NAR Genomics and Bioinformatics, 2024, **6**, lqae055
https://doi.org/10.1093/nargab/lqae055
Advance access publication date: 28 May 2024
Standard Article

OXFORD

Interpretable prediction of mRNA abundance from promoter sequence using contextual regression models

Song Wang [1] **and Wei Wang** [1,2,*]

[1]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA 92093-0359, USA
[2]Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA 92093-0359, USA
[*]To whom correspondence should be addressed. Tel: +1 858 822 4240; Fax: +1 858 822 4236; Email: wei-wang@ucsd.edu

Abstract

While machine learning models have been successfully applied to predicting gene expression from promoter sequences, it remains a great challenge to derive intuitive interpretation of the model and reveal DNA motif grammar such as motif cooperation and distance constraint between motif sites. Previous interpretation approaches are often time-consuming or have difficulty to learn the combinatory rules. In this work, we designed interpretable neural network models to predict the mRNA expression levels from DNA sequences. By applying the Contextual Regression framework we developed, we extracted weighted features to cluster samples into different groups, which have different gene expression levels. We performed motif analysis in each cluster and found motifs with active or repressive regulation on gene expression. By comparing the co-occurrence locations of discovered motifs, we also uncovered multiple grammars of motif combination including communities of cooperative motifs and distance constraints between motif pairs. These results revealed new insights of the regulatory architecture of promoter sequences.

Introduction

Promoters are critical for regulating gene expression. The promoter sequences influence the strength, number and position of transcription factor (TF) binding sites, which in turn regulate the transcriptional levels of genes (1–3). In eukaryotic cells, a promoter sequence can be divided into three regions based on the distance from the transcription start site (TSS): core promoter, proximal promoter, and distal promoter (4). The core promoter includes TSS, key DNA sequence elements such as TATA box, and downstream promoter element (DPE) (5). The proximal promoter is upstream from the core promoter, where transcription factors (TFs) predominantly bind (1). The distal promoter is further upstream from the proximal promoter that often contains weak TF binding sites (6). Uncovering the information encoded in the promoter sequences crucial for transcriptional regulation and unraveling the regulatory rules between gene expression and DNA sequences remain important problems.

Previous studies have analyzed the type, number, location, orientation of TF motifs, combinatorial strategies of different TF motifs, and the surrounding sequences of the TF binding sites in the promoters (7–23). For example, several known sequence motifs such as the TATA box (TATAWAAR) and the initiator sequence (YYANWYY in human) located at the fixed position in the core promoter region have been discovered (24). Distance constraints between motif combinations have also emerged from analyzing various TF binding sites. For example, the ETS:IRF composite element (EICE) prefers two nucleotide-long spacers and ETS:IRF response element (EIRE) prefers three nucleotide-long spacers (25). Furthermore, additional insights are obtained from recent efforts on generating millions of synthetic promoter sequences and measuring their impacts on gene expressions (5,16,18,26). Despite these progresses, there remain great challenges of uncovering the regulatory grammar encoded in the promoter sequences.

Several studies have been reported to predict the transcription strength or mRNA level from promoter or core promoter sequences by using various deep learning models (5,26–29). They found that the gene expression levels are controlled by the entire gene regulatory structure and specific combination of regulatory elements rather than single motifs or genomic regions (29). For example, different combinations of promoter motifs lead to different expression levels (4,16,30). However, the models for studying the regulatory rules are usually complex. Due to the black-box nature of deep learning (18), these models still lack a clear interpretation of the promoter architecture, such as the motif location and identity in the promoter regions that have the highest active or repressive effects on gene expression. Furthermore, they cannot reveal motif grammar such as the interactions of the TF binding sites, motif community, and coregulation effects of the motifs. A few interpretation approaches were reported, such as the perturbation impact map and saliency map (31), but they are often time-consuming or hard to learn the combinatory rules (18).

We have developed a framework called contextual regression to interpret nonlinear models (32,33), which concurrently optimize prediction accuracy and feature weights reflecting their contributions to the prediction. Here, we designed a workflow implementing this framework to predict the mRNA levels from the promoter sequences. The model can uncover the most predictive features by calculating the dot product between the sequence features (i.e. motifs) and their context weights to predict gene expression, and the contextual weight values indicate the importance of the features. The weighted features allow to cluster the promoter sequences into groups with different mRNA levels. By analyzing the extracted

Received: September 19, 2023. **Revised:** April 8, 2024. **Editorial Decision:** May 6, 2024. **Accepted:** May 12, 2024

features in each group, we uncovered the DNA segments with the active or repressive effects on expression, and found the enriched motifs in these segments. This workflow is flexible to be applied to seven different promoter sequence ranges (10.5 kb, 800 bp, 400 bp, 200 bp, 100 bp, 50 bp and 19 bp around TSS) with increasing resolution. Starting from the analysis of 10.5 kb promoter sequences around TSS, we found that there are several discrete sequence stripes showing higher contribution to the gene expression level. By comparing the co-occurrence locations of discovered motifs, we also found the motif grammars including the motif communities and motif pairs with specific distance constraints. Then, we analyzed regions of 800, 400, 200, 100 and 50 bp around TSS to increase resolution of locating the most predictive segments and motifs. Lastly, we focused on the downstream promoter region (DPR, +17 to +35 bp around TSS) sequences and elucidated the differences between synthetic and genomic sequences on regulating gene expression as well as identified several bases strongly preferring guanine in highly expressed genes.

Material and methods

Xpresso is a state-of-the-art model for predicting gene expression using promoter sequences (27). In order to make our results directly comparable to Xpresso's, the promoter sequences and corresponding mRNA expression levels were downloaded from (27). The dataset contains 18 377 genes in 56 human cell types generated by the NIH Roadmap Epigenomics Consortium. Following the same procedure as in (27), the median mRNA expression levels across 56 cell types were used for prediction because mRNA expression levels are highly correlated (average correlation of 0.78) between different cell types (27). For training the contextual regression (CR) models, we selected 1000 genes as the test dataset and another 1000 genes as the independent validation dataset. The remaining 16 377 genes were used as a training dataset. We performed 10 times of cross validation by randomly partitioning the dataset to verify the consistency of our models (Supplemental Table S1). We did DNA sequence analysis by the BLAST-like alignment tool (BLAT) and performed the chromosome hold-out tests to prove that the data leakage does not impact our model (Supplemental Table S9). The synthesized DPR sequences in the core promoter region and their corresponding transcriptional strength were downloaded from (5). Among the 468 069 sequences with measured transcriptional level, 7500 sequences were selected as the test dataset, 20 000 sequences as the validation dataset and 180 000 sequences as the training dataset.

We applied the same model structure with different fine-tuned hyperparameters to seven ranges of DNA sequences: the first one (the CR-1 model) for the 10.5 kb sequences around the TSS (−7 kb, +3.5 kb); the second one (CR-2) for the −400 bp to +400 bp sequences around the TSS; the third one (CR-3) for the −200 bp to +200 bp sequences around the TSS; the fourth one (CR-4) for the −144 bp to +56 bp sequences around the TSS; the fifth one (CR-5) for the −112 bp to −12 bp sequences around the TSS; the sixth one (CR-6) for the −92 bp to −42 bp sequences around the TSS; and the seventh one (CR-7-G for genomic sequences and CR-7-S for synthesized sequences) for DPR in core promoters that are +17 to +35 bp relative to the TSS. In each contextual regression model, the sequence features were extracted by a series of convolutional layers and subsequently several fully connected layers were applied to generate a weight vector with the same dimension of the sequence features. Finally, the model output was obtained by summing the dot product of the feature (CR-1 to CR-6) or one-hot-encoded vector (CR-7-G and CR-7-S) vector and the weight vector (Figure 1A). The models were built and trained on TensorFlow 1.15.2 (34) and Keras 2.2.4. The hyperparameters were slightly adjusted from those used in (27) for a specific dataset and model (Supplemental Table S2). The initial parameters were generated by the Glorot normal initializer (35). The Stochastic Gradient Descent (SGD) optimizer was used to optimize the parameters with a learning rate of 0.0005 and momentum of 0.9.

The motifs are found by the software STREME (36) with p-value threshold of 10^{-3}. The similarities between all pairs of motifs are determined by software TOMTOM (37) with the default parameters. The motifs' occurrence sites are explored by the software FIMO (38) with p-value threshold of 10^{-5}. The GO term enrichment analysis is performed by the software GOMO (39).

Results

Predicting gene expression using promoter sequences

Xpresso is a state-of-the-art model for predicting gene expression using promoter sequences (27), but it does not discover which regions have an active or repressive effect on gene expression. We proposed to use the contextual regression model (32,33) for this task. It first extracted features from the 10 kbp promoter sequences, showing that the promoter sequences are not equally important for predicting gene expressions and particular locations in the upstream of TSS are crucial for regulating transcription. The distinction between highly-expressed and lowly-expressed genes largely comes from sequences close to the TSS. To study the sequence features at finer resolutions, we trained additional models on the sequences of 800 to 50 bp around TSS and also the downstream promoter region (DPR, +17 to +35 bp around TSS).

We first trained a contextual regression model (32,33), referred to as CR-1, to predict gene expression levels using promoter sequences that span from −7 kb to +3.5 kb around the TSS (Figure 1A). The model's input is promoter sequences from 18 377 genes in 56 human cell types generated by the NIH Roadmap Epigenomics Consortium, and the model's output is the median mRNA expression levels across 56 cell types. The model was composed of three convolutional blocks, followed by two fully connected blocks and one fully connected layer. The convolutional filter lengths were 7, 7, 7, and the convolutional filter numbers were 32, 16, 8. The strides of max pooling layers were 50, 2, 2, which resulted in 200 bp bins in the extracted feature vector. The numbers of neurons in the fully connected layers were 64, 2, 410 (52 bins × 8 filters). To avoid overfitting, a dropout layer was added after the first two fully connected layers and the dropout probability was set to 0.00099. In the third fully connected layer, we applied L1 regularization on the weight with a penalty coefficient 0.0001 to make the weighted features more interpretable. The model performed well, with the predicted and measured values closely aligned along the diagonal line in the scatter plots (Figure 1B and C). The Pearson correlations for the training and testing datasets were close to

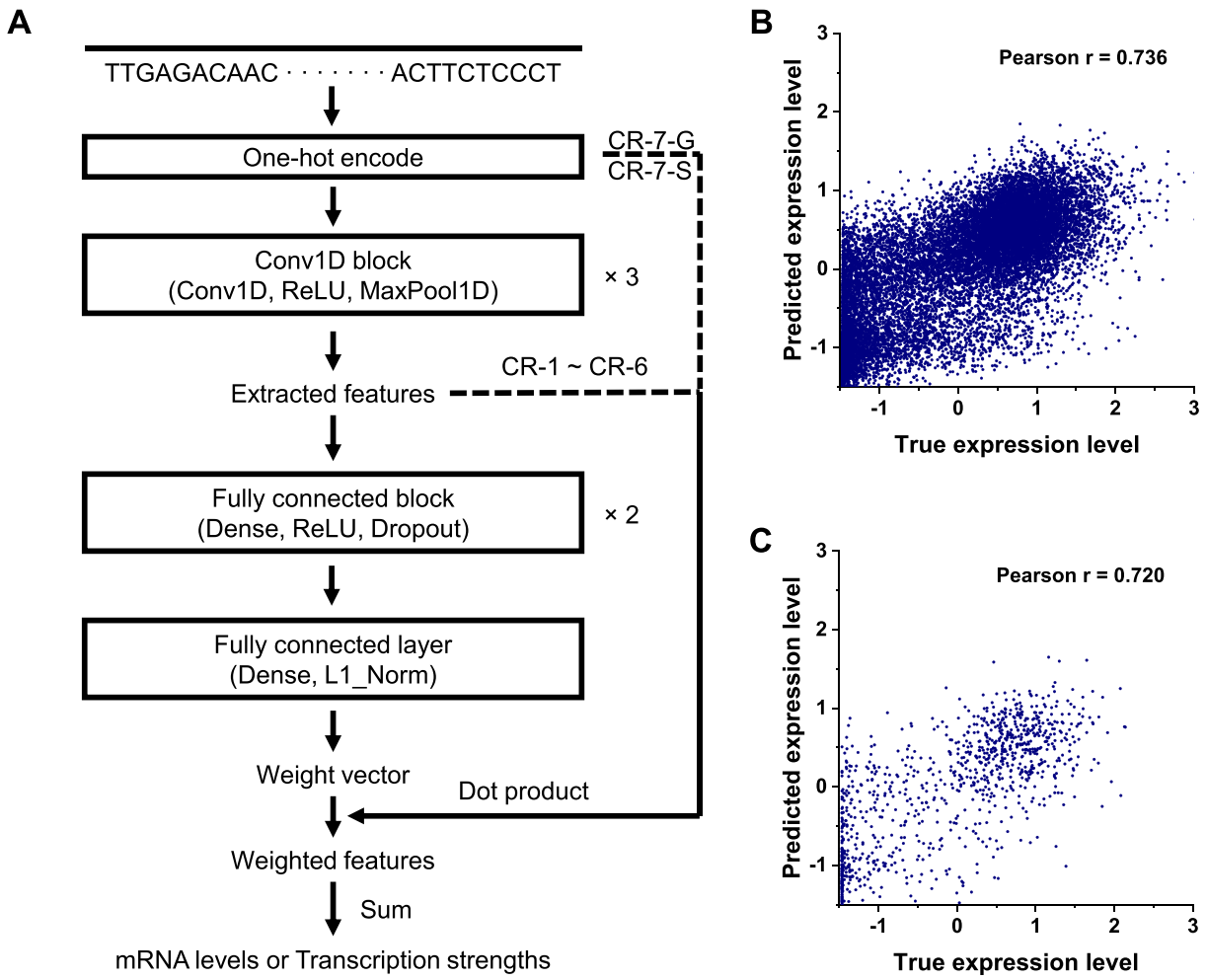

Figure 1. (**A**) The contextual regression model structure. (**B**) Prediction performance of CR-1 for the training dataset. The expression level is defined by the formula $y^{\wedge} = \log_{10}(y + 0.1)$. (**C**) Prediction performance of CR-1 for the testing dataset. The expression level is defined by the formula $y^{\wedge} = \log_{10}(y + 0.1)$.

each other (0.736 and 0.720, respectively), which suggests no overfitting. These correlation values are also comparable with the Xpresso results (27) (Supplemental Table S3).

The model CR-1 successfully uncovered the regions with positive and negative contextual weights (indicating active or repressive effects on gene expression) that are most predictive of gene expression (see below and Figure 2). Interestingly while not surprisingly, the contextual weights of the highly and lowly expressed genes are most distinct on the region next to TSS. As the CR-1 model was 200 bp resolution considering sequence information extraction efficiency and prediction ability, we zoomed into the −400 bp to +400 bp sequences corresponding to the 33rd to 36th bins to retrain the contextual regression model with higher resolution (referred to as CR-2 model). For different resolutions, some hyperparameters such as strides of max pooling layers and neuron numbers of the last fully connected layer were adjusted. The strides of max pooling layers were 4, 2, 2, which resulted in 16 bp-bins in the extracted feature vector. The numbers of neurons in the fully connected layers were 8, 2, 400 (50 bins × 8 filters). We also checked the prediction performance of model CR-2. As shown in Supplementary Figure S1, the Pearson correlations for the training and testing datasets are 0.709 and 0.688, respectively. Since the sequences used in CR-2 are much shorter than CR-1, it is reasonable that the correlation values in CR-2 are slightly lower than CR-1. Such a minor difference suggests that the sequences around the TSS heavily govern the gene expression.

To further improve the resolution of defining the sequence contribution to gene expression with a focus on the region around TSS, we trained four additional models for sequences of lengths 400, 200, 100 and 50 bp around the TSS. The last model was able to locate every base in its 50-bin length weighted feature layer. As shown in Supplementary Figures S2-S5, the Pearson correlation of CR-3 (−200 bp to +200 bp around TSS) was 0.696, only slightly decreasing from CR-1 and comparable with CR-2, while CR-4 to CR-6 with further decreased prediction performance with shorter sequence length. This observation suggested −200 bp to +200 bp sequences around TSS likely contain the majority of the regulatory information of gene expression.

We also analyzed the proximal downstream region around TSS (i.e. the DPR regions which are located in the 26th bin in the model CR-2) and focused on the sequence of +17 bp to +35 bp relative to TSS. This region has been shown to be crucial for transcription by, for example, analyzing gene expressions controlled by random sequences (5). We trained another contextual regression model CR-7-S for this region on

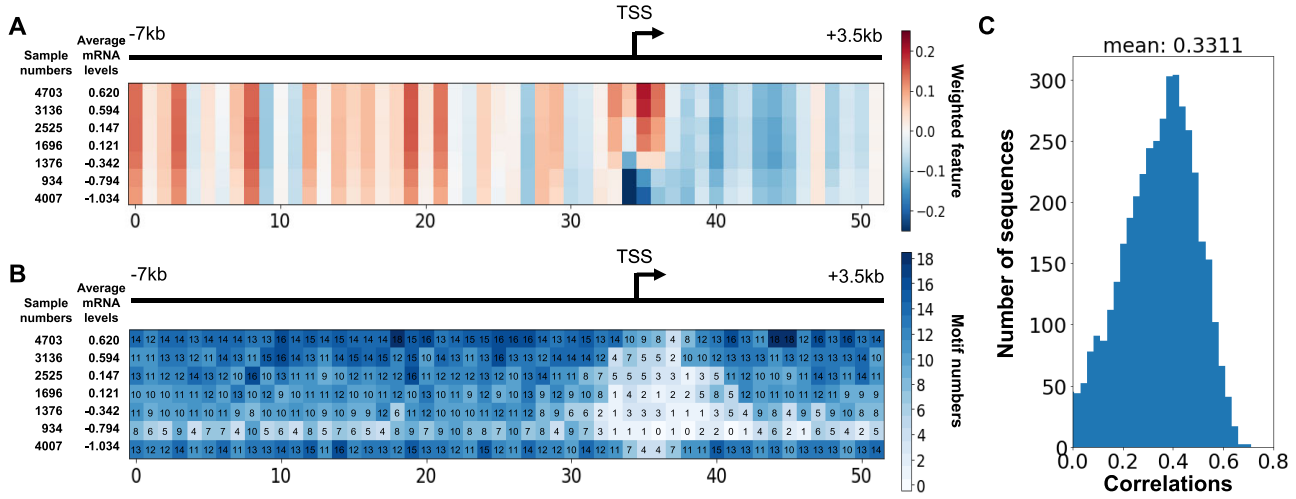

Figure 2. (**A**) The average weighted features of seven groups of samples. Red and blue colors represent positive and negative contributions to the prediction, respectively. The TSS is located in the 35th bin (bin index from 0 to 51). (**B**) The motif numbers in each group and each bin position. In each figure, the top shows the sequence range and TSS location. Left shows the sample numbers and the average mRNA levels ($y^{\wedge} = \log_{10}(y + 0.1)$). The groups are ordered by the descending order of average mRNA levels. (**C**) The distribution of correlations between the weighted features and the non-linear combination of CG content and H3K27ac signal.

the synthetic dataset (5). The model structure was slightly adjusted (Supplementary Figure S6). Instead of using the output from three convolutional layers, the weight layer was multiplied with the one-hot-encoded layer, which let each node in the weighted feature matrix correspond to each base. The strides of max pooling layers were 2, 2, 2. The numbers of neurons in the fully connected layers were 16, 2, 76. The Pearson correlations for the training and testing datasets are 0.896 and 0.836 respectively, indicating that the model successfully captured the regulatory relationship between sequences and gene expressions (Supplementary Figure S7). As a comparison, we also used genomic sequences to train another model CR-7-G for the same region. The Pearson correlations for the training and testing datasets are only 0.486 and 0.483 respectively (Supplementary Figure S8). This observation suggests that the genomic sequences of +17 bp to +35 bp only account for a small portion of random sequences and are insufficient for regulating transcription, highlighting the importance of sequences outside this region in the natural promoters in precise control of gene expressions.

We further analyzed the base preferences of the genomic (CR-7-G) and synthetic (CR-7-S) datasets in DPR. As shown in Supplementary Figure S15, we counted the base frequencies of sequences in DPR and compared them in the two groups with the highest (red line) and lowest (blue line) one percent of expression level. Overall, the genomic sequences have more G and C than A and T consistently in all positions while the synthetic sequences show larger fluctuation and some positions such as +30 have more A/T than C/G. Another striking observation is that G is present at significantly higher percentage in synthetic sequences compared to genomic sequences. Furthermore, the percentage profile of each base is also quite different between the genomic and synthetic sequences. These observations suggest that the genomic DPR sequences lack the additional features included in the synthetic sequences to regulate gene expression, which is consistent with that CR-7-G could not accurately predict gene expression levels using the genomic DPR sequences.

Promoter sequences are not equally important for regulating gene expression

The weighted feature layer of CR-1 comprises 410 neurons, corresponding to 52 bins, i.e. 8 neurons per bin. Before visualizing the features, the 8 weighted features belonging to each bin were summed together, resulting in weighted feature vector 52 bins in length for each 10 500 bp promoter sequence. The similar feature processing procedures were performed for CR-2 to CR-6, leading to 50-bin long weighted feature vectors for each 800 bp to 50 bp promoter sequence. The above six vectors were concatenated together resulting in 302-bin length weighted feature vector for each sequence. Next, we clustered all promoters into seven groups based on their 302-bin weighted feature vectors using the Ward linkage criterion (Supplementary Figure S9). Then, we calculated the average expression levels (Supplementary Figure S10) and averaged weight features for the samples in each group (Figure 2A, Supplementary Figure S11A, and Supplementary Figure S12).

Clearly, the promoter sequences are not equally predictive of gene expression (Figure 2A) and several upstream regions show strong activating effects on transcription (positive contextual weights), such as 0th, 3rd, 8th, 19th and 21st bins. These stripes of high contextual weights suggest that these locations have more contribution to the gene expression regulation than the other locations. The most prominent distinction between the highly and lowly expressed genes is in the bins next to TSS, consistent with the literature that these core promoter regions are crucial for transcriptional regulation. For the model with finer resolutions (Supplementary Figure S11A and Supplementary Figure S12), different promoter regions also exhibit different contributions to gene expression, which helps us gradually zoom in to key regions and eventually reach base-pair resolution. In the DPR sequences, most of positions show positive contributions especially for +20, reflecting its important ability of gene expression regulation.

To find what sequence and chromatin features may contribute to the striping patterns of regulatory importance in promoters, we calculated the correlations between each

sequence's weighted features and several simple sequence features including CG content, H3K27ac, H3K4me1 and H3K4me3. We used the random forest to perform non-linear regressions and found the combination of CG content and H3K27ac from H1-hESC cell line had the high correlation to the weighted feature (Figure 2C, Supplemental Table S7).

Identification of motifs and regulatory grammar of the promoters

The largest contributions came from the bins around the TSS, ranging from activating to repressing transcription consistent with the gene expression levels. This result indicates that the major sequence difference between variant groups is largely around the TSS including the core promoter region. We used the motif finding software STREME (36) to find motifs in each group and each 200 bp bin defined in model CR-1 (200 bp provided a sufficient segment for motif finding) with a p-value cutoff of 0.001 (Figure 2B, Supplementary Figure S11B for CR-2). We found 4003 motifs in total. To remove redundant motifs, we used Tomtom (37) to determine the similarities between all pairs of motifs (Figure 3A and B). Using a *P*-value cutoff of 10^{-5}, we extracted 76 unique motifs.

We analyzed the distribution of these motifs according to the groups or bins identified by our CR models (Supplementary Table S5 and Table S6). There are 24 motifs preferred in the highly-expressed groups, in which 8 of them are TF-associated. On the other hand, there are 6 motifs preferred in the lowly-expressed groups, in which 3 of them are TF-associated. In Table S6, we also noticed that there are 6 exclusive motifs in the 34th and 35th positively-weighted bins, indicating that they are strongly associated with activating gene expression. And there are 3 exclusive motifs in the 31st negatively-weighted bin, indicating that they are associated with repressing the gene expression.

Then, we compared these 76 motifs with the known ones in HOCOMOCO v11 (40) as well as the DNA motifs that are associated with histone modifications (361 motifs) (41), and DNA methylation (313 motifs) (42) using Tomtom with an E-value cutoff of 0.05 (Supplemental Table S4). The 27 matched known motifs include TFs, such as well-known promoter binding factors of TBP and TAF1 (43) as well as SP4 and EGR1, which have been reported to regulate gene expression by binding to the CG rich promoters (44). The majority (55 motifs) of the 76 motifs matched with motifs associated with histone modifications (42 motifs) (41) and DNA methylation (32 motifs) (42) (see details in Supplemental Table S4). The matched histone motifs are associated with histone modifications of H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3 and H3K9me3. This observation highlights the importance of epigenetic modifications and the factors involved in establishing or maintaining these modifications on regulating gene expression.

Next, we investigated whether there is any distance constraint between the co-occurring motifs. We used FIMO (38) to find all motifs' occurrence sites and calculated the cosine similarities with different lags for all pairs of motifs (Supplementary Figure S13). For a pair of motifs, if their average cosine similarity is larger than 0.7 and co-occurrence frequency is larger than a quarter of the total sequence numbers in the group, they were considered to tend to co-occur. Figure 3C–E and Supplementary Figure S16 show that these motifs form three large communities in which the motifs are densely connected to one another. The genome ontology term enrichment analysis was performed by using GOMO for the motifs in these three communities. We found that there are two communities showing enriched GO terms. One of them is associated with 'translational elongation' and 'ribosome', and another is associated with 'transcription' (the full results have been uploaded to GitHub).

For the 64 pairs of motifs that show high cosine similarity (Figure 3C–E), we checked whether they tend to co-occur with certain distance constraint. Figure 4A shows that more than half of them prefer to co-occur in the same 200 bp-bin and interestingly these bins avoid the region around TSS (Supplementary Figure S14). To further reveal distance constraint rules on base-pair resolution, we checked 33 motif pairs preferring to co-occur in the same 200 bp-bin. We found that most of the motif pairs do show preferred distance spacing between their occurrences (Figure 4B, Supplemental Table S8). For example, the distance between motif 17 (AGTGCARTGGYGYGA) and motif 41 (GCTCACTGCAASCTC) prefers to be −20 bp (accounting for 96% of all the occurrences). Another example is motif 1 (CCAGCCTGGSCRACA) and motif 2 (CCTCRGCCTCCCRAR) that mostly prefer to be 47 or 45 bp apart (accounting for 42% of all the pairs).

A trade-off between prediction accuracy and interpretability of deep-learning methods

Compared with other deep-learning methods, such as Enformer (45), Expecto (46) and Basenji (47), the gene expression prediction correlations of CR models are slightly lower (0.750, 0.709, 0.699 of CR-1/2/3 versus 0.85, 0.819, 0.85 of Enformer, Expecto and Basenji), which is likely due to much shorter input sequences to the CR models (10.5 kb promoter sequences in CR-1, 800 bp in CR-2, and 400 bp in CR-3, while 200 kb in Enformer, 40 kb in Expecto, and 131 kb in Basenji to consider both promoter and enhancer). However, our models are interpretable and can uncover regulatory direction (active or repressive) of promoter regions, key motifs and regulation grammars such as motif cooperation and distance constraint between motif sites.

We also compared with other interpretable methods, such as DeepLIFT (48) and GKMexplain (49) that can detect motifs. We have tested them by performing a clustering task based on feature scores, as our CR models have done in Figure 2, Supplementary Figure S10, and Supplementary Figure S11. As shown in Supplementary Figure S17, we found that DeepLIFT failed to cluster sequences to different gene expression levels, while GKMexplain cannot reach a higher prediction accuracy using an SVM model. Additionally, both DeepLIFT and GKMexplain are much slower than CR, and it is hard to use them for larger database and longer sequences.

Discussion

In this study, we trained interpretable neural network models based on the contextual regression (CR) framework. These CR models can predict gene expression from DNA sequences and reveal the key features by using the contextual weight. Their interpretability is reflected in the identification of active or repressive regions. This identification process can be done on DNA sequences of varying lengths. Thus, we can not only narrow down the region of performing motif extraction, but

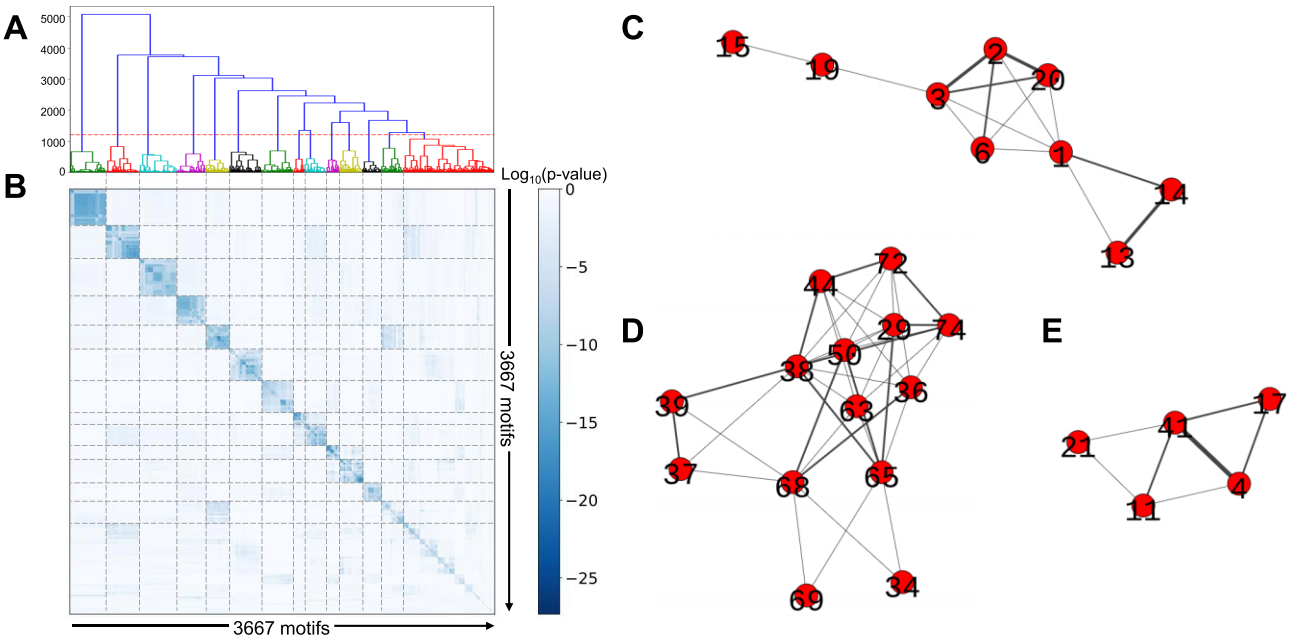

Figure 3. (**A**) The hierarchical clustering of similarity vectors of 3667 motifs by using the Ward variance minimization algorithm. The red dashed line is the distance threshold of 1200. (**B**) The similarities between all pairs of motifs were calculated using Tomtom. (**C–E**) The motif communities for the samples in the group with the highest expression level. The whole version of the figure is shown in Supplementary Figure S13. The layout was generated using the Fruchterman–Reingold force-directed algorithm and the width of the edge represents the cosine similarity score.

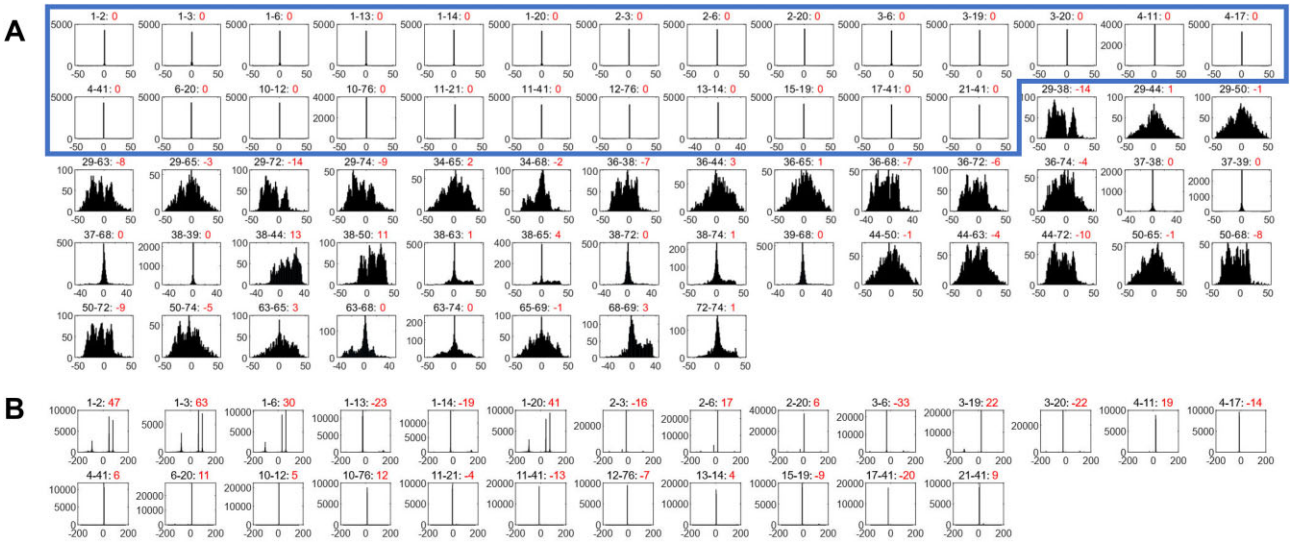

Figure 4. (**A**) The lag distribution for the 64 co-occurring motif pairs. The numbers in red are the median values of the lag. The blue box includes the 25 motif pairs that prefer to co-occur in the same 200 bp-bin. (**B**) The distance constraint for the 25 co-occurring motif pairs. The numbers in red are the median values of distances (bp).

also analyze the active or repressive effect at the base level by CR-7-G and CR-7-S, as shown in Supplementary Figure S15. The first model CR-1, with a resolution of 200 bp, found several stripes that significantly contribute to gene expression levels, either actively or repressively. These stripes are related to the non-linear combination of CpG signals and H3K27ac signals. In particular, the regions around TSS show the most distinction between highly and lowly expressed genes. We thus built higher resolution models (CR-2 to CR-6) zoomed into the regions around TSS to illustrate the most important contributing sequences. A repeating observation is that the promoter sequences are not equally important for gene expression, suggesting the importance of the underlying promoter sequences in regulating transcription. While the overall performance of our model is not as good as that reported in the literature, the difference is marginal. More importantly, this small sacrifice is made in exchange for a huge improvement in model interpretability. The CR model tells us the active or repressive regions, important motifs, and regulation grammars.

By combining CR-1 to CR-6 models with increasing resolution, we found the promoter regions around TSS have the most distinction between highly and lowly expression genes.

Using the contextual weight profiles, we could cluster all the genes into 7 groups with expression levels ranging from high to low, suggesting that the contextual weight reflect the sequence features associated with transcriptional regulation.

An interesting observation is that the CR-7-G model trained on the genomic DPR regions (+17 bp to +35 bp relative to TSS) could not predict gene expression well while the gene expression levels of the synthetic DPR sequences could be accurately predicted, suggesting the necessity of promoter sequences beyond the DPR in the genomic promoters on regulating transcription. Our analysis revealed that the genomic and synthetic sequences differ in multiple ways such as CG content and preference of certain bases in some positions, suggesting possible features lacked in the genomic DPRs for controlling gene expression.

We next discovered 76 unique motifs important for predicting gene expression, among which 27 are matched with known TF motifs including those important for transcription such as TBP and TAF1. Interestingly, 55 out of the 76 motifs (72%) are matched with epigenetic motifs including 42 matched with histone associated motifs, and 32 with DNA methylation associated motifs. While the epigenetic motifs are supposed to be associated with establishing or maintaining epigenetic modifications and their importance in regulating gene expression is not unexpected, their dominance in the 76 unique motifs is still surprising and encourages future studies of the underlying mechanisms.

Our analysis also revealed the motif combination grammars including three motif communities and distance constraint rules. The three communities represent possible collaboration between a set of regulatory proteins. There are 64 pairs of motifs with a high co-occurrence frequency, and about half of them have preferred distance spacing in the same 200 bp-bin. This observation indicates strong cooperation between the regulatory proteins binding to the promoters.

Data availability

The code and intermediate analysis data sets generated in this study are available at GitHub (https://github.com/swang066/CR-for-Promoter), Zenodo (https://doi.org/10.5281/zenodo.11157639) and as Supplemental File.

Supplementary data

Supplementary Data are available at NARGAB online.

Acknowledgements

We would like to thank Mr Chengyu Liu, Dr Lina Zheng and Dr Richard Ainsworth for their helpful comments and discussions. This work was partially supported by NIH (R01HG009626 to W.W.).

Funding

NIH [R01HG009626].

Conflict of interest statement

None declared.

References

1. Huminiecki,Ł. and Horbańczuk,J. (2017) Can we predict gene expression by understanding proximal promoter architecture? *Trends Biotechnol.*, **35**, 530–546.
2. Sanchez,A., Garcia,H.G., Jones,D., Phillips,R. and Kondev,J. (2011) Effect of promoter architecture on the cell-to-cell variability in gene expression. *PLoS Comput. Biol.*, **7**, e1001100.
3. Tang,H., Wu,Y., Deng,J., Chen,N., Zheng,Z., Wei,Y., Luo,X. and Keasling,J.D. (2020) Promoter architecture and promoter engineering in Saccharomyces cerevisiae. *Metabolites*, **10**, 320.
4. Aysha,J., Noman,M., Wang,F., Liu,W., Zhou,Y., Li,H. and Li,X. (2018) Synthetic promoters: designing the cis regulatory modules for controlled gene expression. *Mol. Biotechnol.*, **60**, 608–620.
5. Ngoc,L.V., Huang,C.Y.J., Cassidy,C.J., Medrano,C. and Kadonaga,J.T. (2020) Identification of the human DPR core promoter element using machine learning. *Nature*, **585**, 459–463.
6. Thonpho,A., Rojvirat,P., Jitrapakdee,S. and MacDonald,M.J. (2013) Characterization of the distal promoter of the human pyruvate carboxylase gene in pancreatic beta cells. *PLoS One*, **8**, e55139.
7. White,M.A., Kwasnieski,J.C., Myers,C.A., Shen,S.Q., Corbo,J.C. and Cohen,B.A. (2016) A simple grammar defines activating and repressing cis-regulatory elements in photoreceptors. *Cell Rep.*, **17**, 1247–1254.
8. King,D.M., Hong,C.K.Y., Shepherdson,J.L., Granas,D.M., Maricque,B.B. and Cohen,B.A. (2020) Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. *eLife*, **9**, e41279.
9. Sinha,S., Adler,A.S., Field,Y., Chang,H.Y. and Segal,E. (2008) Systematic functional characterization of cis-regulatory motifs in human core promoters. *Genome Res.*, **18**, 477–488.
10. Fan,K.L., Moore,J.E., Zhang,X.O. and Weng,Z.P. (2021) Genetic and epigenetic features of promoters with ubiquitous chromatin accessibility support ubiquitous transcription of cell-essential genes. *Nucleic Acids Res.*, **49**, 5705–5725.
11. Whitfield,T.W., Wang,J., Collins,P.J., Partridge,E.C., Aldred,S.F., Trinklein,N.D., Myers,R.M. and Weng,Z.P. (2012) Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.*, **13**, R50.
12. Xie,D., Boyle,A.P., Wu,L., Zhai,J., Kawli,T. and Snyder,M. (2013) Dynamic trans-acting factor colocalization in human cells. *Cell*, **155**, 713–724.
13. Xiang,G.J., Keller,C.A., Heuston,E., Giardine,B.M., An,L., Wixom,A.Q., Miller,A., Cockburn,A., Sauria,M.E.G., Weaver,K., *et al.* (2020) An integrative view of the regulatory and transcriptional landscapes in mouse hematopoiesis. *Genome Res.*, **30**, 472–484.
14. Levo,M., Avnit-Sagi,T., Lotan-Pompan,M., Kalma,Y., Weinberger,A., Yakhini,Z. and Segal,E. (2017) Systematic investigation of transcription factor activity in the context of chromatin using massively parallel binding and expression assays. *Mol. Cell*, **65**, 604–617.
15. Weingarten-Gabbay,S. and Segal,E. (2014) The grammar of transcriptional regulation. *Hum. Genet.*, **133**, 701–711.
16. de Boer,C.G., Vaishnav,E.D., Sadeh,R., Abeyta,E.L., Friedman,N. and Regev,A. (2020) Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.*, **38**, 56–65.
17. Cheng,Y., Ma,Z., Kim,B.-H., Wu,W., Cayting,P., Boyle,A.P., Sundaram,V., Xing,X., Dogan,N. and Li,J. (2014) Principles of regulatory information conservation between mouse and human. *Nature*, **515**, 371–375.
18. de Jongh,R.P., van Dijk,A.D., Julsing,M.K., Schaap,P.J. and de Ridder,D. (2020) Designing eukaryotic gene expression regulation using machine learning. *Trends Biotechnol.*, **38**, 191–201.
19. van Arensbergen,J., FitzPatrick,V.D., de Haas,M., Pagie,L., Sluimer,J., Bussemaker,H.J. and van Steensel,B. (2017) Genome-wide mapping of autonomous promoter activity in human cells. *Nat. Biotechnol.*, **35**, 145–153.

20. Perez-Pinera,P., Ousterout,D.G., Brunger,J.M., Farin,A.M., Glass,K.A., Guilak,F., Crawford,G.E., Hartemink,A.J. and Gersbach,C.A. (2013) Synergistic and tunable human gene activation by combinations of synthetic transcription factors. *Nat. Methods*, **10**, 239–242.
21. Haberle,V., Arnold,C.D., Pagani,M., Rath,M., Schernhuber,K. and Stark,A. (2019) Transcriptional cofactors display specificity for distinct types of core promoters. *Nature*, **570**, 122–126.
22. Meyer,P., Siwo,G., Zeevi,D., Sharon,E., Norel,R., Segal,E., Stolovitzky,G. and Consort,D.P.P. (2013) Inferring gene expression from ribosomal promoter sequences, a crowdsourcing approach. *Genome Res.*, **23**, 1928–1937.
23. Won,K.J., Sandelin,A., Marstrand,T.T. and Krogh,A. (2008) Modeling promoter grammars with evolving hidden Markov models. *Bioinformatics*, **24**, 1669–1675.
24. Juven-Gershon,T. and Kadonaga,J.T. (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol.*, **339**, 225–229.
25. Nagy,G. and Nagy,L. (2020) Motif grammar: the basis of the language of gene expression. *Comput. Struct. Biotec.*, **18**, 2026–2032.
26. Vaishnav,E.D., de Boer,C.G., Molinet,J., Yassour,M., Fan,L., Adiconis,X., Thompson,D.A., Levin,J.Z., Cubillos,F.A. and Regev,A. (2022) The evolution, evolvability and engineering of gene regulatory DNA. *Nature*, **603**, 455–463.
27. Agarwal,V. and Shendure,J. (2020) Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Rep.*, **31**, 107663.
28. Zheng,A., Lamkin,M., Zhao,H., Wu,C., Su,H. and Gymrek,M. (2021) Deep neural networks identify sequence context features predictive of transcription factor binding. *Nat. Mach. Intell.*, **3**, 172–180.
29. Zrimec,J., Börlin,C.S., Buric,F., Muhammad,A.S., Chen,R., Siewers,V., Verendel,V., Nielsen,J., Töpel,M. and Zelezniak,A. (2020) Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat. Commun.*, **11**, 6141.
30. Andersson,R. and Sandelin,A. (2020) Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.*, **21**, 71–87.
31. Talukder,A., Barham,C., Li,X. and Hu,H. (2021) Interpretation of deep learning in genomics and epigenomics. *Brief. Bioinform.*, **22**, bbaa177.
32. Liu,C., Liu,Y.-C., Huang,H.-D. and Wang,W. (2019) Biogenesis mechanisms of circular RNA can be categorized through feature extraction of a machine learning model. *Bioinformatics*, **35**, 4867–4870.
33. Liu,C. and Wang,W. (2017) Contextual regression: an accurate and conveniently interpretable nonlinear model for mining discovery from scientific data. arXiv doi: https://arxiv.org/abs/1710.10728, 30 October 2017, preprint: not peer reviewed.
34. Abadi,M., Barham,P., Chen,J., Chen,Z., Davis,A., Dean,J., Devin,M., Ghemawat,S., Irving,G. and Isard,M. (2016) In: *12th USENIX symposium on Operating Systems Design and Implementation (OSDI 16)*. pp. 265–283.
35. Glorot,X. and Bengio,Y. (2010) In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, pp. 249–256.
36. Bailey,T.L. (2021) STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, **37**, 2834–2840.
37. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.
38. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
39. Buske,F.A., Boden,M., Bauer,D.C. and Bailey,T.L. (2010) Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics*, **26**, 860–866.
40. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Sharipov,R.N., Fedorova,A.D., Rumynskiy,E.I., Medvedeva,Y.A., Magana-Mora,A., Bajic,V.B. and Papatsenko,D.A. (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, **46**, D252–D259.
41. Ngo,V., Chen,Z., Zhang,K., Whitaker,J.W., Wang,M. and Wang,W. (2019) Epigenomic analysis reveals DNA motifs regulating histone modifications in human and mouse. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 3668–3677.
42. Wang,M.C., Zhang,K., Ngo,V., Liu,C.Y., Fan,S.C., Whitaker,J.W., Chen,Y., Ai,R.Z., Chen,Z., Wang,J., *et al.* (2019) Identification of DNA motifs that regulate DNA methylation. *Nucleic Acids Res.*, **47**, 6753–6768.
43. Anandapadamanaban,M., Andresen,C., Helander,S., Ohyama,Y., Siponen,M.I., Lundstrom,P., Kokubo,T., Ikura,M., Moche,M. and Sunnerhagen,M. (2013) High-resolution structure of TBP with TAF1 reveals anchoring patterns in transcriptional regulation. *Nat. Struct. Mol. Biol.*, **20**, 1008–1014.
44. Maag,J.L., Kaczorowski,D.C., Panja,D., Peters,T.J., Bramham,C.R., Wibrand,K. and Dinger,M.E. (2017) Widespread promoter methylation of synaptic plasticity genes in long-term potentiation in the adult brain in vivo. *BMC Genomics [Electronic Resource]*, **18**, 250.
45. Avsec,Z., Agarwal,V., Visentin,D., Ledsam,J.R., Grabska-Barwinska,A., Taylor,K.R., Assael,Y., Jumper,J., Kohli,P. and Kelley,D.R. (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, **18**, 1196–1203.
46. Zhou,J., Theesfeld,C.L., Yao,K., Chen,K.M., Wong,A.K. and Troyanskaya,O.G. (2018) Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.*, **50**, 1171–1179.
47. Kelley,D.R., Reshef,Y.A., Bileschi,M., Belanger,D., McLean,C.Y. and Snoek,J. (2018) Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.*, **28**, 739–750.
48. Shrikumar,A., Greenside,P. and Kundaje,A. (2017) Learning important features through propagating activation differences. In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. **70**, pp. 3145–3153.
49. Shrikumar,A., Prakash,E. and Kundaje,A. (2019) GkmExplain: fast and accurate interpretation of nonlinear gapped k-mer SVMs. *Bioinformatics*, **35**, I173–I182.

Received: September 19, 2023. **Revised:** April 8, 2024. **Editorial Decision:** May 6, 2024. **Accepted:** May 12, 2024