ISSN : 2350-0743

International Journal of Recent Advances in Multidisciplinary Research

Vol. 12, Issue 05, pp.11255-11258, May, 2025

RESEARCH ARTICLE

TRANSCRIPTION FACTOR BINDING SITE PREDICTION: TOOLS, TECHNIQUES, AND APPLICATIONS

***Ragini Kushwaha**

Division of Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute,
New Delhi-110012

ARTICLE INFO

Article History

Received 20th February, 2025
Received in revised form
27th March, 2025
Accepted 26th April, 2025
Published online 30th Mav, 2025

Keywords:

TF, TFBS, Bioinformatics tools, Chipseq etc.

****Corresponding author:***
Ragini Kushwaha

ABSTRACT

Transcription factors (TFs) are pivotal regulatory proteins thatbind to DNA sequences called transcription factor binding sites (TFBSs) to regulate the expression of genes. When, where, and how genes are active or repressed are all fundamentally determined by these binding sites. An overview of the biological significance of TFBSs, their role in gene regulation, and the consequences of their failure in disease are given in this article. The study also examines important uses of TFBS prediction in plant and animal biology, such as evolutionary analysis, illness research in animals, and crop stress tolerance. Different methodologies including bioinformatics tools for TFBS identification have been discussed in this article. Then using the JASPAR database and the BertSNR deep learning model, a case study involving the transcription factor ZNF594 is provided, showcasing contemporary computational methods in TFBS prediction.

Citation: Ragini Kushwaha. 2025. "Transcription Factor Binding Site Prediction: Tools, Techniques, and Applications". *International Journal of Recent Advances in Multidisciplinary Research,* 12, (05), 11255-11258.

INTRODUCTION

Understanding the basic concept of Central Dogma of Molecular Biology (CDMB) is crucial to gain insights of transcription factor binding sites. CDMB describes the flow of genetic information: DNA → RNA → Protein. DNA contains the instructions for making proteins and the process by which these instructions are converted from DNA into messenger RNA (mRNA) is called transcription. Then, the mRNA is translated into proteins. From serving as structural elements to catalysing metabolic reactions, proteins carry out a broad range of cellular tasks. However, gene expression needs to be carefully regulated because not all genes are active at the same time. For this reason, a regulatory machinery is present in the cell, containing some factor, which attach to particular DNA sequences to either activate or decrease gene transcription. These factors are called as Transcription factors (TFs). The genetic code of life is like a vast library, and within it, TFs act as master regulators, turning genes on or off at the right time. (Figure 1.) TF are specialized proteins that recognize specific short DNA sequences, usually 6 to 20 base pairs long, found in the promoter or enhancer regions of genes. These sites are known as Transcription factor binding sites (TFBS) (Lenhard *et al.,* 2002), which serve as docking stations for TFs, enabling them to regulate gene transcription by either activating or repressing RNA synthesis. Short, conserved sequences in DNA recognized by specific TFs, are called DNA motifs.A motif logo is a graphical representation used to depict TFBS, columns of lettersrepresents a nucleotide's location. The frequency of nucleotides at that place is indicated by the height of the letters (Fig. 2). This binding determines the rate, timing, and location of gene expression, directly influencing cellular behaviour and function. Enhancers and promoters are located upstream of genes and are critical for determining when, where, and how much a gene is expressed (Li *et al.,* 2022). Although TFBSs are situated in non-coding regions of the genome, they exert a significant influence on the expression of downstream genes. The proper functioning of TFs and their binding to TFBSs is crucial for maintaining normal cellular processes. However, aberrant expression or dysfunction of TFs can lead to the onset and progression of various diseases. For instance, abnormalities in TF binding have been linked to cancer, neurodegenerative disorders, and other complex diseases. Understanding where and how TFs bind to DNA is therefore essential not only for unravelling gene regulation mechanisms but also for developing effective strategies for disease

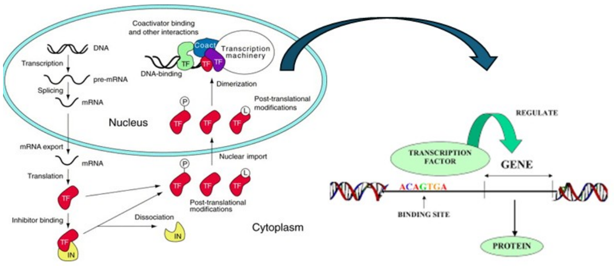


Figure 1. Regulation of TF activity and gene expression (Schwechheimer et al. 1998)

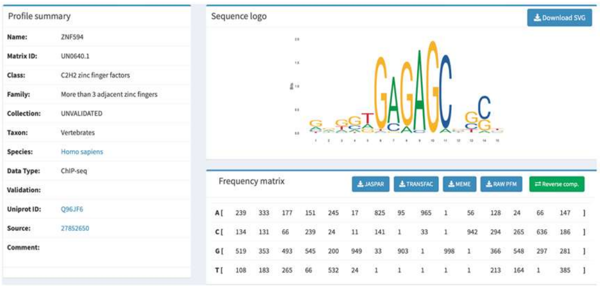
prevention and treatment as studying TFBSs can aid in identifying potential therapeutic targets for gene therapy, drug discovery, and precision medicine. The precise binding of TFs to these sites is fundamental for cellular functions, including development, immune response, and adaptation to environmental changes.

Applications of predicting TFBS: TFBS prediction helps to Understand gene regulation, evolutionary biology, disease mechanisms, and crop improvement. Accurate TFBS prediction in plants and animals has important uses in agriculture, biotechnology, and medicine. Here are a few of the more important uses. TFBS prediction helps uncover regulatory factors in plants that govern how they react to environmental stressors like infections, salt, and drought. Scientists can create genetically engineered crops with increased resistance and yield by comprehending these factors. Deep learning models, for example, have been used to predict TFBSs in plants, which helps identify genes that respond to stress and aids in crop development tactics (Shen *et al.,* 2021) TFBS prediction in animals aids in the comprehension of gene regulation processes and the causes of a number of illnesses (Suryamohan et al., 2015). Finding disease-associated regulatory elements and possible treatment targets requires an understanding of how TFs affect gene expression, which can only be achieved by accurately predicting TFBSs. By facilitating cross-species comparisons, TFBS prediction provides insight into the evolution of gene regulatory networks. Researchers can discover conserved regulatory elements by comparing TFBSs across many taxa, which offers insights into functional genomics and evolutionary links. For example, comparative investigations are made easier by the Animal Transcription Factor Database (AnimalTFDB 4.0), which provides thorough annotations of transcription factors across a variety of animal species (Hu *et al.,* 2019.). Designing artificial promoters and regulatory circuits for plant and animal systems is made possible by an understanding of TFBSs. This information is crucial for creating genetically modified organisms with desirable characteristics, such as improved metabolic pathways or unique biosynthetic capacities. Recent developments have shed new light on the biotechnological uses and mechanisms of transcription factors, emphasizing their critical function in controlling intricate metabolic networks.

Methods for the prediction of TFBS: For the prediction of TFBS, there are mainly two categories of techniques which include experimental techniques and computational approaches. Under the experimental techniques, A number of laboratory-based methods have been created to find TF binding sites both in vitro and in vivo. These techniques offer useful training data for predictions made by computers. Chromatin Immunoprecipitation followed by Sequencing (ChIP-seq), DNAse I Footprinting and Electrophoretic Mobility Shift Assay (EMSA) are the important experimental techniques (Tognon*et al.,* 2023). A potent molecular biology method for examining protein-DNA interactions and locating the binding sites of certain proteins within the genome is ChIP-seq (Furey 2012). In this method, proteins (TF)are linked to DNA and then immunoprecipitation using TF-specific antibodies is performed. The bound DNA fragments then sequenced to identify TFBSs. While ChIP-seq provides high-resolution TFBS data, it requires high-quality TF-specific antibodies and is limited by cell-type specificity. A sizable corpus of information about TFBSs can be produced using ChIP-seq data. In DNAse I Foot printing, DNA is treated with DNAse I, and protected regions where TFs bind remain intact. The pattern is analyzed to identify binding sites. DNase-seq detects regions of open chromatin, which are more accessible to TF binding. This technique provides indirect evidence of potential TFBSs by identifying DNA regions with regulatory potential. However, it does not pinpoint the exact TF binding sites. When DNA fragments and TFs are incubated, EMSA is applied to observe their interaction through gel electrophoresis. While useful for studying binding affinity, they are labor-intensive and do not provide genome-wide insights. Computational methods play a crucial role in predicting TFBSs at a genome-wide scale. The most common computational methods for the prediction of TFBSs include Position Weight Matrices (PWMs) (Alamanova*et al.,* 2010), sequence alignment based methods and Artificial Intelligence (AI)-based methods. PWMs are statistical models that use frequency information from known binding sites to predict potential TFBSs in DNA sequences. These probabilistic models describe the likelihood of nucleotide occurrence at each position in a TFBS. Tools such as JASPAR **(**Rauluseviciute*et al.,* 2024**),**MEME Suite and TRANSFAC store known TF motifs and use them for genome-wide scans (Jayaram et al., 2026). Sequence Alignment-Based MethodsAlign query sequences to known TF binding motifs to identify highly conserved regulatory elements**.** Tools like FIMO (Find Individual Motif Occurrences), PhyloP and PhastCons (for identifying conserved regions) are used for the same (Phan *et al.,* 2024).

The creation of AI-based prediction algorithms targeted at locating TFBSs and detecting TF motifs is made possible by the large dataset generated by experimental techniques discussed above. These algorithms include Machine Learning (ML) and Deep Learning (DL) models having significantly improved TFBS prediction by learning complex patterns from large-scale datasets. Support Vector Machines and Random Forests are two most frequently used ML algorithms. With the recent advancements in deep learning, particularly convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, many approaches have been widely used to model TFBSs by utilizing the wealth of biological sequencing data that is currently available. One of the first techniques to use CNNs for DNA sequence modeling was DeepBind(Alipanahi*et al.,* 2015). DeepTF (Bao *et al.,* 2019) and DeepSNR (Salekin et al.,2018) are another deep learning model to predict the TFBSs. Currently transformer based DL models are getting popularity because of enabling high-resolution identification of TFBSs for example, BertSNR (Bidirectional encoder representation from transformers for Single-Nucleotide Resolution) can identify TFBSs with single-nucleotide resolution (Luo *et al.,* 2024).

TFBS prediction: Here an experiment of TFBS prediction is performed using computational tool.Very firstly a nucleotide sequence encoding for a transcription factor have been selected as per the area of interest.Here, ZNF594 **(**Zinc Finger Protein 594**)** have been taken which is a protein-coding gene in humans and predicted to function as a DNA-binding transcription factor. Then the dataset having sequence id SRR27464678 is searched and downloaded from the NCBI database. Finally, for performing the task of prediction tools like JASPAR and BertSNR is used. BertSNR is a deep learning model which is built upon DNABERT, a pre-trained DNA-specific language model.



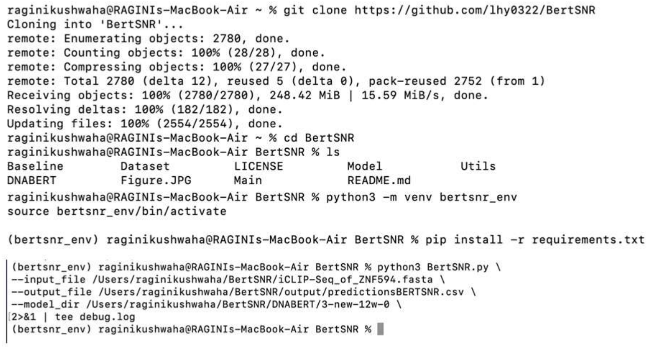
Profile summary

Name:	ZNF594
Matrix ID:	UN0640.1
Class:	C2H2 zinc finger factors
Family:	More than 3 adjacent zinc fingers
Collection:	UNVALIDATED
Taxon:	Vertebrates
Species:	Homo sapiens
Data Type:	ChIP-seq
Validation:	
Uniprot ID:	Q96JF6
Source:	27852650
Comment:	

Frequency matrix

```
A[  239  333  177  151  245  17  825  95  965  1    56   128  24   66   147  ]
C[  134  131  66   239  24   11  141  1   33   1    942  294  265  636  186  ]
G[  519  353  493  545  200  949 33   903 1    998  1    366  548  297  281  ]
T[  108  183  265  66   532  24  1    1   1    1    1    213  164  1    385  ]
```

Fig 3. Here is the result of prediction using JASPAR. Sequence logo is representing a graphical view that is frequently used to illustrate the motif of TFBSs

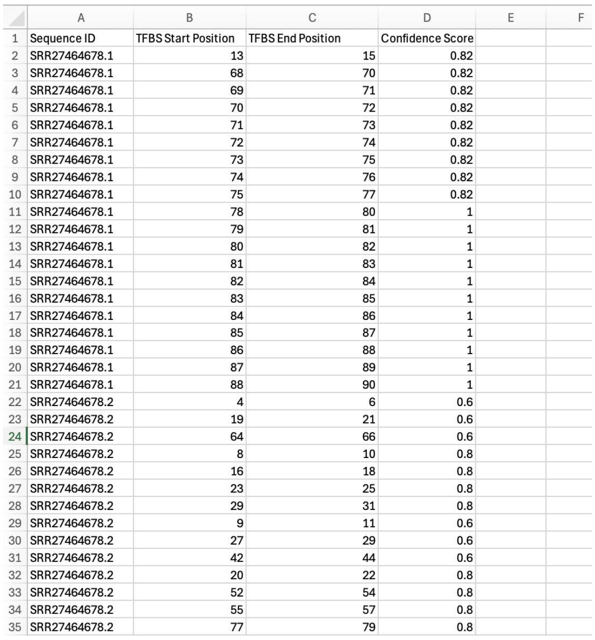


```
raginikushwaha@RAGINIs-MacBook-Air ~ % git clone https://github.com/lhy0322/BertSNR
Cloning into 'BertSNR'...
remote: Enumerating objects: 2780, done.
remote: Counting objects: 100% (28/28), done.
remote: Compressing objects: 100% (27/27), done.
remote: Total 2780 (delta 12), reused 5 (delta 0), pack-reused 2752 (from 1)
Receiving objects: 100% (2780/2780), 248.42 MiB | 15.59 MiB/s, done.
Resolving deltas: 100% (182/182), done.
Updating files: 100% (2554/2554), done.
raginikushwaha@RAGINIs-MacBook-Air ~ % cd BertSNR
raginikushwaha@RAGINIs-MacBook-Air BertSNR % ls
Baseline        Dataset         LICENSE         Model           Utils
DNABERT         Figure.JPG      Main            README.md
raginikushwaha@RAGINIs-MacBook-Air BertSNR % python3 -m venv bertsnr_env
source bertsnr_env/bin/activate

(bertsnr_env) raginikushwaha@RAGINIs-MacBook-Air BertSNR % pip install -r requirements.txt

(bertsnr_env) raginikushwaha@RAGINIs-MacBook-Air BertSNR % python3 BertSNR.py \
--input_file /Users/raginikushwaha/BertSNR/iCLIP-Seq_of_ZNF594.fasta \
--output_file /Users/raginikushwaha/BertSNR/output/predictionsBERTSNR.csv \
--model_dir /Users/raginikushwaha/BertSNR/DNABERT/3-new-12w-0 \
2>&1 | tee debug.log
(bertsnr_env) raginikushwaha@RAGINIs-MacBook-Air BertSNR % █
```

Here the commands have been shown to predict TFBSs using BertSNR.



	A	B	C	D	E	F
1	Sequence ID	TFBS Start Position	TFBS End Position	Confidence Score		
2	SRR27464678.1	13	15	0.82		
3	SRR27464678.1	68	70	0.82		
4	SRR27464678.1	69	71	0.82		
5	SRR27464678.1	70	72	0.82		
6	SRR27464678.1	71	73	0.82		
7	SRR27464678.1	72	74	0.82		
8	SRR27464678.1	73	75	0.82		
9	SRR27464678.1	74	76	0.82		
10	SRR27464678.1	75	77	0.82		
11	SRR27464678.1	78	80	1		
12	SRR27464678.1	79	81	1		
13	SRR27464678.1	80	82	1		
14	SRR27464678.1	81	83	1		
15	SRR27464678.1	82	84	1		
16	SRR27464678.1	83	85	1		
17	SRR27464678.1	84	86	1		
18	SRR27464678.1	85	87	1		
19	SRR27464678.1	86	88	1		
20	SRR27464678.1	87	89	1		
21	SRR27464678.1	88	90	1		
22	SRR27464678.2	4	6	0.6		
23	SRR27464678.2	19	21	0.6		
24	SRR27464678.2	64	66	0.6		
25	SRR27464678.2	8	10	0.8		
26	SRR27464678.2	16	18	0.8		
27	SRR27464678.2	23	25	0.8		
28	SRR27464678.2	29	31	0.8		
29	SRR27464678.2	9	11	0.6		
30	SRR27464678.2	27	29	0.6		
31	SRR27464678.2	42	44	0.6		
32	SRR27464678.2	20	22	0.8		
33	SRR27464678.2	52	54	0.8		
34	SRR27464678.2	55	57	0.8		
35	SRR27464678.2	77	79	0.8		

Here representing the results of prediction: showing only 35 rows out of thousands of rows.

CONCLUSION

TFBSs are essential elements in the regulation of gene expression, playing a central role in stress response, development and disease pathways. Their precise forecast is extremely valuable in a variety of industries, including medicine and agriculture. The development of computational biology and high-throughput sequencing has improved the accuracy and accessibility of TFBS prediction. Basic datasets are provided by experimental methods like ChIP-seq, while genome-wide and high-resolution prediction is made possible by computational tools and deep learning models. Gene therapy, drug development, and synthetic biology are made easier by the combination of biological data and AI-driven tools. Such in silico analyses provide valuable insights that can propel future biological research and innovation.

REFERENCES

Alamanova, D., Stegmaier, P., & Kel, A. (2010). Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies. *BMC bioinformatics*, *11*, 1-15.

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology*, *33*(8), 831-838.

Bao, X. R., Zhu, Y. H., & Yu, D. J. (2019). DeepTF: Accurate prediction of transcription factor binding sites by combining multi-scale convolution and long short-term memory neural network. In *Intelligence Science and Big Data Engineering. Big Data and Machine Learning: 9th International Conference, IScIDE 2019, Nanjing, China, October 17–20, 2019, Proceedings, Part II 9* (pp. 126-138). Springer International Publishing.

Furey, T. S. (2012). ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics*, *13*(12), 840-852.

Hu, H., Miao, Y. R., Jia, L. H., Yu, Q. Y., Zhang, Q., & Guo, A. Y. (2019). AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic acids research*, *47*(D1), D33-D38.

Jayaram, N., Usvyat, D., & R. Martin, A. C. (2016). Evaluating tools for transcription factor binding site prediction. *BMC bioinformatics*, *17*, 1-12.

Lenhard, B., & Wasserman, W. W. (2002). TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics*, *18*(8), 1135-1136.

Li, Q., Zhang, L., Xu, L., Zou, Q., Wu, J., & Li, Q. (2022). Identification and classification of promoters using the attention mechanism based on long short-term memory. *Frontiers of Computer Science*, *16*(4), 164348.

Luo, H., Tang, L., Zeng, M., Yin, R., Ding, P., Luo, L., & Li, M. (2024). BertSNR: an interpretable deep learning framework for single-nucleotide resolution identification of transcription factor binding sites based on DNA language model. *Bioinformatics*, *40*(8), btae461.

Phan, M. H., Zehnder, T., Puntieri, F., Lo, B. W., Lenhard, B., Mueller, F., ... & Ibrahim, D. M. (2024). Conservation of regulatory elements with highly diverged sequences across large evolutionary distances. *bioRxiv*, 2024-05.

Rauluseviciute, I., Riudavets-Puig, R., Blanc-Mathieu, R., Castro-Mondragon, J. A., Ferenc, K., Kumar, V., ... & Mathelier, A. (2024). JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic acids research*, *52*(D1), D174-D182.

Salekin, S., Zhang, J. M., & Huang, Y. (2018). Base-pair resolution detection of transcription factor binding site by deep deconvolutional network. *Bioinformatics*, *34*(20), 3446-3453.

Schwechheimer, C., & Bevan, M. (1998). The regulation of transcription factor activity in plants. *Trends in Plant Science*, *3*(10), 378-383.

Shen, W., Pan, J., Wang, G., & Li, X. (2021). Deep learning-based prediction of TFBSs in plants. *Trends in Plant Science*, *26*(12), 1301-1302.

Suryamohan, K., & Halfon, M. S. (2015). Identifying transcriptional cis-regulatory modules in animal genomes. *Wiley Interdisciplinary Reviews: Developmental Biology*, *4*(2), 59-84.

Tognon, M., Giugno, R., & Pinello, L. (2023). A survey on algorithms to characterize transcription factor binding sites. *Briefings in Bioinformatics*, *24*(3), bbad156.

****