BMC Genomics

RESEARCH

Open Access

iPromoter-Seqvec: identifying promoters using bidirectional long short-term memory and sequence-embedded features

Thanh-Hoang Nguyen-Vo[1], Quang H. Trinh[2], Loc Nguyen[1], Phuong-Uyen Nguyen-Hoang[3], Susanto Rahardja[4,5*] and Binh P. Nguyen[1*]

From The 20th Asia Pacific Bioinformatics Conference (APBC 2022)
Virtual. 26-28 April 2022

Abstract

Background: Promoters, non-coding DNA sequences located at upstream regions of the transcription start site of genes/gene clusters, are essential regulatory elements for the initiation and regulation of transcriptional processes. Furthermore, identifying promoters in DNA sequences and genomes significantly contributes to discovering entire structures of genes of interest. Therefore, exploration of promoter regions is one of the most imperative topics in molecular genetics and biology. Besides experimental techniques, computational methods have been developed to predict promoters. In this study, we propose iPromoter-Seqvec – an efficient computational model to predict TATA and non-TATA promoters in human and mouse genomes using bidirectional long short-term memory neural networks in combination with sequence-embedded features extracted from input sequences. The promoter and non-promoter sequences were retrieved from the Eukaryotic Promoter database and then were refined to create four benchmark datasets.

Results: The area under the receiver operating characteristic curve (AUCROC) and the area under the precision-recall curve (AUCPR) were used as two key metrics to evaluate model performance. Results on independent test sets showed that iPromoter-Seqvec outperformed other state-of-the-art methods with AUCROC values ranging from 0.85 to 0.99 and AUCPR values ranging from 0.86 to 0.99. Models predicting TATA promoters in both species had slightly higher predictive power compared to those predicting non-TATA promoters. With a novel idea of constructing artificial non-promoter sequences based on promoter sequences, our models were able to learn highly specific characteristics discriminating promoters from non-promoters to improve predictive efficiency.

Conclusions: iPromoter-Seqvec is a stable and robust model for predicting both TATA and non-TATA promoters in human and mouse genomes. Our proposed method was also deployed as an online web server with a user-friendly

*Correspondence: susantorahardja@ieee.org; binh.p.nguyen@vuw.ac.nz

[1] School of Mathematics and Statistics, Victoria University of Wellington, Gate 7, Kelburn Parade, 6140 Wellington, New Zealand
[4] School of Marine Science and Technology, Northwestern Polytechnical University, 127 West Youyi Road, 710072 Xi'an, China
Full list of author information is available at the end of the article

interface to support research communities. Links to our source codes and web server are available at https://github.com/mldlproject/2022-iPromoter-Seqvec.

Keywords: DNA, Transcription start site, Promoter, TATA-box, Bidirectional long short-term memory

Background

Promoters are DNA non-coding regions found near and upstream the transcription start site (TSS) of genes or gene clusters [1]. As essential regulatory elements for the initiation and regulation of transcriptional processes, promoters play an important role in determining the direction and pace of DNA transcription and combining with RNA polymerase to facilitate proper initiation of transcription [2]. Understanding their molecular behaviors is also critical in investigating gene structures, assessing gene regulation methods, and annotating functional genes [3]. Besides, the initial step in explaining the transcriptional processes and expression control of genes is to map promoters to genomes [4]. Furthermore, identifying promoters in DNA sequences and genomes significantly contributes to discovering entire structures of genes of interest [5–7]. These eukaryotic transcriptional elements have typical lengths from around 60-120bp to 250bp, extending to downstream regions of the TSS [8]. For prokaryotes, the lengths of promoters extensively vary up to 1000bp [9]. Promoters may be characterized by TSS-upstream regions called TATA-boxes, which can direct other transcriptional factors to recognize the TSS [10]. The name 'TATA-box' comes from the nature of the region accumulating repetitive T and A base pairs (TATA). In human genomes, there are about 25% of known genes having promoters regions containing TATA-boxes [11]. In eukaryotic promoter regions, TATA-boxes are commonly ascertained at approximately 25bp upstream regions of the TSS [12]. The recognition of TATA-boxes indicates not only transcriptional directions but also which DNA strands are for binding [3]. Therefore, exploration of promoter regions is one of the most imperative topics in molecular genetics and biology.

To identify promoters, experimental techniques have been developed to improve determination efficiency and accuracy. Mutational analysis [13] and immunoprecipitation assays [14, 15] have been known as the two most prevalent used techniques. These techniques, however, are not time- and cost-effective and require skilled and experienced workers. Recently, with the extensive growth of the next-generation sequencing (NGS) technology [16], a large number of genomes have been sequenced to provide a huge source of genome data for *in silico* discovery [17–22]. This data availability has motivated researchers to develop computational models to predict promoters besides experimental approaches. So far, computational models have been developed based on signals, contents, and GpG information of sequences. Signal-based models use features extracted from information on RNA polymerase binding sites while neglecting information about neighboring sites so that their performances are usually poor [23–26]. Content-based models focus on features obtained from the calculation of *k*-mer frequencies and *k*-mer-derived features but pay less attention to the serial information of the nucleic acids in the sequence [27–29]. Unlike those two previous approaches, GpG-based models exploit locational information of GpG islands; however, GpG-based features are indistinct if just over a half of promoters possess GpG islands [30–32]. Besides, limited data sources for computational modeling was one of major limitations at that time. In recent years, science and technology have made a big leap in improving computing platforms, data storage, and computational methods to enhance computing efficiency and prediction power. Therefore, today *in silico* models have been developed with considerably elevated performances. Most of the recently developed models employ diverse types of sequence-based features [32–36]. These methods, however, mainly rely on selecting feature engineering techniques to extract sequence's domain knowledge, and combining multiple encoding schemes may unnecessarily increase data dimensionality. Besides, developing models using traditional machine learning algorithms with high-dimensional data requires high computational costs. Deep learning, hence, can be an alternative method to construct prediction models with highly effective feature extraction integrated. Besides known successful applications in image [37], voice [38], and video [39] processing and detection, deep learning has also been widely applied in drug discovery [40], bioinformatics [41], and other scientific fields [42] to address existing shortcomings for a decade. For promoter identification, various studies have been conducted with different objectives [43–46]. In 2018, iPromoter-2L [43] was first developed for bacterial promoter prediction using random forest [47] and pseudo *K*-tuple nucleotide composition features [48]. One year after, iPromoter-2L 2.0 [44] (iPromoter-2L's upgraded version), developed using support vector machines and *k*-mer incorporated with pseudo *K*-tuple nucleotide composition features, was released. In 2019, DeePromoter [45] was developed using convolutional neural networks, a prevalently used deep learning architecture, and one-hot encoding

to predict promoters in human and mouse genomes. In the same year, Lai et al. introduced iProEP [46] for identifying promoters in multiple species, encompassing *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Bacillus subtilis*, and *Escherichia coli*, using support vector machines in combination with pseudo *K*-tuple nucleotide composition features. In 2021, Zhu et al. proposed a cross-species prediction framework called Depicter to determine three distinct types of promoters, including TATA, non-TATA, and unclassified promoters [49]. Despite satisfactory performance obtained, there is still a large room for model improvement to achieve more effective models having higher predictive efficiency, robustness, and stability.

In this study, we introduce a more effective computational model called iPromoter-Seqvec to predict TATA and non-TATA promoters in human (*Homo sapiens*) genome and mouse (*Mus musculus*) genome using bidirectional long short-term memory (Bi-LSTM) incorporated with sequence-embedded features. Long short-term memory, a deep learning architecture, belongs to a group of recurrent neural networks which are widely used in natural language processing and machine translation. For a decade, deep learning has been widely implemented to solve multiple issues in diverse fields, including biology [50], chemistry [51–53], and biochemistry [54–57]. Numerous computational approaches were developed using deep learning to address diverse biological issues [58–64]. The application of the Bi-LSTM architecture on sequence-embedded features promotes effective learning of models in forward and reverse directions with accelerated training speed compared to traditional machine learning algorithms. Sequence-embedded features, inspired by the idea of word embedding, can efficiently represent serial information of biological sequences characterized by orders of the nucleic acids in each sequence. Sequence samples used in our experiments were collected from the Eukaryotic Promoter database [65, 66] and carefully curated to create a training set, a validation set, and a test set. These datasets were controlled to be independent of each other without any repeated or highly similar sequences. To fairly assess the model performance, we compared iPromoter-Seqvec with two state-of-the-art methods: DeePromoter [45] and iProEP [46] that share common characteristics and are relevant to our study.

Results and discussion

Model evaluation

The model performance of iPromoter-Seqvec on the validation sets is provided in Table S1 (Supplementary Information). Since DeePromoter was also developed using 'fake' negative samples like ours, we reimplemented DeePromoter and evaluated its performance on the validation sets to compare the adaptivity of using 'fake' negative samples between iPromoter-Seqvec and DeePromoter. The results show that variation in model performance between the validation sets and the test sets for both methods is relatively small. The area under the receiver operating characteristic curve (AUCROC) and the area under the precision-recall curve (AUCPR) are two key metrics used for model evaluation. For identifying promoters in human and mouse genomes, the models predicting TATA promoters perform better than the models predicting non-TATA promoters in terms of AUCROC and AUCPR. In the aspect of other metrics, the models predicting TATA promoters for both species achieve higher values compared to those predicting non-TATA promoters. The distinct characteristics between promoters and non-promoters somehow can explain the slightly greater performance of models predicting TATA promoters in comparison with those predicting non-TATA promoters. Generally, both methods show high adaptivity to using 'fake' negative samples in the training model.

Comparative analysis

Table 1 compares differences in model performance of iPromoter-Seqvec, iPro-EP, and DeePromoter. Since iPro-EP does not support predicting promoters in mouse genome, we compared the model performance based on the datasets for human genome only. To evaluate the performance of iPro-EP and DeePromoter, the test sets were uploaded to their online web servers to perform prediction tasks and retrieve predicted probabilities. For identifying promoters in human genome, iPromoter-Seqvec obtains AUCROC values of 0.99 and 0.85 for predicting TATA promoters and non-TATA promoters, respectively. The AUCPR values of iPromoter-Seqvec are also higher over those of iPro-EP and DeePromoter with 0.99 and 0.86 for predicting TATA promoters and non-TATA promoters, respectively. For identifying TATA promoters in mouse genome, AUCROC and AUCPR values of iPromoter-Seqvec are also higher than those of DeePromoter. For models predicting non-TATA promoters in mouse genome, both AUCROC and AUCPR values also confirm that iPromoter-Seqvec outperformed DeePromoter. The other metrics were also computed to provide more detailed information on model performance.

iPromoter-Seqvec (our method), iPro-EP, and DeePromoter were developed to predict promoter regions from long DNA sequences. Also, there are other computational tools have been proposed to identify promoter sequences from limited-length DNA sequences. While prediction models like ours can answer whether any promoter region is present in DNA sequences of length up to

Table 1 Model performance on the independent test sets of iPromoter-Seqvec and other state-of-the-art methods

Dataset	Method	AUCROC	AUCPR	BA	SN	SP	PR	MCC	F1
HS-TApro	iPro-EP	0.89	0.87	0.81	0.84	0.78	0.79	0.62	0.81
	DeePromoter	-	-	0.67	0.94	0.39	0.61	0.40	0.74
	iPromoter-Seqvec (Ours)	0.99	0.99	0.94	0.90	0.99	0.99	0.89	0.94
HS-nonTApro	iPro-EP	0.73	0.74	0.65	0.73	0.56	0.63	0.30	0.67
	DeePromoter	-	-	0.51	0.90	0.12	0.51	0.04	0.65
	iPromoter-Seqvec (Ours)	0.86	0.86	0.75	0.62	0.89	0.85	0.53	0.72
MM-TApro	DeePromoter	-	-	0.59	0.84	0.34	0.56	0.21	0.67
	iPromoter-Seqvec (Ours)	0.99	0.99	0.93	0.88	0.98	0.97	0.86	0.92
MM-nonTApro	DeePromoter	-	-	0.64	0.87	0.40	0.59	0.31	0.71
	iPromoter-Seqvec (Ours)	0.91	0.91	0.83	0.74	0.91	0.90	0.67	0.81

300bp, iPromoter-2L [43], as well as similar approaches, can only answer whether any promoter region is present in a DNA sequence of length at 81bp or lower. Nevertheless, iPromoter-2L can determine which type a promoter sequence belongs to. Hence, both approaches have their values and contributions in supporting different purposes and users.

Conclusions

In this study, we proposed iPromoter-Seqvec, an efficient computational model using bidirectional long short-term memory neural networks and sequence-embedding features to identify TATA promoters and non-TATA promoters in human and mouse genomes. Based on evaluation metrics recorded on independent test sets, iPromoter-Seqvec is a stable and robust computational model with high AUCROC and AUCPR values. In comparison with other state-of-the-art methods, iPromoter-Seqvec shows stronger prediction power in recognizing both TATA and non-TATA promoters. Our proposed method was also deployed as an online web server with a user-friendly interface to support research communities.

Methods

Overview

Figure 1 summarizes major steps in developing iPromoter-Seqvec. First, the sequence data, including experimentally verified ('real') promoter and non-promoter sequences, were collected from the Eukaryotic Promoter database [65, 66]. Benchmark dataset Section explains how the datasets were collected and refined. To create a validation set and an independent test set for each dataset, real promoter sequences and real non-promoter sequences were combined at an equal proportion. To create a training set, real promoter sequences were used as templates for building artificial promoter sequences. Each promoter sequence was split into smaller subsequences and then recombined to create one artificial non-promoter sequence. The detailed information on building artificial ('fake') non-promoter sequences is described in Construction of artificial non-promoter sequences Section. The real promoter sequences and the fake non-promoter sequences of each dataset were combined to create a training set. The training sets were used to train models while the validation sets were used for determining at which epoch the training process should be stopped. After obtaining optimal models, the independent test sets were used to evaluate the model performance. To be recognized as the model input, all sequence data were converted to their corresponding index vectors. The index vectors stored indices of triplet sets of consecutive nucleic acids. Sequence-embedded features Section describes the data transformation process.

Benchmark dataset

The sequence samples used for model development and testing were collected from the Eukaryotic Promoter database [65, 66], a high-quality source of promoters. This database contains non-redundant eukaryotic POL II promoters whose TSSs have been experimentally verified. The length of all collected sequences is 300bp which were cut from a location of from -249 to +50bp (+1 refers to TSS) for promoter sequences and from -51 to +350bp for non-promoter sequences. Sequence samples were collected from data sources of both human and mouse genomes with annotated distinguishing groups: TATA promoters and non-TATA promoters. Therefore, four separate datasets, including TATA-promoters of human (HS-TApro), TATA-promoters of human (HS-nonTApro), TATA-promoters of mice (MM-TApro), and TATA-promoters of mice (MM-nonTApro) were obtained. High-similarity sequences in the four datasets were removed using the CD-HIT tool [67] with a sequence identity cut-off of 0.8. The training set of each

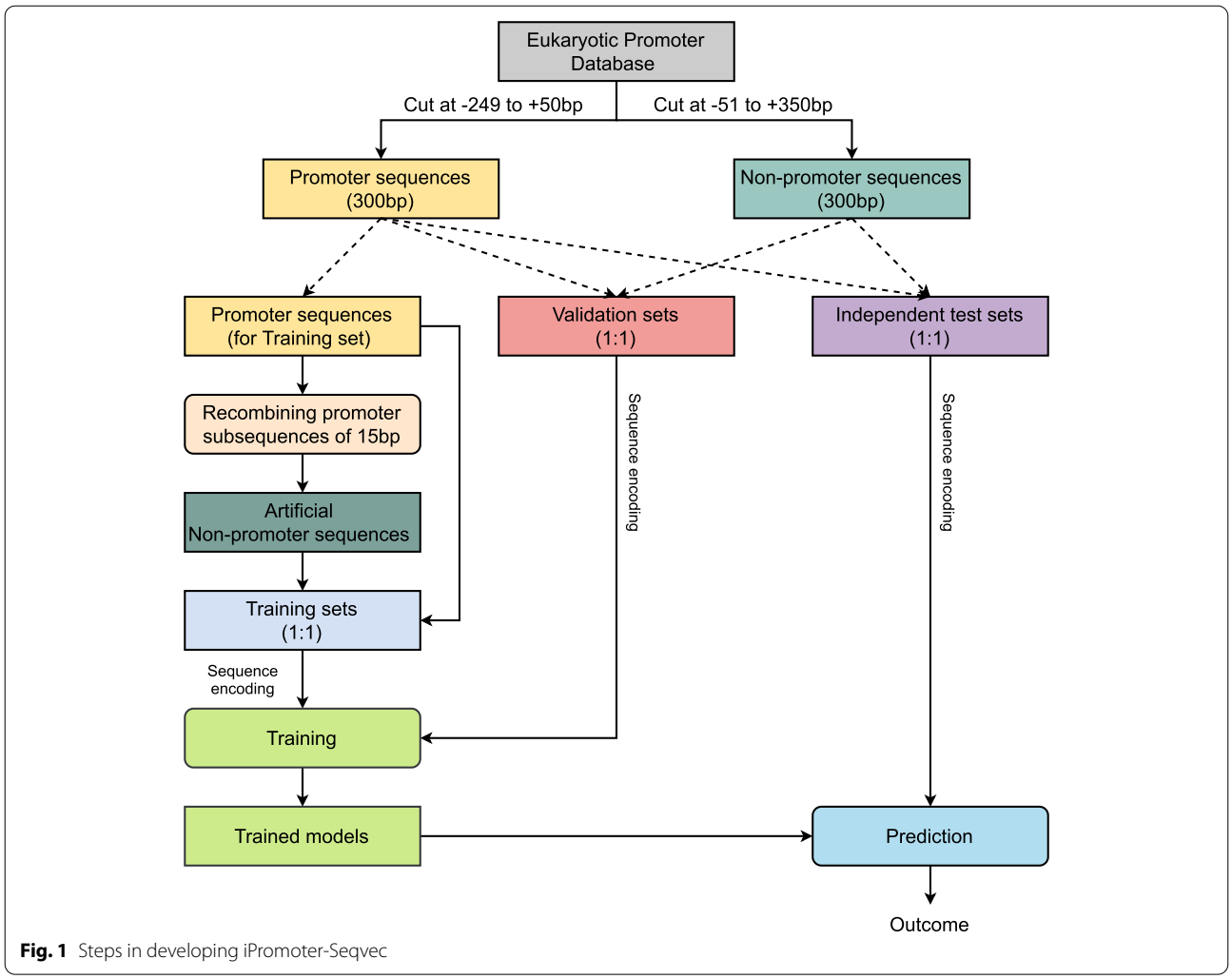

Fig. 1 Steps in developing iPromoter-Seqvec

Table 2 Datasets used for model training and evaluation

Dataset	No. of sequences (Promoters: Non-promoters = 1: 1)			Total
	Training	Validation	Test Set	
HS-TATApro	4958	400	500	5858
HS-nonTATApro	42800	4000	5000	51800
MM-TATApro	5272	400	500	6172
MM-nonTATApro	33892	4000	5000	42892

dataset was designed with an equal number of promoter and artificial non-promoter samples. The reason and processing steps of creating artificial non-promoter samples were described in the next section. The validation and test sets of each dataset contained an equal number of promoter and non-promoter sequences. Information on datasets used for model development and evaluation is provided in Table 2.

Construction of artificial non-promoter sequences

In many *in silico* studies on sequence analysis, negative samples were extracted from significantly different regions. Non-promoter or non-enhancer sequences, for instance, were collected by slicing sequences from distant locations which contain non-relevant nucleic acid contents. Since the nature of positive samples (sequences of interest) and negative samples are highly distinct, models can effortlessly learn to distinguish positives from negatives. The models, therefore, can achieve very high performance but practical applications in future prediction may be limited. As promoter sequences are characterized by highly specific regions, including TATA-box (-30 to -25bp), CAAT-box (-80 to -70bp), and GC-box (-110 to -80bp), non-promoters

having all these elements removed have no actual role but balancing the dataset. A large disparity between the promoters and non-promoters restricts models from learning decisive characteristics to accurately distinguish promoters from non-promoters. Models trained with bad or weak negatives find prediction tasks on genomics sequences challenging because genomic sequences enriched with promoter motifs may not be promoter sequences. The appearance of more 'TATA' motifs along with the genome sequences can confuse models and cause misclassifications. Hence, to develop a stable and robust model, negatives should be rigorously chosen because their features will be learned by the model to decide which class should be assigned for an unknown sample. In 2014, Wei et al. have proved the influence of good negatives on classification tasks in their studies [68]. Oubounyt et al. applied Wei et al.'s idea in developing DeePromoter using non-promoters constructed from original promoters [45]. The idea was to introduce small fragments of functional motifs from promoters to non-promoters to overcome the model's dependency on these motifs.

Figure 2 describes key steps in constructing non-promoter sequences based on their corresponding promoter sequences. For each promoter sequence, we constructed a non-promoter sequence by recombination of some promoter subsequences while keeping other promoter subsequences at their original positions. Promoter subsequences having their positions unchanged are termed 'conservative'. Promoter subsequences having their original positions interchanged by another one are termed 'substitutional' subsequences. Initially, promoter sequences of 300bp were equally split into 20 subsequences of 15bp. For each promoter sequence, 8 in 20 subsequences were randomly selected for recombination while the rest were kept immobile. The picked substitutional subsequences were then randomly filled in the gap positions until no gap remained. Finally, a new recombinant sequence was generated by joining all subsequences. Those artificial sequences

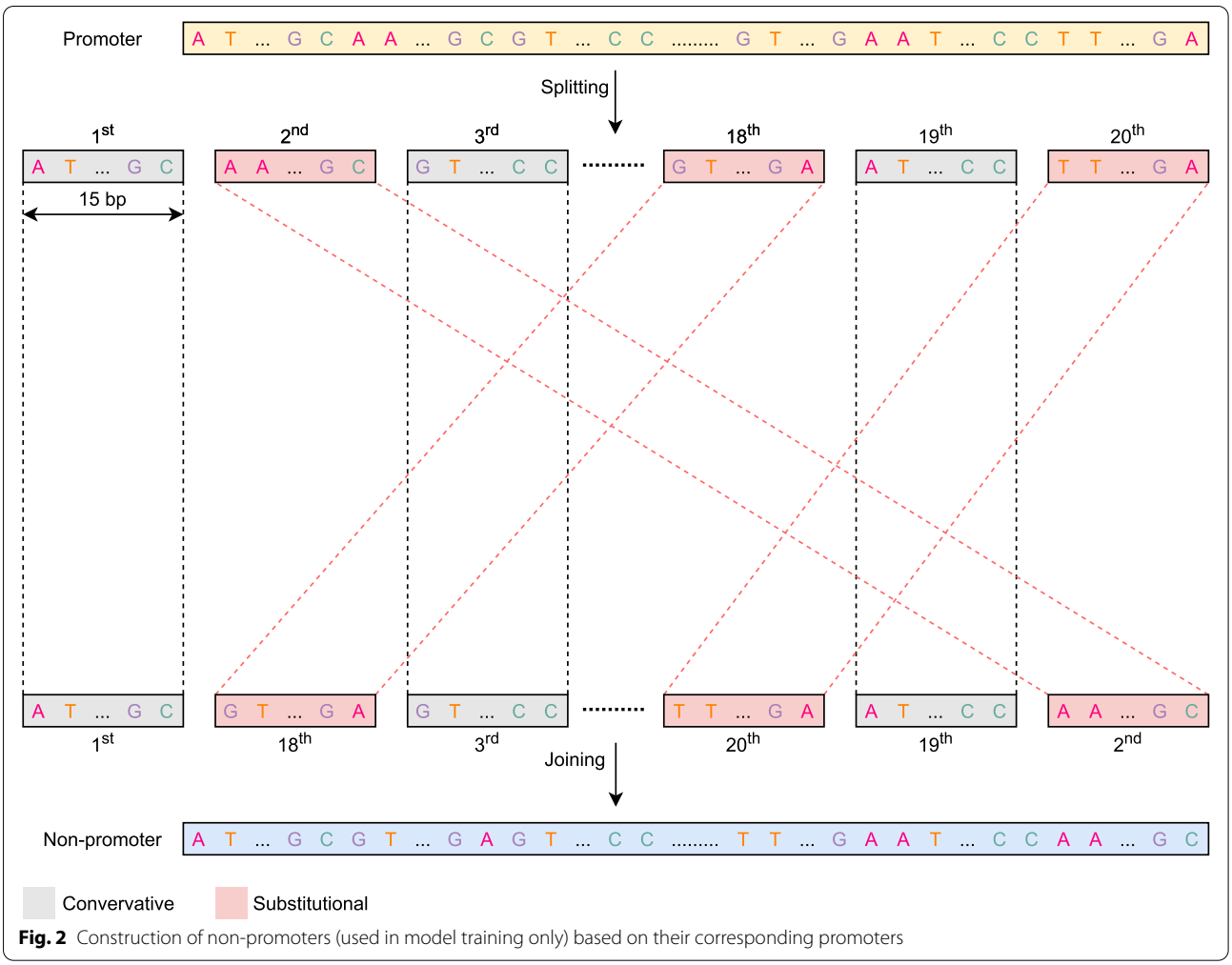

Fig. 2 Construction of non-promoters (used in model training only) based on their corresponding promoters

which share minor structural similarities compared to corresponding promoter sequences were treated as non-promoter sequences for model training only. For each present promoter sequence, a corresponding artificial non-promoter sequence was created so that the ratios of promoters to artificial non-promoters in all datasets were equal (Table 2).

Sequence-embedded features

Figure 3 summarizes the steps involved in constructing index vectors for sequence samples in our study. Initially, an index table for triplet keys was created to store indices of triplet sets of consecutive nucleic acids. For a sequence, a window of 3 was used to read the whole sequence, starting at the first nucleic acid and terminating when reaching the final one. Since the sequence length is 300bp, the maximum number of triplet keys extracted is 298. Each triplet key was then looked up with the index table to get its corresponding index. Subsequently, a list of indices was obtained with a specific order and then joined to create an index vector of 1×298. The index vectors were inputs of our models.

Model architecture

Figure 4 describes the model architecture designed to identify human TATA promoters, human non-TATA promoters, mice's TATA promoters, and mice's non-TATA promoters. The input data of the models are index vectors sized 1×298. The input data first enters the embedding layer with an embedding size of 64 to create embedding matrices sized 298×64 before passing through the batch normalization (BatchNorm) layer. The embedding layer receives data in the form of index vectors storing a series of indices. These indices come from the triplet sets of consecutive nucleic acids. The normalized matrices are the inputs of bidirectional long short-term memory (Bi-LSTM) layers designed with a hidden dimension of 128. Bi-LSTM activates a process of reading sequence information in both directions: forward and backward. Unlike regular LSTM models that use only one stream of input data, the Bi-LSTM model receives input streams in both directions. The Bi-LSTM layers transform normalized matrices sized 298×64 to matrices sized 298×256. These matrices are then flattened and passed through the first fully connected (FC1) layer activated by a Leaky Rectified Linear Unit (Leaky ReLU). After passing layer FC1, vectors sized 1×76288 are converted to vectors sized 1×128 which are gone through layer FC2 and finally activated by the sigmoid function to return probabilities. The loss function used is the binary cross-entropy which is expressed as:

$$Loss = \sum_{i=1}^{n} y_i \times log\hat{y}_i + (1 - y_i) \times log(1 - \hat{y}_i), \quad (1)$$

where y is the true label and $\hat{y}$ is the predicted probability. The prediction threshold was set at 0.5 by default. The validation sets were used to define the stopping epochs for four models. For each model, the stopping epoch was the epoch where the validation loss was minimum. The Adam optimization algorithm [69] was used along with each minibatch of 64 samples. In our experiments, iPromoter-Seqvec was implemented using PyTorch 1.3.1 and trained on Google Colab equipped with 25 GB of RAM and one NVIDIA Tesla T4 GPU. iPromoter-Seqvec was trained over 50 epochs. It took about 15 seconds and 60 seconds to complete one training epoch for models predicting TATA promoters and models predicting non-TATA promoters, respectively. iPromoter-Seqvec requires 0.5 seconds and 3 seconds to complete testing models that predict TATA promoters and models that predict non-TATA promoters, respectively.

Evaluation metrics

To assess the model performance, several metrics including balanced accuracy (BA), sensitivity (SN), specificity (SP), precision (PR), F1 score, Matthews's correlation coefficient (MCC), the area under the receiver operating characteristic curve (AUCROC), and the area under the precision-recall curve (AUCPR) were measured. TP, FP, TN, and FN are abbreviated for True Positives, False Positives, True Negatives, and False Negatives, respectively. The mathematical formulas of these metrics are expressed below.

$$BA = \frac{SN + SP}{2} \quad (2)$$

$$SN = \frac{TP}{TP + FN} \quad (3)$$

$$SP = \frac{TN}{TN + FP} \quad (4)$$

$$PR = \frac{TP}{TP + FP} \quad (5)$$

$$F_1 = 2 \times \frac{PR \times SN}{PR + SN} \quad (6)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (7)$$

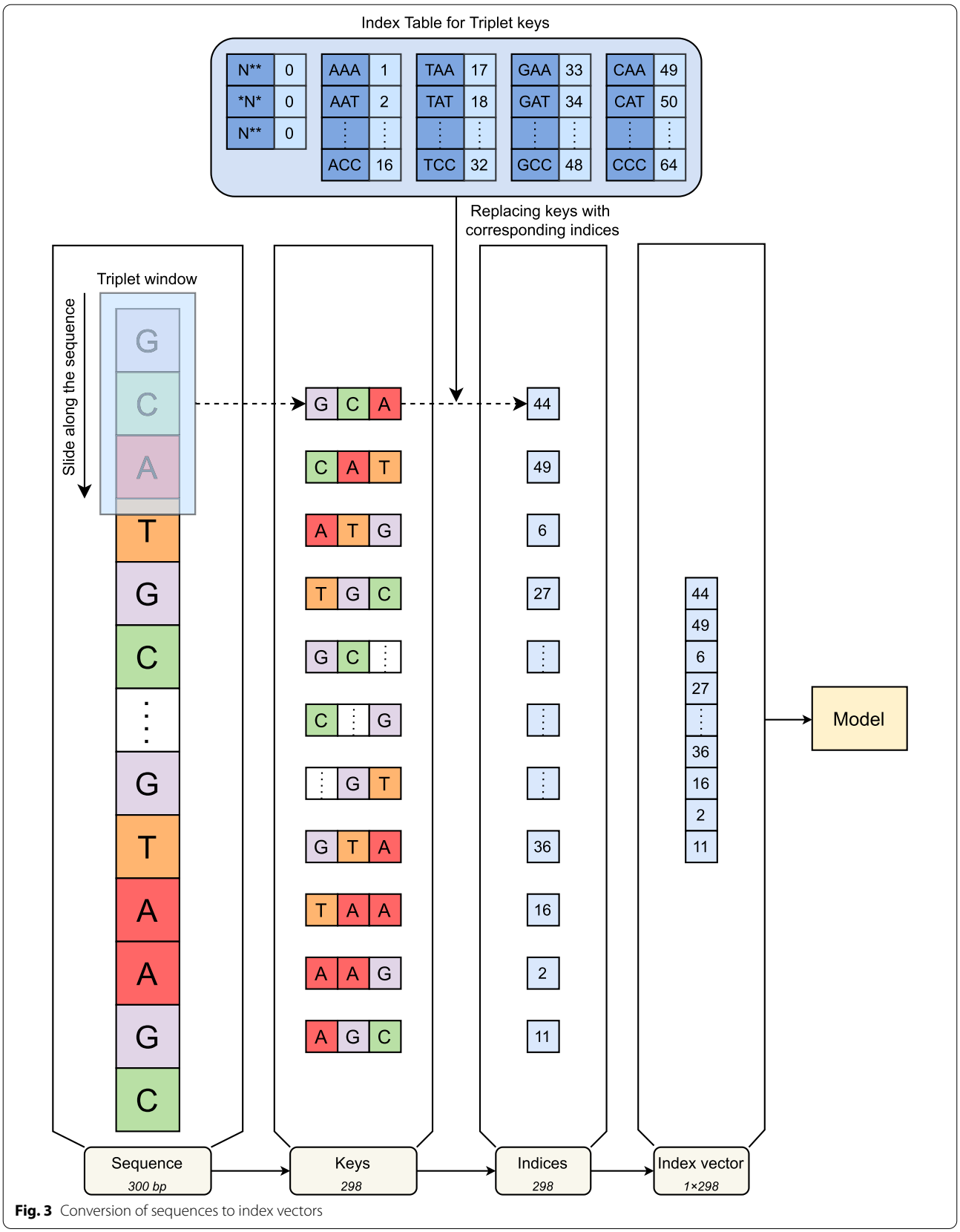

Fig. 3 Conversion of sequences to index vectors

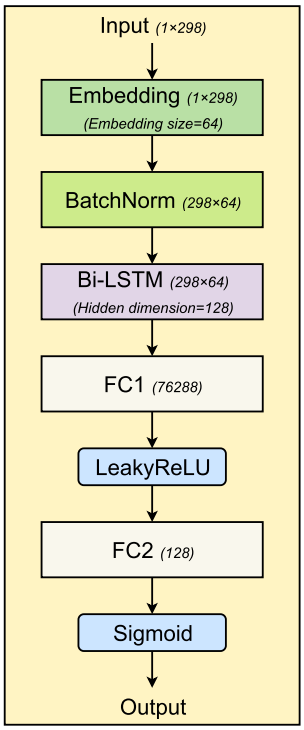

Fig. 4 Model architecture

Software availability

To support research communities to identify promoters, we deployed iPromoter-Seqvec as a user-friendly interface web server which can be accessed via https://github.com/mldlproject/2022-iPromoter-Seqvec. iPromoter-Seqvec supports identifying TATA and non-TATA promoters in human and mouse genomes. Users can follow simple steps described on the web server to perform their predictions task with iPromoter-Seqvec.

Abbreviations

DNA: Deoxyribonucleic Acid; TSS: transcription start site; MCC: Matthew's Correlation Coefficient; ROC: Receiver Operating Characteristic; PR: Precision-Recall; AUCROC: Area under the ROC curve; AUCPR: Area under the PR curve.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08829-6.

Additional file 1: Supplementary Table 1: Model performance of iPromoter-Seqvec and DeePromoter on the validation sets. **Supplementary Figure 1**: ROC curves of iPromoter-Seqvec and iPro-EL on different independent test sets. **Supplementary Figure 2**: PR curves of iPromoter-Seqvec and iPro-EL on different independent test sets.

Acknowledgements

Not applicable.

About this supplement

This article has been published as part of BMC Genomics Volume 23 Supplement 5, 2022 Selected articles from the 20th Asia Pacific Bioinformatics Conference (APBC 2022): genomics. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-23-supplement-5.

Authors' contributions

THNV: conceptualization, methodology, data curation, investigation, formal analysis, validation, visualization, writing original draft, writing review & editing. QHT: investigation, software. LN: investigation, data curation. PUNH: writing original draft, writing review & editing. SR: formal analysis, writing review & editing, supervision. BPN: conceptualization, methodology, formal analysis, visualization, writing review & editing, supervision. All authors read and approved the final manuscript.

Funding

The authors received no specific funding for this work. The publication costs were covered by the authors.

Availability of data and materials

To create the benchmark dataset, original data were collected from the Eukaryotic Promoter database [65, 66] and then independently refined. The benchmark datasets can be downloaded from our project website at https://github.com/mldlproject/2022-iPromoter-Seqvec. A web server implementing the proposed method can be accessed from there as well.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Mathematics and Statistics, Victoria University of Wellington, Gate 7, Kelburn Parade, 6140 Wellington, New Zealand. [2]School of Information and Communication Technology, Hanoi University of Science and Technology, 1 Dai Co Viet, 100000 Hanoi, Vietnam. [3]Computational Biology Center, International University - VNU HCMC, Quarter 6, Linh Trung Ward, Thu Duc District, 700000 Ho Chi Minh City, Vietnam. [4]School of Marine Science and Technology, Northwestern Polytechnical University, 127 West Youyi Road, 710072 Xi'an, China. [5]Infocomm Technology Cluster, Singapore Institute of Technology, 10 Dover Drive, 138683 Singapore, Singapore.

Received: 1 August 2022 Accepted: 8 August 2022
Published online: 03 October 2022

References

1. Haberle V, Lenhard B. Promoter architectures and developmental gene regulation. In: Seminars in Cell & Developmental Biology. vol. 57. Elsevier; 2016. p. 11–23. https://doi.org/10.1016/j.semcdb.2016.01.014.
2. Thomas MC, Chiang CM. The general transcription machinery and general cofactors. Critical Reviews in Biochemistry and Molecular Biology. 2006;41(3):105–78. https://doi.org/10.1080/10409230600648736.
3. Slobodin B, Agami R. Transcription initiation determines its end. Molecular Cell. 2015;57(2):205–6. https://doi.org/10.1016/j.molcel.2015.01.006.
4. Sutherland H, Bickmore WA. Transcription factories: gene expression in unions? Nature Reviews Genetics. 2009;10(7):457–66. https://doi.org/10.1038/nrg2592.
5. Yamasaki T, Nakajima H, Kono N, Hotta K, Yamada K, Imai E, et al. Structure of the entire human muscle phosphofructokinase-encoding gene: a

two-promoter system. Gene. 1991;104(2):277–82. https://doi.org/10.1016/0378-1119(91)90262-a.
6. Vilches C, Gardiner CM, Parham P. Gene structure and promoter variation of expressed and nonexpressed variants of the KIR2DL5 gene. J Immunol. 2000;165(11):6416–21. https://doi.org/10.4049/jimmunol.165.11.6416.
7. Lombardi L, Ciana P, Cappellini C, Trecca D, Guerrini L, Migliazza A, et al. Structural and functional characterization of the promoter regions of the NFKB2 gene. Nucleic Acids Res. 1995;23(12):2328–36. https://doi.org/10.1093/nar/23.12.2328.
8. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol. 2018;19(10):621–37. https://doi.org/10.1038/s41580-018-0028-8.
9. Kristiansson E, Thorsen M, Tamás MJ, Nerman O. Evolutionary Forces Act on Promoter Length: Identification of Enriched Cis-Regulatory Elements. Mol Biol Evol. 2009;26(6):1299–307. https://doi.org/10.1093/molbev/msp040.
10. Watson JD, Baker TA, Bell SP, Gann A, Levine M, Losick R. Molecular Biology of the Gene. 6th ed. Pearson Education. 2008.
11. Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. Gene. 2007;389(1):52–65. https://doi.org/10.1016/j.gene.2006.09.029.
12. Xu M, Gonzalez-Hurtado E, Martinez E. Core promoter-specific gene regulation: TATA box selectivity and Initiator-dependent bi-directionality of serum response factor-activated transcription. Biochim Biophys Acta (BBA) Gene Regul Mech. 2016;1859(4):553–63. https://doi.org/10.1016/j.bbagrm.2016.01.005.
13. Matsumine H, Yamamura Y, Hattori N, Kobayashi T, Kitada T, Yoritaka A, et al. A microdeletion of D6S305 in a family of autosomal recessive juvenile parkinsonism (PARK2). Genomics. 1998;49(1):143–6. https://doi.org/10.1006/geno.1997.5196.
14. Tian X, Jin RU, Bredemeyer AJ, Oates EJ, Błazewska KM, McKenna CE, et al. RAB26 and RAB3D are direct transcriptional targets of MIST1 that regulate exocrine granule maturation. Mol Cell Biol. 2010;30(5):1269–84. https://doi.org/10.1128/MCB.01328-09.
15. Dahl JA, Collas P. A rapid micro chromatin immunoprecipitation assay (ChIP). Nat Protoc. 2008;3(6):1032–45. https://doi.org/10.1038/nprot.2008.68.
16. Behjati S, Tarpey PS. What is next generation sequencing? Arch Dis Child-Educ Pract. 2013;98(6):236–8. http://dx.doi.org/archdischild-2013-304340.
17. Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. J Gen Genomics. 2011;38(3):95–109. https://doi.org/10.1016/j.jgg.2011.02.003.
18. Xu Y, Ding J, Wu LY, Chou KC. iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PloS ONE. 2013;8(2):e55844. https://doi.org/10.1371/journal.pone.0055844.
19. Chen W, Feng P, Ding H, Lin H, Chou KC. iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. Anal Biochem. 2015;490:26–33. https://doi.org/10.1016/j.ab.2015.08.021.
20. Jia J, Zhang L, Liu Z, Xiao X, Chou KC. pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. Bioinformatics. 2016;32(20):3133–41. https://doi.org/10.1093/bioinformatics/btw387.
21. Cheng X, Zhao SG, Xiao X, Chou KC. iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals. Oncotarget. 2017;8(35):58494. https://doi.org/10.18632/oncotarget.17028.
22. Nguyen-Vo TH, Nguyen QH, Do TTT, Nguyen TN, Rahardja S, Nguyen BP. iPseU-NCP: Identifying RNA pseudouridine sites using random forest and NCP-encoded features. BMC Genomics. 2019;20(971). https://doi.org/10.1186/s12864-019-6357-y.
23. Prestridge DS. Predicting Pol II promoter sequences using transcription factor binding sites. J Mol Biol. 1995;249(5):923–32. https://doi.org/10.1006/jmbi.1995.0349.
24. Knudsen S. Promoter2.0: for the recognition of PolII promoter sequences. Bioinformatics. 1999;15(5):356–61. https://doi.org/10.1093/bioinformatics/15.5.356.
25. Reese MG. Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. Comput Chem. 2001;26(1):51–6. https://doi.org/10.1016/s0097-8485(01)00099-7.
26. Down TA, Hubbard TJP. Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res. 2002;12(3):458–61. https://doi.org/10.1101/gr.216102.
27. Hutchinson GB. The prediction of vertebrate promoter regions using differential hexamer frequency analysis. Bioinformatics. 1996;12(5):391–8. https://doi.org/10.1093/bioinformatics/12.5.391.
28. Ohler U, Harbeck S, Niemann H, Nöth E, Reese MG. Interpolated Markov Chains for Eukaryotic Promoter Recognition. Bioinformatics. 1999;15(5):362–9. https://doi.org/10.1093/bioinformatics/15.5.362.
29. Scherf M, Klingenhoff A, Werner T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. J Mol Biol. 2000;297(3):599–606. https://doi.org/10.1006/jmbi.2000.3589.
30. Ioshikhes IP, Zhang MQ. Large-scale human promoter mapping using CpG islands. Nat Genet. 2000;26(1):61–3. https://doi.org/10.1038/79189.
31. Davuluri RV, Grosse I, Zhang MQ. Computational identification of promoters and first exons in the human genome. Nat Genet. 2001;29(4):412–7. https://doi.org/10.1038/ng780.
32. Ponger L, Mouchiroud D. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics. 2002;18(4):631–3. https://doi.org/10.1093/bioinformatics/18.4.631.
33. Lin H, Liang ZY, Tang H, Chen W. Identifying Sigma70 Promoters with Novel Pseudo Nucleotide Composition. IEEE/ACM Trans Comput Biol Bioinforma. 2017;16(4):1316–21.
34. Yang Y, Zhang R, Singh S, Ma J. Exploiting sequence-based features for predicting enhancer-promoter interactions. Bioinformatics. 2017;33(14):i252-60. https://doi.org/10.1093/bioinformatics/btx257.
35. Bharanikumar R, Premkumar KAR, Palaniappan A. PromoterPredict: sequence-based modelling of Escherichia coli σ70 promoter strength yields logarithmic dependence between promoter strength and sequence. PeerJ. 2018;6:e5862. https://doi.org/10.7717/peerj.5862.
36. Xiao X, Xu ZC, Qiu WR, Wang P, Ge HT, Chou KC. iPSW(2L)-PseKNC: A two-layer predictor for identifying promoters and their strength by hybrid features via pseudo K-tuple nucleotide composition. Genomics. 2019;111(6):1785–93. https://doi.org/10.1016/j.ygeno.2018.12.001.
37. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.
38. Sisman B, Yamagishi J, King S, Li H. An overview of voice conversion and its challenges: From statistical modeling to deep learning. IEEE/ACM Trans Audio Speech Lang Process. 2020. https://doi.org/10.1109/TASLP.2020.3038524.
39. Ciaparrone G, Sánchez FL, Tabik S, Troiano L, Tagliaferri R, Herrera F. Deep learning in video multi-object tracking: A survey. Neurocomputing. 2020;381:61–88. https://doi.org/10.1016/j.neucom.2019.11.023.
40. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today. 2018;23(6):1241–50. https://doi.org/10.1016/j.drudis.2018.01.039.
41. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinforma. 2017;18(5):851–69. https://doi.org/10.1093/bib/bbw068.
42. Shinde PP, Shah S. A review of machine learning and deep learning applications. In: Proceedings of the Fourth International Conference on Computing Communication Control and Automation (ICCUBEA 2018). IEEE; 2018. p. 1–6. https://doi.org/10.1109/ICCUBEA.2018.8697857.
43. Liu B, Yang F, Huang DS, Chou KC. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics. 2018;34(1):33–40. https://doi.org/10.1093/bioinformatics/btx579.
44. Liu B, Li K. iPromoter-2L2.0: identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features. Mol Therapy Nucleic Acids. 2019;18:80–7.
45. Oubounyt M, Louadi Z, Tayara H, Chong KT. DeePromoter: robust promoter predictor using deep learning. Front Genet. 2019;10:286. https://doi.org/10.3389/fgene.2019.00286.
46. Lai HY, Zhang ZY, Su ZD, Su W, Ding H, Chen W, et al. iProEP: a computational predictor for predicting promoter. Mol Therapy-Nucleic Acids. 2019;17:337–46. https://doi.org/10.1016/j.omtn.2019.05.028.
47. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.
48. Chen W, Lei TY, Jin DC, Lin H, Chou KC. PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. Anal Biochem. 2014;456:53–60. https://doi.org/10.1016/j.ab.2014.04.001.

49. Zhu Y, Li F, Xiang D, Akutsu T, Song J, Jia C. Computational identification of eukaryotic promoters based on cascaded deep capsule neural networks. Brief Bioinforma. 2021;22(4):bbaa299.
50. Trinh K, Pham D, Le L. Semantic relation extraction for herb-drug interactions from the biomedical literature using an unsupervised learning approach. In: Proceedings of the 18th International Conference on Bioinformatics and Bioengineering (BIBE 2018). IEEE; 2018. p. 334–7. https://doi.org/10.1109/BIBE.2018.00072.
51. Goh GB, Hodas NO, Vishnu A. Deep learning for computational chemistry. J Comput Chem. 2017;38(16):1291–307. https://doi.org/10.1002/jcc.24764.
52. Mater AC, Coote ML. Deep learning in chemistry. J Chem Inf Model. 2019;59(6):2545–59. https://doi.org/10.1021/acs.jcim.9b00266.
53. Debus B, Parastar H, Harrington P, Kirsanov D. Deep learning in analytical chemistry. TrAC Trends Anal Chem. 2021;145:116459. https://doi.org/10.1016/j.trac.2021.116459.
54. Nguyen-Vo TH, Nguyen L, Do N, Le PH, Nguyen TN, Nguyen BP, et al. Predicting drug-induced liver injury using convolutional neural network and molecular fingerprint-embedded features. ACS Omega. 2020;5(39):25432–9. https://doi.org/10.1021/acsomega.0c03866.
55. Nguyen-Vo TH, Trinh QH, Nguyen L, Nguyen-Hoang PU, Nguyen TN, Nguyen DT, et al. iCYP-MFE: Identifying Human Cytochrome P450 Inhibitors using Multitask Learning and Molecular Fingerprint-Embedded Encoding. J Chem Inf Model. 2021. https://doi.org/10.1021/acs.jcim.1c00628.
56. Nguyen-Vo TH, Trinh QH, Nguyen L, Do TTT, Chua MCH, Nguyen BP. Predicting Antimalarial Activity in Natural Products using Pretrained Bidirectional Encoder Representations from Transformers. J Chem Inf Model. 2021. https://doi.org/10.1021/acs.jcim.1c00584.
57. Nguyen L, Nguyen-Vo TH, Trinh QH, Nguyen BH, Nguyen-Hoang PU, Le L, et al. iANP-EC: Identifying Anticancer Natural Products Using Ensemble Learning Incorporated with Evolutionary Computation. J Chem Inf Modeling. 2022. https://doi.org/10.1021/acs.jcim.1c00920.
58. Umarov RK, Solovyev VV. Recognition of Prokaryotic and Eukaryotic Promoters using Convolutional Deep Learning Neural Networks. PloS ONE. 2017;12(2):e0171410. https://doi.org/10.1371/journal.pone.0171410.
59. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. 2017;18(1):1–13. https://doi.org/10.1186/s13059-017-1189-z.
60. Le NQK, Nguyen BP. Prediction of FMN Binding Sites in Electron Transport Chains based on 2-D CNN and PSSM Profiles. IEEE/ACM Trans Comput Biol Bioinforma. 2019. https://doi.org/10.1109/TCBB.2019.2932416.
61. Nguyen QH, Nguyen-Vo TH, Le NQK, Do TTT, Rahardja S, Nguyen BP. iEnhancer-ECNN: Identifying Enhancers and Their Strength using Ensembles of Convolutional Neural Networks. BMC Genomics. 2019;20(951). https://doi.org/10.1186/s12864-019-6336-3.
62. Nguyen BP, Nguyen QH, Doan-Ngoc GN, Nguyen-Vo TH, Rahardja S. iProDNA-CapsNet: Identifying Protein-DNA Binding Residues using Capsule Neural Networks. BMC Bioinforma. 2019;20(634). https://doi.org/10.1186/s12859-019-3295-2.
63. Chaudhari M, Thapa N, Roy K, Newman RH, Saigo H, KC DB. DeepRMethylSite: A deep learning based approach for prediction of arginine methylation sites in proteins. Mol Omics. 2020;16(5):448–54. https://doi.org/10.1039/d0mo00025f.
64. Min X, Ye C, Liu X, Zeng X. Predicting enhancer-promoter interactions by deep learning and matching heuristic. Brief Bioinforma. 2021;22(4):bbaa254.
65. Périer RC, Praz V, Junier T, Bonnard C, Bucher P. The eukaryotic promoter database (EPD). Nucleic Acids Res. 2000;28(1):302–3. https://doi.org/10.1093/nar/28.1.302.
66. Dreos R, Ambrosini G, Périer RC, Bucher P. The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. Nucleic Acids Res. 2015;43(D1):D92-6. https://doi.org/10.1093/nar/gku1111.
67. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010;26(5):680–2. https://doi.org/10.1093/bioinformatics/btq003.
68. Wei L, Liao M, Gao Y, Ji R, He Z, Zou Q. Improved and promising identification of human microRNAs by incorporating a high-quality negative set. IEEE/ACM Trans Comput Biol Bioinforma. 2013;11(1):192–201. https://doi.org/10.1109/TCBB.2013.146.
69. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv. 2014. https://doi.org/10.48550/ARXIV.1412.6980.

Publisher's Note