Journal of Information Systems Engineering and Management
2025, 10(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

EPI-PRED: Leveraging CNN and Bi-LSTM Models for Accurate Prediction of Enhancer-Promoter Interactions

Ms.G.Aruna Arumugam*1, Dr.M.Mohamed Divan Masood*2

[1]Research Scholar, Department of Computer Applications, B.S.Abdur Rahman Crescent Institute of Science and Technology, Vandalur, Tamilnadu, India. arunanet23@gmail.com(*First Author)

2Assistant Professor, Department of Computer Applications, B.S.Abdur Rahman Crescent Institute of Science and Technology, Vandalur, Tamilnadu, India. divan@crescent.education (*Corresponding Author)

ARTICLE INFO

Received: 12 Oct 2024
Revised: 11 Dec 2024
Accepted: 24 Dec 2024

ABSTRACT

The identification of Enhancers-promoter interactions (EPI's) will assist in understanding the genetic regulation mechanisms. EPIs are determined through time-consuming and laborious testing techniques. Several methods are being contributed to deal with this issue. Due to their promising ability to predict, DL-based techniques have been extensively employed in the genome-scale detection of EPIs recently. This study is to employ the CNN and BiLSTM model named as "EPI-PRED" in predicting the EPI's with a set of features trained using the DL models in a python platform and it is successful predicted and evaluated with the SEPT, EPI-Trans and TF-EPI models using sensitivity, Specificity , precision and accuracy . It outperforms the other state of art DL methods and offers a new avenues in medical research.

Keywords: Enhancers-promoter interactions, EPI-PRED .

INTRODUCTION

The regulation of gene expression by enhancer-promoter interactions (EPIs) is crucial for numerous cellular processes. Technological advances in chromatin structure capture have allowed for the profiling of various three-dimensional (3D) structures throughout the genome, even in individual cells. However, the current catalogs of 3D structures are still unreliable and incomplete due to differences in technology and tools and data resolution. These methods usually use DNA sequencing data (k-mers, Transcription Factor Binding Site (TFBS) motifs, etc.), genome annotation data (ChIP-seq, DNAse-seq, etc.), and other genomic properties to find the relationships between genomic features and chromatin interactions[1]. . Enhancers and promoter interactions have consequently emerged as a crucial field of study. In addition to being essential for the initiation and regulation of genes, these interactions provide information about the effects of the three-dimensional arrangement of DNA in the nucleus on the genetic information that cells receive and process [2].

The vast amount of data generated by current technological advances has rendered it feasible to develop sequence-based deep learning algorithms that link DNA structures to the biochemical processes and regulatory aspects that affect the regulation of transcription. These frameworks can be utilized for modeling gene expression, epigenetic marks, and 3D genome organization in specific tissues and cell types It is possible to predict the functional effects of any non-coding alternative in the individual's genome—including rare or previously unknown variants—and systematically describe those effects in a way that goes beyond what can be determined by experiments or quantitative genetics studies alone. Key sequence patterns that are relevant to the estimated tasks have recently been discovered by the development and implementation of interpretation approaches. For example, even though the DNA sequences have not changed, the same pair of enhancer and promoter interactions exist in some cell lines but not in others . In order to tackle this problem, a number of models have been created that use epigenomic signals, such as chromatin accessibility, the binding of particular transcription factors, and

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

histone modification levels, to identify cell-line-specific EPIs. Machine learning techniques provide an alternative for obtaining missing 3D interactions and/or improving resolution. The specificity of the cell line dictates how the enhancer and promoter interact. This interaction is governed by different rules in different cell lines. As a result, a model created using one cell line might not work with another. Every cell line has a different model to train and test. These findings offer new perspectives on the basic biological processes that have been learned and offer opportunities for future enhancement of the models [4].

Still, with regard to of efforts, time, and resources, experimental methods to EPI classification are too costly. In order to solve such problems, an increasing amount of research is being done on the creation of computational methods, especially utilizing deep learning and other machine learning techniques. However, the vast majority of computing methods used today rely on convolutional neural networks, recurrent neural networks, or a combination of these, which ignore long-range interactions among enhancer as well as promoter sequences and contextual data. In this work, a novel transformer-based model named EPI-Trans is provided to overcome the issues in other DL models. The attention mechanism in transformer based model autonomously picks up characteristics that reflect the intricate connections.

The main contributions of the study are as follows

- To predict the EPIs from vast collection of genomic data.
- To test and validate the model performance using the performance metrics

Literature Study

The genome's enhancer–promoter interactions, or EPIs, are essential for controlling transcription. The work in [5] discusses a novel transformer-based model named EPI-Trans is introduced. The transformer model's multi-head attention mechanism autonomously picks up features that show the complex connections between enhancer and promoter patterns. Moreover, a transferable general designs is developed that can be used as a pre-trained model for numerous cell types. Additionally, to enhance efficiency, the generic algorithm's variables are adjusted using a specific cell line dataset. The main drawback in the transformer's model is that it cannot handle the biased training data and long tail phenomenon, Moreover; the attention mechanism suffers from memory complexities.CNN and transformer models is used to predict the enhancer –promoter's interactions in the gene regulation [6]. CNN and Bi-LSTM are used extract the features in the DNA Sequences which identifies the enhancers and its strength [7]. Bacterial Promoter regions were predicted using the ML algorithms like Support Vector machines(SVM), Random Forests(RF) and XGBoost[8] and the interpretability of the model is further improved Explainable Artificial Intelligence(XAI) with SHaply values.

Proformer, a transformer encoder architecture proposed in [9] is to forecast expression values based on DNA sequences. It is a Macaron-like Transformer encoder design, with a separate 1D convolution layer inserted following the first Feed -forward Layer(FFL) and in front of the multi-head attention layer. Each one encoder block had two half-step Feed-forward network (FFN) layers at its start and finish. A study in [10] introduces an EPInformer, an extensible DL technique framework that combines the chromatin interactions, epigenomic messages, and promoter-enhancer interactions to predict gene expression.

A CNN model which learns from both text and graph is suggested in [11] to leverage interactions with sequence features. Several Deep Neural Networks (DNN) models are developed to detect the regulatory elements [12], DNA enhancers [13] , regulatory variants in brain cells[14] and histone marks and predict gene expression[15].

A sequence-based technique known as SEPT that uses transfer learning (TL) and cross-cell information for predicting EPI's in new cell lines. It uses CNN to extract the features of enhancers and promoters from the DNA sequences. On the basis of labeled information from different cell lines, it is capable of

recognizing EPIs in an entirely novel cell line if the precise positions of enhancers and promoters are provided [16].

A novel computational method called "SPEID," employing DL methods is used to predict the enhancer-promoter interactions solely based on sequence-based features if one knows the locations of putative enhancers and promoters in any specific cell type[17] . Many functional genomic and epigenomic features are needed for the majority of machine learning methods currently in use, which restricts their applicability to particular cell lines. A random forest model, HARD (H3K27ac, ATAC-seq, RAD21, and Distance), is used in the study in [18] to predict EPI with just four different kinds of features. Long-range dependencies on the promoter and enhancer sequences are captured by the gated recurrent unit network, while local features are learned using a two-layer CNN. Second, features that are deemed relatively important are focused on using an attention mechanism. Lastly, a matching heuristic technique is presented to investigate the interaction between enhancers and promoters [19].

The study in [20] proposed an attention-based DL technique (Enhancer-LSTMAtt) and bi-directional long-short term memory (Bi-LSTM) for enhancer recognition. An end-to-end deep learning model called Enhancer-LSTMAtt primarily uses feed-forward attention, Bi-LSTM, and deep residual neural networks.

The methods discussed in the literature are detects the EPI's using the various aspects in the genome. Most of the techniques works well only for the training cells, but lacks in predicting the specific cell. The aim of the study is to improve accuracy of the model in predicting the enhancers and promoters.

Table 1: Recent Studies in EPI's Predictions

S.No	Author's	Methods used	Evaluation metrics	Score
1	Ahmed et al 2024[5]	Transformer based models	AUROC AUPR	AUROC Specific Model 94.2% Generic Models 95% Best Models 95.7% AURC Specific Model 80.5% Generic Models 66.1% Best Models 79.6%
2	Ni et al, 2022[6]	CNN and transformer model	AUROC AUPR	AUROC>90% AUPR>70%
3	Liao et al, 2022[7]	CNN, Bi-LSTM and transformer learning	Accuracy, Specificity, Sensitivity, Mathews Correlation Coefficient(MCC) AUROC AUPR	Proposed model outperforms the other advanced models
4	Paul .S et al, 2024[8]	SVM, RF, XGBoost	Accuracy, Precision, Recall, Specificity, F1-Score and MCC Metrics	F1-Score >95%
5	Tenekeci et al, 2024[11]	CNN	F1-Score	3 % higher than the state of art methods
6	Hu et al, 2024[13]	PDCNN model	Accuracy	95%
7	Lu et al, 2024[15]	CNN with Specialized Residual networks	Accuracy	Outperforms the compared models

		–CatLearning		
8	Jing et al, 2020[16]	CNN –SEPT	AUC	Effective and best prediction performance
9	Singh et al, 2019 [17]	SPEID, Deep learning models	AUROC, AUPR and F1.Score	Outperforms the compared methods
10	Zheng et al, 2023	Random Forest	Specificity, sensitivity, Precision, Accuracy AUC	Shows better performance in the chosen dataset.

The table 1 shows some of the recent studies carried out to predict the enhancer and promoter interactions, strength of enhancers, type of enhancers etc. The research is still going in improving the accuracy of the prediction models by modifying the existing models or combining the deep learning models. The dataset is also very limited and in depth knowledge about the DNA sequences is vital to study further exploration in this field. Thus, researching the interactions between enhancers and promoters can help us gain insight into both wellness and disease.

3. Materials and Methods

The method consists of the following steps (i) Data collection from a dataset.(ii)Choosing the epigenomic characteristics relevant to EPI. (iii) Sequence embedding (iv) feature extraction (v) Classifying and predicting the EPI[18]. The inputs are the enhancers and promoters as matrix and fed into the neural network layer for learning. It uses the already learned features to predict the EPI's .The Figure.1 depicts the overall flow of the EPI prediction model.

Figure 1. Architecture of EPI Prediction Model

3.1 Data Preparation

The data's are collected using the benchmark experimental datasets such as CRISPR , GM12878 and ChIA-PET . The duplicate values are then removed and selected RNA data from ChIA-PET and cell lines with a positive and negative samples in a ratio of 1:10. Then 39070 pairs of EPIs are obtained from the GM12878 dataset and 1735 pairs in the HeLa dataset and it is divide into training and testing set for GM 12878 sample. This study uses 70% of the data for training and 30% for testing. The experiments were implemented in python using the appropriate libraries. The Figure 2 depicts the detailed flow diagram of the EPI prediction model.

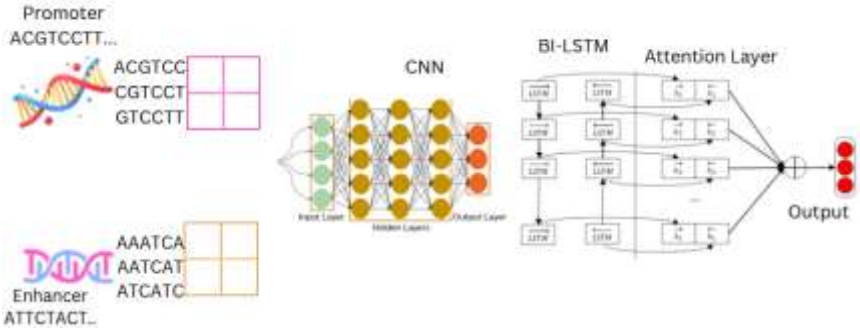

Figure 2: CNN-BiLSTM –Attention Mechanism for the prediction of EPI's

3.2 Sequence Embedding

An efficient technique for analyzing lengthy DNA fragments is to use the k-mer representations. This work utilizes a k-bp window with s as the sliding step size to distinguish the promoters and enhancers based on the k-mer representation. For instance, using k-mer representation, "AGCTGTTC" is subsequently divided into "AGCTGT," "GCTGTT," and "CTGTTC." It is evident that the k-mer representation is easy to calculate and recognize. The one -hot vector encoding suffers from the problem of dimensionality. Consequently, this method employs dna2vec embedding for expressing the DNA sequences using k-mer words. Dna2vec may generate high-quality, low-dimensional vectors that can represent k-mer words.

3.3 Feature Extraction

A combination network architecture with a bidirectional Long short term memory(Bi-LSTM) with attention mechanism and a CNN is used to extract the features, While Bi-LSTM was utilized to identify the long-term dependencies of local characteristics, CNN was utilized for learning the local characteristics of promoters and enhancers. Additionally, to calculate significant features that were assigned a higher weight to represent feature vectors, an attention layer was added[19]. The CNN model captures the unique features for the enhancer and promoter and then it gets combined into a merge layer.

3.4 CNN

CNN with two layers is employed : a max-pooling layer and a convolution layer. The max-pooling layer reduces feature dimensions, while the convolution layer mainly learns the local characteristics of enhancers and promoters. The two CNNs were set up in the experiment one for enhancers and another for promoters. Then the number of filters at 64, the stride at 30, the pooling length of the max-pooling layer at 30, and the filter length of the convolution layer at 60 for the enhancers. Then for the promoter set the stride at 20, the number of filters at 64, the pooling length of the max-pooling layer at 20, and the filter length of the convolution layer at 40

The same hyperparameter setting has to be followed for the comparison models.

3.5 Bi-LSTM

LSTM is a type of RNN suitable for time series data. It is mainly utilized to acquire the links between words in the former and later, but vice versa is not possible. The Bi-LSTM is used to solve this problem by using two LSTM's one for forward to backward and another one for backward to forward. The combined output of both LSTMs will generate the predictions [20].

3.6 Attention Mechanism

In a BiLSTM network with an attention mechanism uses the BiLSTM's implicit state to coordinate the current step's cell state with the input. It additionally makes use of the BiLSTM's most recent cell state. Prediction efficiency and accuracy can be improved by reducing insignificant data and highlighting pertinent data during the learning process. The BiLSTM network's attention layer's output is constructed as follows:

$$M = \tanh(X) \quad (1)$$

$$A = \text{softmax}(W_a M) \quad (2)$$

$$A = X\alpha^{T} \quad (3)$$

In equation (1), X is the matrix that denotes the selected features X=(X1,X2,X3,....Xt), Wa is the weight of the co-efficient matrix of the attention layer. T indicates the transpose operation. The algorithm for the EPI-PRED is as follows.

1. Start the procedure in the unbalanced data
2. Divide the data into training (70%) and testing (30%)
3. Augment the data to balance the dataset
4. Train the model using 10% of the data
5. Stop the CNN and Bi-LSTM with attention mechanisms of the model
6. Continue the training
7. Evaluate

Then perform cross validation and repeat the steps of the dataset splits.

4. Results and Discussions

The proposed model is compared using the same dataset and training process in the other existing models, such as SEPT, EPI-Trans, and TF-EPI. The following metrics were used to estimate the performance of the model. It includes Sensitivity, Specificity, Precision, Accuracy. These formulas for calculating the metrics are as follows,

$$Sensitivity\ (SN) = \frac{TP}{TP+FN} \quad (1)$$

$$Specificity(SP) = \frac{TN}{TN+FP} \quad (2)$$

$$Precision(PR) = \frac{TP}{TP+FP} \quad (3)$$

$$Accuracy(ACC) = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

Table 2. Comparison of Proposed model for EPI prediction

Method	SN	SP	PR	ACC
SEPT	0.677	0.923	0.745	0.780
EPI-TRANS	0.634	0.937	0.754	0.851
TF-EPI	0.644	0.945	0.762	0.879
EPI-PRED	**0.704**	**0.961**	**0.798**	**0.912**

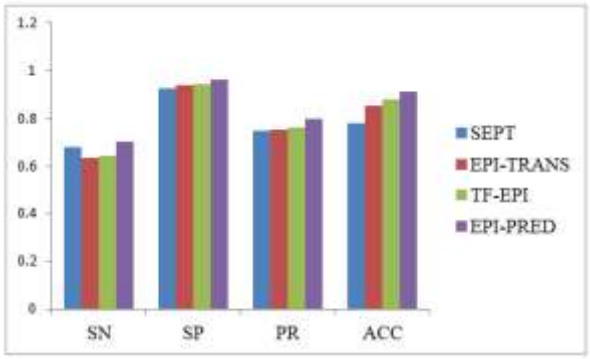

Figure 3 Evaluation of Proposed model

Table 2 and Figure 3 depicts the metrics values of the proposed classifier and other methods considered in this study .Results showed that the EPI-PRED outperforms the other three models. The EPI-PRED achieves a higher SN, SP , PR and ACC values of 0.704, 0.961,0.798,0.912 respectively. The proposed model combines the strength of CNN-BiLSTM and attention mechanisms in predicting the EPI's.Moreover, it is capable for handling the unbalanced data and can be used as one of the classifier.

Since only the site with the highest affinity is known, information regarding the transcription factor binding-site specificity is frequently lacking or skewed by prediction techniques. The learning techniques predict EPI using a vast number of genomic and epigenomic features. The redundancy in the features leads to imperfect outcomes. This study uses small number of epigenomic features to predict the cell line specific EPI's. Moreover, the EPI.s specificity is responsible for the differential expression of the gene.

One of the most captivating phenomena in gene regulation is the long-range interaction between enhancers and promoters. It is now possible to identify putative EPIs genome-wide because of to advances in high-throughput experimental techniques, but it is still unclear whether our genome already contains sequence-level instructions that aid in the identification of EPIs.

Continuous training of the network can be accomplished through data augmentation, which allows the recurrent layer to detect long-range dependencies among these characteristics and the convolutional layers to recognize useful subsequence features. However, classes are extremely unbalanced, identical to the original data, in typical applications of predicting interactions. When the network trained on augmented data is employed carelessly in such instances, the false positive rate is extremely high.

4.1 Enhancer Promotor Interaction (EPI) Score Calculations:

1. Affinity Score – It is used to measure how healthy the enhancer sequence brings into line with a promotor sequence (Enhancer-Promotor match is affinity). It's frequently proportionate to the alignment score (global or local) in a basic model. It can be calculated using the formula

$$Affinity = \frac{Alignment\ Score}{Length\ of\ Enhancer\ Sequence}$$

Where alignment score is obtained from the sequence alignment and the length is the distance of the enhancer sequence is used in the alignment.
If the affinity score is high, it indicates there is a strong interaction between promotor and enhancer region. i.e. the enhancer region is pointedly having regulatory effect on promotor which leads to higher gene expression. On the contrary if the score is low there is a weaker interaction.

Example:

```
ATGCGTACGTAGC-TA-G--CGTAGCT-A
   |||||||||| || |  |||| |  |
---CGTACGTAGCGTACGATCGTA-C-GA
  Score=19

Predicted Affinity Score: 19.0
```

2. Specificity Score – It is used to measure how healthy the enhancer aligns with the promotor sequence relative to the length of the promotor region.

$$Specificity = \frac{Alignment\ Score}{Length\ of\ Promotor\ Sequence}$$

Where alignment score is obtained from the sequence alignment and the length is the distance of the promotor sequence is used in the alignment.

If the specificity score is high, then it indicates the enhancer has the more targeted and stronger interaction with promotor region. Targeted interaction means enhancer exclusively regulate that precise promoter rather than consuming a wide-ranging, non-specific effect. Low specificity score implies less precise with particular promotor or enhancer interaction happens with multiple promotor region.

Example:

```
ATGCGTACGTAGCTAGCTAGCTAGCTA
   |||    |     |  | ||||
---CGT----A-----C--G-TAGC--
  Score=10

ATGCGTACGTAGCTAGCTAGCTAGCTA
   ||||         |  | ||||
---CGTA---------C--G-TAGC--
  Score=10


Specificity Score: 1.00
```

3. Fitness Score – It is the combined measure of affinity and specificity of the Enhancer promotor interaction. Fitness score provide the complete view of how healthy the interaction is.

$$Fitness = \frac{Affinity + Specificity}{2}$$

Where the affinity and specificity scores are already calculated using the above formula. By doing this we can measure the strength of the enhancer promotor interaction.

If the fitness score is high, it represents the well targeted and stronger enhancer promotor interaction. This implies that enhancer not only well binds with promotor but also do so in effective and specific means. Low score indicates weaker and less targeted interactions.

Example:

```
Best Alignment:
ATGCGTACGTAGCTAGCTAGCTAGCTA
   ||||||||||
---CGTACGTAGC--------------
  Score=8.5

Fitness Score: 0.85
```

Table 3. Parameters used to calculate Enhancer promotor interaction

Dataset	Scoring Parameters	Score Calculations
• *promotor_seq* • *enhancer_seq*	• *match_score:* Score for each matching base pair. • *mismatch_penalty:* Penalty for mismatches between the sequences. • *gap_open_penalty:* Penalty for opening a gap in the alignment. • *gap_extend_penalty:* Penalty for extending a gap in the alignment.	calculate_scores(promoter_seq, enhancer_seq) • *Affinity* (EP align) • *Specificity* (EP length match) • *Fitness* (Strength of interaction)

	Promoter	Enhancer	Affinity	Specificity	Fitness
0	ATGCGTACGTAGCTAGCTAGCTAGCTA	CGTACGTAGC	8.5	0.850000	0.229730
1	TACGTAGCTAGGCTAGCTAGCTAGCTA	GCTAGCTA	5.0	0.625000	0.142857
2	GCTAGCTAGCATGCGTACGTAGCTAGC	ATGCGTAC	3.5	0.437500	0.100000
3	GCTAGCGTACGCTAGCTAGCTGACTA	CGTAGCTA	2.5	0.312500	0.073529
4	ATGCGTAGCTGACTGCGTACGTAGCA	GCGTACGT	4.0	0.500000	0.117647
5	CGTACGTAGCTAGCTAGCTAGCATGC	TAGCTAGC	4.0	0.500000	0.117647
6	GCTAGCTAGCTAGCGTACGTAGCTAG	GTACGTA	1.5	0.214286	0.045455
7	ATGCGTACGTAGCTAGCTGACTGCTA	GCTAGCT	1.5	0.214286	0.045455
8	CGTACGTAGCTAGCTAGCATGCGTAG	CTAGCTA	1.5	0.214286	0.045455
9	GCTAGCTAGCTAGCATGCGTACGTAG	GCTAGCTA	5.5	0.687500	0.161765
10	TACGCTAGCTAGCATGCGTACGTAGC	GTAGCTA	0.0	0.000000	0.000000
11	CGTACGTAGCTAGCTGACTGCGTACG	CTAGCT	-1.0	-0.166667	-0.031250
12	GCTAGCTAGCGTAGCTAGCTGACTGC	TACGTAG	0.0	0.000000	0.000000
13	ATGCGTACGTAGCTAGCATGCGTACG	GCTAGC	-1.0	-0.166667	-0.031250
14	CGTACGTAGCTAGCTAGCTAGCTAGC	TAGCTA	-1.0	-0.166667	-0.031250
15	GCTAGCTAGCTAGCGTACGTAGCTAG	CTAGCT	-1.0	-0.166667	-0.031250
16	TACGCTAGCTAGCTAGCTAGCATGCG	TAGCT	-3.5	-0.700000	-0.112903
17	CGTACGTAGCTAGCTAGCGTACGTAG	GCTAGCT	1.5	0.214286	0.045455
18	GCTAGCTAGCTAGCATGCGTAGCTAG	TACGTA	-2.5	-0.416667	-0.078125
19	ATGCGTACGTAGCTAGCTAGCATGCG	CTAGCT	-1.0	-0.166667	-0.031250
20	CGTACGTAGCTAGCTAGCTAGCGTAG	GCTAGC	-1.0	-0.166667	-0.031250
21	GCTAGCTAGCGTACGTAGCTAGCTAG	TAGCT	-3.5	-0.700000	-0.112903
22	TACGCTAGCTAGCATGCGTACGTAGC	GCTAGC	-1.0	-0.166667	-0.031250

Figure 3 Result: Promoter-Enhancer Interactions Strength and Precision

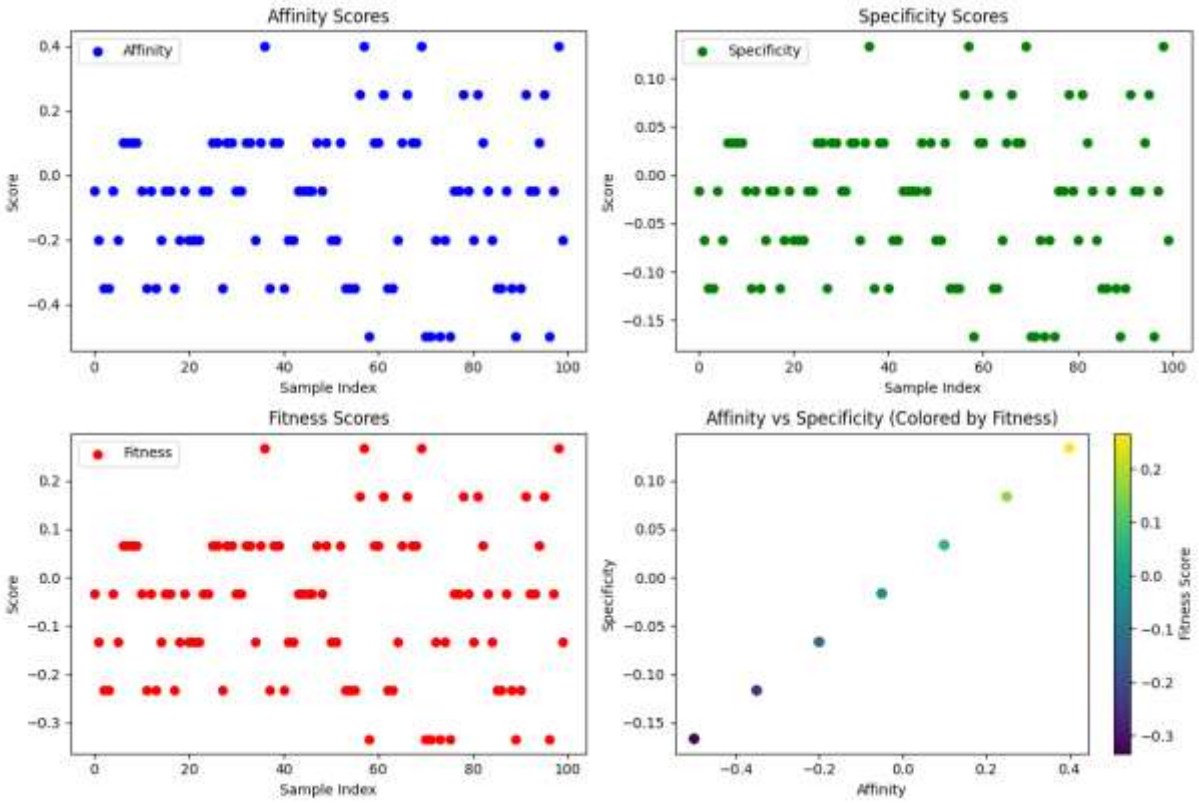

Figure 4 Promoter-Enhancer Interactions score calculations

5. Conclusion

The EPI predictions are a complex procedure and it is automated using the learning approaches. This study aims to increase the accuracy of predictions using deep learning models combining the strength of CNN, Bi-LSTM and attention mechanisms. The major drawback in this study is the lack of dataset with the proper sequence patterns. Moreover, it requires more preprocessing than the dataset used in other applications. Furthermore, the proposed algorithm has a considerable accuracy than present DL models, which suggests that it could be a helpful tool for rapid EPI predictions in genome-wide applications.

References

[1] Wall, B. P., Nguyen, M., Harrell, J. C., & Dozmorov, M. G. (2024). Machine and deep learning methods for predicting 3D genome organization. ArXiv.

[2] Liu, B., Zhang, W., Zeng, X., Loza, M., Park, S. J., & Nakai, K. (2024). TF-EPI: an interpretable enhancer-promoter interaction detection method based on Transformer. Frontiers in Genetics, 15, 1444459.

[3] Sokolova, K., Chen, K. M., Hao, Y., Zhou, J., & Troyanskaya, O. G. (2024). Deep Learning Sequence Models for Transcriptional Regulation. Annual Review of Genomics and Human Genetics, 25.

[4] Liu S, Xu X, Yang Z, Zhao X, Liu S, Zhang W: EPIHC: Improving enhancer-promoter interaction prediction by using hybrid features and communicative learning. IEEE/ACM Trans Comput Biol Bioinform 2022, 19:3435–3443 10.1109/TCBB.2021.3109488.

[5] Ahmed, F.S., Aly, S. & Liu, X. EPI-Trans: an effective transformer-based deep learning model for enhancer promoter interaction prediction. BMC Bioinformatics 25, 216 (2024). https://doi.org/10.1186/s12859-024-05784-9

[6] Ni, Y., Fan, L., Wang, M. et al. EPI-Mind: Identifying Enhancer–Promoter Interactions Based on Transformer Mechanism. Interdiscip Sci Comput Life Sci 14, 786–794 (2022). https://doi.org/10.1007/s12539-022-00525-z

[7] Liao, M., Zhao, Jp., Tian, J. et al. iEnhancer-DCLA: using the original sequence to identify enhancers and their strength based on a deep learning framework. BMC Bioinformatics 23, 480 (2022). https://doi.org/10.1186/s12859-022-05033-x.

[8] Paul, S., Olymon, K., Martinez, G. S., Sarkar, S., Yella, V. R., & Kumar, A. (2024). MLDSPP: bacterial promoter prediction tool using DNA structural properties with machine learning and explainable AI. Journal of Chemical Information and Modeling, 64(7), 2705-2719.

[9] Kwak, I. Y., Kim, B. C., Lee, J., Kang, T., Garry, D. J., Zhang, J., & Gong, W. (2024). Proformer: a hybrid macaron transformer model predicts expression values from promoter sequences. BMC bioinformatics, 25(1), 81.

[10] Lin, J., Luo, R., & Pinello, L. (2024). EPInformer: a scalable deep learning framework for gene expression prediction by integrating promoter-enhancer sequences with multimodal epigenomic data. bioRxiv, 2024-08.

[11] Tenekeci, S., & Tekir, S. (2024). Identifying promoter and enhancer sequences by graph convolutional networks. Computational Biology and Chemistry, 110, 108040.

[12] Chandrashekar, P. B., Chen, H., Lee, M., Ahmadinejad, N., & Liu, L. (2024). DeepCORE: An interpretable multi-view deep neural network model to detect co-operative regulatory elements. Computational and Structural Biotechnology Journal, 23, 679-687.

[13] Hu, W., Li, Y., Wu, Y., Guan, L., & Li, M. (2024). A deep learning model for DNA enhancer prediction based on nucleotide position aware feature encoding. *Iscience*, *27*(6).

[14] Zhou, J., Weinberger, D. R., & Han, S. (2024). Deep learning predicts DNA methylation regulatory variants in specific brain cell types and enhances fine mapping for brain disorders. *bioRxiv*.

[15] Lu, W., Tang, Y., Liu, Y., Lin, S., Shuai, Q., Liang, B., ... & Fang, D. (2024). CatLearning: highly accurate gene expression prediction from histone mark. *Briefings in Bioinformatics*, *25*(5).

[16] Jing F, Zhang S-W, Zhang S: Prediction of enhancer-promoter interactions using the cross-cell type information and domain adversarial neural network. BMC Bioinformatics 2020, 21:507. 10.1186/s12859-020-03844-4

[17] Singh, S., Yang, Y., Póczos, B., and Ma, J. (2019). Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quant. Biol.* 7, 122–137. doi:10.1007/s40484-019-0154-0

[18] Zheng L, Liu L, Zhu W, Ding Y and Wu F (2023) Predicting enhancer-promoter interaction based on epigenomic signals. *Front. Genet*. 14:1133775. doi: 10.3389/fgene.2023.1133775

[19] Xiaoping Min, Congmin Ye, Xiangrong Liu, Xiangxiang Zeng, Predicting enhancer-promoter interactions by deep learning and matching heuristic, *Briefings in Bioinformatics*, Volume 22, Issue 4, July 2021, bbaa254, https://doi.org/10.1093/bib/bbaa254

[20] Huang, G.; Luo, W.; Zhang, G.; Zheng, P.; Yao, Y.; Lyu, J.; Liu, Y.; Wei, D.-Q. Enhancer-LSTMAtt: A Bi-LSTM and Attention-Based Deep Learning Method for Enhancer Recognition. *Biomolecules* **2022**, *12*, 995. https://doi.org/10.3390/biom12070995

Ms. G. Aruna Arumugam completed her B.Sc. in Computer Science at Sarah Tucker College, Palayamkottai, Tirunelveli in 2008. She then obtained her MCA from Manonmaniam Sundaranar University, Tirunelveli in 2011, followed by an M.Phil. from St. Peter's University, Chennai in 2014. With 9 years of teaching experience, she is currently pursuing her Ph.D in the field of machine learning at B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai. Her areas of interest include artificial intelligence, deep learning, machine learning, computational biology, computer networks, and programming languages.

Dr. Mohamed Divan Masood M serves as an Assistant Professor in the Department of Computer Applications at the B.S. Abdur Rahman Crescent Institute of Science & Technology, Chennai a position he has held since February 2019. Dr. Masood earned his Ph.D. in Computer Science and Engineering from Anna University, Chennai, in 2018. He holds an M.Tech in Information Technology from B.S. Abdur Rahman Crescent Institute of Science & Technology, obtained in 2013, and a B.Tech in Information Technology from Anna University, Tirunelveli, completed in 2011.

Dr. Masood's research interests encompass a broad spectrum of cutting-edge areas, including Computational Biology, Artificial Intelligence, Machine Learning, Deep Learning, Big Data, and Social Network Analysis.