EPRA International Journal of Agriculture and Rural Economic Research (ARER)- Peer-Reviewed Journal
Volume: 9 | Issue: 7| July 2021 | Journal DOI: 10.36713/epra0813| Impact Factor SJIF (2021) : 7.604| ISSN: 2321 - 7847

COMPUTATIONAL IDENTIFICATION OF PROMOTER REGIONS IN PROKARYOTES AND EUKARYOTES

[1]Sudheer Menon, [2]Shanmughavel Piramanayakam, [3]Gopal Agarwal
[1&2] Department of Bioinformatics, Bharathiar University, Coimbatore, Tamil Nadu
[3]DBEB, IIT-Delhi

Article DOI: https://doi.org/10.36713/epra7667
DOI No: 10.36713/epra7667

ABSTRACT

Promoters are modular DNA structures that contain complex regulatory elements required for the initiation of gene transcription. Therefore, the use of machine learning methods to identify promoters is very important for improving genome annotation and understanding transcriptional regulation. In recent years, many methods for predicting eukaryotic and prokaryotic promoters have been proposed. However, the performance of these methods is still far from satisfactory. In this article, we have developed a hybrid method (called IPMD) that combines a position correlation score function and diversity increment with modified Mahalanobis Discriminant to predict eukaryotic and prokaryotic promoters. The precise calculation and identification of promoters remains a challenge because these key DNA regulatory regions have variable structures composed of functional motifs that can provide gene-specific transcription initiation. The promoter is a regulatory DNA region, which is very important for gene transcription regulation. It is located near the transcription start site (TSS) upstream of the corresponding gene. In the post-genomics era, the availability of data makes it possible to build computational models to detect promoters robustly, because these models are expected to be helpful to academia and drug discovery. Until recently, the developed model only focused on distinguishing sequences into promoters and non-promoters. However, by considering the classification of weak and strong promoters, promoter predictors can be further improved.

INDEX TERMS—: ***deep learning, DNA sequence analysis, Promoter prediction, Promoters, Promoter elements***

1. INTRODUCTION

Promoters are key regions involved in protein coding and differential transcription regulation of RNA genes. The gene-specific structure of the promoter sequence makes it extremely difficult to design a general computational identification strategy. The 5'flanking region of the promoter may contain many short (5-10 bases long) motifs, which can be used as protein recognition sites to provide the initiation of transcription and specific regulation of gene expression. The smallest eukaryotic promoter region called the core promoter can initiate basic transcription and contains a transcription start site (TSS). Among all known eukaryotic promoters, approximately 30-50% of the TATA box is located ~30 bp upstream of the transcription start site. Many highly expressed genes contain a powerful TATA box in their core promoters. At the same time, a large number of genes including housekeeping genes, some oncogenes and growth factor genes have promoters that do not contain TATA. Among these promoters, Inr (promoter region) or the recently discovered downstream promoter element (DPE) (usually located ~25-30 bp downstream of TSS) can control the exact position of transcription initiation. Promoters are functional regions containing complex regulatory elements and are used to determine the initiation of gene transcription. Therefore, the use of computational techniques to predict promoters is very important for discovering genes missed by gene predictors and designing experiments to understand transcriptional regulation (Abeel et al., 2008a, b). Although many methods for promoter prediction have been developed, the performance of existing methods is still far from satisfactory. It is necessary to develop more effective methods to accurately and quickly predict promoters. It is well known that prokaryotic and eukaryotic promoters use different DNA sequences to regulate gene expression. In prokaryotes, the transcription of most genes is regulated by the r70 promoter. The r70 promoter usually contains three basic regulatory elements (Hawley and McClure 1983): the Pribnow box (or TATA box), which has a consensus TATAAT about -10 bp upstream of the transcription start site (TSS), and the -35 box. The total TTGACA is around -35. The bp upstream of the TSS and the initiator (Inr) around the TSS. In eukaryotes, all protein-coding

Volume: 9 | Issue: 7| July 2021 | Journal DOI: 10.36713/epra0813| Impact Factor SJIF (2021) : 7.604| ISSN: 2321 - 7847

genes and certain small nuclear RNAs are regulated by the pol II promoter. The core region of the pol II promoter usually contains several regulatory motifs (Pedersen et al., 1999; Bajic et al., 2004): The TATA box is located near -25 bp upstream of TSS, and the initiator and downstream promoter elements (DPE) are about 30%. bp downstream TSS. Promoter is a key region involved in protein coding and differential transcription regulation of RNA genes. The gene-specific structure of the promoter sequence makes it extremely difficult to design a general computational recognition strategy. The 5'flanking region of the promoter may contain many short (5-10 bases long) motifs, which can be used as protein recognition sites to provide transcription initiation and specific regulation of gene expression. The smallest eukaryotic promoter region is called the core promoter, which can initiate basic transcription and contains a transcription start site (TSS). Among all known eukaryotic promoters, approximately 30-50% of the TATA boxes are located upstream of the transcription start site * 30 bp. Many highly expressed genes contain a powerful TATA box in their core promoters. At the same time, a large number of genes including housekeeping genes, some oncogenes and growth factor genes have promoters that do not contain TATA. Among these promoters, Inr (promoter region) or the recently discovered downstream promoter element (DPE), usually located downstream of TSS* 25-30 bp, can control the exact position of transcription initiation. Bacterial promoters contain two short conserved sequence elements upstream of the transcription start site, which are approximately -10 and -35 nucleotides. The -10 box is absolutely necessary to start transcription in prokaryotes. The sequence of the -35 box will affect the transcription rate. Although these consensus sequences are conserved on average, they are not complete in most promoters.

Eukaryotic Promoters

Human data are taken from Genbank 90 edition (Benson et al., 1994). Specifically, all human sequences containing the feature key "prim_transcript" were selected. This function key indicates that the sequence is an unprocessed transcript, so it may contain one or more transcription starting points. From these sequences, select a sequence at least 250 bp upstream of the first transcription start point and at least 250 bp downstream, cut out 501 bp symmetrically surrounding the start point, and perform training. This produced a set of 340 sequences, 37 of which contained multiple transcription start points.

Prokaryotic Promoters

The E. coli promoter sequence was taken from the compilation of Lisser and Margalit (Lisser & Margalit 1993). This database contains 300 sequences and is superior to most other E. coli promoter databases available in two respects:

• Each sequence has been compared with the original paper, thereby minimizing the chance of database input errors.

• For each sequence, the assignment of the transcription starting point has been verified through related papers, and the most reliable

Eukaryotic Promoter Architecture

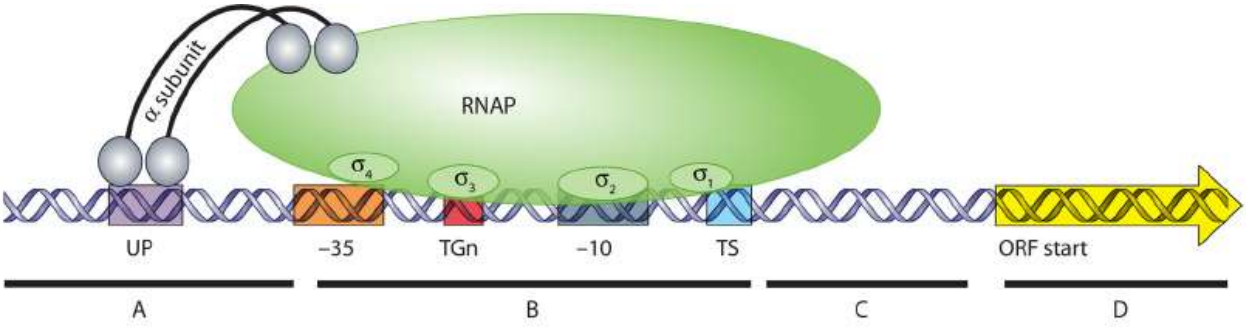
The promoter region is generally defined as any genomic DNA assembled by the transcription machinery and initiate transcription. The promoter region is composed of a protein binding region and a transcription start site (TSS). The structure of promoters in prokaryotes and eukaryotes is different in complexity. In prokaryotes, a single RNA polymerase can transcribe all types of RNA. The promoter region is characterized by the presence of -35 and -10 elements, and in some cases UP elements. In general, in prokaryotes, the regulatory region is located within 100 base pairs of TSS. In eukaryotes, the promoter structure is more complex, and its complexity is increasing from single-celled yeast to mammals. There are several different types of RNA polymerases in eukaryotes (usually three), each of which is responsible for producing a different subset of RNA. RNA polymerase II is responsible for the synthesis of all mRNA and has been thoroughly studied compared with other RNA polymerases. Therefore, only features corresponding to the promoters of genes transcribed by RNA polymerase II are discussed below. In eukaryotes, the promoter region is roughly divided into core promoter, proximal promoter and distal promoter. The actual length of the core promoter region assembled by the basic transcription mechanism is 30-100 nucleotides. These regions are characterized by the presence of sequence motifs, such as TATA boxes and Inr elements. They may also contain downstream elements, such as DPE, MTE (human) and related TSS (Juven-Gershon et al. 2008; Thomas and Chiang 2006). The proximal promoter region is a sequence within 500 base pairs relative to TSS and contains certain proximal promoter elements, including GC box, CAAT box, cis-regulatory module (CRM) (Lenhard and Sandelin 2012), etc. . The accelerator elements include reinforcing agents, insulators and silencers. The remote promoter region has no clearly defined length, and can extend from TSS to 10 kb in the upstream and downstream regions. The distal promoter interacts with the transcription activator to increase the rate of transcription. In vertebrates, 5% of genes are known to encode specific transcriptional activators, which interact with proximal and distal promoter regions.

EPRA International Journal of Agriculture and Rural Economic Research (ARER)- Peer-Reviewed Journal
Volume: 9 | Issue: 7| July 2021 | Journal DOI: 10.36713/epra0813| Impact Factor SJIF (2021) : 7.604| ISSN: 2321 - 7847

Curvature Prediction

With the help of the internal software NUCGEN (49), all the curvature calculations of the promoter sequence studied in this study were carried out. Our previous analysis showed that based on the crystal structure data of oligonucleotides (50,51), a set of dinucleotide parameters (CS) can correctly predict the curvature of synthetic and genomic DNA sequences. Therefore, CS parameters are used in the generation of DNA structures. Other analyses (A. Kanhere and M. Bansal, unpublished data) also show that for reliable curvature prediction, the window size should be at least 50 bp or greater. Therefore, we chose a window size of 75 bp for all curvature calculations. This not only allows us to estimate curvature more reliably, but also helps reduce noise. Therefore, for a promoter sequence with a length of "n" and a window size of "w" = 75 bp, we have obtained (n w +1) DNA fragments. According to (i) the radius of curvature (LSC), (ii) the ratio of the maximum component (Imax) to the minimum component (Imin) of the moment of inertia (Imax / Imin), calculate the curvature of the predicted structure of each of these segments) and (iii)) The ratio of the end-to-end distance'd' along the path traced by the DNA molecule to the contour length'lmax' (d / lmax). Because similar trends are observed for all three parameters, only the parameter d/lmax is discussed in detail.

General Promoter Architecture



Type	Mechanism	Action	TF binding			
			upstream (A)	core promoter (B)	downstream (C)	ORF (D)
repression	Steric hindrance	No RNAP binding		+		
repression	Roadblock	No transcription elongation			±	+
repression	DNA looping	No RNAP binding	+		+	±
repression	Activator modulation	Prevents activator binding	±	±		
activation	Class I	Interaction α subunit RNAP	+			
activation	Class II	Facilitates σ factor binding	+			
activation	DNA conformational change	DNA helix twist		+		
activation	Repressor modulation	Prevents repressor binding	±	±	±	±

Figure : 1 Molecular mechanism of transcription modulation. The main features of four repression and four activation types are presented. +, the TF binds at this location; +-, there are multiple places where the TF could bind. TS signifies the transcription start site, TGn signifies the extended 10 element, and UP signifies the UP element. The ORF is the gene regulated by the promoter.

Some genes are highly transcribed, while others are hardly transcribed or even not transcribed at all. This is largely due to the fact that transcriptional regulation mainly occurs during the initial binding of RNAP to DNA, the isomerization process, and the earliest stages of RNAP development along the DNA duplex (36). Since the supply of both factors and free RNAP is restricted in cells, promoters strongly compete for the binding of RNA complete enzymes (36, 192a). The binding of specific RNAP subunits plays an important role in transcription regulation. -The three main functions of the factor are (i) to ensure the recognition of core promoter elements, (ii) to locate RNAP on the target promoter, (iii) to unwind DNA near the transcription start site (321) (Figure 1)

A genome can encode many different factors. In addition to specific TFs, each factor can also determine the transcriptional response of bacterial cells by directing RNAP to a specific set of target

Volume: 9 | Issue: 7| July 2021 | Journal DOI: 10.36713/epra0813| Impact Factor SJIF (2021) : 7.604| ISSN: 2321 - 7847

genes (111). In general, bacterial housekeeping factors are similar to E. coli 70 70 kDa factors (111, 226) and regulate genes related to cell growth. Several members of the 70 factor family have been described. In addition to 70 (231), Escherichia coli K-12 has five other 70 family factors, and Bacillus subtilis has 17 known variants of 70 (274). Generally, the 70 housekeeping factors bind to 35 and 10 DNA sequence elements in the promoter, which are relatively conserved hexanucleotide sequences, with the consensus sequence TTGACA at position 35 and TATAAT at position 10 (36). The intrinsic strength of the core promoter (except for the effect of binding to other TFs, the level of transcription occurring) depends to a large extent on the matching degree of the core promoter elements with these consensus sequences (154, 157, 289). Substitution factors (including those in the 54 family) usually regulate a set of genes with well-defined functions, but their regulators may also cover a wider range of target genes involved in a variety of biological processes, and significantly overlap with housekeeping genes. Factor (306). There is also a specific factor-subfamily (ECF factor) that directly binds to extracellular environmental signals to regulate transcription (121). The substitution factors can be discussed in detail, and their various functions (111, 121, 151) can be discussed in detail. A variety of factors are usually regulated by anti-factors, which can inhibit its function under certain conditions (139).

II. TRADITIONAL METHODS

A promoter is a basic DNA element located around the transcription start site (TSS) and can regulate gene transcription. Promoter recognition is of great significance in determining transcription units, studying gene structure, analyzing gene regulation mechanisms and annotating gene function information. Many models have been proposed to predict promoters. However, the performance of these methods still needs improvement. In this work, we combined the pseudo-k-tuple nucleotide composition (PseKNC) with the position-related scoring function (PCSF) to form Homo sapiens (H. sapiens), D. melanogaster (D. melanogaster), Caenorhabditis elegans (C. elegans), Bacillus subtilis (B. subtilis) and Escherichia coli (E. coli). [1]

The promoter region is located near the transcription initiation site and regulates the transcription initiation of genes by controlling the binding of RNA polymerase. Therefore, promoter region recognition is an important area of concern in the field of bioinformatics. Many tools for promoter prediction have been proposed. However, the reliability of these tools still needs to be improved. In this work, we propose a powerful deep learning model DeePromoter to analyze the characteristics of short eukaryotic promoter sequences and accurately identify human and mouse promoter sequences. DeePromoter combines Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). [2]

Promoter identification is an important step in understanding drug development and gene transcription regulation in academia. This paper proposes a new computational model that can identify prokaryotic promoters and their strength through deep learning and pseudo-dinucleotide composition. The proposed model has been evaluated on the benchmark data set, and is superior to the current state-of-the-art model in both the promoter identification and promoter strength identification tasks. [3]

In this study, the authors investigated the possibility of predicting prokaryotic promoters by detecting evolutionary conserved motifs. We focused on the possible G-quadruplex structure upstream of AT-rich elements. The basic principle starts with the following evidence: In human, yeast, and bacterial genomes, G-quadruplexes are overexpressed in regions near the promoter [18, 19, 40, 41]. In this study, we showed that among the experimentally identified TSS, in 75.6% and 73.4% of the cases rich in Co.color A3(2) and Pseudomonas aeruginosa PA14, the four G-rich cases The AT-rich elements of the chain motif are within ±50 bp. Genome, respectively (Table 5). These high percentages support the idea that the G-quadruplex is a prototype motif involved in general promoter function/regulation. [4]

This study proved the excellent performance of the CNN model in classifying promoter and non-promoter sequences. However, the accurate identification of promoters in long genome sequences is still a major challenge, not only requires accurate classifiers, but also requires proper selection of unique predictions among multiple overlapping high-scoring genome fragments. In this task, it is also important to consider multiple or alternative promoters for each transcription unit, and it is possible to apply non-parametric methods recently described and tested on the promoter region of the model dicot plant Arabidopsis. Although we have integrated the developed CNN classifier into the promoter recognition program in the genome sequence, we will consider ways to solve many difficult aspects of this task in our follow-up research. [5]

Identifying promoters on a computer is a huge challenge in computer biology. A large number of promoter prediction programs are available, and they differ in the features used to distinguish promoter regions from a large amount of genomic sequence information. Search by structure or ensemble algorithms seem promising because they are applicable to different model systems, while hybrid

Volume: 9 | Issue: 7| July 2021 | Journal DOI: 10.36713/epra0813| Impact Factor SJIF (2021) : 7.604| ISSN: 2321 - 7847

algorithms are usually effective but limited to the availability of auxiliary experimental information (such as epigenetic features and CAGE tag counts) system. With the rapid development of high-throughput technologies, which provide genome-wide information about transcription, our understanding of promoter characteristics is changing. [6]

In prokaryotes, the nature of the problem is different. Since there is usually no splicing, it is usually simple to divide the genome into gene units. However, this cannot make the correct inference of the protein product trivial, but it is difficult to find the correct start codon in the open reading frame (ORF). In this case, although the location of the promoter is useful, it cannot provide key information useful to eukaryotes due to the presence of the polycistronic operon. [7]

This article discusses the application of soft computing technology in the field of gene prediction. Soft computing technologies, especially neural networks, seem to be powerful tools for gene prediction. This seems to be an ideal technique for combining multiple sources of information. However, the success of neural networks as a genetic prediction technology mainly depends on the type of information used as input. Genetic algorithms and hybrid techniques have given encouraging results, but they have been applied in very limited ways. Although the current soft computing technology is very helpful in identifying protein codes and ncRNA genes, since most of the work is done for specific genomes, the output results are still far from perfect. In the future technology, such as fuzzy logic, genetic algorithm, neurofuzzy and neurogenetics all need to be explored. [8]

In this article, the author has developed an effective method for eukaryotic and prokaryotic promoter prediction. Five promoters were used to evaluate the performance of the IPMD method. And achieved a higher prediction accuracy. Although this method shows good performance in promoter prediction, there is still a lot of room to improve prediction accuracy. The current study can be regarded as the first draft of the promoter annotation. Future work will focus on DNA structural information and complete genome prediction. This method can also play an important complementary role with other existing methods for predicting promoters and transcription start sites. [9]

Barrett and Palsson and Covert and colleagues predict that through an iterative model construction strategy in which subsequent iterations of high-throughput experiments and computer simulations can be completed within a few years to elucidate the regulatory network of the model organism Escherichia coli. Such an iterative method is indeed promising, because in this way, future experimental research will be effectively simplified to generate the most dense information. However, if complex regulatory mechanisms (such as those discussed in this review) play a major role in prokaryotes, the prospects given by Barrett and Palsson and Covert and colleagues may be too optimistic. A more complex model may be required to arrive at TRN with a minimum amount of inconsistency. [10]

This communication proposes a simple algorithm with high specificity and sensitivity to determine the promoter region in the human genome sequence. This method relies on non-redundant and experimentally verified promoter datasets from the Eukaryotic Promoter Database (EPD) as training parameters. The technology predicts and computationally satisfies the promoter region surrounding the gene sequence in the NCBI annotation database. [11]

Based on the atomic MD simulation of the physical potential calculated from quantum chemistry, the resulting spiral stiffness parameters reveal the complexity of the DNA deformation mode. Using these intuitive parameters at the genomic level allows us to define promoters as regions with unique deformation characteristics, especially near TSS. Using this pattern of difference, we trained a very simple method based on the Mahalanobis metric, which can locate human promoters with amazing accuracy. [12]

It is predicted that the promoter regions in prokaryotic and eukaryotic genomes have several common structural features compared with their neighboring regions, such as lower stability, higher curvature and less bending. All four sets of promoters considered here are also significantly different from non-promoter regions in single nucleotide, dinucleotide and trinucleotide composition. However, there are also some important differences between the various groups of promoters. In the case of prokaryotic sequences, the unique structural features are restricted to relatively short upstream regions compared to eukaryotic sequences, where they appear to extend in a significantly larger upstream region. In addition, compared with eukaryotic promoters, prokaryotic sequences are expected to be generally less flexible. [13]

Promoters are very complex structures, defined by many different structural features. The actual regulatory elements are usually very short, which makes their clear identification very complicated. As a result, computer simulation prediction of promoters and regulatory motifs is not simple. In addition, our understanding of general transcriptional regulation, especially organism-specific expression regulation, is still very limited. Especially for plants, there is still a need for reliable "intrinsic" genomic data that can be integrated into existing prediction tools. In this regard, we start by

EPRA International Journal of Agriculture and Rural Economic Research (ARER)- Peer-Reviewed Journal
Volume: 9 | Issue: 7| July 2021 | Journal DOI: 10.36713/epra0813| Impact Factor SJIF (2021) : 7.604| ISSN: 2321 - 7847

analyzing the CpG and CpNpG islands that are usually associated with promoters. Although several implementations of detecting such "islands" in vertebrates have been described (Ioshikhes and Zhang, 2000), the parameter settings used to detect these islands in animals cannot be used to find similarities in the Arabidopsis genome. [14]

In this article, the author shows that the hidden Markov model can learn the sequential structure that exists in both prokaryotic and eukaryotic promoter sequences. They significantly enhance the features that are obscured by the strict and gapless alignment of the transcription start point. We further introduced a new method of performing clustering experiments using HMM technology and the need to model sequential structures. This is important in many biological sequence analysis situations. Here we take the study of promoter sequences as an example. The promoter sequence is known for its strong diversity related to the recognition of individual RNA polymerase-related factors. [15]

III. METHODOLOGY

Training and testing data

In this study, in order to prove the universality of the proposed method for the problem of promoter prediction, we selected promoter sequences from a group of organisms far away: two kinds of bacteria, namely human, mouse and plant. Table 1 shows the number of studies of promoter and non-promoter sequences for each organism. We use bacterial promoter and non-promoter sequences with a length of 81 nt (nucleotides). The bacterial non-promoter sequence is taken from the corresponding genomic sequence: we randomly select the fragment of the protein-coding gene and use the opposite (non-coding) strand sequence. The E. coli σ70 promoter sequence was extracted from RegulonDB managed manually. The promoter of Bacillus subtilis is from the described collection. For human, mouse, and Arabidopsis non-promoter sequences (251 nt in size), we use random fragments of genes located after the first exon. The eukaryotic promoter sequence is extracted from the famous EPD database.

S. No.	Organism	#promoter sequences	#non-promoter sequences	Length/Location
1.	Escherichia coli s70	839	3000	81/-60 - +20
2.	Bacillus subtilis	746	2000	81/-60 - +20
3.	Human TATA	1426	8256	251/-200 - +50
4.	Human non-TATA	19811	27731	251/-200 - +50
5.	Mouse TATA	1255	3530	251/-200 - +50
6.	Mouse non-TATA	16283	24822	251/-200 - +50
7.	Arabidopsis TATA	1497	2879	251/-200 - +50
8.	Arabidopsis non-TATA	5905	11459	251/-200 - +50

Table: 1 Training and Testing Data

We used 20% of each set sequence in the test set. 70% of the remaining sequences are used as training and 10% are used as validation set. The training set provides data to generate the parameters of the CNN model, while the validation set is used to find the optimal number of learning periods (periods) that should be limited to avoid overfitting.

Convolutional Networks

The convolutional layer is the core building block of the convolutional network [20-23]. One layer is composed of filters. The filter is a small matrix (W), such as L×L×D, where D is the depth of the input data, and L is called the filter length. These filters are convolved with the input, that is, they move in space on the input and calculate a dot product for each position: W×x + b, where W is our filter and x is a small block of the input, b is the deviation. The local L×L area in the input is called the receiving field, and the distance of each step of the filter sliding on the input is called the stride. Calculating the dot product at each location provides an activation map for our filter. The next layer takes the activation maps of all filters as input. The activation map is actually partially connected neurons, they share the same weight, that is, the weight corresponding to the filter. This weight sharing is an important attribute of convolutional networks. Compared with fully connected layers, it greatly reduces the number of parameters required. The convolutional layer can be followed by another convolutional layer. In this case, the input depth is the number of filters from the previous layer. The convolutional layer is finally the pooling layer. This is a simple layer that runs on each activation map, making it smaller and easier to manage. The most common pooling technique is "Max-Pooling", which selects the largest value among multiple values for further representation. Max-Pooling enhanced convolutional layers are common among many modern deep learners [23]. They can be used to process biological sequences, because convolutional filters can capture information about functional sequence motifs.

EPRA International Journal of Agriculture and Rural Economic Research (ARER)- Peer-Reviewed Journal
Volume: 9 | Issue: 7| July 2021 | Journal DOI: 10.36713/epra0813| Impact Factor SJIF (2021) : 7.604| ISSN: 2321 - 7847

CNN architecture for building promoter recognition models

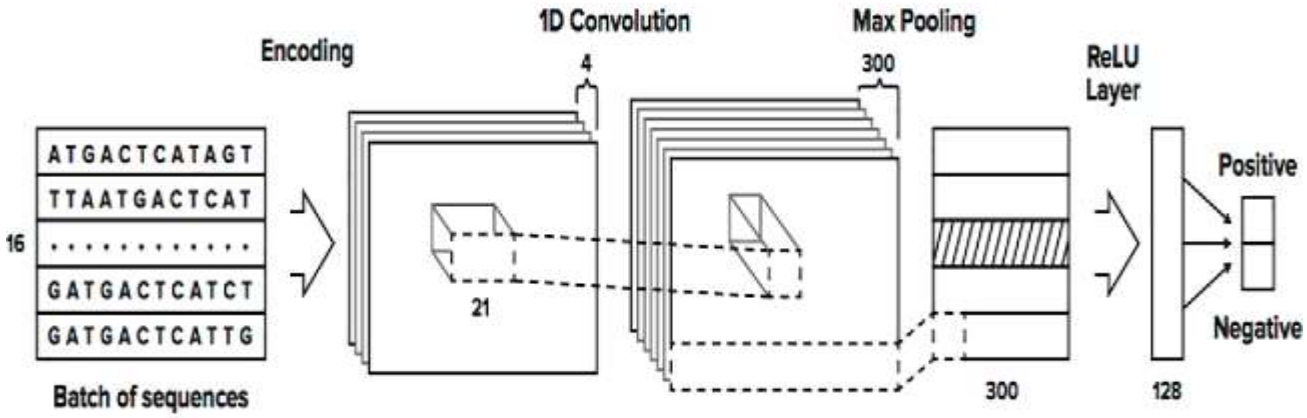


Figure : 2 Basic CNN architecture that was used in building promoter models implemented in the learnCNN.py program (see text for description)

There are many network architectures, and the task is to select the appropriate architecture for a specific research problem. In the learningCNN.py program, we use Keras to implement the CNN model, which is a minimalist and highly modular neural network library written in Python [32]. It uses the Theano library [33, 34] as a backend and utilizes GPU [35] for fast neural network training. Adam optimizer [36] is used for training of classification cross entropy as a loss function. In most cases, our CNN architecture (Figure 1) consists of only one convolutional layer and 200 filters of length 21. After the convolutional layer, we have a standard Max-Pooling layer. The output of the Max-Pooling layer is fed to a standard fully connected ReLU layer with 128 neurons. The combined size is usually 2. Finally, the ReLU layer is connected to the output layer through S-type activation, where neurons correspond to promoter and non-promoter subcategories. The batch size used for training is 16. The input of the network consists of nucleotide sequences, where each nucleotide is coded by a three-dimensional vector A (1,0,0,0), T (0,1,0,0), G (0,0,1 ,0) and C(0,0,0,1). The output is a two-dimensional vector: promoter (1, 0) and non-promoter (0, 1) predictions. Training on GTX 980 Ti GPU takes several minutes. In most cases, we intentionally use one layer of CNN architecture, but sometimes in order to strike the right balance between positive examples (initiators) and negative examples (non-initiators), two or three layers can be applied. A typical example of model calculation is shown in Figure 2.

Integrated Algorithms

For ab initio promoter prediction, it is important to select the feature with the highest discriminative power and the discriminant model (statistical model). Some programs integrate different functions to achieve better predictions (Zeng et al., 2010). ARTS (Sonnenburg et al., 2006), CoreBoost (Zhao et al., 2007), PromoterExplorer (Xie et al., 2006) and SCS (Zeng et al., 2010) are just a few examples of such new-generation algorithms. Two or more features are used to predict the promoter. PPP, such as MetaProm (Wang and Ungar 2007), integrates many algorithms to predict promoters. Compared with the algorithm described earlier, the integrated algorithm is usually a better promoter region identifier.

Hybrid Methods

Hybrid PPP has been developed recently. In addition to the inherent characteristics of the promoter sequence, they also use experimental information, such as gene expression and histone modification data (Wang et al., 2012). CoreBoost_HM (Wang et al., 2009) and the method of enriching data using ChIP-seq Pol-II (Gupta et al., 2010) belong to the category of hybrid PPP. CoreBoost_HM integrates specific histone modification profiles and DNA sequence features (core promoter elements, TFBS, flexibility) to predict the human Pol II promoter. Similarly, another recent method integrates gene expression data of Chip-seq and CAGE methods (average number of tags per million and maximum number of tags) and DNA sequence characteristics (10 sequence composition variables and 22 attribute variables) to predict humans In the promoter region. In terms of sensitivity and specificity, these two methods are superior to earlier methods.

CONCLUSIONS AND FUTURE SCOPE

Identifying promoters on a computer is a huge challenge in computer biology. A large number of promoter prediction programs are available, and they differ in the features used to distinguish promoter regions from a large amount of genomic sequence information. Search by structure or ensemble algorithms seem promising because they are applicable to different model systems, while hybrid algorithms are usually effective but limited to the availability of auxiliary experimental information (such as epigenetic features and CAGE tag counts) system. With the rapid development of high-throughput technologies, which provide genome-wide information about transcription, our understanding of promoter characteristics is changing. Promoter identification is an important step in understanding drug discovery and gene transcription regulation in academia. This paper proposes a new computational model that can identify prokaryotic promoters and their strength through deep learning and pseudo-dinucleotide composition. The proposed model has been evaluated on the benchmark data set, and is better than the current state-of-the-art model in both the promoter identification and promoter strength identification tasks.

REFERENCES

1. *Hong-Yan Lai, Zhao-Yue Zhang, Zhen-Dong Su, Wei Su, Hui Ding, Wei Chen, and Hao Lin "iProEP: A Computational Predictor for Predicting Promoter" American Socity of Gene & Cell Therapy 2019.*
2. *Mhaned Oubounyt, Zakaria Louadi, Hilal Tayara, Kil To Chong "DeePromoter: Robust Promoter Predictor Using Deep Learning" Grontiers in Genetics 2019.*
3. *Hilal Tayaraa , Muhammad Tahira,b , Kil To Chong "Identification of prokaryotic promoters and their strength by integrating heterogeneous features" Elsevier 2019.*
4. *Marco Di Salvo, Eva Pinatel, Adelfia Talà, Marco Fondi, Clelia Peano and Pietro Alifano "an algorithm for predicting transcription promoters in GC-rich bacterial genomes based on AT-rich elements and G-quadruplex motifs" BMC Bioinformatics 2018.*
5. *Ramzan Kh. Umarov, Victor V. Solovyev "Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks" PLOS 2017.*
6. *Venkata Rajesh Yella and Manju Bansal "In silico Identification of Eukaryotic Promoters" Research Gate 2014.*
7. *James W. Fickett and Artemis G. Hatzigeorgiou "Eukaryotic Promoter Recognition" Cold Spring Harbor Laboratory Press 2014.*
8. *Neelam Goel, Shailendra Singh, and Trilok Chand Aseri "A Review of Soft Computing Techniques for Gene Prediction" Hindawi 2013.*
9. *Hao Lin and Qian-Zhong Li "Eukaryotic and prokaryotic promoter prediction using hybrid approach" Theory Biosci 2011.*
10. *Sacha A. F. T. van Hijum, Marnix H. Medema, and Oscar P. Kuipers "Mechanisms and Evolution of Control Logic in Prokaryotic Transcriptional Regulation" Microbiology and Molecular Biology Review 2009.*
11. *Q. M. Alfred, K. Bishayee , P. Roy and T. Ghosh "A computational technique for prediction and visualization of promoter regions in long human genomic sequences" Journal of Bioinformatics and Sequence Analysis 2009.*
12. *J Ramon Goñi, Alberto Pérez, David Torrents and Modesto Orozco "Determining promoter location based on DNA structure first-principles calculations" Genome Biology 2007.*
13. *Aditi Kanhere and Manju Bansal "Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes" Nucleic Acids Research 2005.*
14. *Stephane Rombauts, Kobe Florquin, Magali Lescot, Kathleen Marchal, Pierre Rouze, and Yves Van de Peer "Computational Approaches to Identify Promoters and cisRegulatory Elements in Plant Genomes" American Society of Plant Biologists 2003.*
15. *Anders Gorm Pedersen, Pierre Baldi, Soren Brunak and Yves Chauvin "Characterization of Prokaryotic and Eukaryotic Promoters Using Hidden Markov Models" 1996.*