IET Systems Biology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

SVM-LncRNAPro: An SVM-Based Method for Predicting Long Noncoding RNA Promoters

Guohua Huang[1,2] | Taigan Xue[1] | Weihong Chen[2] | Liangliang Huang[2] | Qi Dai[3] | JinYun Jiang[1]

[1]College of Information Science and Engineering, Shaoyang University, Shaoyang, China | [2]Hunan Provincial Key Laboratory of Finance & Economics Big Data Science and Technology, Hunan University of Finance and Economics, Changsha, China | [3]College of Life Science and Medicine, Zhejiang Sci-Tech University, Hangzhou, China

Correspondence: JinYun Jiang (tjjjy86@ecjtu.edu.cn)

Received: 10 November 2024 | **Revised:** 5 January 2025 | **Accepted:** 12 February 2025

Handling Editor: Renzhi Cao

Funding: The work was supported by the Special Support Plan for High Level Talents in Zhejiang Province (Grant No. 2021R52019), by the Scientific Research Fund of the Hunan Provincial Education Department (24A0701, 24A0694) and by the Shaoyang University Innovation Foundation for Postgraduate (CX2023SY048).

Keywords: bioinformatics | DNA | feature selection | support vector machines

ABSTRACT

Long non-coding RNAs (lncRNAs) are closely associated with the regulation of gene expression, whose promoters play a crucial role in comprehensively understanding lncRNA regulatory mechanisms, functions and their roles in diseases. Due to limitations of the current techniques, accurately identifying lncRNA promoters remains a challenge. To address this challenge, we propose a support vector machine (SVM)–based method for predicting lncRNA promoters, called SVM-LncRNAPro. This method uses position-specific trinucleotide propensity based on single-strand (PSTNPss) to encode the DNA sequences and employs an SVM as the learning algorithm. The SVM-LncRNAPro achieves state-of-the-art performance with reduced complexity. Additionally, experiments demonstrate that this method exhibits a strong generalisation ability. For the convenience of academic research, we have made the source code of SVM-LncRNAPro publicly available. Researchers can download the code and perform the prediction of the lncRNA promoter via the following link: https://github.com/TG0F7/Prom/tree/master.

1 | Introduction

Long noncoding RNAs (LncRNAs) are a class of RNA molecules that do not encode proteins but typically exceed 200 nucleotides in length [1]. LncRNAs play essential roles in various cellular processes, including gene expression [2], transcription [3], splicing [4], chromatin remodelling [5, 6], genome stability [7] and cell cycle regulation [8]. According to their genomic location and function, lncRNAs are classified into long intergenic noncoding RNAs (lincRNAs), antisense lncRNAs, intronic lncRNAs and competitive endogenous RNAs (ceRNAs) [9]. LncRNAs interact with DNA, RNA or proteins to perform diverse and complex functions [2]. Due to their critical roles in biological processes, lncRNAs are considered to be potential therapeutic targets or biomarkers for various diseases [10], including cancer [11], cardiovascular disease [12], neurological disorders [13] and autoimmune diseases [14].

A promoter is a region of the DNA sequence to which the transcription factors bind to begin the transcription of RNA from the DNA sequence [15, 16]. The promoter is typically adjacent to the initial transcript sites of genes and provides the binding sites for transcription factors that recruit RNA polymerase. The promoters serve as critical elements to control the

https://doi.org/10.1049/syb2.70013

gene expression [17]. The lncRNA expression levels and biological functions are regulated by the lncRNA promoters. Therefore, identifying the lncRNA promoter is the key foundation to explore its function. Over the past decades, a number of methods have been proposed for identifying lncRNA promoters, which are categorised into experimental and computational methods. Experimental methods include mutation analysis and chromatin immunoprecipitation (ChIP) [18–20]. Mutation analysis, which often utilises CRISPR-Cas9 [21], introduces genomic mutations into the lncRNA promoter to evaluate their effect on gene expression [22]. Accurately selecting the mutation site is crucial in mutation analysis. If the site was not selected properly or if the mutation site was outside of the functional region of the gene, it would result in invalid or inaccurate experimental results. Additionally, mutation analysis is time-consuming and challenging for large-scale applications. ChIP [19, 23–26] is a technique that identifies protein–DNA interaction sites by chemically cross-linking proteins to DNA, enriching the DNA fragments bound to target proteins using specific antibodies and precisely identifying binding sites through quantitative polymerase chain reaction (qPCR) or high-throughput sequencing (ChIP-seq). However, ChIP faces technical challenges in antibody specificity and data interpretation, which limits its efficiency in high-throughput applications. These limitations of experimental methods have prompted researchers to explore computational approaches as alternatives. In recent years, advancements in computational power and machine learning techniques have significantly promoted the resolution of various challenges in bioinformatics [27–33]. These methods demonstrate flexibility and strong predictive capabilities in addressing various biological problems. In the field of lncRNA promoter identification, both traditional machine learning and deep learning have been widely applied, both achieving significant improvements in predictive accuracy. In 2019, Alam et al. proposed a deep learning model called DeepCNPP based on convolutional neural networks (CNN) for identifying lncRNA promoters [34]. This method was specifically designed for the prediction of human lncRNA promoters. In 2020, Tang et al. developed ncPro-ML [35], whereas Zhang et al. proposed DeepLncPro [18] in 2022. Both methods aimed at identifying human and mouse lncRNA promoters. The ncPro-ML is the lncRNA promoter recognition method based on support vector machine (SVM), employing various feature encodings to represent DNA sequences. The DeepLncPro is a deep learning method that uses three encodings to represent DNA sequences, with two one-dimensional convolutional layers. Although these methods perform well in predicting lncRNA promoters, their reliance on multiple encodings or complex deep learning models makes their design relatively complex, limiting their broader applications. Therefore, it is necessary to develop a simpler method with wider applicability. We proposed a simple and effective method based on SVM for identifying human and mouse lncRNA promoters, called SVM-LncRNAPro. This method used PSTNPss as feature encoding and employed SVM as the learning algorithm. The SVM-LncRNAPro reduced the complexity of the architectures while still maintaining high predictive accuracy. Compared to existing methods, our method not only performed well in identifying lncRNA promoters in humans and mice, but also demonstrated strong generalisation ability across multiple species.

2 | Materials and Methods

2.1 | Dataset

We used the same datasets as DeepLncPro [18] to construct benchmark datasets. The human and mouse noncoding promoter databases were downloaded from the Eukaryotic Promoter Database (EPD, https://epd.epfl.ch) [36]. The promoter is usually located in the upstream of the transcription start sites (TSS). We cut the sequences between n-20 bp upstream from the TSS to 20 bp downstream of the TSS as the positive sample. The negative samples were extracted from 1000 bp downstream of the TSS to (1000+n) bp downstream of the TSS. The human and mouse datasets contained 4678 and 6154 samples, respectively. The ratio of positive to negative samples in the datasets was 1:1. The dataset was divided into training, validation and test set at a ratio of 6:2:2. In the human dataset, the training set contained 1403 positive samples, whereas the validation and the test sets contained 468 positive samples, respectively, with an equal number of negative samples. In the mouse dataset, the training set contained 1846 positive samples, the validation set contained 616 positive samples and the test set contained 615 positive samples, with the negative sample counts also being equal.

2.2 | Method

As shown in Figure 1, the method architecture is divided into three parts. The first part involves feature selection and feature combination. A total of 11 various types of features were computed with iLearn [37] and evaluated by their performance using SVM [38]. The 11 types of features were sorted by their accuracy, and the features with top performance were combined. The second part focuses on determining the optimal feature and the optimal sequence length. A total of 13 different sequence lengths and feature combinations were investigated. On the basis of predictive performance, we optimised the sequence length and chose the optimal feature combination. Finally, we compared six common machine learning algorithms (random forest [RF] [39], logic regression [LR] [40], K-nearest neighbours algorithm [KNN] [41], SVM [38], light gradient boosting machine [LightGBM] [42] and extreme gradient boosting [XGBoost] [43]) and selected the one with the best performance as the learning algorithm in the SVM-LncRNAPro. It should be noted that the method comparison and parameter optimisation processes in this paper were performed on the validation set of the human dataset.

2.3 | PSTNPss

The PSTNPss encoding method utilises statistical rules to encode DNA sequences. The method starts from the beginning of the DNA sequence [44] and moves along in windows of three nucleotides until it reaches the end of the sequence. During this process, the DNA sequence is divided into 64 different trinucleotides, such as ‘AAA’, ‘AAC’, ‘AAG’, ..., until ‘TTT’. Therefore, for a given DNA sequence of length L, its position-specific trinucleotide propensity can be represented by a $64 \times (L-2)$ matrix:

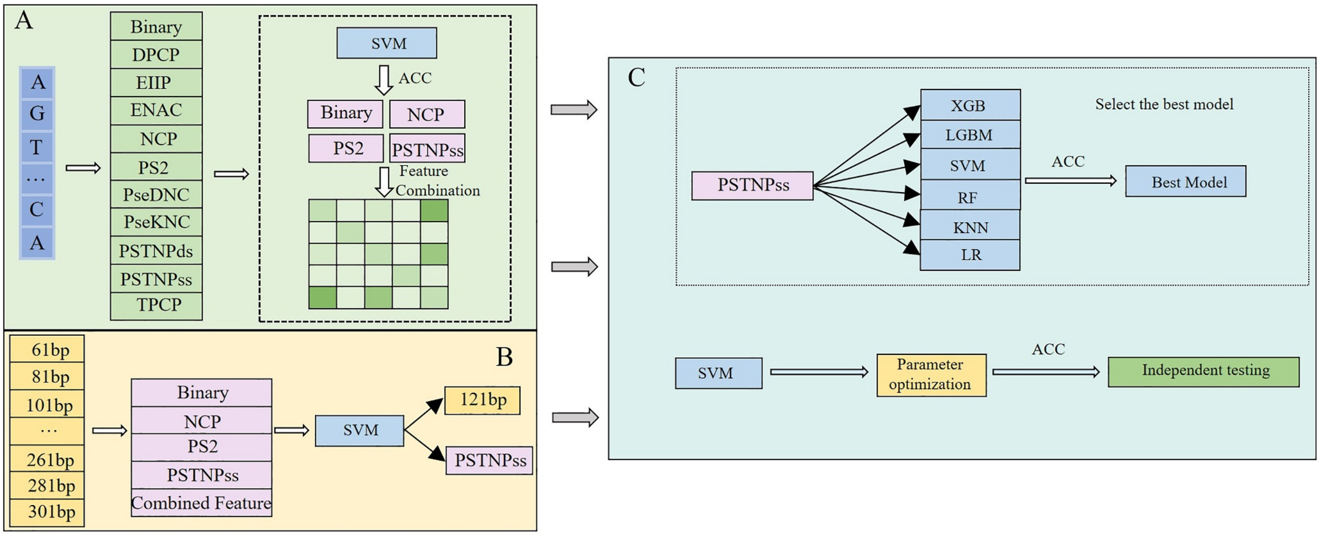

FIGURE 1 | The architecture of the proposed SVM-LncRNAPro model.

$$Z = \begin{bmatrix} Z_{1,1} & \cdots & Z_{1,L-2} \\ \vdots & \ddots & \vdots \\ Z_{64,1} & \cdots & Z_{64,L-2} \end{bmatrix} \tag{1}$$

In Equation (1), $Z_{i,j} = Z_{i,j}^{+} - Z_{i,j}^{-}$, where i represents the ith trinucleotide. When $i = 1$, the corresponding trinucleotide is AAA; when $i = 2$, the corresponding trinucleotide is AAC and so forth. j denotes the position index of trinucleotides. $Z_{i,j}^{+}$ represents the frequency of the ith trinucleotide appearing at position j in the positive samples, and $Z_{i,j}^{-}$ represents the frequency of the ith trinucleotide appearing at position j in the negative samples. Therefore, for a sequence of length L, the position-specific trinucleotide propensity is denoted by its feature vector S:

$$S = [\phi_1, \phi_2, \cdots, \phi_t \cdots, \phi_{L-2}]^T \tag{2}$$

In Equation (2), T denotes the transpose of the vector, and ϕ_t is defined as follows:

$$\phi_t = \begin{cases} Z_{1,t}, \text{ when } N_t N_{t+1} N_{t+2} = AAA \\ \qquad \vdots \qquad\qquad\qquad\qquad (1 \le t \le L-2) \\ Z_{64,t}, \text{ when } N_t N_{t+1} N_{t+2} = TTT \end{cases} \tag{3}$$

In Equation (3), N_t represents the nucleotide (A, T, C or G) of the DNA sequence at the t-th position.

2.4 | Performance Evaluation Metrics

To evaluate the performance of SVM-LncRNAPro fairly, we employed the same metrics as DeepLncPro [18] and DeePromClass [45]: recall/sensitivity (SN), specificity (SP), accuracy (ACC), Matthews correlation coefficient (MCC), precision (Pre) and F1 Score (F1). The definitions of these metrics are as follows:

$$\text{SN} = \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{SP} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5}$$

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

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

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

where TP refers to the number of samples correctly predicted as promoters in the datasets, TN refers to the number of samples correctly predicted as nonpromoters, FP refers to the number of samples incorrectly predicted as promoters and FN refers to the number of samples incorrectly predicted as nonpromoters. Additionally, we employed the area under the curve (AUC) to evaluate the performance of SVM-LncRNAPro. The closer the AUC value is to 1, the stronger the model's classification capability.

3 | Results and Discussion

3.1 | Feature Selection and Sequence Length Selection

We compared 11 commonly used feature encodings (binary [46], dinucleotide physicochemical properties (DPCP) [35], electron-ion interaction pseudopotential (EIIP) [47], enhanced nucleic acid composition (ENAC) [48], NCP [49], PS2 [50], pseudo dinucleotide composition (PseDNC) [51], pseudo K-tuple nucleotide composition (PseKNC) [52], position-specific trinucleotide propensity based on double-strand (PSTNPds) [53], PSTNPss [44] and Trinucleotide Physicochemical Properties (TPCP) [54]). A total of 11 types of features were computed by iLearnPlus [37]. We also compared 13 different lengths (61 bp, 81 bp, 101 bp, 121 bp, 141 bp, 161 bp, 181 bp, 201 bp, 221 bp, 241 bp, 261 bp, 281 bp and 301 bp) of DNA sequences. The encodings and sequence lengths were evaluated by the SVM classifiers over the validations.

As shown in Table 1, the binary reached the best ACC (0.8387) at the sequence length of 121 bp, the DPCP reached the best ACC (0.6880) at the 101 bp, the EIIP reached the best ACC (0.7468) at the 61 bp, the ENAC reached the best ACC (0.7863) at the 61 bp, the NCP reached the best ACC (0.8387) at the 121 bp, the PS2 reached the best ACC (0.8472) at the 121 bp, the PseDNC reached the best ACC (0.6944) at the 221 bp, the PseKNC reached the best ACC (0.7051) at the 161 bp, the PSTNPds reached the best ACC (0.7286) at the 221 bp, the PSTNPss reached the best ACC (0.8857) at the 121 bp and the TPCP reached the best ACC (0.7179) at the 161 bp. Obviously, the PSTNPss performed best, followed by the PS2, then by NCP and the binary. Figure 2 showed SN, SP, ACC and MCC of 11 types of features over the human validation dataset.

Different encodings exhibited varying performances (Figure 2). The PSTNPss, PS2, NCP and binary performed excellently on the whole. Therefore, we further explored the combination of 4 types of encoding. For 4 types of encoding, there are 15 potential combinations. Table 2 showed the performance of 15 combinations. Obviously, the single encoding PSTNPss still reached the best ACC at the 121 bp among 15 combinations. Therefore, we set the sequence length to 121 and chose PSTNPss to encode promoter sequences.

3.2 | Comparison of Different Machine Learning Algorithms

Different machine learning algorithms would perform differently on the same dataset. We compared six widely used machine learning algorithms LightGBM [42], XGBoost [43], RF [39], SVM [38], LR [40] and KNN [41]) and chose the best-

TABLE 1 | Comparison with various features and sequence lengths.

Length													
Encoding	**61 bp**	**81 bp**	**101 bp**	**121 bp**	**141 bp**	**161 bp**	**181 bp**	**201 bp**	**221 bp**	**241 bp**	**261 bp**	**281 bp**	**301 bp**
Binary	0.8259	0.8119	0.8291	**0.8387**	0.8120	0.8184	0.7788	0.7917	0.8098	0.7479	0.7585	0.7809	0.7735
DPCP	0.6806	0.6613	**0.6880**	0.6752	0.6731	0.6838	0.6677	0.6816	0.6816	0.6410	0.6303	0.6517	0.6496
EIIP	**0.7468**	0.7115	0.7286	0.7382	0.7212	0.6987	0.7030	0.7083	0.6987	0.6891	0.6645	0.6944	0.6741
ENAC	**0.7863**	0.7585	0.7618	0.7735	0.7671	0.7607	0.7233	0.7447	0.7564	0.7222	0.6806	0.7479	0.7244
NCP	0.8259	0.8119	0.8291	**0.8387**	0.8120	0.8184	0.7788	0.7917	0.8098	0.7479	0.7585	0.7809	0.7735
PS2	0.8408	0.8258	0.8365	**0.8472**	0.8162	0.8365	0.8013	0.8088	0.8109	0.7607	0.7532	0.7821	0.7874
PseDNC	0.6912	0.6592	0.6795	0.6848	0.6763	0.6870	0.6752	0.6902	**0.6944**	0.6485	0.6293	0.6581	0.6624
PseKNC	0.6966	0.6709	0.6966	0.6998	0.6794	**0.7051**	0.6901	0.7019	0.7008	0.6613	0.6613	0.6720	0.6891
PSTNPds	0.7083	0.6891	0.7041	0.7137	0.6806	0.7137	0.7105	0.7126	**0.7286**	0.6902	0.6774	0.7062	0.7201
PSTNPss	0.8451	0.8558	0.8632	**0.8857**	0.8665	0.8589	0.8462	0.8515	0.8504	0.8483	0.8301	0.8611	0.8483
TPCP	0.6870	0.6581	0.7073	0.7137	0.6795	**0.7179**	0.7019	0.7009	0.7083	0.6603	0.6624	0.6891	0.6720

Note: Bold values indicate the highest value in each encoding method.

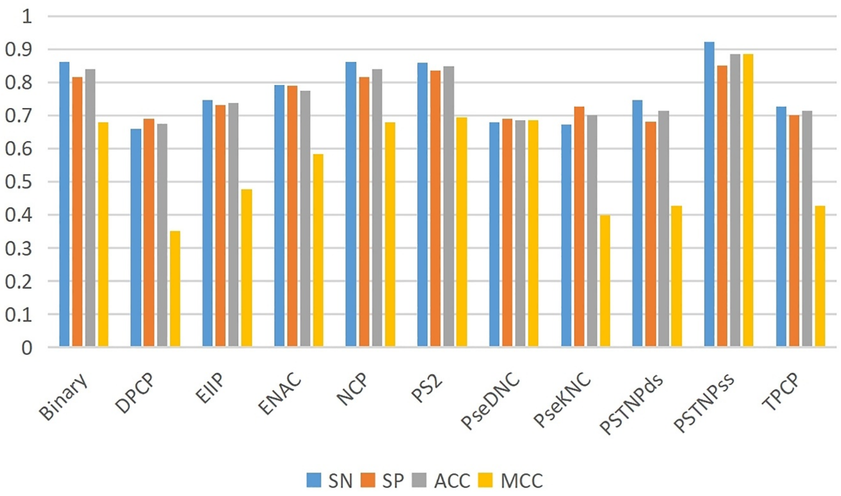

FIGURE 2 | Performance comparison of different feature encodings on the SVM.

TABLE 2 | Comparison with combinations of different encodings.

Length Encoding	61 bp	81 bp	101 bp	121 bp	141 bp	161 bp	181 bp	201 bp	221 bp	241 bp	261 bp	281 bp	301 bp
Binary	0.8259	0.8119	0.8291	0.8387	0.8120	0.8184	0.7788	0.7917	0.8098	0.7479	0.7585	0.7809	0.7735
NCP	0.8259	0.8119	0.8291	0.8387	0.8120	0.8184	0.7788	0.7917	0.8098	0.7479	0.7585	0.7809	0.7735
PS2	0.8408	0.8258	0.8365	0.8472	0.8162	0.8365	0.8013	0.8088	0.8109	0.7607	0.7532	0.7821	0.7874
PSTNPss	0.8451	0.8558	0.8632	**0.8857**	0.8665	0.8589	0.8462	0.8515	0.8504	0.8483	0.8301	0.8611	0.8483
Binary+NCP	0.8226	0.8120	0.8269	0.8387	0.8098	0.8184	0.7767	0.7949	0.8088	0.7511	0.7585	0.7821	0.7703
Binary+PS2	0.8472	0.8280	0.8429	0.8419	0.8184	0.8323	0.7885	0.8056	0.8109	0.7618	0.7543	0.7853	0.7831
Binary+PSTNPss	0.8216	0.8120	0.8333	0.8397	0.8109	0.8184	0.7810	0.7949	0.8109	0.7521	0.7607	0.7831	0.7703
NCP+PS2	0.8440	0.8291	0.8376	0.8429	0.8216	0.8333	0.7981	0.8077	0.8141	0.7660	0.7575	0.7853	0.7842
NCP+PSTNPss	0.8216	0.8141	0.8301	0.8365	0.8098	0.8205	0.7799	0.7895	0.8120	0.7521	0.7606	0.7842	0.7724
PS2+PSTNPss	0.8419	0.8279	0.8376	0.8451	0.8184	0.8365	0.8034	0.8098	0.8109	0.7618	0.7553	0.7821	0.7874
Binary+NCP+ PS2	0.8450	0.8290	0.8387	0.8408	0.813	0.8237	0.7917	0.8045	0.8120	0.7639	0.7543	0.7874	0.7853
Binary+NCP+ PSTNPss	0.8205	0.8120	0.8301	0.8376	0.8098	0.8184	0.7767	0.7927	0.8109	0.7532	0.7607	0.7842	0.7724
Binary+PS2+ PSTNPss	0.8483	0.8280	0.8429	0.8419	0.8205	0.8333	0.7917	0.8077	0.8120	0.7639	0.7543	0.7863	0.7831
NCP+PS2+ PSTNPss	0.8440	0.8290	0.8376	0.8429	0.8216	0.8323	0.7981	0.8088	0.8141	0.7660	0.7585	0.7863	0.7842
Binary+NCP+ PS2+ PSTNPss	0.8429	0.8312	0.8387	0.8408	0.8119	0.8237	0.7917	0.8066	0.8120	0.7650	0.7543	0.7874	0.7863

Note: The bold value indicates the highest value among all encoding combinations.

TABLE 3 | Comparison with different machine learning algorithms.

Method	SN	SP	ACC	MCC
XGBoost	0.8205	0.8141	0.8173	0.6346
LightGBM	0.7885	0.8226	0.8056	0.6115
LR	0.8504	**0.8803**	0.8654	0.7311
SVM	**0.9209**	0.8504	**0.8857**	**0.7733**
KNN	0.8825	0.8162	0.8493	0.7003
RF	0.8910	0.8376	0.8643	0.7297

Note: Bold values indicate the highest value in each metric.

TABLE 4 | Performance comparison of SVM-LncRNAPro and DeepLncPro.

Species	Name	SN	SP	ACC	MCC
Human	DeepLncPro	0.8974	0.8269	0.8622	0.7300
	SVM-LncRNAPro	**0.9017**	**0.8312**	**0.8665**	**0.7347**
Mouse	DeepLncPro	**0.8878**	**0.8488**	**0.8683**	**0.7400**
	SVM-LncRNAPro	0.8685	0.8374	0.8530	0.7063

Note: Bold values indicate the highest value for each metric in each dataset.

performing one as the learning algorithm to identify lncRNA promoters. All samples were set to 121 bp and were encoded into the PSTNPss [44]. For each learning algorithm, we used the grid search to optimise their super-parameters, respectively.

As shown in Table 3, the SVM performs the best in three evaluation metrics (SN, ACC and MCC). The LR performed best only in the SP, whereas the RF, the KNN, the XGBoost and the LightGBM were inferior to the LR. Therefore, we used the SVM as the learning algorithm to predict lncRNA promoters.

3.3 | Comparison With State-of-The-Art Methods

We compared SVM-LncRNAPro with DeepLncPro [18]which is one of the latest methods to predict lncRNA promoters.

As shown in Table 4, the SVM-LncRNAPro obtained competitive performance to the DeepLncPro. In the mouse dataset, DeepLncPro showed slightly superior performance compared to the SVM-LncRNAPro, exceeding 1.93% of SN, 1.14% of SP, 1.53% of ACC and 3.37% of MCC. The latter exhibited superiority over the former in the human dataset, increasing SN by 0.43%, SP by 0.43%, ACC by 0.43% and MCC by 0.47%, respectively.

3.4 | Generalisation Ability Test

To further test the generalisation ability of the SVM-LncRNAPro, we downloaded five datasets from the DeePromClass [45]. These five datasets are from five species: *Saccharomyces cerevisiae* (*S. cerevisiae*), *Caenorhabditis elegans* (*C. elegans*), *Drosophila melanogaster* (*D. melanogaster*), *Mus musculus* and *Homo sapiens*. We used the PSTNPss to encode the promoter and nonpromoter sequences. Because the DeePromClass set the sequence length to 151 bp, the SVM-LncRNAPro used the same length to fairly be compared to it.

As shown in Table 5, the SVM-LncRNAPro exhibited competitive performance to the DeePromClass. In the S. cerevisiae dataset, the SVM-LncRNAPro reached a precision of 0.9324, an F1 score of 0.9176, an ACC of 0.9189 and an AUC of 0.9632, respectively, promoting precision by 0.0424, F1 score by 0.0076, ACC by 0.0089 and AUC by 0.0532 over the DeePromClass. In other species datasets, the SVM-LncRNAPro is slightly inferior to the DeePromClass. For example, in the human dataset, the SVM-LncRNAPro is less by 0.0151 ACC than the DeePromClass.

3.5 | Analysis of Feature Contributions

We used the SHAP (SHapley Additive exPlanations) tool to explore the contributions of features to the predictive performance. The promoter and the nonpromoter sequences are 121 bp in length, which results in $121-3+1 = 119$ PSTNPss due to overlapping windows of trinucleotides. We named each feature PSTNPss-k ($k = 0, 1, 2, \ldots, 117, 118$).

As shown in Figure 3, the top 10 contributing features are PSTNPss-98, PSTNPss-99, PSTNPss-100, PSTNPss-97, PSTNPss-102, PSTNPss-73, PSTNPss-58, PSTNPss-112, PSTNPss-24 and PSTNPss-118. In the SHAP plot, the horizontal axis represents the SHAP values, with each point denoting the SHAP value which reflects how much the feature contributes to the identification of lncRNA promoters. The colour indicates the value of the feature. Red indicates larger values for the samples, whereas blue implies smaller values. These features are located in the *y*-axis rows, ranked from top to low according to the importance. The PSTNPss-98, the PSTNPss-99, the PSTNPss-97, the PSTNPss-100 and the PSTNPss-102 are the top contributing features, whose corresponding trinucleotides are located nearby the TSS. This indicated that the region around the TSS contained strong signals of lncRNA promoters. The PSTNPss-73, the PSTNPss-58, the PSTNPss-112, the PSTNPss-24 and PSTNPss-118 contribute much to lncRNA promoter identification. These implied that there was a signal of promoter both in the upstream and in the downstream of the TSS.

4 | Discussion

We developed an SVM-based method called the SVM-LncRNAPro to identify lncRNA promoters. The SVM-LncRNAPro used the PSTNPss to represent the lncRNA promoters. As shown in Tables 1 and 2, the length of sample sequences would pose a certain effect on the predictive performance of the LncRNAPro. The longer sequences were not sure to increase the performance. The shorter sequences would not include the necessary signal of promoters. On the contrary, the longer sequence would contain the noise which interferes the prediction. This might be the reason why the performance varied when the sequence became long.

As shown in Table 5, the SVM-LncRNAPro exhibited an outstanding generalisation ability. The SVM-LncRNAPro was

TABLE 5 | Generalisation ability test.

Species	Model	Precision	Recall	F1	ACC	AUC
S. cerevisiae	XGBoost	0.9057	0.9102	0.9079	0.9077	**0.9640**
	LR	0.8927	0.8613	0.8767	0.8789	0.9432
	LightGBM	0.8986	0.92675	0.9125	0.9111	0.9631
	KNN	0.9187	0.8281	0.8710	0.8774	0.9249
	RF	0.8898	0.9150	0.9022	0.9008	0.9586
	DeePromClass	0.89	**0.93**	0.91	0.91	0.91
	SVM-LncRNAPro	**0.9324**	0.9033	**0.9176**	**0.9189**	0.9632
C. elegans	XGBoost	0.9219	0.9199	0.9209	0.9210	**0.9731**
	LR	0.9228	0.8735	0.8975	0.9002	0.9594
	LightGBM	0.9119	0.9228	0.9173	0.9168	0.9725
	KNN	0.8900	0.8694	0.8795	0.8810	0.9323
	RF	0.9058	0.9185	0.9121	0.9115	0.9690
	DeePromClass	**0.93**	**0.94**	**0.93**	**0.94**	0.94
	SVM-LncRNAPro	0.9259	0.9129	0.9194	0.9199	0.9705
D. melanogaster	XGBoost	0.9175	0.8940	0.9056	0.9068	**0.9623**
	LR	0.9142	0.8003	0.8535	0.8626	0.9381
	LightGBM	0.9133	0.8940	0.9035	0.9046	0.9615
	KNN	0.9188	0.8162	0.8645	0.8720	0.9266
	RF	0.9017	0.8757	0.8885	0.8901	0.9531
	DeePromClass	**0.93**	**0.91**	**0.92**	**0.92**	0.92
	SVM-LncRNAPro	0.9124	0.8769	0.8943	0.8963	0.9517
Mus musculus	XGBoost	0.8674	0.8527	0.8600	0.8611	0.9274
	LR	0.8716	0.7760	0.8211	0.8309	0.9061
	LightGBM	0.8684	**0.8630**	**0.8657**	0.8661	**0.9319**
	KNN	0.8517	0.7830	0.8159	0.8233	0.8820
	RF	0.8554	0.8505	0.8529	0.8534	0.9214
	DeePromClass	**0.93**	0.79	0.85	**0.87**	0.87
	SVM-LncRNAPro	0.8790	0.8375	0.8578	0.8611	0.9235
Homo sapiens	XGBoost	0.8416	0.8193	0.8303	0.8325	0.9044
	LR	0.8537	0.7562	0.802	0.8133	0.8769
	LightGBM	0.8410	0.8209	0.8308	0.8329	**0.9047**
	KNN	0.7942	0.7595	0.7764	0.7813	0.8446
	RF	0.8328	0.8059	0.8191	0.8220	0.8918
	DeePromClass	0.83	**0.86**	**0.84**	**0.84**	0.84
	SVM-LncRNAPro	**0.8661**	0.7875	0.8249	0.8328	0.8974

Note: Bold values indicate the highest value for each metric in each dataset.

designed initially for lncRNA promoter identification, but it can be applied to common promoter identification. The SVM-LncRNAPro approached the DeePromClass which is one of the latest methods for promoter predictions. Most computational methods or tools are specific in bioinformatics. That is to say, these methods or tools are suitable only to a specific object and otherwise are low effective.

Compared to the DeepLncPro, the SVM-LncRNAPro is much simpler, which uses only the PSTNPss to encode sequences and the SVM to learn a classifier. Therefore, the computation of the SVM-LncRNAPro is very cheap, requiring nearly no computational resources. However, there is still some room to improve for the SVM-LncRNAPro, including the interpretability of methods as well as predictive accuracy, and a more informative representation of promoter sequences.

5 | Conclusion

Although advances in the biotechnology made progress, accurately, and at a large scale, identifying lncRNA promoters

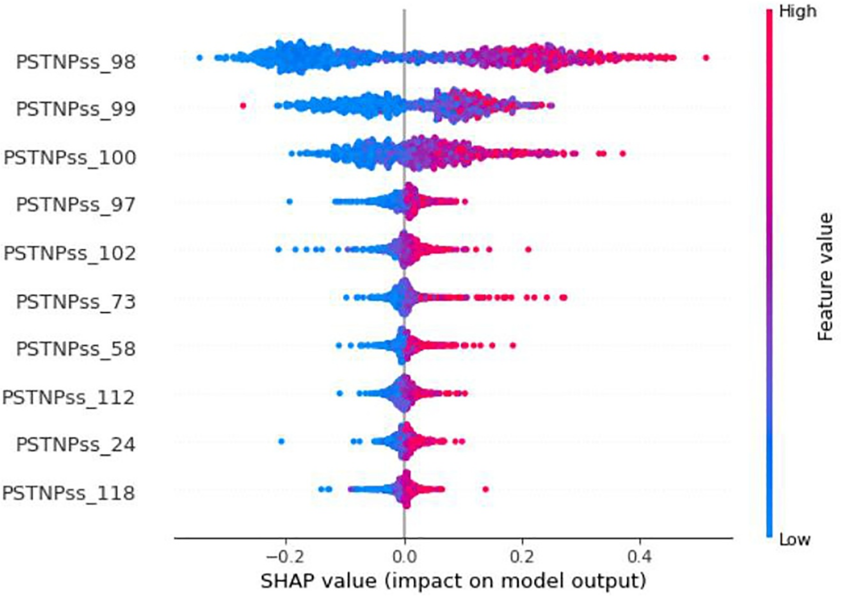

FIGURE 3 | SHAP plot of feature contributions.

remains challenging. We presented a simple but much effective method called SVM-LncRNAPro for computationally predicting lncRNA promoters. The SVM-LncRNAPro reached the state-of-the-art performance. In addition, the SVM-LncRNAPro showed a powerful ability of generalisation and thus can be used for common promoter prediction. In the future, we will explore powerful representations of the promoter sequence to enhance predictive performance. We also shall focus on the large language model to promote the interpretability of results.

Author Contributions

G.H.: conceptualisation, methodology, supervision, writing – review and editing. T.X.: data curation, investigation, formal analysis, methodology, writing – original draft, resources, software, visualisation. W.C.: supervision, methodology, conceptualisation. L.H.: methodology, validation. Q.D.: project administration, validation. J.J.: conceptualisation, project administration, supervision, writing – review and editing. All the authors read and approved the manuscript.

Ethics Statement

The authors have nothing to report.

Consent

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All the experimental data were available at https://github.com/TG0F7/Prom/tree/master.

References

1. K. Nemeth, R. Bayraktar, M. Ferracin, and G. A. Calin, “Non-Coding RNAs in Disease: From Mechanisms to Therapeutics,” *Nature Reviews Genetics* 25, no. 3 (2024): 211–232, https://doi.org/10.1038/s41576-023-00662-1.

2. L. Statello, C. Guo, L. Chen, and M. Huarte, “Gene Regulation by Long Non-coding RNAs and Its Biological Functions,” *Nature Reviews Molecular Cell Biology* 22, no. 2 (2021): 96–118, https://doi.org/10.1038/s41580-021-00330-4.

3. Y. Long, X. Wang, D. T. Youmans, and T. R. Cech, “How Do lncRNAs Regulate Transcription?,” *Science Advances* 3, no. 9 (2017): eaao2110, https://doi.org/10.1126/sciadv.aao211.

4. N. Romero-Barrios, M. F. Legascue, M. Benhamed, F. Ariel, and M. Crespi, “Splicing Regulation by Long Noncoding RNAs,” *Nucleic Acids Research* 46, no. 5 (2018): 2169–2184, https://doi.org/10.1093/nar/gky095.

5. M. Palihati and N. Saitoh, “RNA in Chromatin Organization and Nuclear Architecture,” *Current Opinion in Genetics & Development* 86 (2024): 102176, https://doi.org/10.1016/j.gde.2024.102176.

6. G. Böhmdorfer and A. T. Wierzbicki, “Control of Chromatin Structure by Long Noncoding RNA,” *Trends in Cell Biology* 25, no. 10 (2015): 623–632, https://doi.org/10.1016/j.tcb.2015.07.002.

7. P. Han and C. Chang, “Long Non-Coding RNA and Chromatin Remodeling,” *RNA Biology* 12, no. 10 (2015): 1094–1098, https://doi.org/10.1080/15476286.2015.1063770.

8. M. Heydarnezhad Asl, F. Pasban Khelejani, S. Z. Bahojb Mahdavi, L. Emrahi, A. Jebelli, and A. Mokhtarzadeh, “The Various Regulatory Functions of Long Noncoding RNAs in Apoptosis, Cell Cycle, and Cellular Senescence,” *Journal of Cellular Biochemistry* 123, no. 6 (2022): 995–1024, https://doi.org/10.1002/jcb.30221.

9. S. Dahariya, I. Paddibhatla, S. Kumar, S. Raghuwanshi, A. Pallepati, and R. K. Gutti, “Long Non-Coding RNA: Classification, Biogenesis and Functions in Blood Cells,” *Molecular Immunology* 112 (2019): 82–92, https://doi.org/10.1016/j.molimm.2019.04.011.

10. B. Xu, J. Mei, W. Ji, et al., “LncRNA SNHG3, a Potential Oncogene in Human Cancers,” *Cancer Cell International* 20 (2020): 1–11, https://doi.org/10.1186/s12935-020-01608-x.

11. G. Yang, X. Lu, and L. Yuan, “LncRNA: A Link Between RNA and Cancer,” *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1839, no. 11 (2014): 1097–1109, https://doi.org/10.1016/j.bbagrm.2014.08.012.

12. J. Liu, Y. Deng, Z. Fan, et al., "Construction and Analysis of the Abnormal lncRNA–miRNA–mRNA Network in Hypoxic Pulmonary Hypertension," *Bioscience Reports* 41, no. 8 (2021): BSR20210021, https://doi.org/10.1042/BSR20210021.

13. I. A. Qureshi, J. S. Mattick, and M. F. Mehler, "Long Non-Coding RNAs in Nervous System Function and Disease," *Brain Research* 1338 (2010): 20–35, https://doi.org/10.1016/j.brainres.2010.03.110.

14. J. Hao, W. Sun, and H. Xu, "Pathogenesis of Concanavalin A Induced Autoimmune Hepatitis in Mice," *International Immunopharmacology* 102 (2022): 108411, https://doi.org/10.1016/j.intimp.2021.108411.

15. W. Zeng, F. Wang, Y. Ma, X. Liang, and P. Chen, "Dysfunctional Mechanism of Liver Cancer Mediated by Transcription Factor and Non-coding RNA," *Current Bioinformatics* 14, no. 2 (2019): 100–107, https://doi.org/10.2174/1574893614666181119121916.

16. Y. Wang, S. Tai, S. Zhang, N. Sheng, and X. Xie, "PromGER: Promoter Prediction Based on Graph Embedding and Ensemble Learning for Eukaryotic Sequence," *Genes* 14, no. 7 (2023): 1441, https://doi.org/10.3390/genes14071441.

17. E. D. Vaishnav, C. G. de Boer, J. Molinet, et al., "The Evolution, Evolvability and Engineering of Gene Regulatory DNA," *Nature* 603, no. 7901 (2022): 455–463, https://doi.org/10.1038/s41586-022-04506-6.

18. T. Zhang, Q. Tang, F. Nie, Q. Zhao, and W. Chen, "DeepLncPro: An Interpretable Convolutional Neural Network Model for Identifying Long Non-Coding RNA Promoters," *Briefings in Bioinformatics* 23, no. 6 (2022): bbac447, https://doi.org/10.1093/bib/bbac447.

19. T. Yu, D. T. Tzeng, R. Li, et al., "Genome-Wide Identification of Long Non-Coding RNA Targets of the Tomato MADS Box Transcription Factor RIN and Function Analysis," *Annals of Botany* 123, no. 3 (2019): 469–482, https://doi.org/10.1093/aob/mcy178.

20. M. Ye, L. Xie, J. Zhang, et al., "Determination of Long Non-Coding RNAs Associated With EZH2 in Neuroblastoma by RIP-Seq, RNA-seq and ChIP-seq," *Oncology letters* 20, no. 4 (2020): 1, https://doi.org/10.3892/ol.2020.11862.

21. J. Zhang, J. Li, S. Saeed, et al., "Identification and Functional Analysis of lncRNA by CRISPR/cas9 During the Cotton Response to Sap-Sucking Insect Infestation," *Frontiers in Plant Science* 13 (2022): 784511, https://doi.org/10.3389/fpls.2022.784511.

22. J. Hain, W. D. Reiter, U. Hüdepohl, and W. Zillig, "Elements of an Archaeal Promoter Defined by Mutational Analysis," *Nucleic Acids Research* 20, no. 20 (1992): 5423–5428, https://doi.org/10.1093/nar/20.20.5423.

23. A. M. Schmitt and H. Y. Chang, "Long Noncoding RNAs in Cancer Pathways," *Cancer Cell* 29, no. 4 (2016): 452–463, https://doi.org/10.1016/j.ccell.2016.03.010.

24. W. Wu, S. Zhang, X. Li, M. Xue, S. Cao, and W. Chen, "Ets-2 Regulates Cell Apoptosis via the Akt Pathway, Through the Regulation of Urothelial Cancer Associated 1, a Long Non-Coding RNA, in Bladder Cancer Cells," *PLoS One* 8, no. 9 (2013): e73920, https://doi.org/10.1371/journal.pone.0073920.

25. E. Zhang, D. Yin, M. Sun, et al., "P53-Regulated Long Non-Coding RNA TUG1 Affects Cell Proliferation in Human Non-Small Cell Lung Cancer, Partly Through Epigenetically Regulating HOXB7 Expression," *Cell Death & Disease* 5, no. 5 (2014): e1243, https://doi.org/10.1038/cddis.2014.201.

26. P. M. Das, K. Ramachandran, J. vanWert, and R. Singal, "Chromatin Immunoprecipitation Assay," *Biotechniques* 37, no. 6 (2004): 961–969, https://doi.org/10.2144/04376RV01.

27. F. Zhu, Q. Niu, X. Li, Q. Zhao, H. Su, and J. Shuai, "FM-FCN: A Neural Network With Filtering Modules for Accurate Vital Signs Extraction," *Research: Ideas for Today's Investors* 7 (2024): 0361, https://doi.org/10.34133/research.0361.

28. W. Wang, L. Zhang, J. Sun, Q. Zhao, and J. Shuai, "Predicting the Potential Human lncRNA–miRNA Interactions Based on Graph Convolution Network With Conditional Random Field," *Briefings in Bioinformatics* 23, no. 6 (2022): bbac463, https://doi.org/10.1093/bib/bbac463.

29. J. Wang, L. Zhang, J. Sun, et al., "Predicting Drug-Induced Liver Injury Using Graph Attention Mechanism and Molecular Fingerprints," *Methods* 221 (2024): 18–26, https://doi.org/10.1016/j.ymeth.2023.11.014.

30. J. Xie, P. Xu, Y. Lin, et al., "LncRNA–miRNA Interactions Prediction Based on Meta-Path Similarity and Gaussian Kernel Similarity," *Journal of Cellular* 28, no. 19 (2024): e18590, https://doi.org/10.1111/jcmm.18590.

31. S. Yin, P. Xu, Y. Jiang, et al., "Predicting the Potential Associations Between circRNA and Drug Sensitivity Using a Multisource Feature-Based Approach," *Journal of Cellular and Molecular Medicine* 28, no. 19 (2024): e18591, https://doi.org/10.1111/jcmm.18591.

32. L. Liu, Y. Wei, Q. Zhang, and Q. Zhao, "SSCRB: Predicting circRNA-RBP Interaction Sites Using a Sequence and Structural Feature-Based Attention Model," *IEEE Journal of Biomedical and Health Informatics* 28, no. 3 (2024): 1762–1772, https://doi.org/10.1109/JBHI.2024.3354121.

33. X. Yang, J. Sun, and B. Jin, "Multi-Task Aquatic Toxicity Prediction Model Based on Multi-Level Features Fusion," *Journal of Advanced Research* 68 (2024): 477–489, https://doi.org/10.1016/j.jare.2024.06.002.

34. T. Alam, M. T. Islam, M. Househ, and S. B. Belhaouari, "Deepcnpp: Deep Learning Architecture to Distinguish the Promoter of Human Long Non-Coding Rna Genes and Protein-Coding Genes," in *Health Informatics Vision: From Data via Information to Knowledge*, Vol. 232–235 (IOS press, 2019): 477–489, https://doi.org/10.3233/SHTI190061.

35. Q. Tang, F. Nie, J. Kang, and W. Chen, "ncPro-ML: An Integrated Computational Tool for Identifying Non-Coding RNA Promoters in Multiple Species," *Computational and Structural Biotechnology Journal* 18 (2020): 2445–2452, https://doi.org/10.1016/j.csbj.2020.09.001.

36. P. Meylan, R. Dreos, G. Ambrosini, R. Groux, and P. Bucher, "EPD in 2020: Enhanced Data Visualization and Extension to ncRNA Promoters," *Nucleic Acids Research* 48, no. D1 (2020): D65–D69, https://doi.org/10.1093/nar/gkz1014.

37. Z. Chen, P. Zhao, C. Li, et al., "iLearnPlus: A Comprehensive and Automated Machine-Learning Platform for Nucleic Acid and Protein Sequence Analysis, Prediction and Visualization," *Nucleic Acids Research* 49, no. 10 (2021): e60, https://doi.org/10.1093/nar/gkab122.

38. M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support Vector Machines," *IEEE Intelligent Systems and Their Applications* 13, no. 4 (1998): 18–28, https://doi.org/10.1109/5254.708428.

39. L. Breiman, "Random Forests," *Machine Learning* 45, no. 1 (2001): 5–32, https://doi.org/10.1023/a:1010933404324.

40. I. Ruczinski, C. Kooperberg, and M. LeBlanc, "Logic Regression," *Journal of Computational & Graphical Statistics* 12, no. 3 (2003): 475–511, https://doi.org/10.1198/1061860032238.

41. N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *American Statistician* 46, no. 3 (1992): 175–185, https://doi.org/10.1080/00031305.1992.10475879.

42. G. Ke, Q. Meng, T. Finley, et al., "Lightgbm: A Highly Efficient Gradient Boosting Decision Tree," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)* (Curran Associates Inc., 2017), 3146–3154.

43. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016), 785–794.

44. W. He, C. Jia, and Q. Zou, "4mCPred: Machine Learning Methods for DNA N4-Methylcytosine Sites Prediction," *Bioinformatics* 35, no. 4 (2019): 593–601, https://doi.org/10.1093/bioinformatics/bty668.

45. H. Kari, S. M. S. Bandi, A. Kumar, and V. R. Yella, "Deepromclass: Delineator for Eukaryotic Core Promoters Employing Deep Neural Networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 20, no. 1 (2022): 802–807, https://doi.org/10.1109/TCBB.2022.3163418.

46. A. Wahab, O. Mahmoudi, J. Kim, and K. T. Chong, "DNC4mC-Deep: Identification and Analysis of DNA N4-Methylcytosine Sites Based on Different Encoding Schemes by Using Deep Learning," *Cells* 9, no. 8 (2020): 1756, https://doi.org/10.3390/cells9081756.

47. D. Lalović and V. Veljković, "The Global Average DNA Base Composition of Coding Regions May Be Determined by the Electron-Ion Interaction Potential," *Biosystems* 23, no. 4 (1990): 311–316, https://doi.org/10.1016/0303-2647(90)90013-Q.

48. H. Xu, P. Jia, and Z. Zhao, "Deep4mC: Systematic Assessment and Computational Prediction for DNA N4-Methylcytosine Sites by Deep Learning," *Briefings in Bioinformatics* 22, no. 3 (2021): bbaa099, https://doi.org/10.1093/bib/bbaa099.

49. W. Chen, H. Yang, P. Feng, H. Ding, and H. Lin, "iDNA4mC: Identifying DNA N4-Methylcytosine Sites Based on Nucleotide Chemical Properties," *Bioinformatics* 33, no. 22 (2017): 3518–3523, https://doi.org/10.1093/bioinformatics/btx479.

50. B. Liu, X. Gao, and H. Zhang, "BioSeq-Analysis2. 0: An Updated Platform for Analyzing DNA, RNA and Protein Sequences at Sequence Level and Residue Level Based on Machine Learning Approaches," *Nucleic Acids Research* 47, no. 20 (2019): e127, https://doi.org/10.1093/nar/gkz740.

51. W. Chen, P. Feng, H. Lin, and K. Chou, "iRSpot-PseDNC: Identify Recombination Spots With Pseudo Dinucleotide Composition," *Nucleic Acids Research* 41, no. 6 (2013): e68, https://doi.org/10.1093/nar/gks1450.

52. W. Chen, H. Lin, and K. C. Chou, "Pseudo Nucleotide Composition or PseKNC: An Effective Formulation for Analyzing Genomic Sequences," *Molecular BioSystems* 11, no. 10 (2015): 2620–2634, https://doi.org/10.1039/C5MB00155B.

53. P. Zhang, Y. Wu, H. Zhou, B. Zhou, H. Zhang, and H. Wu, "CLNN-Loop: A Deep Learning Model to Predict CTCF-Mediated Chromatin Loops in the Different Cell Lines and CTCF-Binding Sites (CBS) Pair Types," *Bioinformatics* 38, no. 19 (2022): 4497–4504, https://doi.org/10.1093/bioinformatics/btac575.

54. P. Yang and M. Guo, "Interaction of Some Non-Platinum Metal Anticancer Complexes With Nucleotides and DNA and the Two-Pole Complementary Principle (TPCP) Arising Therefrom," *Metal-Based Drugs* 5, no. 1 (1998): 41–58, https://doi.org/10.1155/MBD.1998.41.