ijlpr 2023; doi 10.22376/ijlpr.2023.13.5.L267-L273

International Journal of Life science and Pharma Research

ISSN 2250-0480

Research Article

Genetic Factors for Neural Network

Features of Genetic Sigma Factors Learning Model Simulating Neural Network

Sasikala S[1*] and Dr. Ratha Jeyalakshmi T[2]

[1] *Research Scholar, Sri Sarada College for Women, Tirunelveli, Tamilnadu, Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli, Tamilnadu, India, PIN - 627012*

[2] *Research Supervisor, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli Tamilnadu, India, – 627 012)*

Abstract: Sigma factors play a crucial role in the gene regulation process, which binds with RNA polymerase to unwind the gene sequence by identifying the recognition of transcription starting motif pattern, a combination of nucleobases (Adenine(A), Cytosine(C), Guanine(G), Thymine(T)). The advancements in DNA analysis are useful for the geneticist to learn different patterns from different perspectives for the identification of mutations in genetic structures, new organisms ranging from unicellular to multicellular, and useful for creating new gene patterns to get relief from hereditary diseases. To meet all these challenging needs, the proposed research work aimed to predict the DNA motif patterns of various sigma factors. Thus the main objective is to create a novel method named "Features of Genetic Sigma Factors Learning Model simulating Neural Network" to predict the prefix motif patterns of major sigma factors such as sigma 70, sigma 32, sigma 24, sigma 19, and sigma 38(σ^{70}, σ^{32}, σ^{24}, $\sigma^{19,}$ *and* σ^{38}). Each of the sigma factors possesses significant functionality, like vegetative growth for the development of nutrients. In the proposed model, the novel idea is a generation of a dictionary of DNA motifs that mimics the n-gram of natural language processing. The proposed model is trained to feed the DNA motif dictionary, which consists of positive and negative motif patterns. The model is tested by an array of K-mer motif patterns taken from the whole *E.Coli* bacterial genome, downloaded from the NCBI website (Escherichia coli str. K-12 substr. MG1655, complete genome ACCESSION: NC_000913). The experimental results of the proposed model yielded 100% accuracy. The model's outcome is a set of patterns that are highly helpful to experts in the biological fields to identify new gene patterns.

Keywords: sigma factors; motif prediction; LSTM; DNA binding motif; pattern learning

***Corresponding Author**

Sasikala S , Research Scholar, Sri Sarada College for Women, Tirunelveli, Tamilnadu, Affiliated to Manonmaniam Sundaranar University, Abishekapatti,Tirunelveli, Tamilnadu, India, PIN - 627012

Received On 17 February, 2023

Revised On 12 April, 2023

Accepted On 3 May, 2023

Published On 1 September, 2023

Funding This research did not receive any specific grant from any funding agencies in the public, commercial or not for profit sectors.

Citation Sasikala S and Dr. Ratha Jeyalakshmi T , Features of Genetic Sigma Factors Learning Model Simulating Neural Network.(2023).Int. J. Life Sci. Pharma Res.13(5), L267-L273 http://dx.doi.org/10.22376/ijlpr.2023.13.5.L267-L273

1. INTRODUCTION

Gene is a region of DNA composed of four nucleobases (A-Adenine, C-Cytosine; G- Guanine, and T-Thymine). There are three regions in a gene: promoter, coding region, and terminator region. A promoter is a short DNA motif of the gene regulation process onto which the transcription mechanism holds and instigates the transcription process. Normally promoters reside near the transcription start site, but the region of the promoter is not a constant one. Identifying promoters is an important biological task, given that they are central to understanding how genes are regulated. Promoters reside upstream of the genes they regulate. Though the promoters fluctuate among prokaryotic genomes, some elements are conserved at the -10 and -35 regions upstream of the beginning site; two promoter consensus motifs are similar across all promoters and various bacterial species. The motif of -10 regions, TATAAT, and the -35 sequence, TTGACA, is recognized and bound by the sigma factor (*σ)*. The Sigma factors represent the specificity of promoter DNA binding and control how efficiently RNA synthesis is started. Once this interaction is made, the subunits of the core enzyme stick to the site. The AT-rich -10 region facilitates the unwinding of the DNA template, and various phosphodiester bonds are made. Various algorithms have been constructed to predict the promoters bound by sigma factors. A motif is a small sequence used to classify a promoter with a sigma factor responsible for gene expression, a protein synthesis process. A layered structure of a self-organizing Neural Network is applied[1] to identify motifs in DNA sequences. CNN-BiLSTM model was proposed[2] to explore the potential contextual relationships of amino acid sequences and to obtain more other features. The triad pattern algorithm took a UP-element, required for interaction with the α subunit, optimally separated patterns of -35 and -10 boxes, required for interaction with the σ^{70} subunit of RNA polymerase, and was developed[3] for predicting strong bacterial promoters. Seven sigma factors were found in *E.Coli* bacteria, and ten were found in Bacillus subtilis bacteria[4]. The consensus binding sites of different sigma factors prefix motifs were discovered as they are vital for transcription. For the gene sequence classification, a machine learning approach named weightily averaged one-dependence estimators (WAODE) was conceived[5]. A generalization of a nonlinear model based on Information Theory[6] that measured two parametric uncertainty estimators for each TFBS, which were the total amount of information change produced by assuming position independence. In contrast, the second estimator measured the total amount of change of per-position mutual information. Three models, CNN, CNN-LSTM, and CNN-Bidirectional LSTM[7], were proposed using Label and k-mer encoding for DNA sequence classification. A method for predicting the main genome sequences of SARS-CoV-2[8] was implemented using the deep learning architecture. An attention-based Bi-LSTM+CNN hybrid model[9] that focused on the advantages of LSTM and CNN with an additional attention mechanism to classify the movie review data and trained the model using the Internet Movie Database (IMDB) movie review data to evaluate the performance of the proposed model, and the test results obtained more accurate classification result. Two different deep learning based methods[10] for identifying DNA-Binding Proteins (DBPs): DeepDBP-ANN and DeepDBP-CNN. The DeepDBP-ANN used a generated set of features trained on a traditional neural network, and DeepDBP-CNN used a pre-learned embedding and Convolutional Neural Network. To predict meaningful labels from small motif sequences, a Deep learning approach[11] was suggested. The algorithm CNN-MGP[12] showed the ability of deep learning to predict genes in meta genomics fragments; for the first time[13], it was attempted to design, implement, and test deep bidirectional long short-term memory based sequence to sequence (Bi-LSTM S2S) regression approach. A long short-term memory deep learning (LSTM) network[14,15] was introduced to recognize emotions using EEG signals. The best-performing architectures by varying CNN width, depth, and pooling designs were introduced[16]. The overfitting problem can be resolved using the dropout technique, which improves significantly over other regularization methods. It was shown that dropout enhances[17] the performance of neural networks on supervised learning methods in vision, document classification, and computational biology. For identifying motifs that abstract the task of finding short conserved sites in genomic DNA. The planted (l, d)-motif problem, PMP, is the mathematical abstraction of finding a substring of length l; an algorithm was proposed[18] that combined the voting algorithm and pattern matching algorithm to find exact motifs. The notion of regulatory motifs was generalized[19] from computational biology and a new methodology with a custom-designed node applied for gene expression prediction[20,21,22]. There are seven sigma factors found[4], each with a specific role, such as sigma 70(σ^{70}) responsible for housekeeping, sigma 54 (σ^{54}) for nitrogen metabolism, sigma 32 (σ^{32}) for heat shock, sigma 24 (σ^{24}) for extreme heat shock, sigma 19(σ^{19}) for iron transport and sigma 38 (σ^{38}) for the stationary phase, sigma 28(σ^{28}) for Flagellar proteins with the respective recognition motifs prefix as 'ttgaca', 'ctggcac', 'cttgaa', 'gaactt', 'ggaaat' and 'ttgaca'.' As the size of the prokaryotic genome varies from 2kb to over 1 Mb, it is challenging to generate a model for the whole genome to perform the gene regulation process. The proposed work aims to create a model to identify the sigma factors prefix patterns in the whole genome. To make the process an effective one, a novel idea is implemented that mimics the natural language process. A dictionary of unigram sigma factors patterns, including positive and negative sequences, is generated, which plays a crucial role in the model implementation. The raw genome sequence is preprocessed and encoded using the unique codes assigned to the prefix motifs. This paper concentrated on the five sigma factors prefix sequences, and the Features of the Sigma Factors Learning Model are constructed using Python for predicting the patterns of small prefix motifs of sigma factors (σ^{70}, σ^{32}, σ^{24}, σ^{19}, and σ^{38}). The whole genome of *E.Coli* bacteria is taken as a dataset for predicting the motif described above sequence of various sigma factors. We achieved 100% accuracy from the experimental results of the proposed model, and our model can be extended for any pattern-matching tasks with different length sequences.

2. MATERIALS AND METHODS

Protein synthesis takes place to transcript the encoded gene for certain functionality. For transcription initiation, RNA polymerase binds with the promoter with the help of the sigma factor. Hence the role of the sigma factor, a subunit of RNA polymerase, is significant for unwinding the gene sequence. The proposed work was done with the whole genome of *E.Coli* bacteria (Escherichia coli str. K-12 substr. MG1655, complete genome) with about 46, 41,652 nucleobases downloaded from the NCBI website.

2.1. Model Construction

The proposed model is heavily focused on model construction in which motif corpus generation plays a vital role. Only the model can learn the sigma factors' motif and non-motif sequence features. The model is constructed in two phases.

2.1.1. Phase I

In phase I, the DNA motif dictionary is constructed to learn the features of sigma factors motif vocabulary, which is shown in Fig. 1. The model considers the recognition motif prefix of five sigma factors of (σ^{70}, σ^{38}, σ^{32}, σ^{24}, and σ^{19}) provided σ^{70} and σ^{38} share a common prefix.

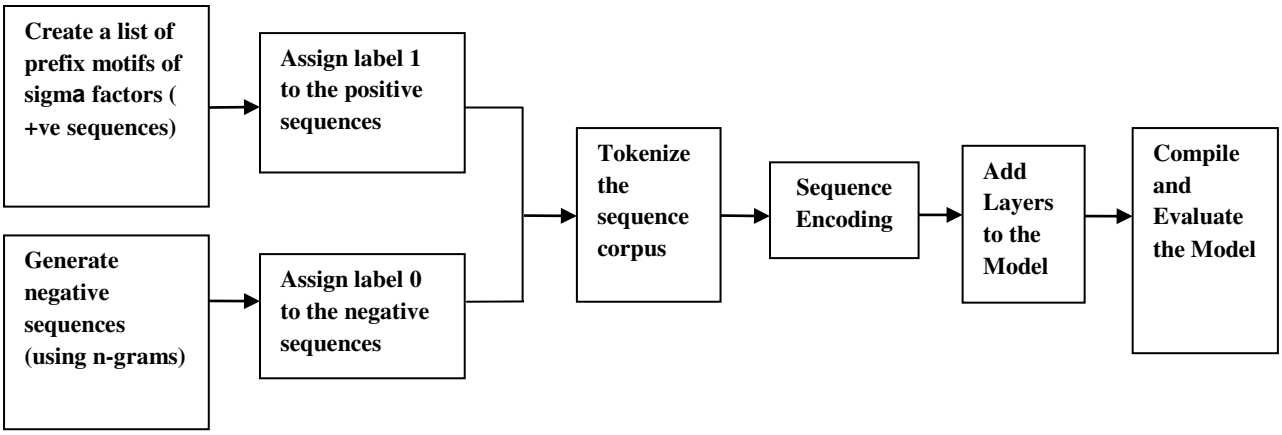


Fig 1: Design of Features of Genetic Sigma Factors Learning Model

The length of each of these prefix motifs is equal and treated as positive and labeled as 1. The negative sequences are synthetically patterned as n-grams that resemble motif sequences' vocabulary. The dictionary generation is crucial for feature identification while implementing the prediction model. The motif sequences are uniquely tokenized to make the model understand the sigma factors patterns. The proposed neural network model is a deep learning model constructed with LSTM (Long Short Term Memory), which involves different layers such as the embedding layer, Spatial1D Dropout layer, LSTM, and dense layer. The following Fig. 2 explains the layered architecture of the proposed model. The sigmoid function classifies whether it is a motif pattern of the sigma factor.

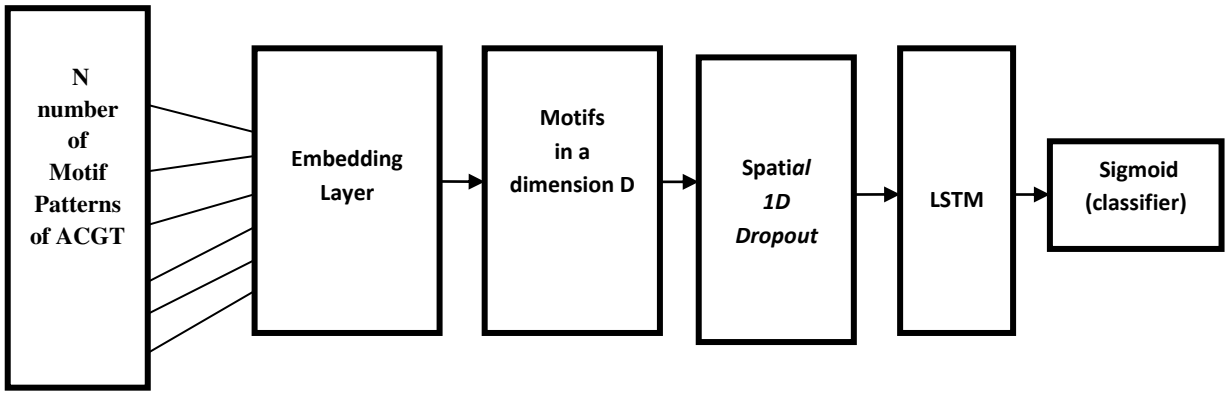


Fig 2: Architecture of Features of Genetic Sigma Factors Learning Model

(1) Embedding Layer

This layer accepts the pre-generated k-mer motif patterns as a motif corpus which consists of both positive and negative patterns. The corpus is encoded using a tokenizer () to make the model for understanding the sequence patterns. The outcome of this layer is a formatted sequence in a specific dimension.

(2) Spatial 1D Dropout Layer

The spatial 1D Dropout layer makes the encoded sequence independent of each other, which ease the classification process.

(3) LSTM Layer

In recent studies, Long Short Term Memory (LSTM) plays a vital role in predicting sequences in deep learning because it remembers important patterns for a long span of time[7,21]. The LSTM layer is defined as a cell state with various components; the forget gate and input gate are used to identify the information to be passed through or not to manage the current information.

(4) Dense Layer

The output of the LSTM is passed to the dense layer, which uses the sigmoid function as a classifier represented in the equation (1). The final output of the proposed model classifies the existence of the five sigma factors.

$$\text{Sigmoid} = \frac{e^x}{e^x + 1} \quad (1)$$

The binary cross entropy loss function is a convenient choice for estimating the loss function used in a binary classification, as shown in equation (2).

$$\text{Binary cross entropy (loss)} = -\frac{1}{N}\sum_{i=1}^{N} y_i . \text{Log}\, \hat{y}_{i+} (1 - y_i) . \log (1 - \hat{y}_i \quad (2)$$

Where N is the size of the data set, yi is i^{th} target output, and ŷi is the i^{th} calculated output value.

2.1.2. Phase 2

In Phase 2, the downloaded bacteria genome is preprocessed to generate an array of patterns P for a k-mer motif using the equation (3).

$$P = \{p_i : p_{i+k-1}\} \quad (3)$$

Where ,$0 \leq i \leq (N-k)$; N - genome size; k - the size of a motif (k-mer sequence)

For example, if the genome sequence length is 40 and the k-mer size is 6, the patterns are generated, as illustrated in Table 1. Then the generated patterns are tokenized and given for testing using the model generated in phase 1.

Table1: Patterns Generation (for 6-mer motif)

Sequence	a c g t a g g c t a g c t t a c g t g c c a a c g a c g t a c t g g t a c c t g
Patterns generated	a c g t a g
	c g t a g g
	g t a g g c
	t a g g c t
	a g g c t a
	g g c t a g
	… … …
	t a c c t g

Here it shows how to split a whole genome into motif patterns of length 6

2.1.3. Data Set

The benchmark dataset used during the current study is available in the following Link (NCBI Website)

https://www.ncbi.nlm.nih.gov/nuccore/NC_000913.3?report=fasta

DEFINITION: Escherichia coli str. K-12 substrate. MG1655, complete genome
ACCESSION: NC_000913
VERSION: NC_000913.3

3. STATISTICAL ANALYSIS

The whole genome of *E.Coli* Bacteria is preprocessed using equation (3) and further processed and evaluated using the model, which is developed in Python.

4. RESULTS AND DISCUSSION

To train the model, the prefix sequences of five sigma factors (σ^{70}, σ^{38}, σ^{32}, σ^{24}, and σ^{19}) of *E.Coli* bacteria are taken as positive motifs, and the negative motifs are synthetically generated. The proposed Sigma Factors Learning Model tests the computationally processed and confirmed *E.Coli* bacteria's Genome to predict trained motif patterns.

4.1. Experimental Outcome

The *E.Coli* bacteria genome sequence is processed to generate *6*-mer motif patterns and encoded, which are evaluated using the Features of the Sigma Factors Learning Model. The model learns the features of sigma factors and identifies them with a maximum accuracy level of 100%. The following are the experimental results of the model.

- ***OUTPUT 1***

Features of Genetics Sigma Factors Learning Model Outcome Summary
Model: "sequential_2"

_____

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 1, 4)	20000
spatial_dropout1d_2 (SpatialDropout1D)	(None, 1, 4)	0
lstm_2 (LSTM)	(None, 50)	11000
flatten_2 (Flatten)	(None, 50)	0
dense_2 (Dense)	(None, 1)	51

Total params: 31,051
Trainable params: 31,051
Non-trainable params: 0

_____

None

Training Accuracy

Accuracy: 100.00

145052/145052 [==============================] - 390s 3ms/step

Metrics of the Features of Genetic Sigma Factors Learning Model

Precision: 100.000

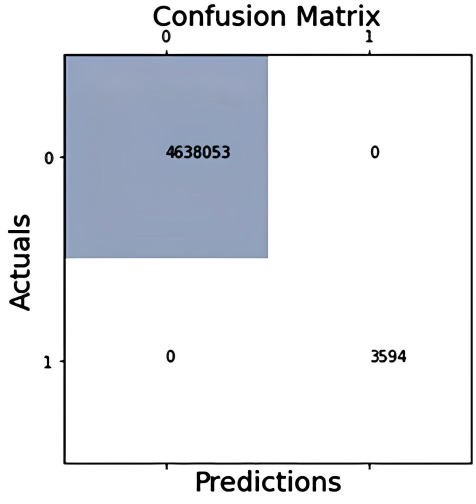
Recall: 100.000

Accuracy: 100.000

F1 Score: 100.000

<Figure size 432x288 with 0 Axes>

- ***OUTPUT 2***



4.2. Performance Analysis

The different metrics, including accuracy, precision, recall, and F1Score, are measured, which are defined based on true positive, false positive, true negative, and false negative predictions as follows.

Accuracy = TP+ TN/(TP+TN+FP+FN)

Precision = TP/ (TP+FP)
Recall = TP/ (TP+FN)
F1Score = (2 x Recall x Precision) / (Recall + Precision)

- ***True Positive (TP)***

It represents several motif patterns that are correctly predicted as true.

- ***True Negative (TN)***

It represents the number of motif patterns that are correctly predicted as false.

- ***False Positive (FP)***

It represents several motif patterns that are incorrectly predicted as true, which are actually to be false.

- ***False Negative (FN)***

It represents several motif patterns that are incorrectly predicted as false, which are actually to be true.

The proposed model has split the whole genome sequence into 4,641647 numbers of 6-mer motifs. The confusion matrix in output 2 showed no false positive and false negative prediction, which means the true motif patterns are correctly predicted as true. The false motif patterns are correctly predicted as false.

Table 2: Metrics of the Proposed Model

Metrics	Features of Genetic Sigma Factors Learning Model
Accuracy	100%
Precision	100%
Recall	100%
F1-Score	100%

The performance metrics of the proposed model have achieved 100%, as shown in Table 2. In addition, table 3 shows the proposed model has gained a higher level of accuracy than the other models[7,10].

Table 3: Performance Comparison

Metrics	DeepDBP-ANN	DeepDBP-CNN	CNN	LSTM	Bidirectional LSTM	Proposed Model
Accuracy	**82.8%**	**84.31 %**	**93.16 %**	**93.02 %**	**93.13 %**	**100 %**

In Fig. 3, it is shown that the proposed Genetic Sigma Factors Learning Model achieved a 100% accuracy level than other methods.

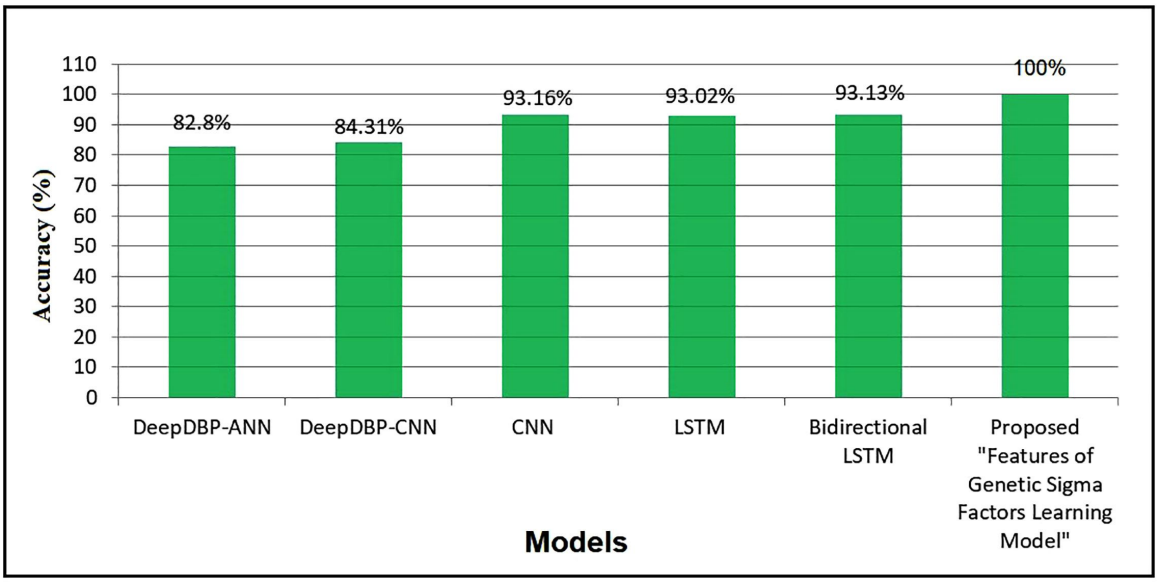


Fig.3 Graph compares the accuracy of the proposed model with other models.

5. CONCLUSION

The Sigma factors play a vital role in locating the specificity of the promoter DNA binding site and determining the efficiency of the RNA synthesis process. Though more algorithms have been suggested earlier, a gap still exists to be resolved. The proposed Features of the Sigma Factors Learning Model is a novel method that predicts the five Sigma factors recognition motif in the *E.Coli* bacterial genome. This model uses the n-grams approach to generate positive and negative motif patterns corpus and effectively applies a tokenizer for encoding, as the neural network works only on numerical data. The whole bacterial genome is pre-processed and split into k-mer motif patterns. Finally, the patterns were tested and showed that this model accurately gained a higher performance than the other classifiers. Geneticists and biologists can use this proposed model to learn different patterns of DNA sequences from different perspectives for the identification of mutations in genetic structures and new organisms and useful for creating new gene patterns for

curing genetic diseases. This model can be extended in the future by taking n-gram patterns and applied for any text pattern prediction.

6. AUTHORS CONTRIBUTION STATEMENT

Sasikala S conceived and carried out the project. Dr, Ratha Jeyalakshmi T supervised the project. Both authors contributed to the analysis of the results and the writing of the manuscript.

7. CONFLICT OF INTEREST

Conflict of interest declared none.

8. REFERENCES

1. Liu D, Xiong X, Dasgupta B, Zhang H. Motif discoveries in unaligned molecular sequences using self-organizing neural networks. IEEE Trans Neural Netw. 2006 Jul 1;17(4):919-28. doi: 10.1109/TNN.2006.875987 ((ISBN: 1045- 9227)). PMID 16856655.
2. Hu S, Ma R, Wang H. An improved deep learning method for predicting DNA-binding proteins based on contextual features in amino acid sequences. PLOS ONE. 2019 Nov 14;14(11):e0225317. doi: 10.1371/journal.pone.0225317, PMID 31725778.
3. Dekhtyar M, Morin A, Sakanyan V. Triad pattern algorithm for predicting strong promoter candidates in bacterial genomes. BMC Bioinformatics. 2008 Dec;9:233. doi: 10.1186/1471-2105-9-233, PMID 18471287.
4. Available from: https://www.sciencedirect.com/topics/neuroscience/sigma-factor#:~:text=The%20seven%20sigma%20factors%20of,that%20are%20encoded%20by %20bacteriophage.
5. Htike ZZ, Win SL. Recognition of promoters in DNA sequences using weightily averaged one-dependence estimators. Procedia Comput Sci. 2013 Jan 1;23:60-7. doi: 10.1016/j.procs.2013.10.009.
6. Maynou J, Pairó E, Marco S, Perera A. Sequence information gain-based motif analysis. BMC Bioinformatics. 2015 Dec;16(1):377. doi: 10.1186/s12859-015-0811-x, PMID 26553056.
7. Gunasekaran H, Ramalakshmi K, Rex Macedo Arokiaraj A, Deepa Kanmani S, Venkatesan C, Suresh Gnana Dhas C. Analysis of DNA sequence classification using CNN and hybrid models. Comp Math Methods Med. 2021 Jul 15;2021:1835056. doi: 10.1155/2021/1835056, PMID 34306171.
8. Lopez-Rincon A, Tonda A, Mendoza-Maldonado L, Mulders DGJC, Molenkamp R, Perez-Romero CA et al. Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning. Sci Rep. 2021 Jan 13;11(1):947. doi: 10.1038/s41598-020-80363-5, PMID 33441822.
9. Jang B, Kim M, Harerimana G, Kang SU, Kim JW. Bi-LSTM model to increase accuracy in text classification: combining Word2vec CNN and attention mechanism. Appl Sci. 2020 Aug 24;10(17):5841. doi: 10.3390/app10175841.
10. Shadab S, Alam Khan MT, Neezi NA, Adilina S, Shatabdi S. DeepDBP: deep neural networks for identifying DNA-binding proteins. Inform Med Unlocked. 2020 Jan 1;19:100318. doi: 10.1016/j.imu.2020.100318.
11. Busia A, Dahl GE, Fannjiang C, Alexander DH, Dorfman E, Poplin R, et al. A deep learning approach to pattern recognition for short DNA sequences. bioRxiv. 2018 Jun 22:353474. doi: 10.1101/353474 Corpus ID: 90436562.
12. Al-Ajlan A, El Allali A. CNN-MGP: convolutional neural networks for metagenomics gene prediction. Interdiscip Sci Comp Life Sci. 2019 Dec;11(4):628-35. doi: 10.1007/s12539-018-0313-4, PMID 30588558.
13. Mughees N, Mohsin SA, Mughees A, Mughees A. Deep sequence to sequence Bi-LSTM neural networks for day-ahead peak load forecasting. Expert Syst Appl. 2021 Aug 1;175:114844. doi: 10.1016/j.eswa.2021.114844.
14. Sakalle A, Tomar P, Bhardwaj H, Acharya D, Bhardwaj A. An LSTM-based deep learning network for recognizing emotions using a wireless brainwave-driven system. Expert Syst Appl. 2021 Jul 1;173:114516. doi: 10.1016/j.eswa.2020.114516.
15. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modeling techniques for genomics. Nat Rev Genet. 2019 Jul;20(7):389-403. doi: 10.1038/s41576-019-0122-6, PMID 30971806.
16. Zeng H, Edwards MD, Liu G, Gifford DK. Convolutional neural network architectures for predicting DNA–protein binding. Bioinformatics. 2016 Jun 15;32(12):i121-7. doi: 10.1093/bioinformatics/btw255, PMID 27307608.
17. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014 Jan 1;15(1):1929-58.
18. Abbass MM, Bahig HM. An efficient algorithm to identify DNA motifs. Math Comput Sci. 2013 Dec;7(4):387-99. doi: 10.1007/s11786-013-0165-6.
19. Syed Z, Stultz C, Kellis M, Indyk P, Guttag J. Motif discovery in physiological datasets: a methodology for inferring predictive elements. ACM Trans Knowl Discov Data (TKDD). 2010 Jan 18;4(1):1-23. doi: 10.1145/1644873.1644875.
20. Eetemadi A, Tagkopoulos I. Genetic Neural Networks: an artificial neural network architecture for capturing gene expression relationships. Bioinformatics. Jul 1, 2019;35(13):2226-34. doi: 10.1093/bioinformatics/bty945, PMID 30452523.
21. Wang H, Li C, Zhang J, Wang J, Ma Y, Lian Y. A new LSTM-based gene expression prediction model: L-GEPM. J Bioinform Comp Biol. 2019 Aug 29;17(04):195002.
22. Dizaji KG, Chen W, Huang H. Deep large-scale multitask learning network for gene expression inference. J Comput Biol. 2021 May 1;28(5):485-500. doi: 10.1089/cmb.2020.0438, PMID 34014778.