frontiers | Research Topics

Computational analysis of promoters in prokaryotic genomes

Edited by
Hao Lin, Yongchun Zuo and Ettayapuram Ramaprasad Azhagiya Singam

Published in
Frontiers in Microbiology
Frontiers in Genetics

ISSN 1664-8714
ISBN 978-2-8325-3065-8
DOI 10.3389/978-2-8325-3065-8

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Computational analysis of promoters in prokaryotic genomes

Topic editors

Hao Lin — University of Electronic Science and Technology of China, China
Yongchun Zuo — Inner Mongolia University, China
Ettayapuram Ramaprasad Azhagiya Singam — University of California, Berkeley, United States

Citation

Lin, H., Zuo, Y., Singam, E. R. A., eds. (2023). *Computational analysis of promoters in prokaryotic genomes*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-3065-8

Table of contents

frontiers | Frontiers in Microbiology

TYPE Editorial
PUBLISHED 10 July 2023
DOI 10.3389/fmicb.2023.1242139

OPEN ACCESS

EDITED AND REVIEWED BY
John R. Battista,
Louisiana State University, United States

*CORRESPONDENCE
Hao Lin
✉ hlin@uestc.edu.cn
Yongchun Zuo
✉ yczuo@imu.edu.cn

RECEIVED 18 June 2023
ACCEPTED 27 June 2023
PUBLISHED 10 July 2023

CITATION
Lin H, Zuo Y and Azhagiya Singam ER (2023) Editorial: Computational analysis of promoters in prokaryotic genomes.
Front. Microbiol. 14:1242139.
doi: 10.3389/fmicb.2023.1242139

Editorial: Computational analysis of promoters in prokaryotic genomes

Hao Lin[1]*, Yongchun Zuo[2]* and Ettayapuram Ramaprasad Azhagiya Singam[3]

[1]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China, [2]College of Life Sciences, Inner Mongolia University, Hohhot, China, [3]Molecular Graphics and Computation Facility, College of Chemistry, University of California, Berkeley, Berkeley, CA, United States

KEYWORDS

promoter, prokaryote, artificial intelligence, sequence, prediction

Editorial on the Research Topic
Computational analysis of promoters in prokaryotic genomes

Promoters are DNA sequence fragments located upstream of structural gene, which start gene transcription by combining with RNA polymerase. It has been found that in Prokaryote, promoters are considered to be key elements for Sigma factor recognition in the transcription process. By changing the promoter sequence, gene expression can be regulated. At present, enough prokaryotic promoter sequences have been accumulated, and multiple prokaryotic promoter databases have been constructed, such as PPD (Su et al., 2021), RegulonDB (Tierrafria et al., 2022), Pro54DB (Liang et al., 2017) and DBTBS (Sierro et al., 2008). The study of prokaryotic promoters will provide more useful information for understanding microbial gene transcription. This Research Topic aims to provide an important scientific communication platform for the analysis of prokaryotic promoters using artificial intelligence and big data techniques, including the development and application of computing methods and technologies for the analysis and research of prokaryotic genome promoters.

In this Research Topic, nine papers were published, five of which are about the use of artificial intelligent techniques to identify the prokaryotic promoter sequence.

Zulfiqar et al. developed a random forest (RF)-based model to predict promotors in *Agrobacterium Tumefaciens* strain C58. In the model, promotor sequences were encoded by accumulated nucleotide frequency, *k*-mer nucleotide composition, and binary encodings, and then optimized by using correlation and the mRMR-based algorithm. They inputted these optimized features into RF classifier to classify promotor sequences. The examination of 10-fold cross-validation (CV) showed that the proposed model could yield an overall accuracy of 0.837. They have also discussed the limitations and the future perspective of this study. Lin Y. et al. also developed a model to predict promotors in *Klebsiella Aerogenes*. In their model, they have utilized pseudo *k*-tuple nucleotide composition and position-correlation scoring function to encode the promotor sequences. They have also utilized mRMR to optimize the encoded features. Afterwards, they inputted the optimized features into support vector machine (SVM)-based classifier to recognize promotor sequences. Results on 10-fold CV showed the overall accuracy of 96.0%. They have also discussed about the future perspectives of this study. Li R. et al. developed a promoter prediction model for *Corynebacterium glutamicum* based on novel feature by calculating statistical parameters

of multiple physicochemical properties (Li H. et al.). Feature dimensionality is effectively reduced by using variance analysis and hierarchical clustering. Finally, they achieved an accuracy of 91.6%. They briefly analyzed the importance of feature selection and validated the robustness of the model. Sumeet et al. focused on sigma70 promoter in *Escherichia coli* K-12 strains. They used over 8000-dimension features to formulate samples (Patiyal et al.). By utilizing SVM as classifier, they achieved the maximum accuracy 97.38% with AUROC 0.99 on training dataset by using 200 most relevant features. They established a webserver for using by wet-experimental scholars. Shujaat et al. designed a powerful computational model to identify phage promoters and their types (Shujaat et al.). Ten distinct feature encoding approaches were investigated in this work. Finally, a 1-D convolutional neural network model combined with one-hot encoding approach was proposed to construct model. They also built a freely web server.

Transcription factors (TFs) are important regulators for gene expression. Zheng et al. presented a capsule network-based method to identify TFs. Their model obtained an accuracy of 0.8820. They also constructed a user-friendly web server for all scientific researchers.

Bo, Sun, Ning et al.; Bo, Sun, Li et al. submitted two works for mRNA splice regulation. They first presented a novel approach to analyze the association characteristics between post-spliced introns and their corresponding mRNA based on binding free energy weighted local alignment algorithm method. They briefly introduce the advantages of binding free energy weighted local alignment algorithm method to analyze the interaction of RNA-RNA, compared with Smith-Waterman local alignment-based algorithm method. They also discussed biological significance and evolutionary mechanism of the interaction between introns and mRNAs. Subsequently, they studied the ubiquitous conservative interaction patterns between post-spliced introns and their mRNAs revealed by genome-wide interspecies comparison. They also discussed show that there are abundant functional units in the introns, and these functional units are correlated structurally with all kinds of sequences of mRNA.

Although previous studies have revealed that introns play an important role in regulating gene expression and participate in gene evolution, but the function of introns is far from clear, and are being studied from different perspectives. In the work of Li R. et al., the characteristics of the optimal matched segments between the first intron and the reverse complementary sequences of other introns of each gene were analyzed, some interesting results had been gotten. The results in this paper showed that the characteristics of the optimal matched segments presented varied regular variation along with the evolution of eukaryotes. It is found that some optimal matched segments may be related to non-coding RNA with special biological functions, just like siRNA and miRNA, they may play an important role in the process of gene expression and regulation. And perhaps the optimal matched segments with special characteristics in the first introns may take part in regulating gene expression by RNA matching competition with other introns or exon.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Liang, Z. Y., Lai, H. Y., Yang, H., Zhang, C. J., Yang, H., Wei, H. H., et al. (2017). Pro54DB: a database for experimentally verified sigma-54 promoters. *Bioinformatics* 33, 467–469. doi: 10.1093/bioinformatics/btw630

Sierro, N., Makita, Y., De Hoon, and Nakai, M. (2008). DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nuc. Acids Res.* 36, D93–96. doi: 10.1093/nar/gkm910

Su, W., Liu, M. L., Yang, Y. H., Wang, J. S., Li, S. H., Lv, H., et al. (2021). PPD: a manually curated database for experimentally verified prokaryotic promoters. *J. Mol. Biol.* 433, 166860. doi: 10.1016/j.jmb.2021.166860

Tierrafria, V. H., Rioualen, C., Salgado, H., Lara, P., Gama-Castro, S., Lally, P., et al. (2022). RegulonDB 11, 0. Comprehensive high-throughput datasets on transcriptional regulation in Escherichia coli K-12. *Microb. Genom.* 8, 833. doi: 10.1099/mgen.0.000833

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 04 November 2022
DOI 10.3389/fmicb.2022.1061122

OPEN ACCESS

EDITED BY
Hao Lin,
University of Electronic Science and Technology of China, China

REVIEWED BY
Leyi Wei,
Shandong University, China
Yongqiang Xing,
Inner Mongolia University of Science and Technology, China

*CORRESPONDENCE
Hilal Tayara
hilaltayara@jbnu.ac.kr
Kil To Chong
kitchong@jbnu.ac.kr

SPECIALTY SECTION
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 04 October 2022
ACCEPTED 18 October 2022
PUBLISHED 04 November 2022

CITATION
Shujaat M, Jin JS, Tayara H and Chong KT (2022) iProm-phage: A two-layer model to identify phage promoters and their types using a convolutional neural network. *Front. Microbiol.* 13:1061122. doi: 10.3389/fmicb.2022.1061122

iProm-phage: A two-layer model to identify phage promoters and their types using a convolutional neural network

Muhammad Shujaat[1], Joe Sung Jin[2], Hilal Tayara[3]* and Kil To Chong[1,4]*

[1]Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju, South Korea, [2]Graduate School of Integrated Energy AI, Jeonbuk National University, Jeonju, South Korea, [3]School of International Engineering and Science, Jeonbuk National University, Jeonju, South Korea, [4]Advances Electronics and Information Research Center, Jeonbuk National University, Jeonju, South Korea

The increased interest in phages as antibacterial agents has resulted in a rise in the number of sequenced phage genomes, necessitating the development of user-friendly bioinformatics tools for genome annotation. A promoter is a DNA sequence that is used in the annotation of phage genomes. In this study we proposed a two layer model called "iProm-phage" for the prediction and classification of phage promoters. Model first layer identify query sequence as promoter or non-promoter and if the query sequence is predicted as promoter then model second layer classify it as phage or host promoter. Furthermore, rather than using non-coding regions of the genome as a negative set, we created a more challenging negative dataset using promoter sequences. The presented approach improves discrimination while decreasing the frequency of erroneous positive predictions. For feature selection, we investigated 10 distinct feature encoding approaches and utilized them with several machine-learning algorithms and a 1-D convolutional neural network model. We discovered that the one-hot encoding approach and the CNN model outperformed based on performance metrics. Based on the results of the 5-fold cross validation, the proposed predictor has a high potential. Furthermore, to make it easier for other experimental scientists to obtain the results they require, we set up a freely accessible and user-friendly web server at http://nsclbio.jbnu.ac.kr/tools/iProm-phage/.

KEYWORDS

DNA promoters, convolutional neural networks, bioinformatics, computational biology, phages

Introduction

Bacteriophages, commonly referred to as phages, are viruses that infect and destroy bacteria (Salmond and Fineran, 2015). The number of sequenced phage genomes has increased exponentially in recent decades, primarily owing to their small size and ability to bacterial infections (Silva and Echeverrigaray, 2012). This richness of genomic data necessitates the development of user-friendly bioinformatics tools to aid biologists in genome analyses. Recognition of regulatory elements is the most difficult phase in phage genome analysis. Promoters are DNA sequences responsible for transcription initiation. These sequences are difficult to identify because they are composed of short, nonconserved components. However, it is essential to comprehend and describe the genetic regulatory networks of phages, which may permit the engineering of improved phages for medicinal or biotechnological applications (Guzina and Djordjevic, 2015).

Several attempts have been made to develop promoter prediction tools for bacterial genomes. The majority of these tools use computational techniques based on-10 and-35 motifs (Sierro et al., 2008; Mishra et al., 2020; Wang et al., 2020). In contrast to these promoters with typical motifs, phage genome promoters are composed of host and phage promoters with varying motifs (Sampaio et al., 2019).

Therefore, existing tools are not suitable for identifying promoters in phages. Computational tools are required to predict promoters in phages. Prediction of phage promoters has seldom been studied. The PHIRE method (Lavigne et al., 2004) systematically scans a bacteriophage genome to determine the frequency of subsequences in a sequence. All sequences are compared, which significantly increases the running time. PromoterHunter (Klucar et al., 2010) is an online tool to identify phage promoters; however, it requires additional information as input, such as weight matrices of the two promoter elements and is limited concerning the size of the input genome sequences. The PhagePromoter tool (Sampaio et al., 2019) can be used to identify promoters across the entire phage genome. It was created using machine learning (ML) methods, such as artificial neural networks or support vector machines, in conjunction with sequence characteristics (size and score of motifs, frequency of adenine and thymine, and free energy value). Additionally, PhagePromoter can distinguish host promoters from phage promoters. However, PhagePromoter has to be used in a deterministic manner with some previous experimental or predictive knowledge, such as phage family, host bacterium species, and phage type (temperature or virulence), which limits the effectiveness of PhagePromoter. DPProm (Wang et al., 2022) is a proposed convolutional neural network (CNN)-based method for predicting phage promoters and their types as phages or hosts. However, the proposed sequence-processing workflow requires a long time for a query sequence.

Significant progress has been achieved in the essential aspects of phage promoter identification, although improvements are required in different aspects. We identified the following shortcomings of prior research:

1. Most of the aforementioned studies only predicted the promoter sequence as phage or non-promoter. Classification of predicted promoter sequences as phages or hosts was rare.
2. Most studies utilized ML models to classify predicted sequences.
3. Not all studies created a user-friendly and publicly available web server, which has proven inconvenient for practical use by experimental scientists.
4. Performance analysis of different feature encoding schemes on different ML and CNN models was not performed.
5. In the previously proposed tools, the number of false positive values for promoter prediction requires further improvement.
6. Previous studies selected non-coding regions as negative dataset, that's makes a very easy task for the classifier on other hand trained model cannot perform well on difficult test datasets.

In this study, we focused on overcoming these drawbacks to improve the prediction capabilities in identifying phage promoters. First, high-quality benchmark datasets were constructed. Subsequently, we extracted the best feature representation vector and model from a variety of encoding techniques, ML, and CNN models. To achieve this, we sequentially fed encoded vector sequences from all encoding methods into various ML and CNN algorithms. Based on performance evaluation, we chose the one-hot encoding technique and CNN algorithm. We investigated the sequence and properties of phage promoters and presented a two-layer model designated "iProm-phage." In the first layer model, the query sequence is identified as a promoter or non-promoter. If it is a promoter sequence, then the second layer classifies the identified sequence as a phage promoter or host promoter. To assess model performance, we measured the accuracy (Acc), sensitivity (Sn), specificity (Sp), and Matthew's correlation coefficient (MCC). All these parameters are frequently used in state-of-the-art methods in computational biology and bioinformatics (Rahman et al., 2019; Ali et al., 2020; Shujaat et al., 2020; Rehman et al., 2021). In addition, we evaluated the model using five-fold cross validation and receiver operating characteristic (ROC) curves. Finally, the iProm-phage web server was built in compliance with the suggested paradigm. The proposed flow diagram of the study is shown in Figure 1.

Materials and methods

Benchmark dataset

While developing an effective biological predictor, it is critical to select an appropriate benchmark dataset to evaluate the proposed predictive model. We prepared separate datasets for

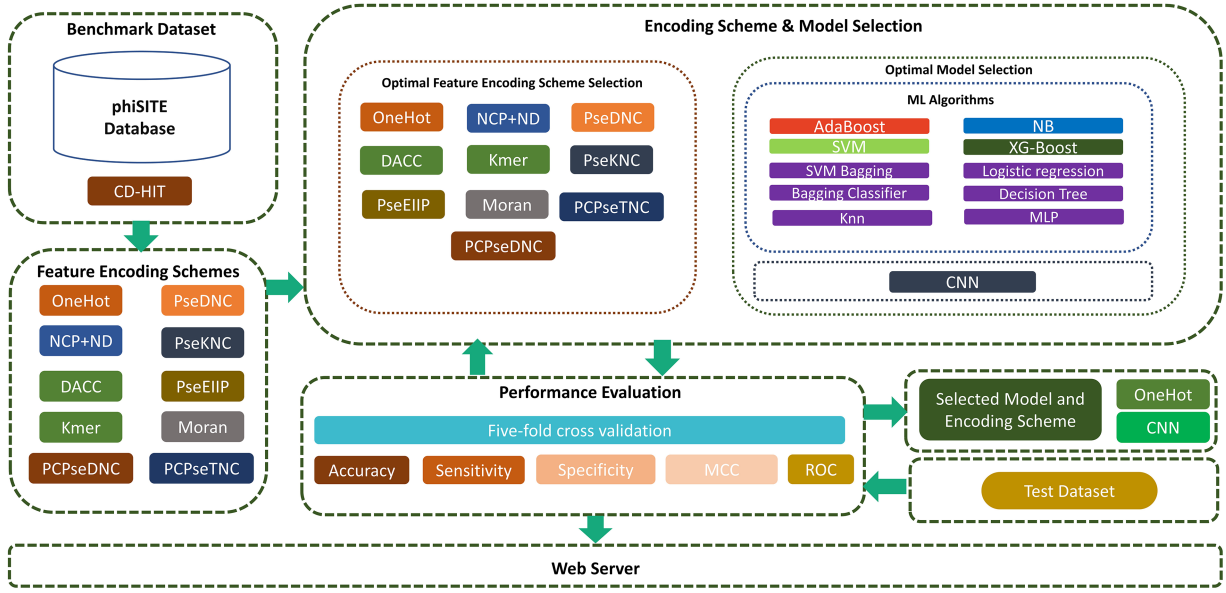

FIGURE 1
Flow diagram of iProm-phage.

TABLE 1 Summary of the Benchmark dataset.

Model Layer	Dataset	Promoter	Non-promoter
First layer	Training	901	901
	Test	198	198
Second layer		**Phage**	**Host**
	Training	111	382
	Test	28	96

each layer of the model, as described in Sections "Dataset for the first layer" and "Dataset for the second layer."

Dataset for the first layer

The promoters of phage genomes have been poorly characterized. Only the phiSITE database has identified the promoters of phage genomes (Klucar et al., 2010). The phage promoter sequence utilized in this study is the same as that used in previous studies (Sampaio et al., 2019; Wang et al., 2022). For the model's first layer, 1,140 promoter sequences from 69 phages were collected and divided into training and test datasets; 901 promoter sequences were utilized as the training dataset and 198 promoter sequences were utilized as the test dataset. Supplementary Table S1 in Supplementary file summarize the promoter sequences from each phage genome.

The selection of a negative dataset is an important step in ensuring model performance. In previous studies, non-promoter regions were randomly selected to build a negative dataset. However, this method tends to be illogical because there is no intersection between positive and negative sets. Consequently, the model immediately detected the key differences between the two groups. Therefore, precision could not be maintained when tested on more difficult datasets. To overcome this problem, we propose a negative dataset generation technique. We created a negative dataset from positive promoter sequences by the following three steps. First, each positive sequence is divided into eight subsequences. Second, five subsequences are randomly selected and placed. Thirdly, the remaining three subsequences are placed at the same position. Using this method, each positive promoter sequence creates one negative sequence with 35–40% conserved portions from the promoter sequence. This proportion is ideal as a reliable predictor of promoter activity.

Dataset for the second layer

To create the positive and negative sets for the second layer of the model, promoter sequence type information as a host or phage was retrieved. The collection contains several promoters of unknown types. Finally, we collected 139 phage promoter-negative and 478 host promoter-positive samples. We randomly chose 80% of these positive and negative samples as the training dataset and 20% as the test dataset. Table 1 lists the dataset parameters for both layers.

Methods

In this section, we briefly explain the proposed model, feature encoding techniques, and baseline models.

Proposed model

The proposed two-layer model is designated "iProm-phage." The model's first layer predicts the query sequence as a phage

promoter or non-promoter. If the predicted sequence is a phage promoter then the model's second layer classifies it as a phage or host. Figure 2 illustrates the proposed model.

Based on performance measures, we opted for the CNN model and one-hot encoding technique for this two-layer predictor. The selection of the model and encoding technique are briefly explained in the performance measure section.

Convolutional neural network model architecture

The CNN is composed of 2 one-dimensional convolutional layers (Conv1D), which are followed by maximum (max) pooling and dropout layers. The filter and kernel sizes of both Conv1D is 16 and 5, respectively. The max pooling size is four with strides of two in both the max pooling layers. A dropout layer is utilized after each max pooling layer, with a value of 0.5. A flattened layer was utilized, followed by a dense layer with 64 nodes. Subsequently, we used a dropout layer with a value of 0.5. The ReLU activation function was utilized in all the Conv1D and dense layers. Finally, the dense layer is employed as an output layer with a single node and sigmoid activation function that classifies the input sequence as positive or negative based on the probability scores. The mathematical expression for the sigmoid activation function is as follows:

$$S(p) = \frac{1}{1+\exp(-p)}$$

We used L2 regularization and bias regularization in the convolution and dense layers to ensure that the model did not overfit. The values for both regularizations were set to 0.0001. The loss function of the model is binary cross-entropy. Adam was used as the optimizer. The batch size was set to 20 with a total of 85 epochs. iProm-phage was created and trained using the Keras framework. The CNN architecture is illustrated in Figure 3.

Feature encoding techniques

A DNA sequence is comprised of the *A*, *C*, *G*, and *T* nucleotides. To perform computational operations, the sequence must be translated into a numerical representation. Feature encoding schemes play a vital role in creating optimal predictors. The input size should be the same for all sequences. We apply the zero-filled method to make every DNA sequence with an equal length of 99 bp. This technique was previously applied by DPProm (Wang et al., 2022). In this study, we find the best feature encoding technique among the 10 different techniques. The details of each encoding scheme are presented below.

One-hot feature encoding

One-hot encoding techniques are used by many state-of-the-art bioinformatics tools (Umarov and Solovyev, 2017; Liu and Li, 2019; Shujaat et al., 2021; Kim et al., 2022). Each nucleotide in a DNA sequence is represented by a four-dimensional vector, which is a vector of zeros with a single one. Nucleotide *A* is encoded as (1,0,0,0), *C* (0,1,0,0), *G* (0,0,1,0), and *T* (0, 0,0,1). Each DNA sequence can be represented by a (99,4) two-dimensional vector.

Nucleotide chemical property feature encoding

The chemical characteristics of the four DNA nucleic acids differ (Jeong et al., 2014). Nucleotides are classified into three types based on their chemical characteristics: hydrogen-bond strength, base type, and functional groups. Purines with two rings are represented by the letters *A* and *G*, whereas pyrimidines with one ring are represented by the letters *C* and *T*. The hydrogen bonds between *A* and *T* are weak, whereas the hydrogen bonds between *C* and *G* are strong. In terms of functional groups, the amino group includes *A* and *C*, whereas the keto group includes *G* and *T*. Each DNA sequence is represented by a three-dimensional vector (*b*, *c*, *p*) based on chemical properties, where n_i denotes the nucleotide *n* at position *i;* hence, *b*, *c*, and, *p* were computed as follows:

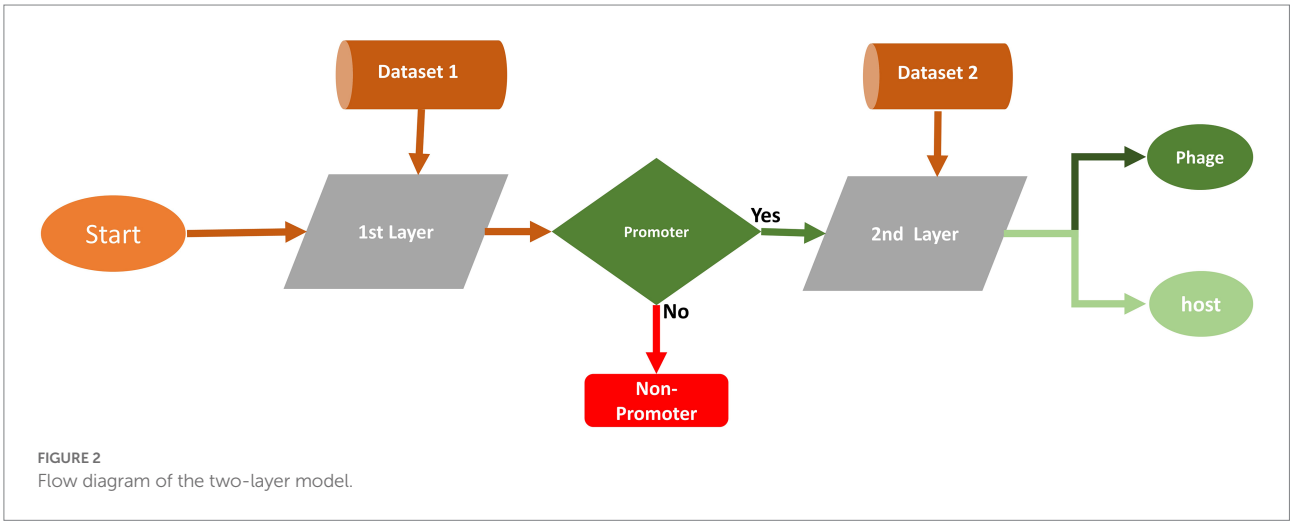

FIGURE 2
Flow diagram of the two-layer model.

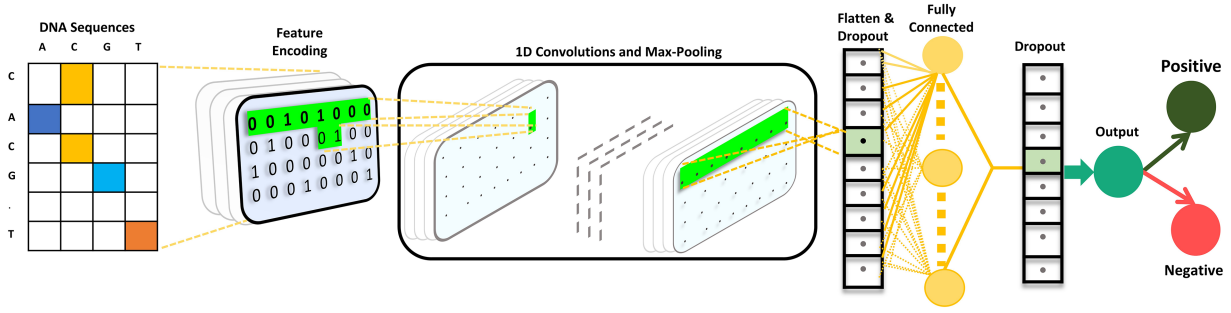

FIGURE 3
iProm-phage CNN architecture.

$$b_i = \begin{cases} 1 \text{ if } n_i \in \{A,C\} \\ 0 \text{ if } n_i \in \{G,T\} \end{cases}, \quad c_i = \begin{cases} 1 \text{ if } n_i \in \{A,G\} \\ 0 \text{ if } n_i \in \{C,T\} \end{cases},$$
$$p_i = \begin{cases} 1 \text{ if } n_i \in \{A,T\} \\ 0 \text{ if } n_i \in \{C,G\} \end{cases}$$

Dinucleotide-based auto-cross covariance feature encoding

DACC is a combination of dinucleotide-based auto-covariance (DAC) and dinucleotide-based cross covariance (DCC) encoding. DAC computes the correlation of the same physicochemical index between two dinucleotides separated by a lag distance along the sequence. DAC is calculated as:

$$\mathrm{DAC}(u,\mathrm{lag}) = \sum_{i=1}^{L-\mathrm{lag}-1} \begin{pmatrix} \left(P_u\left(R_iR_{i+1}\right) - \bar{P}_u\right) \\ \left(P_u\left(R_{i+\mathrm{lag}}R_{i+\mathrm{lag}+1}\right) - \bar{P}_u\right) / \left(L - \mathrm{lag} - 1\right) \end{pmatrix}$$

where u, L represent the physicochemical index and length of the sequence, respectively, and the physicochemical index u for the dinucleotide $\left(R_iR_{i+1}\right)$ at position i is expressed numerically as $P_u\left(R_iR_{i+1}\right)$. $\bar{P}_u$ represents the average value of the physicochemical index u along the whole sequence, and is calculated as:

$$\bar{P}_u = \sum_{j=1}^{L-1} P_u\left(R_jR_{j+1}\right) / \left(L-1\right)$$

The DAC feature vector has a dimension of $N \times \mathrm{LAG}$, where LAG is the maximum lag (lag = 1, 2,…, LAG) and N is the total number of physicochemical indices. DCC computes the correlation of two different physicochemical indices between two dinucleotides along the sequence separated by *lag* nucleic acids. Mathematically, DCC can be represented as

$$\mathrm{DCC}(u_1,u_2,\mathrm{lag}) = \sum_{i=1}^{L-\mathrm{lag}-1} \left(P_{u_1}\left(R_iR_{i+1}\right) - \bar{P}_{u_1}\right) \left(P_{u_2}\left(R_{i+\mathrm{lag}}R_{i+\mathrm{lag}+1}\right) - \bar{P}_{u_2}\right) / \left(L - \mathrm{lag} - 1\right)$$

where u_1, u_2 and L represent the physicochemical indices and length of the nucleotide sequence, respectively, $P_{u_1}\left(R_iR_{i+1}\right)$ is the numerical value of the physicochemical index u_1 for the dinucleotide $\left(R_iR_{i+1}\right)$ at position i, and $\bar{P}_{ua}$ is the average value for the physicochemical index u_a along the whole sequence, calculated as:

$$\bar{P}_{ua} = \sum_{j=1}^{L-1} P_{ua}\left(R_jR_{j+1}\right) / \left(L-1\right)$$

The DCC feature vector has dimensions of $N \times \left(N-1\right) \times \mathrm{LAG}$, where LAG is the maximum lag (lag = 1, 2,.., LAG) and N is the total number of physicochemical indices. Thus, the dimension of the DACC encoding is $N \times N \times \mathrm{LAG}$, where N is the number of physicochemical indices and LAG is the maximum lag (lag = 1, 2, …, LAG).

Pseudo dinucleotide composition

PseDNC encoding incorporates both contiguous local and global sequence order information into a feature vector of the nucleotide sequence. PseDNC is mathematically defined as follows:

$$S = \left[s_1, s_2, \ldots, s_{16}, s_{16+1}, \ldots, s_{16+1}, \ldots, s_{16+\lambda}\right]^T$$

Whereas:

$$s_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j}, \left(1 \le k \le 16\right) \\ \dfrac{w\theta_{k-16}}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j}, \left(17 \le \boldsymbol{k} \le 16+\lambda\right) \end{cases}$$

where f_k $(k = 1, 2, \ldots, 16)$ is the normalized frequency of dinucleotide occurrence in the nucleotide sequence, λ

represents the highest counted rank (or tie) of the correlation along the nucleotide sequence, w is the weight factor ranging from 0 to 1, and θ_j $(j=1,2,\ldots,\ \lambda)$ is the jth correlation factor and is defined as

$$\begin{cases} \theta_1 = \dfrac{1}{L-2}\sum_{i=1}^{L-2}\theta\left(R_iR_{i+1},R_{i+1}R_{i+2}\right) \\ \theta_1 = \dfrac{1}{L-2}\sum_{i=1}^{L-3}\theta\left(R_iR_{i+1},R_{i+2}R_{i+3}\right) \\ \theta_1 = \dfrac{1}{L-2}\sum_{i=1}^{L-4}\theta\left(R_iR_{i+1},R_{i+3}R_{i+3}\right)(\lambda < L) \\ \cdots \\ \theta_\lambda = \dfrac{1}{L-1-\lambda}\sum_{i=1}^{L-1-\lambda}\theta\left(R_iR_{i+1},R_{i+\lambda}R_{i+\lambda+1}\right) \end{cases}$$

The correlation function is given as follows:

$$\theta\left(R_iR_{i+1},R_{j+1}R_{j+1}\right) = \frac{1}{\mu}\sum_{u=1}^{\mu}\left[P_\mu\left(R_iR_{i+1}\right) - P_\mu\left(R_jR_{j+1}\right]^2\right.$$

where physicochemical indices are represented by μ, $P_\mu\left(R_iR_{i+1}\right)$ measures are the numerical values of the u-th ($u=1$, 2, …, μ) physicochemical index of the dinucleotide R_iR_{i+1} at position i and $P_\mu\left(R_jR_{j+1}\right)$ represents the corresponding value of the dinucleotide R_jR_{j+1} at position j .Pseudo k-tupler composition (PseKNC).

PseKNC encoding uses a k-tuple nucleotide composition defined as

$$D = \left[d_1,d_2,\ldots,d_{4^k},d_{4^k+1},\ldots,d_{4^k+\lambda}\right]^T$$

Whereas:

$$\begin{cases} \dfrac{f_u}{\sum_{i=1}^{4^k}f_i + w\sum_{j=1}^{\lambda}\theta_j}, (1 \le u \le 4) \\ \dfrac{w\theta_{u-4^k}}{\sum_{i=1}^{4^k}f_i + w\sum_{j=1}^{\lambda}\theta_j}, \left(4^k \le u \le 4^k+\lambda\right) \end{cases}$$

where λ is the total number of ranks of correlations along a nucleotide sequence, $f_u\left(u=1,2,\ldots,4^k\right)$ is the frequency of oligonucleotides normalized to $\sum_{i=1}^{4^k}f_i = 1$, w is the factor, and θ_j is defined as follows:

$$\theta_j = \frac{1}{L-j-1}\sum_{i=1}^{L-j-1}\Theta\left(R_iR_{i+1},R_{i+j}R_{i+j+1}\right),$$
$$\left(j=1,,,;2,,,;\ldots,,,;\lambda,,,;\lambda < L\right)$$

The correlation function is defined as:

$$\tilde{}\ \left(R_iR_{i+1},R_{i+j}R_{i+j+1}\right) = \frac{1}{\mu}\sum_{v=1}^{\mu}\left[P_v\left(R_iR_{i+1}\right) - P_v\left(R_{i+j}R_{i+j+1}\right)\right]^2$$

where μ represents the physicochemical index. $P_v\left(R_iR_{i+1}\right)$ is a numerical value v-th ($v=1$, 2, …, μ). The physicochemical index of dinucleotide $\left(R_iR_{i+1}\right)$ at position i and $P_v\left(R_{i+j}R_{i+j+1}\right)$ represents the corresponding value of dinucleotide $\left(R_{i+j}R_{i+j+1}\right)$ at position $i+j$.

Electron-ion interaction pseudopotentials of trinucleotide

The values of nucleotides *A*, *G*, *C*, and *T* electron-ion interaction pseudopotentials (EIIP) were determined as previously described using Nair (Lavigne et al., 2004; *A*: 0.1260, *C*: 0.1340, *G*: 0.0806, *T*: 0.1335). Nucleotides in the DNA sequence are directly represented by EIIP using the EIIP value. EIIPA, EIIPT, EIIPG, and EIIPC represent the EIIP values of nucleotides *A*, *T*, *G*, and *C*, respectively, in PseEIIP encoding. A feature vector is created using the mean EIIP value of the trinucleotides in each sample, as follows:

$$V = \left[EIIP_{AAA}\cdot f_{AAA},EIIP_{AAC}\cdot f_{AAC},\ldots,EIIP_{TTT}\cdot f_{TTT}\right]$$

Parallel correlation pseudo dinucleotide composition

Similar to PseDNC, PCPseDNC encoding differs in that it uses 38 default physiochemical indices for DNA instead of the six indices used in PseDNC encoding. Supplementary Table S2 in Supplementary file presents a list of 38 physicochemical indices.

Parallel correlation pseudo trinucleotide composition

PCPseTNC encoding is described as:

$$S = \left[s_1,s_2,\ldots,s_{64},s_{64+1},\ldots,s_{64+\lambda}\right]^T$$

Whereas:

$$s_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{64}f_i + w\sum_{j=1}^{\lambda}\theta_j}, (1 \le k \le 64) \\ \dfrac{w\theta_{k-64}}{\sum_{i=1}^{64}f_i + w\sum_{j=1}^{\lambda}\theta_j}, (65 \le k \le 64+\lambda) \end{cases}$$

where f_k ($k = 1, 2, \ldots, 64$) is the normalized frequency of dinucleotide occurrence in the nucleotide sequence, λ represents the highest counted rank (or tie) of the correlation along the nucleotide sequence, w is the weight factor ranging from 0 to 1, and θ_j ($j = 1,2,\ldots, \lambda$) is the jth correlation factor and is defined as:

$$\begin{cases} \theta_1 = \frac{1}{L-3}\sum_{i=1}^{L-3}\Theta\left(R_iR_{i+1}R_{i+2}, R_{i+1}R_{i+2}R_{i+3}\right) \\ \theta_2 = \frac{1}{L-4}\sum_{i=1}^{L-4}\Theta\left(R_iR_{i+1}R_{i+2}, R_{i+2}R_{i+3}R_{i+4}\right) \\ \theta_3 = \frac{1}{L-5}\sum_{i=1}^{L-5}\Theta\left(R_iR_{i+1}R_{i+2}, R_{i+3}R_{i+4}R_{i+5}\right)\left(\lambda < L\right) \\ \theta_\lambda = \frac{1}{L-2-\lambda}\sum_{i=1}^{L-2-\lambda}\Theta\left(R_iR_{i+1}R_{i+2}, R_{i+\lambda}R_{i+\lambda+1}R_{i+\lambda+2}\right) \end{cases}$$

The correlation function is defined as:

$$\Theta\left(R_iR_{i+1}R_{i+2}, R_{j+1}R_{j+1}R_{j+2}\right) = \frac{1}{\mu}\sum_{u=1}^{\mu}\left[P_u\left(R_iR_{i+1}R_{i+2}\right) - P_u\left(R_jR_{j+1}R_{j+2}\right)\right]^2$$

where physicochemical indices are represented by μ, $P_\mu\left(R_iR_{i+1}R_{i+2}\right)$ measures are the numerical values of the u-th ($u = 1, 2, \ldots, \mu$) physicochemical index of the dinucleotide $R_iR_{i+1}R_{i+2}$ at position i and $P_\mu\left(R_jR_{j+1}R_{j+2}\right)$ represents the corresponding value of the dinucleotide $R_jR_{j+1}R_{j+2}$ at position j.

Moran correlation

The distribution of amino acid characteristics along the sequence is used to create autocorrelation descriptors (Horne, 1988; Feng and Zhang, 2000; Sokal and Thomson, 2006). The amino acid properties used here are different types of amino acid indices retrieved from the AAindex Database (Kawashima et al., 2008) available at http://www.genome.jp/dbget/aaindex.html.

kmer

DNA sequences are represented as the occurrence frequencies of k adjacent nucleic acids in the kmer descriptor, which has been effectively used for human gene regulatory sequence prediction. The kmer descriptor ($k = 3$) is calculated as follows:

$$f(t) = \frac{N(t)}{N}, t \varepsilon \{AAA, AAC, AAG, \ldots, TTT\}$$

where $N(t)$ represents the number of kmer types (t) and N is the length of the sequence.

Baseline models

Selection of the optimal model is a vital step in developing a novel predictor. We have utilized different ML and CNN models and, based on performance measures, selected the best model. ML models include the Adaboost (AdB) classifier, multinomial naive Bayes, extreme gradient boosting (XGboost), gradient boosting (Gboost), logistic regression (LR), K-nearest neighbor, decision tree classifier, support vector machine (SVM), multilayer perceptron classifier, and SVM bagging. A CNN is composed of two convolution layers. We used hyperparameter tuning to determine the best convolution, pooling, dropout, and dense layer parameters.

Performance measures

In this section, we explain the evolution metrics, selection of the best model and feature encoding scheme, model performance, and model comparison.

Evaluation metrics

In the performance assessment matrix, we used the accuracy (Acc), sensitivity (Sn), specificity (Sp), and MCC. These parameters have been used in several cutting-edge studies. The numerical representation of an evaluation matrix is expressed using the following equations:

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Sn} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Sp} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

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

The terms TP, TN, FP, and FN in the aforementioned equations represent the appropriate numbers of true positives, true negatives, false positives, and false negatives, respectively.

Selection of best model and feature encoding

To generate an optimum model, we compared all the encoding strategies stated above to the baseline approaches. Supplementary Tables S3, S4 in Supplementary file, and

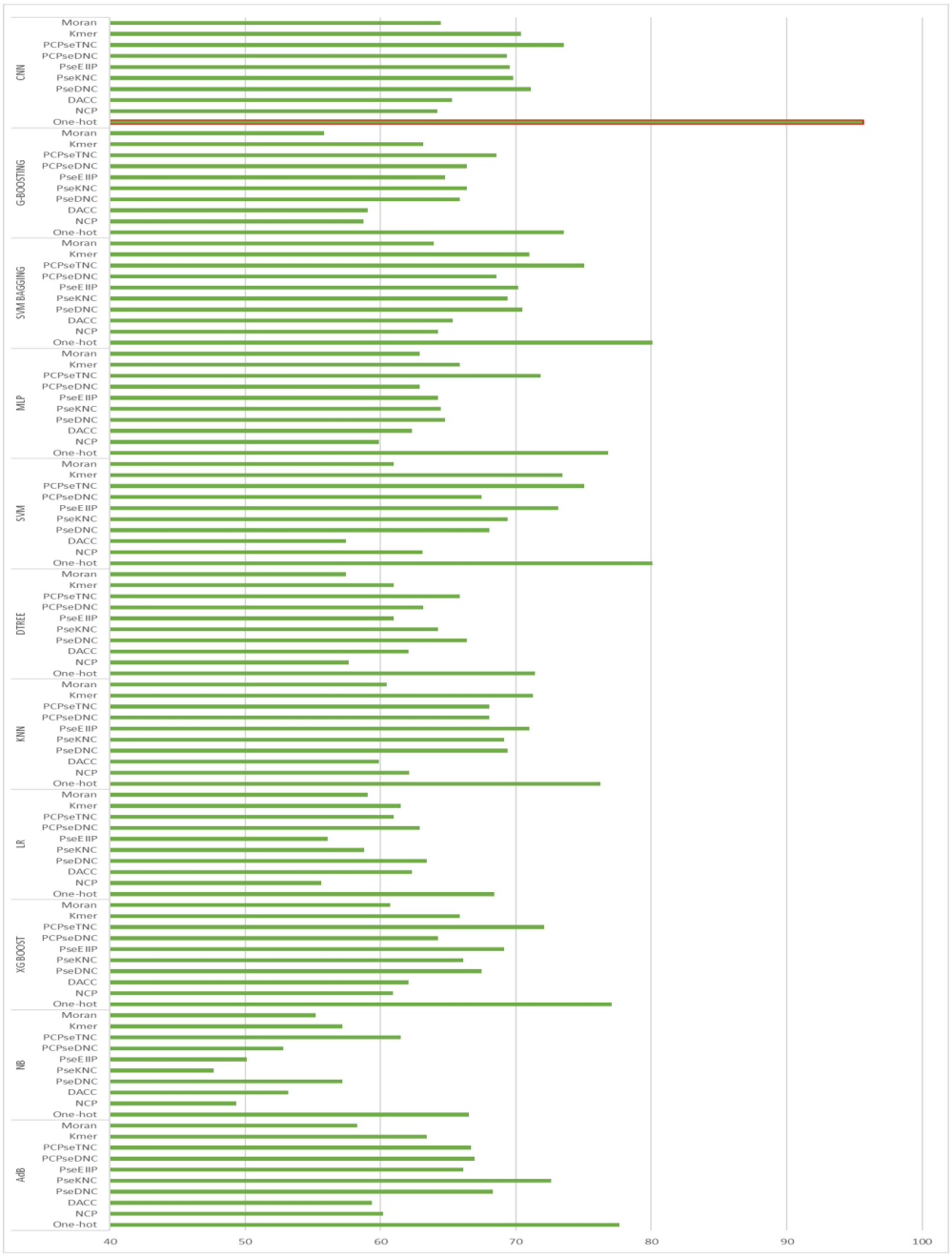

FIGURE 4
Accuracy of First layer baseline models.

Figures 4, 5 illustrate the performance of each method on various encoding schemes for the first and second layers. For the first layer of the model CNN and one-hot encoding outperformed after that AdB performed better on PseKNC feature encoding and for the second layer almost every feature encoding scheme performed good on ML and CNN algorithms, but one-hot and CNN outperformed in the second layer as well. Therefore, based on performance evaluation, we chose the CNN and one-hot encoding technique for both layers and the proposed tool "iProm-phage."

Model performance

The prediction performance of iProm-phage was evaluated using 5-fold cross validation. We employed the same parameters

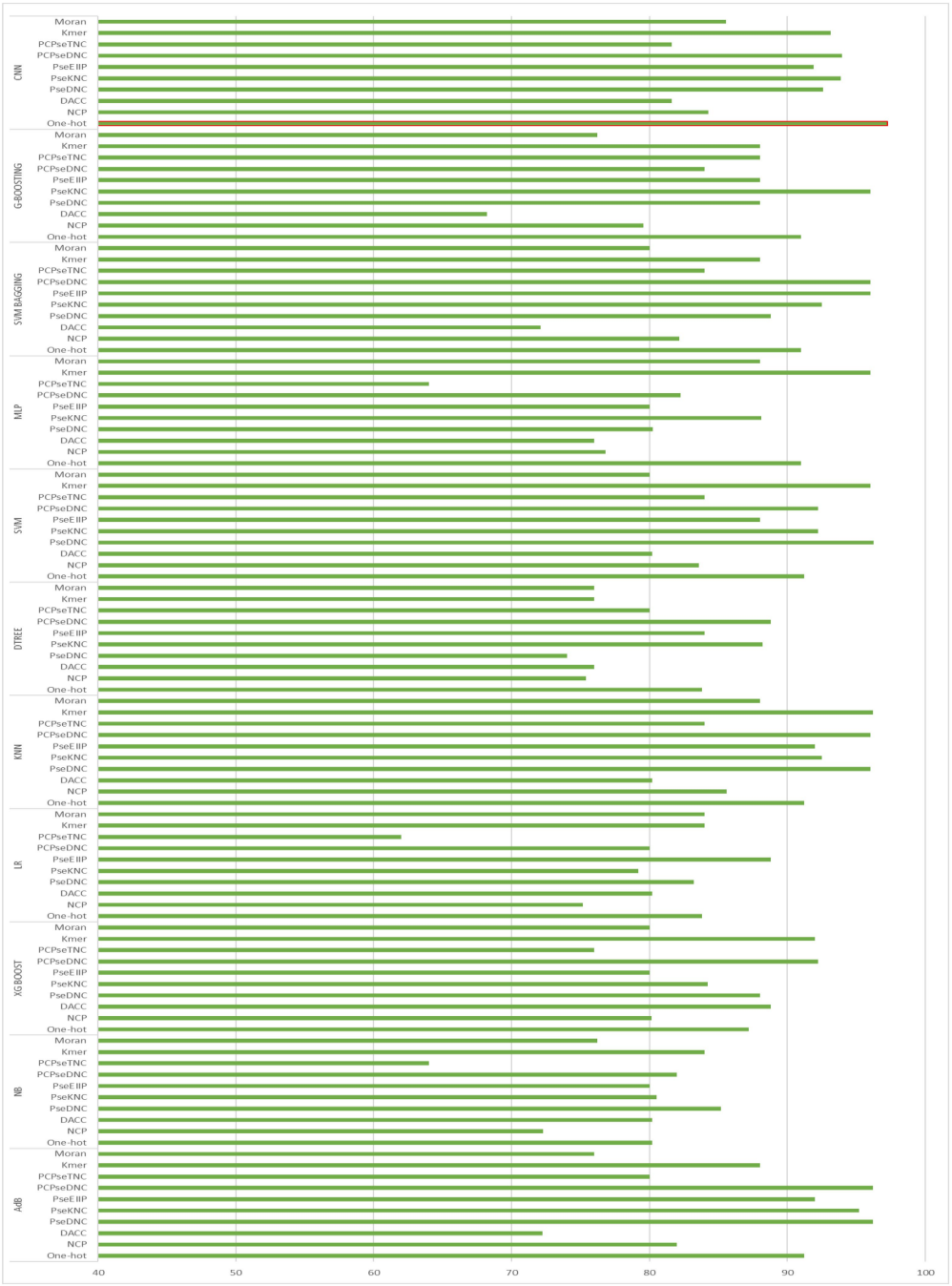

FIGURE 5
Accuracy of Second layer baseline models.

used in choosing the best model and also considered ROC curve data. The first layer of iProm-phage achieved an Acc of 95.68 93.47%, Sn of 96.12%, Sp of 92.63%, MCC of 0.872, and AUROC of 0.99 during cross validation. These findings suggest that our predictor is capable of properly recognizing whether a query sequence is a promoter. The second layer of iProm-Zea achieved values of 97.25, 94.32, 98.5%, 0.8619, and 0.97, respectively. In the test dataset model, the first layer achieved an accuracy of 94.2%, Sn 90%, Sp 90%, and MCC 0.88. The second layer obtained accuracies of 95.2%, 94.37%, 97.14%, and 0.88% for the test dataset. Figures 6, 7 depict the ROC curves for both layers of the iProm-phage model.

Comparison with existing models

We compared iProm-phage with state-of-the-art promoter identification tools PhagePromoter and DPProm for the identification of query sequences as promoters or promoters.

TABLE 2 First layer performance comparison.

Methods	Acc%	Precision%	Recall%
PhagePromoter	92	89	87
DPProm	85.5	88.9	83
iProm-phage	95.68	94.2	93.5

TABLE 3 Second layer performance comparison.

Methods	Acc%	Precision%	Recall%
DPProm	93.0	95.2	96.4
iProm-phage	95.2	96.5	97.2

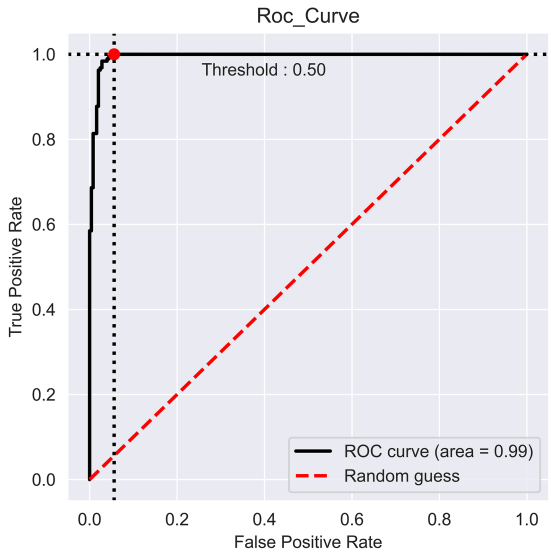

FIGURE 6
First layer ROC curve.

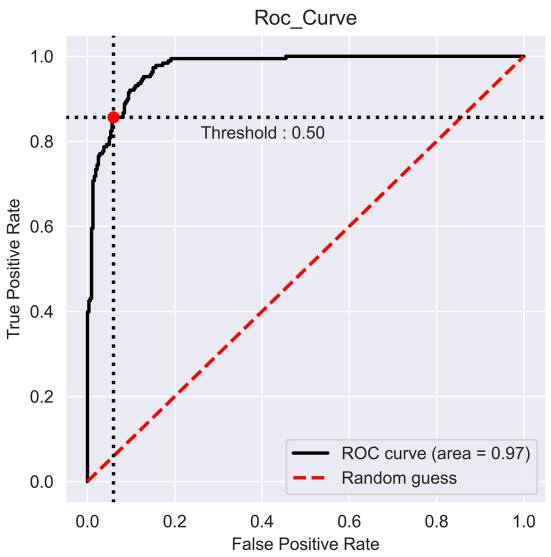

FIGURE 7
Second layer ROC curve.

We measured the precision and recall for both layers to compare them with state-of-the-art methods. The following equations express precision and recall:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Precison} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

A performance comparison of the methods used for promoter identification is presented in Table 2. The superior performance of the proposed iProm-phage tool can be observed in all four performance metrics for this particular task.

We demonstrate the performance comparison between DPProm in Table 3 for promoter classification as a phage or host. The iProm-phage tool was superior to DPProm in performance for all classification tasks. The precision and recall of iProm-phage for promoter identification and classification were higher than those of DPProm, and the values were more consistent. As a result, iProm-phage showed a considerably higher score than the state-of-the-art methods in all cases.

Webserver

A web server hosting the high performance iProm-phage tool is freely available at the following link[1] to enable easy access to the proposed tool for the scientific community. This approach has been adopted by several scholars (Chantsalnyam et al., 2020; Ali SD et al., 2022). iProm-phage is an easy-to-use tool that can be utilized by researchers and specialists in bioinformatics. It consists of two stages first is input and second is output. To input it uses two input methods: direct sequence input and uploading a file containing sequences for prediction. Each sequence should be 99 bp long and contain the letters *A*, *C*, *G*, and *T*. Figures 8, 9 depict web server snippets; Figure 8 is an example of adding sequences for prediction and Figure 9 provides the predictor's output. We also provide an example to better understand how to use the webserver.

1 http://nsclbio.jbnu.ac.kr/tools/iProm-phage/

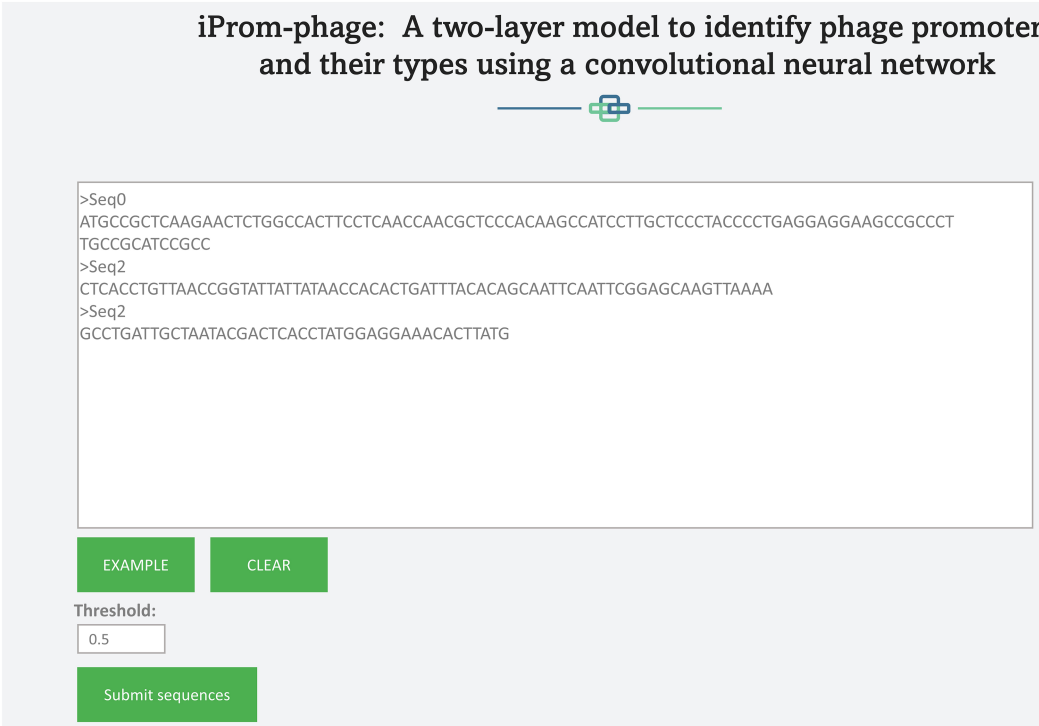

FIGURE 8
Webserver adding query sequence.

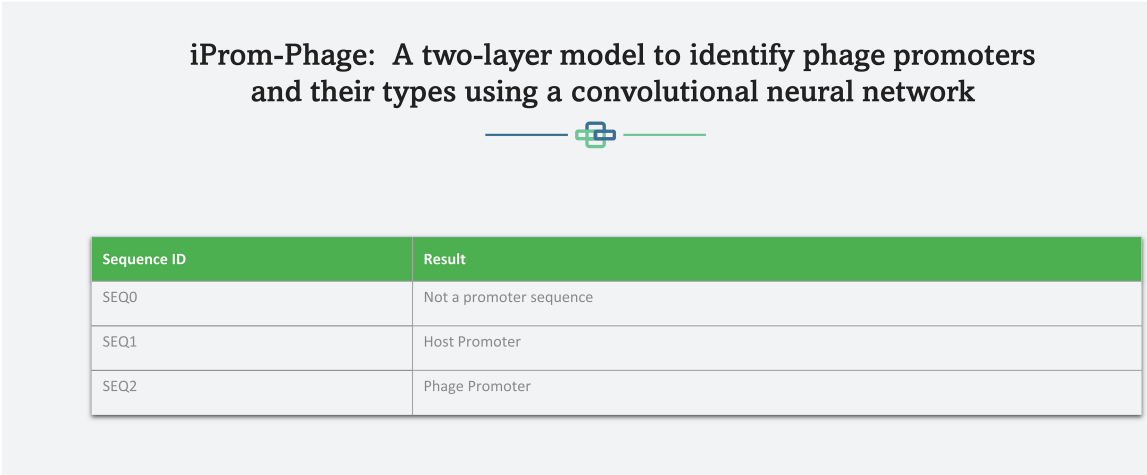

FIGURE 9
Predictor output.

Conclusion

This work presents iProm-phage, a two-layer technique for identifying phage promoters and classifying them as phages or hosts. We developed a new method for generating negative datasets to create a robust model that performs well on tough datasets. Based on cutting-edge performance tests, we also found the best model among several ML and CNN algorithms, as well as the best feature encoding method among the 10 distinct methods. The architecture of the proposed model was evaluated using publicly available datasets. Compared to earlier techniques, the program had superior overall results. Finally, we created a web server that is available online and will be extremely useful to other experimental scientists.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

MS: conceptualization, methodology, software, writing–original draft, and writing–review and editing. JJ: methodology and writing–review and editing. HT: supervision and writing–review and editing. KC: conceptualization, validation, supervision, writing–review and editing, and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT; nos. 2020R1A2C2005612 and 2022R1G1A1004613). This work was supported by "Human Resources Program in Energy Technology" of the Korea Institute of Energy Technology Evaluation and Planning (KETEP), granted financial resource from the Ministry of Trade, Industry & Energy, Republic of Korea (no. 20204010600470).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.1061122/full#supplementary-material

References

Ali, S. D., Alam, W., Tayara, H., and Chong, K. (2020). Identification of functional pi RNAs using a convolutional neural network. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 14:1. doi: 10.1109/tcbb.2020.3034313

Ali, S. D., Alam, W., Tayara, H., and Chong, K. T. (2022). Identification of functional piRNAs using a convolutional neural network. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 19, 1661–1669. doi: 10.1109/TCBB.2020.3034313

Chantsalnyam, T., Lim, D. Y., Tayara, H., and Chong, K. T. (2020). ncRDeep: non-coding RNA classification with convolutional neural network. *Comput. Biol. Chem.* 88:107364. doi: 10.1016/j.compbiolchem.2020.107364

Feng, Z. P., and Zhang, C. T. (2000). Prediction of membrane protein types based on the hydrophobic index of amino acids. *J. Protein Chem.* 19, 269–275. doi: 10.1023/A:1007091128394

Guzina, J., and Djordjevic, M. (2015). Bioinformatics as a first-line approach for understanding bacteriophage transcription. *Bacteriophage* 5:e1062588. doi: 10.1080/21597081.2015.1062588

Horne, D. S. (1988). Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. *Biopolymers* 27, 451–477. doi: 10.1002/bip.360270308

Jeong, B.-S., Golam Bari, A. T. M., Rokeya Reaz, M., Jeon, S., Lim, C.-G., and Choi, H.-J. (2014). Codon-based encoding for DNA sequence analysis. *Methods* 67, 373–379. doi: 10.1016/j.ymeth.2014.01.016

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* 36, D202–D205. doi: 10.1093/nar/gkm998

Kim, J., Shujaat, M., and Tayara, H. (2022). Iprom-zea: a twolayer model to identify plant promoters and their types using convolutional neural network. *Genomics* 114:110384. doi: 10.1016/j.ygeno.2022.110384

Klucar, L., Stano, M., and Hajduk, M. (2010). Phi SITE: database of gene regulation in bacteriophages. *Nucleic Acids Res.* 38, D366–D370. doi: 10.1093/nar/gkp911

Lavigne, R., Sun, W. D., and Volckaert, G. (2004). PHIRE, a deterministic approach to reveal regulatory elements in bacteriophage genomes. *Bioinformatics* 20, 629–635. doi: 10.1093/bioinformatics/btg456

Liu, B., and Li, K. (2019). Ipromoter-2l2. 0: identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features. *Mol. Ther. Nucleic Acids* 18, 80–87. doi: 10.1016/j.omtn.2019.08.008

Mishra, A., Dhanda, S., Siwach, P., Aggarwal, S., and Jayaram, B. (2020). A novel method seprom for prokaryotic promoter prediction based on dna structure and energetics. *Bioinformatics* 36, 2375–2384. doi: 10.1093/bioinformatics/btz941

Rahman, M. S., Aktar, U., Jani, M. R., and Shatabda, S. (2019). iPro70-FMWin: identifying sigma 70 promoters using multiple windowing and minimal features. *Mol. Gen. Genomics.* 294, 69–84. doi: 10.1007/s00438-018-1487-5

Rehman, M. U., Hong, K. J., Tayara, H., and Chong, K. T. (2021). To Chong, m6A-neural tool: convolution neural tool for RNA N6-methyladenosine site identification in different species. *IEEE Access* 9, 17779–17786. doi: 10.1109/ACCESS.2021.3054361

Salmond, G. P., and Fineran, P. C. (2015). A century of the phage: past, present and future. *Nat. Rev. Microbiol.* 13, 777–786. doi: 10.1038/nrmicro3564

Sampaio, M., Rocha, M., Oliveira, H., and Dias, O. (2019). Predicting promoters in phage genomes using phage promoter. *Bioinformatics* 35, 5301–5302. doi: 10.1093/bioinformatics/btz580

Shujaat, M., Lee, S. B., Tayara, H., and Chong, K. T. (2021). Crprom: a convolutional neural network-based model for the prediction of rice promoters. *IEEE Access* 9, 81485–81491. doi: 10.1109/ACCESS.2021.3086102

Shujaat, M., Wahab, A., Tayara, H., and Chong, K. T. (2020). Chong, pc promoter-CNN: a CNN-based prediction and classification of promoters. *Genes (Basel)* 11:1529. doi: 10.3390/genes11121529

Sierro, N., Makita, Y., de Hoon, M., and Nakai, K. (2008). Dbtbs: a database of transcriptional regulation in bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res.* 36, D93–D96. doi: 10.1093/nar/gkm910

Silva, S., and Echeverrigaray, S. (2012). Bacterial promoter features description and their application on *E. coli* in silico prediction and recognition approaches. *Bioinformatics. InTech* 1, 241–260. doi: 10.5772/48149

Sokal, R. R., and Thomson, B. A. (2006). Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. *Am. J. Phys. Anthropol.* 129, 121–131. doi: 10.1002/ajpa.20250

Umarov, R. K., and Solovyev, V. V. (2017). Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS One* 12:e0171410. doi: 10.1371/journal.pone.0171410

Wang, Y., Wang, H., Wei, L., Li, S., Liu, L., and Wang, X. (2020). Synthetic promoter design in escherichia coli based on a deep generative network. *Nucleic Acids Res.* 48, 6403–6412. doi: 10.1093/nar/gkaa325

Wang, C., Zhang, J., Cheng, L., Wu, J., Xiao, M., Xia, J., et al. (2022). DPProm: a two-layer predictor for identifying promoters and their types on phage genome using deep learning. *IEEE J. Biomed. Health Inform.* 26, 5258–5266. doi: 10.1109/JBHI.2022.3193224

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 14 November 2022
DOI 10.3389/fmicb.2022.1042127

OPEN ACCESS

EDITED BY
Yongchun Zuo,
Inner Mongolia University,
China

REVIEWED BY
Jian-Yu Shi,
Northwestern Polytechnical University,
China
Quan Zou,
University of Electronic Science and Technology of China, China
Shravan Sukumar,
Corteva Agriscience™, United States
Masaya Fujita,
University of Houston,
United States
Mario Andrea Marchisio,
Tianjin University,
China

*CORRESPONDENCE
Gajendra P. S. Raghava
raghava@iiitd.ac.in

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 12 September 2022
ACCEPTED 27 October 2022
PUBLISHED 14 November 2022

CITATION
Patiyal S, Singh N, Ali MZ, Pundir DS and Raghava GPS (2022) Sigma70Pred: A highly accurate method for predicting sigma70 promoter in *Escherichia coli* K-12 strains. *Front. Microbiol.* 13:1042127. doi: 10.3389/fmicb.2022.1042127

Sigma70Pred: A highly accurate method for predicting sigma70 promoter in *Escherichia coli* K-12 strains

Sumeet Patiyal[1†], Nitindeep Singh[2†], Mohd Zartab Ali[2†], Dhawal Singh Pundir[2†] and Gajendra P. S. Raghava[1*]

[1]Department of Computational Biology, Indraprastha Institute of Information Technology Delhi, New Delhi, India, [2]Department of Computer Science and Engineering, Indraprastha Institute of Information Technology Delhi, New Delhi, India

Sigma70 factor plays a crucial role in prokaryotes and regulates the transcription of most of the housekeeping genes. One of the major challenges is to predict the sigma70 promoter or sigma70 factor binding site with high precision. In this study, we trained and evaluate our models on a dataset consists of 741 sigma70 promoters and 1,400 non-promoters. We have generated a wide range of features around 8,000, which includes Dinucleotide Auto-Correlation, Dinucleotide Cross-Correlation, Dinucleotide Auto Cross-Correlation, Moran Auto-Correlation, Normalized Moreau-Broto Auto-Correlation, Parallel Correlation Pseudo Tri-Nucleotide Composition, etc. Our SVM based model achieved maximum accuracy 97.38% with AUROC 0.99 on training dataset, using 200 most relevant features. In order to check the robustness of the model, we have tested our model on the independent dataset made by using RegulonDB10.8, which included 1,134 sigma70 and 638 non-promoters, and able to achieve accuracy of 90.41% with AUROC of 0.95. Our model successfully predicted constitutive promoters with accuracy of 81.46% on an independent dataset. We have developed a method, Sigma70Pred, which is available as webserver and standalone packages at https://webs.iiitd.edu.in/raghava/sigma70pred/. The services are freely accessible.

KEYWORDS

sigma70 factor, promoter, machine learning, transcription, prokaryotic genome

Introduction

Promoters and enhancers regulate the fate of a cell by regulating the expression of the genes. Promoters are generally located at the upstream of genes' transcription start sites (TSS) responsible for the switching on or off the respective gene. In prokaryotes, promoters are recognized by the holoenzyme, which is made up of RNA polymerase and a related sigma factor. There are various types of sigma factors responsible for different functions, such as sigma54 controls the transcription of genes responsible for the modulation of cellular nitrogen levels, sigma38 regulates the stationary phase genes, sigma32 regulates heat-shock genes, and sigma24 and sigma18 controls the extra-cytoplasmic functions (Paget, 2015). The number associated with each sigma factor represents the molecular weight. Sigma70 factor is a crucial

factor as it regulates the transcription of most of the housekeeping genes and responsible for the most of the DNA regulatory functions. Sigma70 promoter comprises two well-defined short sequences located at-10 and-35 base pairs upstream of TSS, known as pribnow box and-35 region, respectively (Paget and Helmann, 2003). It is essential to identify the promoter regions in a genome, as it can aid in illuminating the genome's regulatory mechanism and disease-causing variants within cis-regulatory elements. The area of the promoters is of great interest as researchers pay great attention to their importance not only in developmental gene expression but also in environmental response. To control the expression of every gene and transcription unit in the genome, promoters must be precisely identified, and in terms of consensus sequences, promoter sequences may differ and be comparable within and between the different classes of promoters. However, since each promoter often deviates from the consensus at one or more locations, it is still difficult to predict promoters with reliable accuracy (Mrozek et al., 2014, 2016). Moreover, due to the advancement in sequencing technology, the data is growing exponentially, which made the accurate identification of promoter regions in the DNA sequences a difficult task. Of note, the accurate and fast classification of the promoter region is a crucial problem, as the standard experimental procedures are expensive in terms of time, and performance (Bernardo et al., 2009; Lu et al., 2015).

In the past, ample of methods have been developed for predicting sigma70 promoters which are based on different machine-and deep-learning approaches developed using various types of features (Lin and Li, 2011; Song, 2012; He et al., 2018; Liu et al., 2018; Lai et al., 2019; Lin et al., 2019; Liu and Li, 2019; Zhang et al., 2019). IMPD (Lin and Li, 2011), is based on the increment of diversity, which achieved an accuracy of 87.9%. This method was trained on RegulonDB (Gama-Castro et al., 2016) dataset that contains 741 *E. coli* sigma70 promoters. Z-curve-based approach (Song, 2012) attains the maximum accuracy of 96.1% by using a smaller dataset that comprises 576 sigma70 promoters and 1,661 non-promoters. Liu et al. (2018) proposed a two-layer prediction method, named as iPromoter-2L, for the identification and classification of multiple sigma promoters using the multi-window-based pseudo K-tuple nucleotide composition approach and achieved the highest accuracy of 81.68% for sigma70 promoter prediction. 70Propred (He et al., 2018) has incorporated features like position-specific trinucleotide propensity based on single-stranded characteristic (PSTNPss) and electron-ion potential values for trinucleotides (PseEIIP) using benchmark dataset of 741 sigma70 promoters and 1,400 non-promoters from RegulonDB9.0, and reported 95.56% accuracy. iPro70-PseZNC (Lin et al., 2019) is based on a multi-window Z-curve approach and gained the maximum accuracy of 84.5% using the dataset from RegulonDB9.0 (Gama-Castro et al., 2016). iPromoter-2L2.0 (Liu and Li, 2019) is an update of iPromoter-2L, which implemented the combination of smoothing cutting window algorithm and sequence-based features to improve the performance with accuracy 85.94%.

The aforementioned methods are developed using traditional machine learning approaches such as logistic regression (Rahman et al., 2019a), support vector machine (He et al., 2018; Lai et al., 2019; Lin et al., 2019; Liu and Li, 2019; Zhang et al., 2019), random forest (Liu et al., 2018), ensemble of different classifiers (Rahman et al., 2019b). On the other hand, due to the advancement in the computational and sequencing technology, deep convolutional neural network (CNN) based methods have been implemented to develop the prediction methods with the ability to identify the sigma promoters and then determines the different types of sigma promoter sequences such as sigma24, sigma28, sigma32, sigma38, sigma54, and sigma70. Amin et al. proposed a method, iPromoter-BnCNN (Amin et al., 2020), is a branched-CNN based method which utilized the sequence and structural based properties to identify and classify the sigma promoters. Shujaat et al. (2020) introduced pcPromoter-CNN which convert the nucleotide sequence information into one-hot encoding vectors and feed them to convolutional neural network (CNN)-based classifier to predict and determine the sigma promoter classes. Recently, a new method based on the light CNN named as PromoterLCNN was proposed by Hernandez et al. (2022) which also used one-hot encoding representation of nucleotide sequences to predict the sigma promoters using the sequencing information. The correct prediction of sigma70 promoters in the DNA sequences is still a difficult challenge due to the intraclass variation in terms of consensus sequence as sigma70 factor is responsible for the transcription of the most of the regulatory genes. Albeit, number of computational methods are available to predict the sigma70 promoters using the sequence information, but there is a still enough room for the improvement in term of various performance measures.

In the present study, we have developed a computational method called as Sigma70Pred, to classify the sequences in sigma70 promoter and non-promoter. In this study, we have trained and evaluated the prediction model on the benchmark dataset which have been used in ample of previously published methods such as 70Propred, iPro70-FMWin, iPro70-PseZnc, IPMD, iProEP, and iPromoter-FSEn. In order to investigate the validity of the generated model, we have also created the independent dataset with no common sequences with the benchmark dataset. We calculated the performance of the proposed method on the independent dataset and also compared it with the working existing methods. A user-friendly and freely accessible web server and Python and Perl-based standalone software have been developed to serve the scientific community for predicting the sigma70 promoters. Moreover, the same package has also been distributed *via* docker-based technology through GPSRdocker (Agrawal et al., 2019).

Materials and methods

Dataset generation

The choice of a standard benchmark dataset is a crucial first step in developing a prediction method. In this study, we have used the high-quality pre-constructed benchmark dataset, which has been

used previously published studies such as, 70Propred (He et al., 2018), iPro70-FMWin (Rahman et al., 2019a), iPro70-PseZNC (Lin et al., 2019), iProEP (Lai et al., 2019), IPMD (Lin and Li, 2011), and iPromoter-FSEn (Rahman et al., 2019b). We have trained and tested our models using cross-validation, on the benchmark dataset downloaded from RegulonDB9.0 (Gama-Castro et al., 2016), which is one of the best available databases on bacterial gene regulation in the model organism *E. coli.* K-12. It contains 741 sigma70 promoters and 1,400 non-promoters from the *E. coli.* K-12 genome, and each sequence is of length 81 bp. Due to the lack of sufficient experimentally verified negative data (that is, the locations that are identified not to be transcription start site), randomly generated sequences from the same chromosome have been obtained in the benchmark dataset to generate the non-promoter sequences. As shown by Gordon et al., 81% of the transcription start sites are located at the intergenic non-coding regions and 19% are available in the coding region (Gordon et al., 2003). Therefore, number of methods used the middle regions of long coding sequences of *E. coli.* K-12 genome to create the negative/non-promoter dataset (Shujaat et al., 2020; Hernandez et al., 2022), whereas, other methods used both the coding and non-coding regions to extract non-promoter sequences (Lin and Li, 2011; He et al., 2018; Lai et al., 2019; Liu and Li, 2019; Rahman et al., 2019a,b; Amin et al., 2020). In the benchmark dataset used in this study, half of the negative samples or non-promoter sequences were extracted from the coding and rest half were obtained from convergent intergenic spacers (non-coding regions). In order to validate our model on external or independent dataset, we have extracted the data from RegulonDB 10.8, which comprises 1,134 sigma70 and 638 non-promoters. There is no identical sequence in training and independent dataset. The datasets can be downloaded from our server.

Overall workflow

The comprehensive workflow for Sigma70Pred is shown in Figure 1.

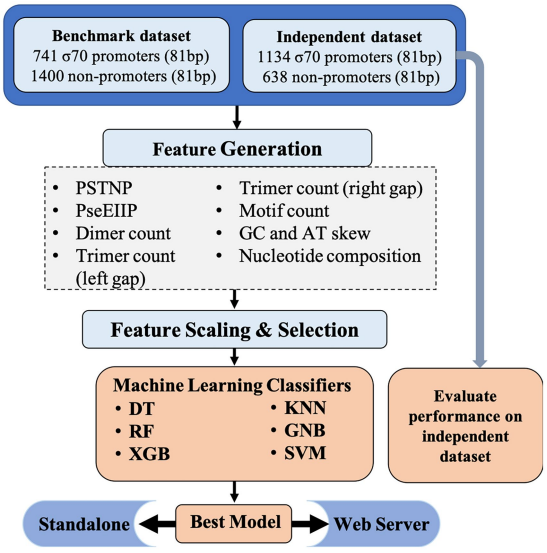

FIGURE 1
Architecture of sigma70Pred.

Feature generation

We have generated a wide range of features like Position-Specific Tri-Nucleotide Propensity (PSTNPP), Electron-Ion Interaction Pseudopotentials of trinucleotide (EIIIP; He et al., 2018), dimer count, trimer count, motif counts, GC and AT skew (Rahman et al., 2019a), Dinucleotide Auto-Correlation (DAC), Dinucleotide Cross-Correlation (DCC), Dinucleotide Auto Cross-Correlation (DACC; Friedel et al., 2009), Moran Auto-Correlation (MAC), Normalized Moreau-Broto Auto-Correlation (NMBAC; Chen et al., 2015), and Parallel Correlation Pseudo Tri-Nucleotide Composition (PC_PTNC; Liu et al., 2014), which resulted in 8465 features. The aforementioned features were calculated using Nfeature webserver (Mathur et al., 2021) available at https://webs.iiitd.edu.in/raghava/nfeature/. Then, we have used the Min-Max scaler from the scikit-learn library (Pedregosa et al., 2011) to scale down the values of the features, we have constructed. Further, we have implemented Recursive Feature Elimination (RFE; Pedregosa et al., 2011) for the feature selection with logistic regression as the estimator and step-size 10. RFE is a wrapper-style technique, i.e., we have used logistic regression algorithm which is wrapped by RFE, to choose features by iteratively considering smaller sets of features progressively. First, the classifier is trained on the initial set of features and importance of each feature is calculated. Further, the features with least importance are eliminated from the current set of features. This process is recursively repeated on the current feature-set until we are left with the desired number of features. Less number of features can make the models developed using machine learning classifiers, more efficient and effective in terms of space and complexity. It also aid the model to achieve the better predictive performance by avoid learning on the irrelevant input features. Details of each feature and processing of the features are explained in the Supplementary File. The comprehensive details of the top-200 features are reported in Supplementary Table S1, where we have provided the description of each feature along with their mean in sigma70-promoter and non-promoter sequences and value of p to check if the difference is significant or not. The features are sorted as per their importance which is calculated using the random forest based classifiers and top-20 features are plotted as per their rank in Supplementary Figure S1.

Model development

In this study, we developed models for predicting sigma70 promoters using wide range of machine learning techniques such

as decision tree (DT), random forest (RF), k-nearest neighbor (KNN), extreme gradient boosting (XGB), gaussian Naïve Bayes (GNB), and support vector machine (SVM; Pedregosa et al., 2011). We got the best performance using SVM based model. Our best model on training dataset was evaluated on independent dataset (obtained from RegulonDB 10.8).

Cross-validation

In order to avoid the biasness and test the prediction models' performance, we have implemented five-fold cross-validation. In this approach, the complete dataset is divided into five parts, the model is trained on four out of five parts, whereas the model is tested on the left part, and the performance is recorded. The same process is iterated five times so that each part gets the chance to be used for the purpose of testing. The overall performance is calculated by taking the mean of all five iterations (Patiyal et al., 2020).

Measures of performance

To assess the performance of generated prediction models, we have used various threshold-dependent and independent parameters. We have considered sensitivity that is, percent of sigma70 samples classified correctly; specificity that is, percent of non-promoter samples classified as negative; accuracy that is, percentage of samples which are correctly predicted by the model; and Matthews correlation coefficient (MCC) that explains the relationship between the observed and predicted value, under threshold-dependent parameters, whereas, in threshold-independent measures, we have considered Area Under the Receiver Operating Characteristics (AUROC) which is the relation between true positive rate and false positive rate. The AUROC was computed using the pROC package (Sachs, 2017) of R. The equations depicting the threshold-dependent parameters are as follows:

$$\text{Sensitivity} = \frac{P_T}{P_T + N_F} \quad (1)$$

$$\text{Specificity} = \frac{N_T}{N_T + P_F} \quad (2)$$

$$\text{Accuracy} = \frac{P_T + N_T}{P_T + P_F + N_T + N_F} \quad (3)$$

$$\text{MCC} = \frac{(P_T * N_T) - (P_F * N_F)}{(P_T + P_F)(P_T + N_F)(N_T + P_F)(N_T + N_F)} \quad (4)$$

where, P_T refers to number of true positives; P_F refers to number of false positives; N_T refers to number of true negatives; and N_F refers to number of false negatives.

Results and discussion

Compositional analysis

In order to assess the proportion of the nucleotides in the sigma70 promoter and non-promoter, we have calculated the mono-nucleotide composition. As shown in Figure 2, nucleic acid adenine and thymine are abundant in sigma70 promoter sequences, whereas cytosine and guanine are higher in percentage in the case of non-promoter sequences.

Position conservation analysis

In this analysis, we explored the preference of each nucleotide at each position of the sigma70 promoter sequences. For the same, we have created the one-sample and two-sample logo using WebLogo (Crooks et al., 2004) and Two Sample Logo (TSL) tool (Vacic et al., 2006). One Sample logo reports the abundance of nucleotides at each position in a single dataset (i.e., positive/negative dataset), whereas TSL takes two files as input (i.e., positive dataset and negative dataset) to exhibits the preference of nucleotides in the positive dataset in comparison to the negative dataset. Therefore, we have provided sigma70 promoter sequences in the FASTA format to WebLogo tool to generate the one-sample logo, and provided both the files, i.e., sigma70 promoter and non-promoter sequences in the FASTA format to TSL tool. Figure 3A represents the one sample sequence logo and Figure 3B exhibits the two-sample logo for sigma70 promoter sequences. In Figure 3A, consensus short sequences "TATAAT" and "TTGACA" at position-10 and-35, respectively, is blurred due to the variability in the spacing between these regions (Shultzaberger et al., 2007), as we have taken all the sequences to generate the sequence logo. However, the region around-10 and-35 is abundant with the nucleotides involve in the consensus sequences at-10 and-35. As shown in Figure 3B, sigma70 promoter sequences are enriched in "A" and "T" nucleotides at most of the positions, whereas, depleted in nucleotides "G" and "C." "T" is most abundant nucleotide at positions −59, −56, −50, −49, −40, −38 to-34, −28, −22, −19, −15, −14, −6, −5, +5, and + 11. Whereas nucleotide "A" is preferable at positions −60, −58, −57, −52, −45, −3, +6, +8, +14, +15, +17, and + 18 in the sigma70 promoter sequences. On the other hand, at positions −13, 0, and + 20 nucleotide "G" is also preferred, and positions −2, −1, and + 1 are also occupied with nucleotide "C." Whereas, on the rest of the positions, both "A" and "T" are the most abundant nucleotides in the sigma70 promoter sequences, as shown in Figure 3B. In order to represent the-10 and-35 consensus sequence, we have generated the motif using MEME software (Bailey et al., 2009) and highlighted the sigma70 promoters' conserved sequences "TATAAT" and "TTGACA" in Supplementary Figure S2.

Performance of machine learning classifiers on benchmark dataset

Initially, we have generated more than 8,000 nucleotide-based features, and then selected 200 most relevant features

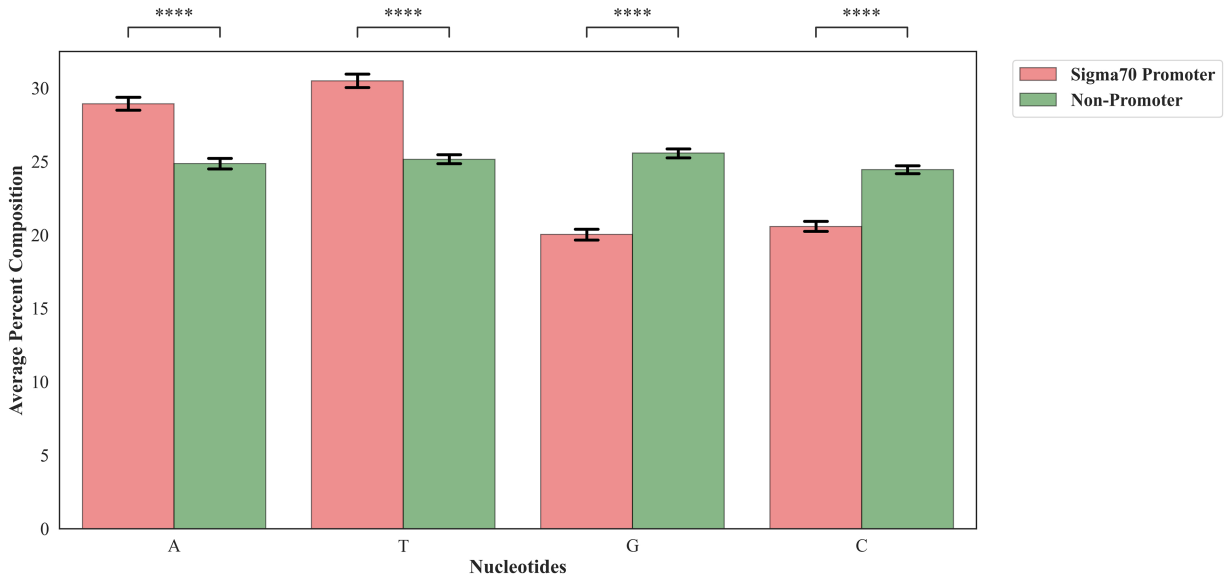

FIGURE 2
Mono-nucleotide composition of sigma70 promoters and non-promoters. ****p<0.0001.

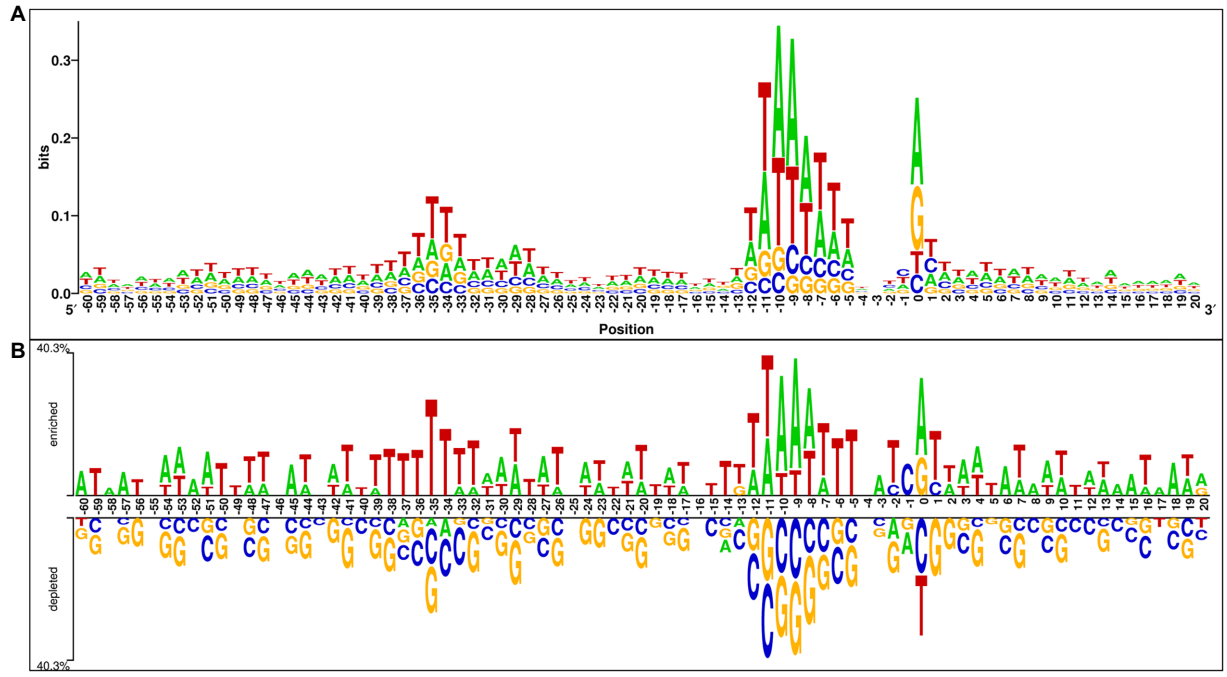

FIGURE 3
Positional preference analysis **(A)** One sample logo exhibiting nucleotide preference in sigma70 promoter sequences at different positions. **(B)** Two sample logo to exhibit the preference of nucleotides at each position in sigma70 promoter sequences with respect to non-promoter sequences.

after applying feature scaling method min-max scaler and feature selection method RFE. Using these selected features, we have generated various models by implementing various machine learning techniques. To compare the performance of each generated model, we have calculated different performance measures as reported in Table 1. The model developed using SVM-based classifier performed best among all the other classifiers with 97.38% accuracy, 0.996 AUROC, and 0.94 MCC on the benchmark dataset.

Performance comparison with existing methods on benchmark dataset

There are ample of methods which are trained and evaluated on the same benchmark dataset such as, 70ProPred (He et al., 2018), iPro70-FMWin (Rahman et al., 2019a), iPro70-PseZNC (Lin et al., 2019), Z-Curve (Song, 2012), IPMD (Lin and Li, 2011), iProEP (Lai et al., 2019), and iPromoter-FSEn (Rahman et al., 2019b). Out of all the considered methods, four methods such as 70Propred, iPro70-PseZnc, Z-curve, and IPMD were not available or working. Therefore, for such methods we have considered the performance reported by the authors in their respective articles for comparison. For rest of the methods, we have predicted the class by providing the benchmark dataset as input and calculated the performance measures based on the predictions made by the respective methods. We have compared the performance of Sigma70Pred with sigma70 promoter prediction methods and found out that Sigma70Pred has outperformed all the considered methods, as shown in Table 2. In terms of AUROC, out of the all the methods developed on the same benchmark dataset, 70Properd attained the second highest performance with AUROC of 0.990, followed by iPro70-FMWin with AUROC of 0.960.

Performance comparison on independent dataset

In order to evaluate the proposed method's robustness and performance, we have also investigate the performance of our proposed model on the independent dataset of DNA sequences extracted from RegulonDB 10.8. We have also considered the existing methods for performance comparison on the independent dataset, which were trained and evaluated on different datasets such as MULTiPly (Zhang et al., 2019), iPromoter-2L (Liu et al., 2018), and, iPromoter-2L2.0 (Liu and Li, 2019). Moreover, to compare the efficiency of our generated model with deep-learning based classifiers, we have compared the performance with methods like iPromoter-BnCNN (Amin et al., 2020), pcPromoter-CNN (Shujaat et al., 2020), and PromoterLCNN (Hernandez et al., 2022). We have calculated the different performance measures for all the working sigma promoter predictors. The results on the independent dataset showed that our proposed model is quite robust towards the unseen data and performs well on it (Table 3). It also implies that our SVM model is significantly free from bias and overfitting on training dataset. As shown in Table 3, method named "MULTiPly" considered for the comparison which is not able to produce the results, therefore we have reported the performance achieved by the authors in this method. For comparison, we have considered the methods developed using machine-learning as well as deep-learning based classifiers. As exhibited in Table 3, SVM-based model developed on top-200 features in Sigma70Pred outperformed all the existing approaches in terms of each performance measure. Two-layer predictor method iPromoter2L-2.0 achieved the second highest accuracy of 83.36% on the independent dataset, followed by light-CNN based method PromoterLCNN with 79.56% accuracy.

Implementation of model in web server

In order to serve the scientific community, we have also developed the webserver Sigma70Pred by implementing our best

TABLE 1 Performance of various machine learning classifiers on benchmark dataset.

Classifier	Sensitivity	Specificity	Accuracy	AUROC	MCC
DT	74.49	87.14	82.77	0.808	0.62
RF	92.04	91.57	91.73	0.977	0.82
XGB	91.90	92.14	92.06	0.980	0.83
KNN	90.15	91.79	91.22	0.958	0.81
GNB	88.66	88.71	88.70	0.955	0.76
SVM	**97.44**	**97.36**	**97.38**	**0.996**	**0.94**

The values in the tables are in bold to represent the best performing classifier or method.

TABLE 2 Comparison of performances of our model with existing method on benchmark dataset evaluated using cross-validation technique.

Methods	Sensitivity	Specificity	Accuracy	AUROC	MCC
Sigma70Pred	**97.44**	**97.36**	**97.38**	**0.996**	**0.943**
iPro70-FMWin	83.81	95.07	91.17	0.960	0.803
70ProPred*	92.40	96.90	95.30	0.990	0.897
iPro70-PseZNC*	80.30	86.80	84.50	0.909	0.663
Z-Curve*	74.60	79.50	77.80	0.848	0.527
IPMD*	82.40	90.70	87.90	–	0.731
iProEP	89.52	64.03	76.88	0.654	0.554

*Reported by the authors in the manuscript. The values in the tables are in bold to represent the best performing classifier or method.

TABLE 3 The performance of existing methods on independent dataset.

Methods	Sensitivity	Specificity	Accuracy	AUROC	MCC
Sigma70Pred	**91.45**	**88.56**	**90.41**	**0.953**	**0.794**
iPro70-FMWin	84.12	86.67	85.04	0.921	0.693
iProEP	84.50	53.83	69.30	0.541	0.404
MULTiPly*	90.43	76.93	84.91	–	0.685
iPromoter-2L	86.21	72.81	79.56	–	0.601
iPromoter-2L2.0	88.72	77.91	83.36	–	0.674
iPromoter-FSEn	68.76	68.16	68.46	0.751	0.369
iPromoter-BnCNN	80.64	72.70	76.71	–	0.543
pcPromoter-CNN	81.44	61.07	71.35	–	0.445
Promoter-LCNN	88.77	70.15	79.54	–	0.604

*Reported by the authors in the manuscript. The values in the tables are in bold to represent the best performing classifier or method.

model to predict the sigma70 promoters. The web server consists of three modules namely "Predict," "Scan," and "Design." Our final model is based on SVC, it calculates SVC score for a sequence. SVC score is proportional to probability of correct prediction to promoter. SVC score varies from 0 to 1, higher the SVC score chances are higher that sequence is a sigma70 promoter. To provide balance between sensitivity and specificity, we provide default threshold. User may select desire threshold depending on their need. The detailed description of each module is as follows:

Predict

This module allows users to classify the submitted sequence as sigma70 promoter or non-promoter. There is a restriction of length in this module, as the model is trained on sequences with length 81 bp, hence if the submitted sequence is having a length less than 81, "A" will be added as the dummy variable and then, the sequence will be classified into one of the class, and if the length is greater than 81, only first 81 nucleotides will be considered for prediction. The user can submit sequences in either FASTA or single line format, and can select the desired threshold as SVC score above which the sequence will be classified as sigma70 promoter, otherwise non-promoter. The user can either provide single or multiple sequences, and can also upload the text file containing sequences. The output page displays the results in the tabular form, which is downloadable in the csv format.

Scan

Scan module allow users to scan or identify the sigma70 promoter region in given genome. This module does not have any length restriction as in the "predict" module. In this module, overlapping patterns of length 81 will be generated from submitted sequences and then used for prediction. The user can provide single or multiple sequences either in FASTA or in single line format. The user is also allowed to upload the sequence file. The output result will exhibit the overlapping patterns of length 81 with the prediction as promoter or non-promoter. The result is downloadable in the csv format.

Design

Design module allow users to identify the minimum mutations that can convert the sigma70 promoter into non-promoter or *vice-versa*. This module also has the restriction of sequence length 81, as it generates all the possible mutants by changing nucleotides at each position and then make the predictions based on the selected threshold. Since, generating all possible mutants is a time and computational expensive process, hence only one sequence is allowed at a time. The output page displays all the possible mutants with its prediction as promoter or non-promoter in tabular form which is downloadable in csv format.

Standalone

We have also developed Python and Perl-based standalone package, which is downloadable from URL: https://webs.iiitd.edu.in/raghava/sigma70pred/stand.html. The advantage of this module is that, it is not dependent on the availability of the internet, the user can download these standalone on their local machines and can use all the aforementioned modules. This module also take the input as single or multiple sequences in a file, in either FASTA or single line format. The output will be stored in the user-defined file in the comma separated value format.

Discussion

The expression of genes decides the cell's fate, which is regulated by the promoter regions present upstream of the transcription start site (Atkinson and Halfon, 2014). The interaction between the promoter region and the holoenzyme, switch on or off the expression of the respective genes. Various sigma factors are associated with the holoenzyme responsible for different functions, such as regulating nitrogen levels, controlling stationary phase genes, etc. (Paget, 2015). One of the essential sigma factors is sigma70, as it regulates the expression of most of the housekeeping genes required for the cell's survival (Paget and Helmann, 2003). The accurate identification of the promoter regions associated with the respective sigma factors may help in

the understanding of the regulatory mechanism, which can further be exploited to treat diseases caused by the disease-causing variants. The recognition of the promoter regions has been an important aspect of gene structure recognition and it is also the fundamental problem in building a network of gene transcriptional regulation. However, the experimental methods to identify the promoters are laborious, expensive, and time-consuming. On the other hand, computational approaches are reliable and fast with equivalent accuracy. Although, several methods have been developed in the past for the prediction of sigma promoters in the DNA sequences based on machine-learning (Lin and Li, 2011; Song, 2012; He et al., 2018; Liu et al., 2018; Lai et al., 2019; Liu and Li, 2019; Zhang et al., 2019) and deep-learning approaches (Amin et al., 2020; Shujaat et al., 2020; Hernandez et al., 2022), but the accurate identification of the sigma promoters remained a strenuous task due to the inter-and intra-class similarities and variations in the different sigma-specific promoter sequences (Zhang et al., 2019). It has been seen in the past that promoter sequences often differ at one or more locations from the consensus sequences (Mrozek et al., 2014, 2016), which makes the task of prediction of sigma70 promoters more difficult as sigma70 factor specific promoters are responsible for the transcription of most of the genes in prokaryotic genome. Moreover, the exponential increase in the data of promoter sequences due to the advancement in the high-throughput sequencing technology, also increased the level of difficulty in the identification of sigma70 promoter regions in the DNA sequences. Therefore, an accurate and robust method is required that can distinguish the sigma70 promoter sequences from the non-promoter sequences.

To understand the preference of nucleotides in the sigma70 promoter sequences, we have conducted the compositional and positional preference analysis for the sigma70 promoter sequences (Figures 2, 3). The compositional analysis showed that nucleotides "A" and "T" are in higher abundance in sigma70 promoter sequences in comparison with non-promoter sequences. For positional preference analysis, we have generated one-sample and two-sample logo using WebLogo and TSL logo tool. In one-sample logo, the preference of nucleotide at each position is shown in Figure 3A, however, the consensus sequences at position-10 and-35 is not clear. As shown by Shultzaberger et al. (2007) the gap between the regions-10 and-35 is not fixed, it varies from promoter to promoter. Therefore, they have shown the consensus sequences in their Figure 2 of the article at-10 and-35 regions in the form of sequence logos by vary the spacing between 21 and 26. On the other hand, we have generated the sequence logo by taking all the sigma70 promoter sequences without considering the variability in the spacing between the-10 and-35 regions. Whereas, in Figure 3B, we have represented the two-sample logo, by considering the sigma70 promoter and non-promoter sequences. It corresponds with the compositional analysis that most of the positions in the sigma70 promoter sequences are abundant in nucleotides "A" and "T" in comparison to the non-promoter sequences.

There are different methods which are specific to the classification of sigma70 promoters (Lin and Li, 2011; Song, 2012; He et al., 2018; Lai et al., 2019; Rahman et al., 2019a,b) whereas others are developed for the identification and classification of different sigma promoters such as sigma24, sigma28, sigma32, sigma38, sigma54, and sigma70 (Liu et al., 2018; Liu and Li, 2019; Zhang et al., 2019; Amin et al., 2020; Shujaat et al., 2020; Hernandez et al., 2022). In this study, we have also developed a bioinformatic-ware to classify the sigma70 promoters using only sequence information. The models were trained and evaluated using the nucleotide sequences of length 81 bp in the benchmark dataset retrieved from RegulonDB9.0 (Gama-Castro et al., 2016), which consists of 741 sigma70 promoters and 1,400 non-promoters. Initially, we calculated more than 8,000 features for each sequence, which were further processed using min-max scaling and top-200 most relevant features were selected using RFE feature selection technique. Further investigation was performed on these selected features. Then, we have implemented six different machine-learning classifiers to develop the prediction models on the selected features. The SVM-based model outperformed all the other classifiers with AUROC of 0.996 on the benchmark dataset (See Table 1). To understand the advantages and disadvantages of a new method, it is important to compare the proposed method with the already existing methods. We have considered already existing methods, some of them were non-functional, hence we have considered the performance reported in their respective articles for those methods. For rest of the methods, we have used the benchmark dataset to evaluate and compare the performance. Our proposed method has outperformed the methods developed on the same benchmark dataset, as shown in Table 2. Further, in order to check the efficiency of the proposed method, the generated model was evaluated and compared with existing methods using the unseen independent dataset, where sigma70pred outperformed the existing working method with AUROC of 0.953 (see Table 3). This comparison signified that our feature-set of 200 features is more effective to identify the sigma70 promoter sequences.

To understand the reason behind the wrong predictions made by our proposed model, we have selected all the sigma70 promoter sequences which were predicted as non-promoter, and provided them to the other existing sigma promoters predicting approaches. We found that most of the selected sequences were also wrongly predicted by other methods. Further, we checked the similarities of these sequences with the benchmark dataset using the "blastn" approach. For that, we have created a customized database using the sequences in the benchmark dataset by implementing the "makeblastdb" module of the BLAST program version 2.1.2. Then, we hit the wrongly predicted sequences to the customized dataset and considered the top-hit for further analysis. We have observed that most of the top-hit were non-promoter sequences, i.e., sigma70 promoter sequences in the independent dataset share similarity with the non-promoter sequences in the benchmark dataset. The negative data in the benchmark dataset used by several studies, was generated randomly from the coding and non-coding regions of *E. coli*. K-12 genome. Therefore, there is a need to develop the experimentally verified non-promoter

sequence dataset to improve the overall performance and efficiency of the prediction methods.

Moreover, Shimada et al. (2014) introduced the whole set of constitute promoters which was defined as the promoters recognized *in vitro* by the RNA polymerase RpoD holoenzyme without needing the additional supporting proteins. They have provided the list of the promoter sequences along with the genes which is controlled by the respective promoters. In order to investigate the efficiency of the our proposed method to classify the constitutive promoters, we have extracted the sequences from RegulonDB (Tierrafria et al., 2022) and colibir (Medigue et al., 1993) and used them for the prediction. We were able to extract the 329 promoter sequences, which were then submitted to the "predict" module Sigma70Pred web server with default parameters. 268 (81.46%) out of 329 were predicted as sigma70 promoters at the default threshold, which was increase to 276 (83.89%) on dropping the threshold to 0.2. The result on each promoter sequence is reported in Supplementary Table S2 along with the SVC score. These results signify that our proposed model is able to classify the constitutive promoters with reliable accuracy.

Sigma70Pred offers a web server and standalone packages to predict the sigma70 promoters using sequence information. This method uses 200 different optimal features, and we assume that our features have more capability to classify sigma70 promoters. Sigma70Pred provides three major modules: predict, scan, and design. As the application of our method, the user can scan the entire prokaryote genome to identify the sigma70 promoter using the scan module. By using the design module, the user can also determine the minimum number of mutations required to exploit the sigma70 promoter regions, i.e., either induce or deteriorate the capability of the sigma70 promoter. As compared to the existing methods of predicting sigma70 promoters, Sigma70Pred produced commending outcomes. We believe that Sigma70Pred will play an essential role in the area of genomic analysis.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://webs.iiitd.edu.in/raghava/sigma70pred/data.html.

Author contributions

GR conceived the idea and supervised the entire project. NS, MA, and DP collected and curated the datasets. SP, NS, MA, and DP wrote all the in-house scripts, performed the formal analysis, and developed the prediction models. SP developed the web interface and standalone. SP and GR prepared all the drafts of manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

We are thankful to funding agencies Department of Biotechnology (DBT), Govt. of India for financial support and fellowships. We are also thankful to Megha Mathur and Anjali Dhall for python scripts to generate features and help in the figure's preparation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.1042127/full#supplementary-material

References

Agrawal, P., Kumar, R., Usmani, S. S., Dhall, A., Patiyal, S., Sharma, N., et al. (2019). GPSRdocker: a Docker-based resource for genomics, proteomics and systems biology. *BioRxiv*, 827766. doi: 10.1101/827766

Amin, R., Rahman, C. R., Ahmed, S., Sifat, M. H. R., Liton, M. N. K., Rahman, M. M., et al. (2020). iPromoter-BnCNN: a novel branched CNN-based predictor for identifying and classifying sigma promoters. *Bioinformatics* 36, 4869–4875. doi: 10.1093/bioinformatics/btaa609

Atkinson, T. J., and Halfon, M. S. (2014). Regulation of gene expression in the genomic context. *Comput. Struct. Biotechnol. J.* 9:e201401001. doi: 10.5936/csbj.201401001

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335

Bernardo, L. M. D., Johansson, L. U. M., Skarfstad, E., and Shingler, V. (2009). Sigma54-promoter discrimination and regulation by ppGpp and DksA. *J. Biol. Chem.* 284, 828–838. doi: 10.1074/jbc.M807707200

Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., and Chou, K.-C. (2015). PseKNC-general: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics* 31, 119–120. doi: 10.1093/bioinformatics/btu602

Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–1190. doi: 10.1101/gr.849004

Friedel, M., Nikolajewa, S., Suhnel, J., and Wilhelm, T. (2009). DiProDB: a database for dinucleotide properties. *Nucleic Acids Res.* 37, D37–D40. doi: 10.1093/nar/gkn597

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muniz-Rascado, L., Garcia-Sotelo, J. S., et al. (2016). RegulonDB version 9.0: high-

level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.* 44, D133–D143. doi: 10.1093/nar/gkv1156

Gordon, L., Chervonenkis, A. Y., Gammerman, A. J., Shahmuradov, I. A., and Solovyev, V. V. (2003). Sequence alignment kernel for recognition of promoter regions. *Bioinformatics* 19, 1964–1971. doi: 10.1093/bioinformatics/btg265

He, W., Jia, C., Duan, Y., and Zou, Q. (2018). 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. *BMC Syst. Biol.* 12:44. doi: 10.1186/s12918-018-0570-1

Hernandez, D., Jara, N., Araya, M., Duran, R. E., and Buil-Aranda, C. (2022). PromoterLCNN: a light CNN-based promoter prediction and classification model. *Genes* 13:1126. doi: 10.3390/genes13071126

Lai, H.-Y., Zhang, Z.-Y., Su, Z.-D., Su, W., Ding, H., Chen, W., et al. (2019). iProEP: a computational predictor for predicting promoter. *Mol. Ther. Nucleic Acids* 17, 337–346. doi: 10.1016/j.omtn.2019.05.028

Lin, H., and Li, Q.-Z. (2011). Eukaryotic and prokaryotic promoter prediction using hybrid approach. *Theory Biosci.* 130, 91–100. doi: 10.1007/s12064-010-0114-8

Lin, H., Liang, Z.-Y., Tang, H., and Chen, W. (2019). Identifying Sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 1316–1321. doi: 10.1109/TCBB.2017.2666141

Liu, B., and Li, K. (2019). iPromoter-2L2.0: identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features. *Mol. Ther. Nucleic Acids* 18, 80–87. doi: 10.1016/j.omtn.2019.08.008

Liu, B., Yang, F., Huang, D.-S., and Chou, K.-C. (2018). iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics* 34, 33–40. doi: 10.1093/bioinformatics/btx579

Liu, B., Zhang, D., Xu, R., Xu, J., Wang, X., Chen, Q., et al. (2014). Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics* 30, 472–479. doi: 10.1093/bioinformatics/btt709

Lu, C., Xie, M., Wendl, M. C., Wang, J., McLellan, M. D., Leiserson, M. D. M., et al. (2015). Patterns and functional implications of rare germline variants across 12 cancer types. *Nat. Commun.* 6:10086. doi: 10.1038/ncomms10086

Mathur, M., Patiyal, S., Dhall, A., Jain, S., Tomer, R., Arora, A., et al. (2021). Nfeature: a platform for computing features of nucleotide sequences. *BioRxiv*, 10.1101/2021.12.14.472723

Medigue, C., Viari, A., Henaut, A., and Danchin, A. (1993). Colibri: a functional data base for the *Escherichia coli* genome. *Microbiol. Rev.* 57, 623–654. doi: 10.1128/mr.57.3.623-654.1993

Mrozek, D., Daniłowicz, P., and Małysiak-Mrozek, B. (2016). HDInsight4PSi: boosting performance of 3D protein structure similarity searching with HDInsight clusters in Microsoft Azure cloud. *Informat. Sci.* 349-350, 77–101. doi: 10.1016/j.ins.2016.02.029

Mrozek, D., Malysiak-Mrozek, B., and Klapcinski, A. (2014). Cloud4Psi: cloud computing for 3D protein structure similarity searching. *Bioinformatics* 30, 2822–2825. doi: 10.1093/bioinformatics/btu389

Paget, M. S. (2015). Bacterial sigma factors and anti-sigma factors: structure, function and distribution. *Biomol. Ther.* 5, 1245–1265. doi: 10.3390/biom5031245

Paget, M. S. B., and Helmann, J. D. (2003). The sigma70 family of sigma factors. *Genome Biol.* 4:203. doi: 10.1186/gb-2003-4-1-203

Patiyal, S., Agrawal, P., Kumar, V., Dhall, A., Kumar, R., Mishra, G., et al. (2020). NAGbinder: an approach for identifying N-acetylglucosamine interacting residues of a protein from its primary sequence. *Protein Sci.* 29, 201–210. doi: 10.1002/pro.3761

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Rahman, M. S., Aktar, U., Jani, M. R., and Shatabda, S. (2019a). iPro70-FMWin: identifying Sigma70 promoters using multiple windowing and minimal features. *Mol. Gen. Genomics* 294, 69–84. doi: 10.1007/s00438-018-1487-5

Rahman, M. S., Aktar, U., Jani, M. R., and Shatabda, S. (2019b). iPromoter-FSEn: identification of bacterial sigma(70) promoter sequences using feature subspace based ensemble classifier. *Genomics* 111, 1160–1166. doi: 10.1016/j.ygeno.2018.07.011

Sachs, M. C. (2017). plotROC: a tool for plotting ROC curves. *J. Stat. Softw.* 79:2. doi: 10.18637/jss.v079.c02

Shimada, T., Yamazaki, Y., Tanaka, K., and Ishihama, A. (2014). The whole set of constitutive promoters recognized by RNA polymerase RpoD holoenzyme of *Escherichia coli*. *PLoS One* 9:e90447. doi: 10.1371/journal.pone.0090447

Shujaat, M., Wahab, A., Tayara, H., and Chong, K. T. (2020). pcPromoter-CNN: a CNN-based prediction and classification of promoters. *Genes* 11:1529. doi: 10.3390/genes11121529

Shultzaberger, R. K., Chen, Z., Lewis, K. A., and Schneider, T. D. (2007). Anatomy of *Escherichia coli* sigma70 promoters. *Nucleic Acids Res.* 35, 771–788. doi: 10.1093/nar/gkl956

Song, K. (2012). Recognition of prokaryotic promoters based on a novel variable-window Z-curve method. *Nucleic Acids Res.* 40, 963–971. doi: 10.1093/nar/gkr795

Tierrafria, V. H., Rioualen, C., Salgado, H., Lara, P., Gama-Castro, S., Lally, P., et al. (2022). RegulonDB 11.0: Comprehensive high-throughput datasets on transcriptional regulation in *Escherichia coli* K-12. *Microb. Genomics* 8, 8:mgen000833. doi: 10.1099/mgen.0.000833

Vacic, V., Iakoucheva, L. M., and Radivojac, P. (2006). Two sample logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* 22, 1536–1537. doi: 10.1093/bioinformatics/btl151

Zhang, M., Li, F., Marquez-Lago, T. T., Leier, A., Fan, C., Kwoh, C. K., et al. (2019). MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters. *Bioinformatics* 35, 2957–2965. doi: 10.1093/bioinformatics/btz016

frontiers | Frontiers in Microbiology

TYPE Methods
PUBLISHED 06 December 2022
DOI 10.3389/fmicb.2022.1048478

OPEN ACCESS

EDITED BY
Hao Lin,
University of Electronic Science and Technology of China, China

REVIEWED BY
Wen Zhang,
Huazhong Agricultural University, China
Jiangning Song,
Monash University, Australia

*CORRESPONDENCE
Guohua Huang
guohuahhn@163.com

SPECIALTY SECTION
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 19 September 2022
ACCEPTED 26 October 2022
PUBLISHED 06 December 2022

CITATION
Zheng P, Qi Y, Li X, Liu Y, Yao Y and Huang G (2022) A capsule network-based method for identifying transcription factors. *Front. Microbiol.* 13:1048478. doi: 10.3389/fmicb.2022.1048478

A capsule network-based method for identifying transcription factors

Peijie Zheng[1], Yue Qi[1], Xueyong Li[1], Yuewu Liu[2], Yuhua Yao[3] and Guohua Huang[1]*

[1]School of Electrical Engineering, Shaoyang University, Shaoyang, China, [2]College of Information and Intelligence, Hunan Agricultural University, Changsha, China, [3]School of Mathematics and Statistics, Hainan Normal University, Haikou, China

Transcription factors (TFs) are typical regulators for gene expression and play versatile roles in cellular processes. Since it is time-consuming, costly, and labor-intensive to detect it by using physical methods, it is desired to develop a computational method to detect TFs. Here, we presented a capsule network-based method for identifying TFs. This method is an end-to-end deep learning method, consisting mainly of an embedding layer, bidirectional long short-term memory (LSTM) layer, capsule network layer, and three fully connected layers. The presented method obtained an accuracy of 0.8820, being superior to the state-of-the-art methods. These empirical experiments showed that the inclusion of the capsule network promoted great performances and that the capsule network-based representation was superior to the property-based representation for distinguishing between TFs and non-TFs. We also implemented the presented method into a user-friendly web server, which is freely available at http://www.biolscience.cn/Capsule_TF/ for all scientific researchers.

KEYWORDS

transcription factors, capsule network, deep learning, LSTM, semantics

Introduction

Transcription factors (TFs) are also sequence-specific DNA-binding factors, a family of proteins that control the expression of target genes (Karin, 1990; Latchman, 1997). The TFs are widely distributed, and their numbers vary with the size of the genome (Nimwegen, 2006). The larger genomes are likely to have a larger number of TFs on average. Approximately 10% of genes in the human genome are conservatively estimated to code for TFs. Consequently, the TFs are the potentially largest family of proteins in humans. The TFs exert regulating roles alone or together with other proteins in a complex by hindering or facilitating the recruitment of RNA polymerase (a type of enzyme) to specific DNA regions (Roeder, 1996; Nikolov and Burley, 1997). The regulation roles of the TFs are either positive or negative. The TFs promote the

recruitment of RNA polymerase function as activators and contrarily ones to hold back recruitment as repressors. The TFs are involved in many important cellular processes including transcription regulation. Some TFs are responsible for cell differentiation (Wheaton et al., 1996), some respond to intercellular signals (Pawson, 1993), and some reply to environmental changes (Shamovsky and Nudler, 2008). Mutations in the TFs are discovered to be implied in many diseases (Bushweller, 2019). The TFs are a control switch to turn on or off to ensure when, where, and how many genes are accurately expressed. Thus, it is a fundamental problem but a therapeutic opportunity for drug discovery and development to accurately identify TFs. Physical or chemical methods (called wet experiments) are a prime alternative to identify TFs. The wet experiments include SELEX-based methods (Roulet et al., 2002), MITOMI (Rockel et al., 2012), and ChIP-based assays (Yashiro et al., 2016). Most known TFs were discovered by wet experiments and deposited in public databases (Wingender et al., 1996; Riaño-Pachón et al., 2007; Zhu et al., 2007; Zhang et al., 2020). The wet experiments accumulated a limited number of TFs at the expense of an enormous amount of time and money. It is only by the wet experiments that it is impossible and insufficient to discover all TFs in all the tissues or species all over the world. With advances in artificial intelligence, it is becoming possible to learn a computational model from these known TFs to recognize new unknown TFs which will be subsequently examined by the wet experiments. The computational methods shrank greatly the numbers of potential TFs that the wet experiments scanned, and thus, save a vast volume of time and money. The computational methods are becoming essentially complementary to the wet experiments, and both are jointly accelerating the exploration of the TFs.

To the best of our knowledge, Liu et al. (2020) pioneered the first computational method for discriminating TFs from non-TFs. Liu et al. extracted three types of sequence features: composition/transition/distribution (CTD) (Tan et al., 2019), split amino acid composition (SAAC), and dipeptide composition (DC) (Ding and Li, 2015). Comprehensively comparing the contribution of features and performances of five frequently used machine learning algorithms: logistic regression, random forest, k-nearest neighbor, XGBoost, and support vector machine (SVM). Liu et al. finally chose 201 optimal features and SVM for building the classifier. Liu et al. opened an avenue to identify TFs. Lately, Li et al. (2022) created a different idea from Liu et al. to distinguish TFs and non-TFs. Instead of designing sophisticated features. Li et al. directly took the sequence as input, split three amino acid residues as a basic unit, and employed long short-term memory (LSTM) for capturing semantic differences between TFs and non-TFs. Li et al. promoted the predictive accuracy to 86.83%. The LSTM is a special recurrent neural network (RNN) which suffered from the long-distance dependency. The capsule network proposed is a novel neural work architecture (Sabour et al., 2017), whose remarkable advantage is to capture relationship between local parts. This just made up for the deficiency of LSTM. Inspired by this, we proposed a capsule network-based method for TFs prediction.

Materials and methods

Data

The training and the testing data were downloaded from the website[1] (Li et al., 2022), which was manually collected by Liu et al. (2020). The original dataset contained 601 human and 129 mouse TFs which preferred methylated DNA (Graves and Schmidhuber, 2005; Wang et al., 2018) and 286 TFs which preferred non-methylated DNA (Yin et al., 2017). Liu et al. (2020) conducted the following steps for improving the quality of the dataset. The sequences containing illegal characters such as "X", "B", and "Z" were first removed. Then, the CD-HIT, which is a clustering tool (Huang et al., 2010; Zou et al., 2020), was used to decrease redundancy between sequences. The cutoff threshold was set to 0.25, meaning that the sequence identity between any two sequences was no more than 0.25. Third, less than 50 amino acid sequences were excluded. A total of 522 TFs were finally preserved as positive samples after the above three processes. Liu et al. sampled the same number of non-TFs from the UniProt database (release 2019_11) which meets the following five requirements: (1) reviewed proteins, (2) proteins with evidence at protein level, (3) proteins in full length and of more than 50 amino acid residues, (4) proteins without DNA-binding TF activities, and (5) Homo sapiens proteins with less than 25% sequence identity in the CD-HIT. Liu et al. divided the data further into the training and the independent test dataset at the ratio of 8:2, with the former containing 406 positive and 406 negative samples, and the latter containing 106 positive and 106 negative samples.

Methods

As shown in **Figure 1**, the proposed method called Capsule_TF is a deep learning-based method. It mainly contains five layers, namely, embedding layer, bidirectional LSTM layer, capsule network layer, and three fully connected layers. The protein sequence as input goes through the embedding layer and is then embedded into low-dimensional vectors. The bidirectional LSTM layer and the capsule network layer are used to extract high-level representations of protein sequences. Three fully connected layers are finally used to discriminate TFs from non-TFs. The Capsule_TF is an end-to-end deep learning model without designing any features.

1 https://bioinfor.nefu.edu.cn/TFPM/

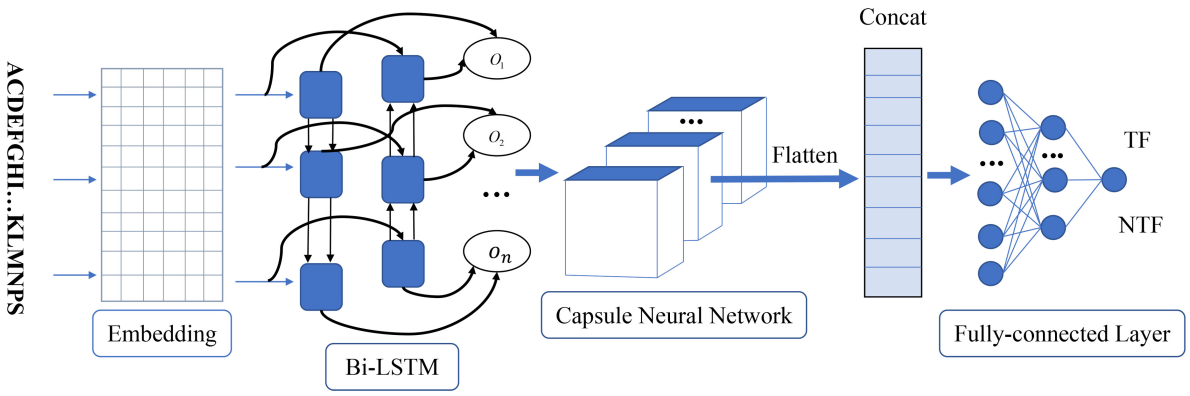

FIGURE 1
The architecture of the proposed capsule network-based method. O_n represents the word vector after the bidirectional LSTM layer. TF and NTF represent the final predicted outcome as a transcription factor or not.

Embedding

It is mandatory for text sequence input to be converted into digital sequences which are suitable to be processed by the subsequent machine learning algorithms. There are many ways of converting text sequences into digital sequences, such as a one-hot encoding scheme (Buckman et al., 2018) and Word2vec (Rong, 2014). The one-hot encoding scheme fails to capture relationships between words and is opt to yield sparse representation when the vocabulary is large. It is a common practice to use embedding to translate text sequences into dense digital vectors. In the field of text analysis by the deep neural network, the embedding is generally the first layer generally defined by

$$\hat{x}_i = W_e x_i \tag{1}$$

where $\hat{x}_i$ denotes the embedding of the word, x_i represents input, and $W_e \in R^{n \times k}$ denotes a lookup table that stores the embedding of words. W_e is the learnable parameter.

Long short-term memory

The LSTM (Hochreiter and Schmidhuber, 1997) belongs to the family of recurrent neural networks (RNNs) (Sherstinsky, 2020), which is typically a neural network sharing parameters at all time steps. The LSTM was pioneered by Hochreiter and Schmidhuber (Hochreiter and Schmidhuber, 1997) and later was continuously improved. The structure of the current LSTM was mainly made up of the cell state, the hidden state, the input, and the output. **Figure 2** demonstrates the structure of the LSTM at the time step t which is identical at all the time steps. The cell state preserved memories for preceding words but was regulated by the gates to determine how much information was conveyed to the next time step. There are three gates in the LSTM: forget gate, input gate, and output gate. The forget gate is defined as

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{2}$$

where h_{t-1} denotes the hidden state at time step $t-1$, x_t is the input at time step t, W_f and b_f are learnable parameters, and σ is the sigmoid function. Obviously, the output of the forget gate falls between 0 and 1. The input gate and the candidate cell are defined, respectively, as

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{3}$$

and

$$\tilde{C}_t = tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \tag{4}$$

where W_i, W_c, b_i, and b_c are learnable parameters. The cell state is updated by

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{5}$$

The preceding information is all forgotten if the forget gate is 0, namely, $f_t = 0$, all the information is born in mind if $f_t = 1$, and part are born if f_t is more than 0 but less than 1. Obviously, the forget gate determines how much memories for preceding words are preserved. The input gate and the candidate cell determine how much new information about the time step is added to the cell state. The contribution of the time step t to the cell state is nearly nothing if the second item in Equation (5) is equal to 0. The hidden states are updated jointly by the cell state and the output gate

$$h_t = O_t * \tanh\left(C_t\right) \tag{6}$$

where O_t denotes the output gate which is computed by

$$O_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{7}$$

Compared with the traditional RNN, the LSTM solved well long-term dependency issues by the cell state conveying memory. To capture both directional dependencies between words, the bidirectional LSTM was used here. Due to its efficiency and effectiveness in sequence analysis, the

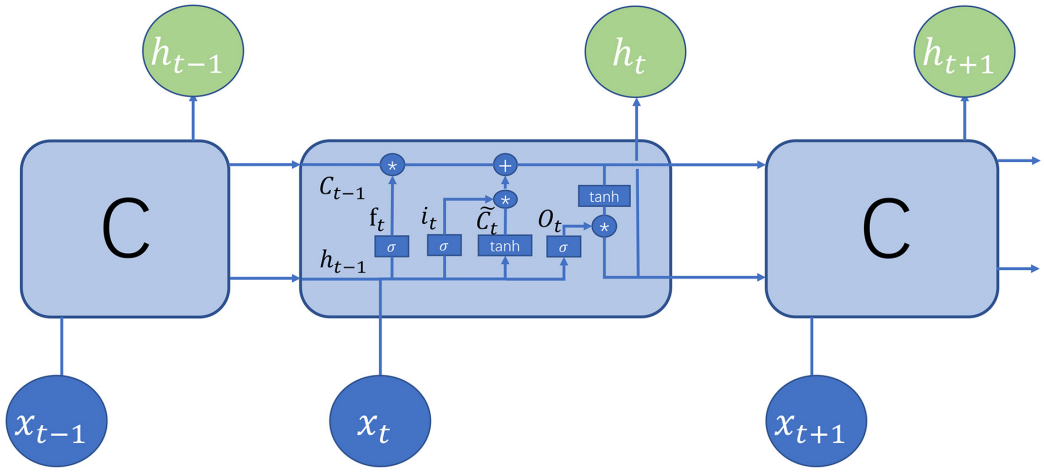

FIGURE 2
Illustration of long short-term memory (LSTM) structure (Olah, 2015).

LSTM has been widely applied to the N6-methyladenosine prediction (Chen et al., 2022), speech recognition (Sak et al., 2015), continuous B-cell epitope prediction (Saha and Raghava, 2006), N4-Acetylcytidine prediction (Zhang et al., 2022), lysine succinylation identification (Huang et al., 2021), sentiment analysis (Arras et al., 2017), and action recognition (Du et al., 2015).

Capsule network

The capsule network is a newly developed neural network in 2017 (Sabour et al., 2017). The capsule network is different from the conventional neural network. The basic unit of the capsule network is capsules which are defined as a set of neurons, while the latter consists of neurons. The neuron is generally a scalar value that represents a single pattern, while the capsules are a multi-dimensional vector, being able to represent multi-patterns. In addition, the capsule network is capable of capturing links between different local properties (Jia and Meng, 2016; Xi et al., 2017), which the convolution neural network (Shin et al., 2016) fail to discover. At the heart of the capsule network lies the dynamic routing as illustrated in **Figure 3**. v_i was assumed to be the capsules in the layer L, whose prediction vectors are defined by

$$u_{j|i} = W_{ij} v_i \tag{8}$$

where W_{ij} is a learnable matrix. The capsule s_j in the layer L+1 denotes a weighted sum over the prediction vectors, which is computed by

$$s_j = \sum_{i=1} c_{ij} u_{j|i} \tag{9}$$

where c_{ij} is the coupling coefficient. The output of the capsule s_j is further activated by a non-linear "squashing" function so that short vectors get shrunk to almost zero length and long vectors get shrunk to a length slightly below 1.

$$a_j = \frac{||s_i||}{1 + ||s_i||^2} \frac{s_i}{||s_i||} \tag{10}$$

The coupling coefficient represents the probability of two capsules to the couple. The more consistent the two capsules, the large the coupling coefficient. The coupling coefficient is initialized as the log prior probabilities that the capsule j was coupled to the capsule i.

$$c_{ij} = \frac{\exp\left(b_{ij}\right)}{\sum_k \exp\left(b_{kj}\right)} \tag{11}$$

The prior probabilities are updated by the dynamic routing algorithm

$$b_{ij} = b_{ij} + a_j u_{j|i} \tag{12}$$

The dynamic routing algorithm is to iterate the Equations (9) to (12).

$$u_i = W_i v_i \tag{13}$$

Metrics

For binary classification, there are four common metrics: sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthews correlation coefficient (MCC), which are defined by

$$Sensitivity = Sn = \frac{TP}{TP + FN} \tag{14}$$

$$Specificity = Sp = \frac{TN}{TN + FP} \tag{15}$$

$$Accuracy = Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP \times FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{17}$$

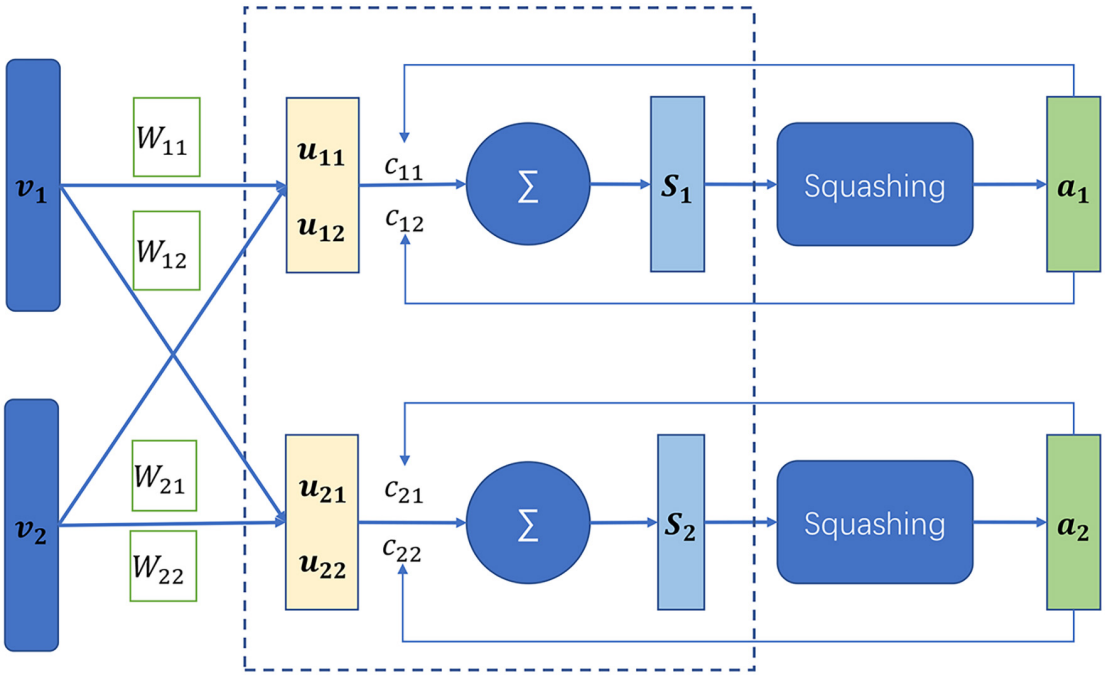

FIGURE 3
Illustration of dynamic routing in the capsule network (Lee, 2018).

where TP and TN are the numbers of correctly predicted positive and negative samples, respectively, as well as FP and FN are the numbers of wrongly predicted positive and negative samples, respectively. In addition, we also employed the receiver operating characteristic (ROC) to evaluate performances. The area under the ROC curve (AUC) lies between 0 and 1. The more the AUC, the better the performance.

Results

There are two state-of-the-art methods for predicting TFs. One is the deep learning-based method by Li et al. (2022), which is called Li's method, and another is the sequence feature-based method by Liu et al. (2020), which is called Liu's method. To examine the Capsule_TF for efficiency and effectiveness in identifying TFs, we compared it with these two methods by the independent test. As shown in **Table 1**, the Capsule_TF is completely superior to the two methods. The Capsule_TF increased the Sn by 0.0283 over Li's and even 0.1132 over Liu's. The Capsule_TF increased MCC by 0.0386 over Li's and even 0.1044 over Liu's.

TABLE 1 Comparison with two states of the art methods in the independent test.

Method	Sn	Sp	Acc	MCC	AUC
Capsule_TF	**0.9151**	**0.8490**	**0.8820**	**0.7658**	**0.9252**
Li et al. (2022)	0.8868	0.8396	0.8663	0.7272	0.9130
Liu et al. (2020)	0.8019	0.8585	0.8302	0.6614	0.9116

The bold highlighted the best values.

Discussion

Effect of position

The length of amino acid sequences varies with TFs. The longest reached 4,834 amino acid residues, the shortest is only 51 residues, and each TFs have an average of 536 residues. It is compulsory that the input is of the unified length in the machine

TABLE 2 Predictive performance of amino acid residues from different positions.

Data	Sn	Sp	Acc	MCC	AUC
Upstream_500	0.9151	0.8490	0.8820	0.7658	0.9252
Centre_500	0.8773	0.8679	0.8726	0.7453	0.9084
Downstream_500	0.9056	0.8396	0.8726	0.7469	0.9149

TABLE 3 Predictive performance of the method without capsule network.

Method	Sn	Sp	Acc	MCC	AUC
Non-Capsule	0.6320	0.8867	0.7594	0.5365	0.8120
With-Capsule	0.9151	0.8490	0.8820	0.7658	0.9252

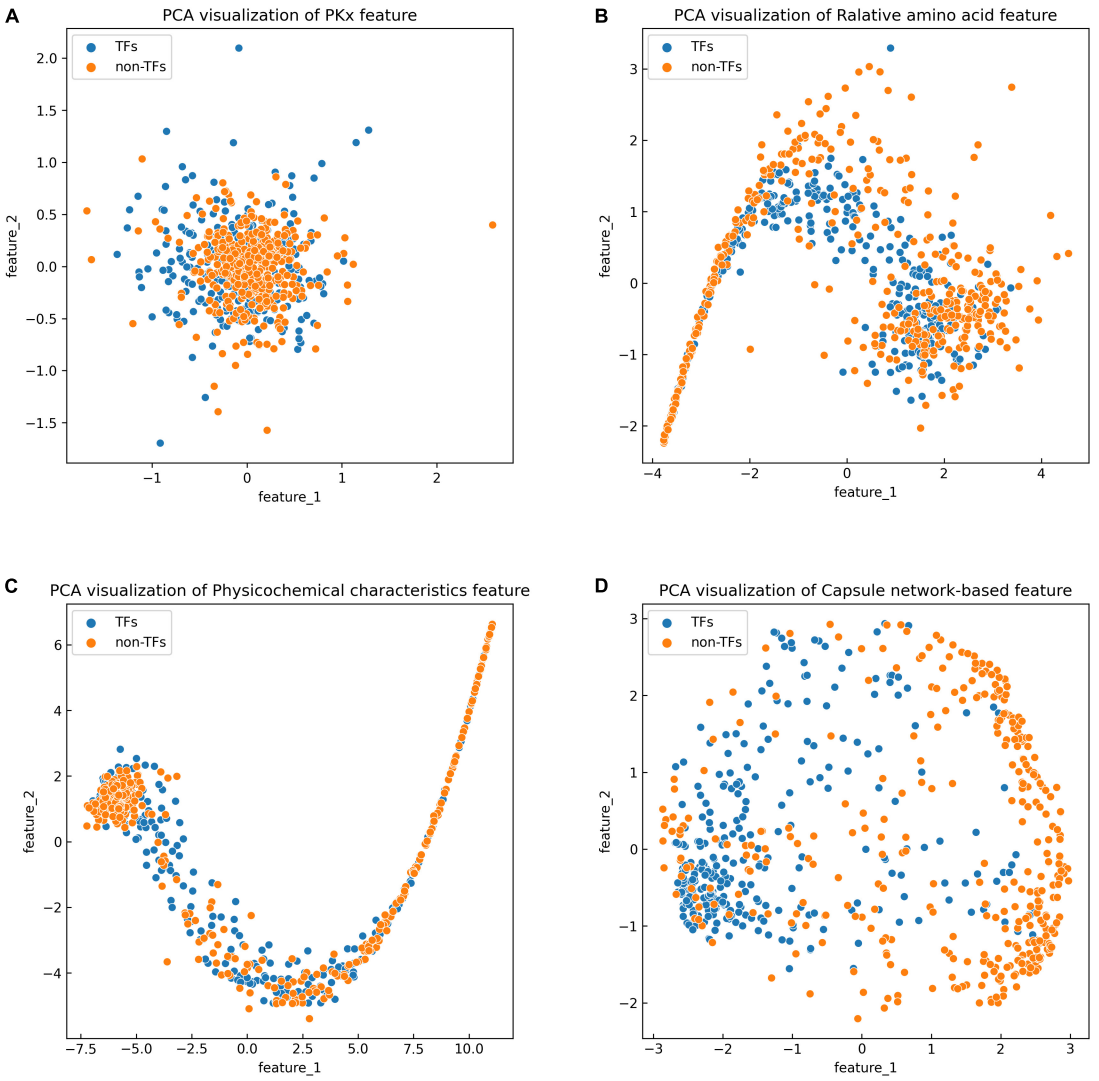

FIGURE 4
Principal component analysis (PCA) visualization about the different features: **(A)** PKx, **(B)** relative amino acid propensity, **(C)** physicochemical characteristics, and **(D)** capsule network-based features. As seen in this image, orange represents non-TFs and blue represents TFs.

learning algorithm. We investigated the effects of the number of amino acid residues at different positions on discriminating TFs from non-TFs. We chose 500 amino acid residues at the start, at the middle, and the end, respectively. As shown in Table 2, their predictive performances are approximately equivalent, meaning that positions have little effect. A potential reason is that 500 amino acid residues might contain sufficient information about TFs.

Contribution of capsule network

In comparison with Li's method, the remarkable characteristic of the Capsule_TF is to utilize the capsule network. In order to investigate the contribution of the capsule network to classifying TFs, we removed it. The predictive performance after excluding the capsule network is listed in Table 3. Obviously, all metrics except Sp. decreased precipitously. Sn decreased from 0.9151 to 0.6320, Acc from 0.8820 to 0.7594, MCC from 0.7658 to 0.5365, and AUC from

TABLE 4 Performance comparison across different features by SVM.

Feature	Sn	Sp	Acc	MCC
PKx	0.5660	0.7452	0.6556	0.3164
Relative amino acid propensity	0.6792	0.7075	0.6933	0.3869
Physicochemical characteristics	0.5283	0.6981	0.6132	0.2297
Capsule network-based feature	**0.9151**	**0.8396**	**0.8773**	**0.7568**

The bold highlighted the best values.

TABLE 5 Performance comparison across different features by logistic regression.

Feature	Sn	Sp	Acc	MCC
Pkx	0.7075	0.5660	0.6368	0.2764
Relative amino acid propensity	0.5943	0.6509	0.6226	0.2457
Physicochemical characteristics	0.6981	0.5849	0.6415	0.2849
Capsule network-based feature	0.9245	0.7924	0.8584	0.7233

TABLE 6 Performance comparison across different features by linear discriminant analysis (LDA).

	Sn	Sp	Acc	MCC
Pkx	0.6981	0.5094	0.6038	0.2113
Relative amino acid propensity	0.6321	0.5472	0.5896	0.1799
Physicochemical characteristics	0.7736	0.5189	0.6462	0.3024
Capsule network-based feature	0.8962	0.7830	0.8396	0.6836

0.9252 to 0.8120. The results indicated that the capsule network contributed much to identifying TFs.

Comparison with feature-based methods

The discriminative features provide a potential explanation to distinguish between both classes of samples. We compared three frequently used property-based features with the capsule network-based features. Three property-based features are PKx, relative amino acid propensity (RAA), and physicochemical characteristics (Li et al., 2008, 2021; Zhang et al., 2019). The output of the capsule layer was considered as the capsule network-based feature. **Figure 4** visualizes the first two components of four types of features. The first two components were computed by PCA (Yang et al., 2004). Obviously, the first two components of the capsule network-based features are more discriminative than those of the other three types of features. We used the SVM (Noble, 2006) to compare the discriminative abilities of these features. As shown in **Table 4**, the capsule network-based feature is superior to the three property-based features. We also compared the logistic regression and LDA with the Capsule_TF. As listed in **Tables 5**, **6**, the Capsule_TF is superior to the logistic regression and the LDA, and the capsule network-based features are superior to the conventional representations.

The previous results indicated that the Capsule_TF outperformed two state-of-the-art methods: Li's method (Li et al., 2022) and Liu's method (Liu et al., 2020). Li's method (Li et al., 2022) is a Bi-LSTM-based method, while Capsule_TF not only employed Bi-LSTM but also utilized a capsule network. The inclusion of a capsule network effectively promoted the representation of protein sequences of TFs. The ablation experiments validated the contribution of the capsule network to the identification of TFs (**Table 3**). Liu's method (Liu et al., 2020) is feature-based. We compared features extracted by Capsule_TF with traditional sequence property-based features. As shown in **Figure 4** and **Table 4**, the capsule network-based feature is more discriminative than the traditional sequence property-based feature. Despite the Capsule_TF obtaining superior performances over the state-of-the-art methods, there were some limitations that need to be improved in the feature. First, the consumption time in dynamic routing is very large. Therefore, Capsule_TF is not suitable to deal with large-scale datasets. Second, the interpretability of Capsule_TF needs to be improved.

Web application

We realized the presented method into a web application which is freely available.[2] The web application is based on the Django framework and utilized python and Tensorflow. The web application is very easy for users to use. The first thing is for the user to upload the predicted protein sequences in the FASTA format to the textbox or the file to the web. Clicking the "submit" button, users will obtain the results. The consuming time is directly proportional to the number of protein sequences. In addition, users could download the training and testing dataset in the experiments.

Conclusion

The TFs are very influential in transcription regulation. It is a challenging task to accurately recognize TFs at present. We presented a capsule network-based method for identifying TFs, which outperformed the state-of-the-art methods in the experiments. The presented method benefits from the inclusion of a capsule network, which captures a more informative representation than the property-based method. We also developed a web application that facilitated the detection of TFs. The method and the web application are helpful to identify TFs and to further explore their roles. The TFs play typically regulating roles in gene expression by binding to short DNA sequences. The roles of TFs depend on their binding to DNA sequences. In the future, we hope to create an effective and efficient method to recognize such binding and interpret its mechanism from the semantics of both protein and DNA sequences.

2 http://www.biolscience.cn/Capsule_TF/

Data availability statement

The datasets presented in this study can be found in online repositories at http://www.biolscience.cn/Capsule_TF/.

Author contributions

PZ: data curation, methodology, software, investigation, and writing. YQ: validation. XL, YL, and YY: conceptualization and writing. GH: conceptualization, funding acquisition, supervision, and writing – reviewing and editing. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the National Natural Science Foundation of China (62272310 and 62162025), the Hunan Province Natural Science Foundation of China (2022JJ50177 and 2020JJ4034), the Scientific Research Fund of Hunan Provincial Education Department (21A0466 and 19A215), and the Shaoyang University Innovation Foundation for Postgraduate (CX2021SY052).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Arras, L., Montavon, G., Müller, K.-R., and Samek, W. (2017). Explaining recurrent neural network predictions in sentiment analysis. *arXiv* [preprint]. doi: 10.48550/arXiv.1706.07206

Buckman, J., Roy, A., Raffel, C., and Goodfellow, I. (2018). "Thermometer encoding: one hot way to resist adversarial examples," in *Proceeding of the international conference on learning representations*.

Bushweller, J. H. (2019). Targeting transcription factors in cancer—from undruggable to reality. *Nat. Rev. Cancer* 19, 611–624. doi: 10.1038/s41568-019-0196-7

Chen, J., Zou, Q., and Li, J. (2022). DeepM6ASeq-EL: prediction of human N6-methyladenosine (m6A) sites with LSTM and ensemble learning. *Front. Comput. Sci.* 16:1–7. doi: 10.1007/s11704-020-0180-0

Ding, H., and Li, D. (2015). Identification of mitochondrial proteins of malaria parasite using analysis of variance. *Amino Acids* 47, 329–333. doi: 10.1007/s00726-014-1862-4

Du, Y., Wang, W., and Wang, L. (2015). "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1110–1118.

Graves, A., and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 602–610. doi: 10.1016/j.neunet.2005.06.042

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Huang, G., Shen, Q., Zhang, G., Wang, P., and Yu, Z. G. (2021). LSTMCNNsucc: a bidirectional LSTM and CNN-based deep learning method for predicting lysine succinylation sites. *BioMed Res. Int.* 2021, 112. doi: 10.1155/2021/9923112

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680–682. doi: 10.1093/bioinformatics/btq003

Jia, X., and Meng, M. Q.-H. (2016). "A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images," in *Proceeding of the 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, (IEEE), 639–642.

Karin, M. (1990). Too many transcription factors: positive and negative interactions. *New Biol.* 2, 126–131.

Latchman, D. S. (1997). Transcription factors: an overview. *Int. J. Biochem. Cell Biol.* 29, 1305–1312. doi: 10.1016/S1357-2725(97)00085-X

Lee, H. Y. (2018). *Capsule*. Available online at: https://www.bilibili.com/video/av16583439/?from=search&seid=4942786181857065642

Li, H., Gong, Y., Liu, Y., Lin, H., and Wang, G. (2022). Detection of transcription factors binding to methylated DNA by deep recurrent neural network. *Briefings Bioinform.* 23:bbab533. doi: 10.1093/bib/bbab533

Li, N., Sun, Z., and Jiang, F. (2008). Prediction of protein-protein binding site by using core interface residue and support vector machine. *BMC Bioinform.* 9:1–13. doi: 10.1186/1471-2105-9-553

Li, Y., Golding, G. B., and Ilie, L. (2021). DELPHI: accurate deep ensemble model for protein interaction sites prediction. *Bioinformatics* 37, 896–904. doi: 10.1093/bioinformatics/btaa750

Liu, M.-L., Su, W., Wang, J.-S., Yang, Y.-H., Yang, H., and Lin, H. (2020). Predicting preference of transcription factors for methylated DNA using sequence information. *Mol. Ther. Nucleic Acids* 22, 1043–1050. doi: 10.1016/j.omtn.2020.07.035

Nikolov, D., and Burley, S. (1997). RNA polymerase II transcription initiation: a structural view. *Proc. Natl. Acad. Sci.U.S.A.* 94, 15–22. doi: 10.1073/pnas.94.1.15

Nimwegen, E. (2006). Scaling laws in the functional content of genomes. *Trends Genet* 19, 236–253. doi: 10.1007/0-387-33916-7_14

Noble, W. S. (2006). What is a support vector machine? *Nat. Biotechnol.* 24, 1565–1567. doi: 10.1038/nbt1206-1565

Olah, C. (2015). *Understanding LSTM networks*. Available online at: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (accessed October 11, 2022).

Pawson, T. (1993). Signal transduction—a conserved pathway from the membrane to the nucleus. *Dev. Genet.* 14, 333–338. doi: 10.1002/dvg.1020140502

Riaño-Pachón, D. M., Ruzicic, S., Dreyer, I., and Mueller-Roeber, B. (2007). PlnTFDB: an integrative plant transcription factor database. *BMC Bioinform.* 8:1–10. doi: 10.1186/1471-2105-8-42

Rockel, S., Geertz, M., and Maerkl, S. J. (2012). MITOMI: a microfluidic platform for in vitro characterization of transcription factor–DNA interaction. *Methods Mol. Biol.* 786, 97–114. doi: 10.1007/978-1-61779-292-2_6

Roeder, R. G. (1996). The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* 21, 327–335. doi: 10.1016/S0968-0004(96)10050-5

Rong, X. (2014). word2vec parameter learning explained. *arXiv* [preprint]. doi: 10.48550/arXiv.1411.2738

Roulet, E., Busso, S., Camargo, A. A., Simpson, A. J., Mermod, N., and Bucher, P. (2002). High-throughput SELEX–SAGE method for quantitative modeling of transcription-factor binding sites. *Nat. Biotechnol.* 20, 831–835. doi: 10.1038/nbt718

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. *Adv. Neural Inform. Proc. Syst.* 2017:30.

Saha, S., and Raghava, G. P. S. (2006). Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins Struct Function Bioinform.* 65, 40–48. doi: 10.1002/prot.21078

Sak, H., Senior, A., Rao, K., and Beaufays, F. (2015). Fast and accurate recurrent neural network acoustic models for speech recognition. *arXiv* [preprint]. doi: 10.48550/arXiv.1507.06947

Shamovsky, I., and Nudler, E. (2008). New insights into the mechanism of heat shock response activation. *Cell. Mol. Life Sci.* 65, 855–861. doi: 10.1007/s00018-008-7458-y

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. D Nonlinear Phenomena* 404:132306. doi: 10.1016/j.physd.2019.132306

Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., et al. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* 35, 1285–1298. doi: 10.1109/TMI.2016.2528162

Tan, J. X., Li, S. H., Zhang, Z. M., Chen, C. X., Chen, W., Tang, H., et al. (2019). Identification of hormone binding proteins based on machine learning methods. *Math Biosci. Eng.* 16, 2466–2480. doi: 10.3934/mbe.2019123

Wang, G., Luo, X., Wang, J., Wan, J., Xia, S., Zhu, H., et al. (2018). MeDReaders: a database for transcription factors that bind to methylated DNA. *Nucleic Acids Res.* 46, D146–D151. doi: 10.1093/nar/gkx1096

Wheaton, K., Atadja, P., and Riabowol, K. (1996). Regulation of transcription factor activity during cellular aging. *Biochem. Cell Biol.* 74, 523–534. doi: 10.1139/o96-056

Wingender, E., Dietze, P., Karas, H., and Knüppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* 24, 238–241. doi: 10.1093/nar/24.1.238

Xi, E., Bing, S., and Jin, Y. (2017). Capsule network performance on complex data. *arXiv* [preprint]. doi: 10.48550/arXiv.1712.03480

Yang, J., Zhang, D., Frangi, A. F., and Yang, J. Y. (2004). Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Analy. Mach. Intelli.* 26, 131–137. doi: 10.1109/tpami.2004.1261097

Yashiro, T., Hara, M., Ogawa, H., Okumura, K., and Nishiyama, C. (2016). Critical role of transcription factor PU.1 in the function of the OX40L/TNFSF4 promoter in dendritic cells. *Sci. Rep.* 6, 1–11. doi: 10.1038/srep34825

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* 356:eaaj2239. doi: 10.1126/science.aaj2239

Zhang, G., Luo, W., Lyu, J., Yu, Z. G., and Huang, G. (2022). CNNLSTMac4CPred: a hybrid model for N4-acetylcytidine prediction. *Int. Sci. Comput. Life Sci.* 14, 439–451. doi: 10.1007/s12539-021-00500-0

Zhang, J., Ma, Z., and Kurgan, L. (2019). Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Briefings Bioinform.* 20, 1250–1268. doi: 10.1093/bib/bbx168

Zhang, Q., Liu, W., Zhang, H.-M., Xie, G.-Y., Miao, Y.-R., Xia, M., et al. (2020). hTFtarget: a comprehensive database for regulations of human transcription factors and their targets. *Geno. Proteomics Bioinform.* 18, 120–128. doi: 10.1016/j.gpb.2019.09.006

Zhu, Q. H., Guo, A. Y., Gao, G., Zhong, Y. F., Xu, M., Huang, M., et al. (2007). DPTF: a database of poplar transcription factors. *Bioinformatics* 23, 1307–1308. doi: 10.1093/bioinformatics/btm113

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2020). Sequence clustering in bioinformatics: an empirical study. *Briefings Bioinform.* 21, 1–10. doi: 10.1093/bib/bby090

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 08 December 2022
DOI 10.3389/fmicb.2022.1091698

OPEN ACCESS

EDITED BY
Yongchun Zuo,
Inner Mongolia University,
China

REVIEWED BY
Hao Wu,
Shandong University,
China
Shiyuan Wang,
Harbin Medical University,
China

*CORRESPONDENCE
Ruifang Li
liruifang@imnu.edu.cn

SPECIALTY SECTION
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 07 November 2022
ACCEPTED 22 November 2022
PUBLISHED 08 December 2022

CITATION
Li RF, Song XW, Gao S and Peng SY (2022) Analysis on the interactions between the first introns and other introns in mitochondrial ribosomal protein genes. *Front. Microbiol.* 13:1091698. doi: 10.3389/fmicb.2022.1091698

Analysis on the interactions between the first introns and other introns in mitochondrial ribosomal protein genes

Ruifang Li*, Xinwei Song, Shan Gao and Shiya Peng

College of Physics and Electronic Information, Inner Mongolia Normal University, Hohhot, China

It is realized that the first intron plays a key role in regulating gene expression, and the interactions between the first introns and other introns must be related to the regulation of gene expression. In this paper, the sequences of mitochondrial ribosomal protein genes were selected as the samples, based on the Smith-Waterman method, the optimal matched segments between the first intron and the reverse complementary sequences of other introns of each gene were obtained, and the characteristics of the optimal matched segments were analyzed. The results showed that the lengths and the ranges of length distributions of the optimal matched segments are increased along with the evolution of eukaryotes. For the distributions of the optimal matched segments with different GC contents, the peak values are decreased along with the evolution of eukaryotes, but the corresponding GC content of the peak values are increased along with the evolution of eukaryotes, it means most introns of higher organisms interact with each other though weak bonds binding. By comparing the lengths and matching rates of optimal matched segments with those of siRNA and miRNA, it is found that some optimal matched segments may be related to non-coding RNA with special biological functions, just like siRNA and miRNA, they may play an important role in the process of gene expression and regulation. For the relative position of the optimal matched segments, the peaks of relative position distributions of optimal matched segments are increased during the evolution of eukaryotes, and the positions of the first two peaks exhibit significant conservatism.

KEYWORDS

local matched alignment, first intron, optimal matched segments, mitochondrial ribosomal protein genes, interaction

Introduction

An intron sequence is regarded as a kind of non-coding sequence of interrupted gene, and its functions are being discovered. A large number of studies have shown that intron can regulate gene expression as a kind of regulatory element (Palmiter et al., 1991; Li et al., 2015; Abou et al., 2020), for example, the heterologous introns can enhance

TABLE 1 Mitochondrial ribosomal protein genes.

Species	The amount of genes	The amount of introns	The amount of first introns
Homo sapiens	114	512	114
Mus musculus	79	351	79
Fugu rubripes	69	266	64
Drosophila melanogaster	75	118	66
Caenorhabditis elegans	74	251	71

expression of transgenes in mice (Palmiter et al., 1991). In recent years, it has also been found that some introns can influence many stages of mRNA metabolism, including initial transcription of a gene, editing of pre-mRNA, and nuclear export, translation and decay of the mRNA (Bo et al., 2019). Furthermore, introns contain kinds of non-coding RNA such as microRNA and snoRNA (Mattick and Gagen, 2001), they also participate in the functional activities of a variety of non-coding RNA (Abou et al., 2020). And it has shown that GC-AG introns are mainly associated with lncRNAs and are preferentially located in the first intron (Abou et al., 2020), additionally, many studies have shown that introns are closely related to various diseases (Sowalsky et al., 2015; Malekkou et al., 2020; Ong and Adusumalli, 2020), for example, a novel mutation deep within intron 7 of the GBA gene can cause Gaucher disease (Malekkou et al., 2020), and increased intron retention is associated with Alzheimer's disease (Ong and Adusumalli, 2020), and metastatic castration-resistant prostate cancer is related to some non-coding RNA (Sowalsky et al., 2015).

The most basic and important interaction among bases is base matching, for example, the formation of a correct codon-anticodon pair is essential to ensure efficiency and fidelity during translation, and circRNA formed by exon cyclization or intron cyclization contains long flanking introns with complementary repeats (Han et al., 2022). Besides, many studies indicated that intron complementary matching fragments are not only the cause of circular RNA, but also the potential factors for the complexity and diversity of gene expression at the transcriptional/post transcriptional level (Zhang et al., 2013, 2014; Jiao et al., 2021). Therefore, it is particularly important to study the circular matching problem of introns.

The first introns have gained increasing attentions in recent years because of their unique features that are located in close proximity to the transcription, and the distinct deposition of epigenetic marks and nucleosome density on the first intronic DNA sequence (Fu et al., 2022; Singh et al., 2022; Spijker et al., 2022; Vosseberg et al., 2022), and it is realized that the first introns play a key role in several mechanisms regulating gene expression. We determined that the matching features between the first introns and the corresponding reverse complementary sequences of other introns must provide many useful information.

The genome consists of an extremely complex network of interactions among functional elements, and its functions are achieved primarily through these interactions. We have known that a complete match between siRNA and targeted genes can lead to targeted genes silencing, and high but incomplete matching between miRNA and targeted genes can suppress gene expression. It means base matching is an important way for non-coding RNA to interact with targeted genes, intron as a kind of non-coding DNA is rich in eukaryote genomes, introns must interact with each other, and the interactions can be embodied by the modes of base matching. Based on this, in this work, the mitochondrial ribosomal protein gene sequences were selected as samples, the characteristics of the optimal matched segments between the first intron and the corresponding reverse complementary sequences of other introns were analyzed, and the variations of the characteristics along with the evolution of eukaryotes were studied.

Materials and methods

Datasets

All the sequences of mitochondrial ribosomal protein genes in the Ribosomal Protein Gene Database (RPG) were selected as our samples, they were from *Homo sapiens, Mus musculus, Fugu rubripes, Drosophila melanogaster* and *Caenorhabditis elegans*. Considering that the mitochondrial ribosomal protein gene has many advantages in biological research as a kind of housekeeping gene, they are involved in the key process of all protein translation and have very good evolutionary conservatism (Yoshihama et al., 2002), thus forming a family of conservative genes. They exist widely in all eukaryotes, and their intron lengths and amounts have little difference in all eukaryotes. We believe that more reliable and functional interactions among introns can be obtained by selecting these conserved genes. Information about the protein genes is given in Table 1.

Matching method

The intron sequences were obtained from the above gene sequences, then, they were transformed into their reverse complementary sequences except the first introns. Next, similar alignments were done by the local similarity matching software called Smith-Waterman.[1] We adopt Ednafull matrix to similarity matching, and parameters chosen as follows, each Gap penalty is 50.0, in the gap each extend penalty is 5.0, thus, we got the optimal similar segment between the first intron and corresponding reverse complementary sequences of other introns in each gene sequence.

1 http://mobyle.pasteur.fr/cgi-bin

The optimal matching frequency

We calculated the length, GC content, matching rate of each optimal matched segment considering that they must provide the basic characteristics of the optimal matched segments, then divided the optimal matching segments into several groups, respectively, according to their lengths, GC contents or matching rates. And then calculated the frequencies of the optimal matched segments with different ranges of lengths, GC contents and matching rates, marked with F_{Lm}, F_{GCm}, and F_{mat} respectively, they were defined as follows,

$$F_{Lmi} = \frac{N_{Lmi}}{\sum_{i=1}^{n_L} N_{Lmi}} \tag{1}$$

$$F_{GCmj} = \frac{N_{GCmj}}{\sum_{j=1}^{n_{GC}} N_{GCmj}} \tag{2}$$

$$Fmatk = \frac{N_{matk}}{\sum_{k=1}^{n_{mat}} N_{matk}} \tag{3}$$

Where, F_{Lmi} is the frequency of the optimal matched segments whose length are within the ith group, N_{Lmi} is the amount of the optimal matched segments in the ith group, and n_L is the amount of the groups divided according to their lengths. F_{GCmj} is the frequency of the optimal matched segments whose GC contents are within the jth group, N_{GCmj} is the amount of the optimal matched segments in the jth group, and n_{GC} is the amount of the groups divided according to their GC contents. F_{matk} is the frequency of the optimal matched segments whose matching rate are within the kth group, N_{matk} is the amount of the optimal matched segments in the kth group, and n_{mat} is the amount of the groups divided according to their matching rates.

The lengths of the first introns in different gene sequences are different, we standardized the first introns as sequences with 100 bp length in order to conveniently compare the relative position distributions of the optimal matched segments. The method of length standardization as follows (Zhang et al., 2016a),

$$n_{ij} = \begin{cases} \lceil 100 * N_{ij} / L_i \rceil & 100 * N_{ij} / L_i \quad is\ integer \\ \lfloor 100 * N_{ij} / Li \rfloor + 1 & 100 * N_{ij} / L_i \ is\ non \llcorner integer \end{cases} \tag{4}$$

Where, L_i is the length of the ith first intron, N_{ij} is the jth base site of the ith first intron, and n_{ij} is its relative position corresponding to the ith standardized first intron. In this way, the first introns are all transformed into 100 bp long sequences.

According to the base site of each optimal matching sequence on the first intron, each base site of the first intron is scored, if in the optimal matching region, base site is scored 1, but if not, it is scored 0, and the definition of matching score as follows (Zhang et al., 2016a),

$$f_{ij} = \begin{cases} 1 & n_{ia} \le j \le n_{ib} \\ 0 & j \langle n_{ia} \ or \ j \rangle n_{ib} \end{cases} \tag{5}$$

Where, f_{ij} is the score of the jth base site on the standardized ith first intron, n_{ia} and n_{ib} are the initiation base relative site and the termination base relative site of the optimal matched segments on the standardized ith first intron. Thus, for each optimal matching sequence, the first intron is transformed into a sequence consisted of 0 and 1. if there are m optimal matching sequences in a gene, we can obtain m sequences consisted of 0 and 1. On this basis, we divided the 100 sites of each number sequences into 10 regions on average, the relative position frequency of the optimal matched segments on each site and in each region are defined as follows,

$$F_{rj} = \sum_{i=1}^{m} f_{ij} / \sum_{i=1}^{m} (N_{ib} - N_{ia} + 1) \tag{6}$$

$$F_{rk} = \sum_{j=p_{ka}}^{p_{kb}} \sum_{i=1}^{m} f_{ij} / \sum_{i=1}^{m} (N_{ib} - N_{ia} + 1) \tag{7}$$

Where, F_{rj} is the relative position frequency of the optimal matched segments on the jth base site of the standardized first intron, F_{rk} is the relative position frequency of the optimal matched segments in the kth region, f_{ij} is the score of jth base site on the standardized first intron, p_{ka} and p_{kb} are the initiation base site and the termination base site of the kth region, N_{ia} and N_{ib} are the initiation base site and the termination base site of the optimal matched segments, and m is the total number of the optimal matched segments in the gene.

Results

The optimal matched segments between the first introns and the reverse complementary sequences of other introns in each mitochondrial ribonucleo protein gene of five species were counted, and the dataset of the optimal matched segments was established. Then, the lengths of the optimal matched segments of each species were counted, and the frequencies of the optimal matched segments with different ranges of lengths were calculated by formula (1). The GC contents of the optimal matched segments of each species were counted, and the frequencies of the optimal matched segments with different ranges of GC contents were calculated by formula (2). The matching rates of the optimal matched segments of each species were calculated, and the frequencies of the optimal matched segments with different ranges

of matching rates were calculated by formula (3). The first introns were standardized to sequences with 100 bp length, and their relative base positions were calculated according to formula (4), and then, according to the base sites of each optimal matching sequence on the first intron, each base site of the first intron is scored according to formula (5), based on this, the relative position frequency were calculated according to formula (6) and (7). On this basis, the characteristics of the optimal matched segments of five species were analyzed. The results are presented in Figure 1.

The length distributions of the optimal matched segments

The lengths of the optimal matched segments of five species are mainly concentrated at 10–50 bp, while some optimal matched segments of *Homo sapiens, Mus musculus* and *Fugu rubripes* are up to 100 bp in length, and the ratio of the optimal matched segments of *Homo sapiens* concentrating at 90–100 bp is up to 24.6 percent. In addition, the length distribution of the optimal matched segments of *Homo sapiens* is similar to that of *Mus musculus*. The results showed that the optimal matched segments of high eukaryotes have a longer length and a wider length distribution than that of the low eukaryotes, and it means the length and the ranges of length distribution of the optimal matched segments are increased along with the evolution of eukaryotes.

The GC content distributions of the optimal matched segments

The distributions of GC content of the optimal matched segments of five species ranged from 0 to 0.9. And comparing the results of the five species, it is found that the peak values of F_{GCm} are decreased along with the evolution of eukaryotes, but the corresponding GC content of the peak values are increased along with the evolution of eukaryotes.

The matching rate distributions of the optimal matched segments

As seen from the matching rate distributions of the optimal matched segments, most of the matching rates are distributed between 60% and 80%. Interestingly, studies showed that the matching rates between miRNA and target mRNA are distributed between 65% and 95% (Cui et al., 2010). Comparing the matching rate ranges of the optimal matched segments with that of siRNA or miRNA with target mRNA (Volpe et al., 2002; Lim et al., 2005; Cui et al., 2010; Zhang et al., 2016b), it is found that there is a high similarity between the optimal matched segments with the most probable matching rates and siRNA or miRNA, this also suggests these optimal matched segments may be related to some non-coding RNAs with special biological functions, just like siRNA and miRNA, they may play an important role in the process of gene expression and regulation.

The relative position distributions of the optimal matched segments in the first introns

It can be seen from Figure 1 that the relative positions of the optimal matched segments vary with the different species. And the peaks with *Fr* values bigger than 10% are analyzed, the results showed that for *Homo sapiens*, there are 3 peaks with Fr values bigger than 10%, which are at 20–30 bp, 50–60 bp and 90–100 bp, and there is two peaks with Fr values bigger than 10%, which are at 20–30 bp and 50–60 bp for *Mus musculus*, 30–40 bp and 60–70 bp for *Fugu rubripes* and *Drosophila melanogaster*, but for *Caenorhabditis elegans*, there is only one peak with Fr value bigger than 10%, which is at 70–80 bp. This also indicates that the relative position frequencies of the optimal matched segments are distinctively differences among the five different species, and the peaks of relative position distributions of optimal matched segments are increased along with the evolution of eukaryotes, but the positions of the first two peaks exhibit significant conservatism.

In order to further confirm the conservatism of the relative position distributions of optimal matched segments, Figure 2 was made according to the calculations by formula (6), which express the relative position frequencies with the base sites of the standardized first intron.

As we can see from Figure 2, the distributions do not accord with normal distribution, so, for the relative position frequency of the optimal matched segments on each site of the standardized first intron, the test for differences between any two species were performed by non-parametric test with level of significance 0.05 using R software, the results are presented in Table 2.

It can be seen from Table 2 that the *p*-values between any two species are >0.05, it indicates that the differences between any two species were not statistically significant, means, in terms of the distributions on each site of the standardized first intron, the optimal matched segments exhibited high conservatism during species evolution. These results further confirmed the conservatism of the relative position distributions of optimal matched segments, which let us be sure that optimal matched segments are organized and functional sequences for each species.

Conclusion

We analyzed the matching features between the first intron and the reverse complementary sequences of other introns in each gene sequence, for the lengths of the optimal matched segments, we found the most probable lengths are distributed between 20 and 30 bp, it can be seen from Figure 1, and we calculated the

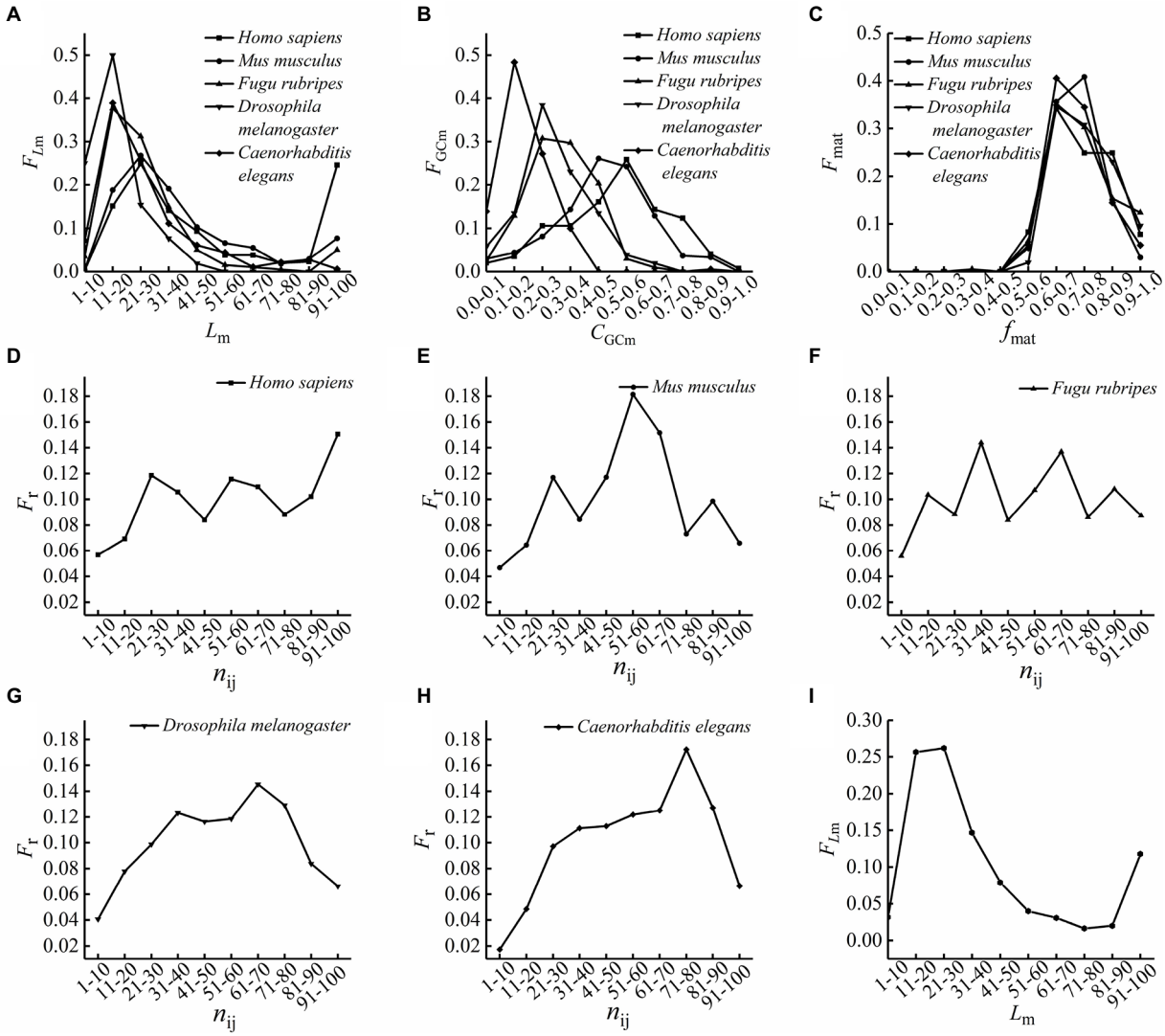

FIGURE 1
The distributions of the optimal matching frequencies. **(A)** The length (L_m) distributions of the optimal matched segments. The values on the *x*-axis are the length ranges of groups divided according to lengths of the optimal matched segments, and those on the *y*-axis are the length frequency of the optimal matched segments. **(B)** The GC content (C_{GCm}) distributions of the optimal matched segments. The values on the *x*-axis are GC content ranges of groups divided according to GC contents of the optimal matched segments, and those on the *y*-axis are the GC content frequencies of the optimal matched segments. **(C)** The matching rate (f_{mat}) distributions of the optimal matched segment of five species. The values on the *x*-axis are matching rate ranges of groups divided according to matching rates of the optimal matched segments, and those on the *y*-axis are the matching rate frequencies of the optimal matched segments. **(D–H)** The relative position distributions of the optimal matched segments on the first intron sequences. The values on the *x*-axis are relative position ranges of groups divided according to relative positions of the optimal matched segments, and those on the *y*-axis are the relative position frequencies of the optimal matched segments. **(D)** *Homo sapiens*, **(E)** *Mus musculus*, **(F)** *Fugu rubripes*, **(G)** *Drosophila melanogaster*, **(H)** *Caenorhabditis elegans*. **(I)** The length (L_m) distributions of the optimal matched segments of 5 species. The calculations of the optimal matched segment lengths of 5 species are combined into one, the optimal matching segments were divided into several groups according to their lengths, the frequency of the optimal matched segments in each group was calculated, the values on the *x*-axis are the length ranges of groups, and those on the *y*-axis are the length frequency of the optimal matched segments.

ratios of the optimal matched segments whose lengths are between 21 and 30 bp, they are 25%, 27%, 31%, 15%, and 26% in *Homo sapiens*, *Mus musculus*, *Fugu rubripes*, *Drosophila melanogaster* and *Caenorhabditis elegans, respectively.* Interestingly, we know that the siRNA, whose length is from 21 to 25 bp, guiding mRNA to silent by perfect complementarity with target mRNA (Volpe et al., 2002; Lim et al., 2005), and the miRNA, whose length is from 18 to 25 bp, restrains transcription and expression of target mRNA by different degree complementarity with target mRNA (Zhang et al., 2016a), the results indicate that the probable lengths of the optimal matched segments are remarkably similar to the lengths of siRNA and miRNA. For the matching rates of the optimal matched segments, we found that most of the matching rates are distributed between 60% and 80%, they are very

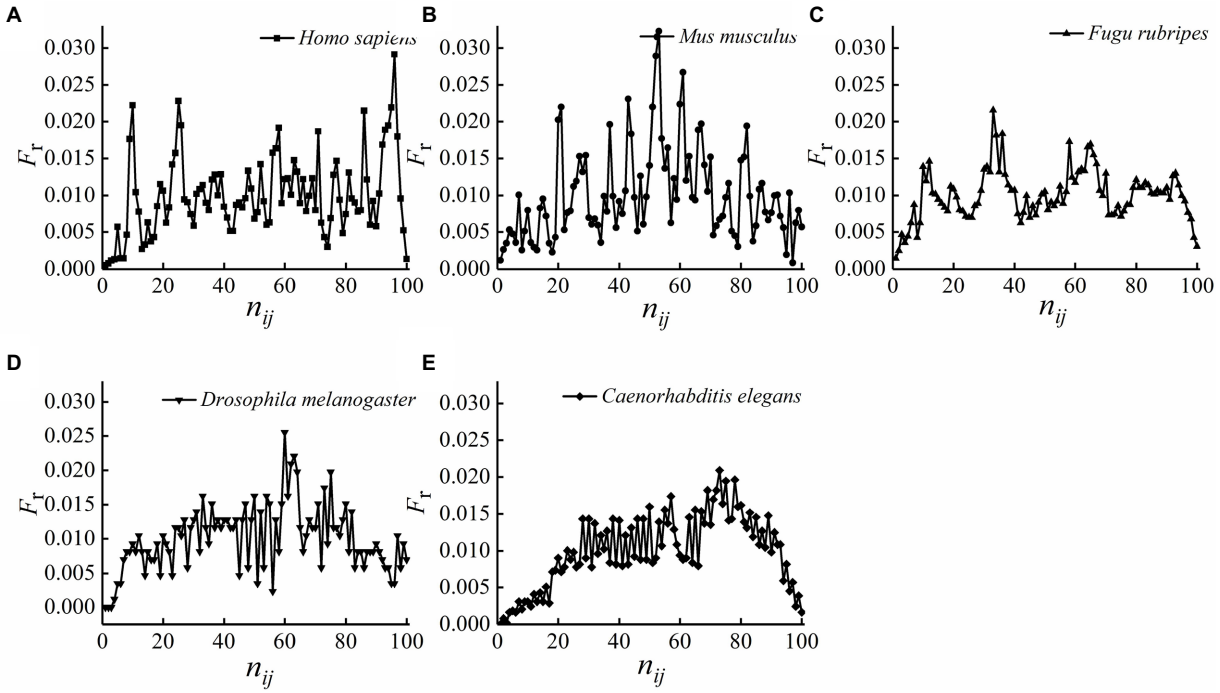

FIGURE 2
(A–E) The distribution of the relative position frequencies with the base sites of the standardized first intron.

TABLE 2 Results of the test for differences of relative position frequency of the optimal matched segments on each site of the standardized first intron.

Species	Value of p	Species	Value of p
Homo sapiens-Mus musculus	0.54	*Mus musculus-Drosophila melanogaster*	0.30
Homo sapiens-Fugu rubripes	0.40	*Mus musculus-Caenorhabditis elegans*	0.33
Homo sapiens-Drosophila melanogaster	0.66	*Fugu rubripes-Drosophila melanogaster*	0.91
Homo sapiens-Caenorhabditis elegans	0.52	*Fugu rubripes-Caenorhabditis elegans*	0.68
Mus musculus-Fugu rubripes	0.10	*Drosophila melanogaster-Caenorhabditis elegans*	0.64

remarkably similar to the matching rate ranges with target mRNA of siRNA or miRNA. It means there is a high similarity between some optimal matched segments and siRNA or miRNA. Is this a coincidence? we do not think so. The basic interaction between introns is base complementary pairing, the matched sequences, especially between the first intron and corresponding complementary sequences of other introns in the same gene, must be related to some elements with special functions. Taking all the analyzes and conclusions above into account, we come to a conclusion that some optimal matched segments may be a kind of non-coding RNA with special biological functions, just like siRNA and miRNA, they are likely to participate in the process of gene expression and regulation. And we think, the optimal matched segments with special characteristics in the first introns may take part in regulating gene expression by RNA matching competition with other introns or exon.

In addition, we have got some interesting results by comparing the results of different species. In terms of the species selected in this work, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Fugu rubripes*, *Mus musculus*, and *Homo sapiens* are listed from lower eukaryotes to higher eukaryotes. Based on this order and the calculations, we tried to analyze the variation law of matching features of introns along with the species evolution. For the lengths of the optimal matched segments, the average length of the optimal matched segments for the high eukaryotes are longer than that of the low eukaryotes. It suggests that the lengths of the optimal matched segments are increased in the evolution of eukaryotes. And the results showed that with the evolution of eukaryotes, the distributions of the length of the optimal matched segment become wider and wider. It means the lengths and the ranges of length distributions of the optimal matched segments are increased along with the evolution of eukaryotes. For the GC content of the optimal matched segments, the peak values of F_{GCm} are decreased along with the evolution of eukaryotes, it suggests

the GC contents of the optimal matched segments are more widely distributed with the evolution of eukaryotes. If some functional elements are related to the optimal matched segments with special GC contents, the result means that higher organisms have more kinds of functional elements than lower organisms. But the corresponding GC content at the peak values are increased along with the evolution of eukaryotes. Both AT and GC basepairs form one set of hydrogen bonds, and it is a truth universally acknowledged that a GC base pair has three hydrogen bonds whereas AT has two, it means that DNA with high GC-content is more stable than DNA with low GC-content. Based on the above theories, it can be concluded that introns of higher organisms interacting with each other though weak bonds binding are more than that of lower organisms, we hypothesized that interactions through weak bonds can ensure the flexibility to take part in gene regulation. For the relative position of the optimal matched segments, the peaks of relative position distributions of optimal matched segments are increased with the evolution of eukaryotes, and the positions of the first two peaks exhibit significant conservatism. We think that some functional elements are related to the optimal matched segments at these proper positions, the results indicated that these elements of higher eukaryotes may have a more specific division of labor.

To conclude, in this work, we analyzed the possibility of interactions between the first introns and the other introns of mitochondrial ribosomal protein genes, then tried to interpret the modes of interactions between introns. We found some universal characteristics of the optimal matched segments between the first introns and the reverse complementary sequences of other introns, and we noticed there is a high similarity between some optimal matched segments and siRNA or miRNA, so, we believe that the characteristics of interactions among introns obtained in this work are the basic characteristics of the RNA–RNA interactions. It means the optimal matched segments are probably functional non-coding RNAs including siRNA and miRNA. At the same time, we found some variation law of the optimal matched segments with the evolution of eukaryotes, which indicates that there may be a great difference in the complexity of the interactions of introns among species at different evolutionary levels, these results are of great significance in explaining the function of non-coding RNA. An increasing number of people are realizing the importance of non-coding RNA in gene expression regulating, but the functions and regulating mechanisms are not very clear, our results indicate that the base matching plays a key role in the interactions among introns. However, the related works have just started, the sample size in this study is relatively small, further large-sample studies are needed to obtain more detailed and clearer mechanism of introns interactions.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

RFL, XWS, SG, and SYP performed the work. RFL has developed the theoretical frame work, finished the data analysis, and wrote the manuscript. XWS, SG, and SYP finished the data collection and the calculations. All authors have read and approved this version of the article.

Funding

This work was supported by the Natural Science Foundation of Inner Mongolia (2019MS03042) and the National Natural Science Foundation of China (31860304).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abou, A. M., Celli, L., Belotti, G., Lisa, A., and Bione, S. (2020). GC-AG introns features in long non-coding and protein-coding genes suggest their role in gene expression regulation. *Front. Genet.* 11:488. doi: 10.3389/fgene.2020.00488

Bo, S., Li, H., Zhang, Q., Lu, Z., Bao, T., and Zhao, X. (2019). Potential relations between post-spliced introns and mature mRNAs in the Caenorhabditis elegans genome. *J. Theor. Biol.* 467, 7–14. doi: 10.1016/j.jtbi.2019.01.031

Cui, J. G., Zhao, Y., Sethi, P., Li, Y. Y., Mahta, A., Culicchia, F., et al. (2010). Micro-RNA-128 (miRNA-128) down-regulation in glioblastoma targets ARP5 (ANGPTL6), Bmi-1 and E2F-3a, key regulators of brain cell proliferation. *J. Neuro-Oncol.* 98, 297–304. doi: 10.1007/s11060-009-0077-0

Fu, L., Crawford, L., Tong, A., Luu, N., Tanizaki, Y., and Shi, Y. B. (2022). Sperm associated antigen 7 is activated by T3 during Xenopus tropicalis metamorphosis via a thyroid hormone response element within the first intron. *Develop. Growth Differ.* 64, 48–58. doi: 10.1111/dgd.12764

Han, Z. P., Chen, H. F., Guo, Z. H., Shen, J., Luo, W. F., Xie, F. M., et al. (2022). Circular RNAs and their role in exosomes. *Front. Oncol.* 12:848341. doi: 10.3389/fonc.2022.848341

Jiao, S., Wu, S., Huang, S., Liu, M., and Gao, B. (2021). Advances in the identification of circular RNAs and research into circRNAs in human diseases. *Front. Genet.* 12:665233. doi: 10.3389/fgene.2021.665233

Li, N. Q., Yang, J., Cui, L., Ma, N., Zhang, L., and Hao, L. R. (2015). Expression of intronic miRNAs and their host gene Igf2 in a murine unilateral ureteral obstruction model. *Braz. J. Med. Biol. Res.* 48, 486–492. doi: 10.1590/1414-431X20143958

Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., et al. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 433, 769–773. doi: 10.1038/nature03315

Malekkou, A., Sevastou, I., Mavrikiou, G., Georgiou, T., Vilageliu, L., Moraitou, M., et al. (2020). A novel mutation deep within intron 7 of the GBA gene causes Gaucher disease. *Mol. Genet. Genom. Med.* 8:e1090. doi: 10.1002/mgg3.1090

Mattick, J. S., and Gagen, M. J. (2001). The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Mol. Biol. Evol.* 18, 1611–1630. doi: 10.1093/oxfordjournals.molbev.a003951

Ong, C. T., and Adusumalli, S. (2020). Increased intron retention is linked to Alzheimer's disease. *Neural Regen. Res.* 15, 259–260. doi: 10.4103/1673-5374.265549

Palmiter, R. D., Sandgren, E. P., Avarbock, M. R., Allen, D. D., and Brinster, R. L. (1991). Heterologous introns can enhance expression of transgenes in mice. *Proc. Nati. Acad. Sci. U. S. A.* 88, 478–482. doi: 10.1073/pnas.88.2.478

Singh, O. P., Mishra, S., Sharma, G., Sindhania, A., Kaur, T., Sreehari, U., et al. (2022). Evaluation of intron-1 of odorant-binding protein-1 of Anopheles stephensi as a marker for the identification of biological forms or putative sibling species. *PLoS One* 17:e0270760. doi: 10.1371/journal.pone.0270760

Sowalsky, A. G., Xia, Z., Wang, L., Zhao, H., Chen, S., Bubley, G. J., et al. (2015). Whole transcriptome sequencing reveals extensive Unspliced mRNA in metastatic castration-resistant prostate cancer. *Mol. Cancer Res.* 13, 98–106. doi: 10.1158/1541-7786.MCR-14-0273

Spijker, H. M. V., Stackpole, E. E., Almeida, S., Katsara, O., Liu, B., Shen, K., et al. (2022). Ribosome profiling reveals novel regulation of C9ORF72 GGGGCC repeat-containing RNA translation. *RNA* 28, 123–138. doi: 10.1261/rna.078963.121

Volpe, T. A., Kidner, C., Hall, I. M., Teng, G., Grewal, S. I. S., and Martienssen, R. A. (2002). Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* 297, 1833–1837. doi: 10.1126/science.1074973

Vosseberg, J., Schinkel, M., Gremmen, S., and Snel, B. (2022). The spread of the first introns in proto-eukaryotic paralogs. *Commun. Biol.* 5:476. doi: 10.1038/s42003-022-03426-5

Yoshihama, M., Uechi, T., Asakawa, S., Kawasaki, K., Kato, S., Higa, S., et al. (2002). The human ribosomal protein genes: sequencing and comparative analysis of 73 genes. *Genome Res.* 12, 379–390. doi: 10.1101/gr.214202

Zhang, Q., Li, H., Zhao, X. Q., Xue, H., Zheng, Y., Meng, H., et al. (2016b). The evolution mechanism of intron length. *Genomics* 108, 47–55. doi: 10.1016/j.ygeno.2016.07.004

Zhang, Q., Li, H., Zhao, X. Q., Zheng, Y., Meng, H., Jia, Y., et al. (2016a). Analysis on the preference for sequence matching between mRNA sequences and the corresponding introns in ribosomal protein genes. *J. Theor. Biol.* 392, 113–121. doi: 10.1016/j.jtbi.2015.12.003

Zhang, X. O., Wang, H. B., Zhang, Y., Lu, X., Chen, L. L., and Yang, L. (2014). Complementary sequence-mediated exon circularization. *Cells* 159, 134–147. doi: 10.1016/j.cell.2014.09.001

Zhang, Y., Zhang, X. O., Chen, T., Xiang, J. F., Yin, Q. F., Xing, Y. H., et al. (2013). Circular Intronic long noncoding RNAs. *Mol. Cell* 51, 792–806. doi: 10.1016/j.molcel.2013.08.017

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 27 February 2023
DOI 10.3389/fgene.2023.1151172

OPEN ACCESS

EDITED BY
Hao Lin,
University of Electronic Science and Technology of China, China

REVIEWED BY
Lei Yang,
Harbin Medical University, China
Hu Meng,
Inner Mongolia University of Science and Technology, China

*CORRESPONDENCE
Zhanyuan Lu,
lzhy2811@163.com
Zhongxian Li,
lizhongxian@immu.edu.cn
Xiaoqing Zhao,
zhaoxq204@163.com

†These authors share first authorship

SPECIALTY SECTION
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics

RECEIVED 25 January 2023
ACCEPTED 15 February 2023
PUBLISHED 27 February 2023

CITATION
Bo S, Sun Q, Ning P, Yuan N, Weng Y, Liang Y, Wang H, Lu Z, Li Z and Zhao X (2023), A novel approach to analyze the association characteristics between post-spliced introns and their corresponding mRNA.
Front. Genet. 14:1151172.
doi: 10.3389/fgene.2023.1151172

A novel approach to analyze the association characteristics between post-spliced introns and their corresponding mRNA

Suling Bo[1†], Qiuying Sun[2†], Pengfei Ning[1], Ningping Yuan[1], Yujie Weng[1], Ying Liang[1], Huitao Wang[1], Zhanyuan Lu[3,4,5,6*], Zhongxian Li[1*] and Xiaoqing Zhao[3,4,5,6*†]

[1]College of Computer Information, Inner Mongolia Medical University, Hohhot, China, [2]Department of Oncology, Inner Mongolia Cancer Hospital and The Affiliated People's Hospital of Inner Mongolia Medical University, Hohhot, China, [3]Inner Mongolia Academy of Agricultural and Animal Husbandry Sciences, Hohhot, China, [4]School of Life Science, Inner Mongolia University, Hohhot, China, [5]Key Laboratory of Black Soil Protection And Utilization (Hohhot), Ministry of Agriculture and Rural Affairs, Hohhot, China, [6]Inner Mongolia Key Laboratory of Degradation Farmland Ecological Restoration and Pollution Control, Hohhot, China

Studies have shown that post-spliced introns promote cell survival when nutrients are scarce, and intron loss/gain can influence many stages of mRNA metabolism. However, few approaches are currently available to study the correlation between intron sequences and their corresponding mature mRNA sequences. Here, based on the results of the improved Smith-Waterman local alignment-based algorithm method (SW method) and binding free energy weighted local alignment algorithm method (BFE method), the optimal matched segments between introns and their corresponding mature mRNAs in *Caenorhabditis elegans* (C.elegans) and their relative matching frequency (RF) distributions were obtained. The results showed that although the distributions of relative matching frequencies on mRNAs obtained by the BFE method were similar to the SW method, the interaction intensity in 5'and 3'untranslated regions (UTRs) regions was weaker than the SW method. The RF distributions in the exon-exon junction regions were comparable, the effects of long and short introns on mRNA and on the five functional sites with BFE method were similar to the SW method. However, the interaction intensity in 5'and 3'UTR regions with BFE method was weaker than with SW method. Although the matching rate and length distribution shape of the optimal matched fragment were consistent with the SW method, an increase in length was observed. The matching rates and the length of the optimal matched fragments were mainly in the range of 60%–80% and 20-30bp, respectively. Although we found that there were still matching preferences in the 5'and 3'UTR regions of the mRNAs with BFE, the matching intensities were significantly lower than the matching intensities between introns and their corresponding mRNAs with SW method. Overall, our findings suggest that the interaction between introns and mRNAs results from synergism among different types of sequences during the evolutionary process.

KEYWORDS

intron, alignment method, interaction, mRNA, evolutionary process

1 Introduction

The past decades have witnessed unprecedented medical breakthroughs. In this respect, the decade-long human genome project, ENCODE (Encyclopedia of DNA Elements) project improved our understanding that the human genome is a complex network system in which individual genes, regulatory elements, and DNA sequences unrelated to coding proteins interact in an overlapping manner to jointly control human physiological activities (The ENCODE Project Consortium, 2007; Zhang et al., 2007). The ENCODE project debunked the concept of "junk DNA", which refers to very small protein-coding genes that are just one of many DNA elements with specific functions. It was also found that 93% of the DNA in the human genome could be transcribed into RNA, and many transcripts were non-coding RNA that could interact with each other (Comeron, 2001; Mattick and Gagen, 2001; Nott et al., 2003; Roy et al., 2003; Gabriel et al., 2005; Gazave et al., 2007).

Intron sequences represent an important and special class of ncRNA transcripts. They are transcribed together with mRNA and spliced to form a relatively independent class of ncRNA. The corresponding mature mRNA is the most important class of transcripts for storing genetic information and performing biological functions. According to the results of ENCODE project, an interaction is present between these two types of transcripts. Although it has been established that intron sequences (especially post-spliced introns) are regulatory elements with biological functions, their functions warrant further systematic study and exploration.

Intron sequences are carriers of important functional elements. It has been found that introns have many important biological functions and actively regulate gene expression. Six definite functions of spliceosome introns have been documented (Fedorova and Fedorov, 2003). Over the years, it has been shown that intron sequences are the vectors of important eukaryotic elements and play important biological functions in eukaryotic gene expression.

Intron loss/gain can affect many stages of mRNA metabolism. The gain and loss of intronic genes can affect the evolution of eukaryotes (Duret, 2001; Maquat and Carmichael, 2001; Jeffares et al., 2006; Nguyen et al., 2006; Roy and Hartl, 2006; Fawcett et al., 2011; Landen et al., 2022). Many experiments have found that introns play important roles in mRNA metabolism, such as transcription, splicing, nuclear transport and translation, as well as in regulating or maintaining the dynamic structure of mRNA (Le Hir et al., 2003; Elmonir et al., 2010; He et al., 2010; MariatiHo et al., 2010). At the transcription level, introns in many genes can significantly improve their transcription efficiency (Alexander et al., 2010; Akaike et al., 2011). In mice, the transcription levels of transgenes containing introns are 10–100-fold higher than those without introns (McKenzie and Brennan, 1996). It has been established that at the level of mRNA editing, introns are directly involved in splicing and contribute to synthesizing the 5'cap and 3'tail of the mRNA. An increasing body of evidence suggests that the cap structure can promote splicing and enhance the excision of its proximal first intron (Komarnitsky et al., 2000; Lewis and Izaurflde, 2004). During mRNA nuclear export, intron splicing is directly related to mRNA export (Le Hir et al., 2000; Gatfield et al., 2001; Kim et al., 2001; Lykke-Andersen et al., 2001). Early experiments have shown that mRNAs transcribed from cDNA cannot exit the nucleus and thus cannot express proteins, whereas mRNAs containing introns can exit the nucleus and express proteins (Ryu and Mertz, 1989; Rafiq et al., 1997). Besides, there is a growing consensus that introns can also affect the translation efficiency of mRNA (Torrado et al., 2009; Li and Pintel, 2012; Rocchi et al., 2012). Interestingly, Braddock et al. found that when a mature mRNA was injected directly into *Xenopus* oocytes, its translation was inhibited. This effect could be abolished by adding a spliceable intron to the 3'UTR of the gene or by injecting the FRGY2 antibody into the cytoplasm (Braddock et al., 1994). Indeed, intron deletion/gain can regulate gene expression at many stages of mRNA metabolism.

Introns can promote cell survival under stress. It is well-established that introns can regulate the survival and apoptosis of biological cells at the cellular level. In 2019, two research groups by Parenteau and Morgan found that yeast cells lack essential nutrients during the growth phase. Intriguingly, introns could accumulate by forming pre-mRNA (the Parenteau research group used pre-mRNA to judge the role of introns) or post-spliced intron (the Morgan research group used post-spliced) intron defines the function of introns) to adjust the rate of cell growth to adapt to this changing environment (Combs et al., 2006; Parenteau et al., 2008; Munding et al., 2013; Wanichthanarak et al., 2015; Awad et al., 2017; Venkataramanan et al., 2017; Wan et al., 2017; Morgan et al., 2019; Parenteau et al., 2019), thereby helping its survival. Results of these studies indicate that the huge family of intron sequences may have many potential functions and unknown binding ways, which warrant further exploration.

The use of binding free energy is an important means of studying RNA-RNA interactions. Based on the binding free energy principle, relative binding free energy calculation represents an effective means to study the interaction between biological macromolecules. During the analysis of the expression of coding RNA and the function of non-coding RNA, the minimum binding free energy method is used to predict its structure and further infer its close association, It has been established that 40%–70% of the known base pairs of RNA below 700bp can be correctly predicted (Deigan et al., 2009). The method of calculating free energy is also widely used in protein folding (Jackson, 1998; Schaefer et al., 1998; Selkoe, 2003), protein structure prediction (Bower et al., 1997; Zhang and Skolnick, 2005; Faraggi et al., 2009), molecular docking (Woo and Roux, 2005; Woo, 2008; Mitomo et al., 2009; Hay and Scrutton, 2012), and analysis of the interaction between biological macromolecules (Tollenaere, 1996; Anderson, 2003; Manke et al., 2003; Thomas et al., 2003; Gao et al., 2004; Martin and MacNeill, 2004; Prathipati et al., 2007). Introns and mRNAs are two types of RNA sequences. The binding free energy principle represents an important way to calculate the sequence interaction (mutual matching).

Based on the Smith-Waterman local alignment method, Li Hong et al. documented interactions between spliced introns and corresponding mRNA/CDS, and the distribution of their preferred interaction regions was universal among species. Since there are obvious differences in the binding free energies of base-base (A-T, C-G) during sequence matching, it is essential to fully consider these differences in binding free energies and to further

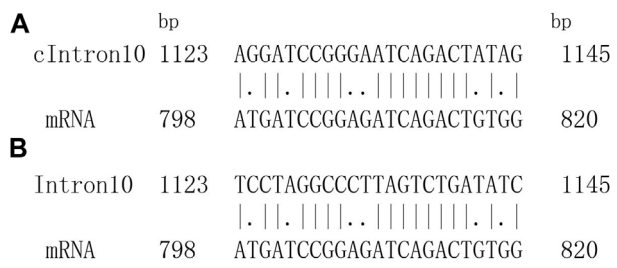

FIGURE 1
One optimal matching segment pair between introns sequence and corresponding mRNA sequence. **(A)** One optimal local alignment of fragment. cIntron10 represents complementary fragment of intron10. **(B)** Authentic matching alignment.

study the matching association between introns and mRNA sequences from the perspective of thermodynamic stability.

Herein, the protein-coding genes in the genome of C. elegans were analyzed. The local high-throughput combined with free energy weighted local alignment method was used to perform local matching analysis of introns and mRNA sequences, to characterize the distribution of preferred regions of intron-associated fragments on mRNA sequences and near functional sites, and to analyze the sequence structure characteristics. We identified the putative biological functions of spliced introns and revealed the evolutionary relationship between introns and corresponding mRNA sequences, which lays the groundwork for exploring the potential biological functions of spliced introns and other ncRNAs.

2 Materials and methods

2.1 The gene sequences

The *C. elegans* genome and its annotation information were downloaded from the Beijing Multi Subnet of Gene Bank (ftp://ftp.cbi. pku.edu.cn/pub/database/genomes). The protein-coding genes of the *C. elegans* genome were selected as our dataset. In this dataset, the genes which contain ncRNAs and/or repetitive elements were excluded first. Next, the genes whose intron lengths are shorter than 40 bp were removed because the 5'splice region (about 8bp) and 3'splice region contain a pyrimidine-rich layer (about 30bp) of introns and functional regions conserved over evolutionary time (Petrov, 2002), and introns below 40 bp do not play other roles. Finally, after genes associated with alternative splicing were excluded, we obtained the *C. elegans* genome consisting of 5736 genes and 24312 introns.

2.2 Matched alignment

If interactions were found between introns and their mRNAs, there were positively matched segment pairs between introns and their mRNAs and *vice versa*. The potential interaction between introns and their mRNAs can be represented by the optimal matched segments (OMS). To obtain the OMS, the introns were first transformed into their complementary sequences. Next, the mRNAs were renamed as tested sequences and the complementary sequences of introns were renamed as aligned sequences; the assessment of similarity between different alignments was performed using an improved Smith-Waterman local alignment software (http://mobyle. pasteur. fr/cgi-bin/). Finally, the optimal similarity segments of the introns were transformed again into their complementary segments, which were the OMSs in the introns. During the similarity aligning process, the Ednafull matrix was used to calculate the OMS using the following parameters: 50.0 for the gap open penalty and 5.0 for the gap extension penalty.

Accordingly, an objective optimal matched segment of a tested sequence and its aligned sequence could be obtained. The local alignment sketch map is shown in Figure 1.

The method based on the weighted comparison of binding free energy involves maximizing the number of hydrogen bonds and predicting the minimum free energy structure according to the negative correlation between the number of hydrogen bonds and the free energy (Zuker and Sankoff, 1984). The effect of base stacking force is not considered for the time being. Suppose the energy obtained by combining A-T/T-A base pair is EA-T/T-A, and the energy obtained by the G-C/C-G base pair is EG-C/C-G, then EA-T/T-A/EGC/CG ≈ 2:3. Due to the different release energy between A-T/T-A base pair and G-C/C-G base pair. In that case, different weights were assigned to them in the specific alignment process. The following principles were adopted during the matching process: If the base pair was correct, +3.0 would be awarded. +2.0 would be added if the base pair was A-T/T-A. If the base pair was G-C/C-G, it increased by +3.0. In this way, A correct matching of base pairs A-T/T-A yields +5.0 and a correct matching of base pairs G-C/C-G yields +6.0. If the base pairing was wrong, the penalty was −4.0. In this paper, the Ednafull matrix was still used to calculate The optimal matched fragments between the intron sequence and its corresponding mRNA sequence by using the binding free energy weighted local alignment method, and the selected parameters were as follows: The gap open penalty was −50.0 and the gap extension penalty was −5.0 for each base site. Finally, an optimal local matching fragment was obtained with the highest probability of interaction between the two sequences.

Definition 1: Sequence length normalization

Due to the different lengths of the tested sequences, they were normalized into 100 to obtain the relative site distributions using the following method.

The relative base site (k) of the *j*th base site in the tested sequence is

$$k = \begin{cases} \left[\left(\frac{100}{L}\right)^{*} j\right] & \left(\frac{100}{L}\right)^{*} j \quad \text{is integer} \\ \left[\left(\frac{100}{L}\right)^{*} j\right] + 1 & \left(\frac{100}{L}\right)^{*} j \quad \text{is non} - \text{integer} \end{cases} \tag{1}$$

Where, j means the *j*th base site of the tested sequence, L is the length of the tested sequence. The square brackets are gauss integer functions which mean to take integer part of a real number. Thus, the different lengths of the tested sequences were normalized to 100.

Definition 2: matched score function

For a tested sequence, the matched score function (fk) is

$$f_k = \begin{cases} 1 & k_s \le k \le k_e \\ 0 & k\pi k_s \quad or \quad k\phi k_e \end{cases} \tag{2}$$

Where, and represent the start base site and the end base site of the optimal matched segment in the normalized tested sequence. The effective value 1 is assigned to each base site within the optimal matched segment, while the ineffective value 0 is assigned to the base sites outside the optimal matched segment. Accordingly, the matched score values are assigned to each base site in the normalized tested sequences.

Definition 3: matched frequency

For the tested sequences, matched frequency function (F) is

$$F = \frac{1}{N}\sum_{i=1}^{N} f_{ik} \tag{3}$$

Where, i means the *i*th tested sequence, N means the number of the tested sequence. F reflects the interacting probability or the potential interaction intensity in the *k*th relative base site of the normalized tested sequences between the tested and aligned sequences.

Definition 4: average matched frequency

The average matched frequency function (<F>) for each base site is

$$\langle F \rangle = \frac{1}{N}\sum_{i=1}^{N} \frac{l_i}{L_i} \tag{4}$$

Where, li is the length of the optimal matched segment for the *i*th tested sequence. Li is the length of the *i*th tested sequence. For our normalized tested sequences, Li = 100. The <F> indicates the average matched frequency of the N-tested sequences, and it is a constant value for each tested set.

Definition 5: relative matched frequency

The relative matched frequency function (RF) of the *k*th base site in N tested sequence is

$$RF = \frac{F}{\langle F \rangle} \tag{5}$$

Where, RF reflects the relative bias of each base site in the N-tested sequences. If RF > 1, it indicates that the interaction in the *k*th base site is preferred, and the regions with RF > 1 are termed optimal matched regions (OMR). RF = 1 represents an average matched frequency of base sites for tested sequences.

2.3 Information entropy analysis

Information entropy can be used to characterize the organizational nature of a sequence. Second-order informational redundancy D2 is a suitable parameter to describe the adjacent base correlation of the sequence (Luo and Li, 1991; Li, 1990).

For an analyzed sequence, the second-order informational redundancy D2 is defined as:

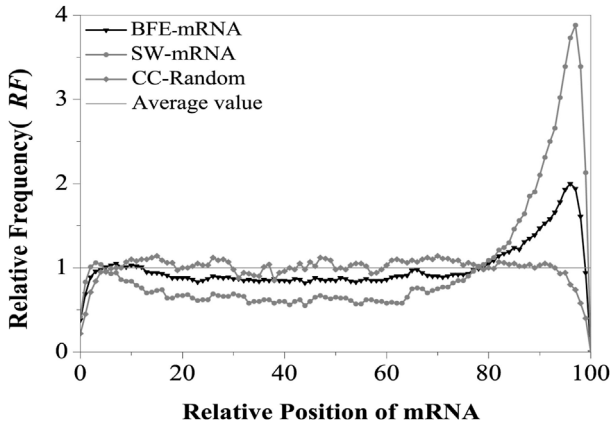

FIGURE 2
Relative Frequency (RF) distributions of mRNA. SW-mRNA means the RF value came from the base matching local alignment method. BFE-mRNA means the RF value came from the binding free energy weighted local alignment method. CC-Random means the local alignment were done between the component constraint random mRNA and their own component constraint random introns. Average value (RF = 1) means the theoretical average value of relative matched frequencies.

$$D_2 = \sum p_{ij}\log_2\left(p_{ij}/p_i p_j\right) \approx {}^{1}/_{2\ln 2}\sum \left(p_{ij} - p_i p_j\right)^2/p_i p_j \tag{6}$$

Where pi or pj is the probability of the base i or j (i, j = A, C, G, U), and pij is the joint probability of the base pair ij in the sequence. A bigger D2 value means that the base correlation is stronger. For a finite sequence of length N, the fluctuation bound (f.b.) of D2 is D2 (f.b.) = 15.65/N (Luo and Li, 1991; Luo, 2004). When D2≥15.65/N, the neighboring bases occur not independently and the correlation does exist at 99% confidence level. Generally, D2≥0. For infinite random sequences, D2 = 0.

3 Results and discussion

3.1 Matched alignment between mature mRNAs and their introns

The relative matched frequency distribution (RF) on the mRNA sequence was assessed using the binding free energy weighted local alignment method and denoted as BFE-mRNA distribution. For the control, the relative matched frequency distribution on the mRNA sequence was assessed using the improved Smith-Waterman local alignment method and denoted as SW-mRNA distribution. The intron sequence was taken as the comparison sequence, and the corresponding mRNA sequence was taken as the test sequence. The optimal matched fragment between the two types of sequences was obtained using the binding free energy weighted local alignment method. Finally, the optimal matched frequency distribution on the BFE-mRNA sequence was obtained (Figure 2).

The results showed that the relative matched frequency (RF) distribution on the BFE-mRNA sequence was similar to the SW-mRNA sequence, and there were two preferred regions at the 5'and

3'ends of mRNAs (Appendix A:Supporting Information S1). The first region was located at about 5%–12% of the 5'end of the mRNA, and its peak value was about 1.05. The second region was located between 80% and 98% of the 3'end of the mRNA, and its peak value was almost 2.0. The relative matched frequency of the 12%–80% region in the middle of the mRNA sequence was relatively low, slightly lower than the theoretical average, and its RF value fluctuated between 0.8 and 0.9. Compared with the CC-Random group (Appendix A:Supporting Information S2), The relative matched frequency (RF) of the BFE-mRNA sequence was more obvious in these two regions, and the difference in RF at the 3'end was highly significant (*t*-test, $p < 0.00002$).

Compared with SW-mRNA (Appendix A:Supporting Information S3), the preference of the relative matched frequency of the BFE-mRNA sequence was relatively weak at the 5'end region, exhibiting only one peak, which shifted slightly downstream. Although the distribution width of the preferred peak area at the 3'end remained unchanged, its peak value was only 1/2 that of the SW method and the difference was highly significant (*t*-test, $p < 0.00003$). The optimal relative matching frequency of the middle region was higher than the SW method, and it was significantly different from the CDS region (*t*-test, $p < 0.00001$) since the binding free energy weighted local alignment algorithm makes the optimal matched fragment combine with CDS with high G + C content.

The improved Smith-Waterman local alignment method and the binding free energy weighted local alignment method were used to describe the interaction between introns sequences and corresponding mRNA sequences. Analysis of the relative matched frequency distribution of mRNA sequence showed a consistent distribution preference by the two types of method. However, the regional difference in relative matched frequency distribution obtained by the base matched method was more obvious. To carefully analyze the distribution characteristics of each part of the mRNA sequence, The relative matched frequency distribution rule of each functional site region was studied next.

3.2 The distribution of relative matching frequency in functional site regions

There are many regions within the transcript that have regulatory functions, Such as translation initiation region, translation termination region and exon-exon junction region. The sequence of these functional domains plays a key role in the accurate expression of eukaryotic protein-coding genes. Therefore, it is necessary to explore the relative matched frequency of functional site regions.

The sites for translation initiation, translation termination and exon-exon junction are important functional regions of mRNA that regulate gene translation. Besides, the sequence of these functional regions is of great significance for the accurate expression of eukaryotic protein-coding genes. In this paper, we selected the ±60 bp regions of the translation start site (AUG), translation termination site (UAA) and exon-exon junction site (EE), which were denoted as AUG regions, UAA regions and EE regions, respectively, to analyze the relative matched frequency distribution of these regions by the BFE method, andcompared with that obtained by the SW method. (Castillo-Davis et al., 2002). Showed a close correlation between intron length and efficient gene expression. Halligan and Keightley et al. (Halligan and Keightley, 2006). Showed that long introns (>80bp) and short intron (≦80bp) distributions were significantly different. Therefore, we used 80bp as the threshold to distinguish between short and long introns.

Next, the introns were divided into three groups: An intron group, a long intron group and a short intron group named as intron, long intron and short intron, respectively. We compared and analyzed the overall differential characteristics of introns and the interactions between long and short introns with mRNA near functional sites. After obtaining The optimal matched fragment on the mRNA sequence, the distribution of the matched rate on the corresponding region was obtained by taking each functional site as the origin of the coordinate without length normalization.

3.2.1 Relative matched frequency distribution of AUG and UAA regions

Analysis of the relative matched frequency distribution of translation initiation and termination regions was conducted to verify whether the matching preference region at both ends of the BFE-mRNA sequence is located in the UTR region. To avoid a boundary effect during comparison, mRNA sequences with 5'UTR of less than 50bp and 3'UTR of less than 80bp were eliminated. Taking the first base of the translation start codon and translation stop codon as the coordinate origin, the relative matched frequency distribution characteristics of the translation start region and translation stop region were obtained (Figure 3).

As shown in Figure 3A, the peak distribution of mRNA was found at the 5'UTR and 3'UTR regions. The relative matched frequency at the 5'end gradually increased from −28bp of the AUG site and peaked at −10bp (RF = 1.3), then decreased to an average value of 10bp (RF = 1.0). Overall, The optimal matched fragments with introns ranged from −28bp to 10bp. The matched frequency of short introns in the AUG region was significantly higher than long introns, suggesting that short introns preferred interacting with the AUG region.

In the UAA region, the relative matched frequency distribution was significantly different from the AUG region (Figure 3B). From −28 bp of the UAA site, the relative matched frequency increased rapidly, the RF value reached 2.8 at the UAA site, peaked at about 28bp (RF = 3.8), and then gradually decreased, but the RF value remained high. In the 3'UTR region, it suggested that the interaction region is longer and much stronger than in the 5'UTR region. In addition, in the 3'UTR region, the interaction intensity of long introns was significantly higher than short introns, which is opposite to the AUG region, suggesting that long introns preferred to interact with the UAA region.

The results obtained by the base matching method (SW method) and the binding free energy weighted method (BFE method) were compared in the AUG and UAA region (Figures 4, 5).

Compared with the base matching method (SW method), the optimal matching frequency distribution trend of the AUG and UAA regions obtained by the binding free energy weighted methe binding free energy weighted local alignment method (BFE method) was similar. In the AUG region, The relative matched frequency distribution of both the whole intron group and the short intron group was slightly lower than that obtained by the base matching

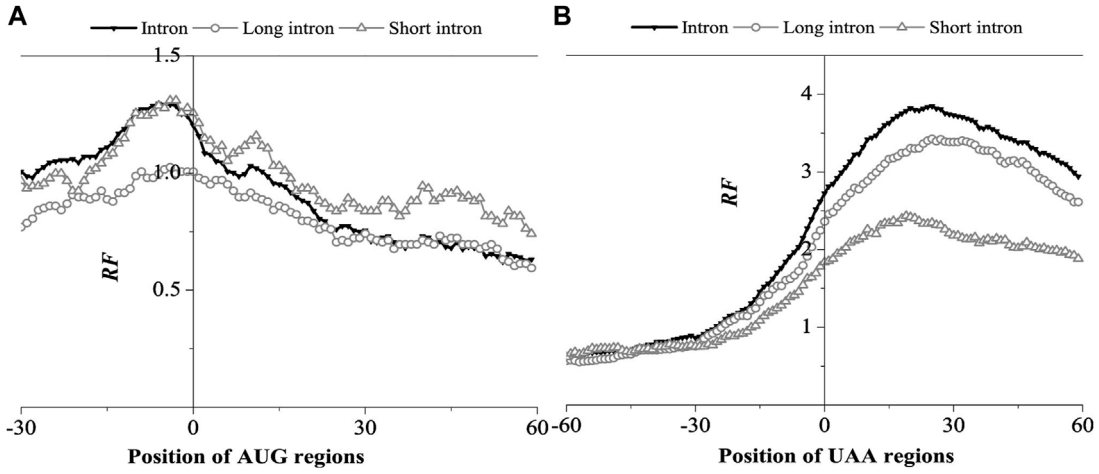

FIGURE 3
Relative Frequency (RF) distributions on AUG region **(A)** and UAA region **(B)** of mRNA. The RF distributions related long introns and short introns are also presented in the figure.

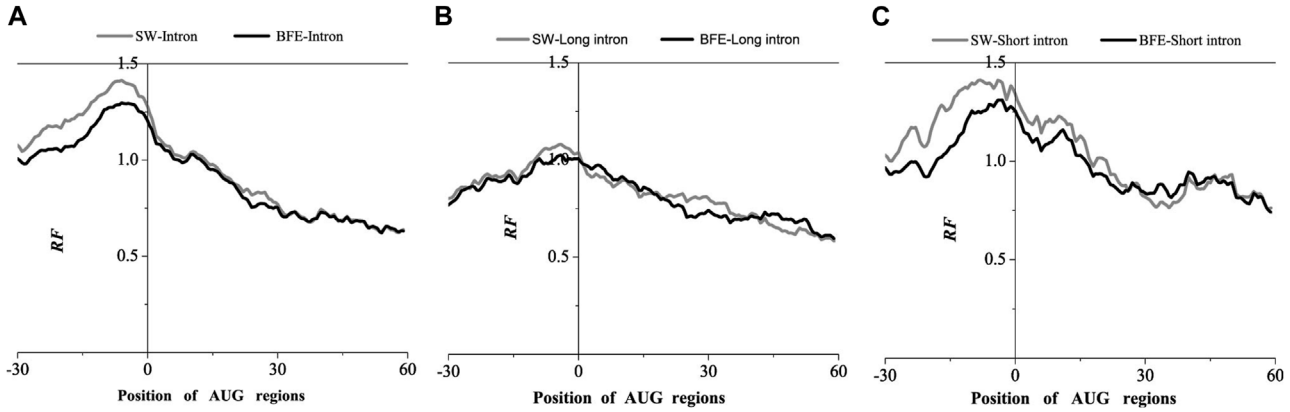

FIGURE 4
Comparations of Relative Frequency (RF) distributions between SW method and BFE method on AUG regions. **(A)** The total introns, **(B)** the long introns and **(C)** the short introns.

method (Figure 4), and the difference was more significant near the −10bp region of the AUG site. For long introns, the distribution was almost the same. In the UAA region, The relative matched frequency of the whole intron and long intron was significantly lower than the SW-mRNA group. Moreover, there was no significant difference in the distribution of short introns (Figure 5).

The analysis results of the two representative interactions indicated a significant preference for intron-mRNA interaction in the UTR region, especially in the 3'UTR region. Short and long introns preferentially acted in the 5'and 3'UTR region, respectively.

3.2.2 Relative matched frequency distribution in EE region

The EE region is divided into three groups: The first exon connection region, the intermediate exon connection region and the last exon connection region, composed of the corresponding exon connection site ±60 bp region. The relative matched frequency distribution was obtained by the binding free energy weighted local alignment method (BFE method), as shown in Figure 6.

The relative matched frequency distribution of EE regions in the three groups was similar. The relative matched frequency of the upstream region of the exon-exon junction site was higher than the downstream region. The difference was more significant in the first and last exon regions and least significant in the middle exon region. The minimum values of the distributions occurred 30bp downstream of the first exon connection point, while it is about 15bp downstream of the last exon connection point. However, there was no obvious difference in the minimum values of the distributions at the middle exon-exon junction. It was also found that the relative matched frequency of short introns was higher than

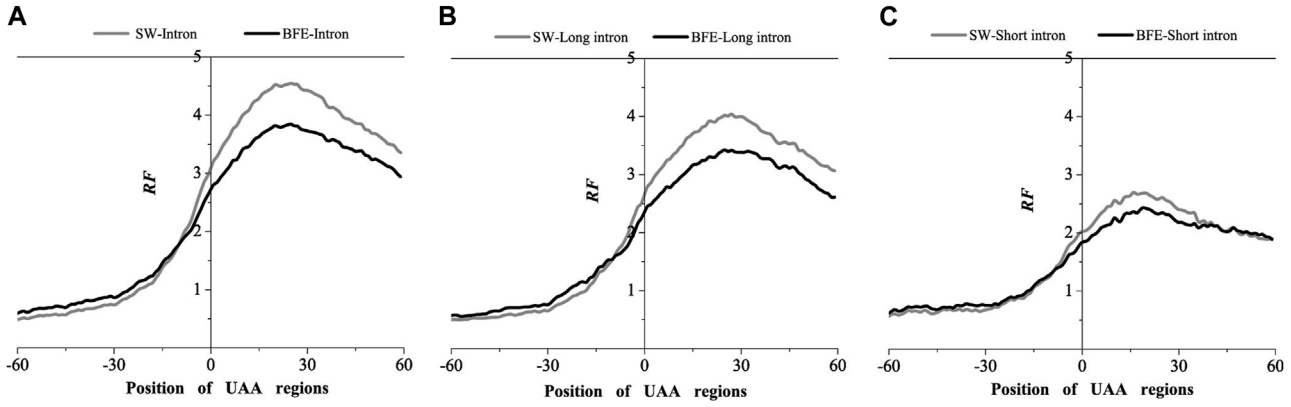

FIGURE 5
Comparations of Relative Frequency (RF) distributions between SW method and BFE method on UAA regions. **(A)** The total introns, **(B)** the long introns and **(C)** the short introns.

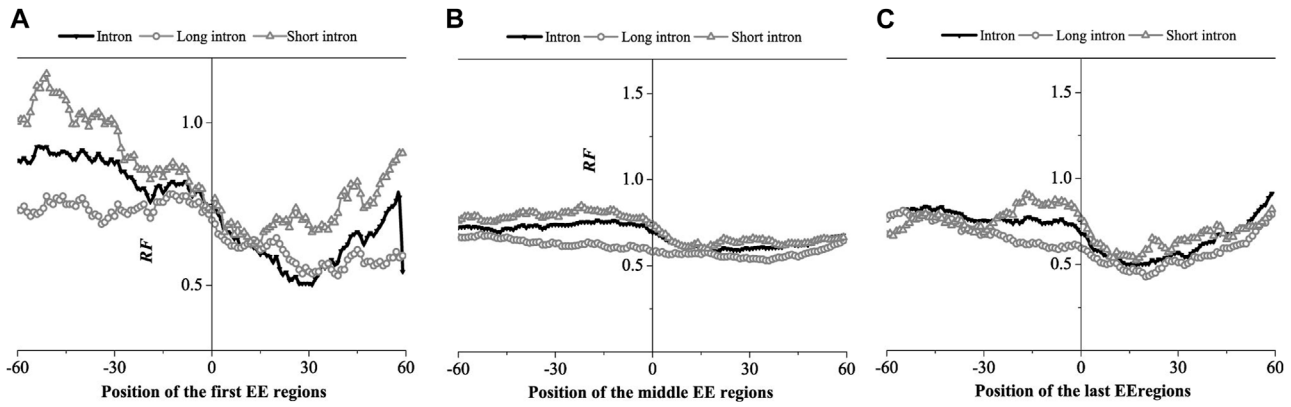

FIGURE 6
Relative Frequency (RF) distributions on the exon-exon junction (EE) regions of the mRNA. **(A)** The first EE regions, **(B)** the middle EE regions, **(C)** the last EE regions. The RF distributions related long introns and short introns are also presented in the figure.

long introns in all three EE regions. Based on the findings of previous studies, we hypothesized that the region with low relative matched frequency might be the protein factor binding region.

We next compared the matched frequency distribution characteristics of the exon-exon junction regions of the mRNA group based on between the improved Smith-Waterman local alignment method and the binding free energy weighted local alignment method (BFE method). The mRNA group based on the improved Smith-Waterman local alignment method was used as the control group. The distribution of the optimal matched frequencies of the whole intron, long intron, and short intron groups on exon junction regions on mRNA based on the binding free energy weighted local alignment method was compared with that of the SW method group. The results were showed in Figures 7–9.

The optimal matched frequency distribution trend of the OMF in the junction region on the mRNA sequence (which is of the corresponding mRNA sequences and the intron sequences) based on the binding free energy weighted local alignment method was comparable to the SW method. In the exon-exon junction regions of the first, last and intermediate exons, although the weighted matched frequency distribution of the whole intron, long intron and short intron groups were slightly higher than the SW method (Figures 7–9), there was no significant difference between them.

These results indicated that the distribution of the matched frequency of exon junction regions obtained by SW method and BFE method is conservative. The matched frequency values of the exon-exon junction regions obtained by the BFE method were larger than those obtained by the SW method, which was caused by the tendency of the binding free energy weighted local alignment algorithm to combine the optimal matched fragment with CDS with high G + C content. The binding preference of intron sequence (especially short introns) and exon connection sites upstream regions suggests a preferred interaction between the intron

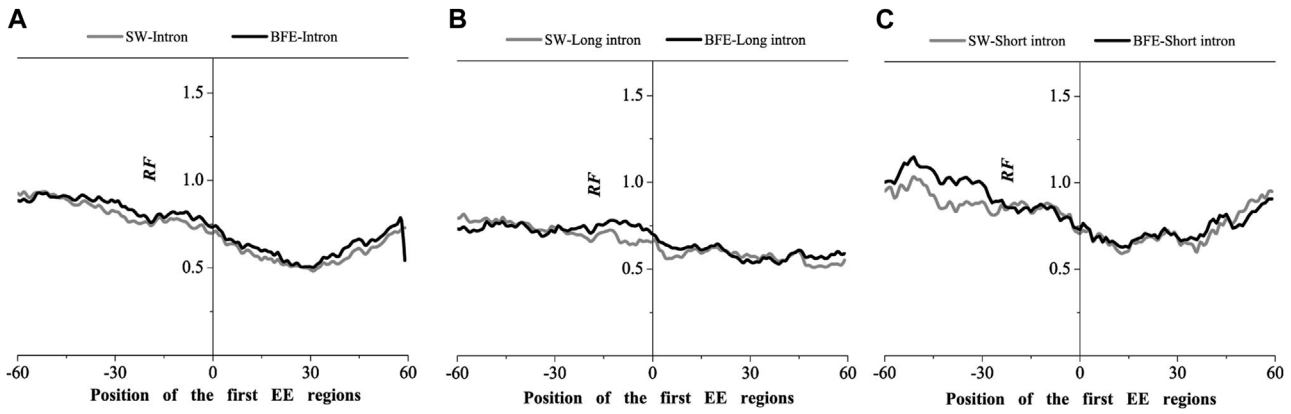

FIGURE 7
Comparations of Relative Frequency (RF) distributions between SW method and BFE method on the first EE regions. **(A)** the total introns, **(B)** the long introns,**(C)** the short introns.

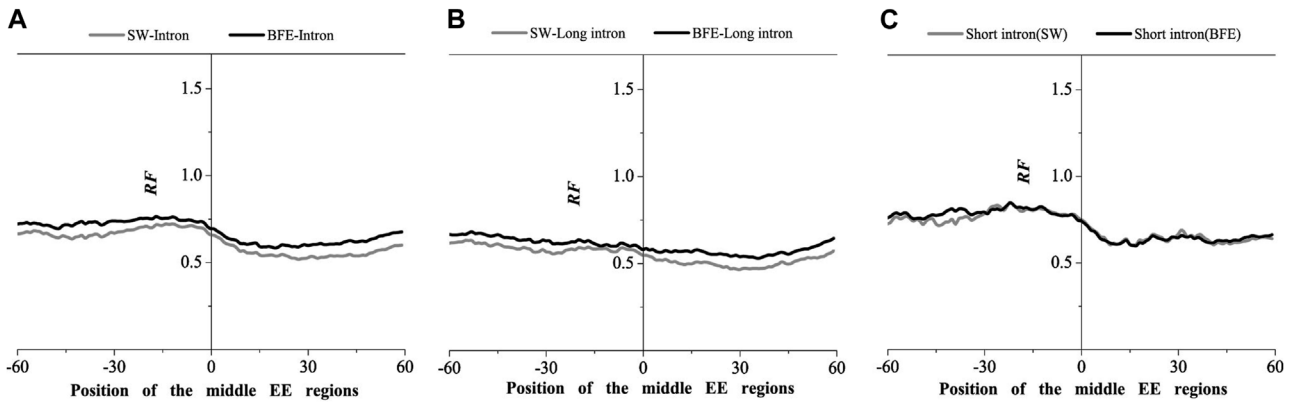

FIGURE 8
Comparations of Relative Frequency (RF) distributions between SW method and BFE method on the middle EE regions. **(A)** the total introns, **(B)** the long introns,**(C)** the short introns.

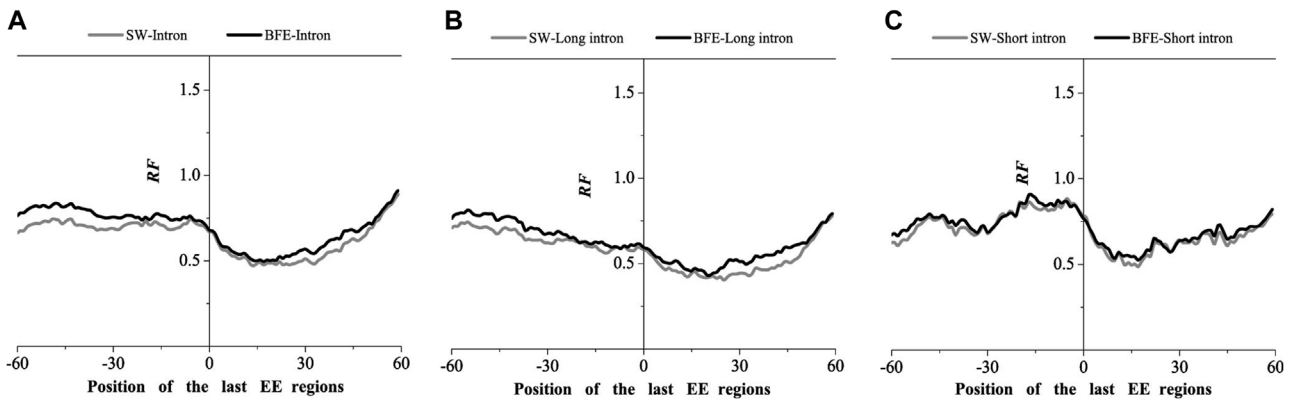

FIGURE 9
Comparations of Relative Frequency (RF) distributions between SW method and BFE method on the last EE regions. **(A)** the total introns, **(B)** the long introns, **(C)** the short introns.

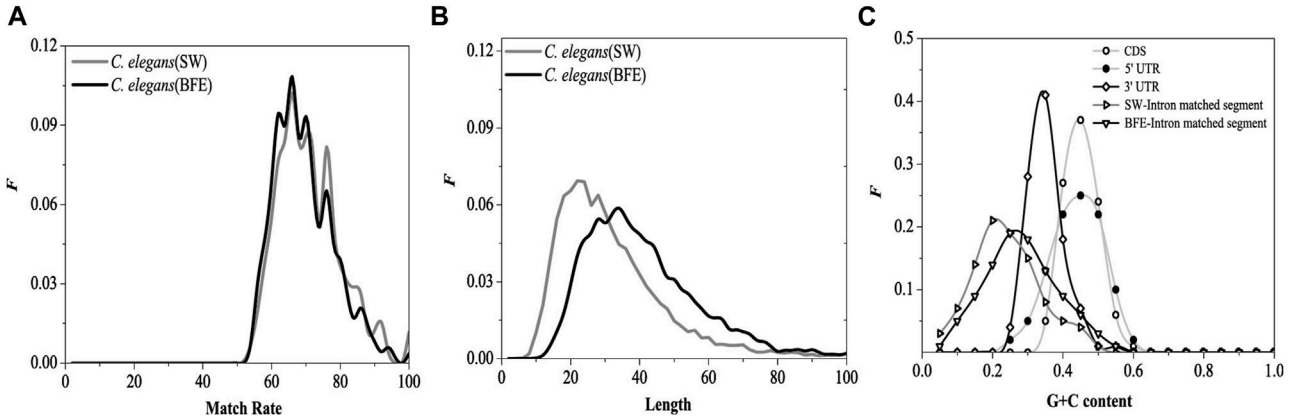

FIGURE 10
Distributions of the match rates **(A)**, the length **(B)** of optimal matched segments. SW represents the result came from SW method and BFE represents the result came from BFE method G + C content distributions **(C)** for CDS, 5, UTR, 3, UTR, optimal matched segments (SW method) and optimal matched segments (BFE method).

sequence and the exon-exon junction region of the mRNA sequence. Besides, the process of short introns is more advantageous, which may be attributed to the fact that the biological function of short introns is mainly related to mRNA splicing or alternative splicing. These interesting results are worth thinking about.

3.3 Sequence characteristics of the optimal matched fragments

We calculated four sequence features of the optimal matched fragment pairs based on the BFE method, including the match rate distribution, length distribution, G + C content distribution and base association (D2 value). The results were compared with those obtained by the SW method.

3.3.1 The distribution of match rate and length

The match rate distribution of the optimal matched fragment of intron obtained by the BFE method is shown in Figure 10A. The distribution of the match rate of the optimal matched fragment obtained by the two methods was very similar, except that the distribution curves have relatively small fluctuations. The length distribution of the optimal matched fragment of intron obtained by the BFE method is shown in Figure 10B.

The functional fragments representing the interaction between introns and mRNA are a class of functional fragments similar to miRNA, and their match rate and the most length should be similar to miRNA fragments. The length of functional segments of siRNA was very conserved, ranging from 21 to 23 bp, while that of miRNA ranged from 18 to 25 bp. The most length of the optimal matching fragment by SW method was 23bp, and its characteristics were similar to miRNA fragments. However, the biologically roles of the interaction between introns and mRNA should be differ from the biologically roles of siRNA and miRNA, we believe that the biologically roles of the interaction between introns and mRNA should be protected mRNA from degradation and be beneficial to transport of mRNA from nucleus to cytoplasm. The interaction strength of between introns and mRNA should be weaker than siRNA and miRNA, and the lengths of the optimal matched segments (OMS) should be longer than siRNA and miRNA. Our results show that the maximum length obtained by BFE method is 36bp, which is quite different from miRNA fragment, and the mated rate obtained by BFE method is lower than SW method and siRNA and miRNA. So, the results by BFE method may have a biological significance.

TABLE 1 D 2 values of different sequences in *Caenorhabditis elegans* protein-coding genes.

	mRNA			Intron	
	CDS	5, UTR	3, UTR	OMS (SW)	OMS (BFE)
D 2	0.029	0.032	0.036	0.066	0.053

Note: OMS indicates The optimal matched fragment of the introns.

3.3.2 G + C content and D2 value

The G + C content distribution of the optimal matched fragment on the introns by the BFE method is shown in Figure 10C. The distribution range of G + C content in the optimal matched fragment of the BFE method was consistent with the SW method, but the peak region of G + C content was about 0.25, which moved toward high G + C content, it increased 0.05 compared with the SW method. The reason for the general increase in G + C content is caused by the fact that was the preference for intron fragments with high G + C content during selecting the optimal matched fragments by the binding free energy weighted local alignment method.

Their D2 values are calculated by formula (6), and the results are shown in Table 1. It can be found that the D2 value of the optimal matched fragment was significantly higher than CDS, 5 and 3'UTR sequences, it suggested the base association in the OMF was significantly stronger than the other three types of sequences, with a strong sequence structure. Besides, the D2 value of the optimal matched fragment by the BFE method was about 20% lower than the SW method, indicating that the former method can document the interaction between the intron and mRNA sequences and characterize their interaction.

4 Conclusion

In the present study, the binding free energy weighted local alignment algorithm method was used to obtain the optimal matched fragment between the post-spliced intron and its corresponding mRNA sequence, and the relative matched frequency distribution on the mRNA and near the functional site. Our results showed that the relative matched frequency distribution obtained by the BFE method was similar to the SW method; there were the region of preference at the UTR region at both ends of the mRNA sequence was identified as a favorable region, especially in the 3'UTR region. However, the suggestion of the combination show that was more in favor of the optimal matched fragments with CDS with high G + C content, which was the weaker interaction in the 5'and 3'UTR regions, and higher in the middle CDS region than the SW method, when the BFE method was applied.

Moreover, we found that the region of preference of theshort introns in the 5'UTR region, and the long introns in the 3'UTR region, which consistent with the SW method. Besides, the relative matched frequency distribution in the exon connection region was similar to the SW method. The interaction intensity of the upstream connection point was greater than that of the downstream, and there was a minimal relative matching frequency distribution of the downstream of the first and last exon connection region, and the interaction of short introns was stronger than long intron sequences.

The match rate distribution and the length distribution shape of The optimal matched fragment were similar to the SW method, although an increase in optimal matching fragment length was observed. When the SW method was applied, the maximum value length was 23bp, and an increase to 36bp was observed with the BFE method. It was still broad (0.05–0.5) with the distribution range of the content of G + C of the optimal matched fragment, but the maximum value of the content of G + C by the SW and BFE methods was 0.2 and 0.25, respectively, which display the content of G + C by the BFE method was generally higher. Although the base correlation of the optimal matched fragment remained strong, it was slightly lower than the D2 value in the SW method. These results substantiate that the optimal matched fragment is a special sequence fragment with a highly structured organization.

Overall, the BFE method and SW method yielded similar results. However, it was the less intensity of the interaction between introns and corresponding mRNA by the BFE method, the length of the optimal matched fragements was longer, and the bases association or sequence structure of the OMF was relatively weaker. Compared with SW, the BFE method is more sensitive than the SW method for representations the RNA-RNA interaction and can avoid the false positives which may occur in SW method.

In conclusion, the BFE method and SW method yielded similar results, the results obtained by the BFE method and SW method were basically the same, indicated that the binding free energy weighted local alignment method can be used to predict the interaction between introns and their corresponding mRNAs. According to the comparison of the matched frequency distribution between introns and corresponding mRNA sequences, the BFE method was more conducive to predict the weak interaction between sequences with high G + C content. The sequence characteristics of the optimal matched fragments obtained by the BFE method implyed that the structures of sequence with longer length, higher G + C content and looser sequence structure are more likely to predict weak interactions between sequences with higher GC content, compared with those calculated by the SW method.

We advocate that using local base matching to characterize the interaction between introns and mRNAs has huge prospects.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Materials, further inquiries can be directed to the corresponding authors.

Author contributions

SB and QS jointly completed the algorithm optimization and paper writing, XZ and ZXL established the theoretical model, ZYL analyzed the theoretical model, PN and YW collected, sorted and refined the data, YL analyzed the sequence characteristics, and HW completed the data summary and results collation. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by grants from the Natural Science Foundation of Inner Mongolia Autonomous Region (2021MS03063); High-end talent training projects of Grassland Talents in Inner Mongolia Autonomous Region, National Natural Science Foundation of China (31500677); The Leading Talent Project of Science and Technology Leading Talent Team Project of Inner Mongolia Autonomous Region (2022LJRC0010).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1151172/full#supplementary-material

References

Akaike, Y., Kurokawa, K., kajita, K., Kuwano, Y., Masuda, K., Nishida, K., et al. (2011). Skipping of an alternative intron in the srsf1 3' untranslated region increases transcript stability. *J. Med. investigation* 58, 180–187. doi:10.2152/jmi.58.180

Alexander, M. R., Wheatley, A. K., Center, R. J., and Purcell, D. F. J. (2010). Efficient transcription through an intron requires the binding of an Sm-type U1 snRNP with intact stem loop II to the splice donor. *Nucleic Acids Res.* 38 (9), 3041–3053. doi:10.1093/nar/gkp1224

Anderson, A. C. (2003). The process of structure-based drug design. *Chem. Biol.* 10 (9), 787–797. doi:10.1016/j.chembiol.2003.09.002

Awad, A. M., Venkataramanan, S., Nag, A., Galivanche, A. R., Bradley, M. C., Neves, L. T., et al. (2017). Chromatin-remodeling SWI/SNF complex regulates coenzyme Q6 synthesis and a metabolic shift to respiration in yeast. *J. Biol. Chem.* 292 (36), 14851–14866. doi:10.1074/jbc.M117.798397

Bower, M. J., Cohen, F. E., and Dunbrack, R. L. (1997). Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *J. Mol. Biol.* 267 (5), 1268–1282. doi:10.1006/jmbi.1997.0926

Braddock, M., Muckenthaler, M., White, M. R., Thorburn, A. M., Sommerville, J., Kingsman, A. J., et al. (1994). Intron-less RNA injected into the nucleus of Xenopus oocytes accesses a regulated translation control pathway. *Nucleic Acids Res.* 22 (24), 5255–5264. doi:10.1093/nar/22.24.5255

Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V., and Kondrashov, F. A. (2002). Selection for short introns in highly expressed genes. *Nat. Genet.* 31, 415–418. doi:10.1038/ng940

Combs, D. J., Nagel, R. J., Ares, M., and Stevens, S. W. (2006). Prp43p is a DEAH-box spliceosome disassembly factor essential for ribosome biogenesis. *Mol. Cell Biol.* 26 (2), 523–534. doi:10.1128/MCB.26.2.523-534.2006

Comeron, J. M. (2001). What controls the length of noncoding DNA? *Curr. Opin. Genet. Dev.* 11 (6), 652–659. doi:10.1016/s0959-437x(00)00249-5

Deigan, K. E., Tian, W., Mathews, D. H., and Weeks, K. M. (2009). Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U. S. A.* 106 (1), 97–102. doi:10.1073/pnas.0806929106

Duret, L. (2001). Why do genes have introns? Recombination might add a new piece to the puzzle. *Trends Genet.* 17 (4), 172–175. doi:10.1016/s0168-9525(01)02236-3

Elmonir, W., Inoshima, Y., Elbassiouny, A., and Ishiguro, N. (2010). Intron 1 mediated regulation of bovine prion protein gene expression: Role of donor splicing sites, sequences with potential enhancer and suppressor activities. *Biochem. Biophysical Res. Commun.* 397 (4), 706–710. doi:10.1016/j.bbrc.2010.06.014

Faraggi, E., Yang, Y., Zhang, S., and Zhou, Y. (2009). Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. *Struct. Lond. Engl. 1993)* 17 (11), 1515–1527. doi:10.1016/j.str.2009.09.006

Fawcett, J. A., Rouze, P., and Van de Peer, Y. (2011). Higher intron loss rate in *Arabidopsis thaliana* than A. Lyrata is consistent with stronger selection for a smaller genome. *Mol. Biol. Evol.* 29 (2), 849–859. doi:10.1093/molbev/msr254

Fedorova, L., and Fedorov, A. (2003). Introns in gene evolution. *Genetics* 118 (2-3), 123–131. doi:10.1023/a:1024145407467

Gabriel, M., Pierre, N., Keightley, P. D., and Charlesworth, B. (2005). Intron size and exon evolution in Drosophila. *Genetics* 170 (1), 481–485. doi:10.1534/genetics.104.037333

Gao, G., Williams, J. G., and Campbell, S. L. (2004). Protein-protein interaction analysis by nuclear magnetic resonance spectroscopy. *Methods Mol. Biol.* 261, 79–92. doi:10.1385/1-59259-762-9:079

Gatfield, D., Le Hir, H., Schmitt, C., Braun, I. C., Kocher, T., Wilm, M., et al. (2001). The DExH/D box protein HEL/UAP56 is essential for mRNA nuclear export in Drosophila. *Curr. Biol.* 11 (21), 1716–1721. doi:10.1016/s0960-9822(01)00532-2

Gazave, E., Marqués-Bonet, T., Fernando, O., Charlesworth, B., and Navarro, A. (2007). Patterns and rates of intron divergence between humans and chimpanzees. *Genome Biol.* 8 (2), R21. doi:10.1186/gb-2007-8-2-r21

Halligan, D. L., and Keightley, P. D. (2006). Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Res.* 16 (7), 875–884. doi:10.1101/gr.5022906

Hay, S., and Scrutton, N. S. (2012). Good vibrations in enzyme-catalysed reactions. *Nat. Chem.* 4 (3), 161–168. doi:10.1038/nchem.1223

He, Y., Wu, Y., Lan, Z., Liu, Y., and Zhang, Y. (2010). Molecular analysis of the first intron in the bovine myostatin gene. *Mol. Biol. Rep.* 38 (7), 4643–4649. doi:10.1007/s11033-010-0598-9

Jackson, S. E. (1998). How do small single-domain proteins fold? *Fold. Des.* 3 (4), 81–91. doi:10.1016/S1359-0278(98)00033-9

Jeffares, D. C., Mourier, T., and Penny, D. (2006). The biology of intron gain and loss. *Trends Genet.* 22 (1), 16–22. doi:10.1016/j.tig.2005.10.006

Kim, V. N., Yong, J., Kataoka, N., Abel, L., Diem, M. D., and Dreyfuss, G. (2001). The Y14 protein communicates to the cytoplasm the position of exon–exon junctions. *EMBO J.* 20 (8), 2062–2068. doi:10.1093/emboj/20.8.2062

Komarnitsky, P., Cho, E. J., and Buratowski, S. (2000). Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes & Dev.* 14 (19), 2452–2460. doi:10.1101/gad.824700

Landen, G., Roy Scott, W., Thornlow, B., Kramer, A., Ares, M., Jr, and Corbett-Detig, R. (2022). Transposable elements drive intron gain in diverse eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 119 (48), e2209766119. doi:10.1073/pnas.2209766119

Le Hir, H., Moore, M. J., and Maquat, L. E. (2000). Pre-mRNA splicing alters mRNP composition: Evidence for stable association of proteins at exon–exon junctions. *Genes & Dev.* 14 (9), 1098–1108. doi:10.1101/gad.14.9.1098

Le Hir, H., Nott, A., and Moore, M. J. (2003). How introns influence and enhance eukaryotic gene expression. *Trends Biochem. Sci.* 28 (4), 215–220. doi:10.1016/S0968-0004(03)00052-5

Lewis, J. D., and Izaurflde, E. (2004). The role of the cap structure in RNA processing and nuclear export. *Eur. J. Biochem.* 247 (2), 461–469. doi:10.1111/j.1432-1033.1997.00461.x

Li, L., and Pintel, D. J. (2012). Splicing of goose parvovirus pre-mRNA influences cytoplasmic translation of the processed mRNA. *Virology* 426 (1), 60–65. doi:10.1016/j.virol.2012.01.019

Lykke-Andersen, J., Shu, M-D., and Steitz, J. A. (2001). Communication of the position of exon-exon junctions to the mRNA surveillance machinery by the protein RNPS1. *Sci. Signal.* 293 (5536), 1836–1839. doi:10.1126/science.1062786

Manke, T., Bringas, R., and Vingron, M. (2003). Correlating protein-DNA and protein-protein interaction networks. *J. Mol. Biol.* 333 (1), 75–85. doi:10.1016/j.jmb.2003.08.004

Maquat, L. E., and Carmichael, G. G. (2001). Quality control of mRNA function. *Cell* 104 (2), 173–176. doi:10.1016/s0092-8674(01)00202-1

Mariati, Ho, S. C. L., Yap, M. G. S., and Yang, Y. (2010). Evaluating post-transcriptional regulatory elements for enhancing transient gene expression levels in CHO K1 and HEK293 cells. *Protein Expr. Purif.* 69 (1), 9–15. doi:10.1016/j.pep.2009.08.010

Martin, I. V., and MacNeill, S. A. (2004). Functional analysis of subcellular localization and protein-protein interaction sequences in the essential DNA ligase I protein of fission yeast. *Nucleic Acids Res.* 32 (2), 632–642. doi:10.1093/nar/gkh199

Mattick, J. S., and Gagen, M. J. (2001). The evolution of controlled multitasked gene networks: The role of introns and other noncoding RNAs in the development of complex organisms. *Mol. Biol. Evol.* 18 (9), 1611–1630. doi:10.1093/oxfordjournals.molbev.a003951

McKenzie, R. W., and Brennan, M. D. (1996). The two small introns of the Drosophila affinidisjuncta Adh gene are required for normal transcription. *Nucleic Acids Res.* 24 (18), 3635–3642. doi:10.1093/nar/24.18.3635

Mitomo, D., Fukunishi, Y., Higo, J., and Nakamura, H. (2009). Calculation of protein-ligand binding free energy using smooth reaction path generation (SRPG) method: A comparison of the explicit water model, gb/sa model and docking score function. *Genome Inf. Int. Conf. Genome Inf.* 23 (1), 85–97. doi:10.1142/9781848165632_0008

Morgan, J. T., Fink, G. R., and Bartel, D. P. (2019). Excised linear introns regulate growth in yeast. *Nature* 565 (7741), 606–611. doi:10.1038/s41586-018-0828-1

Munding, E. M., Shiue, L., Katzman, S., Donohue, J. P., and Ares, M., Jr (2013). Competition between pre-mRNAs for the splicing machinery drives global regulation of splicing. *Mol. Cell* 51 (3), 338–348. doi:10.1016/j.molcel.2013.06.012

Nguyen, H. D., Yoshihama, M., and Kenmochi, N. (2006). Phase distribution of spliceosomal introns: Implications for intron origin. *BMC Evol. Biol.* 6 (1), 69. doi:10.1186/1471-2148-6-69

Nott, A., Meislin, S. H., and Moore, M. J. (2003). A quantitative analysis of intron effects on mammalian gene expression. *RNA* 9 (5), 607–617. doi:10.1261/rna.5250403

Parenteau, J., Durand, M., Veronneau, S., Lacombe, A. A., Morin, G., Guerin, V., et al. (2008). Deletion of many yeast introns reveals a minority of genes that require splicing for function. *Mol. Cell Biol.* 19 (5), 1932–1941. doi:10.1091/mbc.e07-12-1254

Parenteau, J., Maignon, L., Berthoumieux, M., Catala, M., Gagnon, V., and Abou Elela, S. (2019). Introns are mediators of cell response to starvation. *Nature* 565 (7741), 612–617. doi:10.1038/s41586-018-0859-7

Petrov, D. A. (2002). DNA loss and evolution of genome size in Drosophila. *Genetica* 115 (1), 81–91. doi:10.1023/a:1016076215168

Prathipati, P., Dixit, A., and Saxena, A. K. (2007). Computer-Aided Drug Design: Integration of structure-based and ligand-based approaches in drug design. *Curr. Comput. - Aided Drug Des.* 3 (2), 133–148. doi:10.2174/157340907780809516

Rafiq, M., Suen, C. K., Choudhury, N., Joannou, C. L., White, K. N., and Evans, R. W. (1997). Expression of recombinant human ceruloplasmin–an absolute requirement for splicing signals in the expression cassette. *FEBS Lett.* 407 (2), 132–136. doi:10.1016/s0014-5793(97)00325-6

Rocchi, V., Janni, M., Bellincampi, D., Giardina, T., and D'Ovidio, R. (2012). Intron retention regulates the expression of pectin methyl esterase inhibitor (Pmei) genes during wheat growth and development. *Plant Biol.* 14 (2), 365–373. doi:10.1111/j.1438-8677.2011.00508.x

Roy, S. W., Fedorov, A., and Gilbert, W. (2003). Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl. Acad. Sci. U. S. A.* 100 (12), 7158–7162. doi:10.1073/pnas.1232297100

Roy, S. W., and Hartl, D. L. (2006). Very little intron loss/gain in plasmodium: Intron loss/gain mutation rates and intron number. *Genome Res.* 16 (6), 750–756. doi:10.1101/gr.4845406

Ryu, W. S., and Mertz, J. E. (1989). Simian virus 40 late transcripts lacking excisable intervening sequences are defective in both stability in the nucleus and transport to the cytoplasm. *J. virology* 63 (10), 4386–4394. doi:10.1128/JVI.63.10.4386-4394.1989

Schaefer, M., Bartels, C., and Karplus, M. (1998). Solution conformations and thermodynamics of structured peptides: Molecular dynamics simulation with an implicit solvation model. *J. Mol. Biol.* 284 (3), 835–848. doi:10.1006/jmbi.1998.2172

Selkoe, D. J. (2003). Folding proteins in fatal ways. *Nature* 426 (6968), 900–904. doi:10.1038/nature02264

The ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447 (7146), 799–816. doi:10.1038/nature05874

Thomas, A., Cannings, R., Monk, N. A. M., and Cannings, C. (2003). On the structure of protein-protein interaction networks. *Biochem. Soc. Trans.* 31 (6), 1491–1496. doi:10.1042/bst0311491

Tollenaere, J. P. (1996). The role of structure-based ligand design and molecular modelling in drug discovery. *Pharm. world & Sci. PWS* 18 (2), 56–62. doi:10.1007/BF00579706

Torrado, M., Iglesias, R., Nespereira, B., Centeno, A., Lopez, E., and Mikhailov, A. T. (2009). Intron retention generates ANKRD1 splice variants that are co-regulated with the main transcript in normal and failing myocardium. *Gene* 440 (1-2), 28–41. doi:10.1016/j.gene.2009.03.017

Venkataramanan, S., Douglass, S., Galivanche, A. R., and Johnson, T. L. (2017). The chromatin remodeling complex Swi/Snf regulates splicing of meiotic transcripts in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 45 (13), 7708–7721. doi:10.1093/nar/gkx373

Wan, R., Yan, C., Bai, R., Lei, J., and Shi, Y. (2017). Structure of an intron lariat spliceosome from *Saccharomyces cerevisiae*. *Cell* 171 (1), 120–132.e12. doi:10.1016/j.cell.2017.08.029

Wanichthanarak, K., Wongtosrad, N., and Petranovic, D. (2015). Genome-wide expression analyses of the stationary phase model of ageing in yeast. *Mech. Ageing Dev.* 149, 65–74. doi:10.1016/j.mad.2015.05.008

Woo, H. J. (2008). Calculation of absolute protein-ligand binding constants with the molecular dynamics free energy perturbation method. *Methods Mol. Biol. Clift. NJ)* 443, 109–120. doi:10.1007/978-1-59745-177-2_6

Woo, H. J., and Roux, B. (2005). Calculation of absolute protein-ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci. U. S. A.* 102 (19), 6825–6830. doi:10.1073/pnas.0409005102

Zhang, Y., and Skolnick, J. (2005). The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl. Acad. Sci. U. S. A.* 102 (4), 1029–1034. doi:10.1073/pnas.0407152101

Zhang, Z. D., Paccanaro, A., Fu, Y., Weissman, S., Weng, Z., Chang, J., et al. (2007). Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. *Genome Res.* 17 (6), 787–797. doi:10.1101/gr.5573107

Zuker, M., and Sankoff, D. (1984). RNA secondary structures and their prediction. *Bull. Math. Biol.* 46 (4), 591–621. doi:10.1016/s0092-8240(84)80062-2

Appendix A: Supplementary data to this article

1. Instruction about the database. txt (Taking an example).
2. Supporting Information S1-for the optimal matched regions located at UTR. txt.
3. Supporting Information S2-for CC-Random group. txt.
4. Supporting Information S3-for mRNA group. txt.

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 02 March 2023
DOI 10.3389/fmicb.2023.1141227

OPEN ACCESS

EDITED BY
Yongchun Zuo,
Inner Mongolia University,
China

REVIEWED BY
Ran Su,
Tianjin University,
China
Cangzhi Jia,
Dalian Maritime University,
China

*CORRESPONDENCE
Yuming Zhao
✉ zym@nefu.edu.cn
Wen Yang
✉ 13159850336@163.com

SPECIALTY SECTION
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 10 January 2023
ACCEPTED 10 February 2023
PUBLISHED 02 March 2023

CITATION
Li H, Zhang J, Zhao Y and Yang W (2023) Predicting *Corynebacterium glutamicum* promoters based on novel feature descriptor and feature selection technique. *Front. Microbiol.* 14:1141227. doi: 10.3389/fmicb.2023.1141227

Predicting *Corynebacterium glutamicum* promoters based on novel feature descriptor and feature selection technique

HongFei Li[1,2], Jingyu Zhang[3], Yuming Zhao[1,2]* and Wen Yang[4]*

[1]College of Life Science, Northeast Forestry University, Harbin, China, [2]College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, [3]Department of Neurology, The Fourth Affiliated Hospital of Harbin Medical University, Harbin, China, [4]International Medical Center, Shenzhen University General Hospital, Shenzhen, China

The promoter is an important noncoding DNA regulatory element, which combines with RNA polymerase to activate the expression of downstream genes. In industry, artificial arginine is mainly synthesized by *Corynebacterium glutamicum*. Replication of specific promoter regions can increase arginine production. Therefore, it is necessary to accurately locate the promoter in *C. glutamicum*. In the wet experiment, promoter identification depends on sigma factors and DNA splicing technology, this is a laborious job. To quickly and conveniently identify the promoters in *C. glutamicum*, we have developed a method based on novel feature representation and feature selection to complete this task, describing the DNA sequences through statistical parameters of multiple physicochemical properties, filtering redundant features by combining analysis of variance and hierarchical clustering, the prediction accuracy of the which is as high as 91.6%, the sensitivity of 91.9% can effectively identify promoters, and the specificity of 91.2% can accurately identify non-promoters. In addition, our model can correctly identify 181 promoters and 174 non-promoters among 400 independent samples, which proves that the developed prediction model has excellent robustness.

KEYWORDS

promoter, *Corynebacterium glutamicum*, physicochemical properties, analysis of variance, hierarchical clustering, feature selection, random forest

1. Introduction

Corynebacterium glutamicum is a prokaryote, which was first discovered in the 1950s (Sano, 2009). It is mainly responsible for the production of L-glutamic acid and has played a huge potential in the production of amino acids in the industrial field. *C. glutamicum* is considered the best bio-manufacturing substrates by many countries because it can produce amino acids with few nutrients and sufficient capacity (Sun et al., 2011; Vertes et al., 2012). Considering the excellent characteristics of *C. glutamicum*, the genome has been modified to produce a variety of amino acids, organic acids, alcohols, and proteins through biological genetic technology (Okino et al., 2008; Hu et al., 2013). At the beginning of the 20th century, *C. glutamicum* first was published its complete genome sequence, named *C. glutamicum* ATCC 13032. The whole genome consists of a circular chromatin with a length of 3282708 bp, containing 3000 coding protein genes, and the 'C + G' content is 53.8% (Kalinowski et al., 2003). The complete genome sequencing of this species provides convenient conditions for gene editing and regulatory

analysis that can further improve the efficiency of *C. glutamicum* to produce amino acids (Barrangou and Horvath, 2017; Cho et al., 2017; Jiang et al., 2017; Huang et al., 2019). The above biotechnology mainly involves the knockout and inactivation of specific genes, and the key is to locate the starting site of genes and the promoter region of the target gene (Okino et al., 2008; Theron and Reid, 2011; Silar et al., 2016). In Hebert et al. (2018) and Shang et al. (2018) designed a special promoter, which improved the expression level of sucCD and the production of L-lysine. Thus, it is very important to identify and locate the promoter of *C. glutamicum*.

The promoter, as a pivotal regulatory element, is responsible for activating the expression of target genes (Canzio et al., 2019; Xiao et al., 2019; Jeon and Tucker-Kellogg, 2020). In preparation for gene expression, promoters are affected by macromolecular complexes that are produced by the combination of specific transcription factors and regulatory factors to complete the transcription from DNA to RNA (La Fleur et al., 2022; Liu et al., 2022; Rengachari et al., 2022). In industrial systems, the recognition of promoters of *C. glutamicum* requires the help of Sigma factors, which requires the support of gene isolation, polymerase chain reaction (PCR), and gene cloning techniques (Blumenstein et al., 2022; Stepanek et al., 2022). Although the wet lab methods described above can specifically identify promoters, they are time - and labor-consuming, and it is essential to develop a method-based calculating model to rapidly identify promoters. At present, models of promoter recognition already exist for many species (Silar et al., 2016; Bharanikumar et al., 2018; Leemans et al., 2019), but cannot be applied to Corynebacterium because of the large differences in homology between the species. Moreover, these models employed features that do not accurately describe the inherent properties of DNA sequences, resulting in poor overall prediction performance. For example, in the human promoter recognition task, Li et al. (2022b) used five feature descriptors to express DNA sequences, but the final prediction accuracy was only 80%. Hence, it is necessary to design a mathematical prediction model to accurately identify the promoter of *C. glutamicum* for the industrial production of amino acids.

Here, we have collected promoter sequences that have been verified and annotated by experiments (Su et al., 2021), and designed a new feature expression method according to the distribution of multiple physical and chemical properties of sequence DNA. In addition, we have developed a novel feature selection method for redundant information between features. The proposed model has strong robustness by independent set verification.

2. Materials and methods

The following three conditions are indispensable to the excellent properties of the prediction model. First, building a rigorous and proven dataset. Second, designing the corresponding feature descriptor according to the inherent attributes of the sample and the specific distribution. Finally, selecting the machine learning algorithm that conforms to the regular pattern of descriptors. The flow of the whole method is drawn in Figure 1.

2.1. Benchmark dataset

To build a reasonable and interpretable dataset, the promoter of *C. glutamicum* selected comes from the PPD database that collected promoters of 63 eukaryotes, including 129,148 promoter sequences, each of which was confirmed by strict experiments (Su et al., 2021). Therefore, we take 3,581 promoters of *C. glutamicum* ATCC 13032 in the dataset as positive samples. Initially, we filter promoters with incomplete annotation information and the same starting site. Immediately, CD-HIT software was employed to reduce the sequence consisting of the filtered promoters to less than 0.6 (Li and Godzik, 2006; Huang et al., 2010). Finally, we obtained 1,000 promoter sequences with a length of 81 bp. For the selection of negative sample non-promoters, we downloaded the complete genome data from the GenBank database[1], and randomly cut 81 bp from different gene fragments as the original negative samples to enhance the diversity of the sequence. Similarly, the CD-HIT was applied to reduce its sequence consistency to 60%, then we reserved 1,000 non-promoter sequences as negative samples. Aiming to prove the robustness of the model, 2000 samples are randomly divided into the training set and independent set according to the ratio of 8: 2, 800 positive samples and 800 negative samples were used for model fitting and training by five-fold cross-validation, and the remaining 200 positive samples and negative samples are employed to test the model's ability to recognize the unlabeled sample.

2.2. Feature descriptor

The key step in building a model is to accurately describe the inherent attributes and reflect the differences between samples. The combination of promoters with various regulatory elements is inseparable from the physicochemical properties of their bases, such as hydrophilicity and hydrophobicity. Therefore, we design a novel digital feature containing a variety of physical and chemical properties to describe the DNA sequence. First, we found the 90 physical and chemical properties of dinucleotides from published literature. Furthermore, we analyzed the distribution of these physicochemical properties of 16 dinucleotides (Dao et al., 2019). It can be found from Figure 2 that the distribution of 16 kinds of dinucleotides is more remarkable. The minimum value of dinucleotide 'CG' is obtained, while the maximum value of 'TA' is obtained. The ordinate of the violin chart corresponds to the frequency density of data distribution. For example, the distribution of 'GA', 'CT', and 'TC' shows a standard normal distribution, but their wave peaks and widths are different, so they have different mean values and variances. In addition, the area occupied by different dinucleotides also varies greatly, which infers the sum is diverse. Hence, we use the minimum, maximum, variance, mean, and sum of 90 physical and chemical properties to represent the overall physical and chemical property level of 16 dinucleotides, the 90 dimensional physical and chemical properties are replaced by 5

Abbreviations: SVM, support vector machine; RF, random forest; MLP, multi-layer perceptron; KNN, k-nearest neighbors; Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, matthew correlation coefficient; ROC, receiver operating characteristic; AUC, area under receiver operating characteristic (ROC) curve; ANOVA, analysis of various; HC, hierarchical clustering.

1 https://www.ncbi.nlm.nih.gov/nuccore/NC_006958.1

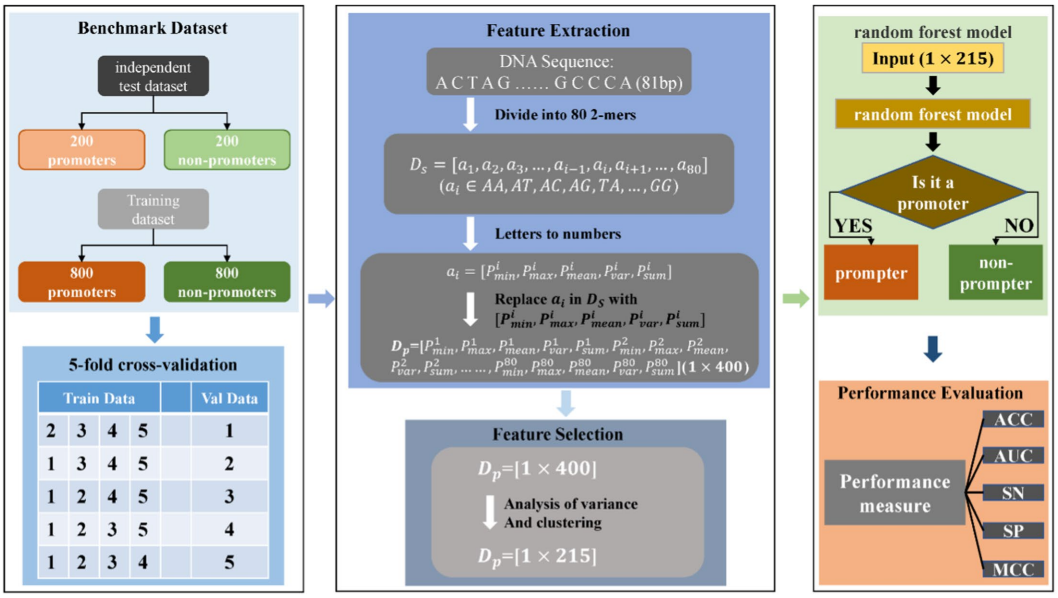

FIGURE 1
The workflow of *Corynebacterium glutamicum* promoter prediction model.

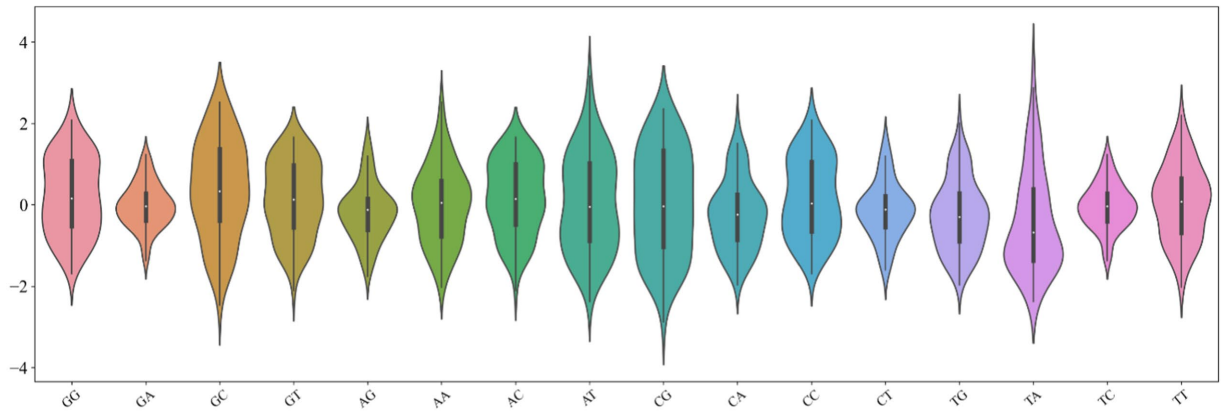

FIGURE 2
Violin chart of physical and chemical properties of 16 dinucleotides.

statistical parameters. The method can not only describe the distribution characteristics of dinucleotides but also greatly reduce the dimensions used to describe the descriptor. Suppose a DNA sequence s with length L, which can contain L-1 dinucleotides, as defined below:

$$D_s = \left[a_1, a_2, a_3, .., a_i, .., a_{L-1}\right]\left(a_i \in AA, AT, AC, AG, TA \ldots, GG\right) \quad (1)$$

where, a_i represents the arrangement of dinucleotides in the sequence, which is one of 16 kinds of dinucleotides because the four bases can form 16 kinds of arrangement combinations in pairs. Dinucleotide ai is converted into five statistical parameters, which are defined as follows:

$$a_i = \left[p^i_{\min}, p^i_{\max}, p^i_{mean}, p^i_{\text{var}}, p^i_{sum}\right] \quad (2)$$

where $p^i_{\min}$, $p^i_{\max}$, p^i_{mean}, p^i_{var}, p^i_{sum} is the minimum, maximum, mean, variance, and sum of 90 physical and chemical properties of the *i*-th dinucleotide. Therefore, the DNA sequence with a length of 81 bp is finally converted into an $(81-1)\times 5 = 400$-dimensional feature vector. Detailed parameters of physical and chemical properties can be downloaded at http://lin-group.cn/server/iORI-PseKNC2.0/download.html.

2.3. Feature selection

Feature selection (Nasi et al., 2018; Zhang et al., 2019; Razzak et al., 2020) is to filter the redundant information in the original feature set to reduce the feature dimension and improve the calculation speed, which can reduce the model learning error caused by noise and improve (Aaron et al., 2019) the accuracy and robustness of the model.

In the process of feature expression, 400-dimensional statistical parameters of physical and chemical properties are used to describe DNA sequences. Due to the similarity between multiple physical and chemical properties and dinucleotide distribution, it is necessary to apply a feature selection algorithm to eliminate highly similar features. Currently, the main feature selection algorithms employed in biological sequence recognition are analysis of variance (ANOVA) (UniProt Consortium, 2012; Hebert et al., 2018; Wu et al., 2020; Moorthy and Gandhi, 2021) and maximum relevance maximum distance (MRMD) (Zou et al., 2016; Ao et al., 2021). ANOVA mainly reflects the contribution of features to the model by calculating the difference between positive and negative samples, then features with less contribution are deleted. MRMD judges the independence between samples and labels through various distance formulas, and features with low independence are filtered. However, the above methods have some defects, ANOVA only measures the difference between positive and negative samples of features, without considering the similarity between features. Oppositely, MRMD lacks the characteristics of analysis of positive and negative samples.

Considering the advantages and disadvantages of MRMD and ANOVA, we propose a novel feature selection method based on ANOVA and hierarchical clustering (HC) (Karna and Gibert, 2022; Zhu et al., 2022). As shown in Figure 3, the method comprehensively considers the similarity between features and the difference between a positive and negative sample of features. The first step is to calculate the F value of each one-dimensional feature, which is obtained by ANOVA of differences between groups and within groups, the 'f_classif' function in the 'sklearn' Python package is used to calculate the F value of each dimension feature. The second step is the hierarchical clustering analysis of features, the 'AgglomerativeClustering' function in 'sklearn' Python package is employed to measure the similarity between features. This algorithm mainly classifies two pairs of features into one cluster according to the distance between features, and we reserve the features with a large F value in each cluster of the first-level clustering results, when the F values are the same, a feature was selected at random. As shown in Figure 3, in the first-level clustering results, F_2 and F_3 are clustered into one cluster. If F_2 is larger than F_3, the feature of F_2 is retained, while F_1 is directly retained for a cluster alone. Therefore, the 3 dimensions feature ultimately remains 2 dimensions feature. In practical application, the 400 dimensions features are selected as the best subset of 215 dimensions for the final model construction.

2.4. Model development

The construction of the prediction model is the process of fitting sample labels according to the distribution of features. Because the feature descriptor designed is based on statistical parameters, it can be seen from Figure 2 that the designed feature distributions are very different, the positive and negative samples of feature subsets after feature selection also have this property. Therefore, the promoter prediction model has superior performance that required to accurately measure the confusion between sample features. The RF algorithm distinguishes the category of samples according to the confusion of feature information, so the algorithm is applied to the construction of the classifier. RF judges the disorder degree of samples according to the 'Gini' coefficient. A small 'Gini' coefficient means that the lower the disorder degree of samples, the greater the probability of correct recognition. The 'RandomForestClassifier' function in the 'sklearn' Python package is used to build the model. In the process of model training, the value range of five parameters is mainly adjusted by grid searching, the 'n_estimators' is 80 to 150 with 5 steps, the 'max_depth' is 15 to 20 with 1 in step, 'min_samples_leaf' is 1 to 8 with 1 in step, 'min_samples_split' is 2 to 5 with 1 in step, and 'max_features' is 0.1 to 1 with 0.1 in step, respectively. The determination of the best combination parameters is based on five-fold cross-validation.

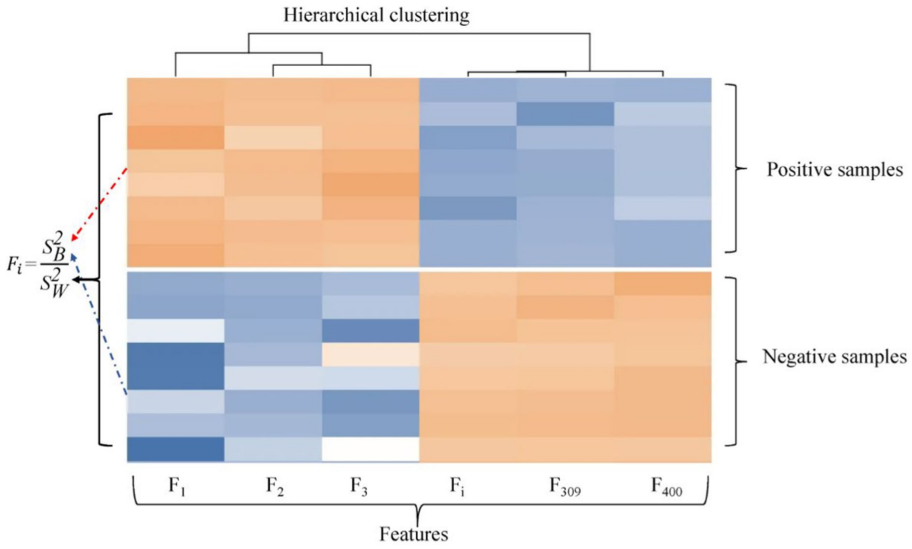

FIGURE 3
Feature Selection Schematic. F_i is the F value of the i-th dimension feature, S_B^2 and S_W^2 are differences between groups and within groups.

2.5. Evaluation parameters

The performance of the model needs to be evaluated by some indicators. For the second classification problem, the most common evaluation parameters (Xu et al., 2018; Chao et al., 2019; Demidova, 2021; Li et al., 2022a,b) are sensitivity (Sn), specificity (Sp), accuracy (Acc), Matthews correlation coefficient (MCC) and area under the Receiver Operating Characteristic (ROC) curve (AUC), which are defined as follows:

$$\begin{cases} Sn = \dfrac{TP}{TP + FN} \\ Sp = \dfrac{TN}{TN + FP} \\ Acc = \dfrac{TP + TN}{TP + FP + TN + FN} \\ MCC = \dfrac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \end{cases} \tag{3}$$

where TP and FP are correctly labeled promoters and incorrectly labeled promoters, and TN and *F* are correctly labeled non-promoters and incorrectly labeled non-promoters. Sn is employed to describe the model's ability to detect promoters, while Sp is employed to describe non-promoters. Acc, MCC, and AUC are used to describe the overall prediction capability of the model.

3. Result and discussion

3.1. Model performance analysis

A model with superior performance can not only accurately fit the sample labels on the training set, but also accurately judge the labels of unknown samples. To prove that the model proposed has the above qualifications, we summarize the results of five-fold cross-validation and independent set validation based on the RF (Zhang et al., 2009; Wei et al., 2017; Ao et al., 2021) prediction model in Table 1. It can be found from the table that in the first cross-validation, Sn, Acc and MCC, respectively, obtained the maximum value of 94.51, 93.13, and 86.26%, and Sp obtained the maximum value of 93.49% at the fourth cross-validation, which shows that different partition strategies of the dataset affect the fitting of the model, so the mean value of five-fold cross-validation is finally regarded as the standard prediction result. In general, the model proposed can accurately identify promoters and non-promoters, with an average Acc of 91.56%, Sn of 91.87%, and Sp of 91.17%. In addition, it can be seen from the ROC curve in Figure 4 that the performance of the model is superior, which shows that the AUC reaches more than 95%. To verify the robustness of the model, we conducted independent set tests and found that the model can also accurately distinguish promoters and non-promoters. In 400 independent samples, the model can correctly identify 181 promoters and 174 non-promoters, which confirms that our proposed model is capable of predicting annotated promoter fragments.

TABLE 1 The prediction performance of different subsets in RF.

Descriptor	Sn (%)	Sp (%)	Acc (%)	MCC (%)
1-th validation	**94.51**	91.67	**93.13**	**86.26**
2-th validation	92.59	91.39	91.88	83.75
3-th validation	91.39	91.72	91.56	83.08
4-th validation	90.73	**93.49**	92.19	84.32
5-th validation	90.12	87.84	89.06	77.99
Mean of validation	91.87	91.17	91.56	83.08
Independent verification	90.50	87.00	88.75	77.55

The bold value represents the maximum value. Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, matthew correlation coefficient.

3.2. Feature composition analysis

The excellent performance of the proposed model is driven by the accurate representation of feature descriptors and the filtering of redundant information by feature selection. It can be seen from Figure 5 that the features marked in red and marked in blue are clustered together and connected by dotted lines. The connected red-blue paired samples have high similarity, and the red samples with low *F* values are removed for noise removal, which horizontal dashed lines represent the points with far distance for dimensions, while vertical dashed lines represent the points with close distance, which proves that our method can filter global features rather than local features. Hence, 370 features are filtered out in half. The black diamond indicates that the samples are grouped into a single category, and they are directly retained. Finally, the feature dimension used to construct the samples is 215. More importantly, the feature accuracy of 400 dimensions has been improved from 90.69 to 91.56% of 215 dimensions, which shows that our feature selection method based on ANOVA and HC can reduce the redundancy of features and improve the model performance to a certain extent.

3.3. Multi-algorithm analysis

In the process of building the model, the RF classification algorithm is selected according to the characteristics of descriptor distribution. Although this algorithm has achieved good prediction

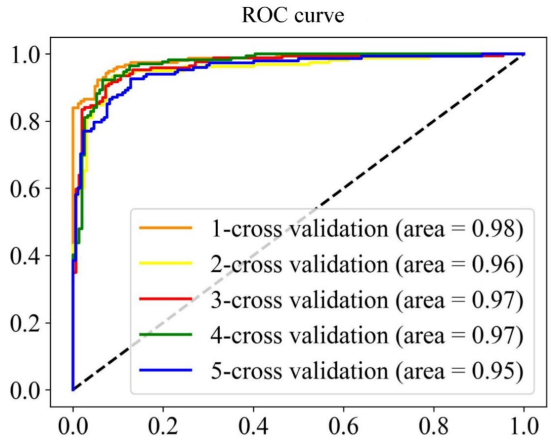

FIGURE 4
ROC curve of cross validation results.

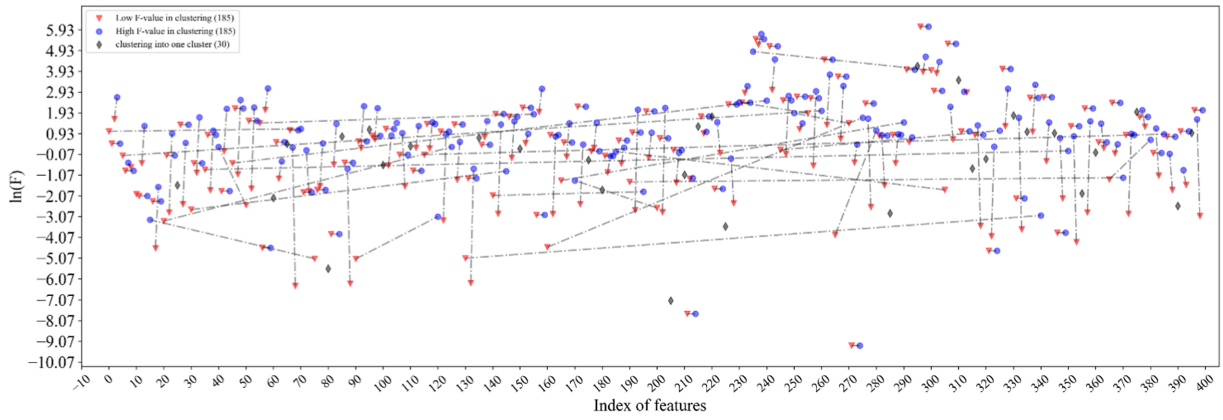

FIGURE 5
Visualization of feature selection results. The features marked in red and marked in blue are clustered together and connected by dotted lines. The black diamond indicates that the samples are grouped into a single category.

TABLE 2 Comparison of different classification algorithms.

Classifier	Verification	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC* (%)
KNN	Five-fold cross-validation	72.98	78.55	75.62	51.58
	Independent testing	67.00	81.00	74.00	48.48
SVM	Five-fold cross-validation	88.59	86.77	87.63	75.31
	Independent testing	82.00	81.50	81.75	63.50
MLP	Five-fold cross-validation	85.25	85.58	85.44	70.85
	Independent testing	79.00	82.50	80.75	61.51
RF	Five-fold cross-validation	**91.87**	**91.17**	**91.56**	**83.08**
	Independent testing	90.50	87.00	88.75	77.55

The bold value represents the maximum value. Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, matthew correlation coefficient; SVM, support vector machine; RF, random forest; MLP, multi-layer perceptron; KNN, k-nearest neighbors.

performance, it is still possible that other classification algorithms have better results, such as K nearest neighbor (KNN) (Wang et al., 2012; Demidova, 2021), Support vector machine (SVM) (Xu et al., 2018; Xiao et al., 2019), Multi-layer perceptron (MLP) (Majidzadeh Gorjani et al., 2021; Lin et al., 2022). Therefore, we compared different classification algorithms based on filtered features. It can be seen from Table 2 that in cross-validation, the performance of the RF is the best. The prediction accuracy of SVM is 87.63%, which is closest to the RF, followed by the MLP with an accuracy of 85%, and the worst KNN accuracy is only 75.62%. The situation of independent verification is consistent with the above situation. And only the accuracy of the RF algorithm has the smallest difference between independent set verification and cross verification, which also proves that the proposed model has strong robustness and small overfitting analysis.

4. Conclusion

In this work, we collected promoter and non-promoter sequences of *C. glutamicum* with annotation information, then designed a feature descriptor based on statistical parameters according to the distribution characteristics of physical and chemical properties. Further, we defined the novel feature selection method to filter redundant information among features. Finally, we successfully built the prediction model based on RF that can accurately identify promoters. In a word, the model we designed can accurately identify the promoter sequences of eukaryotes, and we hope that the feature descriptors and feature selection methods designed can s make positive contributions to other sequence classification problems.

Data availability statement

The original datasets and code used in this study can be found at https://github.com/Hongfeipower/Predicting-Cornebacterium-glutamicum-Promoters.

Author contributions

HL and YZ designed the study. HL and JZ carried out all data collection and drafted the manuscript. WY and YZ revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work has been partially supported by the National Natural Science Foundation of China (61971119, 62272094), and the National Key R&D Program of China (2021YFC2100103).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Aaron, J. S., Taylor, A. B., and Chew, T.-L. (2019). The Pearson's correlation coefficient is not a universally superior colocalization metric. Response to 'Quantifying colocalization: the MOC is a hybrid coefficient - an uninformative mix of co-occurrence and correlation. *J. Cell Sci.* 132:74. doi: 10.1242/jcs.227074

Ao, C., Zou, Q., and Yu, L. (2021). RFhy-m2G: identification of RNA N2-methylguanosine modification sites based on random forest and hybrid features. *Methods* 203, 32–39. doi: 10.1016/j.ymeth.2021.05.016

Barrangou, R., and Horvath, P. (2017). A decade of discovery: CRISPR functions and applications. *Nat. Microbiol.* 2:92. doi: 10.1038/nmicrobiol.2017.92

Bharanikumar, R., Premkumar, K. A. R., and Palaniappan, A. (2018). PromoterPredict: sequence-based modelling of *Escherichia coli* sigma (70) promoter strength yields logarithmic dependence between promoter strength and sequence. *PeerJ* 6:e5862. doi: 10.7717/peerj.5862

Blumenstein, J., Radisch, R., Stepanek, V., Grulich, M., Dostalova, H., and Patek, M. (2022). Identification of *Rhodococcus erythropolis* promoters controlled by alternative sigma factors using in vivo and in vitro systems and heterologous RNA polymerase. *Curr. Microbiol.* 79:55. doi: 10.1007/s00284-021-02747-8

Canzio, D., Nwakeze, C. L., Horta, A., Rajkumar, S. M., Coffey, E. L., Duffy, E. E., et al. (2019). Antisense lncRNA transcription mediates DNA demethylation to drive stochastic Protocadherin alpha promoter choice. *Cells* 177, 639–653.e15. doi: 10.1016/j.cell.2019.03.008

Chao, L., Wei, L., and Zou, Q. (2019). SecProMTB: a SVM-based classifier for secretory proteins of mycobacterium tuberculosis with imbalanced data set. *Proteomics* 19:e1900007. doi: 10.1002/pmic.201900007

Cho, J. S., Choi, K. R., Prabowo, C. P. S., Shin, J. H., Yang, D., Jang, J., et al. (2017). CRISPR/Cas9-coupled recombineering for metabolic engineering of *Corynebacterium glutamicum*. *Metab. Eng.* 42, 157–167. doi: 10.1016/j.ymben.2017.06.010

Dao, F.-Y., Lv, H., Wang, F., Feng, C.-Q., Ding, H., Chen, W., et al. (2019). Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique. *Bioinformatics* 35, 2075–2083. doi: 10.1093/bioinformatics/bty943

Demidova, L. A. (2021). Two-stage hybrid data classifiers based on SVM and kNN algorithms. *Symmetry* 13:13. doi: 10.3390/sym13040615

Hebert, P. D. N., Braukmann, T. W. A., Prosser, S. W. J., Ratnasingham, S., Dewaard, J. R., Ivanova, N. V., et al. (2018). A sequel to sanger: amplicon sequencing that scales. *BMC Genom.* 19:4611. doi: 10.1186/s12864-018-4611-3

Hu, J., Tan, Y., Li, Y., Hu, X., Xu, D., and Wang, X. (2013). Construction and application of an efficient multiple-gene-deletion system in *Corynebacterium glutamicum*. *Plasmid* 70, 303–313. doi: 10.1016/j.plasmid.2013.07.001

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680–682. doi: 10.1093/bioinformatics/btq003

Huang, H., Song, X., and Yang, S. (2019). Development of a RecE/T-assisted CRISPR-Cas9 toolbox for lactobacillus. *Biotechnol. J.* 14:1800690. doi: 10.1002/biot.201800690

Jeon, A. J., and Tucker-Kellogg, G. (2020). Bivalent genes that undergo transcriptional switching identify networks of key regulators of embryonic stem cell differentiation. *BMC Genomics* 21:14. doi: 10.1186/s12864-020-07009-8

Jiang, Y., Qian, F., Yang, J., Liu, Y., Dong, F., Xu, C., et al. (2017). CRISPR-Cpf1 assisted genome editing of *Corynebacterium glutamicum*. *Nat. Commun.* 8:15179. doi: 10.1038/ncomms15179

Kalinowski, J., Bathe, B., Bartels, D., Bischoff, N., Bott, M., Burkovski, A., et al. (2003). The complete *Corynebacterium glutamicum* ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. *J. Biotechnol.* 104, 5–25. doi: 10.1016/S0168-1656(03)00154-8

Karna, A., and Gibert, K. (2022). Automatic identification of the number of clusters in hierarchical clustering. *Neural Comput. Applic.* 34, 119–134. doi: 10.1007/s00521-021-05873-3

La Fleur, T., Hossain, A., and Salis, H. M. (2022). Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria. *Nat. Commun.* 13:5159. doi: 10.1038/s41467-022-32829-5

Leemans, C., Van Der Zwalm, M. C. H., Brueckner, L., Comoglio, F., Van Schaik, T., Pagie, L., et al. (2019). Promoter-intrinsic and local chromatin features determine gene repression in LADs. *Cells* 177:852. doi: 10.1016/j.cell.2019.03.009

Li, W. Z., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Li, H., Gong, Y., Liu, Y., Lin, H., and Wang, G. (2022a). Detection of transcription factors binding to methylated DNA by deep recurrent neural network. *Brief. Bioinform.* 23:23. doi: 10.1093/bib/bbab533

Li, H., Shi, L., Gao, W., Zhang, Z., Zhang, L., Zhao, Y., et al. (2022b). dPromoter-XGBoost: detecting promoters and strength by combining multiple descriptors and feature selection using XGBoost. *Methods* 204, 215–222. doi: 10.1016/j.ymeth.2022.01.001

Lin, R., Zhou, Z., You, S., Rao, R., and Kuo, C. C. J. (2022). Geometrical interpretation and Design of Multilayer Perceptrons. *IEEE Trans. Neural Netw. Learn. Syst.* PP, 1–15. doi: 10.1109/TNNLS.2022.3190364

Liu, Y., Yu, L., Pukhrambam, C., Winkelman, J. T., Firlar, E., Kaelber, J. T., et al. (2022). Structural and mechanistic basis of reiterative transcription initiation. *Proc. Natl. Acad. Sci. U. S. A.* 119:119. doi: 10.1073/pnas.2115746119

Majidzadeh Gorjani, O., Byrtus, R., Dohnal, J., Bilik, P., Koziorek, J., and Martinek, R. (2021). Human activity classification using multilayer perceptron. *Sensors* 21:207. doi: 10.3390/s21186207

Moorthy, U., and Gandhi, U. D. (2021). A novel optimal feature selection technique for medical data classification using ANOVA based whale optimization. *J. Ambient. Intell. Humaniz. Comput.* 12, 3527–3538. doi: 10.1007/s12652-020-02592-w

Nasi, R., Viljanen, N., Kaivosoja, J., Alhonoja, K., Hakala, T., Markelin, L., et al. (2018). Estimating biomass and nitrogen amount of barley and grass using UAV and aircraft based spectral and photogrammetric 3D features. *Remote Sens.* 10:1082. doi: 10.3390/rs10071082

Okino, S., Suda, M., Fujikura, K., Inui, M., and Yukawa, H. (2008). Production of D-lactic acid by *Corynebacterium glutamicum* under oxygen deprivation. *Appl. Microbiol. Biotechnol.* 78, 449–454. doi: 10.1007/s00253-007-1336-7

Razzak, I., Abu Saris, R., Blumenstein, M., and Xu, G. (2020). Integrating joint feature selection into subspace learning: a formulation of 2DPCA for outliers robust feature selection. *Neural Netw.* 121, 441–451. doi: 10.1016/j.neunet.2019.08.030

Rengachari, S., Schilbach, S., Kaliyappan, T., Gouge, J., Zumer, K., Schwarz, J., et al. (2022). Structural basis of SNAPc-dependent snRNA transcription initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.* 29, 1159–1169. doi: 10.1038/s41594-022-00857-w

Sano, C. (2009). History of glutamate production. *Am. J. Clin. Nutr.* 90, 728S–732S. doi: 10.3945/ajcn.2009.27462F

Shang, X., Chai, X., Lu, X., Li, Y., Zhang, Y., Wang, G., et al. (2018). Native promoters of *Corynebacterium glutamicum* and its application in l-lysine production. *Biotechnol. Lett.* 40, 383–391. doi: 10.1007/s10529-017-2479-y

Silar, R., Holatko, J., Rucka, L., Rapoport, A., Dostalova, H., Kaderabkova, P., et al. (2016). Use of in vitro transcription system for analysis of *Corynebacterium glutamicum* promoters recognized by two sigma factors. *Curr. Microbiol.* 73, 401–408. doi: 10.1007/s00284-016-1077-x

Stepanek, V., Dostalova, H., Busche, T., Blumenstein, J., Grulich, M., Plasil, L., et al. (2022). Sigma regulatory network in *Rhodococcus erythropolis* CCM2595. *FEMS Microbiol. Lett.* 369:fnac014. doi: 10.1093/femsle/fnac014

Su, W., Liu, M.-L., Yang, Y.-H., Wang, J.-S., Li, S.-H., Lv, H., et al. (2021). PPD: a manually curated database for experimentally verified prokaryotic promoters. *J. Mol. Biol.* 433:166860. doi: 10.1016/j.jmb.2021.166860

Sun, K., Kasperski, A., Tian, Y., and Chen, L. (2011). Modelling of the *Corynebacterium glutamicum* biosynthesis under aerobic fermentation conditions. *Chem. Eng. Sci.* 66, 4101–4110. doi: 10.1016/j.ces.2011.05.041

Theron, G., and Reid, S. J. (2011). ArgR-promoter interactions in *Corynebacterium glutamicum* arginine biosynthesis. *Biotechnol. Appl. Biochem.* 58, 119–127. doi: 10.1002/bab.15

UniProt Consortium (2012). Reorganizing the protein space at the universal protein resource (UniProt). *Nucleic Acids Res.* 40, D71–D75. doi: 10.1093/nar/gkr981

Vertes, A. A., Inui, M., and Yukawa, H. (2012). Postgenomic approaches to using Corynebacteria as biocatalysts. *Ann. Rev. Microbiol.* 66, 521–550. doi: 10.1146/annurev-micro-010312-105506

Wang, C.-X., Dong, L.-L., Pan, Z.-M., and Zhang, T. (2012). Classification for unbalanced dataset by an improved KNN algorithm based on weight. *Inf. Int. Interdiscipl. J.* 15, 4983–4988.

Wei, L. Y., Xing, P. W., Su, R., Shi, G. T., Ma, Z. S., and Zou, Q. (2017). CPPred-RF: a sequence-based predictor for identifying cell penetrating peptides and their uptake efficiency. *J. Proteome Res.* 16, 2044–2053. doi: 10.1021/acs.jproteome.7b00019

Wu, C., Yan, Y., Cao, Q., Fei, F., Yang, D., Lu, X., et al. (2020). sEMG measurement position and feature optimization strategy for gesture recognition based on ANOVA and neural networks. *Ieee Access* 8, 56290–56299. doi: 10.1109/ACCESS.2020.2982405

Xiao, X., Xu, Z. C., Qiu, W. R., Wang, P., Ge, H. T., and Chou, K. C. (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, 1785–1793. doi: 10.1016/j.ygeno.2018.12.001

Xu, L., Liang, G., Shi, S., and Liao, C. (2018). SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins. *Int. J. Mol. Sci.* 19:1773. doi: 10.3390/ijms19061773

Zhang, G., Li, H., and Fang, B. (2009). Discriminating acidic and alkaline enzymes using a random forest model with secondary structure amino acid composition. *Process Biochem.* 44, 654–660. doi: 10.1016/j.procbio.2009.02.007

Zhang, L., Su, H., and Shen, J. (2019). Hyperspectral dimensionality reduction based on multiscale Superpixelwise kernel principal component analysis. *Remote Sens.* 11:1219. doi: 10.3390/rs11101219

Zhu, S., Xu, L., and Goodman, E. D. (2022). Hierarchical topology-based cluster representation for scalable evolutionary multiobjective clustering. *IEEE Trans. Cybernet.* 52, 9846–9860. doi: 10.1109/TCYB.2021.3081988

Zou, Q., Zeng, J. C., Cao, L. J., and Ji, R. R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354. doi: 10.1016/j.neucom.2014.12.123

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 12 April 2023
DOI 10.3389/fgene.2023.1151703

OPEN ACCESS

EDITED BY
Hao Lin,
University of Electronic Science and Technology of China, China

REVIEWED BY
Wei Chen,
Chengdu University of Traditional Chinese Medicine, China
Ruifang Li,
Inner Mongolia Normal University, China
Guijun Yan,
University of Western Australia, Australia

*CORRESPONDENCE
Zhanyuan Lu,
lzhy2811@163.com
Qiang Zhang,
zhangqiang829@163.com
Xiaoqing Zhao,
zhaoxq204@163.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics

RECEIVED 26 January 2023
ACCEPTED 24 March 2023
PUBLISHED 12 April 2023

CITATION
Bo S, Sun Q, Li Z, Aodun G, Ji Y, Wei L, Wang C, Lu Z, Zhang Q and Zhao X (2023), Ubiquitous conservative interaction patterns between post-spliced introns and their mRNAs revealed by genome-wide interspecies comparison. *Front. Genet.* 14:1151703. doi: 10.3389/fgene.2023.1151703

Ubiquitous conservative interaction patterns between post-spliced introns and their mRNAs revealed by genome-wide interspecies comparison

Suling Bo[1†], Qiuying Sun[2], Zhongxian Li[1], Gerile Aodun[1], Yucheng Ji[1], Lihua Wei[1], Chao Wang[1], Zhanyuan Lu[3,4,5,6*], Qiang Zhang[7*] and Xiaoqing Zhao[3,4,5,6*†]

[1]College of Computer Information, Inner Mongolia Medical University, Hohhot, China, [2]Department of Oncology, Inner Mongolia Cancer Hospital and the Affiliated People's Hospital of Inner Mongolia Medical University, Hohhot, China, [3]Inner Mongolia Academy of Agricultural and Animal Husbandry Sciences, Hohhot, China, [4]School of Life Science, Inner Mongolia University, Hohhot, China, [5]Key Laboratory of Black Soil Protection and Utilization (Hohhot), Ministry of Agriculture and Rural Affairs, Hohhot, China, [6]Inner Mongolia Key Laboratory of Degradation Farmland Ecological Restoration and Pollution Control, Hohhot, China, [7]College of Science, Inner Mongolia Agriculture University, Hohhot, China

Introns, as important vectors of biological functions, can influence many stages of mRNA metabolism. However, in recent research, post-spliced introns are rarely considered. In this study, the optimal matched regions between introns and their mRNAs in nine model organism genomes were investigated with improved Smith–Waterman local alignment software. Our results showed that the distributions of mRNA optimal matched frequencies were highly consistent or universal. There are optimal matched frequency peaks in the UTR regions, which are obvious, especially in the 3′-UTR. The matched frequencies are relatively low in the CDS regions of the mRNA. The distributions of the optimal matched frequencies around the functional sites are also remarkably changed. The centers of the GC content distributions for different sequences are different. The matched rate distributions are highly consistent and are located mainly between 60% and 80%. The most probable value of the optimal matched segments is about 20 bp for lower eukaryotes and 30 bp for higher eukaryotes. These results show that there are abundant functional units in the introns, and these functional units are correlated structurally with all kinds of sequences of mRNA. The interaction between the post-spliced introns and their corresponding mRNAs may play a key role in gene expression.

KEYWORDS

local matched alignment, optimal matched region, interaction patterns, ubiquitous conservative, gene expression

1 Introduction

Since introns, a kind of non-coding DNA, were discovered, there have been many investigations of their functions and evolutionary origin (Roy, 2003). A research study recognized that the main function of introns is alternative splicing, facilitating the expression of multiple proteins from a single gene (Daehyun and Phil, 2005). Recently, it has become

increasingly clear that introns are very important vectors of biological functions (Mattick and Gagen, 2001; Nott et al., 2003; Bianchi et al., 2009; Charital et al., 2009), and the sequence structures of introns and behavior of introns when removed by spliceosomes can influence many stages of mRNA metabolism (Orphanides and Reinberg, 2002; Hir et al., 2003). Many experiments have shown that introns can boost gene expression (Buchman and Berg, 1988; McKenzie and Brennan, 1996). Intron-containing transgenes in mice are transcribed 10–100 times more efficiently than their intron-less counterparts (Brinster et al., 1988), and the transcription of intron-less mRNA *in vivo* directs this mRNA toward translational silencing, while mRNA translational efficiency is dramatically increased by the addition of just one generic intron to the pre-mRNA (Callis et al., 1987). Although some genes contain no introns or their expressions do not require introns, introns can still improve the gene expression of genetically modified organisms (Duncker et al., 1997; Ko et al., 1998). It has also been discovered that the two small introns of the *Drosophila affinidisjuncta* (Adh) gene are required for normal transcriptions (Braddock et al., 1994). Intron mutation can cause many diseases. Besides the mutation at each end (GU and AG), the mutation in the middle of the intron sequences can also cause diseases by activating recessive splice sites (Stover and Verrelli, 2010; Nordin et al., 2012).

An increasing body of evidence shows that there are many introns in the cytoplasm and that they directly regulate gene translational efficiency. Intron sequences are retained in a number of dendritically targeted mRNAs in the cytoplasm (Buckley et al., 2011). Certain spliced mRNAs can be efficiently exported and translated, whereas the same mRNA transcribed from cDNA fails to exit the nucleus and express protein (Ryu and Mertz, 1989; Rafiq et al., 1997; Matsumoto et al., 1998). Removal of an intron from a pre-mRNA, without significantly altering the steady-state cytoplasmic mRNA level, can also affect translational efficiency (Luo and Reed, 1999). Similarly, when a mature mRNA is injected directly into oocyte nuclei, the translation efficiency is repressed and overcome by either adding a spliced intron or injecting FRGY2 protein antibodies into the cytoplasm (Hir et al., 2003). In addition, it is interesting to note that introns can suppress RNA silencing in *Arabidopsis* (Christie et al., 2011).

Many experiments have proved that introns function significantly in all processes of regulating the dynamic structure of mRNAs, their transport and nuclear export, and translation and regulation (Guigó and Ullrich, 2020; Gozashti et al., 2022). However, how introns take part in these biological processes is still unclear. In the past, it was believed that most pre-mRNAs were spliced to liner molecules with only exons. However, circular RNAs (circRNAs) were discovered, showing that the exon–circRNA model is formed by lariat-driven cyclization and intron-paired cyclization (Hansen et al., 2011; Jeck et al., 2013; Sebastian et al., 2013) and circular intronic RNAs can be formed by introns as well (Julia et al., 2012).

Based on these observations, it is believed that introns can directly affect gene expression after splicing by their interactions with the corresponding mRNAs. These kinds of interactions can maintain and regulate mRNA structures. The loss/gain of an intron does affect gene expression after splicing and plays a very important role in the evolution of the eukaryotic genome and the presence of new eukaryotic species (Duret, 2001; Halligan and Keightley, 2006). The interaction between post-spliced introns and their CDS was studied in our early works (Zhao et al., 2013; Zhang et al., 2016; Bo et al., 2019), but the 5′-UTR and 3′-UTR of mRNA are very important to gene expression. It is therefore very meaningful to study the interaction between introns and their corresponding mRNAs and to uncover how introns influence stages of gene expression after splicing by their interactions. Here, we report on the interaction characters between post-spliced introns and their mRNAs in whole genomes.

2 Materials and methods

2.1 Gene sequences

Genes from nine model organism genomes were selected as our dataset. They are *Caenorhabditis elegans*, *Drosophila melanogaster*, *Apis mellifera*, *Anopheles gambiae*, *Arabidopsis thaliana*, *Oryza sativa*, *Danio rerio*, *Mus musculus*, and *Homo sapiens*, and their gene sequences were downloaded from the Beijing Multi Subnet of Gene Bank (ftp://ftp.cbi.pku.edu.cn/pub/database/genomes). In this dataset, the genes that contain more than one mRNA were excluded first. Next, the genes that contain ncRNAs and/or repetitive elements were excluded. Last, introns with lengths shorter than 40 bp were also excluded. The results of the dataset are shown in Table 1.

2.2 Matched alignment

The interaction between introns and their mRNAs was represented by optimal matched segments. The interaction probability was determined by the quality of the optimal matched segments. The mRNAs were renamed as tested sequences, while their corresponding introns were aligned sequences. To obtain the matched alignment segments, introns were transformed into their complementary sequences, and similar alignments were performed using improved Smith–Waterman local alignment software (http://mobyle.pasteur.fr/cgi-bin/). In the alignment process, the EDNAFULL matrix was used for calculating the optimal matched segments with the following parameters: 50.0 for the gap penalty and 5.0 for extend penalty. In this way, the most credible optimal matched segment of the tested sequence and its aligned sequence were obtained. The local alignment sketch map is shown in Figure 1.

For the tested sequence, the matched score function f is defined by Eq. 1.

$$f = \begin{cases} 1 & N_s \le j \le N_e \\ 0 & j < N_s \; or \; j > N_e \end{cases}, \tag{1}$$

where j means the jth base site of the tested sequence ($j = 1,2, \ldots, L$), L means the length of the tested sequence, and N_s and N_e mean the start base site and the end base site of the optimal matched segment in the tested sequence, respectively. The effective value 1 is assigned to each base site within the optimal matched segment, while the ineffective value 0 is assigned to the base sites outside the optimal matched segment. The matched score values are assigned to each base site in the tested sequences.

TABLE 1 Genes of nine eukaryotes.

	Chromosome	Number of genes	Number of introns
Caenorhabditis elegans	1	956	4,052
Drosophila melanogaster	1	1,322	3,846
Arabidopsis thaliana	1	3,311	16,822
Apis mellifera	1	439	2,894
Anopheles gambiae	1	2,238	7,919
Oryza sativa	1	594	2,905
Danio rerio	1	1,005	8,563
Mus musculus	1	1,126	10,117
Homo sapiens	1	1,194	9,265

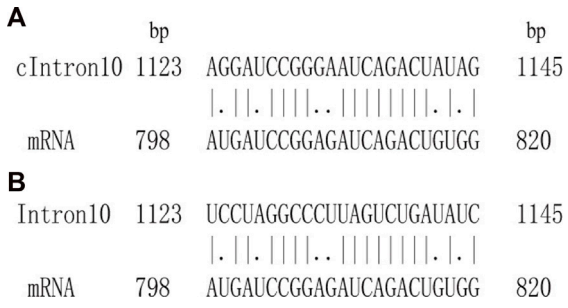

FIGURE 1
Sketch matched map between intron and corresponding CDS. **(A)** Smith–Waterman local alignment. cIntron10 means the complementary segment of intron 10. **(B)** Authentic matched alignment.

For the tested sequences, matched frequency F is defined by Eq. 2.

$$F = 1/m\sum_{i=1}^{m} f_{ij}, \tag{2}$$

where m means the number of the tested sequences, i means the ith tested sequence (i = 1,2, . . . , m), j means the jth base site of the ith tested sequence (j = 1, 2, . . . , L_i), L_i means the length of the ith tested sequence, and f_{ij} means the matched score function of the jth base site of the ith tested sequence. F is a relative matched value at the jth base site in m tested sequences. It reflects the interacting probability or the interaction intensity between the tested and aligned sequences in the jth base site.

The average matched frequency <F> for each base site is also defined by Eq. 3.

$$\langle F\rangle = 1/m\sum_{i=1}^{m} l_i/L_i. \tag{3}$$

Here, i means the ith tested sequence (i = 1, 2, . . . , m), l_i is the length of the optimal matched segment for the ith tested sequence, and L_i is the length of the ith tested sequence. The <F> indicates the average matched frequency of the m tested sequences.

The relative matched frequency RF of the jth base site in the tested sequence is defined by Eq. 4.

$$RF = F/\langle F\rangle. \tag{4}$$

Here, RF reflects the relative bias of each base site in the tested sequences. If RF > 1, it indicates that the interaction in the jth base site is positive in the tested sequence, and the regions with RF > 1 were optimal matched regions. RF = 1 represents an average matched frequency of the base sites for the tested sequences.

To test the significance of our results, we constructed corresponding component constraint random sequences for comparison with real sequences. Component constraint random sequences mean the length and contents of A, C, G, and U are the same as the analyzed sequence, but the order of each base is random. We call them CC-random sequences. The sample of the corresponding component constraint random sequences is 10 times as many as the analyzed sequences, and then the corresponding RF or F distributions are obtained in the same way. When the RF values in the optimal matched regions of the mRNA all are higher than the CC-random sequences and average matched frequency <F>, we call these cases positive tests.

2.3 Sequence normalization

Due to the different lengths of the tested sequences, they are normalized to 100 to obtain the relative site distributions of RF or F by the following method.

We hypothesized that n_{ij} is the jth relative site of the ith normalized tested sequence; the n_{ij} is obtained by the following formulation:

$$n_{ij} = \begin{cases} \left[100N_{ij}/L_i\right] & 100N_{ij}/L_i \text{is integer} \\ \left[100N_{ij}/L_i\right] & 100N_{ij}/L_i \text{is non}-\text{integer} \end{cases}. \tag{5}$$

Here, N_{ij} means the jth base site of the ith tested sequence, and L_i is the length of ith tested sequence ($i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, L_i$). The square brackets are Gaussian integer functions which are meant to take the integer part of a real number. Then, the m tested sequences with different lengths are normalized to 100. In addition, n_{is}, n_{ie}, $n_{ie}^{\cdot}$ $n_{is}+1$, n_{ij} and 100 are used to replace N_{is}, N_{ie}, l_i, N_{ij} and L_i in the formulation (1), (2), (3) and (4) respectively, the normalized relative matching frequency function *RF* or matching frequency function *F* distribution can be obtained.

2.4 Information entropy analysis

Information entropy conception was used to analyze the characters of sequence composition. Second-order informational redundancy D_2 is a suitable parameter to describe the sequence characters; its definition is shown in Eq. 6.

For an analyzed sequence, the second-order informational redundancy D_2 is defined as

$$D_2 = \sum p_{ij}\log_2\left(p_{ij}/p_i p_j\right) \approx 1/2\ln 2\sum\left(p_{ij} - p_i p_j\right)^2/p_i p_j, \quad (6)$$

where p_i or p_j is the probability of the base i or j (i, j = A, C, G, U), and p_{ij} is the joint probability of the base pair ij. D_2 reflects the adjacent base correlation of sequences (Luo and Hong, 1991; Li, 1990). In other words, a bigger D_2 value means that the sequence is more conservative. For a finite sequence of length N, the fluctuation bound (f.b.) of D_2 is D_2(f.b.) = 15.65/N (Luo and Hong, 1991; Luo, 2004). When $D_2 \geq 15.65/N$, the neighboring bases do not occur independently, and the correlation exists at a 99% confidence level. Generally, $D_2 \geq 0$. For infinite random sequences, $D_2 = 0$.

3 Results and discussion

3.1 Distributions of optimal matched regions in mRNA

For the nine model organisms, mRNAs are regarded as the tested sequences, and their corresponding introns are regarded as the aligned sequences. The matched alignments between the mRNAs and their corresponding introns were performed, and the *RF* distributions with base relative sites of mRNA sequences were obtained (Supplementary Appendix SA1). Meanwhile, the local alignments were also performed between the component constraint random mRNAs and their own component constraint random introns, and they were marked as CC-random (Supplementary Appendix SA2). The results are shown in Figure 2.

The relative matched frequency distributions of the mRNA sequences of the nine model organisms were very similar to each other, which meant that the interaction patterns between the introns and their mRNAs are universal. When compared with the CC-random group, their characteristics are as follows: there are high relative matched frequencies in the UTR regions but a relatively low matched degree in the central protein-coding sequence. The relative matched frequency distributions of 3′-UTR are significantly higher than those of 5′-UTR (Supplementary Appendix SA3). It is speculated that the function of post-spliced introns is related to NMD. Compared with those in higher organisms, the matched frequency distributions in the mRNAs of lower eukaryotes are slightly different, and the distribution difference between the coding sequences and the UTR regions in lower organisms is more obvious, which reflects that the interaction modes between introns and their mRNAs in higher organisms are more complex.

Despite the species being very similar to each other for the matched frequency distributions in the mRNAs of the nine model organisms, the distributions of peak regions and peak values are slightly different (Figure 2). For *C. elegans*, the peak regions of the matched frequency distribution in the mRNAs are mainly located in 3%–8% of the 5′-end and between 80% and 98% of the 3′-end of the mRNA; the peaks are approximately 1.1 and 3.9, respectively. The peak regions for *D. melanogaster* are primarily found in 2%–10% of the 3′-end and 85%–99% of the 5′-end of the mRNA; the peak values are about 1.8 and 4.2, respectively. The peak regions for *A. thaliana* are mainly located in 2%–10% of the 5′-end and 85%–98% of the 3′-end of the mRNA, and the peak values are about 1.6 and 2.3, respectively. The peak regions for *A. mellifera* are mainly located in 2%–10% of the 5′-end and 80%–98% of the 3′-end of the mRNA, and the peak values are about 1.5 and 3.7, respectively. The peak regions of the distribution of *A. gambiae* are mainly located in 2%–20% of the 5′-end and 82%–98% of the 3′-end of the mRNA, respectively, and both peak values are about 1.4. The peak regions for *O. sativa* are mainly located in 80%–98% of the 3′-end of the mRNA, and the peak value is about 1.8. The peak regions for *D. rerio* are mainly in the 5%–8% of the 5′-end and 78%–99% of the 3′-end of the mRNA, and the peak values are about 1.1 and 3.1, respectively. The peak regions for *M. musculus* are mainly located in 80%–99% of the 3′-end of the mRNA, and both peak values are about 2.2. The peak regions for *H. sapiens* are distributed in 5%–10% of the 5′-end and 62%–99% of the 3′-end of the mRNA, and the peaks are about 1.1 and 2.0, respectively. It is suggested that introns have a strong preference for interactions with the corresponding mRNAs.

3.2 Matched characteristics of optimal matched segments of introns

It is of great significance when the interaction between introns and the corresponding coding sequences are studied and the sequence characteristics of the optimal matched fragments are analyzed. The sequence paired rate and distribution length of the optimal matched segments of introns between the introns and corresponding mRNA are explored in this section. The results are shown in Figure 3.

From lower eukaryotes to higher eukaryotes, the high consistency of the matched rate of distribution of the optimal matched segments is seen, and the matched rate fluctuates between 60% and 80%. Several clear and conservative peaks are observed, with a clear maximal peak at about 68% and a distinctly sub-maximal peak at about 75%, followed by several discrete peaks with a gradual decrease in distribution (Figure 3).

It is inferred that introns have a precise “quantum state” with the sequences of the optimal matched mRNA segments for each corresponding intron and that each “quantum state” might

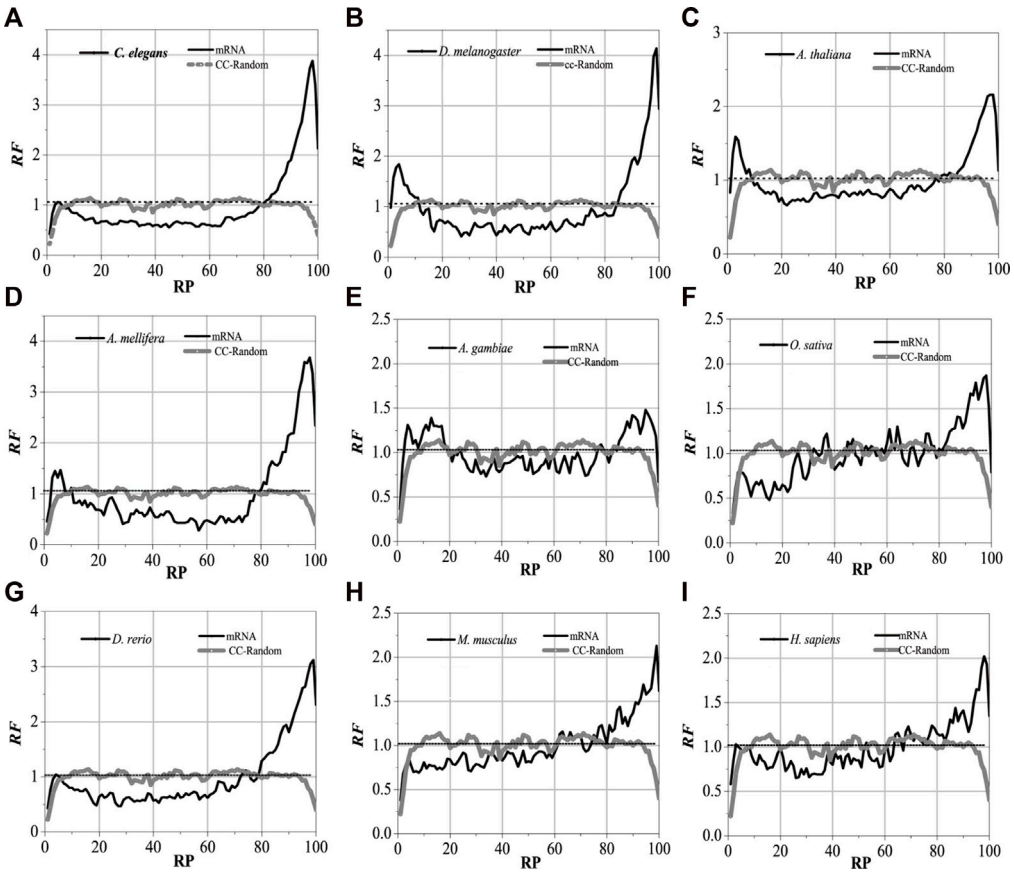

FIGURE 2
RF distributions of mRNA. The X-axis is the relative position of mRNA and the Y-axis represents the *RF* values. CC-Random means the local alignment were done between the component constraint random mRNA and their own component constraint random introns. *RF* = 1 represents the average value of relative match frequencies theoretically.

represent a specific kind of pattern group in which introns control gene expressions. There is no doubt that this distribution is universal from lower eukaryotes to higher eukaryotes. The optimal matched segments between introns and their associated mRNAs have a very noticeable peak in their sequence length distribution. The most probable value, however, varies between the lower and higher eukaryotes; it is roughly 20 bp for lower eukaryotes and 30 bp for higher eukaryotes. It can be implied that higher and lower eukaryotes may differ significantly in the intricacy of gene expression patterns mediated by introns. When compared with siRNA and miRNA, the optimal matched segment from the introns with the associated mRNA appears to have a beneficial impact on gene expression.

3.3 Distributions of optimal matched regions near functional sites

Translation initiation sites, translation termination sites, and exon junction sites exert an irreplaceable role in the normal expression of genes. It is particularly crucial to comprehend the distribution of optimal matched frequency near functional locations. The optimal matched intron segments around each functional site containing the UTR gene are separated, functional sites are set as the origin of coordinates, and the distribution law of the optimal matched regions is counted. These functional sites are translation initiation sites, translation termination sites, the junction sites between the first exon and second exon (the first exon junction site), the junction sites between the last exon with the penultimate exon (the last exon junction site), and the junctions in the middle exon (the middle exon junction site). The results are shown in Figures 4–7 (the results of the junction sites Figures 8, 9 are shown in Supplementary Material S5).

3.3.1 Distribution of optimal matched regions of translation initiation regions

Analytically, the distribution of optimal matched frequency of translation initiation sites of the nine model organisms is revealed to be of high consistency, that is, these model species show good universality for the distributions. The matched frequencies bounded by the translation initiation sites on the sequences near the translation initiation site are significantly altered, it is specifically manifested in the relative matched frequencies of the UTR on the left side of the translation initiation site, and the corresponding intron is generally higher. There is an excellent agreement between their

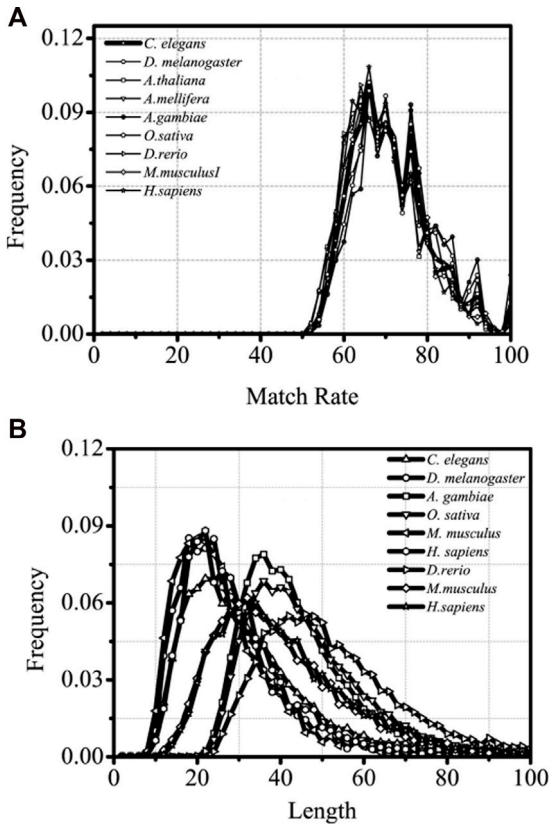

FIGURE 3
Matched rate and Length distributions distributions of different intron optimal matchedsegments, separately. The X-axis is Length and matched rate (%)of intron optimal matched segment,separately and the Y-axis represents the Frequency values.

distribution of optimal matched regions and the 5′-end of the mRNA. When comparing short introns, the distribution of their optimal matched region is consistent with that of long introns. Large differences in the distribution of optimal matched regions are observed in the short introns, but some distributions are very peculiar in that they may be significant and deserve to be studied in depth.

The distribution of the peak regions and peak value are slightly different despite having very similar distributions of optimal matched regions on the flanks of the translation initiation sites of the nine model organisms (see Figure 4). For *C. elegans*, the optimal matched regions between the flanks of the translation initiation site of mRNA and its corresponding introns are mainly located in the range of about −30 to 10 bp, and the peak value is about 1.4. The optimal matched regions of *D. melanogaster* are mainly located in the range of about −30 to 5 bp, and the peak value is about 2.0. The optimal matched regions of *A. thaliana* are mainly located in the range of about −30 to 10 bp, and the peak value is about 1.5. The optimal matched regions of *A. mellifera* are mainly located at about −10 bp, and the peak value is about 2.0. The optimal matched regions of *A. gambiae* are mainly located in the left range of about −30 to 5 bp, and the peak value is about 1.9. The optimal matched regions of *O. sativa* are mainly located at about −20 bp, and the peak value is about 1.2. The optimal matched regions of *D. rerio* are mainly located in the left range of about −30 bp, and the peak value is about 1.0. There is no distinctly optimal matched region for *M. musculus*. The optimal matched regions of *H. sapiens* are mainly located in the range of about −45 to −10 bp, and the peak value is about 1.3. These facts demonstrate that introns do interact with the translation initiation sites flanking their corresponding mRNAs.

3.3.2 Distribution of optimal matched regions of translation termination regions

Analytically, the distribution of optimal matched frequency of translation termination sites of the nine model organisms is revealed to be alike, that is, these model species show good universality for this distribution. The matched frequencies bounded by the translation termination sites are significantly altered; it is specifically manifested in the relative matched frequencies of the UTR on the right side of the translation termination sites, where it is generally higher. There is an excellent agreement between their distribution of optimal matched regions and the 3′-end of the mRNA. When comparing short introns, the distribution of their optimal matched region is consistent with that of long introns. Large differences in the distribution of optimal matched regions are

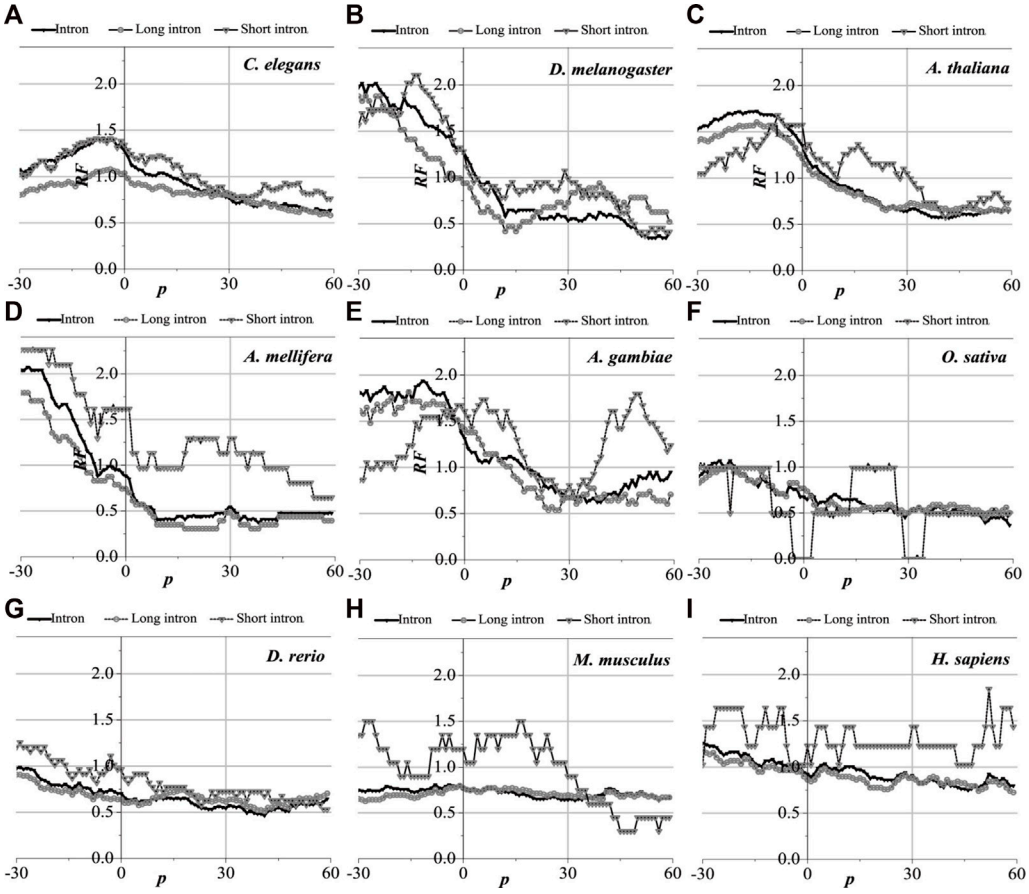

FIGURE 4
RF distributions around translation initiation site. The X-axis is the position of mRNA and the Y-axis represents the *RF* values. *RF* = 1 represents the average value of relative match frequencies theoretically.

observed in the short introns, but some distributions are very peculiar in that they may be significant and deserve to be studied in depth.

The distribution of the peak regions and peak value are slightly different despite being very similar for the distribution of optimal matched regions on the flanks of the translation termination sites of the nine model organisms (Figure 5). The optimal matched regions between the flanks of the translation termination sites of the *C. elegans* mRNA and their corresponding introns are mainly located in the right range of about −20 bp, and the peak value is about 4.5. The optimal matched regions of *D. melanogaster* are mainly located in the right range of about −10 bp, and the peak value is about 3.3. The optimal matched regions of *A. thaliana* are mainly located in the right range of about −20 bp, and the peak value is about 1.8. The optimal matched regions of *A. mellifera* are mainly located in the right range of about −20 bp, and the peak value is about 4.6. The optimal matched regions of *A. gambiae* are mainly located in the right range of about −15 bp, and the peak value is about 1.8. The optimal matched regions of *O. sativa* are mainly located in the range of about −30–60 bp, and the peak value is about 1.6. The optimal matched regions of *D. rerio* are mainly located in the right range of about −18 bp, and the peak value is about 2.6. The optimal matched regions of *M. musculus* are mainly located in the right range of about −10 bp, and the peak value is about 1.5. The optimal matched regions of *H. sapiens* are mainly located in the right range of about 16 bp, and the peak value is about 1.8. These facts demonstrate that introns do interact with the translation termination sites flanked by their corresponding mRNAs.

3.3.3 Distribution of optimal matched regions of exon–exon junction regions

The distribution of the optimal matched regions around the first exon junction sites, the last exon junction sites, and the middle junction sites of the nine model organisms is detected to be alike after analysis, that is, these model species show good universality for this distribution. It is specifically manifested in the generally low relative matched frequencies of the junction flanked by the first exon with the corresponding introns (Figure 6), the last exon junction sites (see Supplementary Figure S1), and the middle junction sites (Supplementary Figure S2). There is an excellent fit between their distributions of the optimal matched regions and CDS regions in all the exon–exon junction regions. When comparing short introns, the distributions of their optimal matched regions are consistent with those of long introns. For short introns, large differences in their

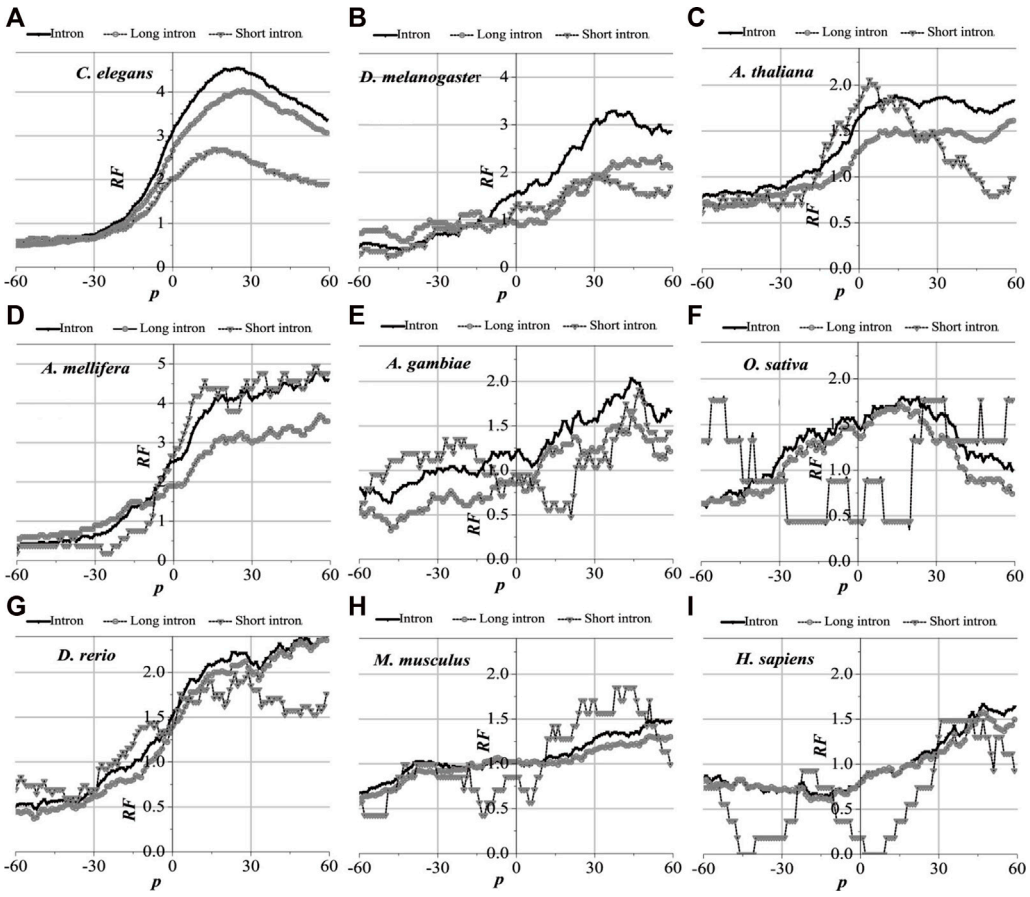

FIGURE 5
RF distributions around translation termination site. The X-axis is the position of mRNA and the Y-axis represents the *RF* values. *RF* = 1 represents the average value of relative match frequencies theoretically.

distributions of the optimal matched regions are observed, but some distributions are very peculiar in that they may be significant and deserve to be studied in depth.

3.4 Sequence feature of optimal matched segment of introns

The UTR regions of the mRNA preferentially interact with the introns, while the CDS regions of the mRNA are poorly matched to the introns. This is probably why the sequence features of the optimal matched segments are similar to the UTR features. GC content and second-order information redundancy D_2 of the optimal matched segment, CDS, 3′-UTR, and 5′-UTR are analyzed, and the correlation between these sequences is discussed. The results are shown in Figure 7 and Table 2.

3.4.1 GC content of optimal matched segment of intron

The distributions in Figure 6 are compared, and the GC content distributions of the optimal matched segments, CDS, 3′-UTR, and 5′-UTR in the nine model organisms are analyzed.

There are significant differences in the distribution center of the GC content in different sequences; however, the GC content of the optimal matched segments shows a special distribution pattern. In addition to having the lowest distribution center when compared to the other three types, the GC content distribution also has a very broad distribution range that almost completely encloses the distribution of the other sequences. It is shown that interactions between introns and mRNAare primarily based on weak bond binding, i.e., AT matching, but that high GC matching canalso occur. The average GC content of the optimal matched segments of introns in the nine model organisms is the closest to that of the 3′-UTR. The average GC content of the optimal matched segments, 3′-UTR, 5′-UTR, and CDS of *C. elegans*, *D. melanogaster*, *A. thaliana*, *A. gambiae*, *A. mellifera*, *D. rerio*, and *H. sapiens* increased gradually. The average GC content of the optimal matched segments, 3′-UTR, CDS, and 5′-UTR in *O. sativa* and *M. musculus* increased gradually. It is an interesting case that there are two clear peaks in the average GC content distributions of the optimal matched segments for *A. gambiae*, *M. musculus*, and *H. sapiens*, and these results show that there are abundant functional units in the introns. Based on the GC content analysis results of the above sequences, similar GC content values may lead to the mRNA

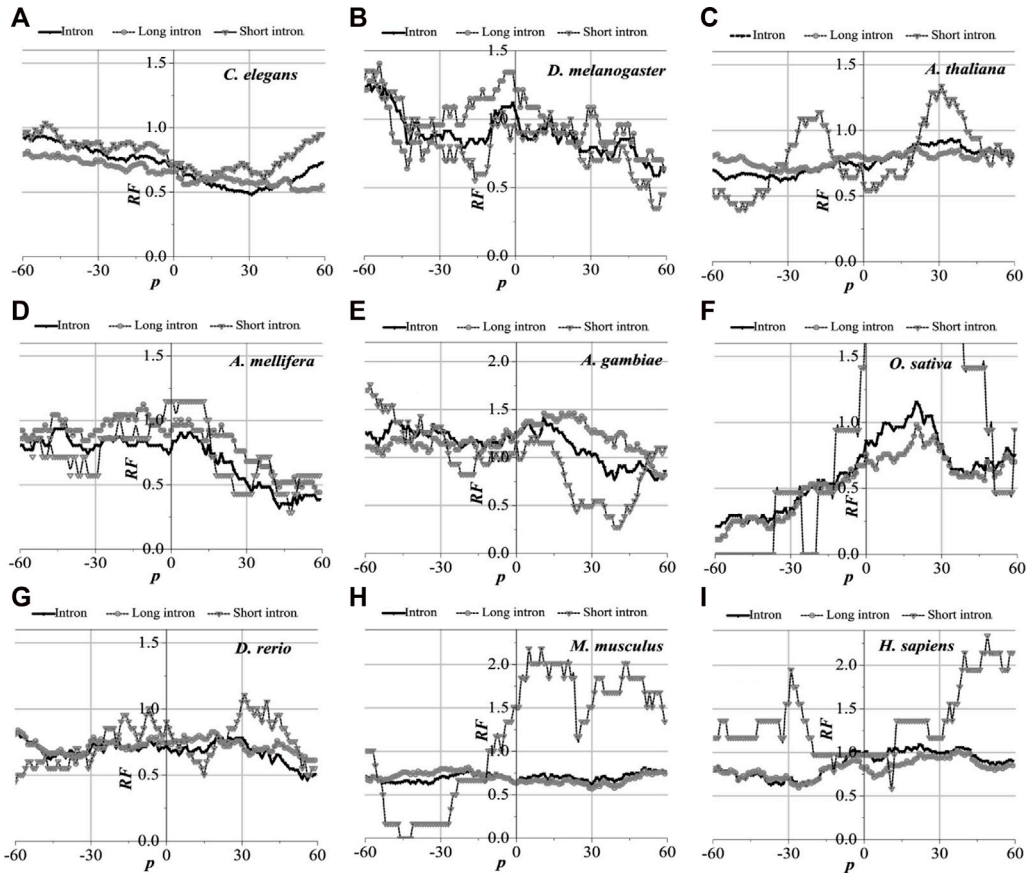

FIGURE 6
RF distributions around the first exon junction site. The X-axis is the position of mRNA and the Y-axis represents the *RF* values. *RF* = 1 represents the average value of relative match frequencies theoretically.

UTR preference for intron interaction, except for *O. sativa* and *M. musculus*.

3.4.2 Second-order information redundancy of optimal matched segments of introns

The intron of each gene is connected with the optimal matched segments of the mRNA, CDS, 3′-UTR (which includes 50 bp downstream of the translation initiation site), and 5′-UTR (which includes 50 bp upstream of the translation termination site) in a new line in sequence and are denoted as the optimal matched segments of the intron, CDS, 3′-UTR, and 5′-UTR. The results of the second-order information redundancy D_2 are shown in Table 2.

By comparing the D_2 values in Table 2, it is seen that the 3′-UTR D_2 values are the closest to those of the optimal matched intron segments in the nine model species. This agrees with the GC content of the optimal matched intron segments. This is one of the reasons why there are optimal matched frequent peaks in the regions of UTR, which is obvious, especially in the 3′-UTR.

4 Conclusion

At the genome-wide level, the optimal matched regions of the introns and their corresponding mRNAs for protein-coding genes in nine model organisms (such as *H. sapiens*) were analyzed. It was observed that the distribution of optimal matched frequencies in the mRNA sequence showed high consistency or universality among the nine model organisms. A peak distribution appeared in the untranslated regions (UTRs) of the mRNA, especially in the 3′-UTR, and the matched frequency in the coding sequence (CDS) was relatively low. It was discovered that introns, particularly the 3′-UTR, have a high preference for interacting with the UTR region of the mRNA. The function of the introns after splicing could be related to NMD. The matched frequencies bounded by functional sites in the sequences near the translation initiation site and translation termination site were different significantly, and the matched frequency of the junction region of the exon was relatively low. The distribution centers of the GC content in different sequences were different, but the GC content in the optimal matched segments showed a special distribution pattern. In addition to having a lower distribution center than the other three types, the GC content also had a very broad distribution range that almost completely enclosed the distribution of the other sequences. The results showed that the interactions between the introns and mRNAs were mainly dominated by weak bond binding, that is, not only AT matching but also in juggling high GC matching. In all nine

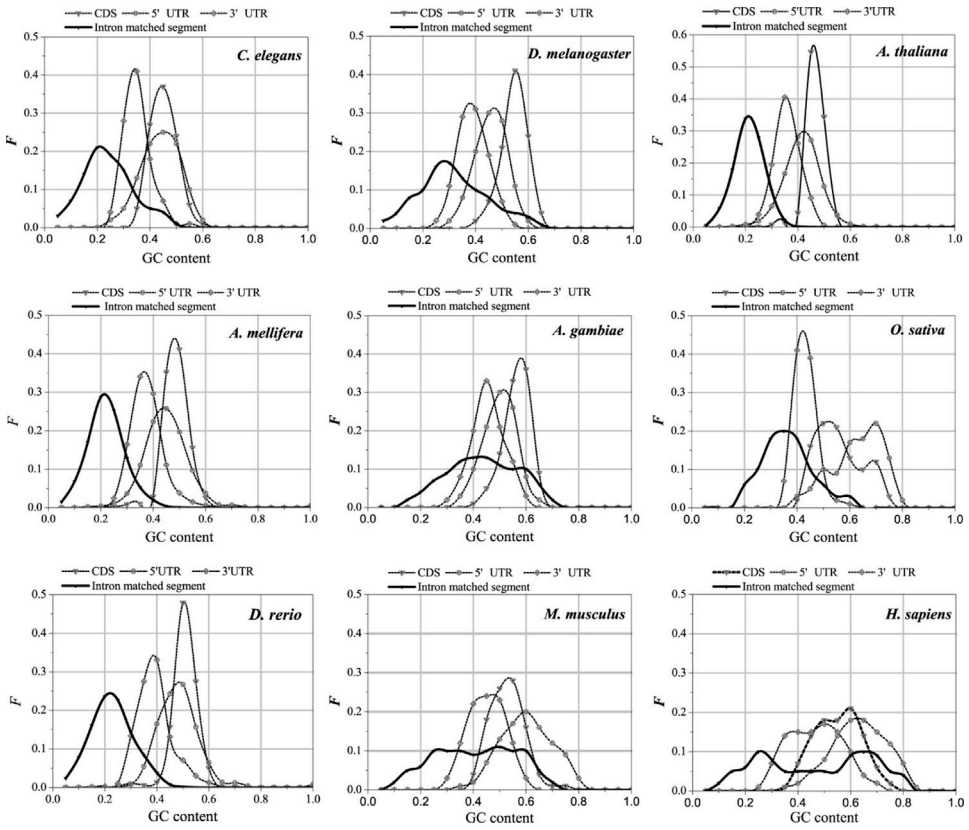

FIGURE 7
GC content distributions of different sequences. The X-axis is GC content and the Y-axis represents the Frequency values.

TABLE 2 D_2 for different sequences of nine eukaryotes.

	D_2			
	CDS	5′-UTR	3′-UTR	Intron matched segment
Caenorhabditis elegans	0.029	0.032	0.32	0.066
Drosophila melanogaster	0.015	0.023	0.009	0.010
Arabidopsis thaliana	0.018	0.026	0.012	0.010
Apis mellifera	0.019	0.021	0.009	0.012
Anopheles gambiae	0.014	0.025	0.010	0.011
Oryza sativa	0.016	0.015	0.015	0.011
Danio rerio	0.021	0.028	0.014	0.013
Mus musculus	0.042	0.031	0.040	0.046
Homo sapiens	0.041	0.028	0.043	0.063

species, a high degree of concordance of the distribution of the matched rate of the optimal matched segments was observed, primarily falling between 60% and 80%. In lower eukaryotes and higher eukaryotes, the most probable value of the optimal matched segment length distribution is around 20 bp and around 30 bp, respectively. These conclusions are in line with the results obtained in the ribonucleoprotein genes. Some peaks of the distribution of matched frequency are conserved for all organisms, and the results reveal the inherent mechanisms of the optimal matched segment composition.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding authors.

Author contributions

SB and QZ analyzed the sequence characteristics and manuscript writing. QS and ZL collected, sorted, and refined the data. YJ and GA jointly completed the algorithm optimization and manuscript writing. LW and CW supervised the entire study. ZL and XZ helped with the design of the experiments and revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by grants from the high-end talent training projects of Grassland Talents in Inner Mongolia Autonomous Region; the Natural Science Foundation of Inner Mongolia Autonomous Region (2021MS03063); the National Natural Science Foundation of China (31500677); and the Leading Talent Project of Science and Technology Leading Talent Team Project of Inner Mongolia Autonomous Region (2022LJRC0010).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, editors, and reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1151703/full#supplementary-material

References

Bianchi, M., Crinelli, R., Giacomini, E., Carloni, E., and Magnani, M. (2009). A potent enhancer element in the 5'-UTR intron is crucial for transcriptional regulation of the human ubiquitin C gene. *Gene* 448 (1), 88–101. doi:10.1016/j.gene.2009.08.013

Bo, S., Li, H., Zhang, Q., Lu, Z., and Bao, T. (2019). Potential relations between post-spliced introns and mature mRNAs in the Caenorhabditis elegans genome. *J. Theor. Biol.* 467, 7-14. doi:10.1016/j.jtbi.2019.01.031

Braddock, M., Muckenthaler, M., White, M. R., Thorburn, A. M., Sommerville, J., Kingsman, A. J., et al. (1994). Intron-less RNA injected into the nucleus of Xenopus oocytes accesses a regulated translation control pathway. *Nucleic Acids Res.* 22 (24), 5255–5264. doi:10.1093/nar/22.24.5255

Brinster, R. L., Allen, J. M., Behringer, R. R., Gelinas, R. E., and Palmiter, R. D. (1988). Introns increase transcriptional efficiency in transgenic mice. *Proc. Natl. Acad. Sci.* 85 (3), 836–840. doi:10.1073/pnas.85.3.836

Buchman, A. R., and Berg, P. (1988). Comparison of intron-dependent and intron-independent gene expression. *Mol. Cell. Biol.* 8, 4395–4405. doi:10.1128/mcb.8.10.4395-4405.1988

Buckley, P. T., Lee, M. T., Sul, J-Y., Miyashiro, K. Y., Bell, T. J., Fisher, S. A., et al. (2011). Cytoplasmic intron sequence-retaining transcripts can be dendritically targeted via ID element retrotransposons. *Neuron* 69 (5), 877–884. doi:10.1016/j.neuron.2011.02.028

Callis, J., Fromm, M., and Walbot, V. (1987). Introns increase gene expression in cultured maize cells. *Genes and Dev.* 1 (10), 1183–1200. doi:10.1101/gad.1.10.1183

Charital, Y. M., Haasteren, G., Massiha, A., Schlegel, W., and Fujita, T. (2009). A functional NF-kappaB enhancer element in the first intron contributes to the control of c-fos transcription. *Gene* 430 (1-2), 116–122. doi:10.1016/j.gene.2008.10.014

Christie, M., Croft, L. J., and Carroll, B. J. (2011). Intron splicing suppresses RNA silencing in Arabidopsis. *Plant J.* 68 (1), 159–167. doi:10.1111/j.1365-313X.2011.04676.x

Daehyun, B., and Phil, G. (2005). Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing[J]. *Proc. Natl. Acad. Sci. U. S. A.* 102 (36), 12813. doi:10.1073/pnas.0506139102

Duncker, B., Davies, P., and Walker, V. (1997). Introns boost transgene expression in *Drosophila melanogaster*. *Mol. General Genet. MGG* 254 (3), 291–296. doi:10.1007/s004380050418

Duret, L. (2001). Why do genes have introns Recombination might add a new piece to the puzzle [J]. *TRENDS Genet.* 17 (4), 172–175. doi:10.1016/s0168-9525(01)02236-3

Gozashti, L., Roy, S. W., Thornlow, B., Kramer, A., Ares, M., and Corbett-Detig, R. (2022). Transposable elements drive intron gain in diverse eukaryotes[J]. *Proc. Natl. Acad. Sci. U. S. A.* 119 (48), e2209766119. doi:10.1073/pnas.2209766119

Guigó, R., and Ullrich, S. (2020). Dynamic changes in intron retention are tightly associated with regulation of splicing factors and proliferative activity during B-cell development. *Nucleic Acids Res.* 48 (3), 1327–1340. doi:10.1093/nar/gkz1180

Halligan, D. L., and Keightley, P. D. (2006). Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Res.* 16 (7), 875–884. doi:10.1101/gr.5022906

Hansen, T. B., Wiklund, E. D., Bramsen, J. B., Villadsen, S. B., Statham, A. L., Clark, S. J., et al. (2011). miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA[J]. *EMBO J.* 30 (21), 4414. doi:10.1038/emboj.2011.359

Hir, H. L., Nott, A., and Moore, M. J. (2003). How introns influence and enhance eukaryotic gene expression. *Trends Biochem. Sci.* 28 (4), 215–220. doi:10.1016/S0968-0004(03)00052-5

Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., et al. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats[J]. *RNA*, 19. 141. doi:10.1261/rna.035667.112

Julia, S., Charles, G., Lincoln, W. P., Lacayo, N., and Brown, P. O. (2012). Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types[J]. *PloS one* 7 (2), e0030733. doi:10.1371/journal.pone.0030733

Ko, C. H., Brendel, V., Taylor, R. D., and Walbot, V. (1998). U-richness is a defining feature of plant introns and may function as an intron recognition signal in maize. *Plant Mol. Biol.* 36 (4), 573–583. doi:10.1023/a:1005932620374

Li, W. (1990). Mutual information functions versus correlation functions. *J. Stat. Phys.* 60 (5-6), 823–837. doi:10.1007/BF01025996

Luo (2004). *Theoretic-physical approach to molecular biology (1 ref)*. Shanghai Scientific and Technical Publishers.

Luo, L., and Hong, L. (1991). The statistical correlation of nucleotides in protein-coding DNA sequences. *Bull. Math. Biol.* 53 (3), 345–353.

Luo, M. J., and Reed, R. (1999). Splicing is required for rapid and efficient mRNA export in metazoans. *Proc. Natl. Acad. Sci.* 96 (26), 14937–14942. doi:10.1073/pnas.96.26.14937

Matsumoto, K., Wassarman, K. M., and Wolffe, A. P. (1998). Nuclear history of a pre-mRNA determines the translational activity of cytoplasmic mRNA. *EMBO J.* 17 (7), 2107–2121. doi:10.1093/emboj/17.7.2107

Mattick, J. S., and Gagen, M. J. (2001). The evolution of controlled multitasked gene networks: The role of introns and other noncoding RNAs in the development of complex organisms. *Mol. Biol. Evol.* 18 (9), 1611–1630. doi:10.1093/oxfordjournals.molbev.a003951

McKenzie, R. W., and Brennan, M. D. (1996). The two small introns of the Drosophila affinidisjuncta Adh gene are required for normal transcription. *Nucleic Acids Res.* 24 (18), 3635–3642. doi:10.1093/nar/24.18.3635

Nordin, A., Larsson, E., and Holmberg, M. (2012). The defective splicing caused by the ISCU intron mutation in patients with myopathy with lactic acidosis is repressed by PTBP1 but can be derepressed by IGF2BP1. *Hum. Mutat.* 33 (3), 467–470. doi:10.1002/humu.22002

Nott, A., Meislin, S. H., and Moore, M. J. (2003). A quantitative analysis of intron effects on mammalian gene expression. *Rna* 9 (5), 607–617. doi:10.1261/rna.5250403

Orphanides, G., and Reinberg, D. (2002). A unified theory of gene expression. *Cell* 108 (4), 439–451. doi:10.1016/s0092-8674(02)00655-4

Rafiq, M., Suen, C. K., Choudhury, N., Joannou, C. L., White, K. N., and Evans, R. W. (1997). Expression of recombinant human ceruloplasmin-an absolute requirement for splicing signals in the expression cassette. *FEBS Lett.* 407 (2), 132–136. doi:10.1016/s0014-5793(97)00325-6

Roy, S. W. (2003). Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain [J]. *Proc. Natl. Acad. Sci.* 100 (12), 7158–7162. doi:10.1073/pnas.1232297100

Ryu, W. S., and Mertz, J. E. (1989). Simian virus 40 late transcripts lacking excisable intervening sequences are defective in both stability in the nucleus and transport to the cytoplasm. *J. virology* 63 (10), 4386–4394. doi:10.1128/JVI.63.10.4386-4394.1989

Sebastian, M., Marvin, J., Antigoni, E., Torti, F., Krueger, J., Rybak, A., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency[J]. *Nature* 495 (7441), 333. doi:10.1038/nature11928

Stover, D. A., and Verrelli, B. C. (2010). Comparative vertebrate evolutionary analyses of type I collagen: Potential of COL1a1 gene structure and intron variation for common bone-related diseases. *Mol. Biol. Evol.* 28 (1), 533–542. doi:10.1093/molbev/msq221

Zhao, X., Li, H., and Bao, T. (2013). Analysis on the interaction between post-spliced introns and corresponding protein coding sequences in ribosomal protein genes. *J. Theor. Biol.* 328, 33–42. doi:10.1016/j.jtbi.2013.03.002

Zhang, Q., Li, H., Zhao, X., Zheng, Y., and Meng, H. (2016). Analysis on the preference for sequence matching between mRNA sequences and the corresponding introns in ribosomal protein genes. *J. Theor. Biol.* 392, 113–121. doi:10.1016/j.jtbi.2015.12.003

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 13 April 2023
DOI 10.3389/fmicb.2023.1170785

OPEN ACCESS

EDITED BY
Ettayapuram Ramaprasad Azhagiya Singam,
University of California, Berkeley, United States

REVIEWED BY
Vijaya Sundar Jeyaraj,
University of Illinois at Urbana-Champaign, United States
Liang Cheng,
Harbin Medical University, China
Hao Wu,
School of Software, Shandong University, China

*CORRESPONDENCE
Hasan Zulfiqar
✉ hasanzulfiqar@uestc.edu.cn
Zhao-Yue Zhang
✉ zyzhang@uestc.edu.cn
Fen Liu
✉ nmlf906@163.com

SPECIALTY SECTION
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 21 February 2023
ACCEPTED 17 March 2023
PUBLISHED 13 April 2023

CITATION
Zulfiqar H, Ahmed Z, Kissanga Grace-Mercure B, Hassan F, Zhang Z-Y and Liu F (2023) Computational prediction of promotors in *Agrobacterium tumefaciens* strain C58 by using the machine learning technique. *Front. Microbiol.* 14:1170785. doi: 10.3389/fmicb.2023.1170785

Computational prediction of promotors in *Agrobacterium tumefaciens* strain C58 by using the machine learning technique

Hasan Zulfiqar[1,2]*, Zahoor Ahmed[1], Bakanina Kissanga Grace-Mercure[2], Farwa Hassan[2], Zhao-Yue Zhang[2]* and Fen Liu[3]*

[1]Yangtze Delta Region Institute (Huzhou), University of Electronic Science and Technology of China, Huzhou, China, [2]School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China, [3]Department of Radiation Oncology, Peking University Cancer Hospital (Inner Mongolia Campus), Affiliated Cancer Hospital of Inner Mongolia Medical University, Inner Mongolia Cancer Hospital, Hohhot, China

Promotors are those genomic regions on the upstream of genes, which are bound by RNA polymerase for starting gene transcription. Because it is the most critical element of gene expression, the recognition of promoters is crucial to understand the regulation of gene expression. This study aimed to develop a machine learning-based model to predict promotors in *Agrobacterium tumefaciens* (*A. tumefaciens*) strain C58. In the model, promotor sequences were encoded by three different kinds of feature descriptors, namely, accumulated nucleotide frequency, *k*-mer nucleotide composition, and binary encodings. The obtained features were optimized by using correlation and the mRMR-based algorithm. These optimized features were inputted into a random forest (RF) classifier to discriminate promotor sequences from non-promotor sequences in *A. tumefaciens* strain C58. The examination of 10-fold cross-validation showed that the proposed model could yield an overall accuracy of 0.837. This model will provide help for the study of promoters in *A. tumefaciens* C58 strain.

KEYWORDS

prokaryotic promotors, feature extraction, *agrobacterium tumefaciens* strain C58, feature selection, algorithms

1. Introduction

Agrobacterium belongs to the family of ubiquitous gram-negative soil bacteria. Infectious strains of agrobacterium such as agrobacterium tumefaciens strain C58 cause hairy root and crown gall diseases in plants (Goodner et al., 2001). Promotors are the genomic regions upstream of a gene on DNA where transcription factor and RNA polymerase bind together to initiate gene transcription (Sawadogo and Roeder, 1985; Zhao et al., 2017; Zhang et al., 2018). The biological process of prokaryotic promotors is shown in Figure 1. The study of promoters is the first step to understanding gene expression.

Correct identification of the promotor sequence could produce vital signs for understanding its mechanism of the regulation (Cao et al., 2022; Li et al., 2022b). Currently, numerous tentative techniques, such as mass spectrometry (Flusberg et al., 2010), reduced-representation bisulfite sequencing (Doherty and Couldrey, 2014), and single-molecule real-time sequencing (Boch and Bonas, 2010), have been developed. Though these procedures are quite helpful in the identification of promotors prediction, they are

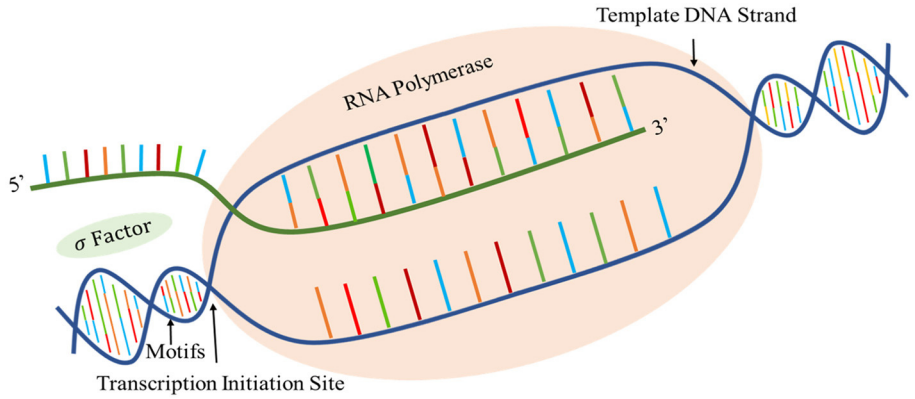

FIGURE 1
Schematic diagram of the prokaryotic promotor structure and its biological processes.

costly when applied to large sequencing data. Thus, a bioinformatics tool to recognize the promotor sequence is urgently needed. At present, some computational tools have been presented to recognize promotors in multiple species, such as PePPer (de Jong et al., 2012) for *Escherichia coli* (*E.coli*) and *Bacillus subtilis* (*B.subtilis*); Promotech for *Bacillus amyloliquefaciens* (*B. amyloliquefaciens*) XH_7 bacterium (Chevez-Guardado and Peña-Castillo, 2021); DeePromotors (Oubounyt et al., 2019) for TATA promotors (Zou et al., 2016) in eukaryotic genomes; iProEP (Lai et al., 2019) for *Homo sapiens* (*H. sapiens*), *Drosophila melanogaster* (*D. melanogaster*), *Caenorhabditis elegans* (*C. elegans*), *B. subtilis*, and *E. coli*; and iPromotor-2L (Liu et al., 2018) for bacterial promotors. However, there is no such model for *A. tumefaciens* C58 strain. To address the above-mentioned problems, we designed an RF-based model to predict promotor sequences in agrobacterium tumefaciens strain C58. Figure 2 illustrates the workflow of the projected model.

Accumulated nucleotide frequency, binary encodings, and *k*-mer nucleotide composition were utilized to convert sequences into numerical features, and then these features were optimized by using correlation and the mRMR-based feature selection algorithm. After this, these optimized features were inputted into a random forest classifier for the identification of promotor sequences on the basis of 10-fold cross-validation. As a result, an ideal model was attained.

2. Materials and methods

A precise and accurate dataset is necessary to establish a prediction model (Liang et al., 2017; Ning et al., 2021a,b; Su et al., 2021). Therefore, we obtained the experimentally verified *Agrobacterium tumefaciens* strain C58 promotors data of 706 sequences from PPD (http://lin-group.cn/database/ppd/index.php) and also collected negative data of 2860 sequences of 81 bp from (http://bioinformatics.hitsz.edu.cn/iPromotor-2L/data). Moreover, we divided the dataset into 80/20 ratios for training and testing the model.

2.1. Feature descriptors

Selecting the feature encodings that are useful and autonomous is a key stage in establishing machine learning-based models (Lv et al., 2021; Zhang D. et al., 2021; Ao et al., 2022a; Li et al., 2022a; Ning et al., 2022; Teng et al., 2022; Wei et al., 2022). Representing the DNA sequences with a mathematical manifestation is very important in functional element identification. Some DNA sequences coding strategies such as accumulated nucleotide frequency, physiochemical properties, binary encodings, nucleotide chemical properties and *k*-tuple nucleotide frequency component, nucleotide pair spectrum encoding, and natural vector have been applied in bioinformatics (Dao et al., 2020; Yang X. et al., 2021; Zhang Y. et al., 2021; Ao et al., 2022b; Ren et al., 2022). The performance of these feature descriptors was good. Here, to extract DNA sequence information as more as possible, accumulated nucleotide frequency, *k*-mer nucleotide composition, and binary encodings were presented to describe the DNA sequences based on their superior performance.

2.1.1. Accumulated nucleotide frequency

The encoding of ANF consists of the distribution and frequency of nucleotides n_i in the sequences. The nucleotide density D_i at any position in the sequence can be calculated as follows:

$$D_i = \frac{1}{|n_i|}\sum_{k=1}^{z} f(n_i), \quad f(g) = \begin{cases} 1 \text{ if } n_i = g \\ 0 \text{ in other case} \end{cases} \tag{1}$$

where z is the sequence length, n_i is the length of the string $\{n_1, n_2, \ldots, n_i\}$ (Li et al., 2022c,d) in the sequence, and $g \in \{A, G, C, T\}$.

2.1.2. k-mer nucleotide composition

k-mer nucleotide composition can reflect short-range nucleotide interaction of sequences (Salimi and Moeini, 2021; Zhang et al., 2022b; Dao et al., 2023). The nucleotide residues can be obtained *via* a sliding window method by setting the window

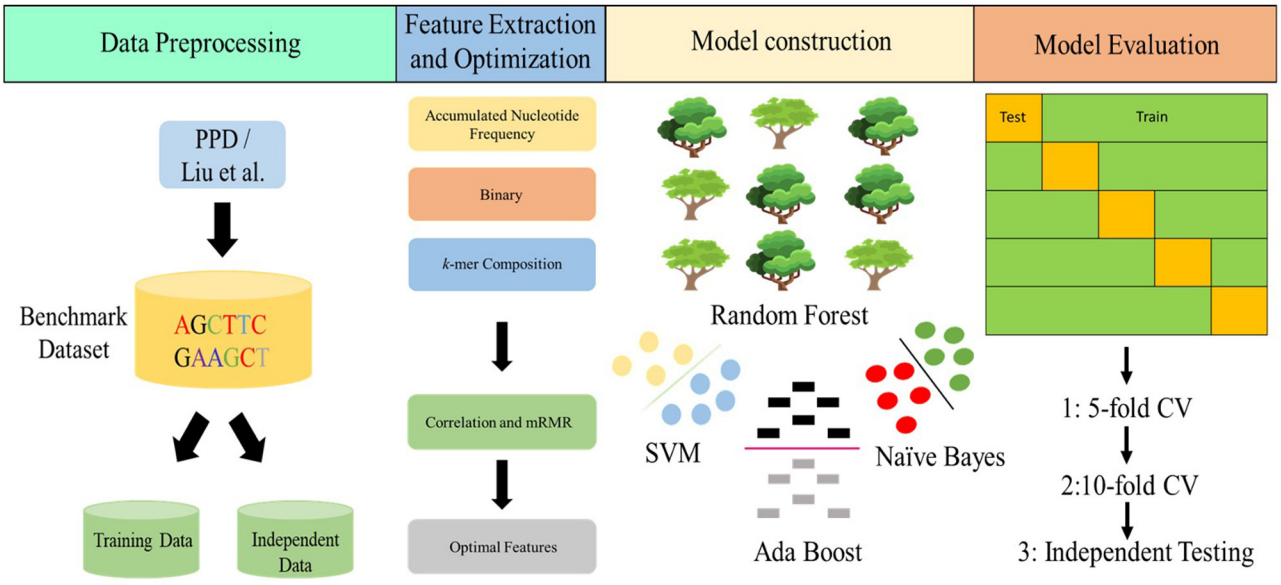

FIGURE 2
Overall workflow of the study.

size of k bp with a step size of 1 bp to examine a sequence with n bp. An arbitrary sample Z with the sequence length of n (where n is 81bp) can be characterized as

$$Z = Q_1\ Q_2\ Q_3 \ldots .. Q_i \ldots .. Q_{(n-1)}\ Q_n \tag{2}$$

where Q_i signifies the nucleotide {A, G, C, T} at the i-th position. The sequences can be transformed into the 4^k D vector using *k-mer* nucleotide composition as follows:

$$Q_k = [p_1^{k-tuple} p_2^{k-tuple} \ldots . p_i^{k-tuple} \ldots .. p_{4^k}^{k-tuple}]^t \tag{3}$$

where t denotes the transposition of the vector, and $p_1^{k-tuple}$ symbolizes the occurrence of the i-th *k-mer* nucleotide composition in the sequence. When $k = 1$, a DNA sample can be decoded into a 4 D vector $Q_1 = [p(A), p(T), p(G), p(C)]^t$. When $k = 2$, the DNA sample can be described by a 16-dimension vector. In this study, the value of k was set as 4 due to the best results. The whole results of k-mer nucleotide composition (k = 1,2,3,4,5,6) on training and independent data are shown in Supplementary Table S1.

2.1.3. Binary encoding

Encoding "0" and "1" can represent any information in the computational work (Zou et al., 2019). Therefore, we can directly convert a DNA sequence into a string of characters, which is consisted of "0" and "1." A = (1,0,0,0), T = (0,1,0,0), G = (0,0,1,0), and C = (0,0,0,1). Thus, a DNA sample of 81 bp length is converted into a 324 (4 × 81) dimension vector in this study.

2.2. Feature selection

2.2.1. Correlation

Feature selection is an important step for improving model performance (Dao et al., 2020). Correlation is a familiar comparison measure between two features. If two features are linearly dependent, then their correlation coefficient will be "±1." If the features are uncorrelated, the correlation coefficient will be "0." There are two comprehensive classes that can be used to measure the correlation between two random variables. One is based on information theory, and the other is classical linear correlation. The most familiar measure is the linear correlation coefficient. The linear correlation coefficient "d" for a pair of (m, n) variables is specified as

$$d = \frac{\sum (m_i - \overline{m}_i)(n_i - \overline{n}_i)}{\sqrt{\sum (m_i - \overline{m}_i)^2}\sqrt{\sum (n_i - \overline{n}_i)^2}} \tag{4}$$

Due to the expansion of the data, the correlation coefficient which is good for a sample may not produce decent outcomes for the whole population. Therefore, it is necessary to determine the significant association between the features, while captivating the whole population. The most commonly used method to examine statistical correlation is the t-test. The procedure used in the projected algorithm is to use the t-test for choosing the most important features from the whole feature set. The formula for calculating the suitable "T" value to test the consequence of a correlation coefficient employs the "T" distribution. The "T" value can be calculated as

$$T = d\sqrt{\frac{i-2}{1-d^2}} \tag{5}$$

where "*i*" is the number of instances and "*d*" is the correlation coefficient for sample data. The significance of the relationship is expressed in probability levels: p (e.g., significant at $p = 0.05$). The degrees of freedom for entering the T-distribution are $i - 2$. If the value of "T" is higher than the threshold value at the 0.05 significant level, then the feature will be significant and selected (Zulfiqar et al., 2022a).

2.2.2. mRMR

mRMR is a very popular feature selection technique, and it has been applied in many bioinformatics and biological applications (He et al., 2020; Zulfiqar et al., 2021b; Su et al., 2023). The compactness functions are described as "*i*" and "*y*," and their corresponding probabilities are $P(i)$ and $P(y)$. The common information between these two functions can be defined as

$$Q_{\min}(f_i, f_y) = \sum_{i \in Q} \sum_{y \in Y} P(fi, fy) \log \frac{P(i,y)}{P(i), P(y)}) \quad (6)$$

If the target is J_i, then calculating the mutual information in relation to the target and can be defined as

$$Q_{\max}(f_i, J_i) = \sum_{fi \in Q} \sum_{Ji \in i} P(fi, Ji) \log \frac{P(fi, Ji)}{P(fi), P(Ji)}) \quad (7)$$

Thus, $mRMR(f_i)$can be calculated as

$$mRMR(f_i) = \frac{Q_{max}(f_i, J_i)}{Q_{min}(f_i, f_y)} \quad (8)$$

2.3. Machine learning classifiers

Naïve Bayes (NB) classifier has been used widely in bioinformatics due to its simplicity (Ye et al., 2021). This classification method totally depends on the Bayes theorems. Ada boost (AB) is another popular machine learning technique. The main idea of AB is to set the classifiers' weights and trained the data in each and every iteration. The support vector machine (SVM) is also very famous and has been used in many bioinformatics and computational biology-related tools (Tao et al., 2020; Ahmed et al., 2022; Manavalan and Patra, 2022; Zou et al., 2022; Bupi et al., 2023; Zulfiqar et al., 2023). It is mostly used to perform binary classification. We implemented these algorithms in Weka version 3. 8.4. by using the default values. RF is a combined knowledge algorithm and is widely used in bioinformatics (Ao et al., 2022c; Zhang et al., 2023). The main idea of this is to combine several weak classifiers and outcomes generated on the basis of voting. The brief description is clearly described by Zulfiqar et al. (2021a). We have used randomized and grid search cross-validations to tune the hyperparameters. We executed this job in the Scikit-learn package version 0.22.2, and its parameters are summarized in Table 1. All experiments were carried out on a Windows operating system with 1.7 GHz intel quad-core i5.

TABLE 1 Best parameters of the proposed model.

Best parameters	
"N-estimators"	80
"Max_depth"	20
"Bootstrap"	True
"Min_samples_leaf"	1
"Min_samples_split"	2

Input: Training data: = H (x_1, x_2, ..., x_k, x_c)
Output: H_{best}
1st Round
1 Start
2 **for** i =1 to k **do**
3 d = calculate correlational coefficient (x_i, x_c)
end
4 let p = 0.05 significant level
5 let ρ = 0 / suppose there is no significant correlation between f_i and f_c
6 **for** i = 1 to k **do**
q = calculate the significance (d, ρ) for x_i / by using the *T-test*
7 **if** T > CV / critical value
8 H_{best} = H_{list}
9 **end**
10 return H_{best}
2nd Round
11 Start
12 By sorting the features
13 **for** each feature *fi* in *Z* **do**
14 By calculating the mutual information in relation to other features as
15 $Q_{\min}(f_i, f_y) = \sum_{i \in Q} \sum_{y \in Y} P(fi, fy) \log \frac{P(i,y)}{P(i), P(y)})$
16 By calculating the mutual information in relation to the target:
17 $Q_{\max}(f_i, J_i) = \sum_{fi \in Q} \sum_{Ji \in i} P(fi, Ji) \log \frac{P(fi, Ji)}{P(fi), P(Ji)})$
18 By calculating the $mRMR(f_i)$ as
19 $mRMR(f_i) = \frac{Q_{\max}(f_i, J_i)}{Q_{\min}(f_i, f_y)}$
20 **end**
21 **for** by sorting the features in descending order
22 By updating the matrix *Z'* with sorted features
23 **end**
24 return *Z'*

Algorithm 1. Correlation and mRMR-based Feature Selection Algorithm.

TABLE 2 Performance of models using different classifiers on the training and independent dataset.

	Training dataset							Independent dataset					
Classifier	FS	*k*	Method	Accuracy	Precision	Recall	F1	AUC	Accuracy	Precision	Recall	F1	AUC
AB	256	4	*k*-mer	0.761	0.772	0.761	0.791	0.812	0.775	0.820	0.801	0.798	0.881
	50	4	*k*-mer	0.799	0.802	0.785	0.789	0.856	0.787	0.824	0.799	0.805	0.872
	324		Binary	0.738	0.742	0.756	0.712	0.786	0.700	0.702	0.700	0.730	0.765
	48		Binary	0.745	0.742	0.698	0.789	0.820	0.720	0.732	0.702	0.726	0.789
	82		ANF	0.684	0.645	0.689	0.743	0.731	0.641	0.692	0.688	0.655	0.699
	38		ANF	0.743	0.726	0.775	0.746	0.796	0.696	0.702	0.698	0.710	0.756
	662		Fusion	0.745	0.732	0.785	0.775	0.799	0.720	0.732	0.775	0.745	0.774
	136		Fusion	0.778	0.768	0.792	0.800	0.845	0.738	0.745	0.765	0.725	0.806
SVM	256	4	*k*-mer	0.761	0.802	0.789	0.799	0.865	0.749	0.838	0.761	0.648	0.860
	50	4	*k*-mer	0.796	0.802	0.802	0.812	0.883	0.753	0.748	0.753	0.756	0.832
	324		Binary	0.744	0.747	0.778	0.765	0.792	0.725	0.755	0.760	0.763	0.786
	48		Binary	0.774	0.775	0.732	0.778	0.815	0.748	0.800	0.778	0.769	0.845
	82		ANF	0.666	0.697	0.732	0.705	0.766	0.612	0.623	0.633	0.605	0.699
	38		ANF	0.755	0.768	0.748	0.759	0.820	0.695	0.703	0.713	0.705	0.806
	662		Fusion	0.710	0.722	0.708	0.709	0.745	0.705	0.700	0.700	0.710	0.740
	136		Fusion	0.752	0.759	0.758	0.768	0.801	0.741	0.750	0.770	0.765	0.810
NB	256	4	*k*-mer	0.748	0.780	0.778	0.719	0.823	0.788	0.801	0.799	0.802	0.884
	50	4	*k*-mer	0.802	0.821	0.823	0.827	0.881	0.792	0.778	0.792	0.802	0.878
	324		Binary	0.737	0.775	0.765	0.789	0.794	0.776	0.770	0.778	0.793	0.835
	48		Binary	0.777	0.789	0.759	0.788	0.864	0.782	0.810	0.815	0.816	0.891
	82		ANF	0.675	0.689	0.720	0.696	0.756	0.665	0.685	0.691	0.701	0.741
	38		ANF	0.735	0.741	0.728	0.733	0.770	0.723	0.715	0.705	0.740	0.762
	662		Fusion	0.712	0.754	0.726	0.745	0.768	0.764	0.777	0.756	0.750	0.788
	136		Fusion	0.778	0.802	0.808	0.810	0.880	0.790	0.807	0.803	0.800	0.892
RF	256	4	*k-mer*	0.809	0.830	0.810	0.74	0.861	0.808	0.841	0.811	0.799	0.897
	50	4	*k*-mer	0.837	0.840	0.841	0.801	0.900	0.831	0.842	0.837	0.818	0.900
	324		Binary	0.792	0.632	0.792	0.701	0.842	0.784	0.804	0.808	0.788	0.887
	48		Binary	0.796	0.653	0.801	0.732	0.865	0.806	0.825	0.811	0.806	0.892
	82		ANF	0.791	0.630	0.791	0.702	0.850	0.788	0.803	0.773	0.778	0.878
	38		ANF	0.795	0.642	0.789	0.743	0.866	0.794	0.726	0.792	0.80	0.868
	662		Fusion	0.792	0.630	0.790	0.708	0.822	0.794	0.771	0.790	0.789	0.856
	136		Fusion	0.801	0.786	0.795	0.800	0.881	0.807	0.799	0.820	0.812	0.889

10.3389/fmicb.2023.1170785

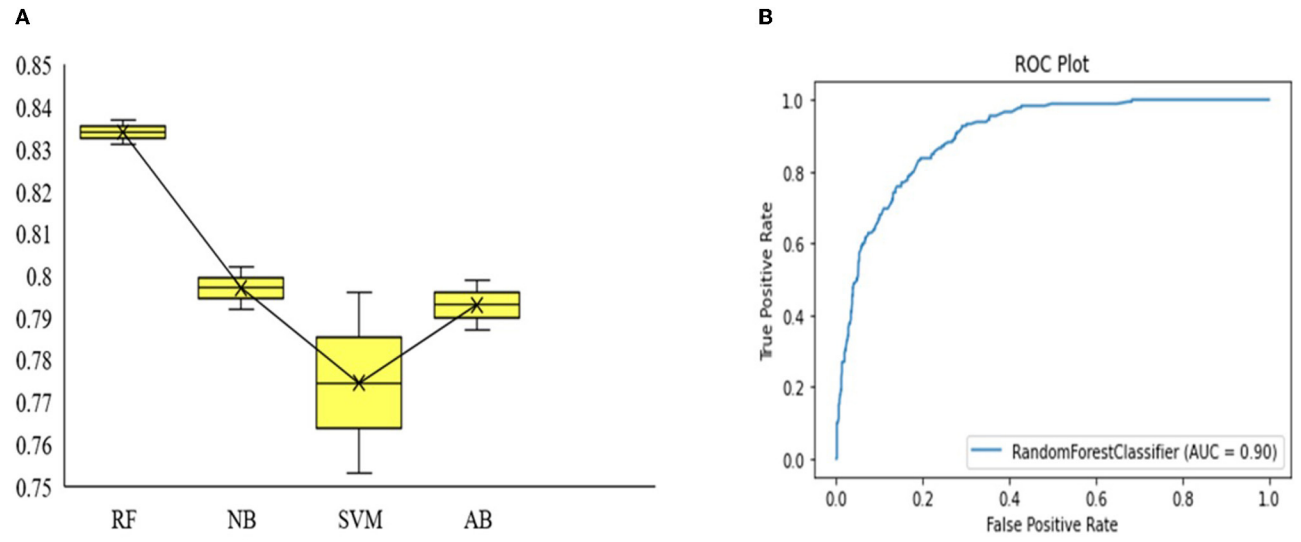

FIGURE 3
Performance comparison on different machine learning classifiers by using training and independent datasets. The higher point represents the training accuracy and the lower point represents the accuracy on independent data **(A).** *AUC* curve of the anticipated model **(B)**.

2.4. Evaluation metrics

Accuracy, precision, recall, and F1 (Hasan et al., 2020; Zhang et al., 2020; Wei et al., 2021b; Shoombuatong et al., 2022; Yang et al., 2022; Zulfiqar et al., 2022b) were employed to assess the performance of the prediction model and are expressed as

$$\begin{cases} Acc = \frac{tp + tn}{tp+fp + tn+fn} \\ Pre = \frac{tp}{tp +fp} \\ Rec = \frac{tp}{tp +fn} \\ F1 = 2 \times \frac{Pre \times Rec}{Pre + Rec} \end{cases} \quad (9)$$

where *tp* symbolizes the correctly predicted promotor sequences and *fp* signifies the non-promotor sequences classified as the promotor sequence. On the other hand, *tn* represents the correctly identified non-promotor sequences, and *fn* demonstrates the promotor sequences, which were classified as the non-promotor sequence.

3. Results and discussion

3.1. Performance evaluation

On the basis of sequence features, we constructed an anticipated model to recognize promotor sequences in *A. tumefaciens* C58 strain. First, the training data were converted into numerical feature vectors using accumulated nucleotide frequency, binary encodings, and *k*-mer nucleotide composition. After this, these features were optimized by using correlation and the mRMR-based algorithm. First, correlation measures and then mRMR were used to select the finest feature subset for the improved prediction outcomes. Afterward, these features were inputted into four machine learning methods. Cross-validation (CV) is a statistical analysis procedure and has been applied in machine learning to evaluate the model's performance (Yang H. et al., 2021; Chen et al., 2022; Liao et al., 2022; Xiao et al., 2022; Zhang et al., 2022a; Yang et al., 2023). In this study, the 10-fold CV test was used to investigate the performance of machine learning methods. In 10-fold CV, the benchmark dataset was randomly separated into ten groups of about equal size. Each group was individually tested by the model which trained with the remaining nine groups. Therefore, the 10-fold CV method was performed 10 times, and the average of the results was the final result (Charoenkwan et al., 2021; Wei et al., 2021a; Hasan et al., 2022). We have trained 32 models on AB, SVM, NB, and RF. At first, we used single encodings and their fusion to train and test the models, and then we optimized the feature encodings and their fusions by using correlation and the mRMR-based algorithm. In this phase, we utilized the *t*-test and picked the significant features by selecting the probability of the significance relation 0.05, and then used mRMR and picked the top features. Moreover, we inputted these features into AB, SVM, NB, and RF and found that the performance of *k*-mer was good as compared to other feature encodings and their fusion. The accuracy of *k*-mer in RF was 3.5%−4.1% higher than the other three classifiers. The *AUC* curve of the anticipated model was 0.900. The *accuracy, precision, recall, and F1* are recorded in Table 2. The performance comparison on different machine learning classifiers by using training and independent datasets and *ROC* plot of the anticipated model is shown in Figures 3A, B.

4. Conclusion

Promotors have a significant role in the transcription process because they are located on upstream of genes where RNA polymerase binds with the transcription factor and initiate the transcription. In this study, an RF model was established to

identify promotors sequences in agrobacterium tumefaciens strain C58. In the proposed model, sequences were encoded using accumulated nucleotide frequency, *k*-mer nucleotide composition, and binary encodings and then optimized with correlation and the mRMR-based algorithm. After this, these optimized features were inputted into the RF-based classifier using the 10-fold CV test and achieved the best model. The estimated outcomes on independent data showed that the projected model provided brilliant performance and oversimplification. We provided the source codes and data freely at https://github.com/linDing-groups/model_promotor. Researchers can yield good results for DNA sequences and recognize their roles by using our freely available source codes. In future, we will further improve the efficiency by using CNN/GNN and release a webserver to make our anticipated model more convenient for users without mathematical and programming knowledge.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

HZ: conceptualization, supervision, methodology, experimentation, visualization, and writing—original draft preparation. ZA and BK: data curation and methodology. FH: data curation. Z-YZ: supervision, methodology, reviewing, and editing. FL: supervision, reviewing, and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This study had been supported by the National Nature Scientific Foundation of China (62102067).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1170785/full#supplementary-material

References

Ahmed, Z., Zulfiqar, H., Tang, L., and Lin, H. (2022). A statistical analysis of the sequence and structure of thermophilic and non-thermophilic proteins. *Int. J. Mol. Sci.* 23, 10116. doi: 10.3390/ijms231710116

Ao, C., Jiao, S., Wang, Y., Yu, L., and Zou, Q. (2022a). Biological sequence classification: a review on data and general methods. *Research* 2022, 0011. doi: 10.34133/research.0011

Ao, C., Zou, Q., and Yu, L. (2022b). NmRF: identification of multispecies RNA 2'-O-methylation modification sites from RNA sequences. *Brief. Bioinform.* 23, bbab480. doi: 10.1093/bib/bbab480

Ao, C., Zou, Q., and Yu, L. (2022c). RFhy-m2G: Identification of RNA N2-methylguanosine modification sites based on random forest and hybrid features. *Methods* 203, 32–39. doi: 10.1016/j.ymeth.2021.05.016

Boch, J., and Bonas, U. (2010). Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Annu. Rev. Phytopathol.* 48, 419–436. doi: 10.1146/annurev-phyto-080508-081936

Bupi, N., Sangaraju, V. K., Phan, L. T., Lal, A., Vo, T. T. B., Ho, P. T., et al. (2023). An effective integrated machine learning framework for identifying severity of tomato yellow leaf curl virus and their experimental validation. *Research* 6, 0016. doi: 10.34133/research.0016

Cao, C., Wang, J. H., Kwok, D., Cui, F. F., Zhang, Z. L., Zhao, D., et al. (2022). webTWAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study. *Nucleic Acids Res.* 50, D1123–D1130. doi: 10.1093/nar/gkab957

Charoenkwan, P., Chiangjong, W., Nantasenamat, C., Hasan, M. M., Manavalan, B., and Shoombuatong, W. (2021). StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides. *Brief. Bioinform.* 22, bbab172. doi: 10.1093/bib/bbab172

Chen, H., Li, D., Liao, J., Wei, L., and Wei, L. (2022). MultiscaleDTA: a multiscale-based method with a self-attention mechanism for drug-target binding affinity prediction. *Methods* 207, 103–109. doi: 10.1016/j.ymeth.2022.09.006

Chevez-Guardado, R., and Peña-Castillo, L. (2021). Promotech: a general tool for bacterial promoter recognition. *Genome Biol.* 22, 1–16. doi: 10.1186/s13059-021-02514-9

Dao, F.-Y., Lv, H., Yang, Y.-H., Zulfiqar, H., Gao, H., and Lin, H. (2020). Computational identification of N6-methyladenosine sites in multiple tissues of mammals. *Comput. Struct. Biotechnol. J.* 18, 1084–1091. doi: 10.1016/j.csbj.2020.04.015

Dao, F. Y., Liu, M. L., Su, W., Lv, H., Zhang, Z. Y., Lin, H., et al. (2023). AcrPred: A hybrid optimization with enumerated machine learning algorithm to predict Anti-CRISPR proteins. *Int. J. Biol. Macromol.* 228, 706–714. doi: 10.1016/j.ijbiomac.2022.12.250

de Jong, A., Pietersma, H., Cordes, M., Kuipers, O. P., and Kok, J. (2012). PePPER: a webserver for prediction of prokaryote promoter elements and regulons. *BMC Genomics* 13, 1–10. doi: 10.1186/1471-2164-13-299

Doherty, R., and Couldrey, C. (2014). Exploring genome wide bisulfite sequencing for DNA methylation analysis in livestock: a technical assessment. *Front. Genet.* 5, 126. doi: 10.3389/fgene.2014.00126

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., et al. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* 7, 461. doi: 10.1038/nmeth.1459

Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Qurollo, B., et al. (2001). Genome sequence of the plant pathogen and biotechnology agent

Agrobacterium tumefaciens C58. *Science* 294, 2323–2328. doi: 10.1126/science.1066803

Hasan, M. M., Schaduangrat, N., Basith, S., Lee, G., Shoombuatong, W., and Manavalan, B. (2020). HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. *Bioinformatics* 36, 3350–3356. doi: 10.1093/bioinformatics/btaa160

Hasan, M. M., Tsukiyama, S., Cho, J. Y., Kurata, H., Alam, M. A., Liu, X., et al. (2022). Deepm5C: a deep-learning-based hybrid framework for identifying human RNA N5-methylcytosine sites using a stacking strategy. *Mol. Ther.* 30, 2856–2867. doi: 10.1016/j.ymthe.2022.05.001

He, S. D., Guo, F., Zou, Q., and Ding, H. (2020). MRMD2.0: a python tool for machine learning with feature ranking and reduction. *Curr. Bioinform.* 15, 1213–1221. doi: 10.2174/2212392XMTA2bMjko1

Lai, H.-Y., Zhang, Z.-Y., Su, Z.-D., Su, W., Ding, H., Chen, W., et al. (2019). iProEP: a computational predictor for predicting promoter. *Mol. Therapy Nucleic Acids* 17, 337–346. doi: 10.1016/j.omtn.2019.05.028

Li, H., Gong, Y., Liu, Y., Lin, H., and Wang, G. (2022a). Detection of transcription factors binding to methylated DNA by deep recurrent neural network. *Brief. Bioinform.* 23, bbab533. doi: 10.1093/bib/bbab533

Li, H., Shi, L., Gao, W., Zhang, Z., Zhang, L., Zhao, Y., et al. (2022b). dPromoter-XGBoost: detecting promoters and strength by combining multiple descriptors and feature selection using XGBoost. *Methods* 204, 215–222. doi: 10.1016/j.ymeth.2022.01.001

Li, Y., Qiao, G., Gao, X., and Wang, G. (2022c). Supervised graph co-contrastive learning for drug-target interaction prediction. *Bioinformatics* 38, 2847–2854. doi: 10.1093/bioinformatics/btac164

Li, Y., Qiao, G., Wang, K., and Wang, G. (2022d). Drug-target interaction predication via multi-channel graph neural networks. *Brief. Bioinform.* 23, bbab346. doi: 10.1093/bib/bbab346

Liang, Z. Y., Lai, H. Y., Yang, H., Zhang, C. J., Yang, H., Wei, H. H., et al. (2017). Pro54DB: a database for experimentally verified sigma-54 promoters. *Bioinformatics* 33, 467–469. doi: 10.1093/bioinformatics/btw630

Liao, J., Chen, H., Wei, L., and Wei, L. (2022). GSAML-DTA: an interpretable drug-target binding affinity prediction model based on graph neural networks with self-attention mechanism and mutual information. *Comput. Biol. Med.* 150, 106145. doi: 10.1016/j.compbiomed.2022.106145

Liu, B., Yang, F., Huang, D.-S., and Chou, K.-C. (2018). iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics* 34, 33–40. doi: 10.1093/bioinformatics/btx579

Lv, H., Dao, F.-Y., Zulfiqar, H., and Lin, H. (2021). DeepIPs: comprehensive assessment and computational identification of phosphorylation sites of SARS-CoV-2 infection using a deep learning-based approach. *Brief. Bioinform.* 22, bbab244. doi: 10.1093/bib/bbab244

Manavalan, B., and Patra, M. C. (2022). MLCPP 2.0: an updated cell-penetrating peptides and their uptake efficiency predictor. *J. Mol. Biol.* 434, 167604. doi: 10.1016/j.jmb.2022.167604

Ning, L., Abagna, H. B., Jiang, Q., Liu, S., and Huang, J. (2021a). Development and application of therapeutic antibodies against COVID-19. *Int. J. Biol. Sci.* 17, 1486–1496. doi: 10.7150/ijbs.59149

Ning, L., Cui, T., Zheng, B., Wang, N., Luo, J., Yang, B., et al. (2021b). MNDR v3.0: mammal ncRNA-disease repository with increased coverage and annotation. *Nucleic Acids Res.* 49, D160–d164. doi: 10.1093/nar/gkaa707

Ning, L., Liu, M., Gou, Y., Yang, Y., He, B., and Huang, J. (2022). Development and application of ribonucleic acid therapy strategies against COVID-19. *Int. J. Biol. Sci.* 18, 5070–5085. doi: 10.7150/ijbs.72706

Oubounyt, M., Louadi, Z., Tayara, H., and Chong, K. T. (2019). DeePromoter: robust promoter predictor using deep learning. *Front. Genet.* 10, 286. doi: 10.3389/fgene.2019.00286

Ren, L., Xu, Y., Ning, L., Pan, X., Li, Y., Zhao, Q., et al. (2022). TCM2COVID: A resource of anti-COVID-19 traditional Chinese medicine with effects and mechanisms. *iMETA* 1, e42. doi: 10.1002/imt2.42

Salimi, D., and Moeini, A. (2021). Incorporating *K-mers* highly correlated to epigenetic modifications for bayesian inference of gene interactions. *Curr. Bioinform.* 16, 484–492. doi: 10.2174/1574893615999200728193621

Sawadogo, M., and Roeder, R. G. (1985). Interaction of a gene-specific transcription factor with the adenovirus major late promoter upstream of the TATA box region. *Cell* 43, 165–175. doi: 10.1016/0092-8674(85)90021-2

Shoombuatong, W., Basith, S., Pitti, T., Lee, G., and Manavalan, B. (2022). THRONE: a new approach for accurate prediction of human RNA N7-methylguanosine sites. *J. Mol. Biol.* 434, 167549. doi: 10.1016/j.jmb.2022.167549

Su, W., Liu, M. L., Yang, Y. H., Wang, J. S., Li, S. H., Lv, H., et al. (2021). PPD: a manually curated database for experimentally verified prokaryotic promoters. *J. Mol. Biol.* 433, 166860. doi: 10.1016/j.jmb.2021.166860

Su, W., Xie, X. Q., Liu, X. W., Gao, D., Ma, C. Y., Zulfiqar, H., et al. (2023). iRNA-ac4C: a novel computational method for effectively detecting N4-acetylcytidine sites in human mRNA. *Int. J. Biol. Macromol.* 227, 1174–1181. doi: 10.1016/j.ijbiomac.2022.11.299

Tao, Z., Li, Y., Teng, Z., and Zhao, Y. (2020). A method for identifying vesicle transport proteins based on LibSVM and MRMD. *Comput. Math. Methods Med.* 2020, 8926750. doi: 10.1155/2020/8926750

Teng, Z., Zhao, Z., Li, Y., Tian, Z., Guo, M., Lu, Q., et al. (2022). i6mA-vote: cross-species identification of DNA N6-methyladenine sites in plant genomes based on ensemble learning with voting. *Front. Plant Sci.* 13, 845835. doi: 10.3389/fpls.2022.845835

Wei, L., He, W., Malik, A., Su, R., Cui, L., and Manavalan, B. (2021a). Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. *Brief. Bioinform.* 22, bbaa275. doi: 10.1093/bib/bbaa275

Wei, L., Ye, X., Sakurai, T., Mu, Z., and Wei, L. (2022). ToxIBTL: prediction of peptide toxicity based on information bottleneck and transfer learning. *Bioinformatics* 38, 1514–1524. doi: 10.1093/bioinformatics/btac006

Wei, L., Ye, X., Xue, Y., Sakurai, T., and Wei, L. (2021b). ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. *Brief. Bioinform.* 22, bbab041. doi: 10.1093/bib/bbab041

Xiao, J., Liu, M., Huang, Q., Sun, Z., Ning, L., Duan, J., et al. (2022). Analysis and modeling of myopia-related factors based on questionnaire survey. *Comput. Biol. Med.* 150, 106162. doi: 10.1016/j.compbiomed.2022.106162

Yang, H., Luo, Y., Ren, X., Wu, M., He, X., Peng, B., et al. (2021). Risk Prediction of Diabetes: Big data mining with fusion of multifarious physical examination indicators. *Inf. Fusion* 75, 140–149. doi: 10.1016/j.inffus.2021.02.015

Yang, K., Li, M., Yu, L., and He, X. (2023). Repositioning linifanib as a potent anti-necroptosis agent for sepsis. *bioRxiv* 9, 57. doi: 10.1038/s41420-023-01351-y

Yang, X., Ye, X., Li, X., and Wei, L. (2021). Idna-mt: identification DNA modification sites in multiple species by using multi-task learning based a neural network tool. *Front. Genet.* 12, 663572. doi: 10.3389/fgene.2021.663572

Yang, Y., Gao, D., Xie, X., Qin, J., Li, J., Lin, H., et al. (2022). DeepIDC: a prediction framework of injectable drug combination based on heterogeneous information and deep learning. *Clin. Pharmacokinet.* 61, 1749–1759. doi: 10.1007/s40262-022-01180-9

Ye, S., Liang, Y., and Zhang, B. (2021). Bayesian functional mixed-effects models with grouped smoothness for analyzing time-course gene expression data. *Curr. Bioinform.* 16, 2–12. doi: 10.2174/1574893615999200520082636

Zhang, D., Chen, H. D., Zulfiqar, H., Yuan, S. S., Huang, Q. L., Zhang, Z. Y., et al. (2021). iBLP: an XGBoost-based predictor for identifying bioluminescent proteins. *Comput. Math. Methods. Med.* 2021, 6664362. doi: 10.1155/2021/6664362

Zhang, S., Wang, Y., Gu, Y., Zhu, J., Ci, C., Guo, Z., et al. (2018). Specific breast cancer prognosis-subtype distinctions based on DNA methylation patterns. *Mol. Oncol.* 12, 1047–1060. doi: 10.1002/1878-0261.12309

Zhang, Y., Liu, T., Hu, X., Wang, M., Wang, J., Zou, B., et al. (2021). CellCall: integrating paired ligand-receptor and transcription factor activities for cell-cell communication. *Nucleic Acids Res.* 49, 8520–8534. doi: 10.1093/nar/gkab638

Zhang, Y.-F., Wang, Y.-H., Gu, Z.-F., Pan, X.-R., Li, J., Ding, H., et al. (2023). Bitter-RF: a random forest machine model for recognizing bitter peptides. *Front. Med.* 10, 1052923. doi: 10.3389/fmed.2023.1052923

Zhang, Z. M., Wang, J. S., Zulfiqar, H., Lv, H., Dao, F. Y., and Lin, H. (2020). Early diagnosis of pancreatic ductal adenocarcinoma by combining relative expression orderings with machine-learning method. *Front. Cell Dev. Biol.* 8, 582864. doi: 10.3389/fcell.2020.582864

Zhang, Z. Y., Ning, L., Ye, X., Yang, Y. H., Futamura, Y., Sakurai, T., et al. (2022a). iLoc-miRNA: extracellular/intracellular miRNA prediction using deep BiLSTM with attention mechanism. *Brief. Bioinform.* 23, bbac395. doi: 10.1093/bib/bbac395

Zhang, Z. Y., Sun, Z.-J., Yang, Y.-H., and Lin, H. (2022b). Towards a better prediction of subcellular location of long non-coding RNA. *Front. Comput. Sci.* 16, 165903. doi: 10.1007/s11704-021-1015-3

Zhao, Y., Wang, F., Chen, S., Wan, J., and Wang, G. (2017). Methods of MicroRNA promoter prediction and transcription factor mediated regulatory network. *Biomed. Res. Int.* 2017, 7049406. doi: 10.1155/2017/7049406

Zou, Q., Wan, S., Ju, Y., Tang, J., and Zeng, X. (2016). Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.* 10, 114. doi: 10.1186/s12918-016-0353-5

Zou, Q., Xing, P. W., Wei, L. Y., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N-6-methyladenosine sites from mRNA. *RNA* 25, 205–218. doi: 10.1261/rna.069112.118

Zou, Y., Ding, Y. J., Peng, L., and Zou, Q. (2022). FTWSVM-SR: DNA-binding proteins identification via fuzzy twin support vector machines on self-representation. *Interdisc. Sci. Comput. Life Sci.* 14, 372–384. doi: 10.1007/s12539-021-00489-6

Zulfiqar, H., Guo, Z., Grace-Mercure, B. K., Zhang, Z. Y., Gao, H., Lin, H., et al. (2023). Empirical comparison and recent advances of computational prediction

of hormone binding proteins using machine learning methods. *Comput. Struct. Biotechnol. J.* 21, 2253–2261. doi: 10.1016/j.csbj.2023.03.024

Zulfiqar, H., Huang, Q.-L., Lv, H., Sun, Z.-J., Dao, F.-Y., and Lin, H. (2022a). Deep-4mCGP: a deep learning approach to predict 4mC sites in Geobacter pickeringii by using correlation-based feature selection technique. *Int. J. Mol. Sci.* 23, 1251. doi: 10.3390/ijms23031251

Zulfiqar, H., Khan, R. S., Hassan, F., Hippe, K., Hunt, C., Ding, H., et al. (2021a). Computational identification of N4-methylcytosine sites in the mouse genome with machine-learning method. *Math. Biosci. Eng.* 18, 3348–3363. doi: 10.3934/mbe.2021167

Zulfiqar, H., Sun, Z.-J., Huang, Q.-L., Yuan, S.-S., Lv, H., Dao, F.-Y., et al. (2022b). Deep-4mCW2V: A sequence-based predictor to identify N4-methylcytosine sites in *Escherichia coli*. *Methods* 203, 558–563. doi: 10.1016/j.ymeth.2021.07.011

Zulfiqar, H., Yuan, S.-S., Huang, Q.-L., Sun, Z.-J., Dao, F.-Y., Yu, X.-L., et al. (2021b). Identification of cyclin protein using gradient boost decision tree algorithm. *Comput. Struct. Biotechnol. J.* 19, 4123–4131. doi: 10.1016/j.csbj.2021.07.013

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 05 May 2023
DOI 10.3389/fmicb.2023.1200678

OPEN ACCESS

EDITED BY
Yongchun Zuo,
Inner Mongolia University, China

REVIEWED BY
Yaser Daanial Khan,
University of Management and Technology, Lahore, Pakistan
Wei Chen,
Chengdu University of Traditional Chinese Medicine, China

*CORRESPONDENCE
Yan Lin
✉ Linyan936@163.com
Hasan Zulfiqar
✉ hasanzulfiqar@uestc.edu.cn
Hongyan Lai
✉ laihy@cqupt.edu.cn

†These authors have contributed equally to this work

RECEIVED 05 April 2023
ACCEPTED 18 April 2023
PUBLISHED 05 May 2023

CITATION
Lin Y, Sun M, Zhang J, Li M, Yang K, Wu C, Zulfiqar H and Lai H (2023) Computational identification of promoters in *Klebsiella aerogenes* by using support vector machine. *Front. Microbiol.* 14:1200678. doi: 10.3389/fmicb.2023.1200678

Computational identification of promoters in *Klebsiella aerogenes* by using support vector machine

Yan Lin[1*†], Meili Sun[2†], Junjie Zhang[1], Mingyan Li[3], Keli Yang[4], Chengyan Wu[5], Hasan Zulfiqar[6*] and Hongyan Lai[7*]

[1]Key Laboratory for Animal Disease-Resistance Nutrition of the Ministry of Agriculture, Animal Nutrition Institute, Sichuan Agricultural University, Chengdu, China, [2]Beidahuang Industry Group General Hospital, Harbin, China, [3]Chifeng Product Quality Inspection and Testing Centre, Chifeng, China, [4]Nonlinear Research Institute, Baoji University of Arts and Sciences, Baoji, China, [5]Baotou Teacher's College, Inner Mongolia University of Science and Technology, Baotou, China, [6]Yangtze Delta Region Institute (Huzhou), University of Electronic Science and Technology of China, Huzhou, Zhejiang, China, [7]Chongqing Key Laboratory of Big Data for Bio Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, China

Promoters are the basic functional cis-elements to which RNA polymerase binds to initiate the process of gene transcription. Comprehensive understanding gene expression and regulation depends on the precise identification of promoters, as they are the most important component of gene expression. This study aimed to develop a machine learning-based model to predict promoters in *Klebsiella aerogenes* (*K. aerogenes*). In the prediction model, the promoter sequences in *K. aerogenes* genome were encoded by pseudo *k*-tuple nucleotide composition (PseKNC) and position-correlation scoring function (PCSF). Numerical features were obtained and then optimized using mRMR by combining with support vector machine (SVM) and 5-fold cross-validation (CV). Subsequently, these optimized features were inputted into SVM-based classifier to discriminate promoter sequences from non-promoter sequences in *K. aerogenes*. Results of 10-fold CV showed that the model could yield the overall accuracy of 96.0% and the area under the ROC curve (AUC) of 0.990. We hope that this model will provide help for the study of promoter and gene regulation in *K. aerogenes*.

KEYWORDS

promoter, pseudo k-tuple nucleotide composition, position-correlation scoring function, feature selection, support vector machine

1. Introduction

Klebsiella aerogenes (*K. aerogenes*) is a ubiquitous Gram-negative bacterium found in a variety of environments, such as soil, sewage, mammalian gastrointestinal tract et al. The *K. aerogenes* can also colonize in human gut and most community-or hospital-acquired bloodstream infections are caused by this common multi-drug resistant pathogen, which is a source of opportunistic infections. Although most of these bacteria are sensitive to the antibiotics targeting them, the drug resistance still exists, and the induced resistance mechanisms are complex (Price and Sleigh, 1970). Promoters are the genomic regions upstream of genes, where RNA polymerase and other transcription factors bind together to initiate genes transcription (Sawadogo and Roeder, 1985). Thus, promoter identification is the first step to understand gene expression mechanism. Thus, a precise identification of promoter sequence could generate dynamic signs for understanding its mechanism of regulation (Zuo and Li, 2010).

In fact, several experimental methods, such as mass spectrometry (Flusberg et al., 2010), reduced-representation bisulfite sequencing (Doherty and Couldrey, 2014), and single-molecule real-time sequencing (Boch and Bonas, 2010), have been developed to recognize promoters. Although these methods are relatively helpful in the identification of promoters, they are exorbitant when implemented to large sequencing data (Hu et al., 2022a). Therefore, a bioinformatics tool to identify promoter sequence is instantly needed.

At present, some machine learning-based methods have been presented to predict promoters in multiple species (Ao et al., 2022). Li and Lin have ever designed a position weight matrix (PWM) method to identify sigma70 promoters in *Escherichia coli* (*E. coli*) (Li and Lin, 2006). Subsequently, they developed a hybrid approach (called IPMD) to identify eukaryotic and prokaryotic promoters (Lin and Li, 2011). PePPER is another webserver for recognizing prokaryote promoter elements and regulons (de Jong et al., 2012). In 2014, Lin et al. proposed a first model called iPro54-PseKNC to predict sigma54 promoters in prokaryotes (Lin et al., 2014). Liu et al. established a friendly tool called iPromoter-2l for the prediction of bacterial promotors. These works mainly used sequence composition to perform prediction. By using Z-curve theory, the bacterial promoters could also be formulated and predicted (Song, 2012; Lin et al., 2019). Combining various of sequence information, Lai et al. built a powerful model named iProEP for the identification of promoters in three kinds of eukaryotes and two kinds of bacteria (Lai et al., 2019). Chevez-Guardado designed a general tool (Promotech) for bacterial promoter recognition (Chevez-Guardado and Peña-Castillo, 2021). Recently, the promoters in two prokaryotes: *Corynebacterium glutamicum* and *Agrobacterium Tumefaciens Strain C58* were studied by using machine learning based models (Zulfiqar et al., 1011; Li et al., 2023). Among them, the sigma70 promoter is the most extensively studied in prokaryotes (Patiyal et al., 2022). iProm-phage is a two-layer model for phage promoters and their types prediction (Shujaat et al., 2022).

Although there are already many prediction models for prokaryotic promoters, due to species specificity and prediction performance limitations, there is a need for trainning more specific promoter prediction models for *K. aerogenes* (Hu et al., 2022b). Thus, in this paper, we designed a SVM-based model to predict the promoters of *K. aerogenes*. The Figure 1 illustrates the workflow of this project, mainly including the core content and key steps. Thereinto, two feature extraction methods, namely PseKNC and PCSF, were employed to convert DNA sequences into numerical features. And then these features were optimized by using mRMR feature selection algorithm based on SVM machine learning model and 5-fold CV. Moreover, the selected optimal feature subset was applied to train a SVM classifier for identifying *K. aerogenes* promoter sequences on the basis of 10-fold CV. As a result, an ideal model with prediction accuracy and AUC of 96.0% and 0.990 was attained.

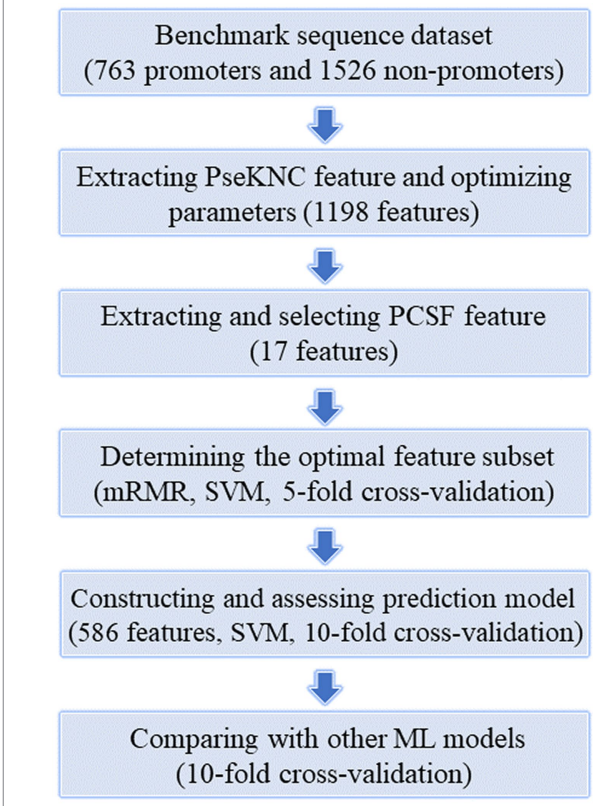

FIGURE 1
The overall workflow of this study.

2. Materials and methods

2.1. Data collection and preprocessing

The construction of a prokaryotic promoter dataset is crucial for obtaining a good promoter model. Prokaryotic Promoter Database (PDD, http://lin-group.cn/database/ppd/) developed by Lin et al. contains comprehensive information on experimentally verified promoters of numerous prokaryotic species and can be freely accessed (Su et al., 2021). The sequence data of 763 *K. aerogenes* promoters were downloaded from the database and defined as positive dataset. Each promoter sequence was composed of 81 nucleotides, including transcription start site (TSS) (namely the 0-th site), upstream 20 bp and downstream 60 bp regions of TSS. In order to generate a reliable negative dataset, we firstly extracted the convergent intergenic (length greater than 81 bp) and coding (length greater than 2000 bp) regions from *K. aerogenes* genome. Secondly, sliding window method with step of 1 bp was applied to generate convergent intergenic and coding sequences, with length of 81 bp. Then, we used CD-HIT program to estimate the sequence similarity of convergent intergenic and coding sequences, and filtered highly similar sequences by setting cutoff value as 0.8. Finally, 763 convergent intergenic sequences and 763 coding sequences were randomly picked out and regarded as negative dataset.

2.2. Feature extraction

Referring to the well-designed eukaryotic and prokaryotic promoter identification tool, iProEP,[1] we also adopted two algorithms, including pseudo k-tuple nucleotide composition (PseKNC) and position-correlation scoring function (PCSF), to transform raw promoter/non-promoter sequence data into suitable numeric features for modeling.

1 http://lin-group.cn/server/iProEP/

In this study, the type II PseKNC method was used to transform each nucleotide sequence into a feature vector of $4^k+\lambda\Lambda$ dimensions (Tang et al., 2021),

$$D_{pseKNC}=\left[d_1\ d_2\cdots d_{4^k}\ d_{4^k+1}\cdots d_{4^k+\lambda}\ d_{4^k+\lambda+1}\cdots d_{4^k+\lambda\Lambda}\right]^T \quad (1)$$

where k means k-tuple nucleotide component, λ is an integer less than $L-k$ (L denotes the length of a DNA sequence). And Λ is the number of physicochemical properties, the value of which is 6 corresponding to the six types of DNA local structural properties included in this work. Each element in D_{pseKNC} is defines as:

$$d_u=\begin{cases}\dfrac{f_u^{k-tuple}}{\sum_{i=1}^{4^k}f_i^{k-tuple}+\omega\sum_{j=1}^{\lambda\Lambda}\tau_j},\left(1\le u\le 4^k\right)\\ \dfrac{\omega\tau_{u-4^k}}{\sum_{i=1}^{4^k}f_i^{k-tuple}+\omega\sum_{j=1}^{\lambda\Lambda}\tau_j},\left(4^k+1\le u\le 4^k+\lambda\Lambda\right)\end{cases} \quad (2)$$

The former 4^k elements are nucleotide composition features, which can reflect local or short-range sequence-order information. The latter $\lambda\Lambda$ factors are pseudo nucleotide composition features corresponding to global or long-range effect. In equation (2), $f_i^{k-tuple}$ represents the normalized frequency of occurrence of the i-th k-tuple nucleotides in the sample sequence. The weight factor ω can adjust the effects of nucleotide composition and local structural properties of DNA. And τ_j indicates the m-tier correlation factor and is formulated with the form of equation (3), the value of which corresponds to the sequence-order correlation between all the m-tier contiguous k-tuple nucleotide component along a promoter/non-promoter sequence.

$$\begin{cases}\tau_1=\dfrac{1}{L-k}\sum_{i=1}^{L-k}J_{i,i+1}^1\\ \tau_2=\dfrac{1}{L-k}\sum_{i=1}^{L-k}J_{i,i+1}^2\\ \cdots\cdots\\ \tau_\Lambda=\dfrac{1}{L-k}\sum_{i=1}^{L-k}J_{i,i+1}^\Lambda\ \lambda<(L-k)\\ \cdots\cdots\\ \tau_{\lambda\Lambda-1}=\dfrac{1}{L-k-\lambda+1}\sum_{i=1}^{L-k-\lambda+1}J_{i,i+1}^{\lambda\Lambda-1}\\ \tau_{\lambda\Lambda}=\dfrac{1}{L-k-\lambda+1}\sum_{i=1}^{L-k-\lambda+1}J_{i,i+1}^{\lambda\Lambda}\end{cases} \quad (3)$$

where

$$\begin{cases}J_{i,i+m}^\xi=H_\xi\left(R_iR_{i+1}\right)\cdot H_\xi\left(R_{i+m}R_{i+m+1}\right)\\ \xi=1,2,\cdots,\Lambda;m=1,2,\cdots,\lambda;i=1,2,\cdots,L-\lambda-1\end{cases} \quad (4)$$

where $H_\xi\left(R_iR_{i+1}\right)$ is the standardized value of the ξ-th DNA local structural properties for the dinucleotide R_iR_{i+1} at position i. The original values of these physicochemical properties are provided by Goñi et al. (2008) and the standardization approach are the same as previously described in iProEP. In addition, the processes of Position-Correlation Scoring Matrix (PCSM) construction and PCSF feature transformation and selection are directly referring to the *E. coli* model in iProEP.

2.3. mRMR

mRMR is a well-known feature selection method and has been used in many computational and biological applications (Zulfiqar et al., 2021; Su et al., 2023). The density functions are described as 'i' and 'y' and their corresponding probabilities are $P(i)$ and $P(y)$. The common information between these two functions can be demarcated as

$$Z_{\min}\left(M_i,M_y\right)=\sum_{i\in Z}\sum_{y\in Y}P\left(Mi,My\right)\log\frac{P(i,y)}{P(i),P(y)} \quad (5)$$

If the target is J_i then calculating the mutual information in relation to the target and can be defined as

$$Z_{\max}\left(M_i,J_i\right)=\sum_{Mi\in Z}\sum_{Ji\in i}P\left(Mi,Ji\right)\log\frac{P(Mi,Ji)}{P(Mi),P(Ji)} \quad (6)$$

So, calculating the *mRMR* as (M_i)

$$\mathrm{mRMR}\left(M_i\right)=\frac{Z_{\max}\left(M_i,J_i\right)}{Z_{\min}\left(M_i,f_y\right)} \quad (7)$$

2.4. Machine learning classifiers

SVM is a well-known classifier and has been utilized in many bioinformatics and computational biology related tools (Basith et al., 2021; Arif et al., 2022; Basith et al., 2022; Bupi et al., 2023; Dao et al., 2023). It is typically used to perform binary classification. Ada boost (AB) is another famous classifier (Wang et al., 2021). The main idea of AB is to set the classifiers weights and trained the data in each and every iteration. Naïve Bayes (NB) classifier has been widely used in bioinformatics due to its simplicity (Naseer et al., 2022; Zulfiqar et al., 2022). This classification method totally depends on the Bayes theorems. Random Forest (RF) is a collective knowledge algorithm and broadly used in bioinformatics (Zhu et al., 2022; Zhang et al., 2023). The main idea of this is to unite multiple weak classifiers and outcome generated on the basis of voting (Zulfiqar et al., 2023). The brief description is clearly described in (Zulfiqar et al., 2021). The k-nearest neighbor (KNN) is a non-parametric and supervised learning classifier, which uses vicinity to make classifications about the grouping of an individual data point. Logistic Regression (LR) is a classification algorithm and used when the value of the target variable is categorical in nature

(Yang et al., 2021). We have executed these algorithms in Weka version 3.8.4. by using the default values.

2.5. Evaluation metrics

Accuracy, sensitivity, specificity (Cao et al., 2017; Tang et al., 2022; Yang et al., 2022; Zhang et al., 2022; Chen et al., 2023) were utilized to evaluate the performance of the prediction model and termed as

$$\begin{cases} Sn = \dfrac{tp}{tp + fn} \\ Sp = \dfrac{tn}{tn + fp} \\ Acc = \dfrac{tp + tn}{tp + fp + tn + fn} \end{cases} \quad (8)$$

where '*tp*' represents the correctly predicted promoter sequences and '*fp*' shows the non-promoter sequences classified as promoter sequence. And the other hand, '*tn*' characterizes the correctly recognized non-promotor sequences and '*fn*' exhibit the promoter sequences which were classified as non-promoter sequence.

3. Results and discussion

In the fields of statistical analysis and machine learning (ML) prediction, cross-validation (CV) strategy has been widely utilized to evaluate the prediction performance of ML models (Hasan et al., 2022; Shoombuatong et al., 2022; Xiao et al., 2022; Yu et al., 2022; Zhang et al., 2022). In this work, 5-fold CV technique was used in the processes of PseKNC parameter optimization and optimal feature subset selection and *10*-fold CV technique was used to assess the performance of the six machine learning methods. In *n*-fold CV, the benchmark dataset was randomly divided into *n* groups with equal size. Each group was individually tested on the model which was trained with the remaining *n*-1 groups. According to this, the *n*-fold CV method was performed *n* times, and the final evaluation result was the average prediction performance of the *n* models.

We constructed a computational model on the basis of sequence features to recognize promoter sequences in *K. aerogenes*. Based on the definition of pseudo nucleotide characteristics, we debugged the parameters k, λ, and ω according to the following range to determine the optimal combination of *k*-mer nucleotide composition information and long-range sequence order information.,

$$\begin{cases} k \in [2,5],\ step = 1 \\ \lambda \in [1,30],\ step = 1 \\ \omega \in [0.1,1],\ step = 0.1 \end{cases}$$

Based on the feature set generated by each combination and the LIBSVM algorithm, we can construct promoter prediction models and evaluate their accuracies using a 5-fold CV method. The final determined values of *k*, λ, and ω were 5, 29, and 0.1, respective. The original vector contains 1,198 features which could produce the prediction accuracy of 88.0%. Then, 17 positional correlation scoring features were calculated based on the most conserved sites in the promoter sequence of the 3-mer nucleotide fragment. After integrating two types of features, the mRMR algorithm was applied to sort all features, and an incremental feature selection (IFS) method was applied to eliminate redundant information to obtain the optimal feature subset for improving the accuracy of the promoter classifier. In the process of IFS, we also used a 5-fold CV method to evaluate the promoter prediction accuracy of each classifier, as shown in Figure 2A. As shown in the figure, the model constructed based on the first 586 features has the highest prediction accuracy of 95.9%.

After determining the optimal subset of features, we further evaluated its promoter prediction ability using a 10-fold CV method

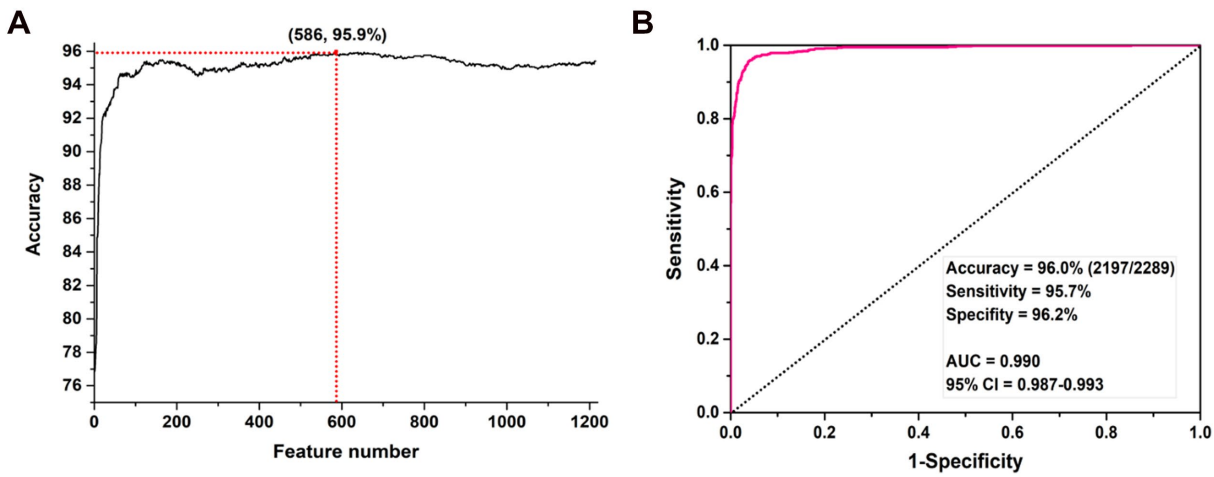

FIGURE 2
The prediction accuracies of SVM models constructed with different numbers of features. **(A)** IFS process for feature selection and **(B)** ROC curve based on the optimal features.

for determining the parameters c and γ in SVM, where $c \in [2^{-5}, 2^{15}]$ with a step of 2, $\gamma \in [2^{3}, 2^{-15}]$ with a step size of 2^{-1}. The final optimal values of c and γ are 2 and 2^{-3}, respectively. The optimal SVM model could produce the best performance with the accuracy of 96.0%, sensitivity of 95.7%, and specificity of 96.2%. The area under the ROC curve (AUC) was 0.990 with 95% confidence interval (CI): 0.987–0.993 (as shown in Figure 2B).

In order to evaluate the performance of this SVM prediction model, we also constructed five models based on LR, KNN, RF, AB and NB for *K. aerogenes* promoter recognition by using the same optimal features. The 10-fold CV results showed that the AUC values of the LR, KNN, RF, and AB models were 0.960, 0.941, 0.939, and 0.959, respectively, as shown in Figure 3. We observed that the sensitivity of the RF model was poor (68.8%), while the overall predictive performance of the NB model was the weakest, with accuracy and AUC values of 81.3% and 0.882 (Table 1). The accuracy of SVM-based model was 96.0% which was 5.6–14.7% higher than the other five classifiers. Overall, identifying *K. aerogenes* promoter sequences based on optimal pseudo nucleotide features and positional correlation scoring features is effective, and the model constructed based on SVM algorithm has the best predictive performance.

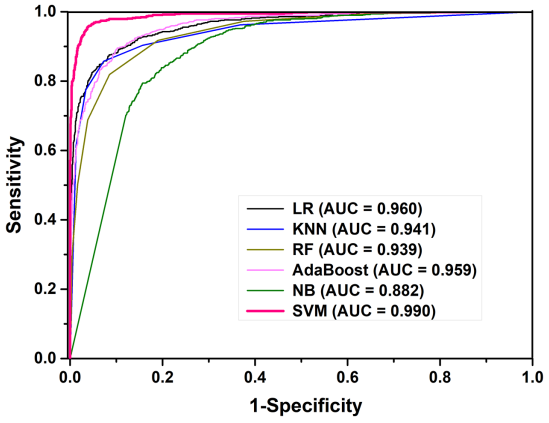

FIGURE 3
The ROC curves for different machine learning models.

TABLE 1 The prediction performance of different machine learning models based on 10-fold cross-validation.

Method	Sn (%)	Sp (%)	Acc (%)	AUC
LR	85.1	93.1	90.4	0.960
KNN	85.7	92.7	90.4	0.941
RF	68.8	96.2	87.1	0.939
AB	84	92.9	89.9	0.959
NB	83.9	79.9	81.3	0.882
SVM	**95.7**	**96.2**	**96.0**	**0.990**

Note: The *K. aerogenes* promoter prediction model constructed with SVM classifier produces the highest accuracy, sensitivity, specificity and AUC,which is the finally determined model.

4. Conclusion

Promoters play an important role in the initiation of transcription, because they are located upstream of genes. RNA polymerase and a quantity of transcription factors bind to promoter to start the transcription. Therefore, studying promoters is crucial for studying gene expression regulation. In this study, we proposed an SVM-based model to identify promoter sequences in *K. aerogenes*. In the proposed model, sequences were encoded using PseKNC and PCSF and then optimized with mRMR and SVM-based algorithm on 5-fold CV. Then, these optimized features were inputted into SVM-based classifier using 10-fold CV and achieved the best model. The results show that our model can predict promoters accurately, suggesting that our feature extraction and selection methods are able to capture the important sequence features. In the future, we will develop more suitable and robust models for more prokaryotic species.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: http://lin-group.cn/database/ppd/.

Author contributions

YL, HZ, and HL project design and oversight, and manuscript writing and revision. MS and HZ sample collection and curation. YL, JZ, HZ, ML, and KY experiment conduction and data analysis. YL and ML table preparation. YL, MS, and CW result interpretation and discussion. YL and JZ funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the grant from the National Natural Science Foundation of Sichuan Province (no. 2022NSFSC0058), the Sichuan Key Science and Technology Project (no. 2021ZDZX0009), the Research Program of Science and Technology at Universities of Inner Mongolia Autonomous Region under Grant NJZZ18381 and the National Natural Science Foundation of China (62262049).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ao, C., Jiao, S., Wang, Y., Yu, L., and Zou, Q. (2022). Biological sequence classification: a review on data and general methods. *Research* 2022:0011. doi: 10.34133/research.0011

Arif, M., Ahmed, S., Ge, F., Kabir, M., Khan, Y. D., Yu, D. J., et al. (2022). StackACPred: prediction of anticancer peptides by integrating optimized multiple feature descriptors with stacked ensemble approach. *Chemom. Intell. Lab. Syst.* 220:104458. doi: 10.1016/j.chemolab.2021.104458

Basith, S., Hasan, M. M., Lee, G., Wei, L., and Manavalan, B. (2021). Integrative machine learning framework for the identification of cell-specific enhancers from the human genome. *Brief. Bioinform.* 22:bbab252. doi: 10.1093/bib/bbab252

Basith, S., Lee, G., and Manavalan, B. (2022). STALLION: a stacking-based ensemble learning framework for prokaryotic lysine acetylation site prediction. *Brief. Bioinform.* 23:bbab376. doi: 10.1093/bib/bbab376

Boch, J., and Bonas, U. (2010). Xanthomonas Avr Bs3 family-type III effectors: discovery and function. *Annu. Rev. Phytopathol.* 48, 419–436. doi: 10.1146/annurev-phyto-080508-081936

Bupi, N., Sangaraju, V. K., Phan, L. T., Lal, A., Vo, T. T. B., Ho, P. T., et al. (2023). An effective integrated machine learning framework for identifying severity of tomato yellow leaf curl virus and their experimental validation. *Research* 6:0016. doi: 10.34133/research.0016

Cao, R., Freitas, C., Chan, L., Sun, M., Jiang, H., and Chen, Z. (2017). Pro Lan GO: protein function prediction using neural machine translation based on a recurrent neural network. *Molecules* 22:1732. doi: 10.3390/molecules22101732

Chen, L., Yu, L., and Gao, L. (2023). Potent antibiotic design via guided search from antibacterial activity evaluations. *Bioinformatics* 39:btad059. doi: 10.1093/bioinformatics/btad059

Chevez-Guardado, R., and Peña-Castillo, L. (2021). Promotech: a general tool for bacterial promoter recognition. *Genome Biol.* 22:318. doi: 10.1186/s13059-021-02514-9

Dao, F. Y., Liu, M. L., Su, W., Lv, H., Zhang, Z. Y., Lin, H., et al. (2023). AcrPred: a hybrid optimization with enumerated machine learning algorithm to predict anti-CRISPR proteins. *Int. J. Biol. Macromol.* 228, 706–714. doi: 10.1016/j.ijbiomac.2022.12.250

de Jong, A., Pietersma, H., Cordes, M., Kuipers, O. P., and Kok, J. (2012). PePPER: a webserver for prediction of prokaryote promoter elements and regulons. *BMC Genomics* 13:299. doi: 10.1186/1471-2164-13-299

Doherty, R., and Couldrey, C. (2014). Exploring genome wide bisulfite sequencing for DNA methylation analysis in livestock: a technical assessment. *Front. Genet.* 5:126. doi: 10.3389/fgene.2014.00126

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., et al. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* 7, 461–465. doi: 10.1038/nmeth.1459

Goñi, J. R., Fenollosa, C., Pérez, A., Torrents, D., and Orozco, M. (2008). DNAlive: a tool for the physical analysis of DNA at the genomic scale. *Bioinformatics* 24, 1731–1732. doi: 10.1093/bioinformatics/btn259

Hasan, M. M., Tsukiyama, S., Cho, J. Y., Kurata, H., Alam, M. A., Liu, X., et al. (2022). Deepm 5C: a deep-learning-based hybrid framework for identifying human RNA N5-methylcytosine sites using a stacking strategy. *Mol. Ther.* 30, 2856–2867. doi: 10.1016/j.ymthe.2022.05.001

Hu, Y., Zhang, Y., Zhang, H., Gao, S., Wang, L., Wang, T., et al. (2022a). Mendelian randomization highlights causal association between genetically increased C-reactive protein levels and reduced Alzheimer's disease risk. *Alzheimers Dement.* 18, 2003–2006. doi: 10.1002/alz.12687

Hu, Y., Zhang, Y., Zhang, H., Gao, S., Wang, L., Wang, T., et al. (2022b). Cognitive performance protects against Alzheimer's disease independently of educational attainment and intelligence. *Mol. Psychiatry* 27, 4297–4306. doi: 10.1038/s41380-022-01695-4

Lai, H.-Y., Zhang, Z.-Y., Su, Z.-D., Su, W., Ding, H., Chen, W., et al. (2019). iProEP: a computational predictor for predicting promoter. *Mol. Ther. Nucleic Acids* 17, 337–346. doi: 10.1016/j.omtn.2019.05.028

Li, Q. Z., and Lin, H. (2006). The recognition and prediction of sigma (70) promoters in *Escherichia coli* K-12. *J. Theor. Biol.* 242, 135–141. doi: 10.1016/j.jtbi.2006.02.007

Li, H., Zhang, J., Zhao, Y., and Yang, W. (2023). Predicting *Corynebacterium glutamicum* promoters based on novel feature descriptor and feature selection technique. *Front. Microbiol.* 14:1141227. doi: 10.3389/fmicb.2023.1141227

Lin, H., Deng, E. Z., Ding, H., Chen, W., and Chou, K. C. (2014). iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.* 42, 12961–12972. doi: 10.1093/nar/gku1019

Lin, H., and Li, Q. Z. (2011). Eukaryotic and prokaryotic promoter prediction using hybrid approach. *Theory Biosci.* 130, 91–100. doi: 10.1007/s12064-010-0114-8

Lin, H., Liang, Z. Y., Tang, H., and Chen, W. (2019). Identifying Sigma70 promoters with novel Pseudo nucleotide composition. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 1316–1321. doi: 10.1109/TCBB.2017.2666141

Naseer, S., Ali, R. F., Khan, Y. D., and Dominic, P. D. D. (2022). iGluK-deep: computational identification of lysine glutarylation sites using deep neural networks with general pseudo amino acid compositions. *J. Biomol. Struct. Dyn.* 40, 11691–11704. doi: 10.1080/07391102.2021.1962738

Patiyal, S., Singh, N., Ali, M. Z., Pundir, D. S., and Raghava, G. P. (2022). Sigma70Pred: a highly accurate method for predicting sigma70 promoter in *Escherichia coli* K-12 strains. *Front. Microbiol.* 13:1042127. doi: 10.3389/fmicb.2022.1042127

Price, D., and Sleigh, J. (1970). Control of infection due to *Klebsiella aerogenes* in a neurosurgical unit by withdrawal of all antibiotics. *Lancet* 296, 1213–1215. doi: 10.1016/S0140-6736(70)92179-3

Sawadogo, M., and Roeder, R. G. (1985). Interaction of a gene-specific transcription factor with the adenovirus major late promoter upstream of the TATA box region. *Cells* 43, 165–175. doi: 10.1016/0092-8674(85)90021-2

Shoombuatong, W., Basith, S., Pitti, T., Lee, G., and Manavalan, B. (2022). THRONE: a new approach for accurate prediction of human RNA N7-Methylguanosine sites. *J. Mol. Biol.* 434:167549. doi: 10.1016/j.jmb.2022.167549

Shujaat, M., Jin, J. S., Tayara, H., and Chong, K. T. (2022). iProm-phage: a two-layer model to identify phage promoters and their types using a convolutional neural network. *Front. Microbiol.* 13:1061122. doi: 10.3389/fmicb.2022.1061122

Song, K. (2012). Recognition of prokaryotic promoters based on a novel variable-window Z-curve method. *Nucleic Acids Res.* 40, 963–971. doi: 10.1093/nar/gkr795

Su, W., Liu, M. L., Yang, Y. H., Wang, J. S., Li, S. H., Lv, H., et al. (2021). PPD: a manually curated database for experimentally verified prokaryotic promoters. *J. Mol. Biol.* 433:166860. doi: 10.1016/j.jmb.2021.166860

Su, W., Xie, X. Q., Liu, X. W., Gao, D., Ma, C. Y., Zulfiqar, H., et al. (2023). iRNA-ac4C: a novel computational method for effectively detecting N4-acetylcytidine sites in human mRNA. *Int. J. Biol. Macromol.* 227, 1174–1181. doi: 10.1016/j.ijbiomac.2022.11.299

Tang, Q., Nie, F., Kang, J., and Chen, W. (2021). mRNALocater: enhance the prediction accuracy of eukaryotic mRNA subcellular localization by using model fusion strategy. *Mol. Ther.* 29, 2617–2623. doi: 10.1016/j.ymthe.2021.04.004

Tang, Q., Nie, F., Zhao, Q., and Chen, W. (2022). A merged molecular representation deep learning method for blood-brain barrier permeability prediction. *Brief. Bioinform.* 23:bbac357. doi: 10.1093/bib/bbac357

Wang, H., Liang, P. F., Zheng, L., Long, C. S., Li, H. S., and Zuo, Y. (2021). eHSCPr discriminating the cell identity involved in endothelial to hematopoietic transition. *Bioinformatics* 37, 2157–2164. doi: 10.1093/bioinformatics/btab071

Xiao, J., Liu, M., Huang, Q., Sun, Z., Ning, L., Duan, J., et al. (2022). Analysis and modeling of myopia-related factors based on questionnaire survey. *Comput. Biol. Med.* 150:106162. doi: 10.1016/j.compbiomed.2022.106162

Yang, Y., Gao, D., Xie, X., Qin, J., Li, J., Lin, H., et al. (2022). DeepIDC: a prediction framework of injectable drug combination based on heterogeneous information and deep learning. *Clin. Pharmacokinet.* 61, 1749–1759. doi: 10.1007/s40262-022-01180-9

Yang, H., Luo, Y., Ren, X., Wu, M., He, X., Peng, B., et al. (2021). Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators. *Inf. Fusion* 75, 140–149. doi: 10.1016/j.inffus.2021.02.015

Yu, L., Zheng, Y. J., and Gao, L. (2022). MiRNA-disease association prediction based on meta-paths. *Brief. Bioinform.* 23:bbab571. doi: 10.1093/bib/bbab571

Zhang, Q., Li, H., Liu, Y., Li, J., Wu, C., and Tang, H. (2022). Exosomal non-coding RNAs: new insights into the biology of hepatocellular carcinoma. *Curr. Oncol.* 29, 5383–5406. doi: 10.3390/curroncol29080427

Zhang, Z. Y., Ning, L., Ye, X., Yang, Y. H., Futamura, Y., Sakurai, T., et al. (2022). iLoc-miRNA: extracellular/intracellular miRNA prediction using deep BiLSTM with attention mechanism. *Brief. Bioinform.* 23:bbac395. doi: 10.1093/bib/bbac395

Zhang, Y. F., Wang, Y. H., Gu, Z. F., Pan, X., Li, J., Ding, H., et al. (2023). Bitter-RF: a random forest machine model for recognizing bitter peptides. *Front. Med.* 10:1052923. doi: 10.3389/fmed.2023.1052923

Zhu, H., Ao, C. Y., Ding, Y. J., Hao, H. X., and Yu, L. (2022). Identification of D modification sites using a random Forest model based on nucleotide chemical properties. *Int. J. Mol. Sci.* 23:3044. doi: 10.3390/ijms23063044

Zulfiqar, H., Guo, Z., Grace-Mercure, B. K., Zhang, Z. Y., Gao, H., Lin, H., et al. (2023). Empirical comparison and recent advances of computational prediction of hormone binding proteins using machine learning methods. *Comput. Struct. Biotechnol. J.* 21, 2253–2261. doi: 10.1016/j.csbj.2023.03.024

Zulfiqar, H., Huang, Q.-L., Lv, H., Sun, Z. J., Dao, F. Y., and Lin, H. (2022). Deep-4mCGP: a deep learning approach to predict 4mC sites in *Geobacter pickeringii* by using correlation-based feature selection technique. *Int. J. Mol. Sci.* 23:1251. doi: 10.3390/ijms23031251

Zulfiqar, H., Khan, R. S., Hassan, F., Hippe, K., Hunt, C., Ding, H., et al. (2021). Computational identification of N4-methylcytosine sites in the mouse genome with machine-learning method. *Math. Biosci. Eng.* 18, 3348–3363. doi: 10.3934/mbe.2021167

Zulfiqar, H., Yuan, S.-S., Huang, Q.-L., Sun, Z. J., Dao, F. Y., Yu, X. L., et al. (2021). Identification of cyclin protein using gradient boost decision tree algorithm. *Comput. Struct. Biotechnol. J.* 19, 4123–4131. doi: 10.1016/j.csbj.2021.07.013

Zulfiqar, H., Zahoor, A., Kissanga Grace-Mercure, B., Hassan, F., Zhang, Z. Y., and Liu, F. (1011). Computational prediction of promotors in *Agrobacterium Tumefaciens* strain C58 by using machine learning technique. *Front. Microbiol.* 14

Zuo, Y. C., and Li, Q. Z. (2010). The hidden physical codes for modulating the prokaryotic transcription initiation. *Phys. A-Stat. Mech. Appl.* 389, 4217–4223. doi: 10.1016/j.physa.2010.05.034

Frontiers in Microbiology

Explores the habitable world and the potential of microbial life

The largest and most cited microbiology journal which advances our understanding of the role microbes play in addressing global challenges such as healthcare, food security, and climate change.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact