IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 30, 2021, accepted May 27, 2021, date of publication June 3, 2021, date of current version June 11, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3086102

Cr-Prom: A Convolutional Neural Network-Based Model for the Prediction of Rice Promoters

MUHAMMAD SHUJAAT[1,2], SEUNG BEOP LEE[3], (Member, IEEE), HILAL TAYARA[3], AND KIL TO CHONG[1,4], (Member, IEEE)

[1]Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju 54896, South Korea
[2]Department of Computer Sciences, Bahria University, Lahore 54600, Pakistan
[3]School of International Engineering and Science, Global Frontier College, Jeonbuk National University, Jeonju 54896, South Korea
[4]Advances Electronics and Information Research Center, Jeonbuk National University, Jeonju 54896, South Korea

Corresponding authors: Hilal Tayara (hilaltayara@jbnu.ac.kr) and Kil To Chong (kitchong@jbnu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant through the Korean Government [Ministry of science and ICT (MSIT)] under Grant 2020R1A2C2005612, and in part by the Brain Research Program of the National Research Foundation (NRF) through the Korean Government (MSIT) under Grant NRF-2017M3C7A1044816.

ABSTRACT The promoter is a regulatory region of the DNA typically located upstream of a gene and plays a key role in regulating gene transcription. Accurate prediction of promoters is crucial for the analysis of gene expression patterns and for the development and understanding of genetic regulatory networks. Genomes of several species have been sequenced, and their gene content has been established to a large extent. Some bioinformatics algorithms have been developed for predicting promoters with high universality for all kinds of plants; however, few studies have been conducted to identify promoters in rice, which might affect the practical applications. Here, we present a rice promoter prediction tool, Cr-Prom. This predictor has been established using a series of sequence-based features and datasets extracted from the PlantProm and RAP-DB databases. We applied a convolutional neural network (CNN)-based strategy to construct a predictor with robust classification performance. To demonstrate our dominance, we ran experiments on a benchmark dataset using 5-fold cross-validation and compared our results with existing techniques using four figure of merits. In addition, CR-Prom was analyzed on an independent dataset. Based on the results, Cr-Prom outperformed the existing rice-specific promoter predictors. The Cr-Prom tool can be freely accessed at: http://nsclbio.jbnu.ac.kr/tools/Cr-Prom/

INDEX TERMS Promoters, convolutional neural network (CNN), computational biology, bioinformatics, rice genome.

I. INTRODUCTION

Rice is a cereal crop that belongs to the Orayza genus, representing a variety of rice, which can be divided into two subspecies, namely indica and japonica. Rice can also be divided into conventional and hybrid rice depending upon its production type. Being a vital direct cash crop, rice is the staple food for majority of the population all over the world. From basic studies to molecular breeding, researchers have played a significant role in boosting rice production worldwide. Owing to the rapid development of biotechnology and genetic engineering technology, scientists started analyzing and collating rice genome in 1998, and by 2002, the entire rice genome map had been interpreted. The rice genome is the most completely sequenced genome among higher organisms [1]. Among the 37,500 genes identified, several have a significant role to play in agricultural production. For example, research on key genes can help increase the yield of rice [2] or change the photoperiod of rice.

Transcription of protein genes and most non-coding RNA genes, as well as that of the DNA regions with uncertain functions, is performed in the nuclear genomes of eukaryotic organisms by RNA polymerase II (Pol II). Transcription controls cellular differentiation and function by initiating expression at specific genomic locations. Changes in gene regulation are the main driving factors for the majority of universal diversification between species [3], [4] and phenotypic diversity within the same species [5]–[7]. Gene regulation has been targeted by a large number of genetic, biological, chemical, and computational studies.

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

Promoters play a significant role in the regulation of gene expression and their accurate prediction has both fundamental and practical significance. The promoter is located essentially upstream of the transcription starting point of the gene and is not involved in the transcription itself [5]. However, some promoters, such as tRNA promoters are located downstream of the transcription starting point [7], and these DNA sequences can be transcribed [8]. The precise identification of transcription sites is key to the successful expression of genes. There are currently several publicly accessible collections of transcripts and promoter sequences with annotated transcription start sites (TSS). The identification and characterization of promoters and TSSs is crucial for the unraveling of mechanisms that control the RNA Pol II transcription. This shows the importance of proper functioning of the promoter in the plant. In the case of rice, the proper functioning of the promoter directly affects agricultural production.

Biological methods for the identification of promoters are usually time-consuming and require a costly procedure to be followed. Although, identical or similar genes exist in different species, there are significant differences in gene expression; this phenomenon is known as biological diversity. The first step to infer whether a gene can be expressed normally is whether it can be transcribed normally, and promoters are the key gene sequences that control transcription. Promoters that show significant differences and diversity along with the transformation of species can control the normal transcription of genes in the coding regions of different species. Therefore, the establishment of a novel predictor for rice promoter prediction is urgently required. Considering the aforementioned drawbacks associated with biological techniques, we believe that employing computational techniques for the prediction of promoters will be a better option.

In recent years machine learning and artificial intelligence have demonstrated great success in different applications [9]–[13]. According to previous studies, genome-wide computational prediction and a two-layer predictor for identifying promoters and their types using a multi-window-based PseKNC [14] can provide us with ideas and technical guidance. A tool for the prediction of plant Pol II, called the TSSPlant, was developed [15]. A previously published plant TATA-box NFM, which was determined using 345 promoters from Plant Prom DB and applying an EM procedure, was used for building the predictor for plant promoters. TSSPlant was developed to predict the promoters of all types of plants. Another tool, called ProRice [16], was proposed, and it used an ensemble learning-based approach to identify promoters from rice genome. The datasets employed in this study were extracted from the PlantProm and RAP-DB databases. ProRice achieved 92.3% accuracy, 95.2% sensitivity, 90% specificity, and 0.857 MCC. Irrespective of these results, the main constraint of ProRice is the extraction of local features.

Both TSSPlant and ProRice are machine learning-based tools that use different numbers of features; for example, ProRice used 11 different features, whereas TSSPlant used 19 different features for model training and evaluation; however, there was considerable scope for improvement in their prediction performances. Notably, a tool that applies neural networks (NN) for the prediction of promoters in rice genome has not yet been developed.

Therefore, this study aimed to develop a tool using a CNN based approach to predict rice promoters. Accurate prediction of rice promoters is of great significance for subsequent studies on molecular design and breeding [17], gene expression regulation, gene transduction, modified genes [18], and for the advancement of other microbiological sciences and technology. We used Chou's [19] five-step rule in our study; several recent publications have used this rule [20]–[22]:

1) Preparing a benchmark dataset
2) Mathematically expressing the sequence
3) Constructing Classification Model
4) Model evaluation
5) Making predictor webserver publicly available

Figure 1 illustrates the graphical representation of the five steps; the remaining text follows the research flow depicted in the steps presented by Chou's rule.

II. BENCHMARK DATASET

The selection of a suitable benchmark dataset is important for the development of an efficient biological predictor and enables the evaluation of the performance of the predictive model. In this study, the promoter dataset used was the same as that used in the development of ProRice. The 4220 rice promoter was extracted from the PlantProm database that contains the TSSs from a variety of plant species. Negative promoter samples of the same size were obtained from the rice annotation project database (RAP-DB) (https://rapdb.dna.affrc.go.jp/download/irgsp1.html) [23]. The length of each sequence screened by the positive and negative samples was 251bp. To ensure robustness of the predictor model, the database must be pre-processed to remove noise [24], for which we used CD-HIT-EST [25], and set the cut-off value of 0.8 to remove mostly similar promoter sequences [26]. Subsequently, we joined the positive and negative promoter sequences and randomly separated them into training and testing sets using k-fold. From the training split 70% is used for training the model while 30% is used for the validation of the model. A numerical summary of the promoter and non-promoter datasets for rice is listed in Table 1. The dataset used in this study is freely available and can be downloaded from https://github.com/Shujaatmalik/Cr-Prom.

TABLE 1. Summary of the dataset.

Classes	Benchmark Dataset	Sequence length
Promoter	4220	251 bp
Non-Promoter	4220	251 bp

III. PROPOSED METHODOLOGY

A. ENCODING SCHEME

A DNA sequence is comprised of four nucleotides, namely adenine (A), cytosine (C), guanine (G), and thymine (T).

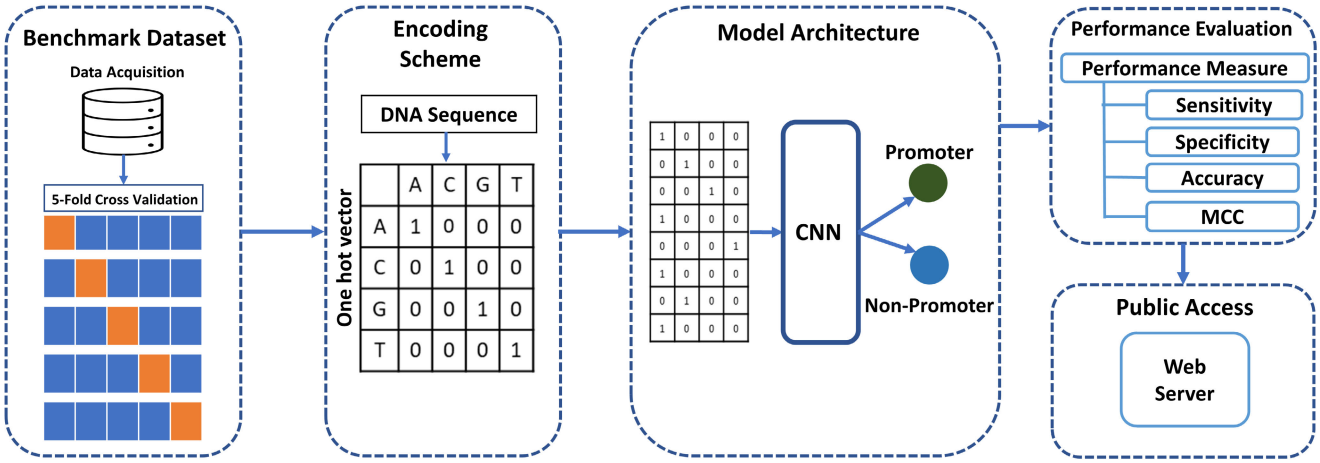


FIGURE 1. Illustrate the graphical representation of five steps.

A numerical representation is required to perform computational operations on DNA sequences. The one-hot encoding scheme shows great effectiveness when applied in deep learning; a number of recent studies in the domain of computer science [27]–[29] and bioinformatics [30]–[32] have applied the one-hot encoding scheme. Using this method, a single nucleotide is converted to a binary vector of four dimensions, where one element is denoted as 1 and all the remaining elements are represented by 0; for each nucleotide, the mathematical representation is as follows:

$$A \Longrightarrow (1, 0, 0, 0)$$
$$C \Longrightarrow (0, 1, 0, 0)$$
$$G \Longrightarrow (0, 0, 1, 0)$$
$$T \Longrightarrow (0, 0, 0, 1)$$

Given that the length of the promoter sequence was 251, each sample was converted into a numerical vector of 251×4 using one-hot encoding.

B. MODEL ARCHITECTURE

CNNs are computational models that use different layers to assimilate features from a dataset with various degrees of deliberation. The ongoing advancements in CNNs has made them highly reliable, and these networks have achieved novel results in various fields. CNNs have also achieved remarkable results in the area of medical image processing [29], [33], [34] and bioinformatics [35], [36]. However, numerous notable examples use CNNs to build predictors that can detect the variation occurred in genetic sequence. The foremost benefit of a CNN is that it does not necessitate preliminary feature extraction. A CNN-based model can directly derive features from the input. This research, utilizes a CNN based model for the prediction of promoters.

A crucial step in promoter detection is to find the precise positions in the promoter region where promoter elements such as the TATA-box, CAAT-box, and GC-box, are localized [37]. Although such positional details are essential for the identification of promoters, the maximum pooling layer or average pooling layer used in CNNs cause the sequence location information to deteriorate to a certain degree [38]. We herein propose a CNN-based architecture the Cr-Prom, for the prediction of rice sequence as a promoter or a non-promoter. Figure 2 illustrates the proposed architecture of the CR-Prom, which consists of two one-dimensional convolution layers. The encoded one-hot sequence is transferred to the input layer of the model.

The batch-normalization and dropout layers are preceded by the first convolution layer, and the max-pooling and dropout layers are preceded by the second convolution layer. The characteristics derived from the convolution layers were flattened and utilized for classification, using the dense layer. Hyper-parameter tuning is performed to select the optimum parameters for convolution, pooling, dropout, and dense layers. The variety of hyperparameters used for tuning purposes is displayed in Table 2.

TABLE 2. Hyper-parameter tuning parameters.

Convolution Filters Size	16,32,64
Kernel Size	3,5,7
Pooling Layer Kernel Size	2,4
Dropout Ratio	0.15,0.20,0.25,0.30,0.35,0.40
Dense Layer Neurons	8,16,32,64,100

The first and second convolution layers use filters of sizes 16 and 32, respectively, and the kernel for both of these layers was 5. Batch normalization was performed on the extracted features by the first convolutional layer, while max pooling was performed using a pool of size 2 with two strides on the features extracted by the second convolutional layer. For feature selection, the dropout layer discharged 25% of the features at the end of each convolution layer. The ReLU activation function was used in both the convolutional layers.

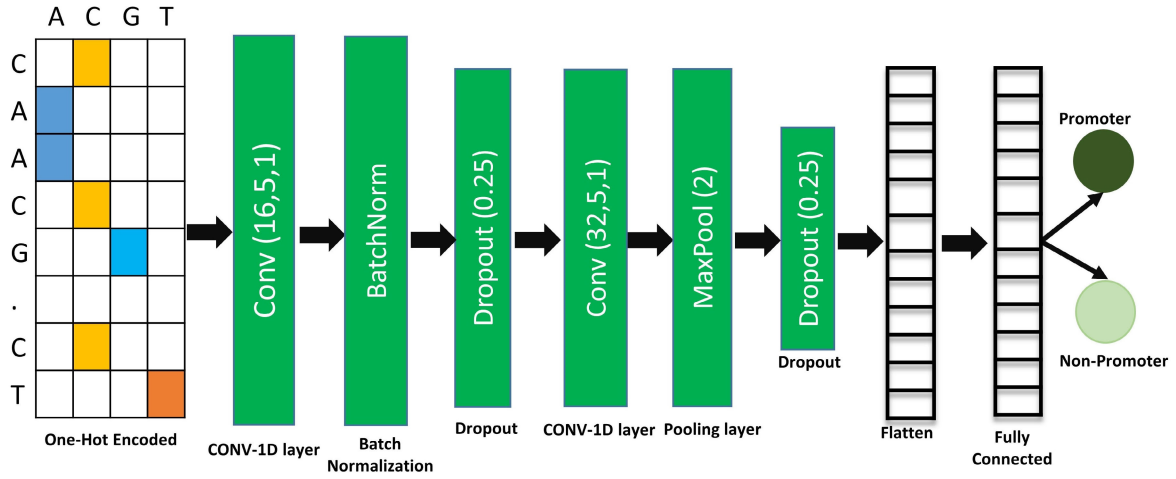


FIGURE 2. Proposed CNN architecture.

Mathematically ReLU can be represented as follows:

$$R(x) = max(0, x) \tag{1}$$

The two fully-connected layers had 16 and 1 neurons, respectively. The ReLU was used as an activation function in the first fully connected layer, whereas the second fully connected layer used a single neuron with a sigmoid activation function, and can be represented as

$$F(s) = \frac{1}{1 + exp(-s)} \tag{2}$$

To avoid overfitting, we used L2 regularization in all the convolution and dense layers. The optimizer used in this model was the stochastic gradient descent (SGD) with a learning rate of 0.007 and momentum of 0.96. Binary cross entropy (BCE) was used in the loss function and can be mathematically represented as

$$BCE(t, p) = -(t * log(p) + (1 - t) * log(1 - p)) \tag{3}$$

IV. RESULTS AND DISCUSSION

In this section we discussed the evaluation parameters, achieved performance by the proposed Cr-Prom and its comparison with state-of-the-art methods.

A. EVALUATION PARAMETER

We used broadly applied methodological measures [14], [36], [39]–[41] to comprehensively analyze the efficiency of the promoter's prediction. These include Matthew's correlation coefficient (MCC), accuracy (Acc), sensitivity (Sn), specificity (Sp), and ROC curve. These metrics can be mathematically expressed as follows:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

$$Acc = \frac{TN + TP}{TP + TN + FP + TN} \tag{5}$$

$$Sp = \frac{TN}{TN + FP} \tag{6}$$

$$Sn = \frac{TP}{TP + FN} \tag{7}$$

True positives, true negatives, false positives, and false negatives are denoted by TP, TN, FP, and FN, respectively. The number of correctly classified promoters is denoted by TP while the number of correctly classified non-promoters is denoted by TN. FN represents the number of incorrectly classified non-promoters and FP represents the number of incorrectly classified non-promoters. Therefore, Sn (also known as the true positive rate) measures the percentage of correctly classified promoters, and Sp measures the percentage of analogously correctly identified non-promoters. The balance quality of the positive and negative data was represented by MCC. Moreover, to determine the overall classification performance, the receiver operating characteristic (ROC) curve is also assessed.

TABLE 3. Classification performance.

Methods	Sn	Sp	Acc	Mcc
TSSPlant	0.822	0.925	0.874	0.752
ProRice	0.952	0.903	0.923	0.857
CR-PROM	0.9857	0.999	0.991	0.9839

B. PERFORMANCE EVALUATION

We performed 5-fold cross validation, to assess the achieved performance by the proposed model. Similarly, the experiments were conducted using ProRice and TSSPlant, which are state-of-the-art diagnostic and classification methods. Table 3 demonstrates the Cr-Prom results for the prediction of promoters and non-promoters. As can be seen, a significant improvement was observed when compared with the results of state-of-the-art methods.

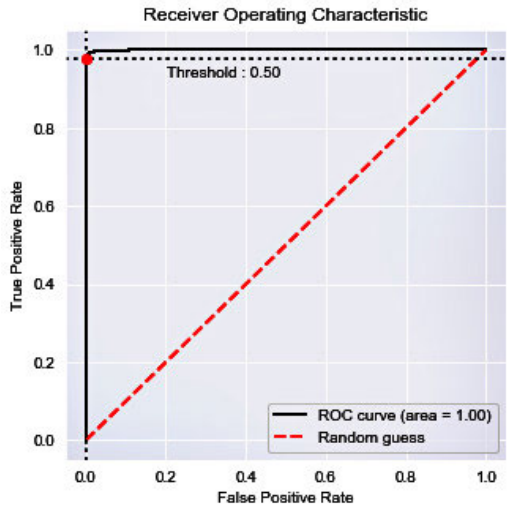


FIGURE 3. ROC curve.

Cr-Prom achieved a sensitivity of 98.57%, specificity of 99.9%, accuracy of 99.1%, and an MCC of 0.9839. All these values were greater than those of ProRice and TSSPlant. The results therefore indicate that Cr-Prom is more suitable for the rice-specific promoter classification problem. A substantial 12.9% increase in the MCC value demonstrates the reliability of the proposed methodology in identifying the classes of the promoters and non-promoters. The ROC curve for the estimation of the promoter and non-promoter regions is shown in Figure 3. An AUC value closer to 1 indicates that the ROC curve is positioned closer to the upper-left corner, indicating a highly accurate prediction.

C. MODEL EVALUATION ON INDEPENDENT DATASET

We performed model evaluation on an independent dataset, for which we used promoters of the plants class that were recently updated on the database. Eukaryotic Promoter Database (EPD) consists of the following two plant species: Arabidopsis thaliana and Zea mays. Table 4 lists the total number of promoters present in each class.

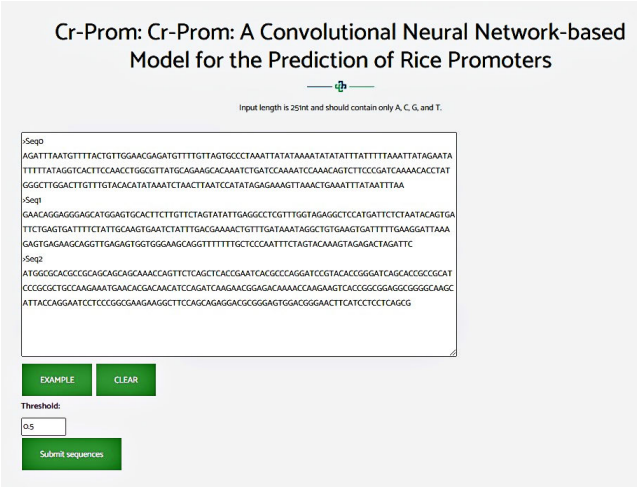
TABLE 4. Details for independent dataset.

Class	Number of Promoters
Arabidopsis thaliana	22702
Zea mays	8264

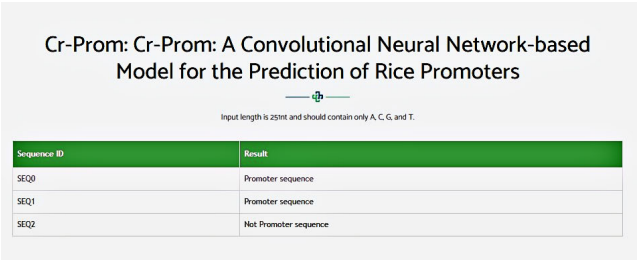
TABLE 5. Evaluation on independent dataset.

Parameter	Arabidopsis thaliana	Zea mays
True Positive	22435	8073
False Negative	267	191

The independent test dataset did not have any non-promoters. Therefore, we only reported the values of true positives and false negatives. Table 5 presents the performance results for Cr-Prom for both the species.



(a) Input to the Webserver



(b) Output from the Webserver

FIGURE 4. WebServer.

V. WEBSERVER

Following the procedure followed by numerous researchers, a web server that hosts our Cr-Prom tool is publicly available to provide easy access to this tool for the research community [28], [42]. Cr-Prom is an easy-to-use platform for researchers and professionals in the areas of bioinformatics and biology. This webserver accepts two types of input data: direct sequence input and uploading a file that contains the sequences which needs to be evaluated. Each sequence containing A, C, G, and T must be 251 nt long, with a limit of 1000 sequences available for prediction when uploading a file.. A web-server snippet can be seen in Figure 4. Figure 4a shows an example of inserting prediction sequences, whereas Figure 4b shows the predictor's output. In addition, the Cr-Prom webserver is made available at: http://nsclbio.jbnu.ac.kr/tools/Cr-Prom/

VI. CONCLUSION

We developed a sequence-based CNN method, called the Cr-Prom, to address the major challenges encountered when uncovering the promoters from a large number of DNA sequences in rice generated in the postgenomic period. A successful discrimination output between the promoter and non-promoter DNA sequences, specifically for rice, was obtained using the proposed method. A single encoding scheme was used by this CNN-based tool, and the proposed architecture

was evaluated using a publicly accessible dataset as well as on independent dataset. Overall, the tool achieved superior results in comparison to the existing techniques.

ACKNOWLEDGMENT

(Muhammad Shujaat and Seung Beop Lee contributed equally to this work.)

REFERENCES

[1] I. R. G. S. Project, "The map-based sequence of the rice genome," *Nature*, vol. 436, no. 7052, pp. 793–800, Aug. 2005.
[2] Z. Li, S. R. M. Pinson, W. D. Park, A. H. Paterson, and J. W. Stansel, "Epistasis for three grain yield components in rice (Oryxa sativa L.)," *Genetics*, vol. 145, no. 2, pp. 453–465, Feb. 1997.
[3] P. J. Wittkopp and G. Kalay, "Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence," *Nature Rev. Genet.*, vol. 13, no. 1, pp. 59–69, Jan. 2012.
[4] R. S. Young, Y. Hayashizaki, R. Andersson, A. Sandelin, H. Kawaji, M. Itoh, T. Lassmann, P. Carninci, W. A. Bickmore, A. R. Forrest, and M. S. Taylor, "The frequent evolutionary birth and death of functional promoters in mouse and human," *Genome Res.*, vol. 25, no. 10, pp. 1546–1557, Oct. 2015.
[5] P. Khaitovich, S. Pääbo, and G. Weiss, "Toward a neutral evolutionary model of gene expression," *Genetics*, vol. 170, no. 2, pp. 929–939, Jun. 2005.
[6] S. McCarroll, "Murphy CT, Zou S, Pletcher SD, Chin CS, Jan YN, Kenyon C, Bargmann CI, Li H," *Comparing Genomic Expression Patterns Across Species identifies Shared Transcriptional Profile Aging. Nat. Genet*, vol. 36, pp. 197–204, Jan. 2004.
[7] I. Tirosh, A. Weinberger, M. Carmi, and N. Barkai, "A genetic signature of interspecies variations in gene expression," *Nature Genet.*, vol. 38, no. 7, pp. 830–834, Jul. 2006.
[8] C. R. Landry, B. Lemos, S. A. Rifkin, W. J. Dickinson, and D. L. Hartl, "Genetic properties influencing the evolvability of gene expression," *Science*, vol. 317, no. 5834, pp. 118–121, Jul. 2007.
[9] M. U. Rehman, Z. Abbas, S. H. Khan, S. H. Ghani, and Najam, "Diabetic retinopathy fundus image classification using discrete wavelet transform," in *Proc. 2nd Int. Conf. Eng. Innov. (ICEI)*, Jul. 2018, pp. 75–80.
[10] T. Ilyas, A. Khan, M. Umraiz, and H. Kim, "SEEK: A framework of superpixel learning with CNN features for unsupervised segmentation," *Electronics*, vol. 9, no. 3, p. 383, Feb. 2020.
[11] A. Khan, H. Kim, and L. Chua, "PMED-net: Pyramid based multi-scale encoder-decoder network for medical image segmentation," *IEEE Access*, vol. 9, pp. 55988–55998, 2021.
[12] I. F. Nizami, M. U. Rehman, M. Majid, and S. M. Anwar, "Natural scene statistics model independent no-reference image quality assessment using patch based discrete cosine transform," *Multimedia Tools Appl.*, vol. 79, nos. 35–36, pp. 26285–26304, Sep. 2020.
[13] I. F. Nizami, M. Majid, M. U. Rehman, S. M. Anwar, A. Nasim, and K. Khurshid, "No-reference image quality assessment using bag-of-features with feature selection," *Multimedia Tools Appl.*, vol. 79, nos. 11–12, pp. 7811–7836, Mar. 2020.
[14] B. Liu, F. Yang, D.-S. Huang, and K.-C. Chou, "IPromoter-2L: A two-layer predictor for identifying promoters and their types by multi-window-based PseKNC," *Bioinformatics*, vol. 34, no. 1, pp. 33–40, Jan. 2018.
[15] I. A. Shahmuradov, R. K. Umarov, and V. V. Solovyev, "TSSPlant: A new tool for prediction of plant pol II promoters," *Nucleic Acids Res.*, vol. 45, no. 8, p. e65, 2017.
[16] Y. Zuo, H. Zhou, and Z. Yue, "Prorice: An ensemble learning approach for predicting promoters in rice," in *Proc. 4th Int. Conf. Comput. Sci. Appl. Eng.*, 2020, pp. 1–5.
[17] L. De la Rosa, E. Zambrana, and E. Ramirez-Parra, "Molecular bases for drought tolerance in common vetch: Designing new molecular breeding tools," *BMC Plant Biol.*, vol. 20, no. 1, pp. 1–18, Dec. 2020.
[18] W. Jiang, B. Yang, and D. P. Weeks, "Efficient CRISPR/Cas9-mediated gene editing in Arabidopsis thaliana and inheritance of modified genes in the T2 and T3 generations," *PLoS ONE*, vol. 9, no. 6, Jun. 2014, Art. no. e99225.
[19] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *J. Theor. Biol.*, vol. 273, no. 1, pp. 236–247, Mar. 2011.
[20] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, and K.-C. Chou, "IRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC," *Mol. Therapy-Nucleic Acids*, vol. 7, pp. 155–163, Jun. 2017.
[21] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "ISuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset," *Anal. Biochem.*, vol. 497, pp. 48–56, Mar. 2016.
[22] S. D. Ali, W. Alam, H. Tayara, and K. Chong, "Identification of functional piRNAs using a convolutional neural network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Oct. 29, 2020, doi: 10.1109/TCBB.2020.3034313.
[23] Y. Kawahara, M. de la Bastide, J. P. Hamilton, H. Kanamori, W. R. McCombie, S. Ouyang, and T. Matsumoto, "Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data," *Rice*, vol. 6, no. 1, pp. 1–10, 2013.
[24] W. He and C. Jia, "EnhancerPred2.0: Predicting enhancers and their strength based on position-specific trinucleotide propensity and electron–ion interaction potential feature selection," *Mol. Biosyst.*, vol. 13, no. 4, pp. 767–774, 2017.
[25] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT suite: A Web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, Mar. 2010.
[26] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, "Sequence clustering in bioinformatics: An empirical study," *Briefings Bioinf.*, vol. 21, no. 1, pp. 1–10, Sep. 2018.
[27] Z. Yu, Z. Niu, W. Tang, and Q. Wu, "Deep learning for daily peak load forecasting—A novel gated recurrent neural network combining dynamic time warping," *IEEE Access*, vol. 7, pp. 17184–17194, 2019.
[28] M. U. Rehman, K. J. Hong, H. Tayara, and K. T. Chong, "M6A-NeuralTool: Convolution neural tool for RNA N6-methyladenosine site identification in different species," *IEEE Access*, vol. 9, pp. 17779–17786, 2021.
[29] M. ur Rehman, S. H. Khan, Z. Abbas, and S. M. D. Rizvi, "Classification of diabetic retinopathy images based on customised CNN architecture," in *Proc. Amity Int. Conf. Artif. Intell. (AICAI)*, Feb. 2019, pp. 244–248.
[30] M. U. Rehman and K. T. Chong, "DNA6 mA-MINT: DNA-6 mA modification identification neural tool," *Genes*, vol. 11, no. 8, p. 898, Aug. 2020.
[31] M. Oubounyt, Z. Louadi, H. Tayara, and K. T. Chong, "DeePromoter: Robust promoter predictor using deep learning," *Frontiers Genet.*, vol. 10, p. 286, Apr. 2019.
[32] 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.
[33] M. U. Rehman, S. Cho, J. H. Kim, and K. T. Chong, "BU-net: Brain tumor segmentation using modified U-net architecture," *Electronics*, vol. 9, no. 12, p. 2203, Dec. 2020.
[34] M. U. Rehman, S. Cho, J. Kim, and K. T. Chong, "BrainSeg-net: Brain tumor MR image segmentation via enhanced encoder–decoder network," *Diagnostics*, vol. 11, no. 2, p. 169, Jan. 2021.
[35] Z. Abbas, H. Tayara, and K. T. Chong, "SpineNet-6 mA: A novel deep learning tool for predicting DNA N6-methyladenine sites in genomes," *IEEE Access*, vol. 8, pp. 201450–201457, 2020.
[36] M. Shujaat, A. Wahab, H. Tayara, and K. T. Chong, "PcPromoter-CNN: A CNN-based prediction and classification of promoters," *Genes*, vol. 11, no. 12, p. 1529, Dec. 2020.
[37] R. Umarov, H. Kuwahara, Y. Li, X. Gao, and V. Solovyev, "Promoter analysis and prediction in the human genome using sequence-based deep learning models," *Bioinformatics*, vol. 35, no. 16, pp. 2730–2737, Aug. 2019.
[38] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.
[39] R. Amin, C. R. Rahman, S. Ahmed, M. H. R. Sifat, M. N. K. Liton, M. M. Rahman, M. Z. H. Khan, and S. Shatabda, "IPromoter-BnCNN: A novel branched CNN-based predictor for identifying and classifying sigma promoters," *Bioinformatics*, vol. 36, no. 19, pp. 4869–4875, Dec. 2020.
[40] M. S. Rahman, U. Aktar, M. R. Jani, and S. Shatabda, "IPromoter-FSEn: Identification of bacterial $\sigma 70$ promoter sequences using feature subspace based ensemble classifier," *Genomics*, vol. 111, no. 5, pp. 1160–1166, Sep. 2019.

[41] M. Zhang, F. Li, T. T. Marquez-Lago, A. Leier, C. Fan, C. K. Kwoh, K.-C. Chou, J. Song, and C. Jia, "MULTiPly: A novel multi-layer predictor for discovering general and specific types of promoters," *Bioinformatics*, vol. 35, no. 17, pp. 2957–2965, Sep. 2019.
[42] Z. Abbas, H. Tayara, and K. T. Chong, "4mCPred-CNN—Prediction of DNA N4-methylcytosine in the mouse genome using a convolutional neural network," *Genes*, vol. 12, no. 2, p. 296, Feb. 2021.

MUHAMMAD SHUJAAT received the B.S. and M.S. degrees in computer sciences from Bahria University, Islamabad, Pakistan, in 2016 and 2018, respectively. For several years, he served as a Junior Lecturer with the Computer Science Department, Bahria University, Lahore, Pakistan. He is currently serving as a Research Scholar with Jeonbuk National University, Jeonju, South Korea. His research interests include artificial intelligence, medical imaging, and bioinformatics.

SEUNG BEOP LEE (Member, IEEE) received the B.S. degree from the Department of Information and Control Engineering, Kwangwoon University, Seoul, South Korea, in 2010, and the M.S. and Ph.D. degrees from the Cho Chun Shik Graduate School of Green Transportation, KAIST, Daejeon, South Korea, in 2013 and 2018, respectively. He is currently an Assistant Professor with the School of International Engineering and Science, Graduate School of Integrated Energy-AI, Jeonbuk National University, Jeonju, South Korea. His expertise covers design optimization ranging from the component-level to system-level, based on computational analysis. His current research interests include developing new design methods for new drugs, solar power systems, and wireless power transfer systems by using artificial intelligence and design optimization.

HILAL TAYARA received the B.Sc. degree in computer engineering from Aleppo University, Aleppo, Syria, in 2008, and the M.S. and Ph.D. degrees in electronics and information engineering from Jeonbuk National University, Jeonju, South Korea, in 2015 and 2019, respectively. He served as a Researcher with Jeonbuk National University, where he is currently serving as an Assistant Professor with the School of International Engineering and Science. His research interests include bioinformatics, machine learning, and image processing.

KIL TO CHONG (Member, IEEE) received the Ph.D. degree in mechanical engineering from Texas A&M University, in 1993. He is currently a Professor with the School of Electronics and Information Engineering, Jeonbuk National University, Jeonju, South Korea, and the Head of the Advanced Information and Electronics Research Center, Jeonbuk National University. He is also the President of the Korean Electronics Engineering Society, Systems and Control. His research interests include bioinformatics, artificial intelligence, brain disease, and new drug discovery.