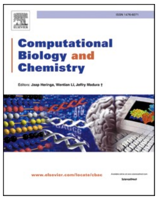
Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac



BERT-Promoter: An improved sequence-based predictor of DNA promoter using BERT pre-trained model and SHAP feature selection

Nguyen Quoc Khanh Le[a,b,c,*,1], Quang-Thai Ho[d,e], Van-Nui Nguyen[f], Jung-Su Chang[g,h]

[a] *Professional Master Program in Artificial Intelligence in Medicine, College of Medicine, Taipei Medical University, Taipei City 106, Taiwan*
[b] *Research Center for Artificial Intelligence in Medicine, Taipei Medical University, Taipei City 106, Taiwan*
[c] *Translational Imaging Research Center, Taipei Medical University Hospital, Taipei 110, Taiwan*
[d] *College of Information & Communication Technology, Can Tho University, Viet Nam*
[e] *Department of Computer Science and Engineering, Yuan Ze University, Chung-Li 32003, Taiwan*
[f] *University of Information and Communication Technology, Thai Nguyen University, Thai Nguyen, Viet Nam*
[g] *School of Nutrition and Health Sciences, College of Nutrition, Taipei Medical University, Taipei 110, Taiwan*
[h] *Graduate Institute of Metabolism and Obesity Sciences, College of Nutrition, Taipei Medical University, Taipei 110, Taiwan*

ARTICLE INFO

Keywords:
Promoter region
Contextualized word embedding
BERT multilingual cases
EXtreme Gradient Boosting
SHAP
Explainable artificial intelligence

ABSTRACT

A promoter is a sequence of DNA that initializes the process of transcription and regulates whenever and wherever genes are expressed in the organism. Because of its importance in molecular biology, identifying DNA promoters are challenging to provide useful information related to its functions and related diseases. Several computational models have been developed to early predict promoters from high-throughput sequencing over the past decade. Although some useful predictors have been proposed, there remains short-falls in those models and there is an urgent need to enhance the predictive performance to meet the practice requirements. In this study, we proposed a novel architecture that incorporated transformer natural language processing (NLP) and explainable machine learning to address this problem. More specifically, a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model was employed to encode DNA sequences, and SHapley Additive exPlanations (SHAP) analysis served as a feature selection step to look at the top-rank BERT encodings. At the last stage, different machine learning classifiers were implemented to learn the top features and produce the prediction outcomes. This study not only predicted the DNA promoters but also their activities (strong or weak promoters). Overall, several experiments showed an accuracy of 85.5 % and 76.9 % for these two levels, respectively. Our performance showed a superiority to previously published predictors on the same dataset in most measurement metrics. We named our predictor as BERT-Promoter and it is freely available at https://github.com/khanhlee/bert-promoter.

1. Introduction

A promoter is a sequence of DNA that initializes the process of transcription and regulates whenever and wherever genes are expressed in the organism (Le et al., 2019). In several eukaryotic genes, one of the most frequently promoter sequence is TATA box that is bound from transcription factors to cause the RNA polymerase transcription complex to form and promote transcription (Oubounyt et al., 2019). Promoters can range in length from 100 to 1000 base pairs. There are three types of promoters in eukaryotic cells including core promoters, proximal promoters, and distal promoters. They each play a unique role in DNA transcription and RNA polymerase. A widely studies have shown that DNA promoter malfunction is correlated with many human diseases, particularly cancer (Vlahopoulos et al., 2008), diabetes (Ionescu-Tîrgovişte et al., 2015), or Asthma (Hobbs et al., 1998).

Regarding their functional importance in facilitating gene expression, promoter identification and classification are currently an area of great interest and hot topics not only for biological research but also for computational research. In the initial studies, promoters were identified by wet experiments based on the biological and genetic characteristics,

[*] Corresponding author at: Professional Master Program in Artificial Intelligence in Medicine, College of Medicine, Taipei Medical University, Taipei City 106, Taiwan.

E-mail address: khanhlee@tmu.edu.tw (N.Q.K. Le).

[1] https://twitter.com/KhanhLe72426129

https://doi.org/10.1016/j.compbiolchem.2022.107732
Received 16 June 2022; Accepted 12 July 2022
Available online 14 July 2022

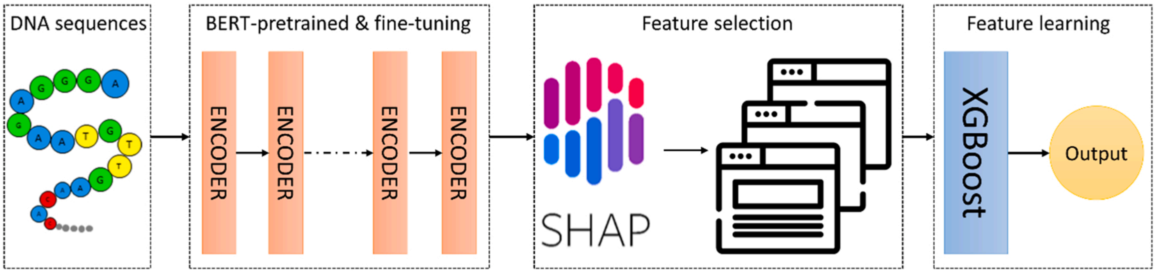


Fig. 1. A workflow of predicting DNA promoter and their strength from sequence. Input of model is DNA sequence, then the BERT-base multilingual-case pre-trained models are employed to extract the features from DNA sequence. SHAP analysis is then applied to select the important features generated from BERT model. Finally, XGBoost is used to learn the features and generate the output prediction.

such as chromatin immunoprecipitation assays (promoters combined with transcription factors) (Gade and Kalvakolanu, 2012), in vivo retroviral transduction (Rettinger et al., 1994), and retroviral plasmid library-based system (Khambata-Ford et al., 2003). However, these biological and genetic features covered only a portion of promoters because not all promoters are occupied by transcription factors. Moreover, most wet lab experiments were costly and time-consuming.

To address these issues, several computational models have been developed to early predict promoters over the past decade. For instance, Davuluri et al (Davuluri et al., 2001). developed a decision tree model to identify promoters and first exons in the human genome. Later, Kanhere and Bansal (Kanhere and Bansal, 2005) took advantage of DNA stability in providing more information to detect promoters correctly. Latter years were the exposure time for creating different promoter predictors with different sequencing features such as pseudo nucleotide composition (Lin et al., 2018), orthologous genomic (Solovyev and Shahmuradov, 2003), ENCODE regions (Bajic et al., 2006), or sequence alignment kernel (Gordon et al., 2003). Furthermore, the development of deep learning also contributed to the improvements of this promoter identification problem via long short-term memory (LSTM) (Oubounyt et al., 2019) and convolutional neural network (CNN) (Umarov and Solovyev, 2017). The above computational methods are mainly designed to discriminate promoters from other regulatory DNA fragments, but they lack the capability for promoter subgroup classification, i.e., the strength of promoters.

For that reason, Xiao et al. proposed a benchmark dataset and predictor not only DNA promoter identification but also promoter strength classification (Xiao et al., 2019). This dataset was then used in later studies aiming to improve the predictive performance. Different sequencing features have been proposed to improve the predictive performance such as heterogeneous features (Tayara et al., 2020), natural language processing (NLP) features (Le et al., 2019; Tahir et al., 2020), or position-specific of nucleotide composition (Lyu et al., 2020). Moreover, deep learning models have been also implemented to enhance the performance, i.e., CNN (Le et al., 2019) or LSTM (Oubounyt et al., 2019). In terms of the performances of the latest predictors, the best one has been achieved in the study of Le et al. (2019). with an accuracy of 85.41 % and 73.1 % for promoter identification and promoter strength classification, respectively. Although these computational predictors have yielded acceptable results, the prediction performance for promoter strength classification is still less than desirable, and there is still room for improvement. The purpose of this study is to address this problem.

2. Materials and methods

The main workflow of our predictor is illustrated in Fig. 1 and can be explained as follows: First, a Bidirectional Encoder Representations from Transformers (BERT) pre-trained model (BERT-cased multilingual) (Devlin et al., 2019) is employed to extract 768 × 81 = 62,208 features. Next, the dimension of features is reduced using SHapley Additive exPlanations (SHAP) analysis (Lundberg and Lee, 2017) before entering different machine learning algorithms. Hyperparameter tuning is conducted during the modeling to ensure the optimal model is reached. After that, the final model is used to predict DNA promoters as well as promoter's strength.

2.1. Benchmark dataset

In this study, we aimed to identify both promoters (classification: promoters or non-promoters) and their activity (classification: strong or weak). The training benchmark dataset collected by Liu's study (Xiao et al., 2019) was used to facilitate the comparison. This dataset was also used in the construction of other predictors such as (Le et al., 2019; Tayara et al., 2020). It was originally collected and verified at RegulonDB (Gama-Castro et al., 2015). In this dataset, each sequence fragment was divided into 81-bp fragments to accord with the biological features of nucleosome and linker DNA. After eliminating redundant sequences (via CD-HIT with a threshold of 0.85) and randomly selecting unbalance subsets, 3382 promoters (1591 strong promoter samples and 1791 weak promoter samples) and 3382 non-promoters were obtained. All of these samples were used in the next steps for feature extraction and model implementation.

2.2. Multilingual pre-trained BERT implementation

To build a robust and reliable bioinformatics tool, sufficient feature information should be incorporated into the model. In this study, BERT pre-trained model (Devlin et al., 2019) was employed for nucleotide sequence encoding. Unidirectionality is a significant issue for word display models, limiting the flexibility of architectures during pre-training. Thus, BERT was created to pre-train a bidirectional representation of the unlabeled text in all layers using a combined reverse and forward context. As a result, it can only be designed to create cutting-edge applications for a wide range of tasks recently (Devlin et al., 2019). This is a modern system of pre-training language representations that provides state-of-the-art performance on a wide variety of NLP tasks. Aiming at interpreting the information of DNA sequences, we also implemented BERT for our purpose.

Many researchers have utilized BERT to represent biological or biomedical information such as biomedical relation statement (Lai and Lu, 2020), biomedical text mining (Lee et al., 2020), peptide sequences (Charoenkwan et al., 2021), or even DNA sequences (Ji et al., 2021; Le et al., 2021a). Thus in this study, we aimed to take the advantages of this architecture to represent our DNA sequences to classify promoter regions. Specifically, since a lot of pre-trained BERT models have been released with different layers, we would like to use the latest one – the BERT-based multilingual cased pre-trained model. This model is the latest one and is recommended by the Google BERT team to achieve optimal performance in different NLP tasks. A strength of this pre-trained model is that it is built on the top 104 languages with the largest Wikipedia using a masked language modeling (MLM) objective.

Therefore, it could be used to interpret the information of different languages and in our opinion, it would be useful for DNA sequence representation also. This BERT model included 12 layers, 768-hidden, 12 heads, and 110,000,000 parameters. Because our input DNA sequences were at 81-bps length, the BERT model generated 81 × 768 = 62,208 features for each DNA sequence.

2.3. SHAP feature selection

After feature extraction, a common practice is to use all features as input to train a predictive model, but this may lead to dimensional disaster and information redundancy. To refine the distinctive features, we introduced our two-step framework, in which the Spearman correlation coefficient (SCC) and the cutting-edge SHAP values analysis (Lundberg and Lee, 2017) were taken into account:

(a) SCC > 0.8: The feature subtraction began with the calculation of SCC. Primarily, we would estimate the correlation rates between every two features. Via this way, features had high correlation (>0.8) with other features were excluded from our feature set. For the next stage, we statistically compared the frequency of each separated feature in the inclusion and exclusion set.

(b) SHAP analysis with SHAP cut-off values > 0: After the preliminary feature filtration by using the SCC of 0.8, the abundant number of kept features were fed into SHAP analysis to obtain the most contributed ones. SHAP was first developed by Lundberg et al., 2017 (Lundberg and Lee, 2017). SHAP values help in demonstrating the importance of each feature to the model's outcome (i.e., prediction). The application of SHAP values in machine learning has recently emerged as a new approach to measure the performance of any decision-tree-based model in many fields.

2.4. Machine learning implementation

The current study considered seven machine learning algorithms including k-Nearest Neighbors (kNN), Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF), Adaptive Boosting (AdaBoost), Multilayer Perceptron (MLP), and eXtreme Gradient Boosting (XGBoost) as the classifiers. These algorithms were applied because of their interpretability of important features in the decision-making process. We implemented these algorithms in *scikit-learn* (v 0.22.1) library in Python. Hyperparameters were tuned via 10-fold cross-validation on the corresponding part of training data, with classification accuracy as the optimization goal.

2.5. Quality measures

Our proposed model was evaluated using 10-fold cross-validation. The quality measures included sensitivity, specificity, accuracy, receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC). They were used to calculate how well a binary classification test correctly identifies (both promoter and non-promoter samples).

$$Sensitivity = \frac{True \quad Positive}{True \quad Positive + False \quad Negative} \tag{1}$$

$$Specificity = \frac{True \quad Negative}{True \quad Negative + False \quad Positive} \tag{2}$$

$$Accuracy = \frac{True \quad Positive + True \quad Negative}{Total \quad patients} \tag{3}$$

AUC is the area under the curve in which the x-axis denotes the false positive rate (FPR) and the y-axis denotes the true positive rate (TPR). The reason why we adopted AUC in evaluating our models is that AUC has been a widely used metric in evaluating binary classifiers without reliance on the decision threshold set on predicted class probability.

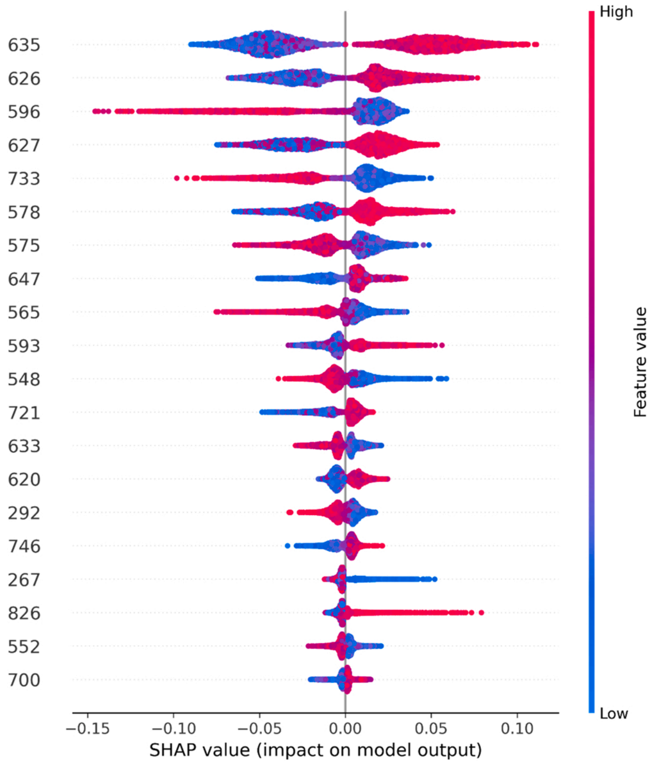


Fig. 2. Top 20 BERT features generated from SHAP analysis. There are some high-impact features at the end of BERT sequence features, i.e., from feature of 500.

3. Results and discussion

3.1. SHAP feature selection

From 62,208 BERT features, we used SCC and SHAP analysis (Lundberg and Lee, 2017) to find the important features of our machine learning model. After this step, we reduced the number of features to 653 features and this amount will be used in further analyses. The results in Fig. 2 showed that the most important BERT features came from the end of the feature list. It means that if we only selected a few features with the first BERT layers, we could not reach the best performance as well as not take all advantages of BERT models in learning this data. Furthermore, we would like to see the distribution of BERT features when plotting in a specific dimension. UMAP analysis then showed that our feature set could help to separate data points into two groups of promoters and non-promoters (Fig. 3A), promoter strong and weak (Fig. 3B). This analysis strongly claimed that we could find a significant feature set from whole BERT features to address this problem efficiently.

3.2. Assessment of different machine learning algorithms in predicting DNA promoter

In this step, we inserted 653 optimal BERT features into different machine learning algorithms to assess their performance. It could be observed that the XGBoost-based model (on BERT features) outperformed other classifier algorithms in both layers (accuracy of 85.5 % and 76.9 % for two layers, respectively) (Table 1). This was slightly higher than the 1st layer's accuracies scored by the other classifiers (improved at about 2–5 %) and was significantly better compared to the others for the 2nd layer (improved from 5–10 %). Thus, the XGBoost classifier was considered to validate and sort out the most significant out of the 653 primary features. Interestingly, these findings might be

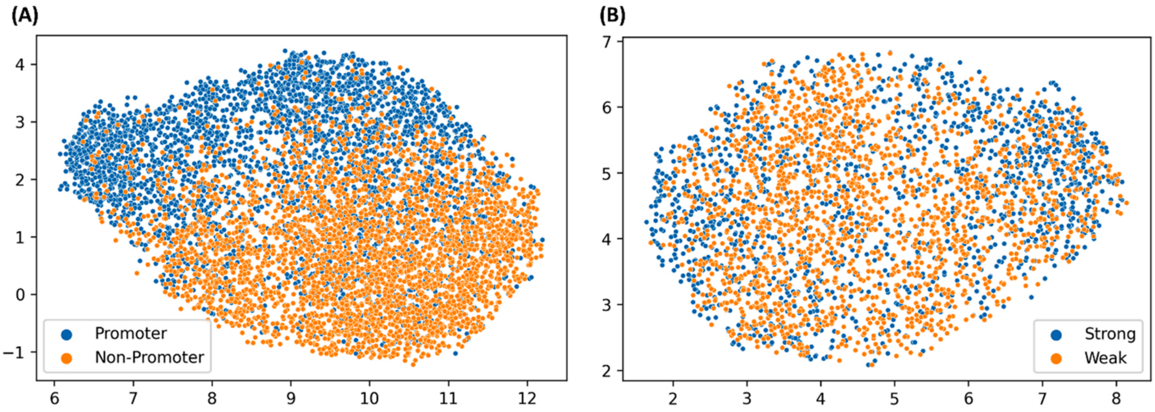


Fig. 3. UMAP projection of optimal feature representation. (A) 1st layer for promoter identification, (B) 2nd layer for promoter's strength classification.

Table 1
Assessment of different machine learning algorithms in predicting DNA promoter using BERT features.

	Promoter identification			Promoter's strength classification		
	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)
LR	80.5	83.8	82.1	61.7	72.4	67.8
kNN	76.0	84.0	80.0	52.5	77.7	66.7
RF	81.2	85.5	83.4	50.2	84.6	69.6
NB	79.9	86.6	83.2	70.8	72.5	71.8
AdaBoost	80.4	83.6	82.0	58.5	76.7	68.8
MLP	84.2	81.8	83.0	61.7	74.0	68.6
XGBoost	84.3	86.6	85.5	70.8	81.6	76.9

LR: Logistic Regression, kNN: k-Nearest Neighbors, RF: Random Forest, NB: Naïve Bayes, AdaBoost: Adaptive Boosting, MLP: Multilayer Perceptron, XGBoost: eXtreme Gradient Boosting.

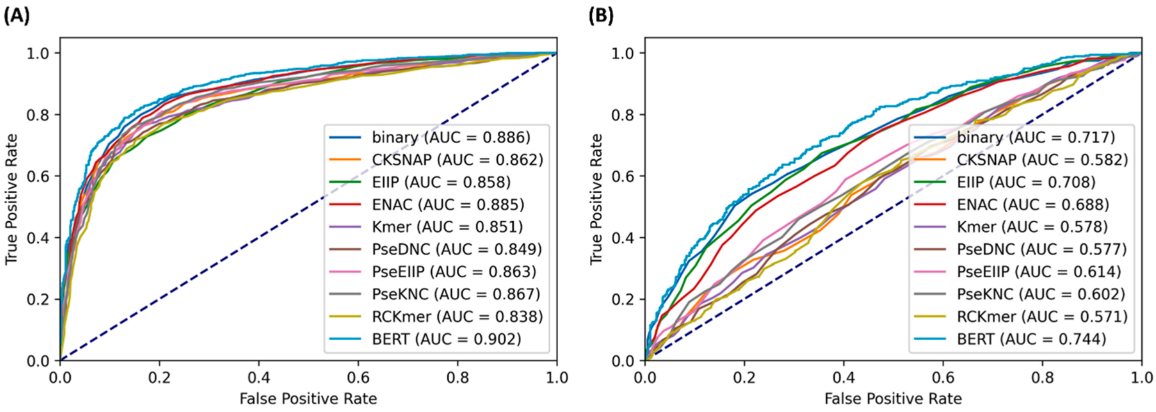


Fig. 4. Assessment of different features in predicting DNA promoter using sequence. (A) 1st layer for promoter identification, (B) 2nd layer for promoter's strength classification.

related to the previous bioinformatics studies where they claimed the efficiency of XGBoost in learning sequencing data (Do and Le, 2020; Le et al., 2021b). This observation may support the hypothesis that we did not need a deep network to learn such hand-craft features from the BERT model. A simple ensemble learning model could learn such features well without time-consuming to create a deep learning model.

3.3. Assessment of different sequence features in predicting DNA promoter

Nine sequence-based descriptors (e.g., binary, Composition of k-spaced Nucleic Acid Pairs (CKSNAP), Electron-ion interaction pseudopotentials of trinucleotide (EIIP), Enhanced Nucleic Acid Composition (ENAC), Kmer, Pseudo Dinucleotide Composition (PseDNC), Electron-ion interaction pseudopotentials of trinucleotide (PseEIIP), Pseudo k-tupler Composition (PseKNC), and Reverse Compliment Kmer (RCKmer)) were employed for sample feature encoding and compared to our BERT-based features. All of them are well-known sequencing features and they helped to achieve promising performance in a variety of bioinformatics studies (Chen et al., 2020). Fig. 4 shows the comparative performance among different features in identifying promoters and their activities. In terms of AUC values, it can be seen that a lot of encoding schemes achieved high performance, especially in promoter identification. Some traditional encodings were better in promoter strength classification such as binary, EIIP, or ENAC. However, an important finding is that our BERT features showed the best performance in both layer problems (AUC of 0.902 and 0.744 for 1st and 2nd layers, respectively). These results indicate that BERT feature is significantly

Table 2
Comparison to previously published predictors.

	Promoter identification			Promoter's strength classification		
	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)
iPSW(2L)-PseKNC	81.37	84.89	83.13	62.23	79.17	71.2
Le et al.	82.76	88.05	85.41	69.40	76.40	73.10
iPSW(PseDNC-DL)	83.34	86.83	85.1	65.81	78.16	72.35
Ours	84.34	86.56	85.45	70.85	81.63	76.92

superior for promoter identification when compared with the other well-known features in this field. This finding also supports evidence from previous observations on the use of BERT-based features in learning DNA (Ji et al., 2021; Le et al., 2021a) or peptide sequences (Charoenkwan et al., 2021). In addition, a novelty was added by using multilingual BERT-based pre-trained model that helped to boost the predictive performance rather than simple BERT models.

3.4. Comparison to previously published predictors

Since we used a benchmark dataset, our performance was compared to the other state-of-the-art predictors, including iPSW(2 L)-PseKNC (Xiao et al., 2019), Le et al (Le et al., 2019)., iPSW(PseDNC-DL) (Tayara et al., 2020) to claim our efficiency. Among the aforementioned predictors, iPSW(2 L)-PseKNC and iPSW(PseDNC) aimed to address the problem using state-of-the-art features according to nucleotide compositions. On the other hand, Le et al. tried to address this problem with the combination of deep learning and NLP-based fastText model. Table 2 provides details of the comparative analysis. It is observed that our proposed model is superior to the aforementioned methods in most measurement metrics, with accuracy improved at about 1–2 % and 3–5 % for promoter identification (the 1st layer) and for promoter activity classification (the 2nd layer), respectively. It can thus be suggested that the BERT-based pre-trained model was significant even with the use of a conventional machine learning model (XGBoost-based BERT model). Further research should be undertaken to investigate the potential of deep learning (Le et al., 2021a) in learning such BERT features.

4. Conclusion

The main goal of the current study was to present a two-layer prediction framework for identifying promoters and their activities. To build an efficient model, the feature representation learning scheme was employed using BERT pre-trained model and filtered using SHAP analysis. Subsequently, different machine learning algorithms were employed to assess the predictive performance and generate the final prediction model. The significant findings to emerge from this study are the potential of BERT-based multilingual NLP model in representing DNA sequences as well as the potential of XGBoost to learn such features. Compared with traditional feature representations, our BERT features showed significant performance improvement and model generalization. Furthermore, through a series of analyses, we saw that our predictor was superior to the state-of-the-art predictors in this field of promoter identification and classification. This approach will prove useful in expanding our understanding of how pre-trained NLP could fit biological sequences and solve bioinformatics problems with promising results.

Declaration of Competing Interest

The authors have no conflicts of interest to disclose.

Acknowledgements

This work has been supported by the Ministry of Science and Technology, Taiwan [MOST 110-2221-E-038-001-MY2] and [MOST 111-2628-E-038-002-MY3].

References

Bajic, V.B., et al., 2006. Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment. Genome Biol. 7 (1), S3.
Charoenkwan, P., et al., BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. Bioinformatics, 2021.
Chen, Z., 2020. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Brief. Bioinform. 21 (3), 1047–1057.
Davuluri, R.V., Grosse, I., Zhang, M.Q., 2001. Computational identification of promoters and first exons in the human genome. Nat. Genet. 29 (4), 412–417.
Devlin, J., et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1 (Long and Short Papers). 2019.
Do, D.T., Le, N.Q.K., 2020. Using extreme gradient boosting to identify origin of replication in Saccharomyces cerevisiae via hybrid features. Genomics 112 (3), 2445–2451.
Gade, P. , D.V. Kalvakolanu, Chromatin Immunoprecipitation assay as a tool for analyzing transcription factor activity. In: Vancura, A., (Ed.), Transcriptional Regulation: Methods and Protocols, 2012, Springer, New York, NY., 85–104.
Gama-Castro, S., et al., 2015. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res. 44 (D1), D133–D143.
Gordon, L., 2003. Sequence alignment kernel for recognition of promoter regions. Bioinformatics 19 (15), 1964–1971.
Hobbs, K., et al., 1998. Interleukin-10 and transforming growth factor- β promoter polymorphisms in allergies and asthma. Am. J. Respir. Crit. Care Med. 158 (6), 1958–1962.
Ionescu-Tîrgovişte, C., Gagniuc, P.A., Guja, C., 2015. Structural properties of gene promoters highlight more than two phenotypes of diabetes. PLoS One 10 (9) e0137950.
Ji, Y., et al., DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics, 2021.
Kanhere, A., Bansal, M., 2005. A novel method for prokaryotic promoter prediction based on DNA stability. BMC Bioinform. 6 (1), 1.
Khambata-Ford, S., et al., 2003. Identification of promoter regions in the human genome by using a retroviral plasmid library-based functional reporter gene assay. Genome Biol. 13 (7), 1765–1774.
Lai, P.-T., Lu, Z., 2020. BERT-GT: cross-sentence n-ary relation extraction with BERT and Graph Transformer. Bioinformatics 36 (24), 5678–5685.
Le, N.Q.K., et al., 2019. Classifying promoters by interpreting the hidden information of DNA sequences via deep learning and combination of continuous fasttext N-grams. Front Bioeng. Biotechnol. 7, 305.
Le, N.Q.K., et al., A transformer architecture based on BERT and 2D convolutional neural network to identify DNA enhancers from sequence information. Brief Bioinform, 2021a.
Le, N.Q.K., 2021b. A sequence-based prediction of Kruppel-like factors proteins using XGBoost and optimized features. Gene 787, 145643.
Lee, J., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36 (4), 1234–1240.
Lin, H., 2018. Identifying sigma70 promoters with novel pseudo nucleotide composition. IEEE/ACM Trans. Comput. Biol. Bioinform. 1-1.
Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
Lyu, Y., 2020. iPro2L-PSTKNC: a two-layer predictor for discovering various types of promoters by position specific of nucleotide composition. IEEE J. Biomed. Health Inf. 1-1.
Oubounyt, M., 2019. DeePromoter: robust promoter predictor using deep learning. Front. Genet. 10, 286.
Rettinger, S.D., et al., 1994. Liver-directed gene therapy: quantitative evaluation of promoter elements by using in vivo retroviral transduction. Proc. Natl. Acad. Sci. USA 91 (4), 1460–1464.
Solovyev, V.V., Shahmuradov, I.A., 2003. PromH: promoters identification using orthologous genomic sequences. Nucleic Acids Res. 31 (13), 3540–3545.
Tahir, M., et al., 2020. An intelligent computational model for prediction of promoters and their strength via natural language processing. Chemom. Intell. Lab. Syst. 202, 104034.

Tayara, H., Tahir, M., Chong, K.T., 2020. Identification of prokaryotic promoters and their strength by integrating heterogeneous features. Genomics 112 (2), 1396–1403.

Umarov, R.K., Solovyev, V.V., 2017. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. PLoS One 12 (2) e0171410.

Vlahopoulos, S.A., et al., 2008. The role of ATF-2 in oncogenesis. BioEssays 30 (4), 314–327.

Xiao, X., et al., 2019. iPSW(2L)-PseKNC: a two-layer predictor for identifying promoters and their strength by hybrid features via pseudo K-tuple nucleotide composition. Genomics 111 (6), 1785–1793.