

2017

Development of Computational Techniques for Regulatory DNA Motif Identification Based on Big Biological Data

Jinyu Yang

Follow this and additional works at: <https://openprairie.sdstate.edu/etd>

 Part of the [Statistics and Probability Commons](#)

Recommended Citation

Yang, Jinyu, "Development of Computational Techniques for Regulatory DNA Motif Identification Based on Big Biological Data" (2017). *Theses and Dissertations*. 2155.
<https://openprairie.sdstate.edu/etd/2155>

This Thesis - Open Access is brought to you for free and open access by Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. For more information, please contact michael.biondo@sdstate.edu.

DEVELOPMENT OF COMPUTATIONAL TECHNIQUES FOR REGULATORY DNA
MOTIF IDENTIFICATION BASED ON BIG BIOLOGICAL DATA

BY

JINYU YANG

A thesis submitted in partial fulfillment of the requirements for the

Master of Science

Major in Statistics

South Dakota State University

2017

DEVELOPMENT OF COMPUTATIONAL TECHNIQUES FOR REGULATORY DNA
MOTIF IDENTIFICATION BASED ON BIG BIOLOGICAL DATA

JINYU YANG

This thesis is approved as a creditable and independent investigation by a candidate for the Master of Science in Statistics degree and is acceptable for meeting the thesis requirements for this degree. Acceptance of this thesis does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Qin Ma, Ph.D.

Thesis advisor

Date

Kurt Cogswell, Ph.D.

Head, Math and Statistics

Date

Dean, Graduate School

Date

ACKNOWLEDGMENTS

Firstly, I would like to thank my academic advisor Dr. Qin Ma, who has helped me a lot in the last two years. His brilliant insights in bioinformatics, especially in motif prediction, have a profound influence on my research. I have learned a lot from him, such as how to prepare a paper, how to give a clear and informative presentation in public, and how to solve bioinformatics problems. He also gave very detailed suggestions and ideas for this thesis.

I would like to thank my committee members Dr. Xijin Ge, Dr. Brady Phelps, and Dr. Erliang Zeng for their insightful comments and useful suggestions in terms of my slides and research plan.

I would also like to thank all Bioinformatics and Mathematical Biosciences Lab (BMBL) members Adam McDermaid, Juan Xie, Anjun Ma, Shihan Wu, Yiran Zhang, and Minxuan Sun for their substantial help and discussion.

Finally, I would like to thank my family for their endless support.

CONTENTS

LIST OF FIGURES	vi
ABSTRACT.....	vii
CHAPTER 1. Introduction.....	1
1.1 Regulatory DNA motif.....	1
1.2 Motif representation	1
1.3 Motif signal detection techniques and performance evaluation.....	4
1.4 ChIP-seq	6
1.5 DNA shape and shape motif.....	6
1.6 DL.....	7
1.7 Outline.....	8
CHAPTER 2. Motif identification.....	9
2.1 Motif identification from promoter sequences.....	9
2.2 Motif identification from ChIP-seq data	10
2.3 DL in motif identification	11
CHAPTER 3. Previous work	13
3.1 BOBRO	13
3.2 BoBro 2.0	13
3.3 DMINDA	14
3.4 MP ³	14

CHAPTER 4. DMINDA 2.0.....	16
4.1 Introduction.....	16
4.2 Methods and results.....	17
4.3 Conclusion.....	19
CHAPTER 5. DESSO.....	21
5.1 Introduction.....	21
5.2 Dataset.....	22
5.3 Methods.....	23
5.3.1 DESSO.....	23
5.3.2 GCNN.....	25
5.4 Results.....	26
5.4.1 DNA shape has strong predictive power in TF-DNA binding specificity prediction.....	26
5.4.2 Identification of sequence and shape motifs.....	27
5.4.3 GCNN captures motif dependencies of long DNA sequences.....	29
5.5 Conclusion.....	29
CHAPTER 6. Discussion.....	30

LIST OF FIGURES

Figure 1. Motif instances and logo.	47
Figure 2. Representation of a motif	48
Figure 3. CNN in image classification.....	49
Figure 4. Existing algorithms and tools for motif identification.	50
Figure 5. DL framework for motif identification.....	51
Figure 6. Schematic overview of the BOBRO.....	52
Figure 7. Workflow of DMINDA.....	53
Figure 8. An outline of the MP3 framework.....	54
Figure 9. Comparison of DMINDA 2.0 and six motif analyses webservers	55
Figure 10. Workflow of DMINDA 2.0.....	56
Figure 11. Identified motifs from DMINDA 2.0	57
Figure 12. (A) Result page of motif scanning.....	58
Figure 13. Result page of motif comparison.....	59
Figure 14. Result page of motif co-occurrence analysis.....	60
Figure 15. Result page of motif finding by MP3	61
Figure 16. (A) A Cytoscape-like network visualization of predicted regulons	62
Figure 17. Workflow of DESSO.....	63
Figure 18. Workflow of GCNN	64
Figure 19. Performance of DESSO.....	65
Figure 20. Analysis of identified motifs	66
Figure 21. HelT motif and sequence logo of MAFF	67
Figure 22. (A) Classification accuracy of CNN with different peak lengths.....	68

ABSTRACT

DEVELOPMENT OF COMPUTATIONAL TECHNIQUES FOR REGULATORY DNA
MOTIF IDENTIFICATION BASED ON BIG BIOLOGICAL DATA

JINYU YANG

2017

Accurate regulatory DNA motif (or *motif*) identification plays a fundamental role in the elucidation of transcriptional regulatory mechanisms in a cell and can strongly support the regulatory network construction for both prokaryotic and eukaryotic organisms. Next-generation sequencing techniques generate a huge amount of biological data for motif identification. Specifically, Chromatin Immunoprecipitation followed by high throughput DNA sequencing (ChIP-seq) enables researchers to identify motifs on a genome scale. Recently, technological improvements have allowed for DNA structural information to be obtained in a high-throughput manner, which can provide four DNA shape features. The DNA shape has been found as a complementary factor to genomic sequences in terms of transcription factor (TF)-DNA binding specificity prediction based on traditional machine learning models. Recent studies have demonstrated that deep learning (DL), especially the convolutional neural network (CNN), enables identification of motifs from DNA sequence directly.

Although numerous algorithms and tools have been proposed and developed in this field, (1) the lack of intuitive and integrative web servers impedes the progress of making effective use of emerging algorithms and tools; (2) DNA shape has not been

integrated with DL; and (3) existing DL models still suffer high false positive and false negative issues in motif identification.

This thesis focuses on developing an integrated web server for motif identification based on DNA sequences either from users or built-in databases. This web server allows further motif-related analysis and Cytoscape-like network interpretation and visualization. We then proposed a DL framework for both sequence and shape motif identification from ChIP-seq data using a binomial distribution strategy. This framework can accept as input the different combinations of DNA sequence and DNA shape. Finally, we developed a gated convolutional neural network (GCNN) for capturing motif dependencies among long DNA sequences.

Results show that our developed web server enables providing comprehensive motif analysis functionalities compared with existing web servers. The DL framework can identify motifs using an optimized threshold and disclose the strong predictive power of DNA shape in TF-DNA binding specificity. The identified sequence and shape motifs can contribute to TF-DNA binding mechanism interpretation. Additionally, GCNN can improve TF-DNA binding specificity prediction than CNN on most of the datasets.

CHAPTER 1. Introduction

1.1 Regulatory DNA motif

Motifs are usually conserved short DNA sequences, which tend to be 8-20 base pairs (bp) long [1]. Typically, they are TF binding sites (TFBSs) and play significant roles in regulating transcription rates of nearby genes and further control their expression levels. Hence, *de-novo* motif prediction and related analysis (e.g., motif scan and motif comparison) provide a solid foundation for the inference of gene transcriptional regulatory mechanisms in both prokaryotic and eukaryotic organisms [2, 3]. Moreover, these techniques also substantially contribute to some system-level studies, such as regulon modeling and regulatory network construction [2, 4, 5]. With the rapidly growing availability of sequenced genomes and advanced biotechnologies, substantial computational techniques have been carried out to identify motifs from query DNA sequences. Nevertheless, the variations among motifs and their short length make their discovery a very challenging problem.

1.2 Motif representation

A motif represents a set of DNA segments with the same length, which are binding sites for the same TF. The segments of a motif can be aligned to form motif logo (Figure 1), where each of them is called an instance. Different instances of the same motif tend to be similar to each other on sequence level (Figure 2A) [6]. A representation model of a motif, to demonstrate the similarity of its instances, is expected to accurately capture the characteristics of protein-DNA binding activity of its corresponding TF [7].

The most straightforward model to denote the binding preference of a TF on each position along a motif is the *consensus* sequence (e.g., AGTCA or AGTCG for the motif

in Figure 2A), which is composed of the concatenation of the most frequent nucleotide on each position. It can be seen as the ancestor of the binding sites of the same TF, with an assumption that these sites evolved from it. Although the consensus presents the characteristics of a motif in each position in a simple and clear way, the variations in this motif are absent in this model. The *degenerate consensus* was proposed to fill this gap, using IUPAC (International Union of Pure and Applied Chemistry) wildcards to replace the exact nucleotides (A, G, C, and T). For example, W means both A and T in this position could be recognized by the TF of this motif (Figure 2B) [8].

A more accurate and most commonly used model is the *motif profile*. A profile is built by aligning the available instances of a motif M and counting the frequency of each nucleotide at each position ($f_{i,j}$). These frequencies give rise to a typical matrix representation of a motif profile ($M_f = \{f_{i,j}\}_{4 \times l}$, Figure 2A), called a *position weight matrix* (PWM). An alternative way of constructing the PWM is using the probability distribution to replace frequencies ($M_p = \{p_{i,j}\}_{4 \times l}$). Specifically, these frequencies will be divided by the number of binding sites of this motif, and such a representation of the PWM in Figure 2A is shown in formula (1).

$$M_p = \{p_{i,j}\} = \begin{bmatrix} 0.6 & 0 & 0 & 0.2 & 0.4 \\ 0 & 0.6 & 0 & 0 & 0.4 \\ 0 & 0.4 & 0 & 0.8 & 0.2 \\ 0.4 & 0 & 1 & 0 & 0 \end{bmatrix} \quad (1)$$

Taking the background frequencies of each nucleotide into consideration, the PWM in formula (1) can be further modified as $M_g = \{g_{i,j}\}$, where $g_{i,j} = \log(f_{i,j}/b_i)$ and b_i is the probability of the i th nucleotide of (A, G, C, T) appearing in the background

sequences. Based on this version of PWM, we can calculate the *Information Content* (IC) to evaluate how conserved this motif is, i.e., formula (2).

$$I(M) = \sum_{j=1}^l \sum_{i=1}^4 p_{i,j} \log \frac{p_{i,j}}{b_i} \quad (2)$$

The matrix can also be used to evaluate how a given DNA segment s , with length l , is consisted with motif M by calculating the score:

$$\text{score}(s) = \sum_{j=1}^l \sum_{i=1}^4 p_{i,j} \log \frac{p_{i,j}}{b_i} \times \delta_{i,j} \quad (3)$$

where $\delta_{i,j} = 1$, if the j th nucleotide of s is the i th nucleotide of (A, G, C, T), $\delta_{i,j} = 0$ otherwise. A problem with this model is that the probability $f_{i,j}$ could be zero for a small set of binding sites, giving rise to negative infinity in formulas (2&3). A common method to avoid this bias is adding a certain value (pseudocount) for each position of the motif [9]. Through system simulation and analysis, K. Nishida *et al.* found that the optimal pseudocount value is correlated with the entropy of a motif profile [10]. Specifically, the less conserved motif profiles prefer larger pseudocount value, and 0.8 is suggested in general.

As shown above, multiple approaches for modeling TF-DNA binding specificity have been developed. A systematic comparison of these approaches can provide substantially valuable information for further motif identification algorithm design. DREAM5 consortium organized a competition on motif representation models by applying 26 approaches to *in vitro* protein binding microarray data [11]. These approaches adopt various strategies, including but not limited to k -mers model, PWM, Hidden Markov Model, and dinucleotides. The k -mers and PWM are the two main

strategies, which show similar average performance on multiple data sets. However, they have substantial differences in terms of individual performance on a few data sets, indicating that motif identification is sensitive for model selection. Another interesting observation is that the IC of a motif may not fully represent its accuracy. It is obviously contrary to the basic principle of general motif identification tools, thus deeper investigation into this area is still needed to improve the motif representation models.

These motif representation models are still not perfect with a common disadvantage that they ignore the correlation among different positions in a motif. For example, the motif in Figure 2C has the same PWM as the motif in Figure 2A, but the first two positions in this motif are correlated and dependent on each other. Hence, a high order Markov model is suggested to be integrated into PWM matrix [12]. Meanwhile, it is unsure whether a known PWM can fairly represent the whole population scenario, as the frequencies of nucleotides in each position are calculated only from the known binding sites of a TF. A motif profile built on partially identified binding sites of a TF may induce bias when it is used to interpret the global binding preference, especially when this profile is used to model the orthologous binding sites from various species. A more fundamental debate is: do the nucleotides with lower frequencies imply lower binding ability? At the time of this writing, there are still no clear answers to this question, and deeper thought about above concerns will bring potential ways to improve existing motif representation models.

1.3 Motif signal detection techniques and performance evaluation

The basic computational assumption of motif identification is that they are overrepresented as conserved patterns in given sequences. The scattered instances of a

motif are not perfectly identical but similar to each other. Once identified and well aligned, they will show significance in conservation compared to background sequences. Therefore, identifying and aligning these instances are the primary issues in motif identification.

Motif identification methods mainly fall into two categories: word-based methods (i.e., consensus-based) and profile-based methods [8, 13]. Word-based methods usually enumerate and compare nucleotides starting from a consensus sequence with a fixed length and a tolerance of mutations. Theoretically, this strategy can identify optimal global solutions but suffers from high computational complexity when it is applied to large-scale input data or to-be-identified motifs which are relatively long or with a large number of mutations [13]. The profile-based methods usually start with some aligned patterns, either randomly chosen [14] or enumerated in a limited subset of input data [14], and refined based on some criteria on the whole data. These criteria are designed to evaluate the overrepresented significance of aligned profiles from the input sequences. Improvements are mostly conducted in a heuristic way, e.g. neighboring improvement (add or delete patterns to see if the profile goes better, similar to a hill-climbing method) or iterative statistical methods (Gibbs sampling or Expectation Maximization). The profile-based methods usually run faster than word-based methods and have better performance in predicting motifs with complex mutations. However, such methods tend to fail in detecting of multiple motifs, especially when the data size is large, as the iterative procedure they adopted often falls into local optimizations, which is difficult to escape [15].

1.4 ChIP-seq

The rapid development of high-throughput biotechnologies [16-25] has provided new insight and powerful support for regulatory mechanism analysis and genome-scale regulatory network elucidation. In particular, ChIP-seq provides massive protein-DNA interactive information and has been successfully applied to genome-wide analyses of TF binding, histone modification markers and polymerase binding [16, 26]. This technology can be summarized as follows: proteins are cross-linked to whole genome sequences [27, 28]; then DNA strands are sheared and immunoprecipitated to obtain sequence segments [16]; finally, these segments will be sequenced into short reads [29, 30]. These reads could be mapped onto their reference genome, if available, using Bowtie [31], BWA [32], etc. Based on the mapping results, the motif-enriched genomic regions could be identified by peak-calling tools [33], such as SPP [34], MACS[35], CisGenome [36], FindPeaks [37], QuEST [38], and PeakRanger [39]. These regions will be served as potential binding sites for motif identification.

1.5 DNA shape and shape motif

Four distinct DNA shape features can be derived in a high-throughput manner directly from DNA sequences based on the Monte Carlo simulation, which are Minor Groove Width (MGW), Propeller Twist (ProT), Helix Twist (HelT), and Roll. These features can provide structural information of DNA sequences and have predictive power in TF-DNA binding specificity. Recent studies have highlighted the complementary role of DNA shape and sequences in quantitatively modeling the TF-DNA binding specificity and motif prediction both *in vitro* and *in vivo* across multiple experimental assays and

diverse TF families [40-44]. In most cases, DNA shape-augmented models consistently improve the binding specificity prediction than models based on sequences alone.

Most recently, DNA shape features have been investigated individually using a two-step algorithm which is named ShapeMF [45]. This algorithm can be used to discovery *de novo* shape motifs based on shape-data, where shape motifs represent DNA shape patterns that can be recognized by TFs. The authors found that shape motifs are prevalent and recognized by many TFs, which is consistent with previous studies. Besides, some TFs enable to recognize shape motifs independently but without recognizing sequence motifs. This indicates that shape motif plays an important role in TF-DNA binding and makes a further influence in regulatory mechanisms. Rather than interpreting co-bound TFs use “tethering” mechanism only, ShapeMF revealed that some TFs extensively use shape-specific binding to form complexes with other TFs. Most importantly, the authors discovered that TFs with the same DNA binding domain have different shape motifs, which can interpret the phenomenon that such TFs recognize distinct binding regions in the human genome.

1.6 DL

Motivated by the hierarchical structure of animal’s visual system (from the retina to visual association cortex), DL has achieved the state-of-the-art in various machine learning fields, including visual object classification, natural language processing, and recommendation systems [46, 47]. Unlike traditional machine learning methods which need well-designed features, DL can learn feature representations automatically by classifying or fitting input data. Much of this interest is attributed to its multiple processing layers which can be used to learn representations of input data with multiple

levels of abstraction. Most importantly, DL can combine local features to form higher order features.

As one of the most important methods in DL, CNN has been successfully applied to image classification. Generally, CNNs are composed of the convolutional layer, pooling layer, and fully-connected layer (Figure 3). The convolutional layer is used to capture local features in given images, pooling layer enables reducing feature size and number of parameters, and the fully-connected layer is used for classification. Compared to previous algorithms in image classification, CNN alleviates the need for careful and time-consuming feature extraction.

1.7 Outline

The rest of this thesis is organized as follows: Chapter 2 will discuss motif identification from the promoter and ChIP-seq data, along with the application of DL in motif identification; Chapter 3 will introduce several motif-related works of BMBL. In Chapter 4, I developed a web server, DMINDA 2.0, which can provide integrative motif analysis. In Chapter 5, I proposed a DL framework, DESSO, which can be used to identify motifs from ChIP-seq data. Chapter 6 will discuss the conclusion of this thesis.

CHAPTER 2. Motif identification

2.1 Motif identification from promoter sequences

Substantial efforts have been devoted to seeking a reliable and efficient way for motif identification over the past few decades. Since the 1980s, identifying motifs in provided promoters has been one of the most prevalent approaches and numerous tools have been developed [8, 13, 48-53], such as Align ACE, BioProspector, CONSENSUS, MDscan, MEME, and BOBRO (Figure 4) [12, 13, 51, 52, 54-63]. Some of these tools have been successfully applied to various organisms for regulatory network construction [2, 5]. The underlying mechanism is that the co-regulated genes should exhibit overrepresented common motifs in their promoter regions. Although considerable efforts have been made, one non-negligible limitation is the high false positive rates in predictions [8, 64-66]. Under the assumption that the motifs in promoters tend to evolve at a lower rate and therefore be more conserved than non-functional surrounding sequences, some phylogenetic footprinting-based algorithms have been developed to reduce the false positive rate, such as PhyloGibbs, Footprinter, PhyloCon and MicroFootprinter [54, 67-71]. The phylogenetic footprinting strategy was firstly proposed in 1988 [72, 73] and has significantly improved the state-of-the-art performance in this field. However, the majority of programs under phylogenetic footprinting did not make full use of the phylogenetic relationship of query promoter sequences from various genomes [61]. Due to this limitation, some promoters from highly divergent species could be included and the motif instances are not conserved enough to carry out motif prediction [74-76]. Most recently, Liu *et al.* developed two computational pipelines aiming to break this bottleneck [4, 77]. Specifically, they extracted phylogenetic

relationships from regulatory sequences using a combinatorial framework based on 216 selected representative genomes to refine the orthologous promoter set. It is noteworthy that all the methods mentioned above could be potentially improved by integrating additional experimental data.

2.2 Motif identification from ChIP-seq data

Recent studies suggest that ChIP-seq can be effectively integrated into and benefit TFBS discovery tools [34, 38, 78-88]. It provides high-throughput motif signals and allows genome-scale discovery in a cell. More accurate binding regions (peaks) can be derived from ChIP-seq experiments, thus leading to more reliable prediction performance [8]. However, the peaks detected from ChIP-seq data can be up to a few hundred bps while the documented motifs are usually only as long as 8-20 bps [89]. Therefore, an *ab initio* motif discovery method is still indispensable to (i) identify the accurate binding sites from these ChIP-seq peaks, and (ii) build conserved motif profiles for further study in transcriptional regulation. Unfortunately, some widely used motif discovery tools, e.g. MEME and WEEDER [90], cannot be directly used on ChIP-seq peaks, since they are designed for co-regulated promoter sequences with limited size. Recently, some efforts have been made to rectify this problem by modifying traditional motif identification tools to adapt to the ChIP-seq data [83, 89, 91] or designing specific strategies for ChIP-seq-based motif identification [88, 92]. The computational challenges of these tools include, but not limited to, (i) huge amounts of sequenced ChIP-seq reads can make motif identification a computationally infeasible problem [8]; (ii) failure to identify the motifs associated with cofactors of the ChIP-ed TF [88] or *cis*-regulatory modules (CRMs) [93]; (iii) lack of insight in integration of ChIP-seq datasets from multiple TFs [94]; (iv) the

traditional false positive issue in motif prediction, caused by the noise in ChIP-seq technology [89]; (v) lack of an efficient way to determine the correct lengths of motifs except exhaustively enumerating each length within an interval [58, 95, 96]; and (vi) weak support in elucidation of the mutual interactions among multiple motifs from larger ChIP-ed datasets [97-99], which is very important in disease diagnosis through gene regulatory network construction.

2.3 DL in motif identification

Recent publications demonstrate that DL has improved the state-of-the-art performance in motif identification [100-102]. For example, DeepBind has utilized CNN to predict TF-DNA binding specificity on various genomic data types and has achieved the best performance [100]. The motif detectors in the trained model were then used to identify motifs. Compared to traditional methods, DeepBind enables extracting more complex patterns owing to its multi-layer architecture (Figure 5).

However, existing CNN models are limited by their ability to capture the long-range dependencies among motifs. Inspired by the recurrent neural network (RNN), which enables capturing the unbounded context in natural language, the models combining CNN and RNN have achieved a significant improvement in identifying more complex motif patterns [103, 104]. The downside of RNN is its inability to parallelize over sequential inputs, resulting in substantial processing steps as the length of input increases. Alternatively, the GCNN has been proposed and performs competitively on benchmarks [105]. It allows parallelization by stacking convolutions but still has the capability of capturing long-range dependencies of inputs.

Although existing DL-based methods equipped with various network architectures have been successfully employed in sequence motif-related problems, the sequence motifs have not been adequately considered and comprehensively analyzed [100, 101, 104, 106]. Moreover, DL has not been organically integrated with DNA shape in shape motif identification. Therefore, a reliable and efficient DL framework for motif identification based on ChIP-seq data and DNA shape is expected to be developed to improve the state-of-the-art performance.

CHAPTER 3. Previous work

3.1 BOBRO

BOBRO (BOttleneck BROken) was proposed in 2011 for identifying motifs in prokaryotes [107]. The performance of BOBRO has been demonstrated on large-scale datasets and identifies motifs more efficiently and accurately (at publication) than the best available tools such as MEME [108]. This appealing performance is mainly achieved by (i) a two-stage alignment strategy for reliably assessing the possibility for each position in each promoter to be the start of a conserved motif (Figure 6A); (ii) a dynamic way for constructing an unweighted graph to represent a list of potential motifs and their pairwise sequence similarities (Figure 6B); (iii) a novel method for identifying all the significant cliques which typically corresponds to the core part of the conserved motif in this graph (Figure 6C); and (iv) a highly reliable way to recognize actual motif incidences from the accidental ones based on the concept of ‘motif closure’ (Figure 6D).

3.2 BoBro 2.0

BoBro 2.0 is an integrated toolkit for motif identification and analysis [12]. This toolkit can (i) reliably identify statistically significant motifs at a genome-scale; (ii) accurately scan for all motif instances of a query motif in specified genomic regions using a novel method for *P*-value estimation; (iii) provide highly reliable comparisons and clustering of identified motifs, which takes into consideration the weak signals from the flanking regions of the motifs; and (iv) analyze co-occurring motifs in the regulatory regions.

We have carried out systematic comparisons between motif predictions using BoBro2.0 and the MEME package. The comparison results on *Escherichia coli* K12

genome and the human genome show that BoBro2.0 can identify the statistically significant motifs at a genome-scale more efficiently, identify motif instances more accurately and get more reliable motif clusters than MEME. In addition, BoBro2.0 provides correlational analyses among the identified motifs to facilitate the inference of joint regulation relationships of TFs.

3.3 DMINDA

DMINDA (DNA motif identification and analyses) is an integrated web server for motif identification (Figure 7) [109]. Key features of this server include (i) a high-performance web service for motif prediction and analyses, powered by a computer cluster with 150 computing nodes; (ii) identification and evaluation of conserved motifs at a genome scale (for prokaryotes) along with estimated statistical significance scores; (iii) an operon database DOOR, in support of prokaryotic motif identification in particular; (iv) accurate scan for all instances of a query motif in specified genomic sequences along with estimated statistical significance scores; (v) motif comparison and clustering for identified motifs, which takes into consideration the weakly conserved signals in the flanking regions of the motifs; and (vi) correlational analyses among the identified motifs to facilitate inference of joint regulatory relationships among TFs.

3.4 MP³

Motif prediction based on phylogenetic footprinting (MP³) is a new framework [110], aiming to develop new methods and strategies for (i) integrating the sequence-similarity and functional association information, (ii) promoter scoring and pruning through motif voting by a set of complementary predicting tools, (iii) motif signal cross-validation using a curve fitting way. Meanwhile, MP³ has been applied to the whole

genome of *E. coli* K12, which has plenty of documented TFBSs in RegulonDB [111]. Its performance was evaluated and compared with other seven existing tools. Specifically, the authors followed Tompa's strategy [64], which uses various statistics defined at the nucleotide level and at the binding site level to assess the correctness of the motif prediction. The comparison of statistics calculated on these tools shown that MP³ has significantly improved performance over other existing tools.

Such remarkable performance mainly benefits from four components of MP³ algorithms: reference promoter set (RPS) preparation from sequenced prokaryotic genomes, candidate binding region (CBR) detection by motif voting strategy and peak finding, candidate binding region clustering based on a graph model, and motif profiles identification through curve fitting (Figure 8). It is noteworthy that MP³ has the following unique features: (i) fully consideration of the operon structures; (ii) a new promoters collection method following a principle named as *huge data source, small final set*, which not only takes advantage of high throughput genomic data but also considers the computational efficiency; (iii) extracting phylogenetic information from regulatory sequences to refine the orthologous promoter set. Unlike in vertebrates, the lateral gene transfer and operon structure widely exist in prokaryotic genomes. Therefore, direct use of the species tree and the phylogenetic tree inferred from the targets genes isn't the best choice for prokaryotic genomes [61]; (iv) pruning promoters to generate CBRs based on the weighting score on each nucleotide, which is generated by a voting strategy on six popular motif identification tools; and (v) a curve-fitting method to identify optimal motif profiles. Here, these strategies with above features are different with all previously used ones thus will facilitate the application of phylogenetic footprinting.

CHAPTER 4. DMINDA 2.0

4.1 Introduction

Despite a lot of algorithms and tools that have been proposed and developed in the past few decades, most mainly focused on motif identification without integrating associated motif analyses [112]. Several web servers are available in the public domain, including the MEME Suite, PATLOC, AIMIE, Melina II, MotifSampler, and STAMP [113-118]. However, phylogenetic footprinting-based algorithms have not been fully considered. The identification and visualization of the relationship among identified motifs (or corresponding genes) remain unexplored. Hence, integrated web servers enabling reliable identification, comprehensive analyses, and intuitive visualization of motifs are still needed.

We have developed an updated version of the DMINDA motif analysis web server [109], DMINDA 2.0 [119], which is available at <http://bmbl.sdstate.edu/DMINDA2> and will be updated on a regular basis. Besides *de-novo* motif identification, motif scanning, motif comparison, and motif co-occurrence analysis, DMINDA 2.0 integrates two newly-published algorithms [4, 110], 2,125 complete genome sequences, and visualization and interpretation functionalities. DMINDA 2.0 has several key features, namely, (i) identification of motifs at a genome scale (for prokaryotes) along with estimated statistical significance values [107]; (ii) accurate scan for all motif instances of a query motif in specified genomic regions, and comparison and correlational analyses among the identified motifs to facilitate the inference of joint regulatory relationships among TFs [120]; (iii) 53 eukaryotic genomes downloaded from the Ensembl and JGI databases as of 01/12/2016 (including human,

mouse, and all the plant genomes) and genome-scale operons for 2,072 prokaryotes with complete genomes retrieved from the DOOR2 operon database [121], in support of the above motif-based analysis; (iv) an integrative phylogenetic footprinting framework for *de-novo* motif identification in prokaryotic genomes based on a global orthologous gene mapping algorithm [110, 122]; and (v) bacterial regulon (co-regulated operons by the same TF) prediction based on a new motif analysis framework and a novel graph model [4], along with a Cytoscape-like network interpretation and visualization function. A systematic comparison between DMINDA 2.0 and other six webservers indicates that DMINDA 2.0 and the MEME Suite can provide the most comprehensive motif identification and analysis functionalities (Figure 9).

4.2 Methods and results

There are six motif analysis functions in DMINDA 2.0 (Figure 10): (i) motif finding; (ii) motif scanning; (iii) motif comparison; (iv) motif co-occurrence analysis; (v) motif prediction by MP³; and (vi) regulon prediction.

The input data for (i) and (v) are DNA sequences in the FASTA format; motif alignments (or their PWMs) are required for (ii), (iii) and (iv); and species name along with operon/gene IDs are needed in (vi). These input data can be uploaded manually or selected from our underlying database by users.

The outputs of each function are: (i) aligned motif instances along with their motif logos and related sequence details; (ii) query motif logo and identified motif instances; (iii) similarity score, heat-map, and clustering tree of query motifs; (iv) identified co-occurrence motifs and their locational mapping to query genome sequences; (v) voting score curve and candidate binding regions along with same output in (i); and (vi)

identified regulons and their network visualization. All the outputs can be easily downloaded or converted for further computational analysis. The description of these six functions is shown below.

(i) *De novo motif finding* identifies a set of statistically significant motifs (if any) in a set of provided promoters (Figure 11). The backend algorithm, BOBRO [107], has been demonstrated on genome-scale datasets and does so more efficiently and accurately than the best available tools such as MEME [113].

(ii) *Motif scanning* scans for all motif instances of a query motif in given genomic sequences (Figure 12). The implemented tool, BBS (BoBro-based motif Scanning tool), has been shown to perform better than the MEME in accuracy on *E. coli* K12 and human genomes.

(iii) *Motif comparison* compares the similarity among the query motifs, and clusters similar motifs into groups (Figure 13). The implemented tool, BBC (BoBro-based motif Comparison and Clustering tool), identifies more accurate motif groups with a competitive sensitivity on synthetic datasets compared to MEME.

(iv) *Motif co-occurrence analysis* identifies co-occurring motifs which may regulate the same set of genes, in given regulatory sequences (Figure 14). The implemented tool, BBA (BoBro-based motif correlation Analysis tool), enables statistically significant TF pairs to be identified among 12,561 pairs of *E. coli* K12, with some of them have been fully or partially proven in the published literature.

The integration of the phylogenetic footprinting strategy and the systematic combination of motif-associated analyses have been integrated into a phylogenetic

footprinting framework for motif identification and bacterial regulon prediction in our server, respectively.

(v) *MP³* identifies novel motifs (if any) in prokaryotic genomes based on an integrative phylogenetic footprinting framework (Figure 15). Compared with seven prevalent programs on *E. coli* K12 genomes, *MP³* consistently achieved distinct improvement in motif identification accuracy. It mainly benefits from a new reference promoter preparation strategy, a promoter refining and pruning method, and the integration of six widespread motif identification tools serving as a candidate TFBSs search engine (Figure 16D).

(vi) *Regulon prediction* models and predicts regulons in given bacterial genomes (Figure 16A-C). Evaluated through documented regulons and co-expressed modules derived from *E. coli*, this method outperforms other algorithms across a wide variety of experiments. This remarkable performance is mainly achieved through the use of a novel computational framework and a graph model, integrating motif identification, motif comparison and clustering (i.e., functions (i), (iii), and (v)). To intuitively illustrate the predicted regulons, a Cytoscape-like visualization method was also implemented in support of further studies.

4.3 Conclusion

Motif identification and analyses provide a solid foundation to infer gene regulatory mechanism in a genome. Our previously published studies showed that, compared to the best available tools such as MEME, our implemented methods could identify and analyze statistically significant motifs equally, sometimes even better at a genome scale. We believe that our web server provides a highly useful and easy-to-use

platform for motif identification and analyses complementary to the existing web servers and tools, and benefits the genomic research community in general and prokaryotic genome researchers in particular. Until now, DMINDA 2.0 has been accessed about 5,000 times and cited by two published papers. Furthermore, approximately 1,000 jobs have been submitted by users.

Although DMINDA 2.0 enables motif identification from promoter sequences, it was limited in its ability to identify motif at a genome-scale based on ChIP-seq data. Existing ChIP-seq-based algorithms, however, suffer severe false positive and false negative issues, which is a big room to improve.

CHAPTER 5. DESSO

5.1 Introduction

Recent publications suggest that DL can also be extended in computational biology with unprecedented performance [123], particularly in motif identification [100, 101, 106]. Much of this interest is attributed to the PWM-like motif detectors in convolution module and fully-connected network for extracting higher-level motifs in prediction module [124]. The basic idea is to train a DL model to classify a huge amount of TF-bound sequences and unbound sequences. Each motif detector in the first convolutional layer represents a pattern which contributes to classification performance.

Although existing DL-based methods equipped with various network architectures have been successfully employed in sequence motif-related problems, the sequence motifs have not been fully considered and comprehensively analyzed [100, 101, 104, 106]. Currently, activation maximization is the most widely used strategy in sequence motif identification based on trained models. This strategy either aligns sequence fragments having maximum activation in each sequence [100, 104] or aligns sequence fragments having activation which are larger than half of maximum activation of motif detector on a set of sequences [101, 106]. This strategy based on the assumption that sequence fragments enable activating a motif detector are more likely being motif instances of the corresponding sequence motif. Such method, however, results in both severe false positive and false negative issues. In addition, the fact that several motif detectors cooperatively describe a pattern complex has not been taken into account [125]. Factorbook provides us integrative motif analysis of ChIP-seq data from ENCODE [126],

but the MEME-ChIP used in this study for motif identification is limited by its computational capability.

Here we introduce a DL framework, DESSO (DEep Sequence and Shape mOtif), which can be used to identify both sequence and shape motifs from ChIP-seq data.

5.2 Dataset

All 690 ChIP-seq datasets of uniform TFBS based on March 2012 ENCODE data freeze were downloaded from ENCODE Analysis Data at UCSC (<https://genome.ucsc.edu/ENCODE/downloads.html>). These datasets represent 161 unique TFs (generic and sequence-specific factors) and cover 91 human cell types [127]. Each dataset contains ranked peaks (ranked by their signal scores) which are derived from the SPP peak caller [34] and de-noised by the Irreproducible Discovery Rate (IDR) [128] based on signal reproducibility among biological replicates. The peaks range in number from 101 to 92,358.

We followed the same strategy of DeepBind [100] to split the peaks in each dataset into training data and test data. For each dataset, we define positive sequences as 101 bps centered on each peak summit, each of which has a label of 1. To overcome overfitting in model training, for a dataset with less than 10,000 peaks, we repeatedly generate random peaks with replacement from training data until having 10,000 positive sequences. Rather than generate negative sequences using dinucleotide-preserving shuffling, we randomly pick the same number of 101bp sequence bins from the hg19 human genome. These sequences are labeled as 0, which are deemed as unbounded sequences. The four normalized DNA shape feature (i.e., HelT, MGW, ProT, and Roll)

vectors of each sequence above are generated by an easy-to-use R package, DNashapeR [129].

DNase I Digital Genomic Footprinting (DNase-DGF) in a raw signal format derived from ENCODE/University of Washington were downloaded from UCSC (<http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeUwDgf>). They provide the footprint landscape of human genome for different cell lines using deep sequencing technique, based on the fact that unbound regions of regulatory factors in nucleosome-depleted chromatin are more sensitive to cleavage of DNase I. We only considered DNase-DGF of cell line K562 and A549 whose reads were mapped to the hg19 human genome.

5.3 Methods

5.3.1 DESSO

DESSO is a CNN-based framework for sequence and shape motif identification (Figure 17). Without loss of generality, here we use DNA sequence as an example to illustrate the sequence motif generation process. For each experiment, let M be the sequences of m top-ranked peaks, where each sequence is 101bp in size centered at each peak summit and $m = \min(500, \text{the total number of peaks})$. Define the *activation* score matrix M'_i as the activation values between a motif detector d_i (each has length L) and M by feeding M into convolution and ReLU layer of its corresponding trained model, and A_i the maximum score in M'_i . A sequence segment (L bp) with the largest activation score in each sequence is defined as an *activation* segment, if its activation score is larger than an activation cutoff C . A motif instance set, denoted as $\Omega(M, \lambda)$, is all activation segments with $C = \lambda \cdot A_i$ in M , where λ is a parameter ranging from 0 to 1. The value of

λ could be determined by a P -value strategy which is based on the assumption that the number of activation segment containing sequences using random selection with replacement in the human genome follows a binomial distribution. To estimate the “success” probability p of each random selection, we divided the human genome into non-overlapping bins with length 101bp, then randomly selected $n = 500,000$ bins as background sequence set H .

Let X be a random variable representing the number of activation segment containing bins with $C = \lambda \cdot A_i$ in H , $f(x) = P(X = x)$ be the probability function, and $F(t) = P(X \geq t)$ be the cumulative distribution function. It is assumed that $f(x)$ can be approximated by a binomial distribution $X \sim \text{Binomial}(n, p)$, where $p = \frac{x}{n}$ is a maximum likelihood estimate. Therefore, the P -value of $\Omega(M, \lambda)$ is given by:

$$F(|\Omega(M, \lambda)|) = P(X \geq |\Omega(M, \lambda)|) \quad (4)$$

For each motif detector d_i , we can obtain the optimal motif instance

$$\Omega(M, \lambda)_i = \operatorname{argmin}_{0 < \lambda < 1} F(|\Omega(M, \lambda)|)$$

and the corresponding P -value. Only $\Omega(M, \lambda)_i$ with P -value less than 1×10^{-4} and $|\Omega(M, \lambda)| > 2 * x$ were considered as true motif instances, based on the fact that motif should be more statistically significant and be observed more frequently in M . The derived motif instances were aligned as motif profiles and visualized using WebLogo 2.8.2 [130]. Each identified motif was compared with Homo sapiens motifs in JASPAR [131], TRANSFAC [132] and HOCOMOCO [133] using TOMTOM [134] with significance threshold $\text{FDR} < 0.05$. Shape motif generation follows almost the same

strategy as above, except that M and H should be replaced with corresponding DNA shape.

5.3.2 GCNN

The main feature of this proposed GCNN model is its “all-convolution” structure, including a convolutional layer, a recurrent convolution-gating block (CGB), and two fully connected layers (Figure 18). Concretely, the convolutional layer aims to detect motifs, the CGB captures long-term dependencies among identified motifs, and the fully connected layers account for binary classification. GCNN require as input the digit vectors, each DNA sequence is firstly transformed to a $n \times 4$ matrix M in one-hot format with $A = [1, 0, 0, 0]$, $T = [0, 1, 0, 0]$, $G = [0, 0, 1, 0]$, and $C = [0, 0, 0, 1]$. This input matrix M is then fed into a one-dimensional convolutional layer with multiple kernels E , where k indicates the number of kernels used. Each kernel is a $l \times 4$ weight matrix, which can be viewed as a motif detector. The core algorithm is summarized below.

Step 1: Slide each kernel in E along M with step size 1 to obtain the matched score on each position: $C = ReLU(conv_E(M))$. Here, C is an $(n - l + 1) \times 1 \times k$ matrix, where $ReLU(x) = \max(0, x)$ indicating rectified linear unit which is a widely-used activation function.

Step 2: Downsample the input C with pooling window size $h \times 1$ and step size h : $P = pool(C)$. P is a $d \times 1 \times k$ matrix, where $d = \lfloor \frac{n-l+1}{h} \rfloor$.

Step 3: Reshape P to a $d \times k \times 1$ matrix indicated by X . The hidden layers H_i for $i = 0, \dots, L$ can be obtained: $H_i = conv_W(X) \otimes \sigma(conv_V(X))$, where L is number of the

hidden layers in CGB, W and V are two convolutional kernels, σ is the sigmoid function ($\sigma(x) = \frac{1}{1+e^{-x}}$), and \otimes is used to calculate the element-wise product.

Step 4: Feed the output of CGB into a fully connected layer. The prediction is then transformed by the sigmoid function, indicated by $y \in [0, 1]$.

5.4 Results

5.4.1 DNA shape has strong predictive power in TF-DNA binding specificity prediction

To predict TF-DNA binding specificity of each TF in different cell types, we constructed DESSO to distinguish bound and unbound regions by learning patterns which are embedded in these regions. This framework was applied to 690 *in vivo* ENCODE ChIP-seq datasets, each of which contains TF uniform peaks derived from a uniform processing pipeline [127]. For each dataset, the top 500 even-number peaks are served as test data, and the remaining peaks are used in model training, and all these peaks represent the positive class. As a binary classification problem, the corresponding negative class is also required for model training. Based on our observation, the performance of trained models is heavily dependent on the choice of negative sequences preparation. To make it more accurate and biologically meaningful, we randomly pick unbound regions in the genome as negative sequences, as opposed to using dinucleotide-preserving shuffle strategy (Figure 19A).

In addition to these DNA sequences, we also applied this framework to their four DNA shape features to evaluate the predictive power of DNA shape on TF-DNA binding specificity. In spite of essential role of DNA shape in TF-DNA recognition suggested by recent studies, it remains incompletely understood to what extent DNA sequence and

DNA shape can quantitatively contribute to this process. To investigate this, the DESSO was also applied to the combination of DNA sequence and DNA shape. All these models were evaluated on held-out test data using the area under the receiver operating characteristic curve (AUC) (Figure 19B). Results show that DNA shape has strong predictive power in TF-DNA binding specificity prediction, and this performance can even be more enhanced when all four shape features cooperate with each other. The models based on sequence alone achieve the best performance. Unlike previous work, incorporation of DNA shape cannot improve the predictive performance compared with using sequence alone. This may be because DL enables extracting DNA shape features from sequences.

5.4.2 Identification of sequence and shape motifs

We next identified both sequence and shape motifs in each experiment using top 500 peaks (if there are more than 500 peaks, otherwise, all peaks were used) by feeding these peaks into the trained model. Rather than choose motif instances using a subjective cutoff, we introduced a *P*-value strategy based on binomial distribution [135]. Only the significant motifs ($P\text{-value} < 1 \times 10^{-4}$) which are more enriched in these 500 sequences than random sequences were retained for further analysis. The retained motif with the lowest *P*-value was defined as a primary motif, and the others are defined as secondary motifs [126]. Redundant sequence motifs were merged based on their similarity score (> 0.9) from BBC [120].

Finally, a total of 82 unique primary sequence motifs were identified, 65 of which can be found in the JASPAR [131] or TRANSFAC [132]. We computed the DNaseI Digital Genomic Footprinting [136] and evolutionary conservation [137] of identified

motif instances (Figure 20A and 20B). To check the enrichment of the identified motifs in all ranked peaks of each experiment, we scanned motif occurrence using BBS [120] and computed enrichment score (ES) for each motif using GSEA (Gene Set Enrichment Analysis software) (Figure 20C) [138]. It showed that identified motifs are more enriched in top-ranked peaks, and primary motifs have dramatic left-skewed trend indicating their predominant role in the discovery of these peaks. The percentage of peaks covered by identified motifs achieved 0.85 in average for only primary motif and 0.91 for both primary and secondary motifs. This revealed the complementary role of secondary motifs in TF-DNA binding, which means secondary motifs may bound by some cofactors.

We also identified 35, 62, 59, and 55 unique primary shape motifs for HelT, MGW, ProT, and Roll, respectively. 322 TFs have at least one shape motif, and MGW is the most prevalent one. This disclosed the shape preference of TF, which is mainly determined by TFs' DNA binding domain. It obviously demonstrated that TFs with identical sequence motifs can have distinct shape motifs. We generated sequence logo for sequences of shape motif instances, some of them are not conserved and do not have matched motif, but some of them corresponds to TF's motif. For example, MAFF recognizes Maf recognition element [TGCTGAC(G)TCAGCA]. One of the sequence logos of the identified HelT motif corresponds to its sequence motif (Figure 21A), but another one is not (Figure 21B). Additionally, most of these sequence logo have low IC compared to their corresponding sequence motifs, indicating that shape motifs are generally not well-conserved at the sequence level.

5.4.3 GCNN captures motif dependencies of long DNA sequences

Based on our observation, the classification accuracy of CNN can be improved significantly as peak length increases (Figure 22A). To investigate the performance of GCNN, we applied our GCNN model on DNA sequences with length 1001 bps. The results show that GCNN outperforms CNN on most of the datasets (Figure 22B). This remarkable improvement mainly benefits from GCNN's capability in capturing motif dependencies.

5.5 Conclusion

In this chapter, we developed a DL framework to identify sequence and shape motifs from ChIP-seq data. Unlike previous work using a solid threshold for motif identification, here we introduced a binomial distribution to select the optimal threshold. For long DNA sequences, we developed a GCNN model to capture motif dependencies. To broadly facilitate motif-related analysis in this field, we also provide an integrated web server DESSO, which is freely available at <http://bmbl.sdstate.edu/DESSO>. In addition to showing derived results based on 690 ENCODE ChIP-seq datasets, DESSO enables a comprehensive analysis of user-provided DNA sequences, along with a 2-dimensional convolutional network visualization to exhibit its actual behavior [139].

CHAPTER 6. Discussion

TFBSs play critical roles in regulating transcription rates and expression levels of their target genes. The knowledge of genome-scale TFBSs can greatly help the elucidation of gene regulatory mechanisms in a cell. Hence, *de-novo* motif identification and associated computational analyses (e.g., motif scanning and comparison) play an important role in regulatory network construction in all organisms.

Although substantial algorithms and tools have been developed in the past few decades, phylogenetic footprinting-based algorithms have not been fully considered. Additionally, no such work has considered the relationship and visualization among identified motifs (or corresponding genes). Existing DL models use a subjective threshold in motif identification, which incurs severe false positive and false negative issues. Furthermore, DNA shape has not been integrated with DL in shape motif identification.

To overcome these limitations, we have made two main contributions as follows: We have developed an integrated web server, DMINDA 2.0, which contains: (i) five motif prediction and analysis algorithms, including a phylogenetic footprinting framework; (ii) 2,125 species with complete genomes to support the above five functions, covering animals, plants, and bacteria; and (iii) bacterial regulon prediction and visualization. Compared to other existing web servers, DMINDA 2.0 provides comprehensive motif analysis functions. DMINDA 2.0 is freely available at <http://bmb1.sdstate.edu/DMINDA2>.

We have proposed a DL framework, DESSO, which is used to identify both sequence and shape motifs from ChIP-seq data. To optimize the threshold in motif identification, we introduced a binomial distribution to select the best threshold based on

a *P*-value strategy. Results show that DNA shape also has strong predictive power in TF-DNA binding specificity prediction. In addition, shape motifs are prevalent and can help interpret why TFs with the same sequence motif bind to distinct genome regions.

Compared to CNN, the GCNN model proposed in this study can improve TF-DNA binding specificity prediction on long DNA sequences. This performance mainly benefits from GCNN's capability in capturing motif dependencies.

REFERENCES

1. D'Haeseleer, P., *What are DNA sequence motifs?* Nature Biotechnology, 2006. **24**(4): p. 423-5.
2. Brohee, S., et al., *Unraveling networks of co-regulated genes on the sole basis of genome sequences.* Nucleic Acids Res, 2011. **39**(15): p. 6340-58.
3. Davidson, E. and M. Levin, *Gene regulatory networks.* Proc Natl Acad Sci U S A, 2005. **102**(14): p. 4935.
4. Liu, B., et al., *Bacterial regulon modeling and prediction based on systematic cis regulatory motif analyses.* Scientific reports, 2016. **6**.
5. Baumbach, J., *On the power and limits of evolutionary conservation--unraveling bacterial gene regulatory networks.* Nucleic Acids Res, 2010. **38**(22): p. 7877-84.
6. Liu, B., et al., *An algorithmic perspective of de novo cis-regulatory motif finding based on ChIP-seq data.* Briefings in Bioinformatics, 2017: p. bbx026.
7. Stormo, G.D., *DNA binding sites: representation and discovery.* Bioinformatics, 2000. **16**(1): p. 16-23.
8. Zambelli, F., G. Pesole, and G. Pavesi, *Motif discovery and transcription factor binding sites before and after the next-generation sequencing era.* Brief Bioinform, 2013. **14**(2): p. 225-37.
9. Nishida, K., M.C. Frith, and K. Nakai, *Pseudocounts for transcription factor binding sites.* Nucleic Acids Res, 2009. **37**(3): p. 939-44.
10. Nishida, K., M.C. Frith, and K. Nakai, *Pseudocounts for transcription factor binding sites.* Nucleic acids research, 2008. **37**(3): p. 939-944.

11. Weirauch, M.T., et al., *Evaluation of methods for modeling transcription factor sequence specificity*. Nat Biotechnol, 2013. **31**(2): p. 126-34.
12. Ma, Q., et al., *An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale*. Bioinformatics, 2013. **29**(18): p. 2261-8.
13. Das, M.K. and H.K. Dai, *A survey of DNA motif finding algorithms*. BMC Bioinformatics, 2007. **8 Suppl 7**: p. S21.
14. Bailey, T.L. and C. Elkan. *Fitting a mixture model by expectation maximization to discover motifs in biopolymers*. in *International Conference on Intelligent Systems for Molecular Biology*. 1994.
15. Ikebata, H. and R. Yoshida, *Repulsive parallel MCMC algorithm for discovering diverse motifs from large sequence sets*. Bioinformatics, 2015. **31**(10): p. 1561-8.
16. Park, P.J., *ChIP-seq: advantages and challenges of a maturing technology*. Nature Reviews Genetics, 2009. **10**(10): p. 669-680.
17. Lingyun Song, G.E.C., *DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells*. Cold Spring Harbor Protocols, 2010. **2010**(2).
18. Wang, Z., et al., *RNA-Seq: a revolutionary tool for transcriptomics*. Nature Reviews Genetics, 2008. **10**(1): p. 57-63.
19. Tsankov, A.M., et al., *Transcription factor binding dynamics during human ES cell differentiation*. Nature, 2015. **518**(7539): p. 344-9.
20. Wu, F., B.G. Olson, and J. Yao, *DamID-seq: Genome-wide Mapping of Protein-DNA Interactions by High Throughput Sequencing of Adenine-methylated DNA Fragments*. Journal of Visualized Experiments Jove, 2015(107).

21. Maragkakis, M., et al., *CLIPSeqTools-a novel bioinformatics CLIP-seq analysis suite*. RNA (New York, N.Y.), 2015. **22**(1).
22. Hafner, M., et al., *PAR-CliP - A Method to Identify Transcriptome-wide the Binding Sites of RNA Binding Proteins*. Journal of Visualized Experiments, 2010. **41**(41): p. e2034-e2034.
23. Ingolia, N.T., *Ribosome profiling: new views of translation, from single codons to genome scale*. Nature Reviews Genetics, 2014. **15**(3): p. 205-13.
24. Giresi, P.G., et al., *FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin*. Genome Research, 2007. **17**(6): p. 877-85.
25. Nutiu, R., et al., *Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument*. Nature Biotechnology, 2011. **29**(7): p. 659-64.
26. Collas, P. and J.A. Dahl, *Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation*. Frontiers in Bioscience A Journal & Virtual Library, 2008. **13**(4): p. 929-943.
27. Kimura, H. and Y. Sato, *DNA Replication and Histone Modification*. 2016: Springer Japan.
28. Suganuma, T. and J.L. Workman, *Histone modification as a reflection of metabolism*. Cell Cycle, 2016.
29. Qu, H. and X. Fang, *A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project*. Genomics Proteomics Bioinformatics, 2013. **11**(3): p. 135-41.

30. Consortium, E.P., *The ENCODE (ENCyclopedia Of DNA Elements) Project*. Science, 2004. **306**(5696): p. 636-40.
31. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*. Genome Biology, 2009. **10**(3): p. 1-10.
32. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows–Wheeler transform*. Bioinformatics, 2010. **26**(5): p. 589-595.
33. Szalkowski, A.M. and C.D. Schmid, *Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts*. Briefings in Bioinformatics, 2011. **12**(6): p. 626-633(8).
34. Kharchenko, P.V., M.Y. Tolstorukov, and P.J. Park, *Design and analysis of ChIP-seq experiments for DNA-binding proteins*. Nature biotechnology, 2008. **26**(12): p. 1351-1359.
35. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS)*. Genome Biol, 2008. **9**(9): p. R137.
36. Jiang, H., et al., *CisGenome Browser: a flexible tool for genomic data visualization*. Bioinformatics, 2010. **26**(14): p. 1781-2.
37. Fejes, A.P., et al., *FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology*. Bioinformatics, 2008. **24**(15): p. 1729-30.
38. Valouev, A., et al., *Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data*. Nature Methods, 2008. **5**(9): p. 829-834.
39. Xin, F., R. Grossman, and L. Stein, *PeakRanger: a cloud-enabled peak caller for ChIP-seq data*. BMC Bioinformatics, 2011. **12**(10): p. 139-139.

40. Zhou, T., et al., *Quantitative modeling of transcription factor binding specificities using DNA shape*. Proceedings of the National Academy of Sciences, 2015. **112**(15): p. 4654-4659.
41. Yang, L., et al., *Transcription factor family-specific DNA shape readout revealed by quantitative specificity models*. Molecular Systems Biology, 2017. **13**(2): p. 910.
42. Mathelier, A., et al., *DNA shape features improve transcription factor binding site predictions in vivo*. Cell systems, 2016. **3**(3): p. 278-286. e4.
43. Zentner, G.E., et al., *ChEC-seq kinetics discriminates transcription factor binding sites by DNA sequence and shape in vivo*. Nature communications, 2015. **6**.
44. Abe, N., et al., *Deconvolving the recognition of DNA shape from sequence*. Cell, 2015. **161**(2): p. 307-318.
45. Samee, M.A.H., B. Bruneau, and K. Pollard, *Transcription Factors Recognize DNA Shape Without Nucleotide Recognition*. bioRxiv, 2017: p. 143677.
46. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning*. Nature, 2015. **521**(7553): p. 436-444.
47. Wang, H., B. Raj, and E.P. Xing, *On the Origin of Deep Learning*. arXiv preprint arXiv:1702.07800, 2017.
48. Lawrence, C.E., et al., *Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment*. Science, 1993. **262**(5131): p. 208-14.
49. Pevzner, P.A. and S.H. Sze, *Combinatorial approaches to finding subtle signals in DNA sequences*. Proc Int Conf Intell Syst Mol Biol, 2000. **8**: p. 269-78.

50. Nakaki, R., J. Kang, and M. Tateno, *A novel ab initio identification system of transcriptional regulation motifs in genome DNA sequences based on direct comparison scheme of signal/noise distributions*. *Nucleic Acids Res*, 2012. **40**(18): p. 8835-48.
51. Li, G., et al., *A new framework for identifying cis-regulatory motifs in prokaryotes*. *Nucleic Acids Res*, 2011. **39**(7): p. e42.
52. Chen, X., et al., *W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data*. *Bioinformatics*, 2008. **24**(9): p. 1121-8.
53. Sinha, S., *PhyME: a software tool for finding motifs in sets of orthologous sequences*. *Methods Mol Biol*, 2007. **395**: p. 309-18.
54. Wang, T. and G.D. Stormo, *Combining phylogenetic data with co-regulated genes to identify regulatory motifs*. *Bioinformatics*, 2003. **19**(18): p. 2369-80.
55. Liu, X., D.L. Brutlag, and J.S. Liu, *BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes*. *Pac Symp Biocomput*, 2001: p. 127-38.
56. Hertz, G.Z. and G.D. Stormo, *Identifying DNA and protein patterns with statistically significant alignments of multiple sequences*. *Bioinformatics*, 1999. **15**(7-8): p. 563-77.
57. Liu, X.S., D.L. Brutlag, and J.S. Liu, *An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments*. *Nat Biotechnol*, 2002. **20**(8): p. 835-9.

58. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching*. Nucleic Acids Res, 2009. **37**(Web Server issue): p. W202-8.
59. Olman, V., D. Xu, and Y. Xu, *CUBIC: identification of regulatory binding sites through data clustering*. J Bioinform Comput Biol, 2003. **1**(1): p. 21-40.
60. Li, X. and W.H. Wong, *Sampling motifs on phylogenetic trees*. Proc Natl Acad Sci U S A, 2005. **102**(27): p. 9481-6.
61. Blanchette, M. and M. Tompa, *Discovery of regulatory elements by a computational method for phylogenetic footprinting*. Genome Res, 2002. **12**(5): p. 739-48.
62. Blanchette, M. and M. Tompa, *FootPrinter: A program designed for phylogenetic footprinting*. Nucleic Acids Res, 2003. **31**(13): p. 3840-2.
63. Li, G., B. Liu, and Y. Xu, *Accurate recognition of cis-regulatory motifs with the correct lengths in prokaryotic genomes*. Nucleic Acids Res, 2010. **38**(2): p. e12.
64. Tompa, M., et al., *Assessing computational tools for the discovery of transcription factor binding sites*. Nat Biotechnol, 2005. **23**(1): p. 137-44.
65. McCue, L.A., et al., *Factors influencing the identification of transcription factor binding sites by cross-species comparison*. Genome Res, 2002. **12**(10): p. 1523-32.
66. Simcha, D., N.D. Price, and D. Geman, *The limits of de novo DNA motif discovery*. PLoS One, 2012. **7**(11): p. e47836.
67. Siddharthan, R., E.D. Siggia, and E. van Nimwegen, *PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny*. PLoS Comput Biol, 2005. **1**(7): p. e67.

68. Blanchette, M., B. Schwikowski, and M. Tompa, *Algorithms for phylogenetic footprinting*. J Comput Biol, 2002. **9**(2): p. 211-23.
69. Neph, S. and M. Tompa, *MicroFootPrinter: a tool for phylogenetic footprinting in prokaryotic genomes*. Nucleic Acids Res, 2006. **34**(Web Server issue): p. W366-8.
70. Carmack, C.S., et al., *PhyloScan: identification of transcription factor binding sites using cross-species evidence*. Algorithms Mol Biol, 2007. **2**: p. 1.
71. Zhang, S., et al., *Genome-wide de novo prediction of cis-regulatory binding sites in prokaryotes*. Nucleic Acids Res, 2009. **37**(10): p. e72.
72. Katara, P., A. Grover, and V. Sharma, *Phylogenetic footprinting: a boost for microbial regulatory genomics*. Protoplasma, 2012. **249**(4): p. 901-7.
73. Tagle, D.A., et al., *Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints*. J Mol Biol, 1988. **203**(2): p. 439-55.
74. Borneman, A.R., et al., *Divergence of transcription factor binding sites across related yeast species*. Science, 2007. **317**(5839): p. 815-819.
75. Odom, D.T., et al., *Tissue-specific transcriptional regulation has diverged significantly between human and mouse*. Nature genetics, 2007. **39**(6): p. 730-732.
76. Boyle, A.P., et al., *Comparative analysis of regulatory information and circuits across distant species*. Nature, 2014. **512**(7515): p. 453-456.

77. Bingqiang Liu, C.Z., Hanyuan Zhang, Guojun Li, Guanghui Wang, Anne Fennell, Yu Kang, Qi Liu and Qin Ma, *An integrative and applicable phylogenetic footprinting framework for cis-regulatory motifs identification in prokaryotic genomes*. BMC Genomics, 2016.
78. Kuan, P.F., et al., *A statistical framework for the analysis of ChIP-Seq data*. Journal of the American Statistical Association, 2011. **106**(495): p. 891-903.
79. Mathelier, A. and W.W. Wasserman, *The next generation of transcription factor binding site prediction*. PLoS Comput Biol, 2013. **9**(9): p. e1003214.
80. Cheng, C., R. Min, and M. Gerstein, *TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles*. Bioinformatics, 2011. **27**(23): p. 3221-3227.
81. Wu, S., et al., *ChIP-PaM: an algorithm to identify protein-DNA interaction using ChIP-Seq data*. Theoretical biology and medical modelling, 2010. **7**(1): p. 1.
82. van Heeringen, S.J. and G.J.C. Veenstra, *GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments*. Bioinformatics, 2011. **27**(2): p. 270-271.
83. Machanick, P. and T.L. Bailey, *MEME-ChIP: motif analysis of large DNA datasets*. Bioinformatics, 2011. **27**(12): p. 1696-7.
84. Kulakovskiy, I.V., et al., *Deep and wide digging for binding motifs in ChIP-Seq data*. Bioinformatics, 2010. **26**(20): p. 2622-3.
85. Jothi, R., et al., *Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data*. Nucleic acids research, 2008. **36**(16): p. 5221-5231.

86. Mercier, E., et al., *An integrated pipeline for the genome-wide analysis of transcription factor binding sites from ChIP-Seq*. PLoS One, 2011. **6**(2): p. e16432.
87. Hu, M., et al., *On the detection and refinement of transcription factor binding sites using ChIP-Seq data*. Nucleic Acids Res, 2010. **38**(7): p. 2154-67.
88. Bailey, T.L., *DREME: motif discovery in transcription factor ChIP-seq data*. Bioinformatics, 2011. **27**(12): p. 1653-9.
89. Jia, C., et al., *A new exhaustive method and strategy for finding motifs in ChIP-enriched regions*. Plos One, 2014. **9**(9): p. e86044.
90. Pavesi, G., et al., *Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes*. Nucleic Acids Res, 2004. **32**(Web Server issue): p. W199-203.
91. Thomas-Chollier, M., et al., *RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets*. Nucleic Acids Res, 2012. **40**(4): p. e31.
92. Tran, N.T. and C.H. Huang, *A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data*. Biol Direct, 2014. **9**: p. 4.
93. Jun Ding, H.H., Xiaoman Li, *SIOMICS: a novel approach for systematic identification of motifs in ChIP-seq data*. Nucleic Acids Research, 2014. **42**(5): p. 1635-1645.
94. Boeva, V., et al., *De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis*. Nucleic Acids Res, 2010. **38**(11): p. e126.

95. Bailey, T.L., et al., *MEME: discovering and analyzing DNA and protein sequence motifs*. Nucleic Acids Research, 2006. **34**(2): p. 369-73.
96. Holger, H., et al., *P-value-based regulatory motif discovery using positional weight matrices*. Genome Research, 2013. **23**(1): p. 181-194.
97. Niu and Meng, *De novo prediction of cis-regulatory modules in eukaryotic organisms*. Dissertations & Theses - Gradworks, 2014.
98. Bolouri, H. and W.L. Ruzzo, *Integration of 198 ChIP-seq datasets reveals human cis-regulatory regions*. J Comput Biol, 2012. **19**(9): p. 989-97.
99. Sun, H., et al., *Unveiling combinatorial regulation through the combination of ChIP information and in silico cis-regulatory module detection*. Nucleic Acids Res, 2012. **40**(12): p. e90.
100. Alipanahi, B., et al., *Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning*. Nature biotechnology, 2015. **33**(8): p. 831-838.
101. Kelley, D.R., J. Snoek, and J.L. Rinn, *Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks*. Genome research, 2016. **26**(7): p. 990-999.
102. Zhou, J. and O.G. Troyanskaya, *Predicting effects of noncoding variants with deep learning-based sequence model*. Nature methods, 2015. **12**(10): p. 931-934.
103. Lanchantin, J., et al., *Deep Motif Dashboard: Visualizing and Understanding Genomic Sequences Using Deep Neural Networks*. arXiv preprint arXiv:1608.03644, 2016.

104. Quang, D. and X. Xie, *DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences*. Nucleic acids research, 2016. **44**(11): p. e107-e107.
105. Dauphin, Y.N., et al., *Language modeling with gated convolutional networks*. arXiv preprint arXiv:1612.08083, 2016.
106. Angermueller, C., et al., *DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning*. Genome Biology, 2017. **18**(1): p. 67.
107. Li, G., et al., *A new framework for identifying cis-regulatory motifs in prokaryotes*. Nucleic acids research, 2011. **39**(7): p. e42-e42.
108. Li, G., et al., *A new framework for identifying cis-regulatory motifs in prokaryotes*. Nucleic acids research, 2010. **39**(7): p. e42-e42.
109. Ma, Q., et al., *DMINDA: an integrated web server for DNA motif identification and analyses*. Nucleic acids research, 2014: p. gku315.
110. Liu, B., et al., *An integrative and applicable phylogenetic footprinting framework for cis-regulatory motifs identification in prokaryotic genomes*. BMC genomics, 2016. **17**(1): p. 578.
111. Gama-Castro, S., et al., *RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation*. Nucleic Acids Res, 2008. **36**(Database issue): p. D120-4.
112. Tompa, M., et al., *Assessing computational tools for the discovery of transcription factor binding sites*. Nature biotechnology, 2005. **23**(1): p. 137-144.

113. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching*. Nucleic acids research, 2009: p. gkp335.
114. Mrázek, J. and S. Xie, *Pattern locator: a new tool for finding local sequence patterns in genomic DNA sequences*. Bioinformatics, 2006. **22**(24): p. 3099-3100.
115. Mrázek, J., et al., *AIMIE: a web-based environment for detection and interpretation of significant sequence motifs in prokaryotic genomes*. Bioinformatics, 2008. **24**(8): p. 1041-1048.
116. Okumura, T., et al., *Melina II: a web tool for comparisons among several predictive algorithms to find potential motifs from promoter regions*. Nucleic acids research, 2007. **35**(suppl 2): p. W227-W231.
117. Thijs, G., et al., *A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes*. Journal of Computational Biology, 2002. **9**(2): p. 447-464.
118. Mahony, S. and P.V. Benos, *STAMP: a web tool for exploring DNA-binding motif similarities*. Nucleic acids research, 2007. **35**(suppl 2): p. W253-W258.
119. Yang, J., et al., *DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses*. Bioinformatics, 2017.
120. Ma, Q., et al., *An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale*. Bioinformatics, 2013. **29**(18): p. 2261-2268.
121. Mao, X., et al., *DOOR 2.0: presenting operons and their functions through dynamic and integrated views*. Nucleic acids research, 2014. **42**(D1): p. D654-D659.

122. Li, G., et al., *Integration of sequence-similarity and functional association information can overcome intrinsic problems in orthology mapping across bacterial genomes*. Nucleic acids research, 2011. **39**(22): p. e150-e150.
123. Angermueller, C., et al., *Deep learning for computational biology*. Molecular systems biology, 2016. **12**(7): p. 878.
124. Park, Y. and M. Kellis, *Deep learning for regulatory genomics*. Nat Biotechnol, 2015. **33**(8): p. 825-6.
125. Ching, T., et al., *Opportunities And Obstacles For Deep Learning In Biology And Medicine*. bioRxiv, 2017: p. 142760.
126. Wang, J., et al., *Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors*. Genome research, 2012. **22**(9): p. 1798-1812.
127. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome*. Nature, 2012. **489**(7414): p. 57.
128. Li, Q., et al., *Measuring reproducibility of high-throughput experiments*. The annals of applied statistics, 2011. **5**(3): p. 1752-1779.
129. Chiu, T.-P., et al., *DNASHapeR: an R/Bioconductor package for DNA shape prediction and feature encoding*. Bioinformatics, 2015. **32**(8): p. 1211-1213.
130. Crooks, G.E., et al., *WebLogo: a sequence logo generator*. Genome research, 2004. **14**(6): p. 1188-1190.
131. Mathelier, A., et al., *JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles*. Nucleic acids research, 2016. **44**(D1): p. D110-D115.

132. Matys, V., et al., *TRANSFAC®: transcriptional regulation, from patterns to profiles*. Nucleic acids research, 2003. **31**(1): p. 374-378.
133. Kulakovskiy, I.V., et al., *HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models*. Nucleic acids research, 2016. **44**(D1): p. D116-D125.
134. Gupta, S., et al., *Quantifying similarity between motifs*. Genome biology, 2007. **8**(2): p. R24.
135. Heinz, S., et al., *Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities*. Molecular cell, 2010. **38**(4): p. 576-589.
136. Neph, S., et al., *An expansive human regulatory lexicon encoded in transcription factor footprints*. Nature, 2012. **489**(7414): p. 83.
137. Pollard, K.S., et al., *Detection of nonneutral substitution rates on mammalian phylogenies*. Genome research, 2010. **20**(1): p. 110-121.
138. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles*. Proceedings of the National Academy of Sciences, 2005. **102**(43): p. 15545-15550.
139. Harley, A.W. *An interactive node-link visualization of convolutional neural networks*. in *International Symposium on Visual Computing*. 2015. Springer.



Figure 1. Motif instances and logo.

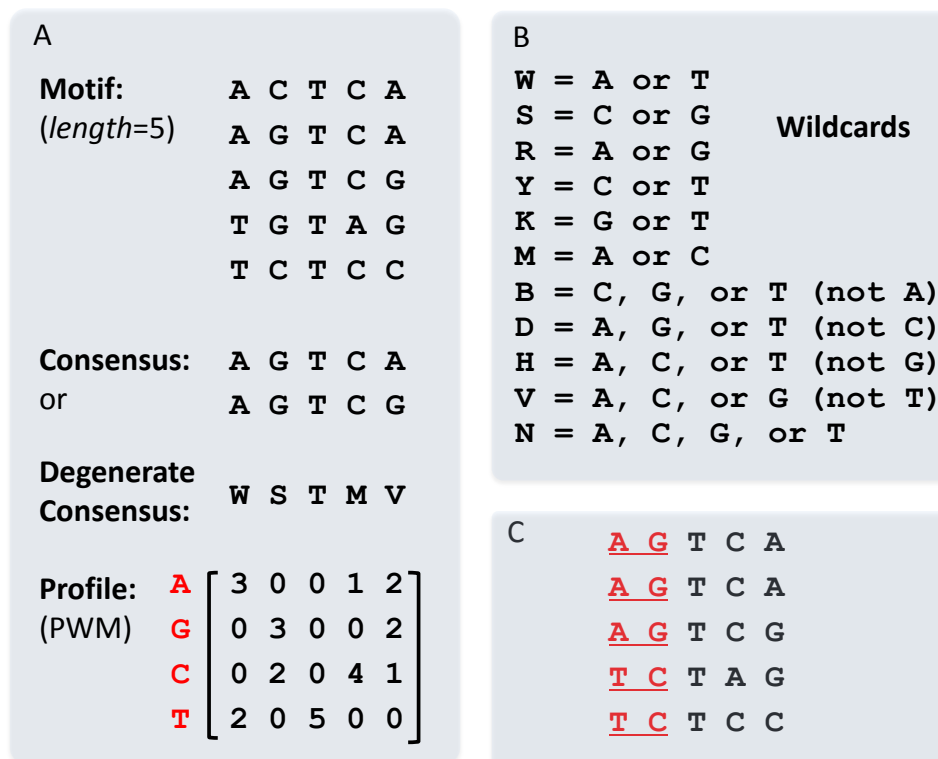


Figure 2. Representation of a motif. (A) An example of motif consensus, degenerate consensus, and profile. (B) A full list of wildcards in the degenerate consensus. (C) A different motif but has same profile with the motif in A.

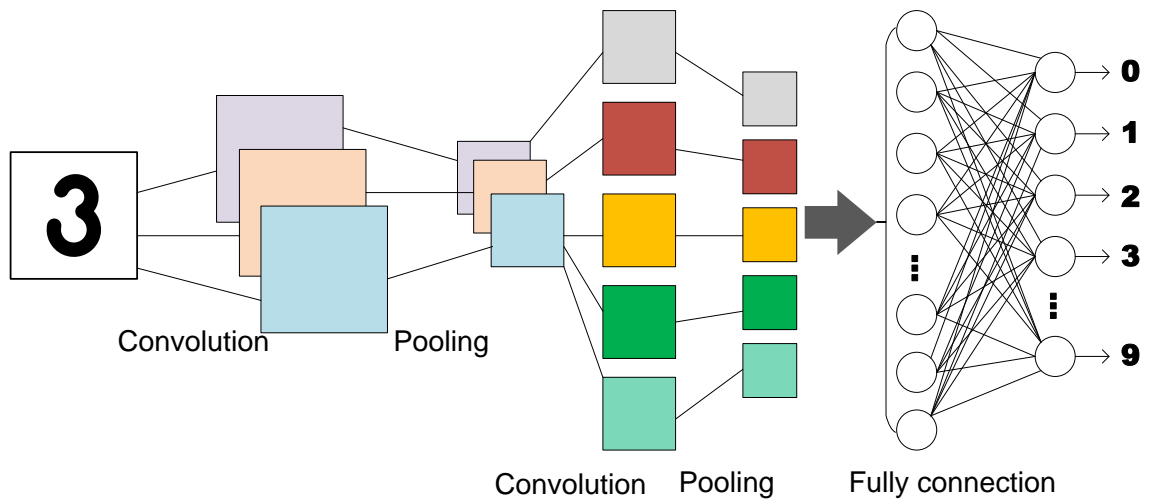


Figure 3. CNN in image classification based on the multi-layer structure: convolutional layer, pooling layer, and fully-connected layer.

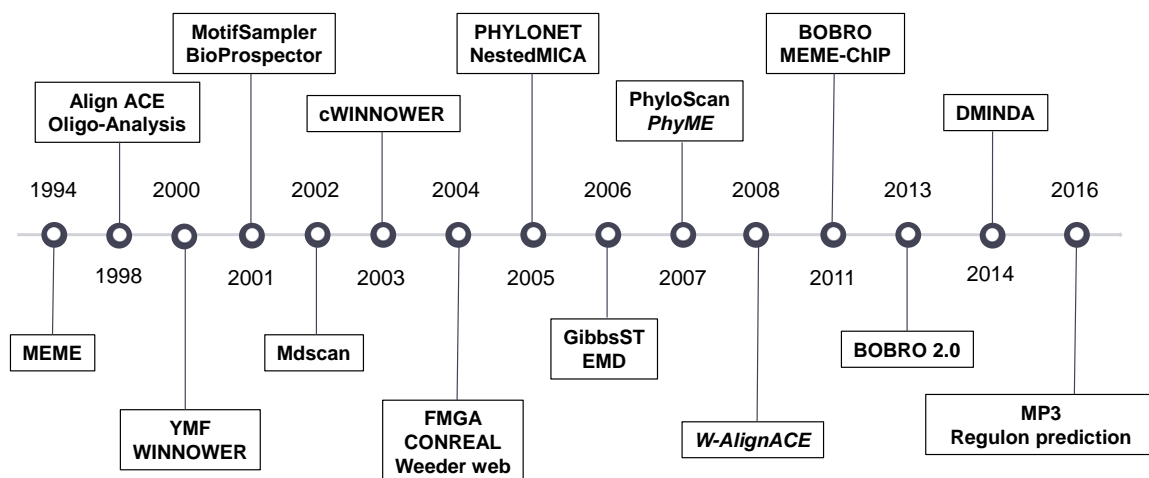


Figure 4. Existing algorithms and tools for motif identification.

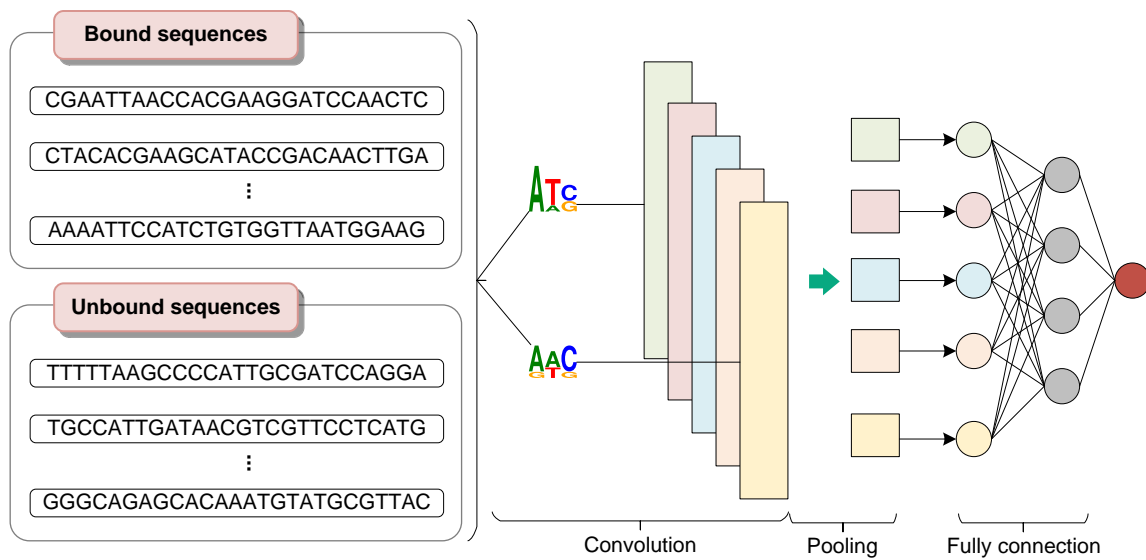


Figure 5. DL framework for motif identification, including bound and unbound sequences as training data, convolutional layer, pooling layer, and fully-connected layer.

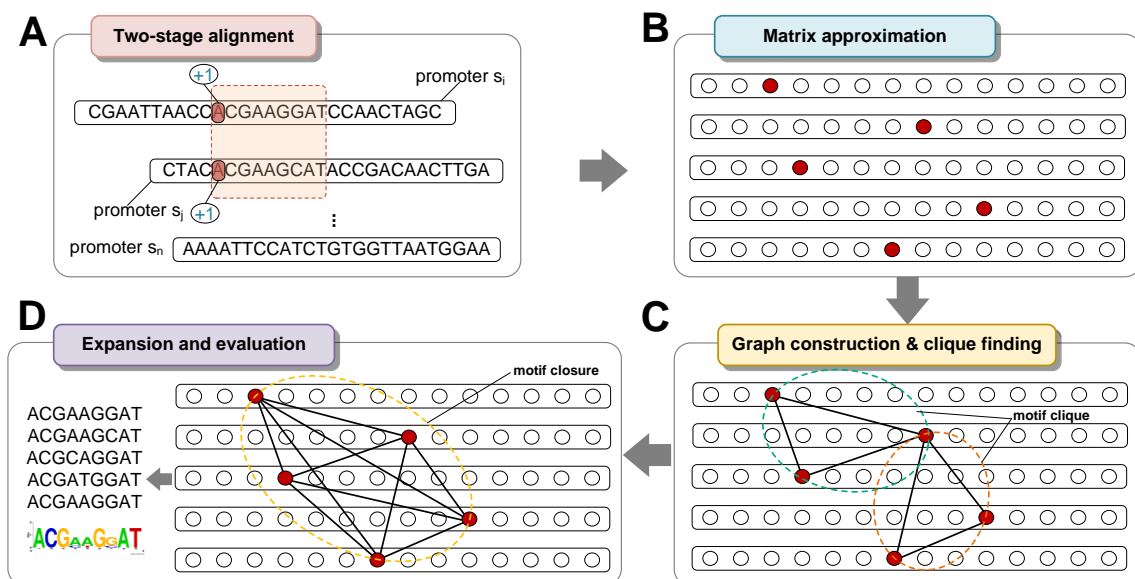


Figure 6. Schematic overview of the BOBRO, including (A) a two-stage alignment, (B) matrix approximation (each red circle represents an identified motif starting position), (C) graph construction and clique finding (each clique corresponds to the core part of a conserved motif pattern), and (D) expansion and evaluation (each motif closure represents an identified motif by refining and expanding corresponding motif cliques).

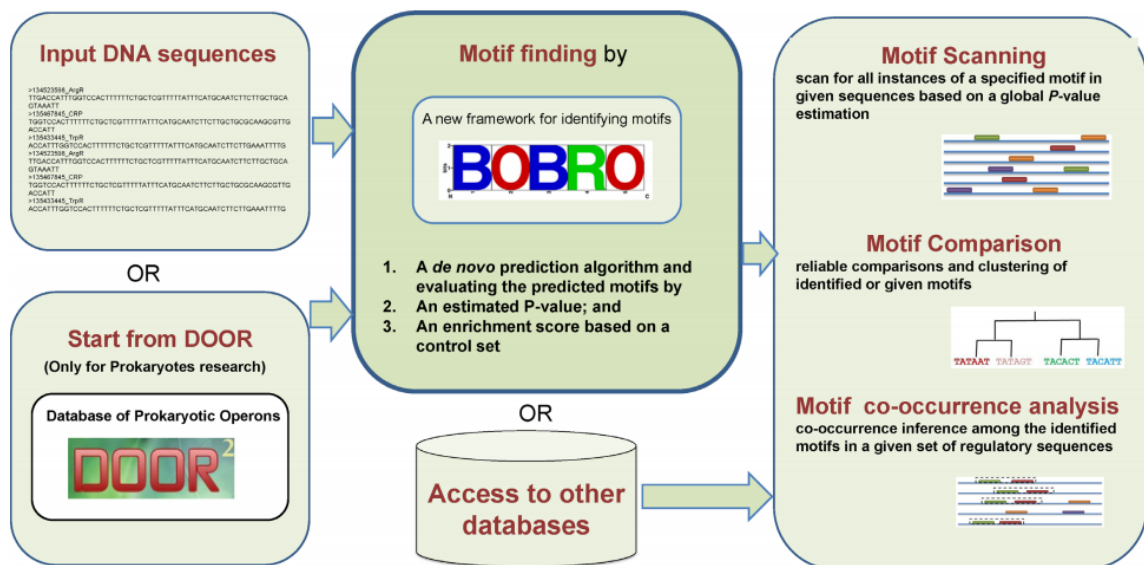


Figure 7. Workflow of DMINDA. Four motif analysis functionalities are accessible by the following clickable buttons on the front page of DMINDA: Motif finding, Motif scanning, Motif comparison and Motif co-occurrence analysis. And 21 motif databases are integrated into Access to other databases.

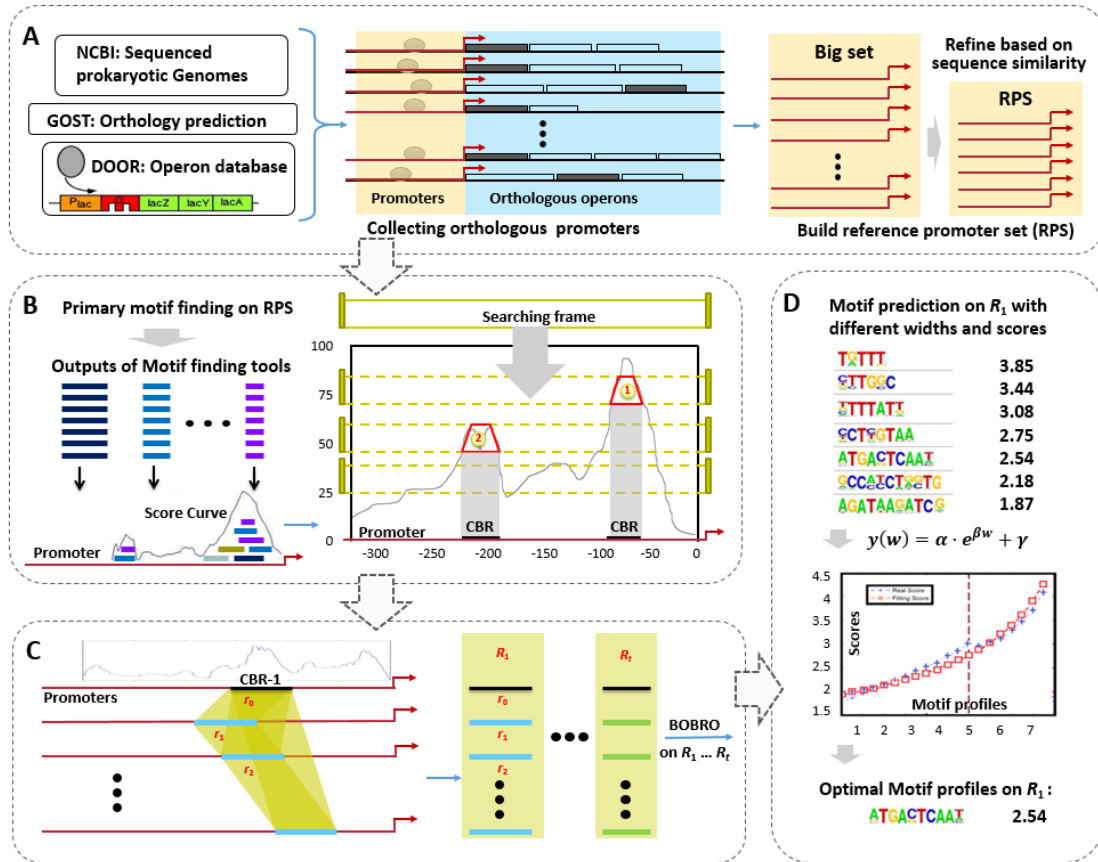


Figure 8. An outline of the MP3 framework. (A) RPS preparation based on the sequenced genome from NCBI, operon information retrieved from DOOR, and identified orthologous genes for a target gene using GOST. The promoters of orthologous operons are generated and then are refined to build RPS. (B) CBR detection by voting strategy and peak finding. The predicted motifs by six tools (short sequences) are mapped back on promoter sequences and generate score curves. The peaks on the curve are identified as CBR by a peak calling method. (C) CBR clustering based on a new graph model. r_0, r_1, \dots are CBRs on promoters, which are clustered together as a related CBR set R_1 . The motif finding will be performed on these clusters (R_1, R_2, \dots, R_t) again to build motif profiles. (D) Motif profiles identification and motif width optimization through curve fitting.

Webservers		DMINDA 2.0	MEME Suite	PATLOC	AIMIE	Melina II	MotifSampler	STAMP
Database	Genome database for motif finding	√			√		√	
	Genome database for motif scanning	√	√	√		√		
	Built-in TFBS database		√					√
Motif finding	<i>De-novo</i> motif finding	√	√		√	√	√	
	Phylogenetic Footprinting Framework	√						
	Gapped-motif finding		√					
	Motif finding based on ChIP-seq		√					
Motif analyses	Motif scanning	√	√	√	√	√		
	Motif comparison	√	√					√
	Motif enrichment analysis	√	√					
	Motif co-occurrence analysis	√						
	Regulon prediction	√						
	Gapped-motif scanning		√					
Visualization	Cytoscape-like network for regulons	√						

Figure 9. Comparison of DMINDA 2.0 and six motif analyses webservers. A check mark indicates that the corresponding functionality is provided by the specific web server.

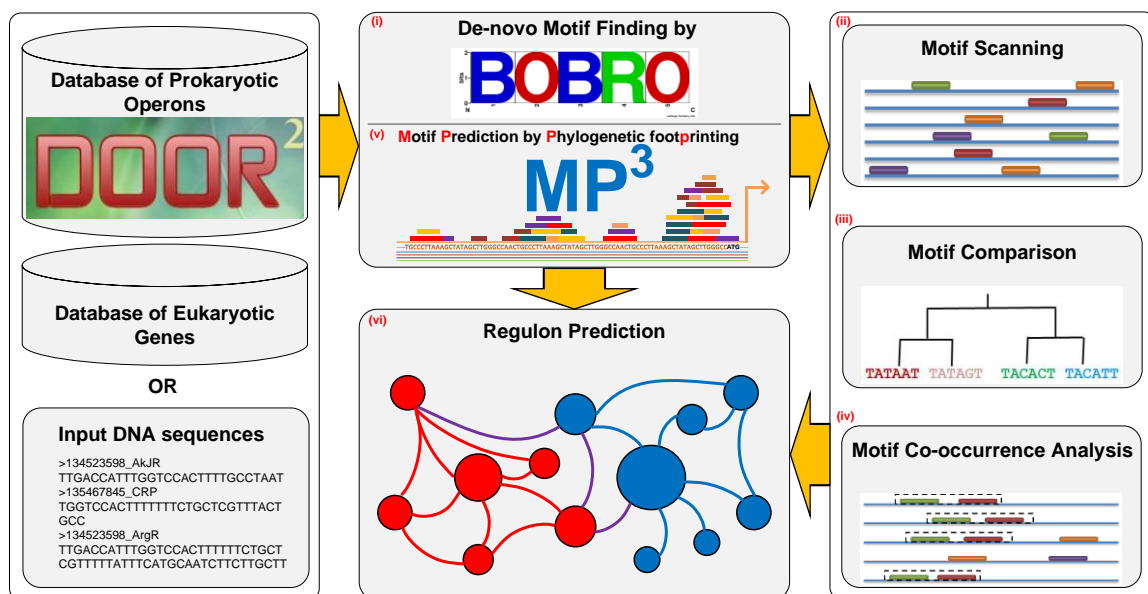


Figure 10. Workflow of DMINDA 2.0, including (i) de-novo motif finding, (ii) motif scanning, (iii) motif comparison, (iv) motif co-occurrence analysis, (v) de-novo motif finding based on phylogenetic footprinting strategy, and (vi) regulon prediction.



All <input type="checkbox"/>	Motif logo	Length <input type="text" value="12"/>	Pvalue <input type="text" value="8.11e-27"/>	Number <input type="text" value="7"/>	Motifs <input type="button" value="ⓘ"/>																																				
<input type="checkbox"/>	 <p>Motif-19</p>	12	8.11e-27	7	<table border="1"> <thead> <tr> <th>Seq</th> <th>Start</th> <th>End</th> <th>Motif</th> <th>Score</th> <th>Info</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>46</td> <td>57</td> <td>CTAGGCAAGAAA</td> <td>21.14</td> <td>1</td> </tr> <tr> <td>2</td> <td>68</td> <td>79</td> <td>CTAGGCAAGAAA</td> <td>21.14</td> <td>2</td> </tr> <tr> <td>11</td> <td>75</td> <td>86</td> <td>CTAGGCAAGAAA</td> <td>21.14</td> <td>11</td> </tr> <tr> <td>16</td> <td>83</td> <td>94</td> <td>CTAGGCAAGAAA</td> <td>21.14</td> <td>16</td> </tr> <tr> <td>7</td> <td>205</td> <td>216</td> <td>CTAGGCATGAAA</td> <td>20.83</td> <td>7</td> </tr> </tbody> </table>	Seq	Start	End	Motif	Score	Info	1	46	57	CTAGGCAAGAAA	21.14	1	2	68	79	CTAGGCAAGAAA	21.14	2	11	75	86	CTAGGCAAGAAA	21.14	11	16	83	94	CTAGGCAAGAAA	21.14	16	7	205	216	CTAGGCATGAAA	20.83	7
Seq	Start	End	Motif	Score	Info																																				
1	46	57	CTAGGCAAGAAA	21.14	1																																				
2	68	79	CTAGGCAAGAAA	21.14	2																																				
11	75	86	CTAGGCAAGAAA	21.14	11																																				
16	83	94	CTAGGCAAGAAA	21.14	16																																				
7	205	216	CTAGGCATGAAA	20.83	7																																				
<input type="checkbox"/>	 <p>Motif-20</p>	12	3.02e-22	7	<table border="1"> <thead> <tr> <th>Seq</th> <th>Start</th> <th>End</th> <th>Motif</th> <th>Score</th> <th>Info</th> </tr> </thead> <tbody> <tr> <td>5</td> <td>200</td> <td>211</td> <td>ATGAAGGTGTCT</td> <td>20.16</td> <td>5</td> </tr> <tr> <td>8</td> <td>222</td> <td>233</td> <td>ATGAAGGTGTCT</td> <td>20.16</td> <td>8</td> </tr> <tr> <td>8</td> <td>238</td> <td>249</td> <td>ATGAAGGTGTCT</td> <td>20.16</td> <td>8</td> </tr> <tr> <td>12</td> <td>47</td> <td>58</td> <td>ATGAAGGTGTCT</td> <td>20.16</td> <td>12</td> </tr> <tr> <td>13</td> <td>264</td> <td>275</td> <td>ATGAAGGTGTCT</td> <td>20.16</td> <td>13</td> </tr> </tbody> </table>	Seq	Start	End	Motif	Score	Info	5	200	211	ATGAAGGTGTCT	20.16	5	8	222	233	ATGAAGGTGTCT	20.16	8	8	238	249	ATGAAGGTGTCT	20.16	8	12	47	58	ATGAAGGTGTCT	20.16	12	13	264	275	ATGAAGGTGTCT	20.16	13
Seq	Start	End	Motif	Score	Info																																				
5	200	211	ATGAAGGTGTCT	20.16	5																																				
8	222	233	ATGAAGGTGTCT	20.16	8																																				
8	238	249	ATGAAGGTGTCT	20.16	8																																				
12	47	58	ATGAAGGTGTCT	20.16	12																																				
13	264	275	ATGAAGGTGTCT	20.16	13																																				

Figure 11. Identified motifs from DMINDA 2.0, including motif logo, motif length, P -value, number of motif instances, and detailed information of motif instances.

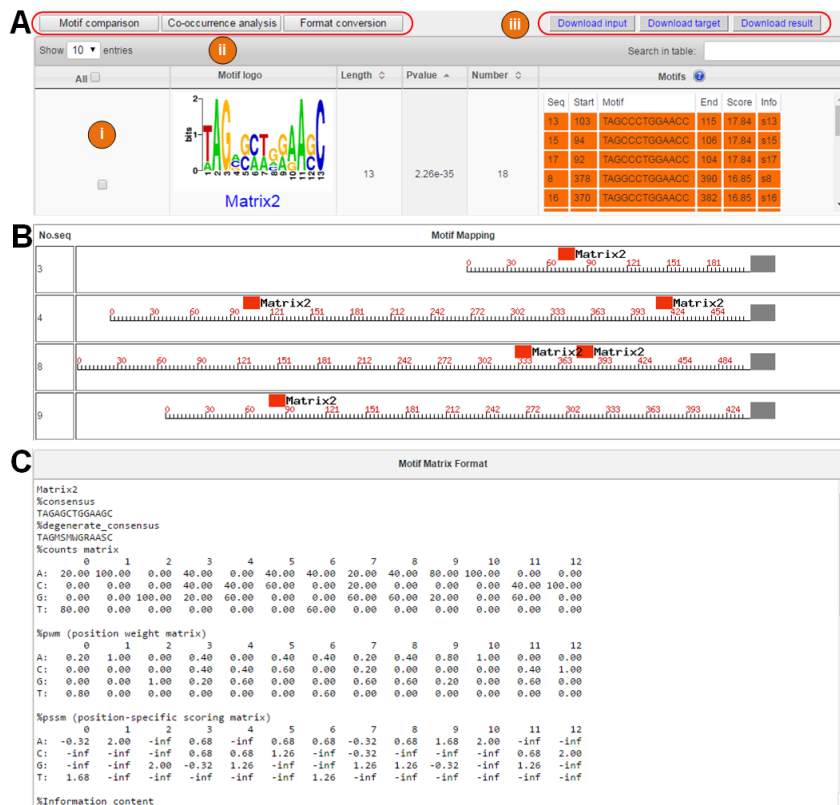


Figure 12. (A) Result page of motif scanning, including (i) query motifs and related sequence details; (ii) three follow-up motif analysis functions; and (iii) options for downloading the submitted motif alignments, query genome sequences and predicted results. (B) The locational mapping of identified motif instances to the corresponding query sequences. (C) The consensus, PWM, position-specific scoring matrix (namely PSSM), IC, and other formats (e.g., MEME and UniPROBE) of the query motifs.

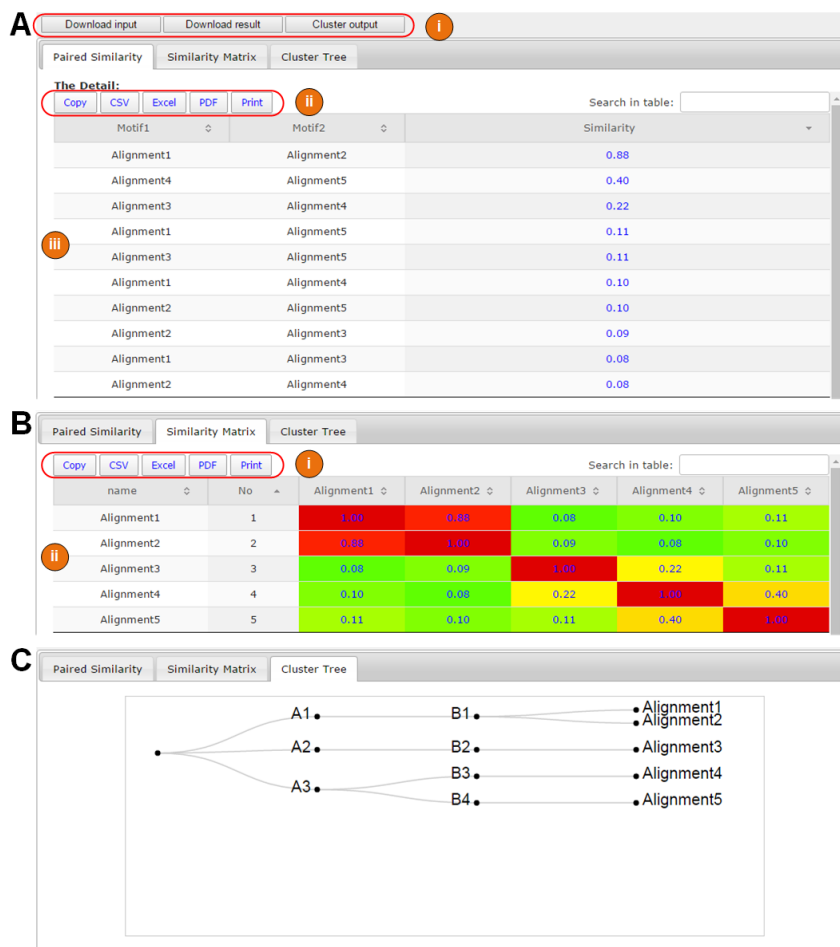


Figure 13. Result page of motif comparison. (A) The paired similarity of query motifs, including (i) options for downloading the submitted motif alignments, similarity matrix, and clustering results; (ii) options for printing, copying, and downloading the paired similarity in multiple formats; and (iii) the paired similarity between submitted motifs. (B) The similarity matrix of query motifs, including (i) options for printing, copying, and downloading the similarity matrix; and (ii) similarity matrix. (C) The clustering tree.



Figure 14. Result page of motif co-occurrence analysis. (A) Identified co-occurring motifs, including (i) options for printing, copying, and downloading the co-occurring motifs in multiple formats; and (ii) *P*-values for each pair of co-occurring motifs. (B) Locational mapping of query motifs to query genome sequences.

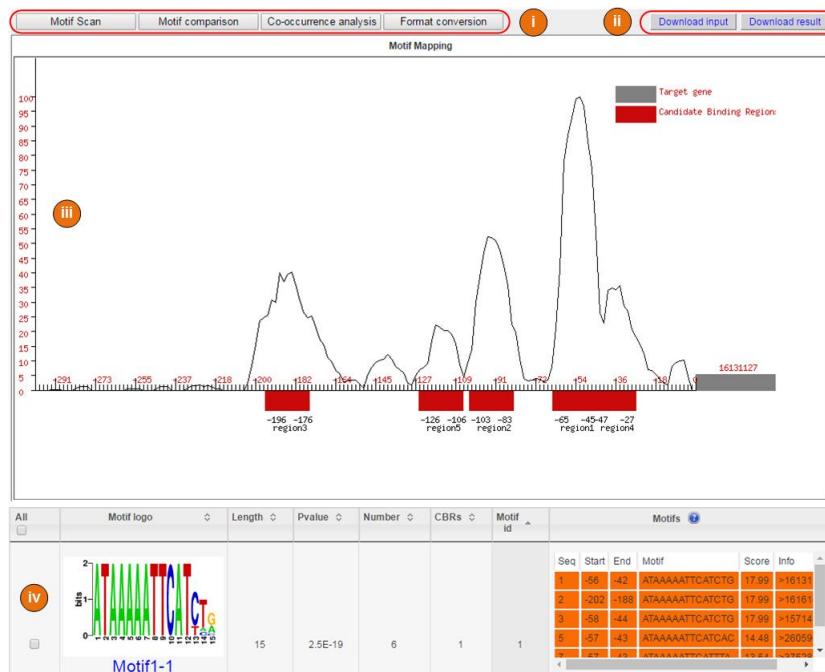


Figure 15. Result page of motif finding by MP3, including (i) four follow-up motif analysis functions; (ii) options for downloading the submitted query sequences and predicted results; (iii) voting score curve and predicted candidate binding regions; and (iv) identified motifs and related sequence details.

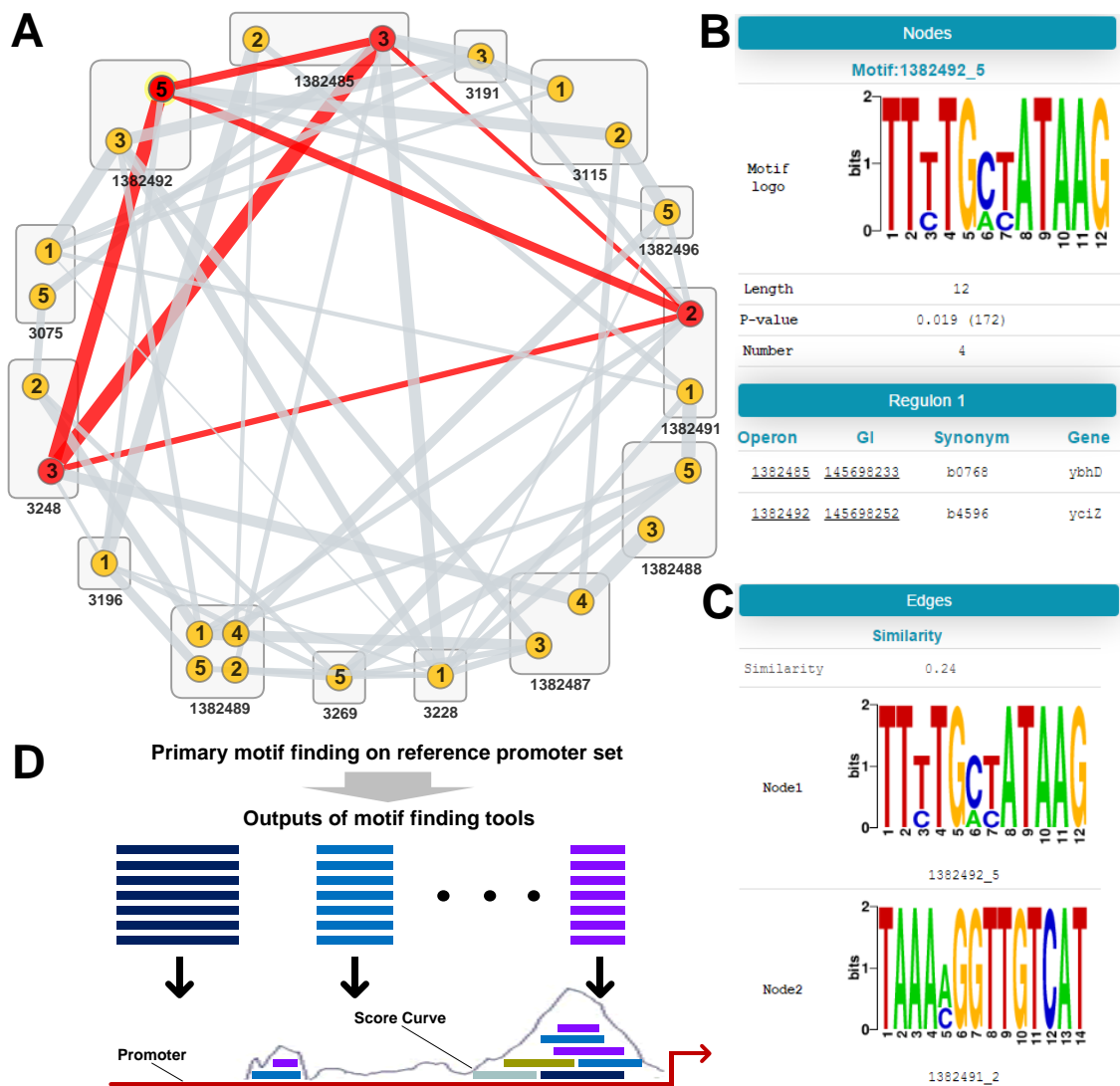


Figure 16. (A) A Cytoscape-like network visualization of predicted regulons. The rounded rectangles indicate operons, orange circles represent identified motifs, and network in red highlights the selected regulon. (B) Details of the selected node (1382492_5) in (A). (C) Details of the selected edge (1382492_5-1382491_2) in (A). (D) The voting strategy in MP³ for generation of reliable TF binding regions.

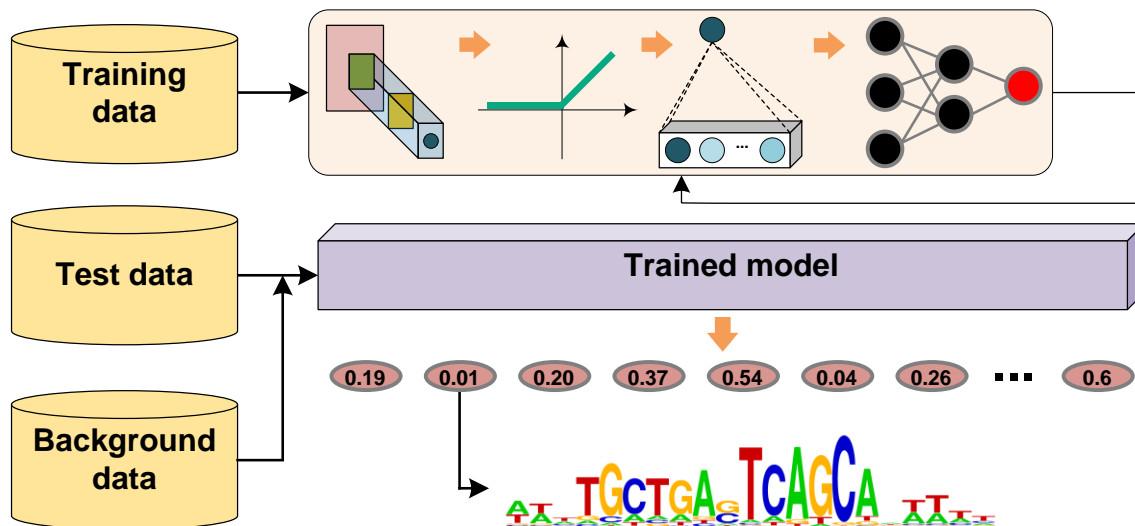


Figure 17. Workflow of DESSO, including a DL model for data training and a statistical model for motif identification.

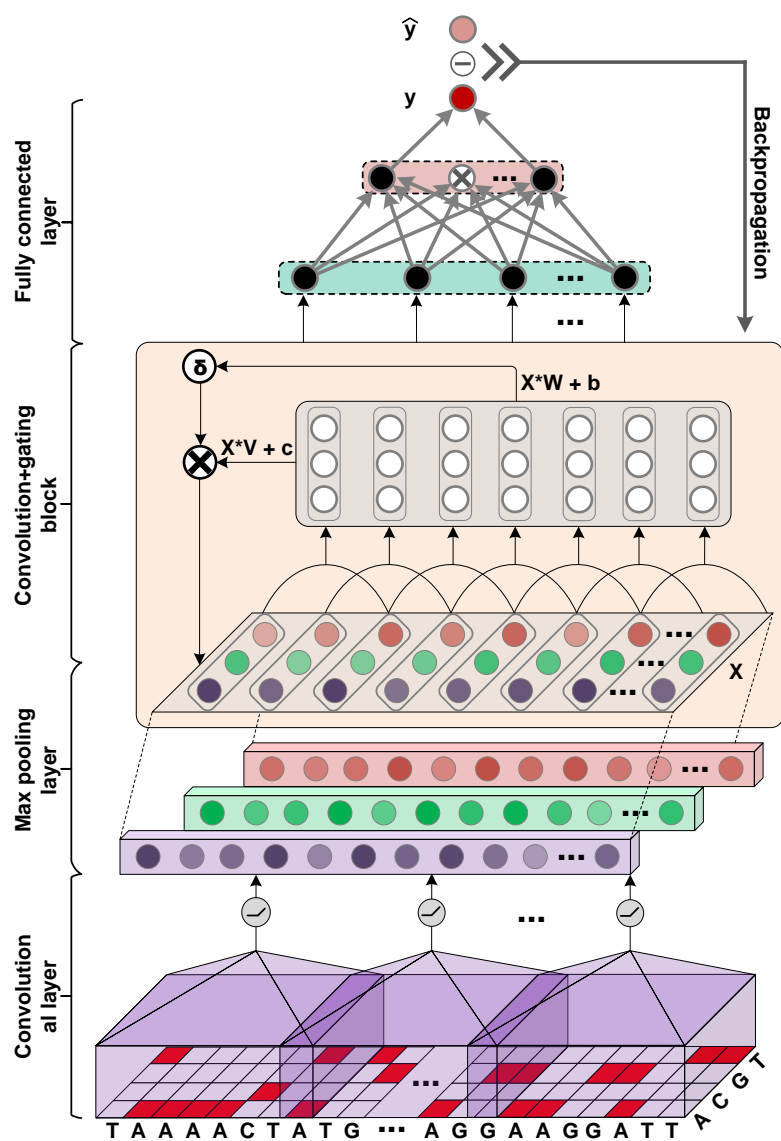


Figure 18. Workflow of GCNN , including a convolutional layer, max pooling layer, CGB, and fully-connected layer.

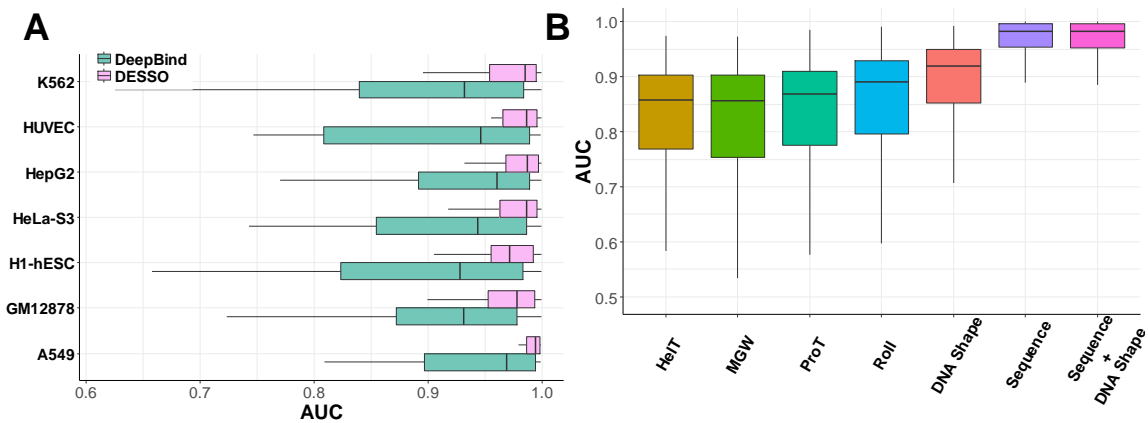


Figure 19. Performance of DESSO. (A) Comparison of DESSO and DeepBind on classification accuracy. (B) Classification accuracy of different inputs.

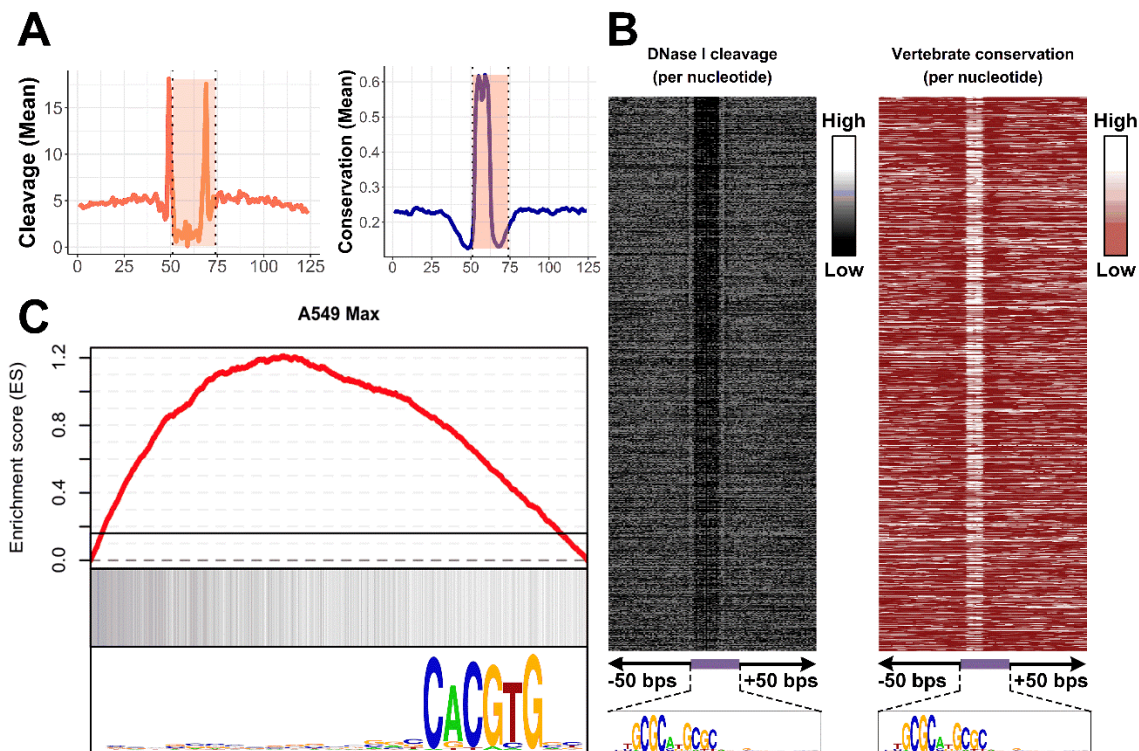


Figure 20. Analysis of identified motifs. (A) Mean value of DNase I cleavage and evolutionary conservation around identified motif instances. (B) Heat map of DNase I cleavage and evolutionary conservation around identified motif instances. (C) Enrichment analysis of identified motifs.

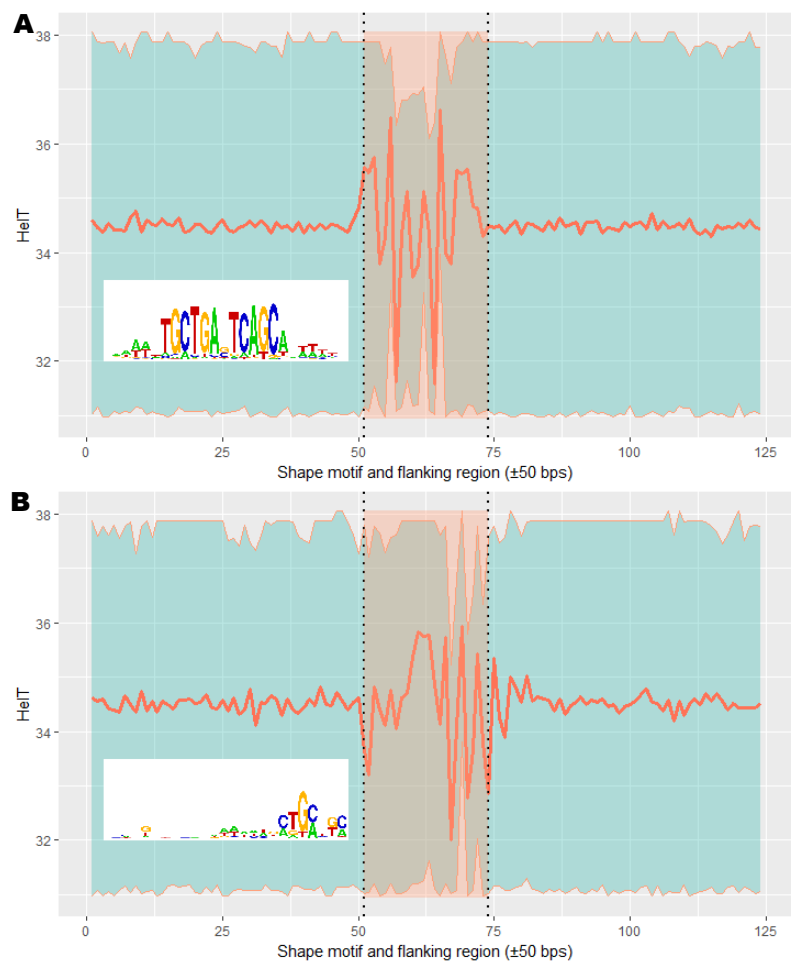


Figure 21. HeIT motif and sequence logo of MAFF . The orange curve represents the mean value of HeIT around shape motif instances (orange shadow), and the motif logo indicates the underlying sequences.

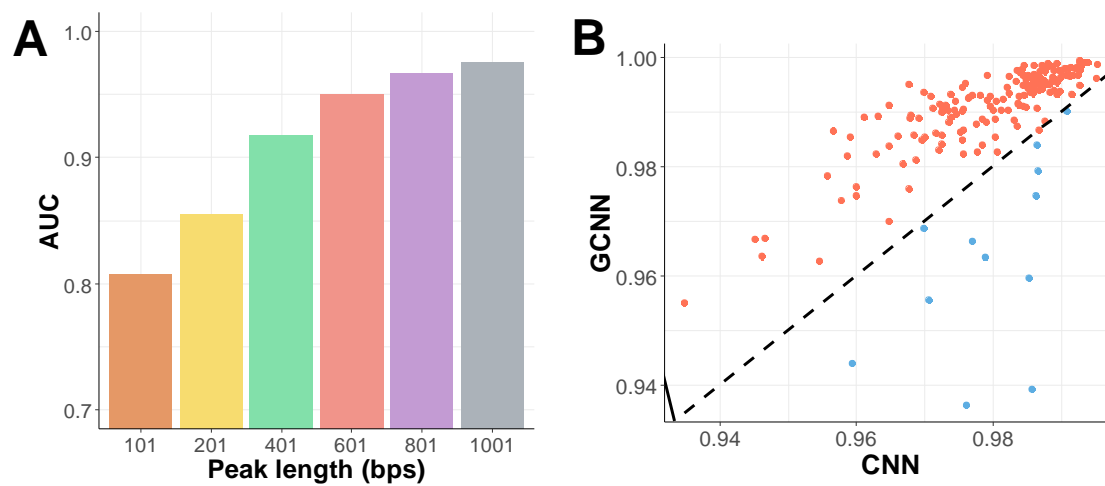


Figure 22. (A) Classification accuracy of CNN with different peak lengths. (B) Comparison of GCNN and CNN on classification accuracy.