaler, and A. Akalin, "Dorina 2.0 upgrading the dorina database of RNA interactions in post-transcriptional regulation," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D160–D167, 2015.
[60] J. König, K. Zarnack, N. M. Luscombe, and J. Ule, "Protein–rna interactions: new genomic technologies and perspectives," *Nature Rev. Genetics*, vol. 13, no. 2, 2012, Art. no. 77.
[61] J.-H. Yang, J.-H. Li, P. Shao, H. Zhou, Y.-Q. Chen, and L.-H. Qu, "starBase: A database for exploring microRNA–mRNA interaction maps from argonaute clip-seq and degradome-seq data," *Nucleic Acids Res.*, vol. 39, no. suppl_1, pp. D202–D209, 2010.
[62] R. L. Skalsky *et al.*, "The viral and cellular microRNA targetome in lymphoblastoid cell lines," *PLoS Pathogens*, vol. 8, no. 1, 2012, Art. no. e1002484.
[63] E. Gottwein *et al.*, "Viral microRNA targetome of KSHV-infected primary effusion lymphoma cell lines," *Cell Host Microbe*, vol. 10, no. 5, pp. 515–526, 2011.
[64] D. D. Licatalosi *et al.*, "Hits-clip yields genome-wide insights into brain alternative rna processing," *Nature*, vol. 456, no. 7221, 2008, Art. no. 464.
[65] Z. Wang *et al.*, "iCLIP predicts the dual splicing effects of TIA-RNA interactions," *PLoS Biol.*, vol. 8, no. 10, 2010, Art. no. e1000530.
[66] J. R. Tollervey *et al.*, "Characterizing the RNA targets and position-dependent splicing regulation by TDP-43," *Nature Neurosci.*, vol. 14, no. 4, 2011, Art. no. 452.
[67] S. Hussain *et al.*, "NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs," *Cell Rep.*, vol. 4, no. 2, pp. 255–261, 2013.
[68] O. Rossbach *et al.*, "Crosslinking-immunoprecipitation (iCLIP) analysis reveals global regulatory roles of hnRNP l," *RNA Biol.*, vol. 11, no. 2, pp. 146–155, 2014.
[69] J. König *et al.*, "iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution," *Nature Structural Mol. Biol.*, vol. 17, no. 7, 2010, Art. no. 909.
[70] K. Zarnack *et al.*, "Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of alu elements," *Cell*, vol. 152, no. 3, pp. 453–466, 2013.
[71] C. Sievers, T. Schlumpf, R. Sawarkar, F. Comoglio, and R. Paro, "Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in mov10 par-clip data," *Nucleic Acids Res.*, vol. 40, no. 20, pp. e160–e160, 2012.
[72] J. I. Hoell *et al.*, "RNA targets of wild-type and mutant fet family proteins," *Nature Structural Mol. Biol.*, vol. 18, no. 12, 2011, Art. no. 1428.
[73] S. Kishore, L. Jaskiewicz, L. Burger, J. Hausser, M. Khorshid, and M. Zavolan, "A quantitative analysis of clip methods for identifying binding sites of RNA-binding proteins," *Nature Methods*, vol. 8, no. 7, 2011, art. no. 559.
[74] R. L. Boudreau *et al.*, "Transcriptome-wide discovery of microRNA binding sites in human brain," *Neuron*, vol. 81, no. 2, pp. 294–305, 2014.
[75] J. R. Sanford *et al.*, "Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts," *Genome Res.*, vol. 19, no. 3, pp. 381–394, 2009.
[76] J. Saulière *et al.*, "Clip-seq of eif4aiii reveals transcriptome-wide mapping of the human exon junction complex," *Nature Structural Mol. Biol.*, vol. 19, no. 11, 2012, Art. no. 1124.
[77] Y. LeCun L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[78] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735
[79] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014 in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Association for Computational Linguistics, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.
[80] A. J. Stewart, S. Hannenhalli, and J. B. Plotkin, "Why transcription factor binding sites are ten nucleotides long," *Genetics*, vol. 192, no. 3, pp. 973–985, 2012. [Online]. Available: https://www.genetics.org/content/192/3/973
[81] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[82] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.
[83] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[84] S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble, "Quantifying similarity between motifs," *Genome Biol.*, vol. 8, no. 2, 2007, Art. no. R24.
[85] S. Budach and A. Marsico, "pysster: Classification of biological sequences by learning sequence and structure motifs with convolutional neural networks," *Bioinformatics*, vol. 34, no. 17, pp. 3035–3037, 2018.
[86] D. Chicco, "Ten quick tips for machine learning in computational biology," *BioData Mining*, vol. 10, no. 1, 2017, Art. no. 35.
[87] D. Ray *et al.*, "A compendium of rna-binding motifs for decoding gene regulation," *Nature*, vol. 499, no. 7457, 2013, Art. no. 172.

Hilal Tayara received the BSc degree in computer engineering from Aleppo University in Aleppo, Syria, in 2008, and the MS and the PhD degrees in electronics and information engineering from Chonbuk National University in Jeonju, South Korea, in 2015 and 2019. He is currently a researcher at Chonbuk National University. His research interests include bioinformatics, machine learning, and image processing.

Kil To Chong received the PhD degree in mechanical engineering from Texas A&M University, in 1995. Currently, he is a professor with the School of Electronics and Information Engineering, Chonbuk National University in Jeonju, Korea, and head of the Advanced Research Center of Electronics. His research interests include areas of machine learning, signal processing, motor fault detection, network system control, and time-delay systems.