**Partitioning-based DNA Motif Prediction using CNN**

1. **Introduction**

The structural trend that is matched by the deoxyribonucleic acid DNA nucleotide sections that are bound to a certain protein is referred to as a DNA motif. In the analysis of the control of gene activation, the identification of motifs within a given DNA genome dataset performs an essential contribution [1]. It is widespread knowledge that transcription factors, (TFs), play a significant role in the genesis of disease by controlling the activity of genes in a given environment [2]. TFs can produce distinct gene outputs by connecting to particular sequences of DNA or ribonucleic acid (RNA), which are referred to as transcription factor binding sites (TFBSs). In an account of this, the discovery of TFBSs and the empirical verification of the actions of the respective TFs are both extremely beneficial to studying human wellness [3].

The study of motifs is vital to the development of new medical treatments and the comprehension of the processes that occur within cells [4]. Motif investigations using chromatin immunoprecipitation (ChIP) paired with large concurrent DNA sequencing accompanied by statistical forecast have enabled quick genome-wide place estimation of numerous high potential motif locations [5]. The creation of analytical algorithms has been presented with several obstacles as a function of the rising complexity of the genetic lookup space [6].

Deep learning (DL) is now the most effective approach for machine learning (ML), and it has been effectively deployed in a variety of industries, including bioinformatics [7]. It is currently utilized as a classification system for DNA sequences. It has been determined that the amount of information utilized to develop the DL model for estimating the motif length was adequate [8].

Databases such as ENCODE maintain a vast number of freely available DNA pattern datasets containing specified motifs. These datasets contain specific patterns [9]. This information can be utilized to generate an adequate number of training examples for motif length prediction. Utilizing techniques such as pairwise and numerous sequence alignment, it is possible to locate functional components that are conserved through the process of conducting conservation research between the sequences of orthologous or paralogous species [10]. Figure 1 demonstrated a DNA motif discovery using a deep learning model and janggu data set.

Diagram

Description automatically generated

Figure 1 A DNA motif discovery using deep learning [11]

1. **Evaluation of DNA motif**

An arrangement of nucleic acid sequences that has some sort of biological relevance is referred to as a DNA motif. DNA attachment spots for a governing protein, also known as a TF, are examples of DNA motifs [12]. Motifs can appear on either strand of DNA. Furthermore, TF bind to the double-stranded DNA in a direct manner [13]. Discovering patterns or motifs in the genomic DNA patterns of organisms is one of the most significant and difficult challenges in the fields of bioinformatics and data science. In a wide variety of different biological applications, the discovery of DNA motifs acts as a crucial step in the process [14]. In Figures 2a, 2b, and 2c the structures of d(TC5) and its intermolecular I-motifs are depicted.

Diagram

Description automatically generated

Figure 2 Structure of the d(TC5) intermolecular i-motif b) Top view of d(TC5) c) A Hemi

protonated cytosine – cytosine+ base pair [15]

Structural DNA Nanotechnology builds patterns and provisions of objectives with the help of unusual DNA motifs [16]. These peculiar motifs are produced as an outcome of the mutual interchange of DNA backbones, which leads to branched systems comprising a vast number of strands and several different helical domains. It is possible to merge the motifs by a process known as sticky ending cohesion, which involves hydrogen coupling or covalent bonding [17].

1. **DNA motif finding algorithm**

The gene is specified as a piece of base patterns that serves as a blueprint for the procedure of transcription. More motif searching algorithms have been developed and deployed to a wide range of motif structures in the last decade [18]. While several of these motif-finding algorithms are effective in yeast and other lower species, they are much less effective in higher species. Cross-species genomic comparisons and evolutionary footprints are being used to circumvent this problem in new motif discovery methods [19]. Figure 3 demonstrates the General block diagram of the motif discovery technique.

**Clean**

**Motif**

**Representation**

**Objective**

**function**

**Search**

**Strategy**

**Evaluate**

**Extraction**

**Clustering**

**Pre-processing**

**Discovering**

**Post-processing**

**Assemble**

Figure 3 General block diagram of motif discovery technique [20]

In recent years, methods have been developed that incorporate DNA sequence information from coregulated genes with phylogenetic footprinting. These algorithms have markedly increased motif discovery from genomic sequences [21]. When applied with efficient data structures such as suffix trees, word-based approaches, make them an excellent alternative for locating entirely confined motifs [22]. Positioning weight matrices are frequently represented graphically as a pictogram, with each position being denoted by a vertical column of letters whose elevation is inversely proportionate to the amount of information associated with that position [23].

1. **Reviewed of literature**

The research described a DNA motif discovery using deep learning. Several researchers provided the following descriptions of their findings:

**Zhang et al., (2022) [24]** demonstrated that the Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) can detect and break DNA patterns in open chromatin areas by using DNase-I and Tn5. When compared with ChIP-seq and DNase-I sequencing, ATAC-seq can collect a greater variety of TF binding locations. Convolutional neural networks (CNNs) have been successful in recent years in the field of bioinformatics, specifically in the areas of theme identification and gene control systems. ATAC-seq data can be used to predict TFBSs using the GraphPred model established by this study. There are two CNN layers in the GraphPred model: the first layer learns the anchoring of k-Mer nodes and the second layer learns how pattern nodes are embedded. ATAC-seq data is fed into GraphPred, which generates plenty of motifs. For each given sequence, GraphPred calculates how likely it is that one or more k-Mer nodes can be used to find multiple TFBSs, and it also shows how closely these motifs match up with those that have been experimentally validated. The hybrid graph is made up of three diverse kinds of edges and two distinct kinds of nodes. In the end, it concluded that GraphPred had superior efficiency than a few other approaches that were considered to be state of the art. Finding motifs in ATAC-seq data is significantly aided by this study's extensive contributions.

**Yu et al., (2022) [25]** revealed that DNA sequence regions that attach to a certain protein all share a common pattern motif, and this motif is known as a DNA motif. In the study of the modulation of gene expression, the identification of motifs within a given DNA fragment dataset plays an essential role. The size of the DNA motif, which is a crucial aspect of the motif itself, has a direct impact on the quality of the motifs that are detected. During motif identification, it introduced a new motif size estimation method called the Motif Len model by employing supervised ML. The goal of this scheme was to calculate the motif size with a high degree of precision. Motif Len is a model that can estimate the size of motifs, and its general architecture is comprised of three sections: the development of observations, the creation of a forecasting model, and the deployment of the projection model. In most cases, the relative entropy of the columns that correspond to the motif and those that do not correspond to the motif differ significantly from one another. Furthermore, the motif's components in the comparative entropy vector are limited and periodic. Deep learning models such as CNN excel in localized characteristic retrieval and can recover abstract features from the original input. According to this study's findings, the Motif Len model's predictive accuracy on the validation set is more than 90%, whereas the comparison techniques are much less accurate. Additionally, certain existing motif discovery methods can benefit from Motif Len’s to increase the time efficiency of some present motif finding techniques, as well.

**Hejaji et al., (2022) [26]** mentioned that the effective cis-acting components construct the genome efficiency platform. This efficiency basis is raised from the interplay of cis and trans-acting units. Because of the identification of interaction sequences in DNA, RNA, and protein, several essential locations, such as active viewing domains, activator regions of genes, and the position of RNA breakdown signals, have been pinpointed. The TF, which are the most significant controllers of gene production and function as trans components by attaching to regulators and other regions of DNA, is responsible for tuning the transcription pace of the genes. Stimulants and noise barriers of transcription both serve as binding sites for the DNA-binding domains of TFs. These components are made up of similar motifs grouped in clusters, and they can be found in the DNA pattern by looking for other clusters that are similar to them. A new technique for detecting groups or clusters with changing attributes can be a tremendous benefit. New algorithms for looking for clusters of certain functional blocks, with user-defined constraints and a web tool called FMS Cluster Finder were reported in this investigation. In the conclusion, it was determined that the FMS Cluster Finder online application, which represented the specified algorithm, is an application that is easy to use and has a fast execution speed.

**Amit Mishra (2022) [27]** demonstrated that ML algorithms are becoming better, but that does not mean they are problem-free. One of their most significant challenges is that are often sensitive to the information that is provided to them and how it is altered. ML algorithms continue to be useful in assisting businesses in making sense of vast volumes of data. Bayesian optimization algorithm (BOA) is a method that can be used to assist in the development of ML algorithms that have an exceptionally high rate of precision. An example of a black box optimization technique is the BOA. The primary benefit of utilizing these techniques is that they make it possible for a computer to autonomously learn and adjust to distinct types of input without having to be specifically programmed to do so. A long-standing challenge in the field of computational biology is the identification of the most appropriate motifs for a given group of genes. The difficulty of discovering motifs is typically tackled by employing a greedy algorithm, which locates the most effective motifs for a given group of genes. Finding patterns using a structured BOA. In this method, a group of genes is initially separated into two groups: one of which has a required motif, and the other of which does not. After that, a support vector machine (SVM) is employed to divide the genes into the group of the aforementioned two categories. In the end, it concluded that this method is efficient because it provides a path to create a classifier more reliable and accurate.

**Vahed et al., (2022) [28]** emphasized the fact that TF is a necessary regulatory pattern that regulates gene expression. Understanding the roles that TFBS plays in regulating gene expression is essential. Within the past three decades, numerous algorithms have been created to solve the problem of finding TFBS, which has been an ongoing difficulty in the field of quantitative biology. The interacted motifs for a given block of TFBS as well as the motifs for two blocks of TFBS that are separated by varying gaps need to be identified. Bipartite motifs come in a wide variety of forms and can be found in both prokaryotes and eukaryotes. The analysis of gene regulation relies heavily on the identification and characterization of different motifs. The inability to use themes effectively due to a dearth of web servers that are user-friendly and integrated to make motif configuration easier for users while still getting the best results, this analysis describes bipartite motifs learning (BML), a parameter-free web server that uses high-throughput sequence alignment info as input to deliver a user-friendly gateway for online finding and analysis of sequence motifs. BML uses positioning and dinucleotide weight matrix, which expresses base dependencies. EM solves this optimization challenge. In the end, it was determined that even if naive users are unsure about the numbers of the input variables, they can still use BML as a PF web server and retrieve consistent patterns.

**Schafer and Lesser (2022) [29]** investigated that the time series (TS) is a series of real values arranged along a given aspect, with time being the most essential aspect in the investigation. The process of looking for patterns like this in a set of input data is referred to as motif discovery (MD). These motifs frequently replicate underlying patterns in the operation of creating the TS, like heart rate in an electrocardiogram (ECG) recording, and k-Complexes in electroencephalogram (EEG) sleeping information. Other examples include a recording of a sleep electroencephalogram (EEG). MD is also essential in the TS process since it serves as a pre-processing phase before categorization, grouping, anomaly identification, and rule creation. There are many distinct methods available since there is no universally accepted description of what a motif is. Instead of asking the operator to specify the threshold r, k-Motiflets use the length k that the person wants for the motif group as the input. This leads to a significant reduction in the amount of time and effort required for exploratory investigation. An acceptable pruning strategy is utilized by the technique to cut down the number of training data by utilizing an upper limit on the finest extent d. The exact algorithm's algorithmic notion is based on an evaluation of all groups of substrings of T of size k paired with aggressive pruning. In the end, it concluded that the estimated approach creates superior motifs than all of its other counterparts at shorter runtimes.

**Zhang et al., (2022) [30]** demonstrated that DNA has been put forward as a candidate for use as an effective substrate for the storage of intellectual data. The fast expansion of human society has led to an explosion in the overall amount of data, which continues to grow at an exponential rate. The conversion of digital data into DNA molecules serves as the foundation of DNA-based storage's fundamental operating principle. Therefore, the coding method is one of the most fundamental phases in DNA-based storage, but it is also one of the most important steps. In DNA-based storage, a few sophisticated biochemical procedures were implemented to add additional capabilities such as photo preview. DNA operability and usefulness are provided via DNA strand displacement, hybridization, and transcription. DNA biochemical processes are not as robust as head-on-disk operations. The limitations of currently available approaches are mostly to blame for the occurrence of errors in DNA-based storage. Because of more local biochemical restrictions and their mixtures, the currently available coding techniques cannot be applied or are unreliable. In this analysis, constructed a graph-based architecture and give it the name SPIDER-WEB. Its purpose is to generate graph-based algorithms conforming to arbitrary local biochemical restrictions. In the end, it concluded that SPIDER-WEB could be utilized directly by configuring the program with the characteristics of the local biochemical restrictions.

**Udayakumar and Pushplatha (2022) [31]** indicated that the word hereditary designing is a conventional term that is used to encompass a broad variety of distinct types of hereditary component constraints, which demonstrates that the term is extremely broad and customary. Several distinct kinds of computations are capable of being carried out with the assistance of these controllers. For disentanglement, the term nucleotide is frequently substituted for DNA. It is a fundamental component of all living cells and possesses an intriguing structure that allows it to fulfill its two primary functions: coding for the production of proteins and self-replication, which ensures that genetic information is accurately transmitted from one generation to the next. It is a significant challenge in the field of computational biology to locate sequences of genetic codes that are similar to one another. The problem presented in this Resemblance provided the efficient computation to identify arrangement similarity among hereditary genetic codes by making use of the most basic aftereffect concern. The calculation enhances the presentation while maintaining the benefits that were gained with the modified distance calculation. The grouping correlations of DNA genetic codes can be used to locate matches among hereditary codes, beginning with one DNA arrangement and moving on to the adjacent succession. To fine-tune the Time and Space complexity, a dynamic programming issue was used in conjunction with the Longest Common Subsequence (LCS) Algorithm. In the end, it was determined that the LCS computation has advantages over the one that is currently being used by reducing the amount of time that is required for the computation and, as a response, enhancing the performance of a variety of lengths.

1. **Comparison of the reviewed literature**

Table 1 shows the comparison of reviewed literature of different authors.

Table 1 Comparison of the reviewed literature

|  |  |  |
| --- | --- | --- |
| **Authors Name** | **Technique used** | **Outcomes** |
| **Zhang et al., (2022) [24]** | GraphPred method | GraphPred had superior efficiency than a few other approaches that were considered to be state of the art. |
| **Yu et al., (2022) [25]** | MotifLen model | MotifLen optimizes motifs found by existing motif discovery techniques and improves their speed. |
| **Hejaji et al., (2022) [26]** | FMS Cluster finder | The FMS Cluster Finder web application, which implements the specified algorithm, is a user-friendly tool that runs quickly. |
| **Amit Mishra (2022) [27]** | Bayesian Optimization | This method is useful because it allows the classifier to become more robust and accurate. |
| **Vahed et al., (2022) [28]** | BML | When naive users are unsure of the input settings, it can utilize BML as a PF web server to get consistent patterns. |
| **Schafer and Lesser (2022) [29]** | k-motiflet model | At reduced runtimes, the approximation approach yields better motifs than all of its competitors. |
| **Zhang et al., (2022) [30]** | SPIDER-WEB | SPIDER-WEB can be utilized directly by configuring the program's properties of local biochemical restrictions. |
| **Udayakumar and Pushplatha (2022) [31]** | LCS | The LCS calculation reduces time complexity and improves length performance. |

1. **Background study**

The study of motifs is vital to the development of new medical treatments and the comprehension of the processes that occur within cells. Individual genes, huge genetic areas, complete chromosomes, and entire genomes can all be analyzed using DNA sequencing. It came up with a better method for predicting DNA motifs. To begin with, just a subset of the input patterns are used in the three-stage technique for motif recognition, and the resulting motifs are then used to find potential sites in the input sequences that do not overlap. The current approach is not reliable because the motifs found in the initial subset are shown as position weight matrices. This means that there are a lot of false positives. method, which is known as called Deep Finder, uses deep learning neural networks with binding site-related features to build a motif model. Deep Finder uses deep learning neural networks with binding site-related features to build a motif model. Furthermore, for a greater number of accurate predictions, the first motif prediction procedure makes use of numerous prediction techniques. Binding sites were used to create these characteristics, which were then supplemented with specificity data from known TF protein recognition sites [32].

1. **Research objectives**

This heading contains quantifiable and feasible objectives that would be accomplished during research.

## Following are the Research objectives:

* To examine the binding locations of transcription factors in a cell-specific manner.
* To find cis-regulatory motifs by finding TF binding sites and figuring out how any organism's genes are controlled.
* To locate their potential binding sites, which should stand out as being overrepresented in the sequences.
* To identify the duration of time motifs that are required by a process that only has room for one parameter by decreasing time series with the assistance of matrix profiles.

1. **Problem formulation**

The potential to discover TF binding regions or motifs in the genome is one of the solutions to unraveling the riddle of how genes are controlled. Motifs are sequence patterns that are repeated several times in a genome. These patterns serve as binding sites for TF, which are essential for the control of protein synthesis in cells. DNA binding factors are defined in great portions by binding site motifs, which are distinctive signatures that range in size from 5 to 15 DNA base pairs (bp) and bind to TFs. Because of this, the classification of such DNA motifs has developed into an essential issue within the biological community. A technique for identifying motifs from experimental data on TF-gene interactions that is reliant on ML and has the potential to both complement and enhance existing approaches. This method, which goes by the name Deep Finder, involves training deep neural networks to recognize elements that are connected to binding sites to create a motif model

1. **Scope of work**

Motifs, small recurring patterns in biological patterns, are critical to understanding how living organisms regulate their genetics. However, the topic is NP-hard and offers a barrier in computational biology because of weak motifs.

The next logical measure after discovering an interesting group of motifs is to determine the biological function of these sequence properties. In particular experimental settings, it can be feasible to link motifs to particular observable effects like an increase or decrease in gene expression. Associating individual TF with the patterns to which they interact can provide further biological information to regulatory networks. These interpretation concerns are not immediately addressed by standard motif discovery technologies. In recent years, however, methods for investigating these issues have been established.

1. **Research methodology**

This section represents the proposed methodology. The current research works on the discovery of DNA motifs by using CNN and clustering techniques.

1. **Research technique**

Two technologies are used in this proposed methodology which is Coevolutionary Neural Networks (CNN) and the SVM Light feature extraction.

* **Convolutional neural network (CNN)**

CNNs are neural networks that consist of one or more convolutional layers and are primarily used for image processing, separation, classifications, and other auto-correlated data analysis. CNN's are a kind of deep neural network that is most frequently used to evaluate visual information. Convolution is the term used to describe the process of moving a filter over an input signal. It can do classification and regression on large amounts of high-dimensional raw data without the need for feature engineering before understanding the capabilities of each feature [33]. CNNs are meant to learn spatial hierarchies of features automatically and adaptively by backpropagation, and they do so by using a variety of building pieces, including fully connected layers, pooling layers, and convolution layers. The structure of the CNN is shown in Figure 2.

A picture containing diagram

Description automatically generated

Figure 4: Schematic structure of CNN [34]

As shown in Figure 4, the fundamental components of CNN design may be separated into five categories. The following sections provide a full explanation of each component.

**Input layer:** It is possible to send the raw set of data for the input nodes straight into the input layer itself. When entering a single photo into the input layer, the pixel value of that picture is utilized.

**Convolutional layer**: Each convolutional layer has a convolutional kernel, which pulls different features from the data based on the characteristics extracted from the input data. Various convolutional kernels are used to extract different attributes from the same input data set, and this is shown in figure 4.

**Fully connected layer:** The characteristics of the preceding convolutions are integrated and normalized to provide a probability for different scenarios in this layer.

**Output layer:** It is governed by the individual conditions as to how many neurons should be present in this layer.

* **SVM Light feature extraction**

Machine learning is a feasible way of finding binding genes and binding locations of TFs using SVM Light feature extraction algorithms. Discrimination and regression issues are often solved using this technique in machine learning.  Its ability is still quite promising even though they are not among the most efficient. The SVM can extract patterns from data sets that contain promoters of genes that are known to bind a certain TF. The memory needs of the technique are scalable, and it can deal effectively with issues involving tens of thousands of support vectors [35]. Figure 5 illustrated deep feature extraction and SVM classification

Diagram

Description automatically generated

Figure 5: An illustration for deep feature extraction and SVM classification[36]

1. **Proposed Methodology**

This section contains the explanation of the proposed methodology for designing of neural network for the discovery of motif sequences. In the context of the suggested methodology designed architecture in Figure 6. It is presented partitioning-based DNA motif prediction using CNN and clustering techniques.

Dataset

Data Partitioning

SVM Light feature extraction

Merge the features and divide using clustering

System learning using CNN

Feature representation

Dense representation

Prediction

MEME

Bio Prospector

Motif Sampler

MD Scan

Extract the feature from top 3 Motifs

Figure 6 Structural design of Proposed Methodology

The proposed methodology shown in figure 6 is explained below step by step:

**Step 1:** Firstly, it began by taking the data set and separating it into four different parts i.e., MEME, Bioprospector, MDscan, and Motifsampler. The characteristics of the techniques would be retrieved, and the top three characteristics of the motifs would be chosen.

**Step 2:** Following the conclusion of step 1, the SVM light retrieved the dataset consisting of promoters of genes that are known to bind a certain DNA TF with a range of errors that can be tolerated in the input data.

**Step 3:** During this stage of the process, the topmost three motifs that were generated by each tool are combined and partitioned with the help of a clustering algorithm.

**Step 4:** Through the use of a CNN, this step of the system training has been accomplished. Convolutional layers were used for data removal and presentation in the input layer, which receives training data with labels and sample features.

**Step 5:** Oncogenic virus integration sites (VISs) in human genomes can be effectively predicted using a CNN model that learns informatic traits and important genomic locations from raw DNA genomes.

**References**

Suter, David M. "Transcription factors and DNA play hide and seek." Trends in cell biology 30, no. 6 (2020): 491-500.

Lin Quy Xiao X, Thieffry D, Jha S, et al. TF regulome R reveals transcription factors’ context-specific features and functions. Nucleic Acids Res 2019;48: e10–0.

Zhang, Shuangquan, Anjun Ma, Jing Zhao, Dong Xu, Qin Ma, and Yan Wang. "Assessing deep learning methods in cis-regulatory motif finding based on genomic sequencing data." Briefings in bioinformatics 23, no. 1 (2022): bbab374.

Maston, Glenn A., Sara K. Evans, and Michael R. Green. "Transcriptional regulatory elements in the human genome." Annu. Rev. Genomics Hum. Genet. 7 (2006): 29-59.

Zambelli, Federico, Graziano Pesole, and Giulio Pavesi. "Motif discovery and transcription factor binding sites before and after the next-generation sequencing era." Briefings in bioinformatics 14, no. 2 (2013): 225-237.

Madrid, Frank, Shima Imani, Ryan Mercer, Zachary Zimmerman, Nader Shakibay, and Eamonn Keogh. "Matrix profile xx: Finding and visualizing time series motifs of all lengths using the matrix profile." In 2019 IEEE International Conference on Big Knowledge (ICBK), pp. 175-182. IEEE, 2019.

Xu, Haodong, Peilin Jia, and Zhongming Zhao. "DeepVISP: deep learning for virus site integration prediction and motif discovery." Advanced Science 8, no. 9 (2021): 2004958.

Jin, Shuting, Xiangxiang Zeng, Feng Xia, Wei Huang, and Xiangrong Liu. "Application of deep learning methods in biological networks." Briefings in bioinformatics 22, no. 2 (2021): 1902-1917.

Wang, Jie, Jiali Zhuang, Sowmya Iyer, XinYing Lin, Troy W. Whitfield, Melissa C. Greven, Brian G. Pierce et al. "Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors." Genome Research 22, no. 9 (2012): 1798-1812.

Poliakov, Alexandre, Justin Foong, Michael Brudno, and Inna Dubchak. "GenomeVISTA—an integrated software package for whole-genome alignment and visualization." Bioinformatics 30, no. 18 (2014): 2654-2655.

Kopp, Wolfgang, Remo Monti, Annalaura Tamburrini, Uwe Ohler, and Altuna Akalin. "Deep learning for genomics using Janggu." Nature communications 11, no. 1 (2020): 1-7.

Rombauts, Stephane, Patrice Déhais, Marc Van Montagu, and Pierre Rouzé. "PlantCARE, a plant cis-acting regulatory element database." Nucleic acids research 27, no. 1 (1999): 295-296.

Wingender, Edgar, Xin Chen, Reinhard Hehl, Holger Karas, Ines Liebich, Volker Matys, T. Meinhardt, M. Prüß, Ingmar Reuter, and Frank Schacherer. "TRANSFAC: an integrated system for gene expression regulation." Nucleic acids research 28, no. 1 (2000): 316-319.

Kurtz, Stefan, Jomuna V. Choudhuri, Enno Ohlebusch, Chris Schleiermacher, Jens Stoye, and Robert Giegerich. "REPuter: the manifold applications of repeat analysis on a genomic scale." Nucleic acids research 29, no. 22 (2001): 4633-4642.

Day, Henry A., Pavlos Pavlou, and Zoë AE Waller. "i-Motif DNA: structure, stability and targeting with ligands." Bioorganic & medicinal chemistry 22, no. 16 (2014): 4407-4418.

Seeman, Nadrian C. "An overview of structural DNA nanotechnology." Molecular Biotechnology 37, no. 3 (2007): 246-257.

Voigt, Niels V., Thomas Tørring, Alexandru Rotaru, Mikkel F. Jacobsen, Jens B. Ravnsbæk, Ramesh Subramani, Wael Mamdouh et al. "Single-molecule chemical reactions on DNA origami." Nature nanotechnology 5, no. 3 (2010): 200-203.

Das, Modan K., and Ho-Kwok Dai. "A survey of DNA motif finding algorithms." BMC bioinformatics 8, no. 7 (2007): 1-13.

Tagle, Danilo A., Ben F. Koop, Morris Goodman, Jerry L. Slightom, David L. Hess, and Richard T. Jones. "Embryonic ε and γ globin genes of a prosimian primate (Galago crassicaudatus): Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints." Journal of molecular biology 203, no. 2 (1988): 439-455.

Hashim, Fatma A., Mai S. Mabrouk, and Walid Al-Atabany. "Review of different sequence motif finding algorithms." Avicenna Journal of medical biotechnology 11, no. 2 (2019): 130.

Siddharthan, Rahul, Eric D. Siggia, and Erik Van Nimwegen. "PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny." PLoS computational biology 1, no. 7 (2005): e67.

Sagot, Marie-France. "Spelling approximate repeated or common motifs using a suffix tree." In Latin American Symposium on Theoretical Informatics, pp. 374-390. Springer, Berlin, Heidelberg, 1998.

Schneider TD, Stephens RM: Sequence logos: a new way to display consensus sequence. Nucleic Acids Res 1990, 18:6097-6100.

Zhang, Shuangquan. "GraphPred: An approach to predict multiple DNA motifs from ATAC-seq data using graph neural network and coexisting probability." bioRxiv (2022): 2022-05.

Yu, Qiang, Xiao Zhang, Yana Hu, Shengpin Chen, and Liying Yang. "A Method for Predicting DNA Motif Length Based on Deep Learning." IEEE/ACM Transactions on Computational Biology and Bioinformatics (2022).

Hejazi, Mohammad Mahdi, Faegheh Golabi, Mohsen Bahrami, Houman Kahroba, and Mohammad Saeid Hejazi. "FMSClusterFinder: A new tool for detection and identification of clusters of sequential motifs with varying characteristics inside genomic sequences." bioRxiv (2022).

Mishra, Amit. "Practical Bayesian Optimization of Machine Learning Algorithms."

Vahed, Mohammad, Majid Vahed, and Lana X. Garmire. "BML: a versatile web server for bipartite motif discovery." Briefings in bioinformatics 23, no. 1 (2022): bbab536.

Schäfer, Patrick, and Ulf Leser. "Motiflets--Fast and Accurate Detection of Motifs in Time Series." arXiv preprint arXiv:2206.03735 (2022).

Zhang, Haoling, Zhaojun Lan, Wenwei Zhang, Xun Xu, Zhi Ping, Yiwei Zhang, and Yue Shen. "SPIDER-WEB enables stable, repairable, and encryptable algorithms under arbitrary local biochemical constraints in DNA-based storage." arXiv preprint arXiv:2204.02855 (2022).

Udayakumar, U., and K. Pushpalatha. "A Review Study on DNA Sequence Similarity Using Longest Common Subsequence Algorithm."

Lee, Nung Kion, Farah Liyana Azizan, Yu Shiong Wong, and Norshafarina Omar. "DeepFinder: An integration of feature-based and deep learning approach for DNA motif discovery." Biotechnology & Biotechnological Equipment 32, no. 3 (2018): 759-768

Ashok, T., & Suresh Varma, P. (2020). Crop prediction is based on environmental factors using machine learning ensemble algorithms. In Intelligent Computing and Innovation on Data Science (pp. 581-594). Springer, Singapore

Gu, H., Wang, Y., Hong, S., & Gui, G. (2019). Blind channel identification aided generalized automatic modulation recognition based on deep learning. IEEE Access, 7, 110722-110729.

Kon, Mark A., Yue Fan, Dustin Holloway, and Charles DeLisi. "Svmotif: A machine learning motif algorithm." In Sixth International Conference on Machine Learning and Applications (ICMLA 2007), pp. 573-580. IEEE, 2007.

Deniz, Erkan, Abdulkadir Şengür, Zehra Kadiroğlu, Yanhui Guo, Varun Bajaj, and Ümit Budak. "Transfer learning based histopathologic image classification for breast cancer detection." Health information science and systems 6, no. 1 (2018): 1-7.