Study on detection of bipartite motifs of DNA by considering base interdependencies

（連続２塩基の特性を考慮した Bipartite モチーフの検出手法に関する研究）

Chiba University
Graduate School of Medical and
Pharmaceutical Sciences,
Frontier Medicine and Pharmacy

2019

(Chief Prof. Tetsuichiro Saito)

MOHAMMAD VAHED

Contents

Abstract

It is of great importance to find where transcription factors are bound to, and accurately infer transcription factor binding sites (TFBSs). Some of TFBFs are proposed as bipartite motifs known as two-block motifs separated by gap sequences with variable lengths.

While position weight matrix (PWM) is commonly used for not only representation and prediction of TFBSs, dinucleotide weight matrix (DWM) enables to express the interdependencies of neighboring bases. Incorporating DWM into detection of the bipartite motifs, I have developed a novel tool for *ab initio* motif detection, DIpartite (bi**partite** motif detection tool based on **di**nucleotide weight matrix) using a Gibbs sampling strategy, and minimization of Shannon's entropy. DIpartite predicts the bipartite motifs by taking into account the interdependencies of neighboring positions, i.e., DWM. I performed the comparison of DIpartite by using test datasets, i.e., CRP in *E. coli*, sigma factors in *B. subtilis* and the promoter sequences in human.

I have developed DIpartite for detecting TFBSs, in particular the bipartite motifs. DIpartite enables *ab initio* prediction of the conserved motif based on not only PWM, but also DWM. I evaluated the performance of DIpartite compared with freely available tools, i.e., MEME, BioProspector, BiPad, and AMD. Taken together, DIpartite performs equivalent or better than those in the cases of the bipartite motifs with the fixed and variable gaps like promoter sequences in human and variable gaps. DIpartite requires users to specify the motif lengths, gap length, and PWM or DWM. DIpartite can be found at https://github.com/Mohammad-Vahed/DIpartite.

1. Introduction

Gene expression can be often regulated by transcription factors (TFs). TFs bind to specific DNA-binding sites and modulate the expression of the genes. Therefore, an accurate inference of Transcription Factor Binding Sites (TFBSs) is of fundamental importance for understanding the complex transcriptional regulations. High throughput ChIP-seq widely used to study TF-DNA interactions provides the sequences of binding regions [1, 2]. TFBSs for a specific TF, the binding priority is usually demonstrated as a position weight matrix (PWM). PWM is a common way to pattern transcription factor binding sites. When a PWM made, it can be used to scan sequences for considered binding sites using the PWM to score how good each sequence segment matches the PWM. To model this variation, the PWM has emerged as a likely construct of popular election. The PWM specifies the frequency distribution of nucleotides at each position of the binding sites and is considered to be related to the energy of binding of the transcription factor to the DNA. TFBSs can be determined as the most over-represented motif in a given set of DNA sequences [3, 4]. Discovery of the motifs in the DNA sequence is high practical importance in the study of gene regulation. The motif finding problem is to find a PWM representing binding sites of an unknown transcription factor, ab initio from sequence data. The PWM scores for specific motifs have been discovered to be beneficial as a measure of the motif strength, for example, PWMS for specific connect sites has been useful as a proxy of connecting performance in prokaryotes and eukaryotes.

Bipartite motif is defined as an extension of one-block TFBS, that is, two conserved motifs separated by variable gaps (Figure 1). A couple of the bipartite motifs have been proposed in both prokaryotes and eukaryotes [3, 4]. Shultzaberger et al. [3] have proposed the bipartite model of ribosome binding sites composed of Shine-Dalgarno region and the initiation region in *Escherichia coli* [3]. In *Bacillus subtilis*, principal sigma factor in vegetative growth SigA binds to the bipartite motif separated by the variable gaps, TGACA<spacer>TATAAT [5-7]. Baichoo and Helmann [8] have determined the bipartite motif, TGATAAT<spacer>ATTATCA, of ferric uptake repressor Fur [8, 9]. It has been reported that global regulator AbrB could recognize the bipartite motif [10-12]. As the case of eukaryotes, the bipartite motifs of yeast TFs, e.g., ABF1 and GAL4, are accepted [13, 14]. It has been reported that around 30% of the promoter sequences contain the bipartite motifs with the constant gaps in human [15]. The conservation score for the motif M4 (ACTAYRNNNCCCR) was much higher than those for most known motifs. Similarly, the TFs CAR and RXR bind to the bipartite motifs in human [4]. Thus, it is conceivable that TFs work in a cooperative manner and recognize the bipartite motifs to regulate the gene expressions [16, 17]. A few tools, e.g., BioProspector [18], BiPad [19, 20] and AMD [21], are available for *ab initio* prediction of the bipartite motif for a set of DNA sequences, while many tools have been developed for prediction of the one-block TFBS, e.g., Consensus [22], Gibbs Sampler [23], and MEME [24]. BioProspector based on Gibbs sampling [18] and BiPad based on entropy minimization method [19, 20] enable to identify the bipartite motifs with the variable gaps. AMD identifies the bipartite motifs with the constant gaps by comparing the target sequences with the background sequences regardless of whether the motifs are long or short, gapped or contiguous [21].

One question, if the independence hypothesis is enough, is nearest-neighbor dinucleotides good for TFBSs detect? Probably, the query is made intricate by the result of sequence on DNA structure and bendability, which expects that the DNA-protein contact interactions are not the just factor at play.

Position weight matrix (PWM) is commonly used for finding and representing of TFBSs [25]. PWMs are based on the assumption that each nucleotide independently participates in the TF-DNA interaction. However, it has long been known that interactions between neighboring DNA bases affect the TF-DNA interactions. For example, a single amino acid interacts with multiple bases simultaneously [26]. Zhao et al. [27] clearly show the existence of dinucleotide dependency in TFs [27, 28]. Indeed, PWMs perform well in modeling TFBS properties but are insufficient for considering position interdependencies. The interdependencies exist between neighboring positions of the binding sites of CRP and LexA in *E. coli* [29]. It has been reported that the method based on dinucleotide weight matrix (DWM) outperformed that based on PWM for yeast datasets [30]. In fact, Weirauch et al. [28] observed the improvement of performance of motif detection by incorporating the dinucleotide interactions [28]. Although BioProspector and BiPad predict the bipartite motifs, those are based on the assumption of independencies of each bases, i.e., PWM.

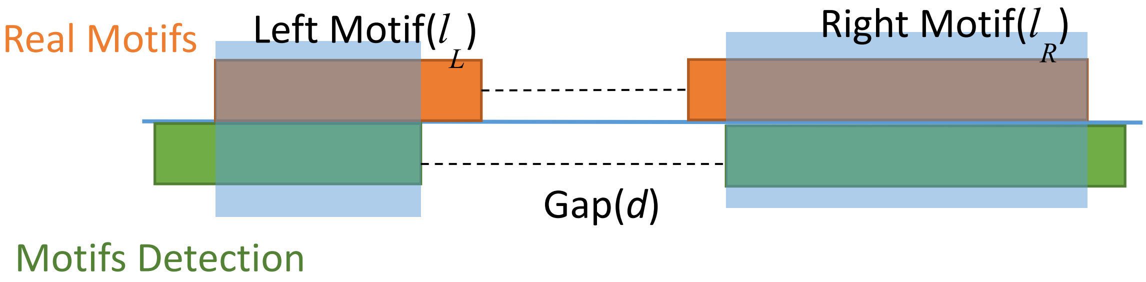


Figure 1. Bipartite patterns on double DNA motifs. A bipartite module is an independent functional unit on the upstream/downstream of a regulated gene and recognized. I assume that the two subunits cooperatively bind to the module with constrained spacers. A bipartite pattern can be expressed as l_L<D>l_R. D is the gap range as defined in the text.

Here I present a novel bipartite motif detection tool, DIpartite (bipartite motif detection tool based on dinucleotide weight matrix). DIpartite predicts the bipartite motif by taking into account interdependencies of neighboring positions, i.e., DWM. I performed the comparison of DIpartite by using test datasets of prokaryote and eukaryote, i.e., CRP in *E. coli*, sigma factors in *B. subtilis* and the promoter motifs in human.

2. Implementation

2.1 A novel method for predicting bipartite motifs by incorporating base-pair dependencies

DIpartite identifies the bipartite motifs with variable gaps based on PWM or DWM from the input sequences (Figure 2). Since it is reported that the bipartite motif represents well by Shannon's entropy [3, 19, 20], I set the objective function to minimize the entropy. Similar to BiPad [19, 20], the algorithm of DIpartite is based on Gibbs sampling and the minimization of information content (IC) by a greedy algorithm. DIpartite adopts the Gibbs sampling strategy which initializes the motif positions for all input sequences at random, and iteratively improves the entropy of PWM or DWM by updating the motif position.

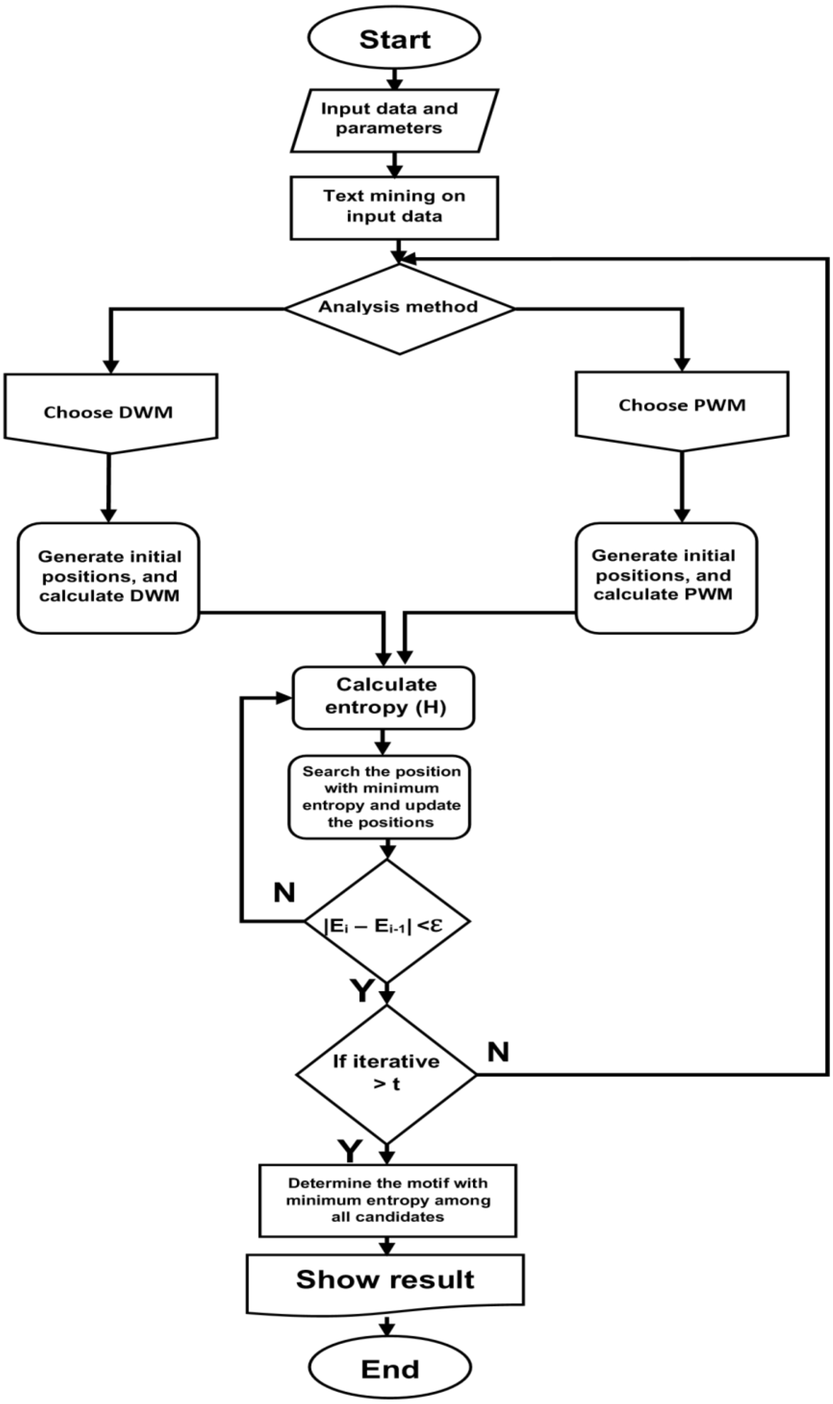

Start
Input data and parameters
Text mining on input data
Analysis method
Choose DWM
Choose PWM
Generate initial positions, and calculate DWM
Generate initial positions, and calculate PWM
Calculate entropy (H)
Search the position with minimum entropy and update the positions
N
|Ei – Ei-1| <ε
Y
If iterative > t
N
Y
Determine the motif with minimum entropy among all candidates
Show result
End

Figure 2. Flowchart of DIpartite. DIpartite proposes the bipartite motif based on PWM or DWM. Each iteration starts from the randomly generated positions. The convergence of each iteration is judged by the differences of the entropy, i.e., ε. I set $\varepsilon = 10^{-8}$. E_i and E_{i-1} correspond to the ith and $i-1$th entropy, respectively.

2.2 Objective function

Input data have *N* sequences for prediction of the bipartite motifs separated by gaps. Similar to BiPad [19, 20], the bipartite motifs are expressed as $l_L<d>l_R$, where l_L and l_R are the widths of left and right motifs, respectively, and *d* is gap length. I set the objective function to minimize Shannon's entropy for PWM or DWM of the concatenated motif of the left and right motifs, in Equation 1:

$$\widehat{M}_{LR} = \mathrm{argmin}_{M_{LR}}\left(IC_{M_{LR}}\right) \ (1)$$

where M_{LR} is the concatenated motif, and $IC_{M_{LR}}$ is the entropy for the motif M_{LR}. So that, $IC_{M_{LR}}$ is given by:

$$IC_{M_{LR}} = \sum_{i}^{j}\sum_{x\in X} -p_i(x) \times \log\left\{\frac{p_i(x)}{b(x)}\right\}, i = \begin{cases}1, \text{PWM} \\ 2, \text{DWM}\end{cases}, X = \begin{cases}\{\text{A,C,G,T}\}, & \text{PWM} \\ \{\text{AA,AC,}\cdots\text{,TT}\}, & \text{DWM}\end{cases} \ (2)$$

where $p_i(x)$ and $b(x)$ are the composition of x in the motif sites and the background sites (not motif sites), respectively. x is one of the mononucleotides, or dinucleotides for PWM, or DWM, respectively. j is the sum of the lengths of the left and right motifs. $p_i(x)$ and $b(x)$ are given by:

$$p_i(x) = \frac{f_i(x) + \beta/k}{N + \beta}, k = \begin{cases}4, & \text{PWM} \\ 16, & \text{DWM}\end{cases} \ (3)$$

$$b(x) = \frac{g(x) + \beta/k}{n + \beta} \ (4)$$

where *N* is a total of input sequences. $f_i(x)$ is the frequency of x at the position i, i.e., the mononucleotide at position i for PWM, or the dinucleotide at position $i-1, i$ for DWM. k is the number of the patterns, i.e., $k = 4$ for PWM or $k = 16$ for DWM. *n* is a total of the mononucleotides for PWM or the dinucleotides of the background for DWM, which do not locate the motif sites. *β* is the total pseudo-count. $g(x)$ is the frequency of *x* in the background sites. I set *β*=1.

2.3 Overview of the algorithm

The algorithm of DIpartite works through an iterative process of calculating entropy. DIpartite was implemented in C++. Fasta and text formats are allowed as input files. Users can specify the lengths of the left and right motifs, the gap length, and PWM for the mononucleotide or DWM for the dinucleotide. The software works for OOPS (one occurrence per sequence) or ZOOPS (Zero or one bipartite occurrence per sequence).

2.4 Expectation-Maximization expressions for the ZOOPS model

The EM algorithm is used in the context of motif discovery as follows: The initial values for the motif model (that is, the initial PWM parameters) are estimated. In the OOPS and ZOOPS models, this estimation is often carried out by choosing a motif start point at random for each input sequence and then counting the numbers of each nucleotide at each motif position, creating a consensus model from these start points. The ZOOPS model supposes that each input sequence either includes identically one appearance of the motif or no appearance of the motif. The ZOOPS model description for this by introducing an additional index variable which indicates whether a specific input sequence includes a motif appearance or not [36-38] .The EM Q function is the expected value of the complete data (that is, {X, Z}) log-likelihood function. The EM algorithm depend on the Q function: the E-step of the algorithm requires calculating the parameters of the Q function. The novel index variable *Qi* is specific as *Qi* $=\sum_{j=1}^{M}$ Zi, j. That is, *Qi* = 1 if sequence i includes a motif appearance and 0 otherwise. The OOPS model then becomes a specific instance of the ZOOPS model where all input sequences include a motif appearance. If sequence i includes a motif appearance, the conditional probability of i given the hidden variables is the same as in the OOPS model:

$$p(X_i|Z_{i,j} = 1, \theta) \triangleq \prod_{l\in\Delta_{i,j}} \prod_{k\in L} \theta_{0,k}^{I(X_{i,l}=k)} \quad \prod_{m=1}^{W} \prod_{k\in L} \theta_{0,k}^{I(X_{i,j+m-1}=k)}. \quad (5)$$

As in the OOPS model, the equation (6) is the outcome of probabilities over the W positions in the motif and the remaining background positions. The conditional probability for a sequence which does not include a motif, the incidence is as well as specific as the product of probabilities, this time using background probabilities for all positions within sequence i:

$$p(X_i|Q_i = 0, \theta) \triangleq \prod_{l=1}^{L_i} \prod_{k\in L} \theta_{0,k}^{I(X_{i,l}=k)} \text{ , } \quad (6)$$

the *Li* is the length of the input sequence *i*. As in the OOPS model, the same distribution of beginning sites within a sequence is supposed. If the previous probability of a sequence including a motif appearance is specific as γ, it follows from the hypothesis of equal input sequence length that the previous probability of any location being a motif start point is:

$$\lambda \triangleq p\big(Z_{i,j} = 1|\theta\big) = \frac{\gamma}{M}. \quad (7)$$

For ease, the model parameters are now collected and define as φ = (θ, γ). It is noted that the model parameters now contain the previous probability of a sequence including a motif appearance, also to the motif and background models from the OOPS model. The full data common probability can be written as:

$$p(\mathrm{X},\mathrm{Z}|\phi) \triangleq \prod_{i=1}^{\mathrm{N}} p(X_i, Z_i|\phi) \quad (8)$$

$$p(\mathrm{X},\mathrm{Z}|\phi) \triangleq \prod_{i=1}^{\mathrm{N}} p(X_i, Z_i|\phi)\; p(Z_i|\phi)$$

$$p(\mathrm{X},\mathrm{Z}|\phi) \triangleq \left[\left(\prod_{i=1}^{\mathrm{M}} p\big(X_i, Z_{i,j} = 1, \theta\big)^{Z_{i,j}}\right) \times\; p(X_i|Q_i = 0, \theta)^{(1-Q_i)} \;\times\; \lambda^{Q_i} \;\times\; (1-\gamma)^{(1-Q_i)}\right]. \quad (9)$$

As in the OOPS model, (9) takes benefit of the fact that all $Z_{i,j}$ in a sequence will be 0 apart from one. The first part in (9) is the expression for a sequence, including a motif appearance (6). The second part is the expression for a sequence without a motif appearance (7). Just one of these parts will be applied, depending on the amount of *Q* for sequence i. If $Q_i = 0$, then all $Z_{i,\,j}$ will be 0, canceling the first part and applying the second. If $Q_i = 1$, then $Z_{i,\,j} = 1$ for some j and the first part is applied while the second part is canceled. This canceling tasks similarly for the previous parts. Note that the previous term for locations in a sequence was 1/M before as any sequence had a motif appearance. Now the just γ of sequences include a motif, the previous on locations within a sequence is γ/M = λ (and the previous part for a sequence not including a motif is 1−γ). The log-likelihood function for all data can be written:

$$p(\mathrm{X},\mathrm{Z}|\phi) \triangleq \sum_{i=0}^{N}\left(\sum_{j=1}^{M} Z_{i,j} \ln p\big(X_i|Z_{i,j}=1,\theta\big)\right)$$

$$+\sum_{i=1}^{N}(1-Qi)\ln p(X_i|Q_i=0,\theta)$$

$$+\sum_{i=1}^{N} Qi\ \ln\lambda + \sum_{i=1}^{N}(1-Qi)\ln(1-\ \gamma) \qquad (10)$$

the $Z_{i,j}^{(t)}$ is specific as the predicted probability of a motif beginning point at location j in sequence i:

$$Z_{i,j}^{(t)}\ \triangleq\ \mathrm{E}_{\mathrm{Z}|\mathrm{X},\phi^{(t)}}\ [Z_{i,j}]$$

$$= 1\,.p\big(\,Z_{i,j}=1\,\big|X_i,\phi^{(t)}\,\big) + 0\,.p\big(\,Z_{i,j}=0\,\big|X_i,\phi^{(t)}\,\big)$$

$$=\ p\big(\,Z_{i,j}=1\,\big|X_i,\phi^{(t)}\,\big)\,, \qquad (11)$$

as *Qi* is related on $Z_{i,j}$, $Q_i^{(t)}$ is specific as the expected probability of sequence *i* including a motif appearance (this decreases to a sum of the appropriate $Z_{i,j}^{(t)}$values) :

$$Q_i^{(t)}\ \triangleq\ \mathbb{E}_{\mathrm{Z}|\mathrm{X},\phi^{(t)}}\ [Q_i]$$

$$Q_i^{(t)} = \sum_{j=1}^{M}\mathbb{E}_{\mathrm{Z}|\mathrm{X},\phi^{(t)}}\ \ [Z_{i,j}]$$

$$Q_i^{(t)} = \sum_{j=1}^{M} Z_{i,j}^{(t)}\ . \qquad (12)$$

The *Q* function is the predicted amount of the log-likelihood function (10), concerning the conditional distribution of Z given X below the current evaluation of parameters $\theta^{(t)}$:

$$Q\ (\phi\ |\phi^{(t)} =\ \mathbb{E}_{\mathrm{Z}|\mathrm{X},\phi^{(t)}}\ [\ln p\,(X,Z\mid\phi)]$$

$$= \mathbb{E}_{\mathrm{Z|X,\phi}^{(t)}} \left[\sum_{i=0}^{N} \left(\sum_{j=1}^{M} Z_{i,j} \ln p\left(X_i | Z_{i,j} = 1, \theta\right) \right) + \right]$$

$$+ \sum_{i=1}^{N} (1 - Qi) \ln p(X_i | Q_i = 0, \theta)$$

$$+ \sum_{i=1}^{N} Qi \ \ln \lambda + \sum_{i=1}^{N} (1 - Qi) \ln (1 - \gamma)$$

$$= \sum_{i=0}^{N} \left(\sum_{j=1}^{M} \mathbb{E}_{\mathrm{Z|X,\phi}^{(t)}} \left[Z_{i,j}\right] \ln p\left(X_i | Z_{i,j} = 1, \theta\right) \right)$$

$$+ \sum_{i=1}^{N} (1 - \mathbb{E}_{\mathrm{Z|X,\phi}^{(t)}} [Q_i]) \ln p(X_i | Q_i = 0, \theta)$$

$$+ \sum_{i=1}^{N} \mathbb{E}_{\mathrm{Z|X,\phi}^{(t)}} [Q_i] \ln \lambda$$

$$+ \sum_{i=1}^{N} (1 - \mathbb{E}_{\mathrm{Z|X,\phi}^{(t)}} [Q_i]) \ln(1 - \mathrm{\gamma})$$

$$= \sum_{i=0}^{N} \left(\sum_{j=1}^{M} Z_{i,j}^{(t)} \ln p\left(X_i | Z_{i,j} = 1, \theta\right) \right)$$

$$+ \sum_{i=1}^{N} (1 - Q_i^{(t)}) \ln \ p(X_i | Q_i = 0, \theta)$$

$$+ \sum_{i=1}^{N} Q_i^{(t)} \ \ln \lambda + \ \sum_{i=1}^{N} \left(1 - Q_i^{(t)}\right) \ln (1 - \mathrm{\gamma}) \qquad (13)$$

where (11) and (12) have been substituted as necessary. This is equal to the phrase given by Bailey and Elkan and by Keles, et al (36-42).

2.5 E-step

As in the OOPS model, the E-step need the assessment of the probability of the hidden data $p(\mathrm{Z}|\mathrm{X},\theta)$, that is, $Z_{i,j}^{(t)}$ for each position. Then, Bayes' theorem is applied to specify $Z_{i,j}^{(t)}$ in parts of (6) and (7):

$$Z_{i,j}^{(t)} = \frac{p\left(X_i|Z_{i,j} = 1, \theta^{(t)}\right) \lambda^{(t)}}{p(X_i|Q_i = 0, \theta^{(t)})\,(1 - \gamma^{(t)}) + \sum_{j=1}^{M} p\left(X_i|Z_{i,j} = 1, \theta^{(t)}\right) \lambda^{(t)}} \qquad (14)$$

For all $i \in \{1, \dots, N\}$ and $j \in \{1, \dots, M\}$.

2.6 Generalizing ZOOPS expressions for Expectation-Maximization

The generalization of the ZOOPS sequence model explanation applied in deterministic EM for motif discovery that removes the need that input sequences should be of equal size. In particular, removing the limitation of equal input sequence size is basic in successfully implementing the cut heuristic that allows detection of multiple appearances of a motif within a single input sequence, a method that fulfills a similar task as the TCM model in MEME. Removing the assumption that completely input sequences are the identical size increases flexibility at the expense of several increases mathematics; however, removing this hypothesis does not fundamentally alter the computation necessary in the E-step for the ZOOPS structure.

The phrase for the conditional probability of a sequence with and without a motif incidence (6 and 7) stay similar as in the ZOOPS model. The γ is certain as the previous probability of a sequence including a motif appearance. The previous determination for the previous probability of a location being a motif start location (λ) becomes problematic in the total setting expand here. In the ungeneralized ZOOPS model, λ could be used as a mathematical convenience as a previous for whole sequences; now assuming that input sequences need not have equal size means that the previous model on any sequence will be different and a single previous is unfit. The easy solution is to replacement $L_i - W + 1 = M$, so:

$$\lambda = \frac{\gamma}{L_i - W + 1} \qquad (15)$$

The dependence of the previous step on the size of the input sequence *Li* is now clear, that if Li is similar for each input sequence and *M* is set to *Li −W + 1,* this is equal to the previous determination. The determination of *Qi* is improved in order to account for the possibility of different sequence size: $Q_i \triangleq \sum_{j=1}^{L_i-W+1} Z_{i,j}$. As pervious step, *Qi* = 1 if sequence *i* includes a motif appearance and 0 differently. Following the novel determination of λ, the generalized phase for the all data joint possibility becomes:

$$p(\mathrm{X},\mathrm{Z}|\phi) \triangleq \prod_{i=1}^{\mathrm{N}} p(X_i, Z_i|\phi)$$

$$p(\mathrm{X},\mathrm{Z}|\phi) \triangleq \prod_{i=1}^{\mathrm{N}} p(X_i, Z_i|\phi)\; p(Z_i|\phi)$$

$$p(\mathrm{X},\mathrm{Z}|\phi) \triangleq \left[\left(\prod_{i=1}^{\mathrm{M}} p\big(X_i, Z_{i,j} = 1, \theta\big)^{Z_{i,j}}\right) \times\ p(X_i|Q_i = 0, \theta)^{(1-Q_i)}\ \times\ (\frac{\gamma}{L_i - W + 1})^{Q_i}\right.$$

$$\left. \times\ (1-\gamma)^{(1-Q_i)}\right]. \qquad (16)$$

The log-likelihood function for all data is, so:

$$ln\, p(\mathrm{X},\mathrm{Z}|\phi) \triangleq \sum_{i=0}^{N}\left(\sum_{j=1}^{L_i-W+1} Z_{i,j} \ln p\big(X_i|Z_{i,j} = 1, \theta\big)\right)$$

$$+\sum_{i=1}^{N}(1-Qi)\ln p(X_i|Q_i = 0, \theta)$$

$$+\sum_{i=1}^{N} Qi\ \ln(\frac{\gamma}{L_i - W + 1}) +$$

$$+\sum_{i=1}^{N}(1-Qi)\ln(\,1-\ \gamma\,) \qquad (17)$$

As the determination of $Z_{i,j}^{(t)}$ remains the same as pervious step, the determination of $Q_i^{(t)}$ is updated:

$$Q_i^{(t)} \triangleq \mathbb{E}_{\mathrm{Z|X,\phi^{(t)}}} [Q_i]$$

$$Q_i^{(t)} = \sum_{j=1}^{L_i-W+1} \mathbb{E}_{\mathrm{Z|X,\phi^{(t)}}} [Z_{i,j}]$$

$$Q_i^{(t)} = \sum_{j=1}^{L_i-W+1} Z_{i,j}^{(t)} . \qquad (18)$$

Finally, the *Q* function is generalized, using the updated determination above:

$$Q\ (\phi\ |\phi^{(t)} = \mathbb{E}_{\mathrm{Z|X,\phi^{(t)}}} [\ln p\,(X, Z \mid \phi)]$$

$$= \mathbb{E}_{\mathrm{Z|X,\phi^{(t)}}} \left[\sum_{i=0}^{N} \left(\sum_{j=1}^{L_i-W+1} Z_{i,j} \ln p(X_i|Z_{i,j} = 1, \theta) \right) + \right]$$

$$+ \sum_{i=1}^{N} (1 - Qi) \ln p(X_i|Q_i = 0, \theta)$$

$$+ \sum_{i=1}^{N} Qi\ \ln(\frac{\gamma}{L_i - W + 1}) + \sum_{i=1}^{N} (1 - Qi) \ln (\ 1 - \gamma\ )$$

$$= \sum_{i=0}^{N} \left(\sum_{j=1}^{L_i-W+1} \mathbb{E}_{\mathrm{Z|X,\phi^{(t)}}} [Z_{i,j}] \ln p(X_i|Z_{i,j} = 1, \theta) \right)$$

$$+ \sum_{i=1}^{N} (1 - \mathbb{E}_{\mathrm{Z|X,\phi^{(t)}}} [Q_i]) \ln p(X_i|Q_i = 0, \theta)$$

$$+ \sum_{i=1}^{N} \mathbb{E}_{\mathrm{Z|X,\phi^{(t)}}} [Q_i] \ln(\frac{\gamma}{L_i - W + 1})$$

$$+ \sum_{i=1}^{N} (1 - \mathbb{E}_{\mathrm{Z|X,\phi^{(t)}}} [Q_i]) \ln(1 - \gamma)$$

$$= \sum_{i=0}^{N} \left(\sum_{j=1}^{L_i-W+1} Z_{i,j}^{(t)} \ln p(X_i|Z_{i,j} = 1, \theta) \right)$$

$$+\sum_{i=1}^{N}(1-Q_i^{(t)})\ln\ p(X_i|Q_i=0,\theta)$$

$$+\sum_{i=1}^{N}Q_i^{(t)}\ \ln(\frac{\gamma}{L_i-W+1})+\ \sum_{i=1}^{N}\left(1-Q_i^{(t)}\right)\ln\left(1-\ \gamma\right) \qquad (19)$$

2.7 Generalized E-Step

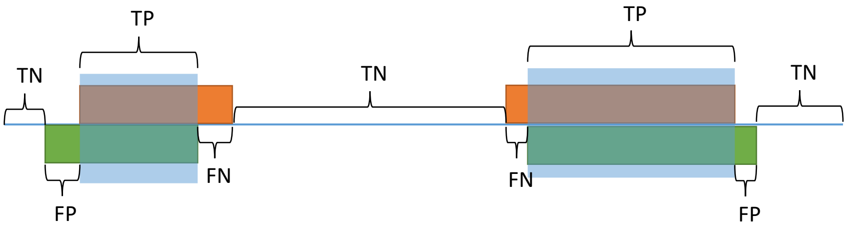
The new determination of λ is applied in the generalization of the E-step. The probability of the hidden data *p* (Z|X, θ) is evaluated for any location:

$$Z_{i,j}^{(t)}=\frac{p\left(X_i|Z_{i,j}=1,\theta^{(t)}\right)(\frac{\gamma}{L_i-W+1})}{p(X_i|Q_i=0,\theta^{(t)})\ (1-\ \gamma^{(t)})+\ \sum_{j=1}^{M}p\left(X_i|Z_{i,j}=1,\theta^{(t)}\right)(\frac{\gamma}{L_i-W+1})} \qquad (20)$$

2.8 Performance evaluation

The nucleotide-level correlation coefficient (*nCC*) was used to evaluate the performance of each tools for the same input data [31] (Figure 3). *nCC* is given by:

$$nCC = \frac{nTP \times nTN - nFN \times nFP}{\sqrt{(nTP + nFN)(nTN + nFP)(nTP + nFP)(nTN + nFN)}} \tag{21}$$



🟧 Real module Motif
🟩 Predicted by DIpartite

TP is the number of nucleotides in a sequence that are correctly predicted by a program as belonging to a module
TN is the number of nucleotides correctly identified as background
FN is the number of true module nucleotides incorrectly classified as background
FP is the number of background nucleotides incorrectly classified as belonging to a module.

Figure 3. The similarity score of nucleotide-level correction coefficient (nCC) for motifs measurements of prediction accuracy.

where *nTP* is the number of nucleotide positions in both known sites and predicted sites, *nFN* is the number of nucleotide positions in known sites but not in predicted sites, *nFP* is the number of nucleotide positions not in known sites but in predicted sites, and *nTN* is the number of nucleotide positions in neither known sites nor predicted sites. I adopted the combined *nCC* by adding *nTP*, *nFN*, *nFP*, and *nTN* over the data sets.

3. Materials and Methods

3.1 CRP

CRP binding sites in *E. coli* were retrieved from Regulon DB as “TF binding sites” (Release: 9.4 Date: 05-08-2017) [32]. For example, the motif sequences of two ECK125158203 entries were identical although the transcription unit was different, i.e., fumA and fumAC. Out of 374 sequences of CRP binding sites, 323 unique sequences ranging from 36 bp to 42bp were filtered and used for the performance comparison. The binding site lengths consisted of 16 bp (11 binding sites), 17 bp (one binding site), 20 bp (one binding site), 22 bp (308 binding sites), and 23 bp (two binding sites).

3.2 Promoter motifs in human

Xie et al. [15] proposed the 1,460 motifs in human. I sought the motifs with the gap lengths greater than or equal to the lengths of left and right motifs. Among of them, I selected 46 motifs with more than 4-nt gaps as the test datasets of two-block motifs. The promoter sequences around the positions of each motifs (500 bp upstream to 500 bp downstream) were retrieved as the target sets.

3.3 Sigma factor

As the dataset of bipartite motifs with variable gap lengths, the sigma factor dataset in *B. subtilis* from DBTBS [7] was used. The nine of the bipartite sigma transcription factors in *B. subtilis* were used. The minimum and maximum gap lengths of sigma factors were determined based on all identified binding sites: σ^A (344 sequences ranging from 38 bp to 93 bp, 6<[11,23]>6), σ^B (64 sequences ranging from 39 bp to 64 bp, 6<[12,18]>6), σ^D (30 sequences ranging from 44 bp to 57 bp, 4<[12,18]>8), σ^E (70 sequences ranging from 41 bp to 58 bp, 7<[12,18]>8), σ^F (25 sequences ranging from 41 bp to 71 bp, 5<[13,19]>10), σ^G (55 sequences ranging from 40 bp to 76 bp, 5<[15,20]>7), σ^H (25 sequences ranging from 41 bp to 60 bp, 7<[9,18]>5), σ^K (53 sequences ranging from 38 bp to 85 bp, 4<[9,17]>9), and σ^W (34 sequences ranging from 38 bp to 53 bp, 10<[13,17]>6).

3.4 Other programs used for comparison

Four popular tools, namely MEME (ver. 5.0.3), BioProspector (release 2), AMD, and BiPad (ver. 2), were compared with DIpartite.

For the CRP dataset, MEME was executed with the options “-mod oops”, “-dna”, “-w 22”, “-minw 22”, and “-maxw 22”. BioProspector was executed with the options “-n 50”, and “-n 3”. AMD was executed with the options “-MI” and “-T 1”. BiPad was executed with the options “-l 22”, “-r 0”, “-a 0”, “-b 0”, “-i”, and “-y 500”. AMD was executed with the option “-T 2” for two sigma datasets, i.e., σ^E and σ^F. I used the background sequences for AMD: the 200 bp upstream regions of 4,314 genes in *E. coli* K-12 (NC_000913.3); the promoter sequences of all human genes (hg17: upstream1000.fa.gz); the 200 bp upstream regions of 4,448 genes in *B. subtilis* 168 (NC_000964.3).

4. Results

4.1 Interdependencies of neighboring DNA bases in CRP

CRP is one of the seven main transcription factors that influences transcriptional networks in *E. coli* [33]. It has been shown that there are interdependencies among neighboring DNA bases in CRP binding sites [29]. More than 300 binding sites for CRP have been registered in Regulon DB as “TF binding sites” (Release: 9.4) [32]. The CRP binding sites are separated by a 6-nt gap (Figure 4A). I measured the interdependency of CRP using the mutual information proposed by Salama and Stekel [29]. Strong correlations between neighboring bases were observed, for example, among positions 1, 2, and 6–8, and among positions 16–19 (Figure 4B). In addition, I observed the higher mutual information between the distant positions in 7, 16 and 8, 17 among the palindromic positions, followed by the position in 6 and 19. This suggests that the palindromic features of CRP binding sites would be incomplete.

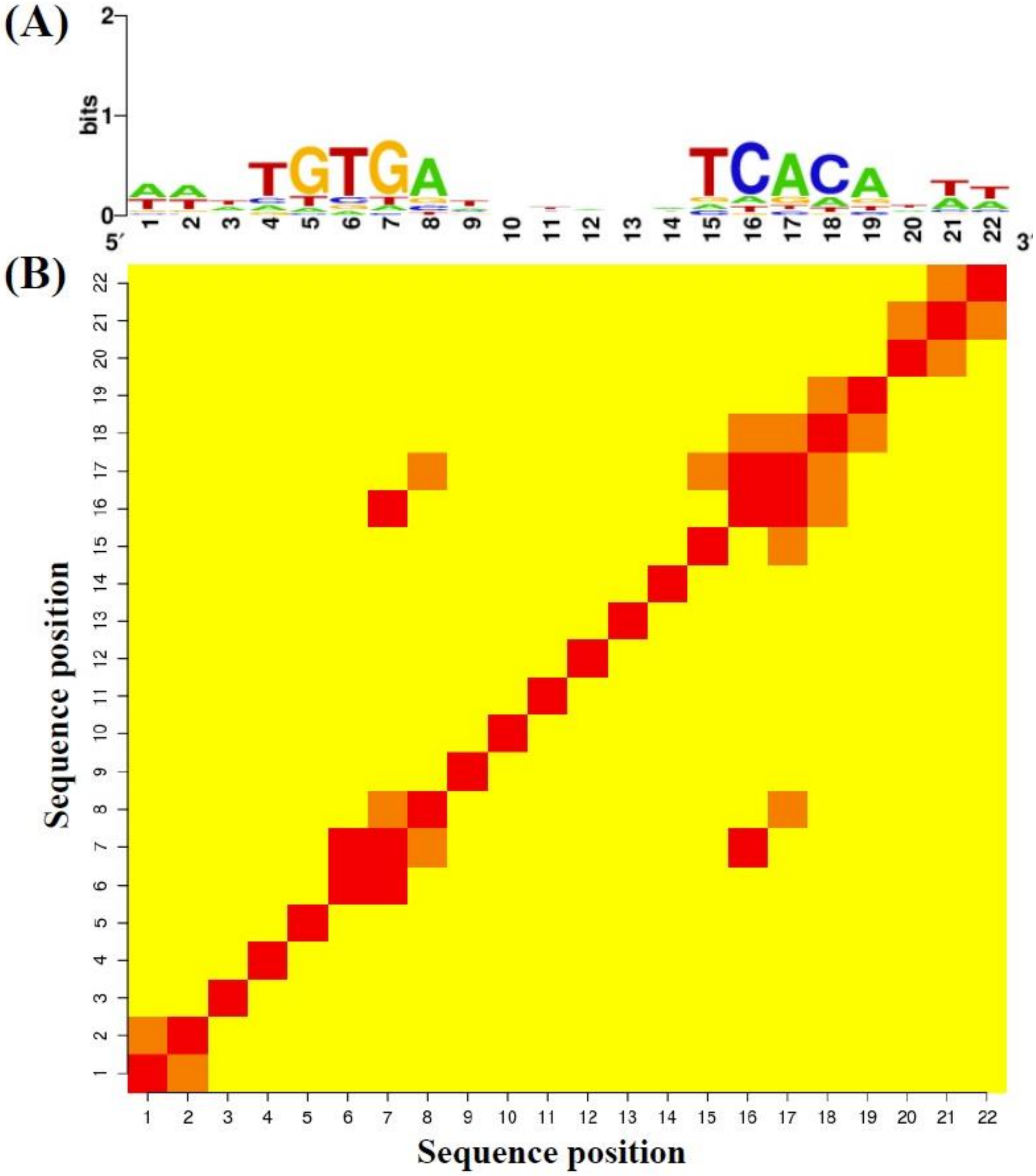


Figure 4. Sequence logo and heat map of CRP. Out of 374 CRP motifs, 308 sequences with the 22 bp motif were used. (A) Sequence logo for CRP using 308 sequences [33]. (B) Heat map of CRP.

4.2 Performance for CRP dataset

I evaluated the performance of DIpartite by using the TF binding sites of CRP. Out of 374 sequences of CRP binding sites, 323 unique sequences were used as the test dataset. Jensen and Liu (2004) analyzed the CRP binding sites as a bipartite motif and proposed the consensus sequence, tGTcA<6,8>CAcattt [19, 35]. I conducted motif prediction by using MEME (ver. 5.0.2), BioProspector (release 2), AMD, BiPad (ver. 2), and DIpartite for these 323 sequences of CRP binding sites (Table 1 and Figure 5). DIpartite with the "PWM" or "DWM" options is referred to as DIpartite PWM or DIpartite DWM, respectively. Although DIpartite PWM performed best among the tested software for the one-block model, namely, the 22-bp motif, the performance was comparable among MEME, BioProspector, BiPad, and DIpartite. AMD exhibited a combined *nCC* value of less than 0.9. I assessed the performance of DIpartite by randomly sampling 100 datasets with 100 sequences from the CRP binding sites. DIpartite DWM slightly outperformed other tested tools for 100 datasets (Table 2 and Figures 6).

Table 1. Peformance comparison for the 323 sequences of CRP binding sites.

sites	Search Pattern	MEME	BioProspector	Dipartite (PWM)	Dipartite (DWM)	BiPad	AMD
323	22	0.852	0.924	0.936	0.934	0.928	0.883
	6<[10]>6		0.775	0.837	0.839	0.742	
	6<[8]>8		0.857	0.9	0.899	0.891	
	8<[6]>8		0.909	0.932	0.932	0.925	

Colored cells indicates the highest performance among tested software.

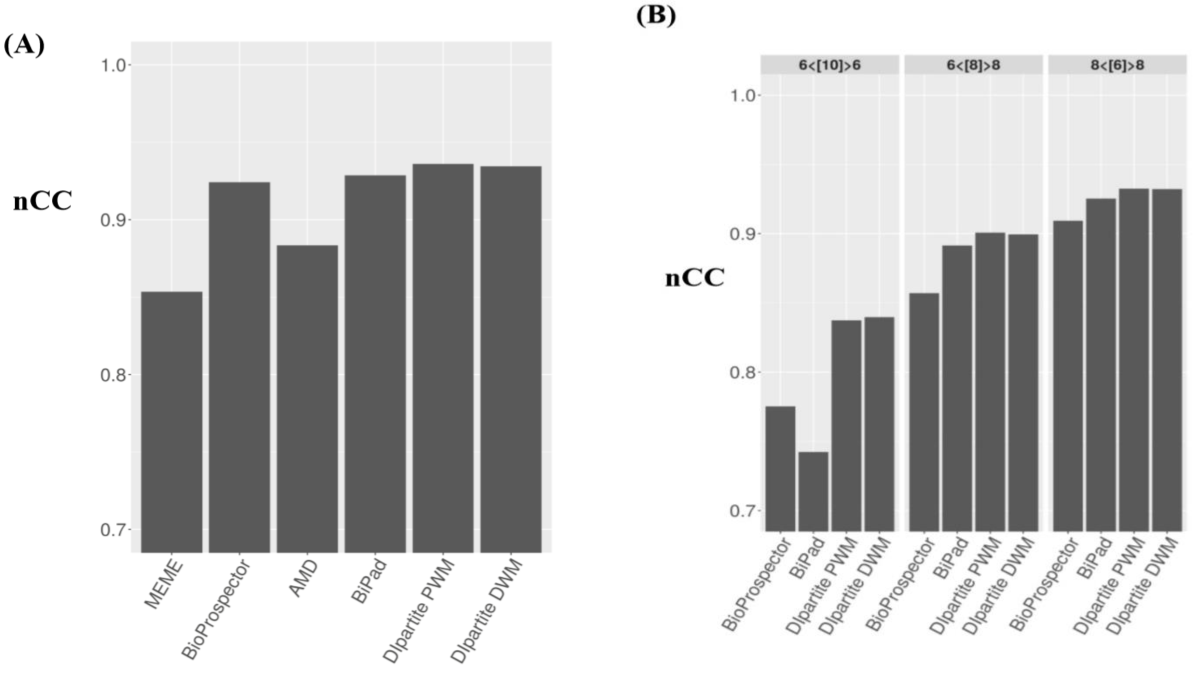


Figure 5. The performance comparison for 323 CRP sequences. The combined nCC values were plotted. (A) Summary of the results for searching the one-block motif, i.e., the 22 bp motif, by MEME, BioProspector, AMD, BiPad, DIpartite PWM and DIpartite DWM. (B) Summary of the results for searching the bipartite motifs, i.e., 6<[10]>6, 6<[8]>8, and 8<[6]>8, by BioProspector, BiPad, DIpartite PWM and DWM.

Table 2. Peformance comparison for randomly sampling 100 datasets with 100 sequences from the CRP binding sites.

#of sites	Search Pattern	MEME	BioProspector	DIpartite (PWM)	DIpartite (DWM)	BiPad	AMD
323	22	0.8153	0.8687	0.8976	**0.8993**	0.8887	**0.8393**
	6<[10]>6		0.7782	**0.8266**	0.8049	0.8048	
	6<[8]>8		0.8156	**0.8724**	0.8532	0.8594	
	8<[6]>8		0.9061	**0.9274**	0.9205	0.9205	

Colored cells indicates the highest performance among tested software.

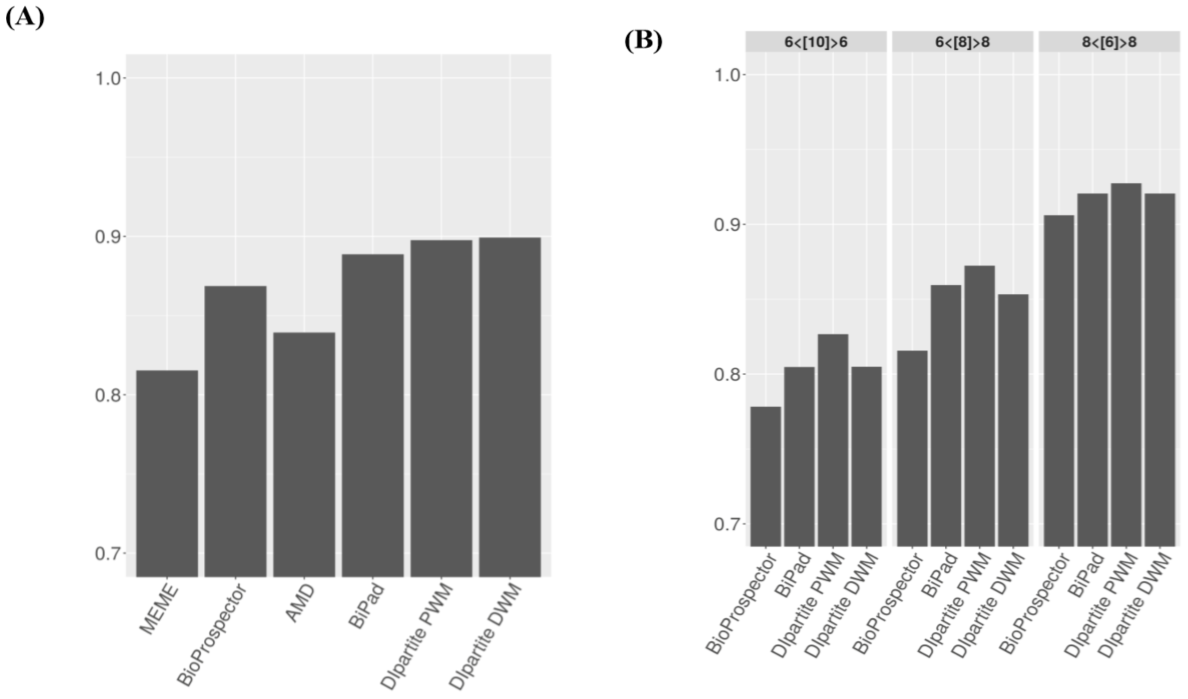


Figure 6. The performance comparison for 100 CRP datasets. 100 datasets consisting of 100 sequences were generated by randomly sampling the CRP datasets. (A) Summary of the results for searching the one-block motif, i.e., the 22 bp. (B) Summary of the results for searching the bipartite motifs, i.e., 6<[10]>6, 6<[8]>8, and 8<[6]>8.

For the bipartite motif, I compared BioProspector, BiPad, DIpartite PWM, and DIpartite DWM. The performance of searching the bipartite motifs was lower than that of searching the one-block model, i.e., 0.936 by DIpartite PWM. For all three types of the bipartite motifs, DIpartite PWM and DIpartite DWM were superior to BioProspector and BiPad. DIpartite DWM was superior to DIpartite PWM in the case of 6<[10]>6. I conducted the performance comparison by using 100 datasets with 100 sequences. DIpartite PWM outperformed other tested tools. Although the implementation of DIpartite PWM is similar to that of BiPad, DIpartite PWM slightly outperformed BiPad. This might be because DIpartite takes into consideration the background sites (not motif sites) unlike BiPad, that is, $b(x)$ in Equation (2). Taken together, DIpartite successfully detected the binding sites of the one-block or bipartite motifs.

In addition, I tested the running time by using the CRP dataset. Although BioProspector was the fastest software among tested software, DIpartite was comparable with BiPad (Table 3 and Figure7).

Table 3. Running times. The datasets consisting of 20, 50, 100, 200, 500 and 1,000 sequences were generated by randomly sampling the CRP sequences.

	DIpartite PWM	**DIpartite DWM**	**BiPad**	**BioProspector**
20 seq	3.686	9.789	19.022	1.237
50 seq	14.279	46.307	66.682	2.937
100 seq	47.752	188.995	175.036	5.474
200 seq	169.444	607.807	388.853	10.288
500 seq	928.439	3165.046	1093.804	26.103
1000 seq	3252.165	10888.849	2502.075	52.446

Run time is indicated as second.

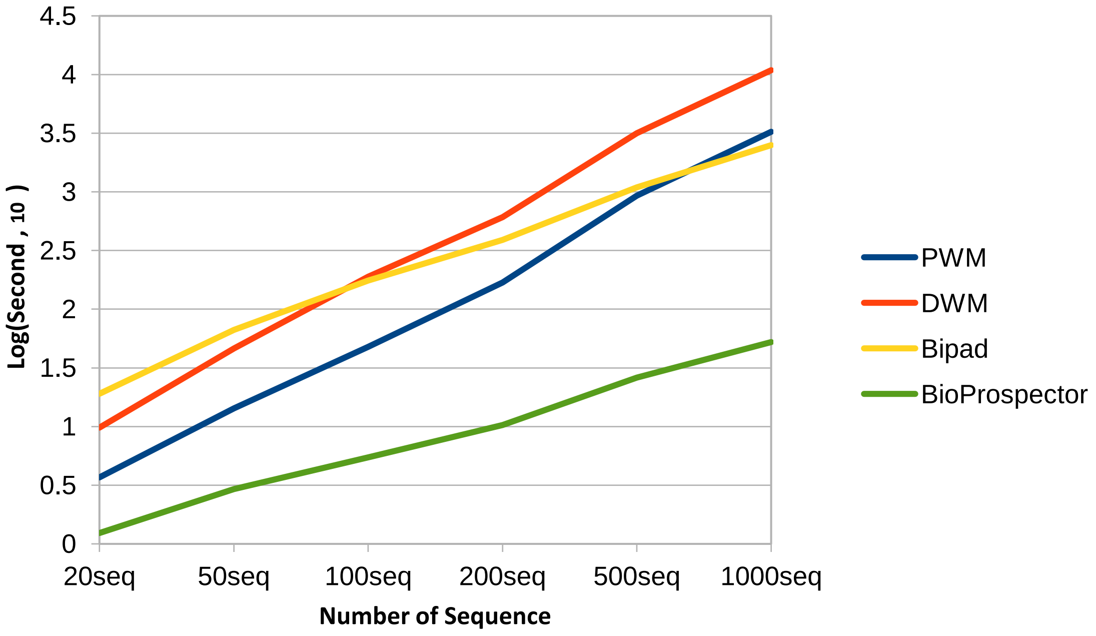


Figure 7. Running times. The datasets consisting of 20, 50, 100, 200, 500 and 1,000 sequences were generated by randomly sampling the CRP sequences. X-axis and Y-axis correspond to the number of sequences, and the running time [s] on a log scale. BioProspector, BiPad, DIpartite PWM (as PWM), and DIpartite DWM (as DMW) were tested.

4.3 Performance for human dataset

I selected the human promoter sequences as the bipartite motifs with the constant gaps in eukaryote [15]. Of 1,460 motifs, 46 motifs with the more than 4-nt gaps were filtered. The promoter sequences around the positions of each motifs (500 bp upstream to 500 bp downstream) were retrieved as the target sets. Since AMD did not detect any motifs for six motifs, i.e., RGGANNNNNAKTCC (54 sequences), RKCTGNNNNNRMTTA (21 sequences), TTGRNNNNNNTCCAR (21 sequences), YMATCNNNNNGCGM (50 sequences), YTGGANNNNNNYCAA (26 sequences), and YTTGRNNNNNNGCCNR (50 sequences), these were excluded, and 40 datasets were evaluated for the performance of DIpartite. I assessed the performance for 40 motif datasets (Figure 8A). DIpartite DWM exhibited the highest performance (50%), followed by DIpartite PWM (48%), BioProspector (38%), MEME (20%), BiPad (8%), and AMD (3%) (Table 4 and Figure 9), indicating that DIpartite performs equivalently to or better than the other tools for detecting DIpartite motifs. In addition to the result of CRP 6<[10]>6, DIpartite DWM outperformed other tested tools, suggesting that DWM might improve the bipartite motif detection. Apparently, MEME and BiPad exhibited the larger interquartile range (Figure 8B), indicating that MEME and BiPad outperformed for the particular motifs but underperformed for the other motifs.

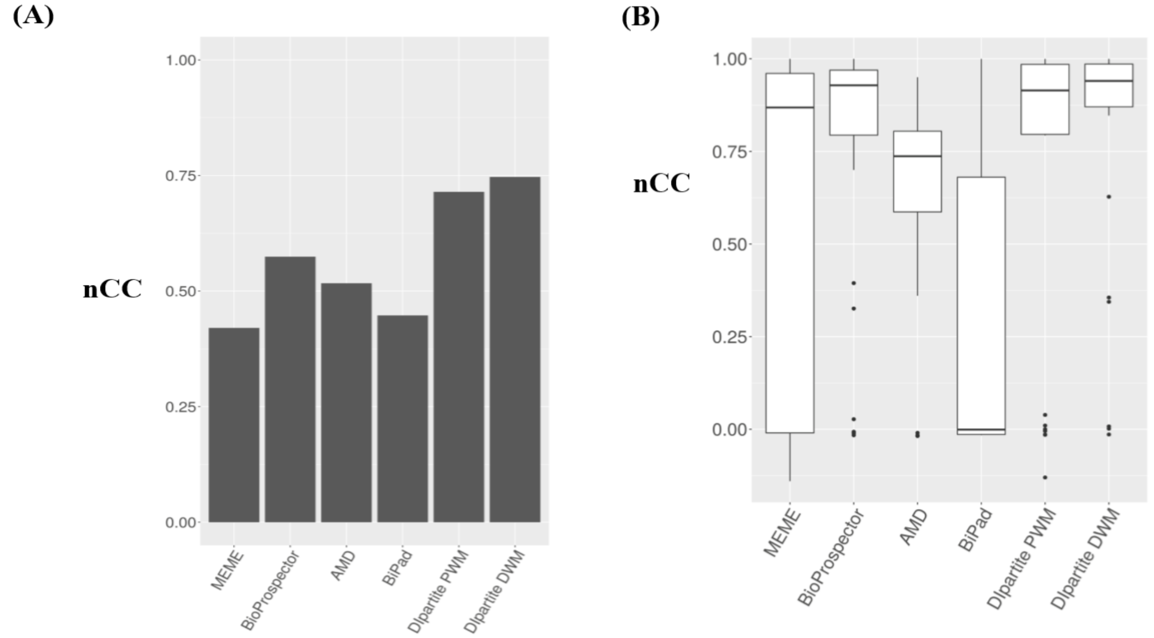


Figure 8. The performance comparison for human promoter datasets. (A) Summary of the results of all 40 human promoter datasets. (B) Boxplots of the nCC values for each 40 human promoter datasets.

Table 4. The performance comparison for 40 human data (colored cells indicates the highest performance).

	#of Site	AMD	BioProspector	MEME	DIpartite-PWM	DIpartite-DWM	BiPad
AATNNNNNNCAGAYR	11	-0.0182	**1**	-0.015	**1**	**1**	-0.015
AATNNNNNNCAGCNG	19	0.8043	**1**	0.9465	**1**	**1**	0.0277
AWCTTNNNNNNNGGG	56	0.5797	**1**	0.9456	**1**	-0.014	-0.014
CCCNNNNNNAAGWT	71	0.3603	**0.8857**	0.8714	0.8571	0.6276	-0.014
CGCNNNNNATTGK	47	0.6814	0.8226	**1**	**1**	**1**	0.0035
CGCNNNNNNATGAY	35	0.6949	**0.942**	**0.942**	0.8737	0.8468	-0.014
CNGCTGNNNNNNATT	23	0.8043	**1**	**1**	**1**	**1**	-0.015
CRTCANNNNNNNGCGMC	44	0.8564	0.9306	0.9537	0.9184	**0.9768**	0.0591
GCCNNNNNNNATTRK	45	0.7677	**0.9538**	-0.0104	**0.9538**	**0.9538**	-0.0134
GCGCNNNNNNATGNM	31	0.676	0.9017	-0.015	-0.015	**0.9345**	-0.015
GCGNNNNNTTTRA	57	0.8472	0.9644	0.9644	**0.9822**	**0.9822**	-0.0061
GGAMTNNNNNTCCY	89	0.5891	0.8746	0.7981	**0.8974**	0.886	-0.014
GGCNNNNNKCCAR	252	0.8236	0.3259	-0.0095	0.9143	0.9143	**0.9183**
GGCNNNNNNATTGK	55	0.839	**0.9262**	0.8656	-0.004	0.9078	0.0004
GKCGCNNNNNNNTGAYG	36	0.8148	0.9435	0.9435	0.887	**0.9717**	0.8604
GTCNNNNNRNCAAC	47	0.7077	0.9567	**1**	-0.014	**1**	-0.014
GTTGNYNNNNNGAC	57	0.7261	0.9466	**0.9644**	0	0.0076	0.0393
GTTNMNNNNNNAAC	144	0.6169	**0.9648**	0.8143	-0.13	0.9226	-0.013
GTTNNNNNKNAAC	151	0.5716	0.953	**0.9597**	-0.0047	0.9262	0.004
KCGCNNNNNGATKR	37	0.7881	0.9725	0.8003	0.0388	**1**	0.6221
KNCATNNNNNNGCGC	53	0.5714	0.6999	0.9042	**0.9616**	0.9425	-0.0035
KYTGCYNNNNNRACA	33	0.7544	**0.9938**	0.963	0.9876	**0.9938**	0.0014
MCAATNNNNNGCG	64	0.7711	**0.8782**	-0.013	0.0101	0.0076	0.0101
MCAATNNNNNNNGCC	38	-0.0165	**1**	0.9733	**1**	**1**	-0.014
MYAATNNNNNNNNGGC	77	**0.7051**	0.3945	0.0174	-0.0009	0.0016	0.0482
RTCATNNNNNNGCG	49	0.7797	0.9172	0.9172	0.793	0.9379	-0.0021
RYAAAKNNNNNNTTGW	44	0.9418	**1**	**1**	0.9148	**1**	**1**
TAAKYNNNNNCAGMY	14	-0.0169	**1**	-0.015	0.797	0.9275	-0.015
TCTGNNNNNTGTMR	35	0.516	0.9006	0.971	**1**	0.8737	0.8737
TGGNNNNNNKCCAR	214	0.8969	-0.009	-0.0133	**0.9194**	0.344	0.9052
TGTYNNNNNRGCARM	37	-0.0097	0.8628	0.8079	**0.9451**	**0.9451**	0.8902
TTTNNNNNAACW	210	0.7976	-0.011	0.0024	**0.8602**	**0.8602**	0.8554
TYAAANNNNNCGC	44	0.748	-0.007	0.8158	**1**	**1**	-0.013
WCAANNNNNNMTTTRY	18	0.8052	**1**	**1**	**1**	0.9435	-0.016
WGTTNNNNNAAA	226	0.7805	-0.006	-0.0011	0.8522	0.8522	**0.8567**
YKACANNNNNCAGA	60	0.695	0.7078	0.8986	0.802	**0.9831**	-0.014
YNGGCNNNNNNYCAAR	41	0.7066	-0.016	-0.016	**0.9752**	0.9024	-0.0036
YRTCTGNNNNNNATT	23	0.5433	0.9117	-0.015	**0.9558**	**0.9558**	-0.015
YTGGMNNNNNGCC	275	0.8724	0.0272	-0.013	0.8611	**0.9424**	0.9277
YTGGMNNNNNNCCA	192	0.9502	0.9683	-0.14	**0.9841**	0.3556	0.9788
Number of the highest performance		**1**	**15**	**8**	**19**	**20**	**3**
Percentages of motifs with highest nCC		**3%**	**38%**	**20%**	**48%**	**50%**	**8%**

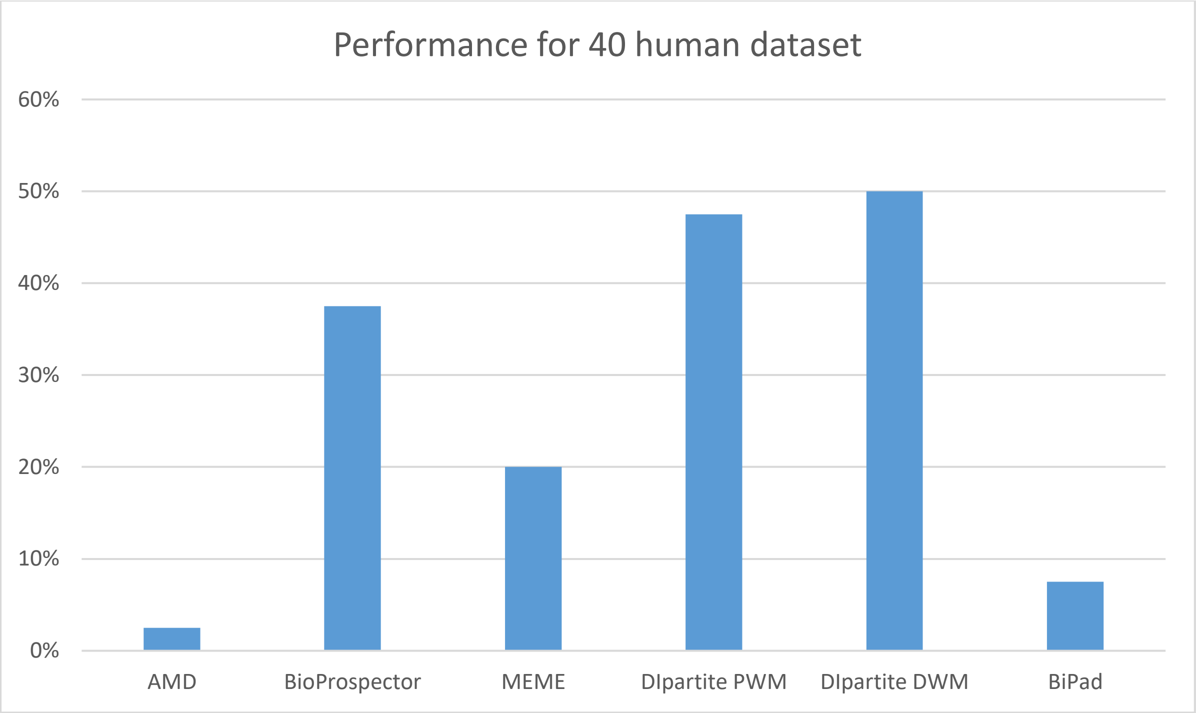
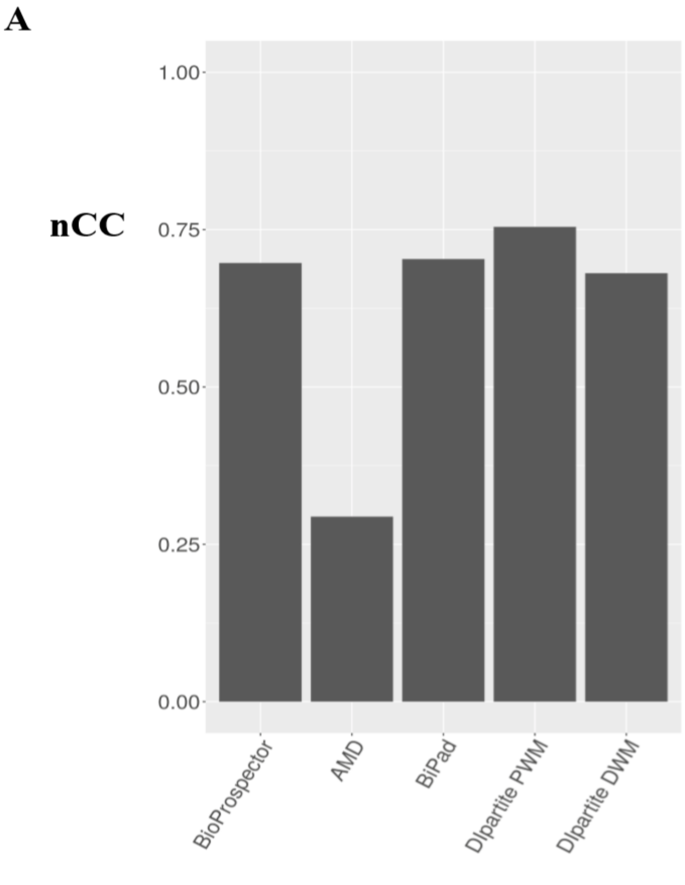


Figure 9. The proportion of performance comparison for 40 human data, AMD, BioProspector, MEME, Dipartite PWM, Dipartite DWM, and BiPad.

4.4 Performance for sigma factor dataset

I compared the performance of DIpartite with BioProspector, AMD and BiPad for the bipartite motifs with the variable gaps. I adopted the nine of the bipartite sigma transcription factors in *B. subtilis*, i.e., σ^A (344 sequences), σ^B (64 sequences), σ^D (30 sequences), σ^E (70 sequences), σ^F (25 sequences), σ^G (55 sequences), σ^H (25 sequences), σ^K (53 sequences), and σ^W (34 sequences) from DBTBS [7] as the test datasets (Figure 10 and Table 5). DIpartite PWM performed better than BioProspector, BiPad, AMD and DIpartite DWM (Figure 10A) for six sigma factors other than σ^D, σ^E and σ^H. While the performance of DIpartite PWM was excellent for two sigma factors, i.e., σ^A and σ^F, that of DIpartite DWM was comparable in four sigma factors, i.e., σ^B, σ^G, σ^K, and σ^W. AMD exhibited the relatively lower *nCC* values for all nine datasets unlike as the results for the human promoter sequences, suggesting that the variable gap lengths could affect the performance of AMD because AMD has been developed for detecting the bipartite motifs with the constant gaps. AMD with the option "-T 1" did not detect any motifs for two sigma datasets, i.e., σ^E and σ^F (Figure 10B).



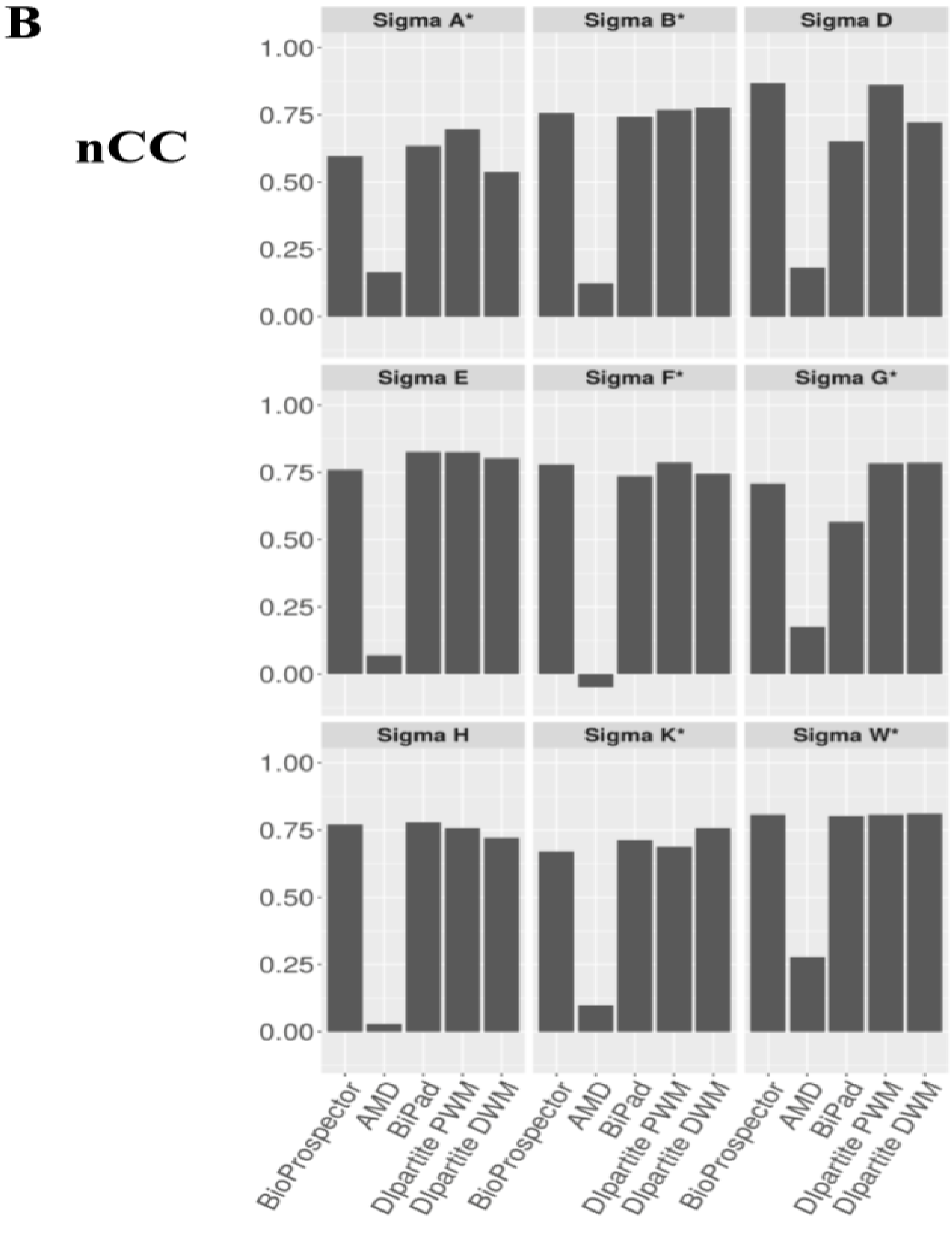


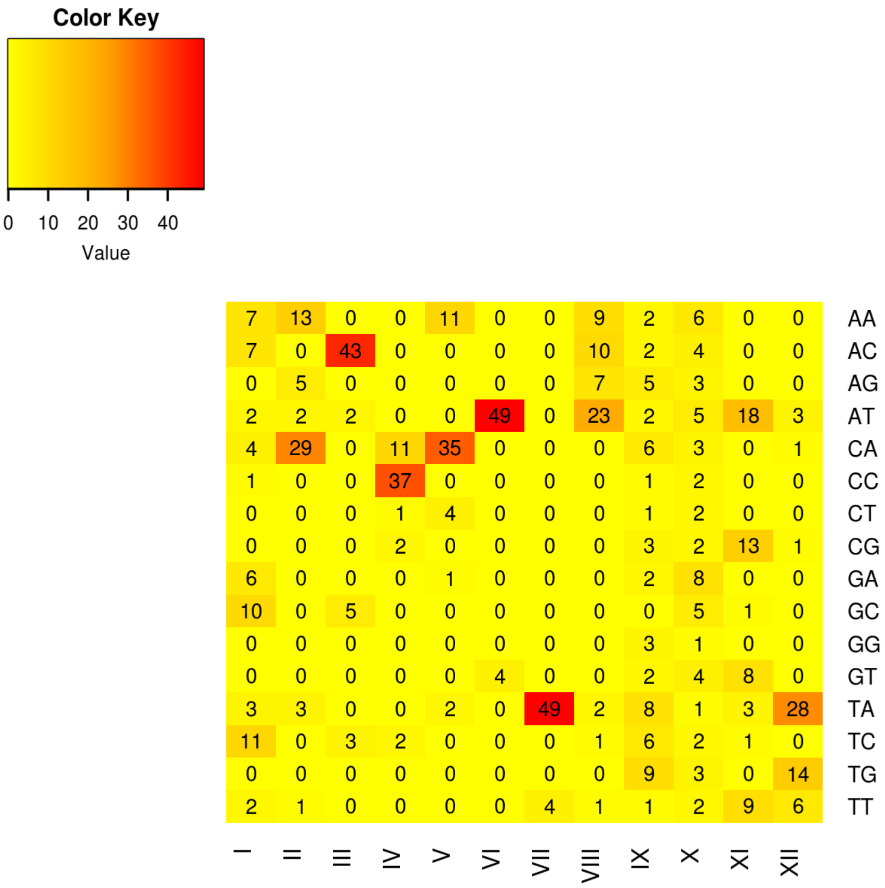
Figure 10. The performance comparison for B. subtilis datasets. (A) Summary of the results of all sigma datasets. (B) Summary of the results of each sigma datasets. σ^A, σ^B, σ^D, σ^E, σ^F, σ^G, σ^H, σ^K, and σ^W consist of 344, 64, 30, 70, 25, 55, 25, 53, and 34 sequences, respectively. The asterisks indicate if DIpartite performed better than BioProspector, AMD, and BiPad.

Table 5. The performance comparison for B. subtilis datasets. The all sigma datasets parameter and result.

Sigma Factors	**sites**	**Search Pattern**	**BioProspector**	**DIpartite (PWM)**	**DIpartite (DWM)**	**BiPad**	**AMD**
SigmaA	**344**	**6<[11,23]>6**	0.596	**0.697**	0.537	0.635	0.165
SigmaB	**64**	**6<[12,18]>6**	0.757	0.769	**0.776**	0.744	0.124
SigmaD	**30**	**4<[12,18]>8**	**0.868**	0.861	0.722	0.652	0.181
SigmaE	**70**	**7<[12,18]>8**	0.760	0.827	0.803	**0.828**	0.070
SigmaF	**25**	**5<[13,19]>10**	0.780	**0.788**	0.746	0.738	-0.050
SigmaG	**55**	**5<[15,20]>7**	0.709	0.785	**0.787**	0.567	0.175
SigmaH	**25**	**7<[9,18]>5**	0.770	0.757	0.721	**0.779**	0.028
SigmaK	**53**	**4<[9,17]>9**	0.670	0.688	**0.757**	0.712	0.098
SigmaW	**34**	**10<[13,17]>6**	0.808	0.808	**0.811**	0.802	0.278

Colored cells indicates the highest performance among tested software.

Among four sigma factors with the highest performance coefficients by DIpartite DWM, the *nCC* value for σ^K was greatly improved by DIpartite DWM, i.e., 0.757, indicating that the base interdependencies could exist in the motif of σ^K (Figure 11). I observed that the left motif of DIpartite DMW was shifted and "AC" was more over-represented, indicating that the left motif of σ^K might be improved. The position 7 was 'T' in all 53 sequences (Figure 12A), consistent with the known motif in DBTBS. In similar, the highest frequency of the dinucleotide, "AT" and "TA", were observed at the position of 6 and 7, 7 and 8, respectively (Figure 12B).



I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	
7	13	0	0	11	0	0	9	2	6	0	0	AA
7	0	43	0	0	0	0	10	2	4	0	0	AC
0	5	0	0	0	0	0	7	5	3	0	0	AG
2	2	2	0	0	49	0	23	2	5	18	3	AT
4	29	0	11	35	0	0	0	6	3	0	1	CA
1	0	0	37	0	0	0	0	1	2	0	0	CC
0	0	0	1	4	0	0	0	1	2	0	0	CT
0	0	0	2	0	0	0	0	3	2	13	1	CG
6	0	0	0	1	0	0	0	2	8	0	0	GA
10	0	5	0	0	0	0	0	0	5	1	0	GC
0	0	0	0	0	0	0	0	3	1	0	0	GG
0	0	0	0	0	4	0	0	2	4	8	0	GT
3	3	0	0	2	0	49	2	8	1	3	28	TA
11	0	3	2	0	0	0	1	6	2	1	0	TC
0	0	0	0	0	0	0	0	9	3	0	14	TG
2	1	0	0	0	0	4	1	1	2	9	6	TT

Figure 11. The frequency graph matrix of sigmaK data by DIpartite-DWM.

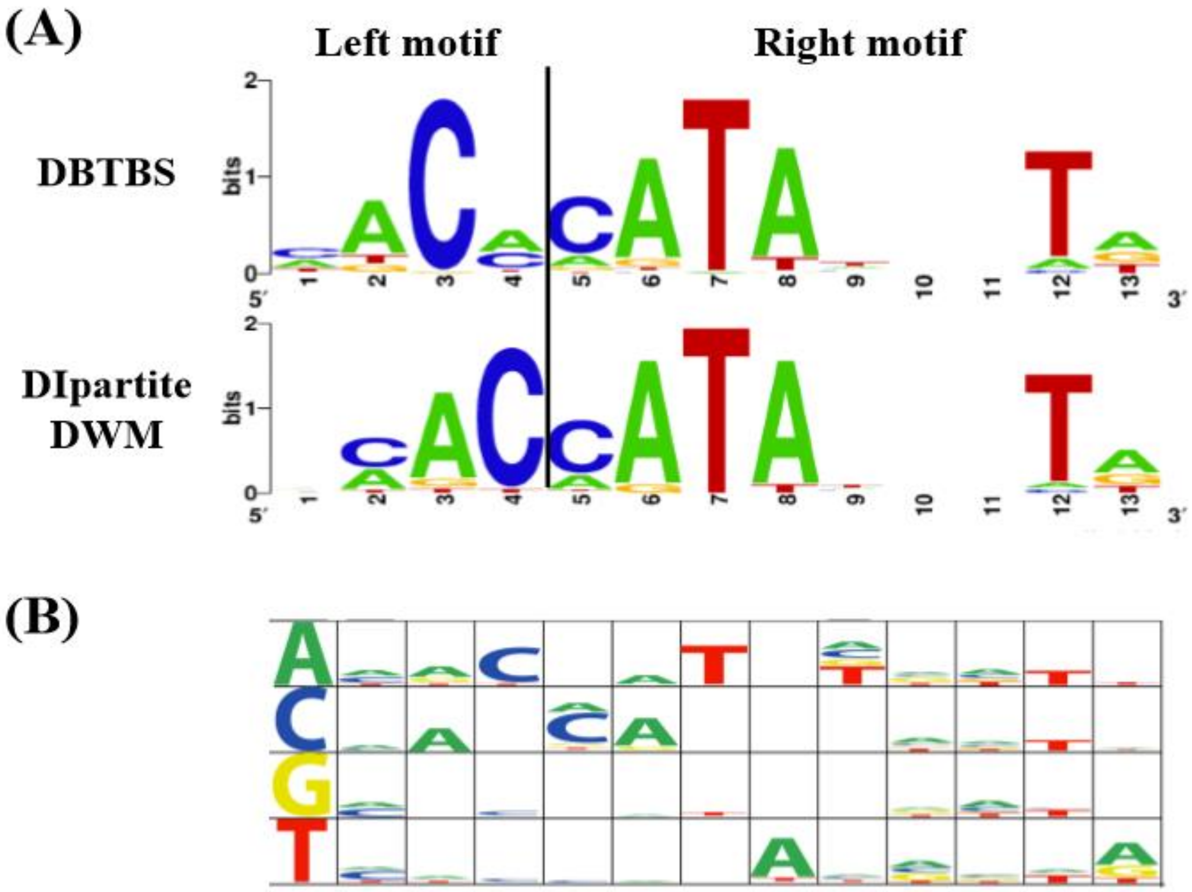


Figure 12. Sequence logo for σ^K by DIpartite DWM. (A) Sequence logos generated by DBTBS and DIpartite DWM. The border between the left and right motifs, i.e., position 4, 5, is indicated as the vertical line. (B) Sequence logo for the probability of each dinucleotides. One base before was depicted in first column. Size of each logo was proportional to the probability of the dinucleotides.

The *nCC* value of σ^A was greatly improved by DIpartite PWM, i.e., 0.697. While the sequence logo generated from the result of BioProspector was similar to that generated from the result of DIpartite DWM, those of BiPad and DIpartite PWM was different with those (Figure 13). In particular, DIpartite PWM exhibited the conserved base, “T”, in position 1. This result was consistent with the motif, TTGACA<>tgnTATAAT, proposed by DBTBS [7]. DIpartite PWM showed the sequences with the minimum entropy.

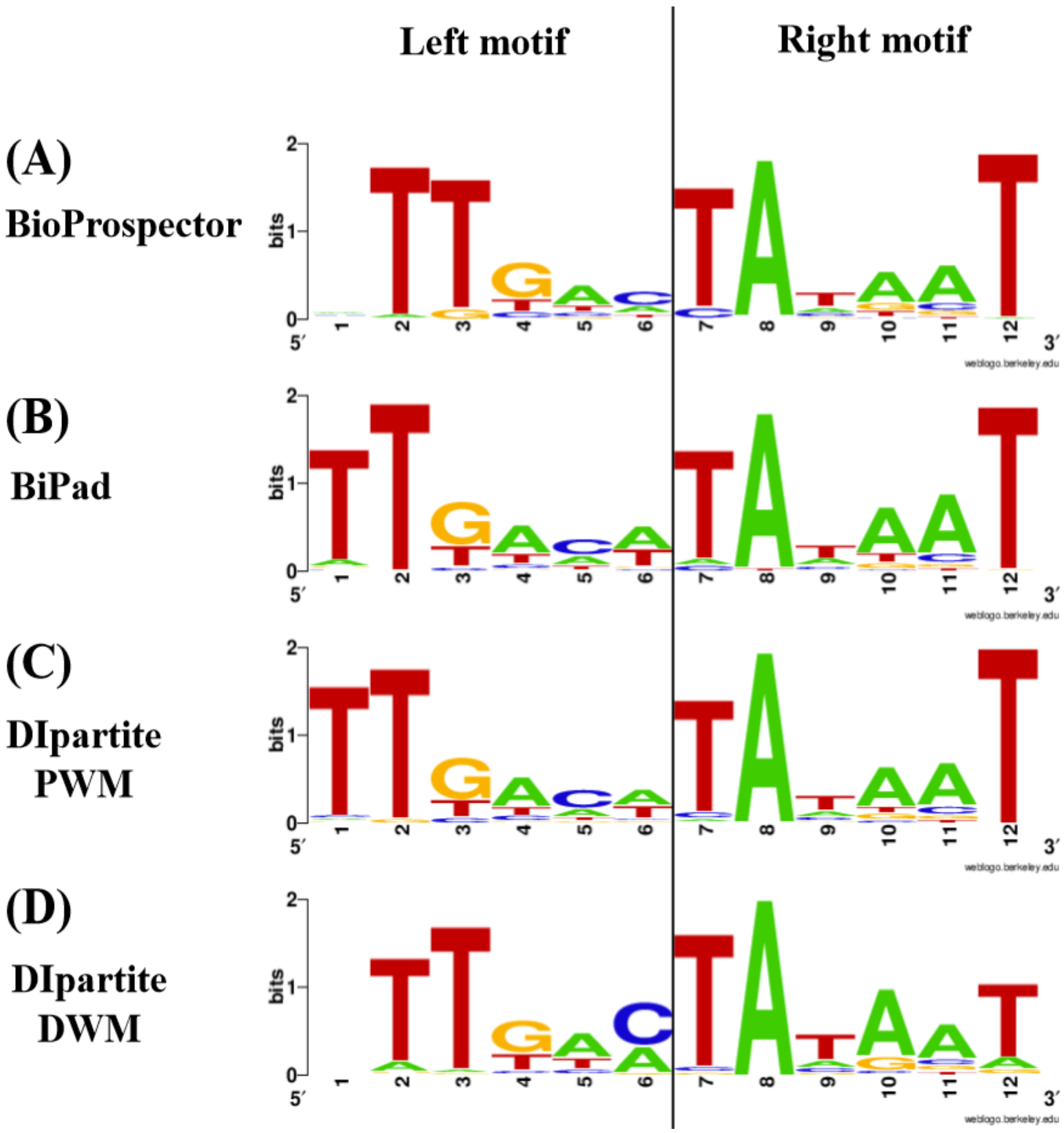


Figure 13. Sequence logos for σ^A *by results of (A) BioProspector, (B) BiPad, (C) DIpatrite PWM, and (D) DIpartite DWM.*

I assessed the performance of DIpartite DWM in terms of the sizes of the input datasets. By randomly sampling the sequences of σ^A in *B. subtilis*, I generated 100 datasets for each including 10, 20, 50, 100, 150, 200, and 300 sequences (Figure 14). Upon increasing the size of the datasets, DIpartite PWM and DWM exhibited better performance. Notably, DIpartite underperformed for the datasets with 10 and 20 sequences, suggesting that DIpartite could perform well for data including more than 50 sequences. The variances of DIpartite PWM for the datasets with 200 and 300 sequences were relatively smaller than those of DIpartite DWM. One potential reason for this is that DWM consists of the frequencies of 16 dinucleotides (Equation 3).

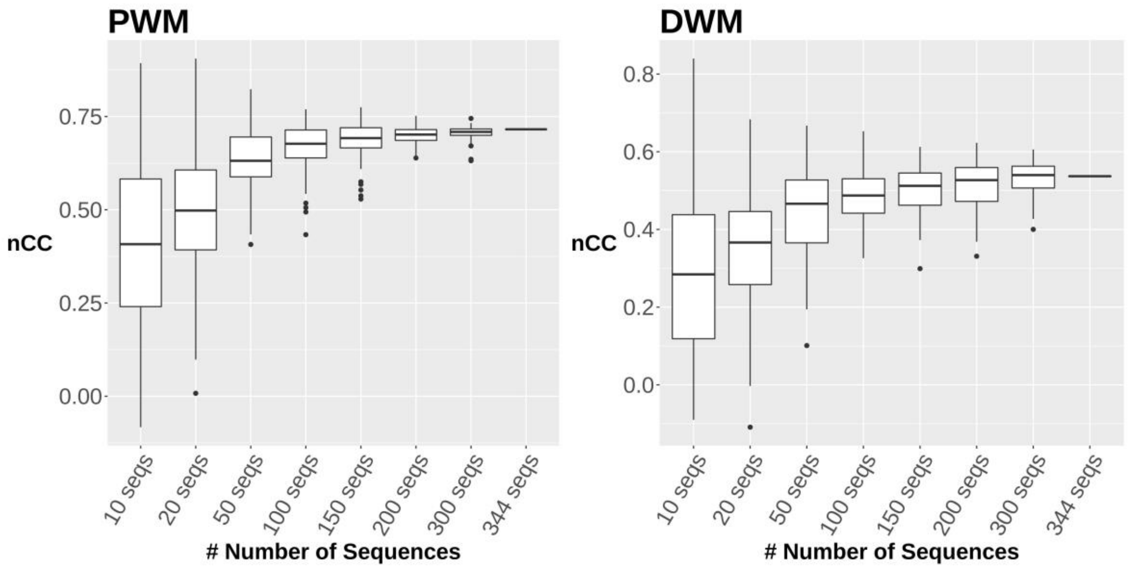


Figure 14. The performance of DIpartite: (A) PWM; (B) DWM. 100 datasets were generated by sampling of the σA dataset. The sizes of the dataset were 10, 20, 50, 100, 150, 200, 300 sequences.

4.5 Performance for the dataset with noise sequences

I assessed the performance for the dataset with noise sequences. DIpartite allows the users to search the motifs for the dataset with noise sequences, known as ZOOPS. I evaluated the accuracy of detection noise sequences by using the datasets with noise sequences. I chose the CRP datasets and human dataset as the test datasets of the one- and two-block motifs. I compared the performance of noise detection by DIpartite with that by MEME for the CRP dataset (Table 6). DIpartite exhibited the TPRs (true positive rate), i.e., 0.835, 0.863, and 0.876 for the datasets with 25%, 50%, and 100% noise sequences, respectively. This indicates that DIpartite ZOOPS could be well tolerated with the noise sequences. Indeed, MEME exhibited the lower FPRs, but lower TPRs, suggesting that DIpartite ZOOPS would be comparable with MEME ZOOPS.

Table 6. The performance of noise detection for the one-block motif.

	MEME			**DIpartite**		
Noise Percentage	25%	50%	100%	25%	50%	100%
Number of TP, TN	(TP:323, TN: 81)	(TP:323, TN: 162)	(TP:323, TN: 323)	(TP: 323, TN: 81)	(TP:323, TN: 162)	(TP: 323, TN: 323)
FPR	0.061	0.030	0.024	0.172	0.172	0.167
TPR	0.798	0.777	0.739	0.835	0.863	0.876

Noise sequences were sampled from the genome sequence of E. coli.

TPR: True positive rate, FPR: False positive rate.

Finally, I compared the performance of noise detection for the two-block dataset, i.e., RYAAAKNNNNNNTTGW consisting of 44 sequences. BioProspector (*nCC*=1) and BiPad (*nCC*=1) outperformed DIpartite (*nCC*=0.914). Increasing the noise sequences, BioProspector and BiPad exhibited lower *nCC* values. DIpartite exhibited higher *nCC* values even adding the noise sequences, suggesting that DIpartite could work well for both one- and two-block motifs with noise sequences (Table 7).

Table 7. The performance of nCC score after added the noise to dataset and compare PWM with BioProspector and Bipad tools, Noise sequences were sampled from the genome sequence of human.

Noise Percentage	0%	25%	50%	100%
Number of TP, TN	(TP: 44, TN:0)	(TP: 44, TN:11)	(TP: 44, TN:22)	(TP: 44, TN:44)
BioProspector	1	0.907	−0.16	−0.16
BIpad	1	1	1	−0.16
DIpartite PWM	0.9148	1	1	1

5. Conclusions

I have developed DIpartite for detecting TFBSs, consisting of the bipartite motifs. DIpartite enables *ab initio* prediction of the conserved motif based on not only PWM, but also DWM. I evaluated the performance of DIpartite compared with freely available tools, i.e., MEME, BioProspector, AMD and BiPad. Both of DIpartite PWM and DWM perform equivalent or better than those in the cases of the bipartite motifs with the fixed and variable gaps like promoter sequences in human and sigma factors in *B. subtilis*. The prediction of σ^K was greatly improved by taking into consideration base interdependencies. DIpartite can be found at https://github.com/Mohammad-Vahed/DIpartite.

6. Usage of DIpartite

6.1 Getting started

Download from: https://github.com/Mohammad-Vahed/DIpartite

```
$ git clon https://github.com/Mohammad-Vahed/DIpartite

$ cd DIpartite

$ make
```

6.2 Make file

```
CC = g++

CFLAGS = -Wall

DIpartite: DIpartite.cxx

        $(CC) $(CFLAGS) -o DIpartite DIpartite.cxx


clean:

        rm Dipartite
```

6.3 Example of usage

<<PWM base prediction>>

For one block motif

```
./DIpartite -i <fasta> -n 1 -p 2 -m 6 -M 0 -g 0 -G 0 -o <output>
```

==> Find the motif of width 6bp one block motif length from the sequence file (FASTA), and search both the given and reverse complement strands of DNA.

For two block motif

```
./DIpartite -i <fasta> -n 2 -p 2 -m 6 -M 6 -g 0 -G 0 -o <output>
```

==> Find the motif of width 6bp for left motif and 6bp for right motif length from the sequence file (FASTA), and search both the given and reverse complement strands of DNA and method DWM.

6.4 Arguments

-i input file

-m left motif width (default 6)

-M right motif width (default 6)

-g min gap between two motif blocks (default 0)

-G max gap between two motif blocks (default 0)

-t number of times trying to repeat process to find best motif (default 30)

-o output file (default output.txt)

-f 1 for fasta file, or 2 for text file (no header) (default 1)

-p 1 for the given strand, or 2 for both the given and reverse complement strands (default 1)

-n 1 for PWM, or 2 for DWM (default 1)

-s 1 for one occurrence motif site per sequence (oops), or 2 for any number of repetitions (anr) or 3 Zero or One Occurrence per Sequence (zoops) (default 1)

7. DIpartite-Learning:

7.1 Web service

I implemented a new web service (by ASP.NET, Java, HTML, and CSS), sa called "DIpartite-Learning", for discovery the best motif site and size without any set parameters. Machine learning includes statistical modeling methods that automatically learn beneficial knowledge from input data and derive unknowns based on a set of knowns. Therefore, these data-driven intelligent algorithms appear as key software for the accurate recognition of CREs (Cis-regulatory elements) (43-52).

In the web service, users can use the DIpartite tool as well as the DIpartite-Learning. The most significant motifs found by DIpartite-Learning are displayed graphically on the main results page with a table containing summary statistics for each motif. Detailed motif information, including the sequence logo, PWM and DWM, consensus sequence.

I added the new option DI-Logo and Mono-Logo, user can input sequences data and select the "Just show web logo" option, then show Weblogo (DI or Mono).

7.2 Method

The DIpartite-Learning approach combines the output of three algorithms, each designed to identify a particular class of motifs. DIpartite-Learning methods include Machine learning, Dynamic programming, and Gibbs sampling strategy.

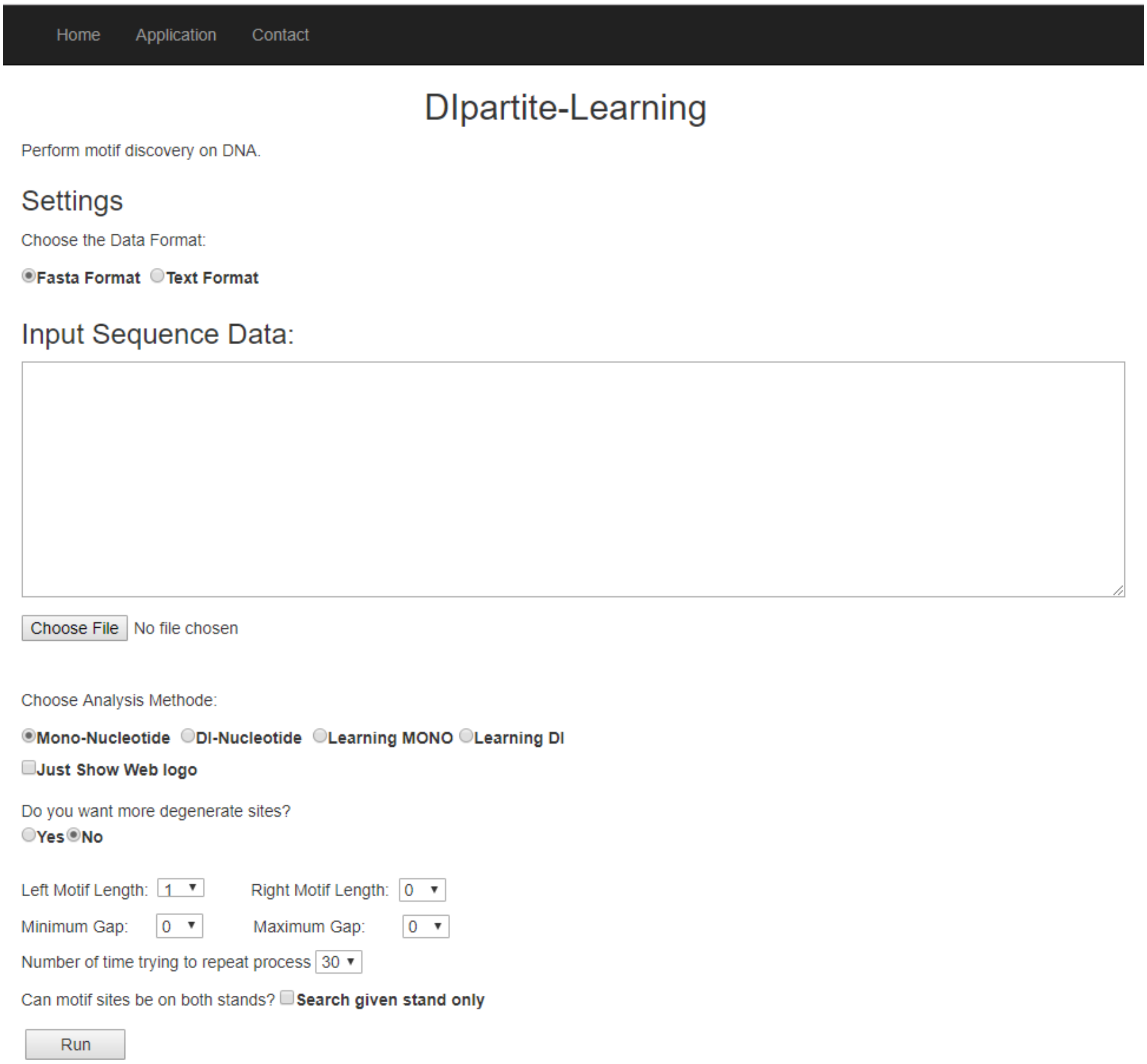
The base of DIpartite-Learning tool is PWM and DWM. Since the users do not have any parameters for find motifs, the parameter space becomes very large. Therefore, I adopted the method for data analysis to gain the best size and location of motifs.

7.3 Algorithm

- Text mining of input file
- Find the minimum and maximum of motif and gap lengths
- Randomly generate the parameters to the process
- Make the PWM(or DWM) and calculate entropy for each step as DIpartite
- Analysis of repeating pattern of DNA sequence and compare with the previous step
- If improve the score, learn to an algorithm to keep this pattern and test new parameters with the same model until finding the better score and come back to calculate entropy step.
- Replace previous parameters value with new ones.
- Learn to the algorithm does not test the pattern with a low score and limit the search range
- Update score and location of motif

Users can be use DIpartite tool on web site that is very user friendly and easy to run and get result, user just should input data and set the parameters (Figure 15), then Run.

In this part, results completely same as old version of DIpartite. However, I made graphical method for show the result and text format (Figures 15~20).



DIpartite-Learning

Perform motif discovery on DNA.

Settings

Choose the Data Format:

● **Fasta Format** ○ **Text Format**

Input Sequence Data:

No file chosen

Choose Analysis Methode:

● **Mono-Nucleotide** ○ **DI-Nucleotide** ○ **Learning MONO** ○ **Learning DI**

□ **Just Show Web logo**

Do you want more degenerate sites?
○ **Yes** ● **No**

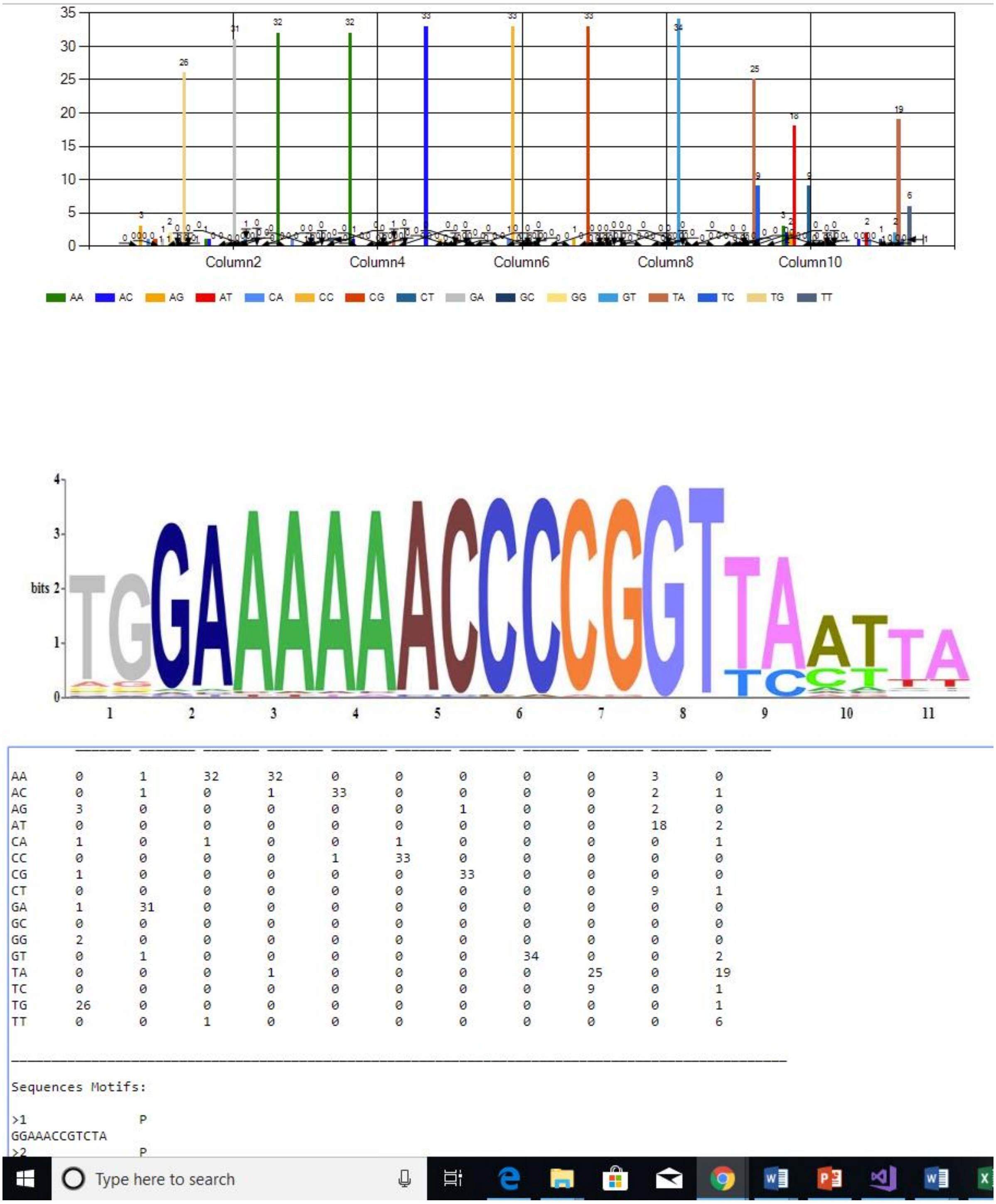
Left Motif Length: 1 ▾ Right Motif Length: 0 ▾

Minimum Gap: 0 ▾ Maximum Gap: 0 ▾

Number of time trying to repeat process 30 ▾

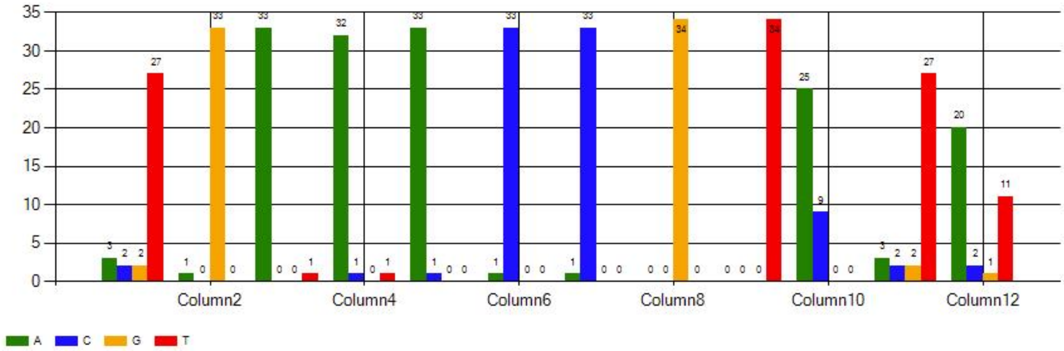
Can motif sites be on both stands? □ **Search given stand only**

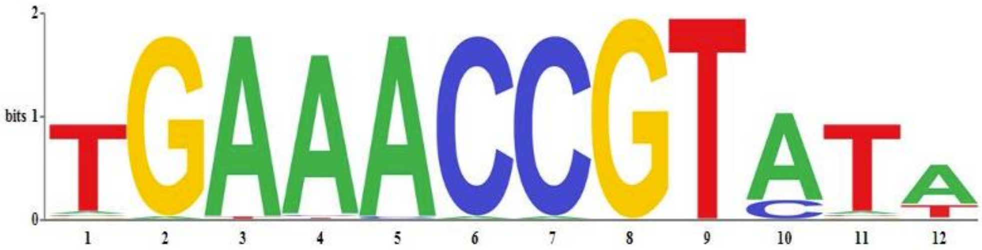
Figure 15. DIpartite-Learning web service schema. The web base of DIpartite is user-friendly and easy access. Users easily can upload input data file and set parameters and finally receive the text or graphically formats as the results.



```
        ______  ______  ______  ______  ______  ______  ______  ______  ______  ______  ______

AA      0       1       32      32      0       0       0       0       0       3       0
AC      0       1       0       1       33      0       0       0       0       2       1
AG      3       0       0       0       0       0       1       0       0       2       0
AT      0       0       0       0       0       0       0       0       0       18      2
CA      1       0       1       0       0       1       0       0       0       0       1
CC      0       0       0       0       1       33      0       0       0       0       0
CG      1       0       0       0       0       0       33      0       0       0       0
CT      0       0       0       0       0       0       0       0       0       9       1
GA      1       31      0       0       0       0       0       0       0       0       0
GC      0       0       0       0       0       0       0       0       0       0       0
GG      2       0       0       0       0       0       0       0       0       0       0
GT      0       1       0       0       0       0       0       34      0       0       2
TA      0       0       0       1       0       0       0       0       25      0       19
TC      0       0       0       0       0       0       0       0       9       0       1
TG      26      0       0       0       0       0       0       0       0       0       1
TT      0       0       1       0       0       0       0       0       0       0       6

____________________________________________________________________________________________

Sequences Motifs:

>1              P
GGAAACCGTCTA
>2             P
```

Figure 16. DWM result on SigmaB dataset by set the 6bp left motif length, 6bp right motif length, 12bp~18bp gap range between two block motifs, and set delft other parameters.





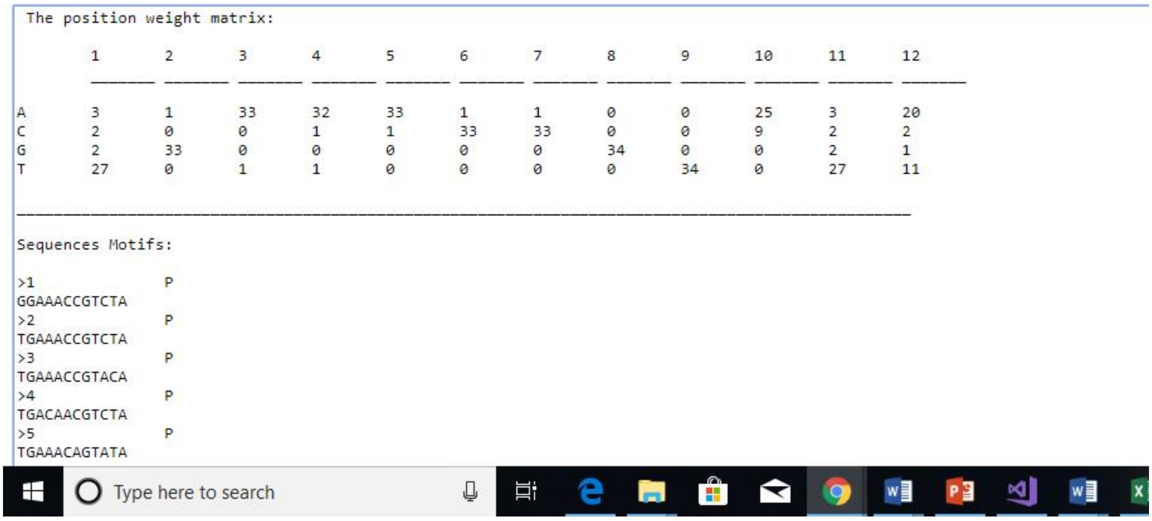
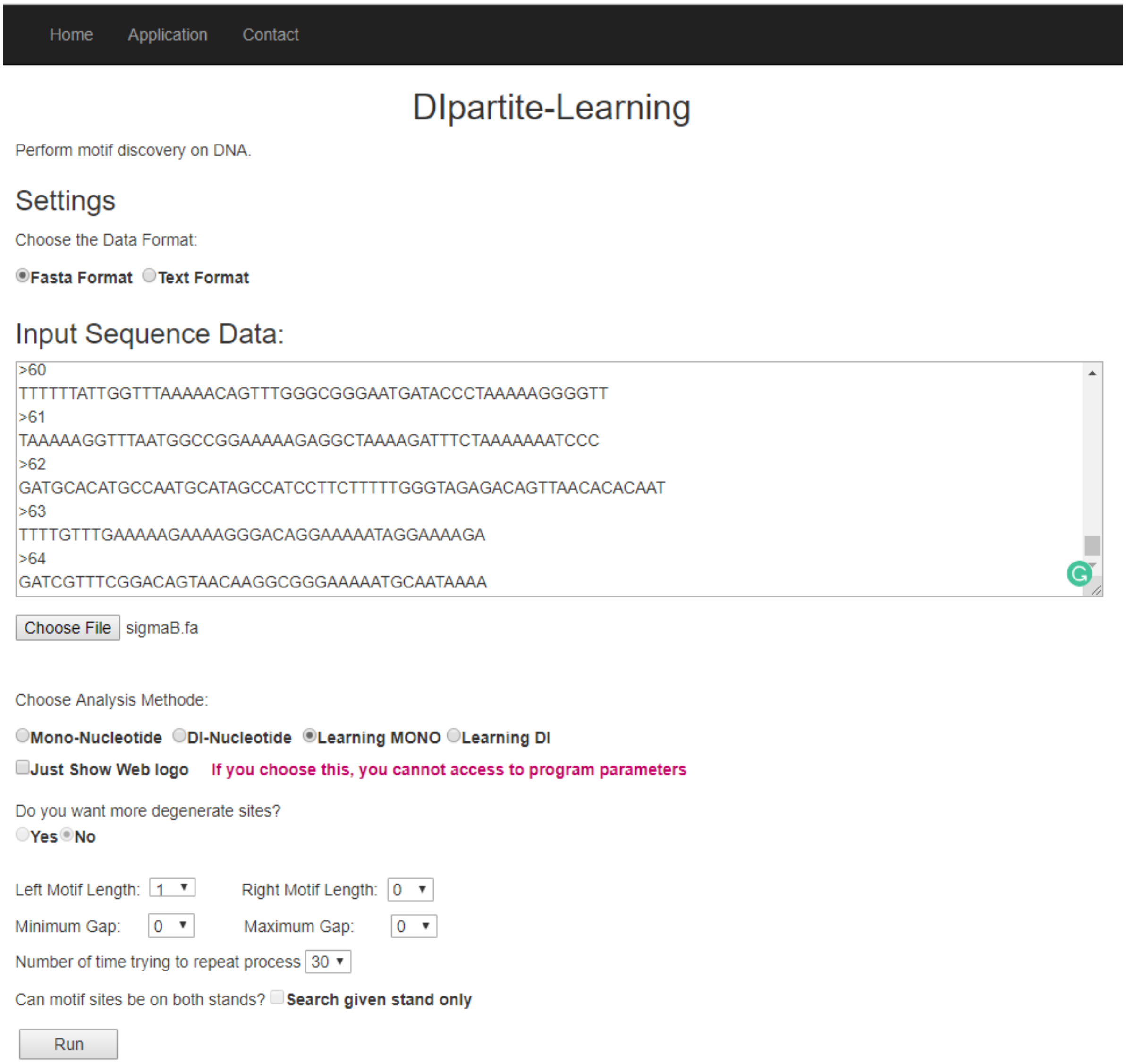

```
 The position weight matrix:

       1       2       3       4       5       6       7       8       9       10      11      12
       _______ _______ _______ _______ _______ _______ _______ _______ _______ _______ _______ _______

A      3       1       33      32      33      1       1       0       0       25      3       20
C      2       0       0       1       1       33      33      0       0       9       2       2
G      2       33      0       0       0       0       0       34      0       0       2       1
T      27      0       1       1       0       0       0       0       34      0       27      11

__________________________________________________________________________________________

Sequences Motifs:

>1                P
GGAAACCGTCTA
>2                P
TGAAACCGTCTA
>3                P
TGAAACCGTACA
>4                P
TGACAACGTCTA
>5                P
TGAAACAGTATA
```

Figure 17. PWM result on SigmaB dataset by set the 6bp left motif length, 6bp right motif length, 12bp~18bp gap range between two block motifs, and set delft other parameters.



DIpartite-Learning

Perform motif discovery on DNA.

Settings

Choose the Data Format:

● Fasta Format ○ Text Format

Input Sequence Data:

```
>60
TTTTTTATTGGTTTAAAAACAGTTTGGGCGGGAATGATACCCTAAAAAGGGGTT
>61
TAAAAAGGTTTAATGGCCGGAAAAAGAGGCTAAAAGATTTCTAAAAAAATCCC
>62
GATGCACATGCCAATGCATAGCCATCCTTCTTTTTGGGTAGAGACAGTTAACACACAAT
>63
TTTTGTTTGAAAAAGAAAAGGGACAGGAAAAATAGGAAAAGA
>64
GATCGTTTCGGACAGTAACAAGGCGGGAAAAATGCAATAAAA
```

sigmaB.fa

Choose Analysis Methode:

○ Mono-Nucleotide ○ DI-Nucleotide ● Learning MONO ○ Learning DI

□ Just Show Web logo If you choose this, you cannot access to program parameters

Do you want more degenerate sites?

○ Yes ● No

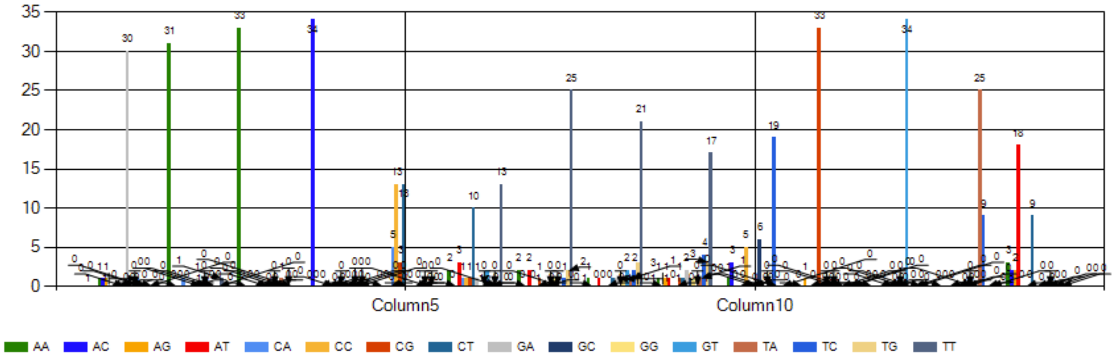
Left Motif Length: 1 ▾ Right Motif Length: 0 ▾

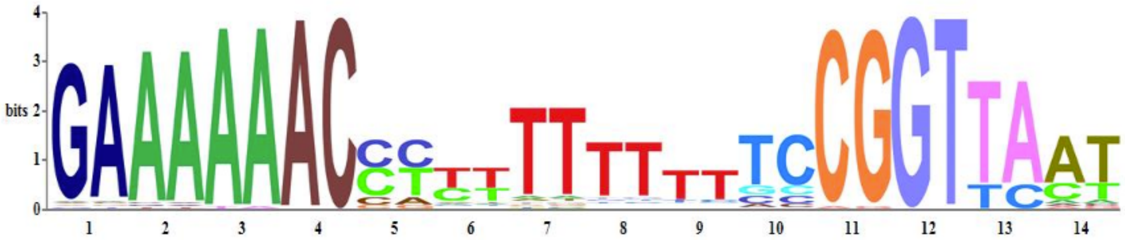
Minimum Gap: 0 ▾ Maximum Gap: 0 ▾

Number of time trying to repeat process 30 ▾

Can motif sites be on both stands? □ Search given stand only

Figure 18. DIpartite-Learning method, in this format, user does not need set parameter just input dataset and select PWM (learning MONO) or DWM (learning DI) structure. DIpartite-Learning set automatically best motifs length and gap range.





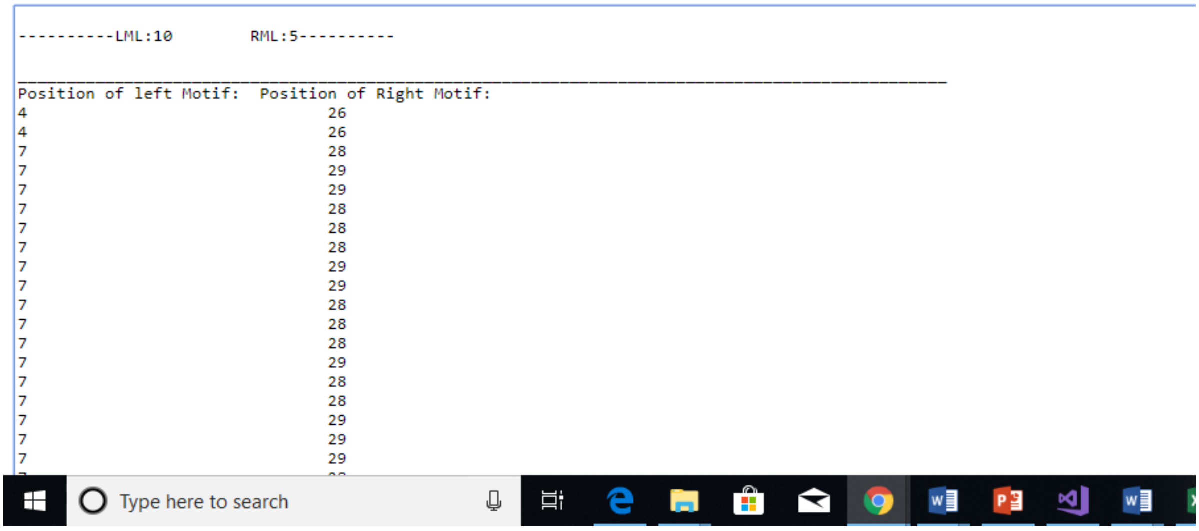
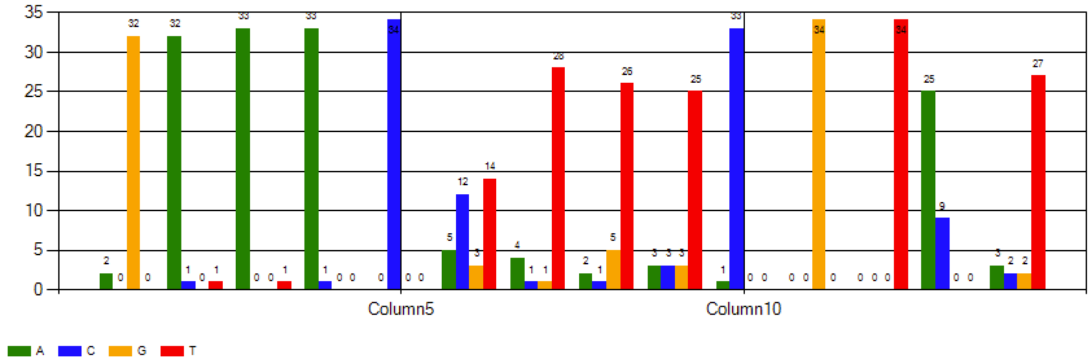
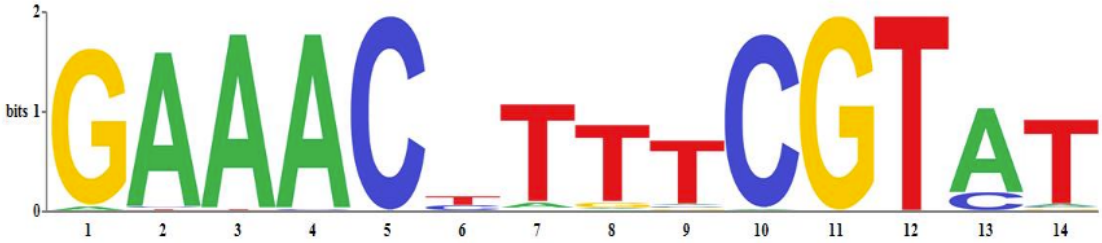

```
----------LML:10        RML:5----------

____________________________________________________________________________________________________
Position of left Motif:  Position of Right Motif:
4                                26
4                                26
7                                28
7                                29
7                                29
7                                28
7                                28
7                                28
7                                29
7                                29
7                                28
7                                28
7                                28
7                                29
7                                28
7                                28
7                                29
7                                29
7                                29
```

Figure 19. DIpartite-Leaning DI (DWM) method of result on SigmaB dataset without any set parameters.





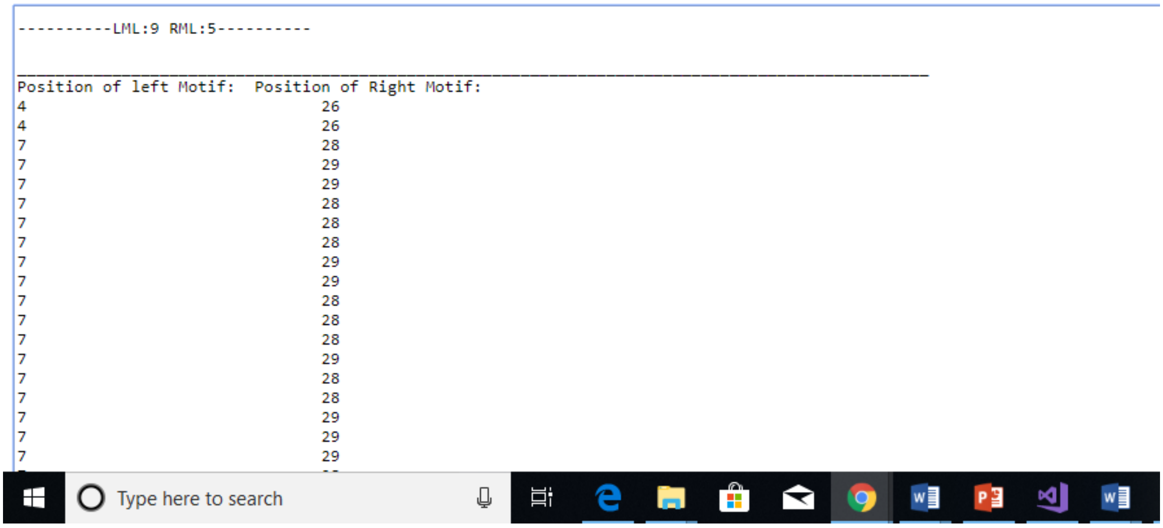

```
----------LML:9 RML:5----------

________________________________________________________________________________
Position of left Motif:  Position of Right Motif:
4                                26
4                                26
7                                28
7                                29
7                                29
7                                28
7                                28
7                                28
7                                29
7                                29
7                                28
7                                28
7                                28
7                                29
7                                28
7                                28
7                                29
7                                29
7                                29
```

Figure 20. DIpartite-Leaning MONO (PWM) method of result on SigmaB dataset without any set parameters.

References

1. Boeva V. Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. Front Genet. 2016;7:24.
2. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316(5830):1497-502.
3. Shultzaberger RK, Bucheimer RE, Rudd KE, Schneider TD. Anatomy of Escherichia coli ribosome binding sites. J Mol Biol. 2001;313(1):215-28.
4. Bi C, Leeder JS, Vyhlidal CA. A comparative study on computational two-block motif detection: algorithms and applications. Mol Pharm. 2008;5(1):3-16.
5. Haldenwang WG. The sigma factors of Bacillus subtilis. Microbiol Rev. 1995;59(1):1-30.
6. Moran CP Jr., Lang N, LeGrice SF, Lee G, Stephens M, Sonenshein AL, et al. Nucleotide sequences that signal the initiation of transcription and translation in Bacillus subtilis. Mol Gen Genet. 1982;186(3):339-46.
7. Makita Y, Nakao M, Ogasawara N, Nakai K. DBTBS: database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics. Nucleic Acids Res. 2004;32(Database issue):D75-7.
8. Baichoo N, Helmann JD. Recognition of DNA by Fur: a reinterpretation of the Fur box consensus sequence. J Bacteriol. 2002;184(21):5826-32.
9. Chumsakul O, Anantsri DP, Quirke T, Oshima T, Nakamura K, Ishikawa S, et al. Genome-Wide Analysis of ResD, NsrR, and Fur Binding in Bacillus subtilis during Anaerobic Fermentative Growth by In Vivo Footprinting. J Bacteriol. 2017;199(13).
10. Chumsakul O, Takahashi H, Oshima T, Hishimoto T, Kanaya S, Ogasawara N, et al. Genome-wide binding profiles of the Bacillus subtilis transition state regulator AbrB and its homolog Abh reveals their interactive role in transcriptional regulation. Nucleic Acids Res. 2011;39(2):414-28.
11. Strauch MA. In vitro binding affinity of the Bacillus subtilis AbrB protein to six different DNA target regions. J Bacteriol. 1995;177(15):4532-6.

12. Xu K, Strauch MA. Identification, sequence, and expression of the gene encoding gamma-glutamyltranspeptidase in Bacillus subtilis. J Bacteriol. 1996;178(14):4319-22.

13. Chen CY, Tsai HK, Hsu CM, May Chen MJ, Hung HG, Huang GT, et al. Discovering gapped binding sites of yeast transcription factors. Proc Natl Acad Sci U S A. 2008;105(7):2527-32.

14. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 2017;46(D1):D260-D266.

15. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature. 2005;434(7031):338-45.

16. Handschin C, Meyer UA. Induction of drug metabolism: the role of nuclear receptors. Pharmacol Rev. 2003;55(4):649-73.

17. GuhaThakurta D, Stormo GD. Identifying target sites for cooperatively binding factors. Bioinformatics. 2001;17(7):608-21.

18. Liu X, Brutlag DL, Liu JS. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput 2001;127-38.

19. Bi C, Rogan PK. Bipartite pattern discovery by entropy minimization-based multiple local alignment. Nucleic Acids Res. 2004;32(17):4979-91.

20. Lu R, Mucaki EJ, Rogan PK. Discovery and validation of information theory-based transcription factor and cofactor binding site motifs. Nucleic Acids Res. 2017;45(5):e27.

21. Shi J, Yang W, Chen M, Du Y, Zhang J, Wang K. AMD, an automated motif discovery tool using stepwise refinement of gapped consensuses. PLoS One. 2011;6(9):e24576.

22. Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics. 1999;15(7-8):563-77.

23. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science. 1993;262(5131):208-14.

24. Bailey TL, Elkany C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994;2:28–36.

25. Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000;16(1):16-23.

26. Luscombe NM, Laskowski RA, Thornton JM. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res. 2001;29(13):2860-74.

27. Zhao Y, Ruan S, Pandey M, Stormo GD. Improved models for transcription factor binding site identification using nonindependent interactions. Genetics. 2012;191(3):781-90.

28. Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, et al. Evaluation of methods for modeling transcription factor sequence specificity. Nat Biotechnol. 2013;31(2):126-34.

29. Salama RA, Stekel DJ. Inclusion of neighboring base interdependencies substantially improves genome-wide prokaryotic transcription factor binding site prediction. Nucleic Acids Res. 2010;38(12):e135.

30. Siddharthan R. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. PLoS One. 2010;5(3):e9722.

31. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, et al. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol. 2005;23(1):137-44.

32. Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS, et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res. 2016;44(D1):D133-43.

33. Martínez-Antonio A, Collado-Vides J. Identifying global regulators in transcriptional regulatory networks in bacteria. Curr Opin Microbiol. 2003;6(5):482-9.

34. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator. Genome Res. 2004;14:1188-90.

35. Jensen ST, Liu JS. BioOptimizer: a Bayesian scoring function approach to motif discovery. Bioinformatics. 2004;20(10):1557-64.

36. Bembom O, Keles S, and Laan M. Supervised detection of conservedmotifs in DNA sequences with cosmo. Statistical Applications in Genetics and Molecular Biology, 2007; 6(1).

37. Keles S, Laan M, Dudoit S, Xing B, and Eisen MB. Supervised detection of regulatory motifs in DNA sequences. Statistical Applications in Genetics and Molecular Biology, 2003; 2(1).

38. Bailey TL. DREME: Motif discovery in transcription factor ChIP-seq data. Bioinformatics, 2011; 27(12):1653–1659.

39. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, and Noble WS. MEME Suite: tools for motif discovery and searching. Nucleic Acids Research, 2009; 37(suppl 2):W202–W208.

40. Bailey TL and Elkan C. The value of prior knowledge in discovering motifs with MEME. In Proceedings of the 3rd International Conference on Intelligent Systems for Molecular Biology, 1995; 21–29.

41. Bailey TL and Elkan C. Unsupervised learning of multiple motifs in biopolymers using Expectation Maximization. Machine Learning, 1995; 21:51–80.

42. Bailey TL and Gribskov M. Combining evidence using p-values: application to sequence homology searches. Bioinformatics, 1998; 14(1):48–54.

43. Bengio Y. Learning deep architectures for AI. Found. Trends Mach Learn. 2009; 2(1), 1–127.

44. Park JH, Kim DH, Kim SS, Lee DJ, and Chun MG. C-ANFIS based fault diagnosis for voltage-fed PWM motor drive systems, 2004; IEEE NAFIPS proc, 379-383.

45. Gyuon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines Machine Learning,2002; 46:389-422.

46. Jung SH. Sample size for FDR-control in microarray data analysis. Bioinformatics, 2005; 21:3097-3104.

47. Park CY, Koo JY, Kim S, Sohn I, Lee JW. Classification of gene functions using SVM for time-course gene expression data. Computational Statistics and Data Analysis, 2008; 52:2578-2587.

48. Lin HT, Lin CJ, and Weng RC.A noteon Platt’s probabilistic outputs for support vector machines. machine learning. Mach. Learn. 2003; 68:267-276.

49. Sonnenburg S, Zien A, Philips P. Ratsch G. POIMs: positional oligomer importance matrices understanding Support vector machine-based signal detectors. Bioinformatics, 2008; 24:i6-i14.

50. Platt JC. Probablistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola A, Bartlett P, Scholkopf B, and Schuurmans D. (eds). Advances in Large Margin Classifers. MIT Press, Cambridge, Mass. 1999; 67-74.

51. Karchin R, Karplus K, Haussler D. Classifying G-protein coupled receptors with Support vector machines. Bioinformatics, 2002; 18:147-159.

52. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 2000; 16:906-914.

PLoS One,
14(8): e0220207, Published: August 30, 2019
DOI: 10.1371/journal.pone.0220207

Computational Biology and Chemistry,
80, 168-176, Published: March 17, 2019
DOI: 10.1016/j.compbiolchem.2019.03.007