PLOS ONE

RESEARCH ARTICLE

DIpartite: A tool for detecting bipartite motifs by considering base interdependencies

Mohammad Vahed[1], **Jun-ichi Ishihara**[1], **Hiroki Takahashi**[1,2]*

1 Medical Mycology Research Center, Chiba University, Chiba, Japan, **2** Molecular Chirality Research Center, Chiba University, Chiba, Japan

* hiroki.takahashi@chiba-u.jp

Citation: Vahed M, Ishihara J-i, Takahashi H (2019) DIpartite: A tool for detecting bipartite motifs by considering base interdependencies. PLoS ONE 14(8): e0220207. https://doi.org/10.1371/journal.pone.0220207

Editor: Jun-Tao Guo, University of North Carolina at Charlotte, UNITED STATES

Received: April 1, 2019

Accepted: July 10, 2019

Published: August 30, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: This work was partly supported by MEXT KAKENHI (16K18671) to HT, AMED under Grant Number JP19fm0208024 to HT, the Tenure Tracking System Program of MEXT to HT, the Institute for Global Prominent Research, Chiba University, to HT, and MEXT KAKENHI (16H06279) to HT and JI. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

It is extremely important to identify transcription factor binding sites (TFBSs). Some TFBSs are proposed to be bipartite motifs known as two-block motifs separated by gap sequences with variable lengths. While position weight matrix (PWM) is commonly used for the representation and prediction of TFBSs, dinucleotide weight matrix (DWM) enables expression of the interdependencies of neighboring bases. By incorporating DWM into the detection of bipartite motifs, we 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 the minimization of Shannon's entropy. DIpartite predicts the bipartite motifs by considering the interdependencies of neighboring positions, that is, DWM. We compared DIpartite with other available alternatives by using test datasets, namely, of CRP in *E. coli*, sigma factors in *B. subtilis*, and promoter sequences in humans. We have developed DIpartite for the detection of TFBSs, particularly bipartite motifs. DIpartite enables *ab initio* prediction of conserved motifs based on not only PWM, but also DWM. We evaluated the performance of DIpartite by comparing it with freely available tools, such as MEME, BioProspector, BiPad, and AMD. Taken the obtained findings together, DIpartite performs equivalently to or better than these other tools, especially for detecting bipartite motifs with variable gaps. DIpartite requires users to specify the motif lengths, gap length, and PWM or DWM. DIpartite is available for use at https://github.com/Mohammad-Vahed/DIpartite.

Introduction

Gene expression is often regulated by transcription factors (TFs). TFs bind to specific DNA-binding sites and modulate the expression of genes. Therefore, to understand transcriptional regulations, given its complexity, it is extremely important to make accurate inferences about transcription factor binding sites (TFBSs). High-throughput ChIP-seq, which is widely used to study TF–DNA interactions, provides the sequences of binding regions [1,2]. TFBSs can be determined as the most over-represented motif in a given set of DNA sequences.

Bipartite motifs are defined as extensions of one-block TFBSs, that is, two conserved motifs separated by variable gaps. Several different types of bipartite motifs have been proposed in both prokaryotes and eukaryotes [3,4]. Shultzaberger et al. (2001) proposed the bipartite

Competing interests: The authors have declared that no competing interests exist.

model of ribosome binding sites, in which they are composed of a Shine–Dalgarno sequence and an initiation region in *Escherichia coli* [3]. In *Bacillus subtilis*, the principal sigma factor in vegetative growth, SigA, binds to the bipartite motif separated by variable gaps, TGA-CA<spacer>TATAAT [5–7]. Baichoo and Helmann (2002) determined the bipartite motif, TGATAAT<spacer>ATTATCA, of the ferric uptake repressor Fur [8,9]. It has been reported that the global regulator AbrB can recognize bipartite motifs [10–12]. As in the case of eukaryotes, the existence of bipartite motifs of yeast TFs, such as ABF1 and GAL4, has been confirmed [13,14]. It has been reported that around 30% of the promoter sequences contain bipartite motifs with constant gaps in humans [15]. The level of conservation of the motif M4 (ACTAYRNNNCCCR) was reported to be much higher than those for most known motifs. Similarly, the TFs CAR and RXR bind to bipartite motifs in humans [4]. Thus, it is conceivable that TFs work in a cooperative manner and recognize bipartite motifs to regulate gene expression [16,17]. Several tools such as BioProspector [18], BiPad [19,20], and AMD [21] are available for the *ab initio* prediction of bipartite motifs for a set of DNA sequences, while many tools have been developed for the prediction of one-block TFBSs, such as Consensus [22], Gibbs Sampler [23], and MEME [24]. BioProspector based on Gibbs sampling [18] and BiPad based on the entropy minimization method [19,20] enable the identification of bipartite motifs with variable gaps. AMD identifies bipartite motifs with constant gaps by comparing the target sequences with the background sequences regardless of whether the motifs are long or short, gapped or contiguous [21].

Position weight matrices (PWMs) are commonly used to find and represent TFBSs [25]. They 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 TF–DNA interactions. For example, a single amino acid interacts with multiple bases simultaneously [26]. Zhao et al. (2012) clearly showed the existence of dinucleotide dependency in TFs [27,28]. Indeed, PWMs perform well in modeling TFBS properties, but are inadequate for considering position interdependencies. There are interdependencies 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. (2013) observed an improvement of performance of motif detection upon incorporating dinucleotide interactions [28]. Although BioProspector and BiPad predict bipartite motifs, they are based on the assumption of independencies among bases, namely, PWM.

Here, we present a novel bipartite motif detection tool, DIpartite (bi**partite** motif detection tool based on **di**nucleotide weight matrix). DIpartite predicts the bipartite motif by considering interdependencies of neighboring positions, namely, DWM. We compared DIpartite with other available alternatives by using test datasets from prokaryote and eukaryote, namely, of CRP in *E. coli*, sigma factors in *B. subtilis*, and promoter motifs in humans.

Materials and methods

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 (S1 Fig). Since it is reported that the bipartite motif represents well by Shannon's entropy [3,19,20], we 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.

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. We set the objective function to minimize Shannon's entropy for PWM or DWM of the concatenated motif of the left and right motifs, in Eq 1:

$$\hat{M}_{LR} = \mathrm{argmin}_{M_{LR}}(IC_{M_{LR}}) \tag{1}$$

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

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

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:

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

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

where N is the total number of input sequences. $f_i(x)$ is the frequency of x at the position i, that is, the mononucleotide at position i for PWM, or the dinucleotide at position $i-1$ and i for DWM. k is the number of the patterns, that is, $k = 4$ for PWM or $k = 16$ for DWM. n is the total number of the mononucleotides for PWM or dinucleotides for DWM that are not located at the motif sites. β is the total pseudo-count. $g(x)$ is the frequency of x in the background sites. We set $\beta = 1$.

Overview of the algorithm

The algorithm of DIpartite works through an iterative process of calculating entropy. DIpartite is implemented in C++ and available under the CNU v3 license. 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), ZOOPS (Zero or one bipartite occurrence per sequence), or ANR (any number of repetitions).

Performance evaluation

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

$$nCC = \frac{nTP \times nTN - nFN \times nFP}{\sqrt{(nTP + nFN)(nTN + nFP)(nTP + nFP)(nTN + nFN)}} \tag{5}$$

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. We adopted the combined *nCC* by adding *nTP*, *nFN*, *nFP*, and *nTN* over the data sets.

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).

Promoter motifs in human

Xie et al. [15] proposed the 1,460 motifs in human. We sought the motifs with the gap lengths greater than or equal to the lengths of left and right motifs. Among of them, we 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.

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).

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 1000". AMD was executed with the option "-T 2" for two sigma datasets, i.e., σ^E and σ^F. We 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), and the 200 bp upstream regions of 4,448 genes in *B. subtilis* 168 (NC_000964.3).

Results

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 (Fig 1A). We 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 (Fig 1B). In addition, we 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.

Performance for CRP dataset

We 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]. We 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 (Fig 2A). 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. We 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 (S2A Fig). In addition, we tested the running time by using the CRP dataset. Although BioProspector was the fastest software among tested software, DIpartite was comparable with BiPad (S3 Fig).

For the bipartite motif, we compared BioProspector, BiPad, DIpartite PWM, and DIpartite DWM (Fig 2B). 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. We conducted the performance comparison by using 100 datasets with 100 sequences (S2B Fig). 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 Eq (2). Taking the findings together, DIpartite successfully detected the binding sites of the one-block or bipartite motifs.

Performance for human dataset

We selected the human promoter sequences as bipartite motifs with constant gaps in eukaryotes [15]. Of 1,460 motifs, 46 motifs with gaps larger than 4 nt were filtered. The promoter sequences around the positions of each motif (500 bp upstream to 500 bp downstream) were retrieved as the target sets. Since AMD did not detect any motifs for six motifs, namely, RGG ANNNNNAKTCC (54 sequences), RKCTGNNNNNRMTTA (21 sequences), TTGRNNNN NNTCCAR (21 sequences), YMATCNNNNNGCGM (50 sequences), YTGGANNNNNNY

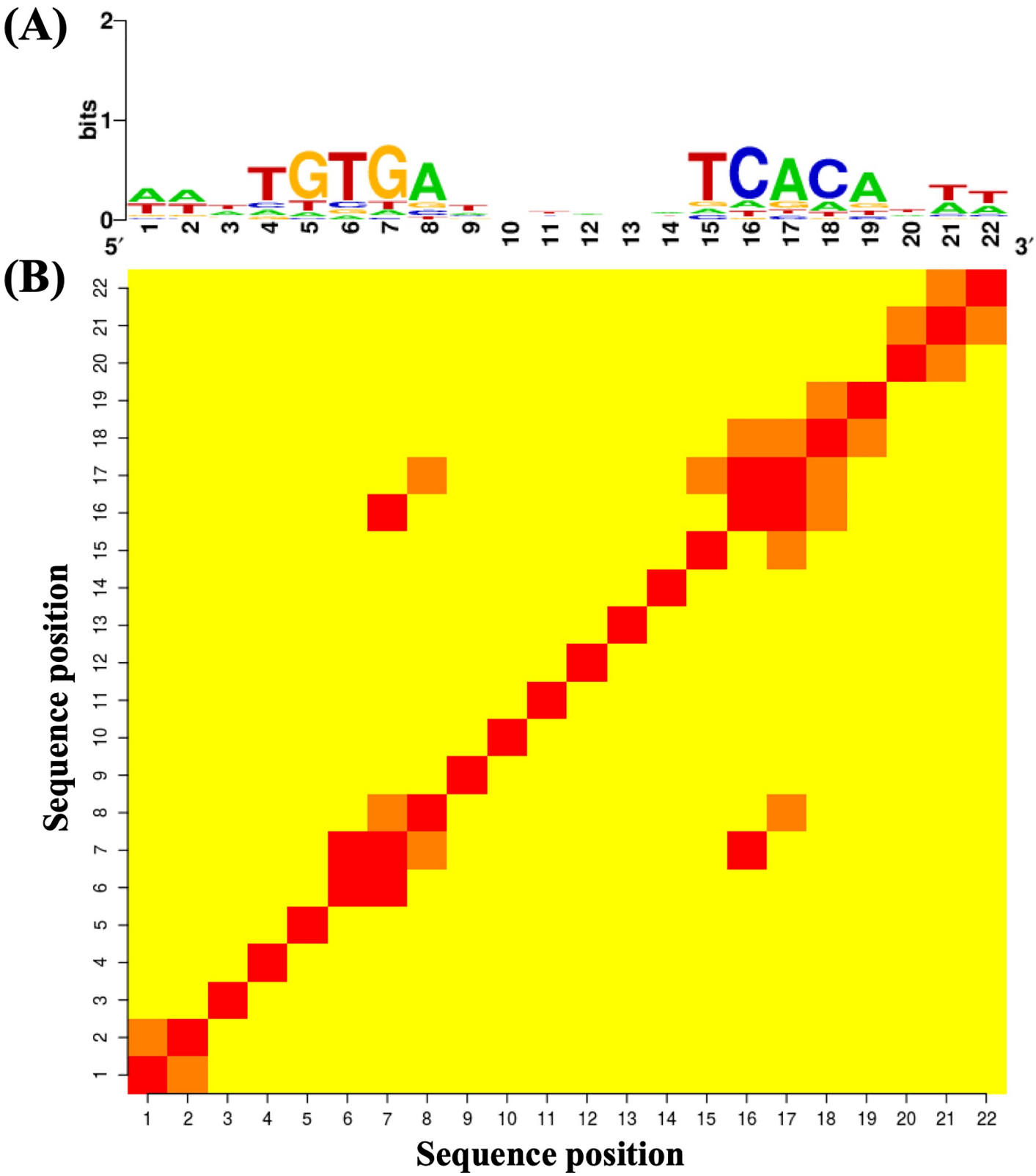

Fig 1. Sequence logo and heat map of CRP. Out of 374 CRP motifs, 308 sequences with the 22-bp motif were used. (A) Sequence logo for CRP using 308 sequences [34]. (B) Heat map of CRP.

https://doi.org/10.1371/journal.pone.0220207.g001

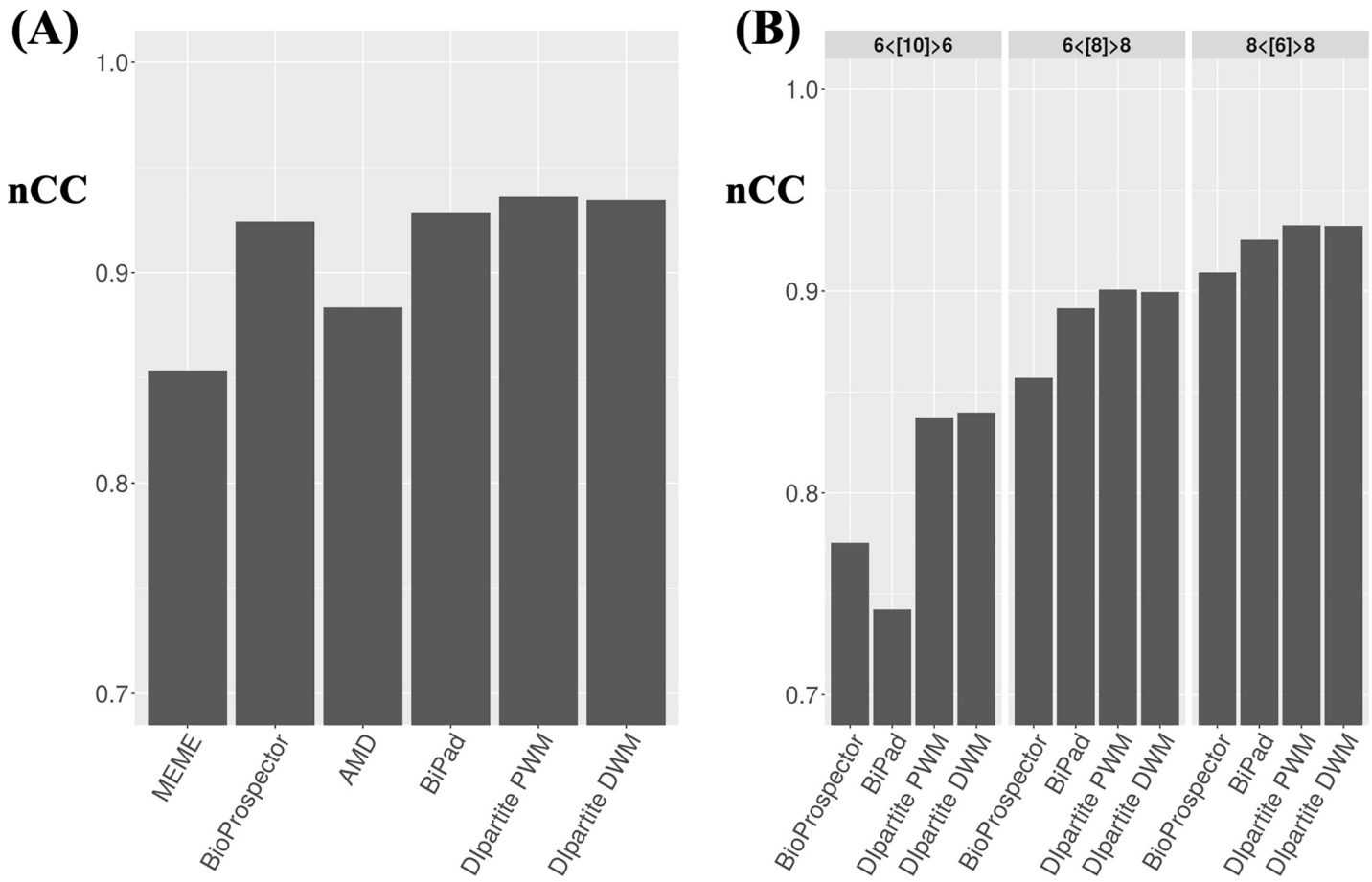

Fig 2. The performance comparison for 323 CRP sequences. The combined *nCC* values were plotted. (A) Summary of the results for searching the one-block motif, i.e., the 22 bp motif, by MEME, BioProspector, AMD, BiPad, DIpartite PWM and DIpartite DWM. (B) Summary of the results for searching the bipartite motifs, i.e., 6<[10]>6, 6<[8]>8, and 8<[6]>8, by BioProspector, BiPad, DIpartite PWM and DIpartite DWM.

https://doi.org/10.1371/journal.pone.0220207.g002

CAA (26 sequences), and YTTGRNNNNNNGCCNR (50 sequences), these were excluded, and 40 datasets were evaluated for the performance of DIpartite. We assessed the performance for 40 motif datasets (Fig 3A). DIpartite DWM exhibited the highest performance (50%), followed by DIpartite PWM (48%), BioProspector (38%), MEME (20%), BiPad (8%), and AMD (3%) (S1 Table), 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 larger interquartile ranges (Fig 3B), indicating that these tools outperformed DIpartite for particular motifs, but were outperformed by it for the other motifs.

Performance for sigma factor dataset

We compared the performance of DIpartite with those of BioProspector, AMD, and BiPad for bipartite motifs with variable gaps. We adopted the nine bipartite sigma transcription factors in *B. subtilis*, namely, 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 (Fig 4A). DIpartite PWM performed

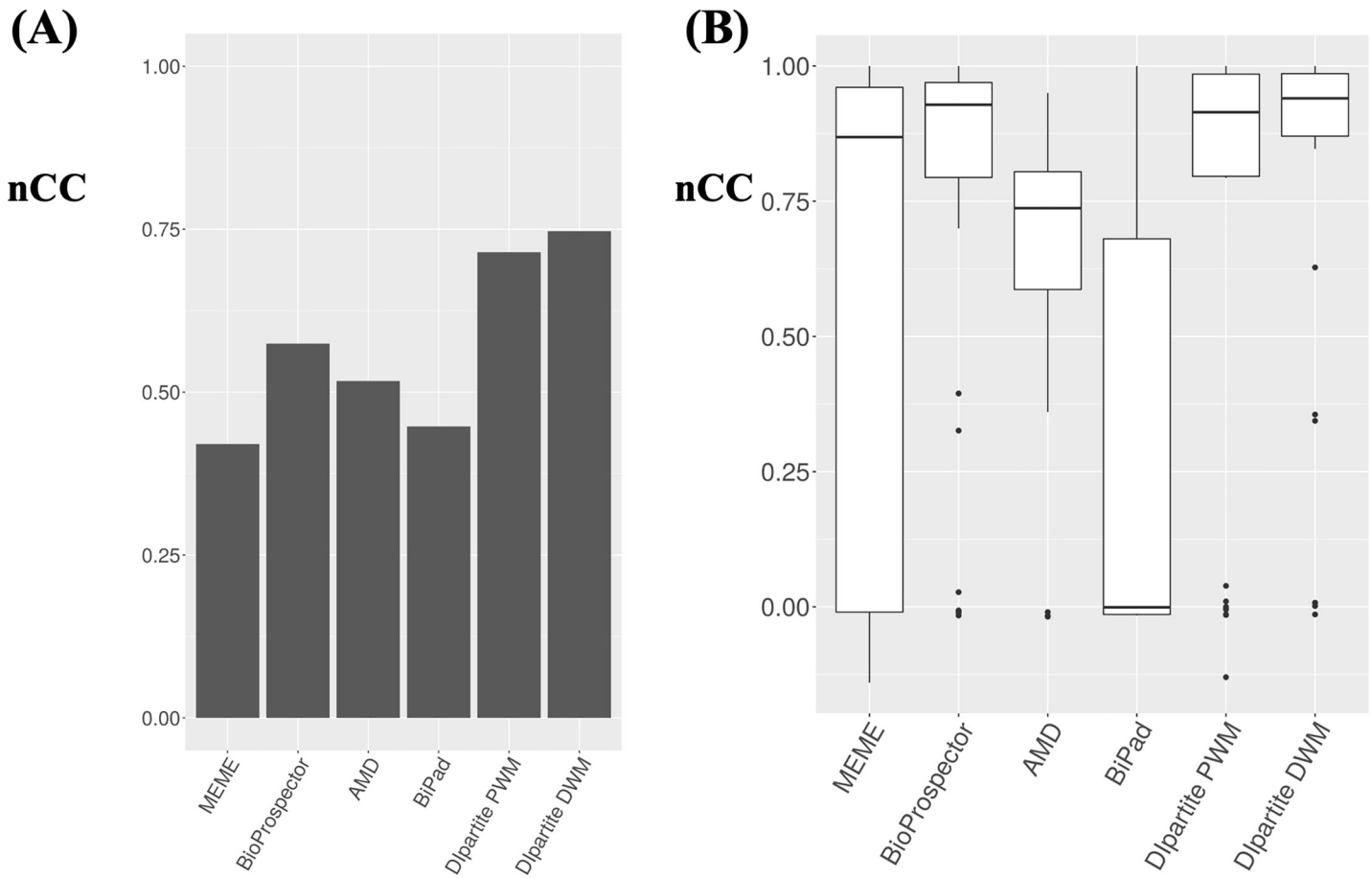

Fig 3. The performance comparison for human promoter datasets. (A) Summary of the results of all 40 human promoter datasets. The combined *nCC* values were calculated by using a total of 3,054 sequences. (B) Boxplots of the *nCC* values for each 40 human promoter datasets. All values are shown in S1 Table.

https://doi.org/10.1371/journal.pone.0220207.g003

better than BioProspector, BiPad, AMD, and DIpartite DWM for six sigma factors, with the exceptions being σ^D, σ^E and σ^H (Fig 4B). While the performance of DIpartite PWM was excellent for two sigma factors (σ^A and σ^F), that of DIpartite DWM was remarkable for four sigma factors (σ^B, σ^G, σ^K, and σ^W). AMD exhibited relatively low *nCC* values for all nine datasets (Fig 4A), unlike the results for human promoter sequences, suggesting that the variable gap lengths could affect its performance. This is reasonable because AMD was developed for detecting bipartite motifs with constant gaps. AMD with the option "-T 1" did not detect any motifs for two sigma datasets, i.e., σ^E and σ^F.

Among four sigma factors with the highest performance coefficients for DIpartite DWM, the *nCC* value for σ^K was greatly improved by DIpartite DWM, namely, to 0.757, indicating the presence of base interdependencies in the motif of σ^K. We observed that the left motif of DIpartite DWM was shifted and "AC" was more over-represented, indicating that the left motif of σ^K might be improved. Position 7 was "T" in all 53 sequences (Fig 5A), consistent with the known motif in DBTBS. Similarly, the highest frequencies of the dinucleotides "AT" and "TA" were observed at positions 6 and 7, and 7 and 8, respectively (Fig 5B).

The *nCC* value of σ^A was greatly improved by DIpartite PWM, namely, to 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 from them (S4 Fig). In particular, DIpartite PWM exhibited the conserved base "T", at position 1. This result

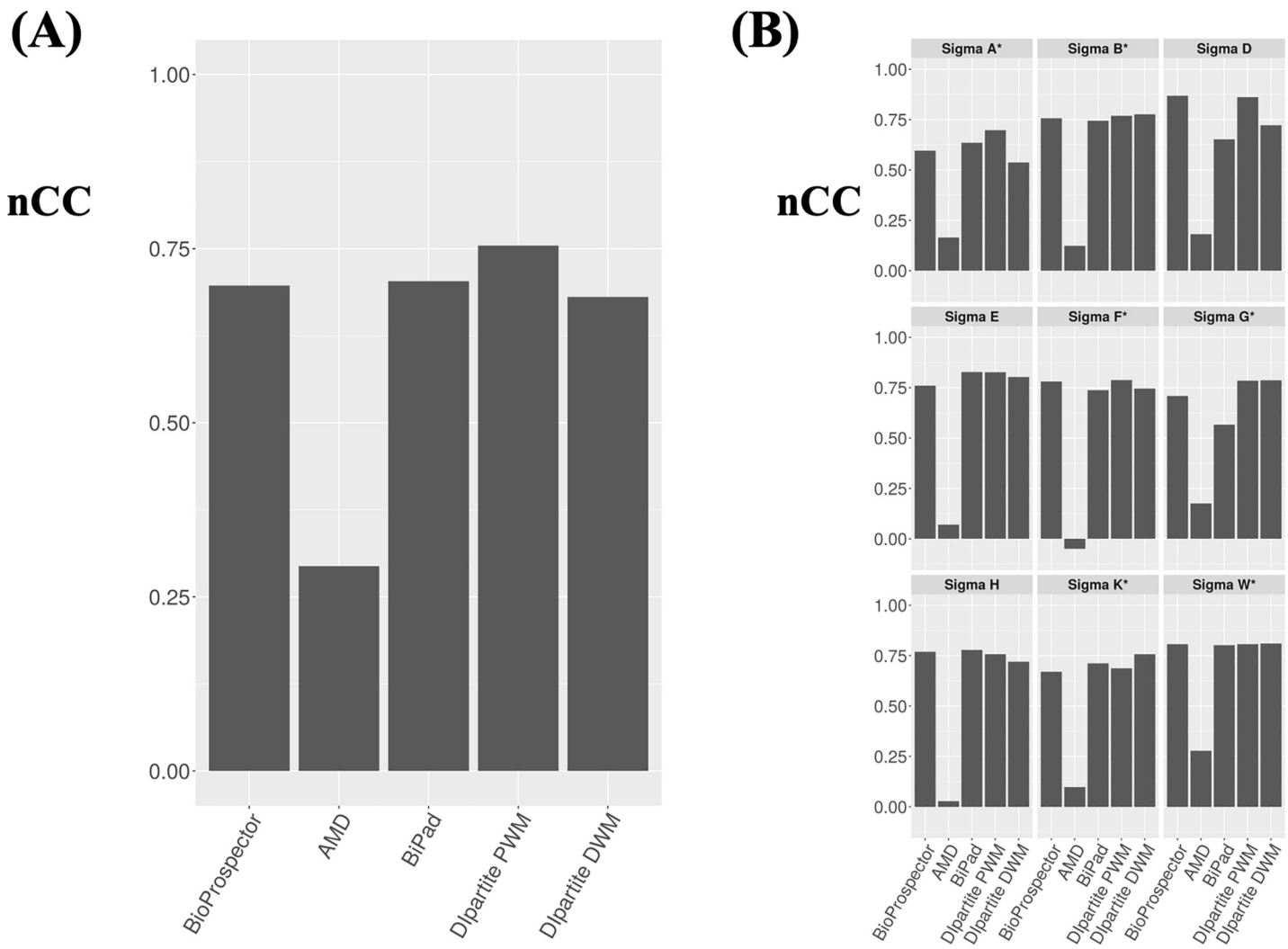

Fig 4. The performance comparison for *B. subtilis* datasets. (A) Summary of the results of all sigma datasets. (B) Summary of the results of each sigma datasets. σ^A, σ^B, σ^D, σ^E, σ^F, σ^G, σ^H, σ^K, and σ^W consist of 344, 64, 30, 70, 25, 55, 25, 53, and 34 sequences, respectively. The asterisks indicate if DIpartite performed better than BioProspector, AMD, and BiPad.

https://doi.org/10.1371/journal.pone.0220207.g004

is consistent with the motif TTGACA<>tgnTATAAT proposed by DBTBS [7]. DIpartite PWM showed the sequences with minimum entropy.

We 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*, we generated 100 datasets for each including 10, 20, 50, 100, 150, 200, and 300 sequences (Fig 6). 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 (Eq 3).

Performance for the dataset with noise sequences

We assessed the performance for the datasets with noise sequences. DIpartite allows the users to search the motifs for the datasets with noise sequences, known as ZOOPS. We evaluated the

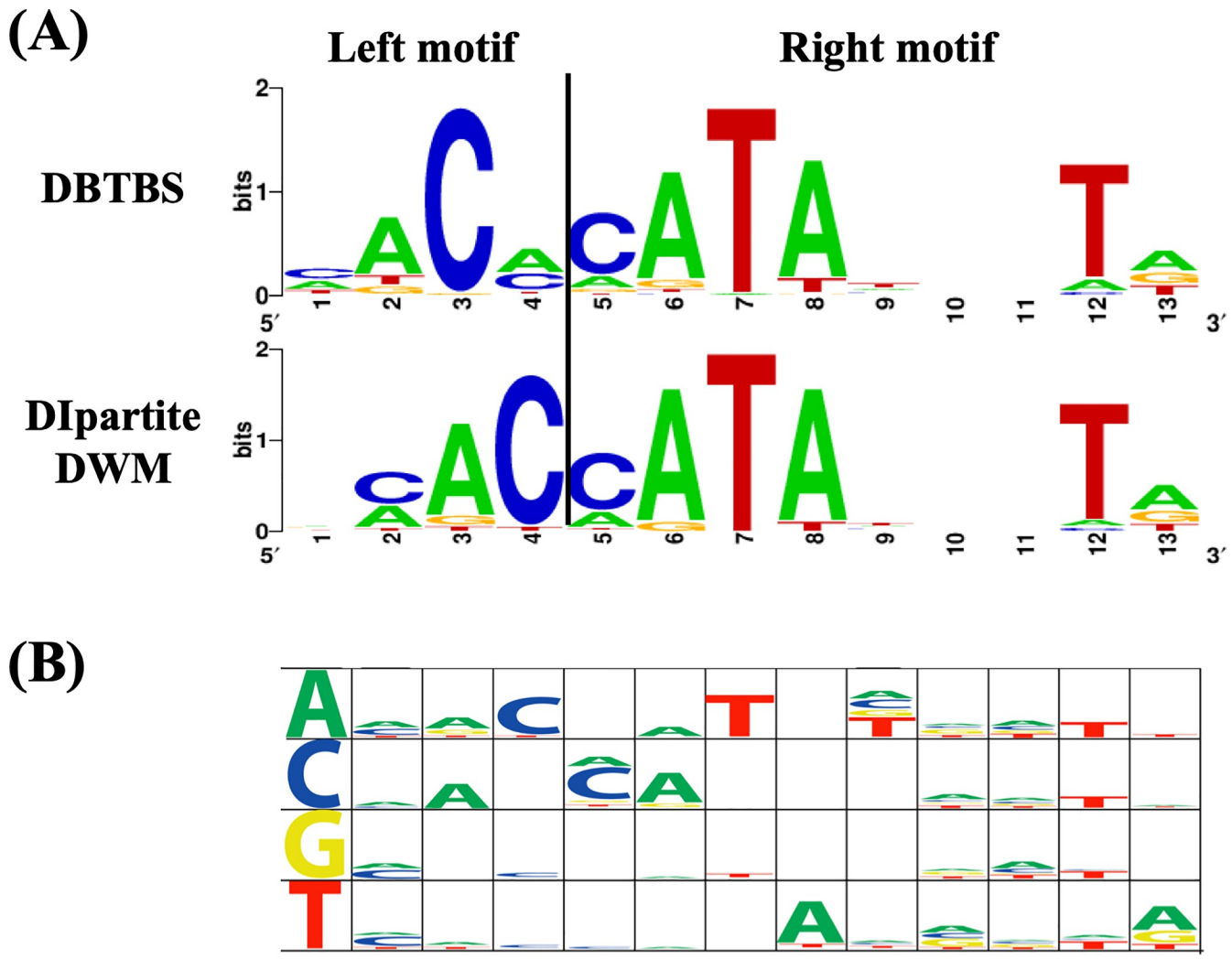

Fig 5. Sequence logo for σ^K by DIpartite DWM. (A) Sequence logos generated by DBTBS and DIpartite DWM. The border between the left and right motifs, i.e., position 4, 5, is indicated as the vertical line. (B) Sequence logo for the probability of each dinucleotides. One base before was depicted in first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

https://doi.org/10.1371/journal.pone.0220207.g005

accuracy of detecting noise sequences by using the datasets with noise sequences. We chose the CRP datasets and human dataset as the test datasets of the one- and two-block motifs. We compared the performance of noise detection by DIpartite with that by MEME for the CRP datasets (Table 1). 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.

Finally, we compared the performance of noise detection for the two-block dataset, i.e., RYAAAKNNNNNNTTGW consisting of 44 sequences (S1 Table). BioProspector ($nCC = 1$) and BiPad ($nCC = 1$) outperformed DIpartite PWM ($nCC = 0.914$). Increasing the noise sequences, BioProspector and BiPad exhibited lower nCC values (Table 2). 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.

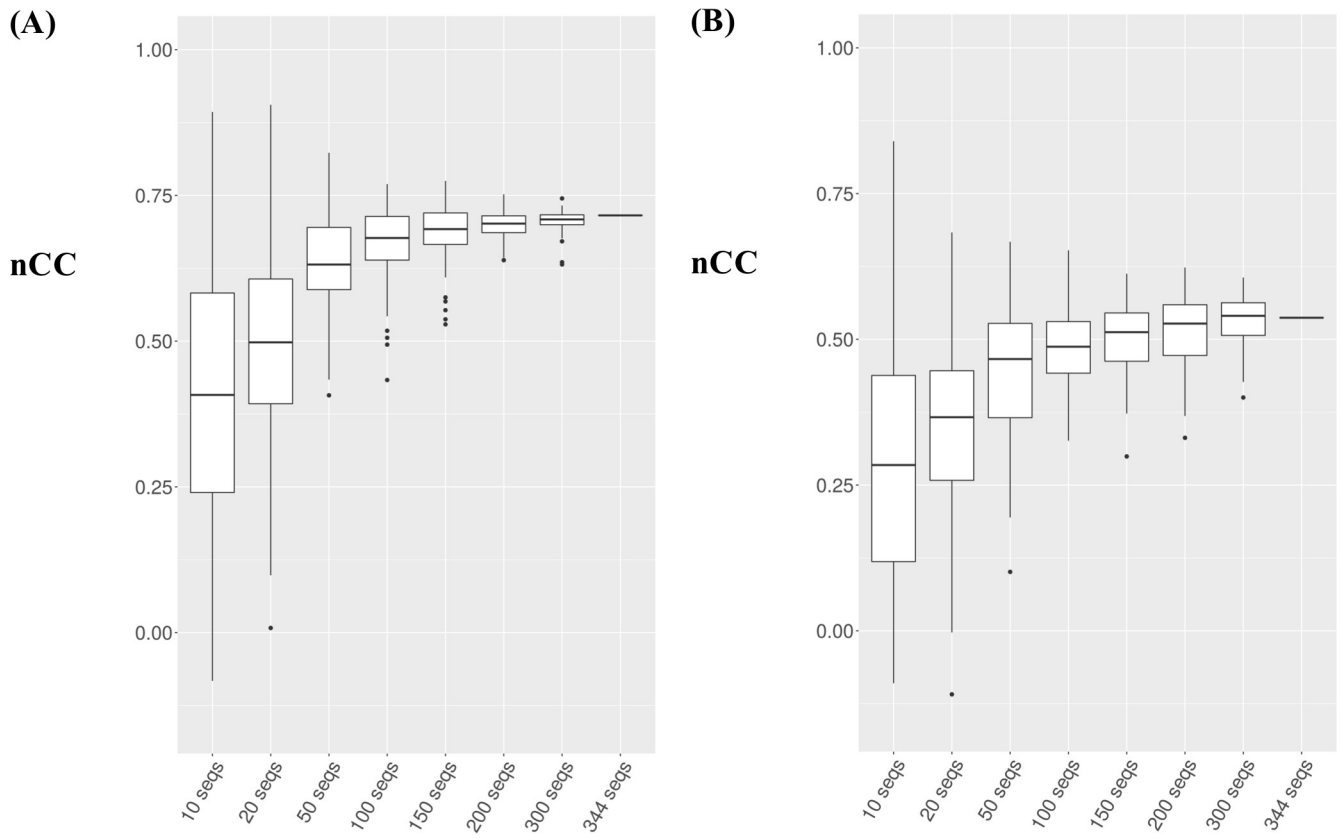

Fig 6. The performance of DIpartite: (A) PWM; (B) DWM. 100 datasets were generated by sampling of the σ^A dataset. The sizes of the dataset were 10, 20, 50, 100, 150, 200, 300 sequences.

https://doi.org/10.1371/journal.pone.0220207.g006

Table 1. The performance of noise detection for the one-block motif.

	MEME			DIpartite		
	CRP_25	CRP_50	CRP_100	CRP_25	CRP_50	CRP_100
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*. CRP_25 consists of 323 CRP and 81 (25%) noise sequences. CRP_50 consists of 323 CRP and 162 (50%) noise sequences. CRP_100 consists of 323 CRP and 323 (100%) noise sequences.
TPR: True positive rate, FPR: False positive rate.

https://doi.org/10.1371/journal.pone.0220207.t001

Table 2. The performance of noise detection for the two-block motif.

	TF_0	TF_25	TF_50	TF_100
BioProspector	1	0.907	−0.16	−0.16
BiPad	1	1	1	−0.16
DIpartite PWM	0.914	1	1	1

The combined *nCC* values were indicated. Noise sequences were sampled from the genome sequence of human. TF_0 consists of 44 RYAAAKNNNNNNTTGW sequences. TF_25 consists of 44 RYAAAKNNNNNNTTGW and 11 noise sequences. TF_50 consists of 44 RYAAAKNNNNNNTTGW and 22 noise sequences. TF_100 consists of 44 RYAAAKNNNNNNTTGW and 44 noise sequences.

https://doi.org/10.1371/journal.pone.0220207.t002

Conclusions

We have developed DIpartite for the detection of TFBSs, consisting of bipartite motifs. DIpartite enables *ab initio* prediction of conserved motifs based on not only PWM, but also DWM. We evaluated the performance of DIpartite compared with freely available tools, namely, MEME, BioProspector, AMD, and BiPad. Both DIpartite PWM and DWM performed equivalently to or better than these alternatives, especially in the case of the bipartite motifs with variable gaps, like for sigma factors in *B. subtilis*. The prediction of σ^K was greatly improved by taking into consideration base interdependencies. DIpartite is available for use at https://github.com/Mohammad-Vahed/DIpartite.

Supporting information

S1 Fig. Flowchart of DIpartite. The input data is the sequence file including *N* sequences. DIpartite proposes bipartite motifs based on PWM or DWM. Each iteration starts from randomly generated positions. The convergence of each iteration is judged by the differences of the entropy, that is, ε. We set $\varepsilon = 10^{-8}$. E_i and E_{i-1} correspond to the *i*th and *i*−1th entropy, i.e., $IC_{M_{LR}}$ (Eq 2), respectively.
(TIFF)

S2 Fig. 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.
(TIFF)

S3 Fig. 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 (designated as Bio), BiPad, DIpartite PWM (designated as PWM), and DIpartite DWM (designated as DWM) were tested.
(TIFF)

S4 Fig. Sequence logos for σ^A from the results of (A) BioProspector, (B) BiPad, (C) DIpatrite PWM, and (D) DIpartite DWM.
(TIFF)

S1 Table. The performance comparison for 40 motifs in human.
(XLSX)

Acknowledgments

We would like to thank Rafik Salam (University of Oxford, UK) and Dov Stekel (University of Nottingham, UK) for their fruitful discussions. MV would like to thank Fujii Medical International Exchanging Foundation for financial support. We also thank Edanz (www.edanzediting.co.jp) for editing the English text of a draft of this manuscript.

Author Contributions

Data curation: Mohammad Vahed, Hiroki Takahashi.

Funding acquisition: Hiroki Takahashi.

Investigation: Mohammad Vahed.

Methodology: Mohammad Vahed, Hiroki Takahashi.

Project administration: Hiroki Takahashi.

Software: Mohammad Vahed, Hiroki Takahashi.

Supervision: Hiroki Takahashi.

Validation: Mohammad Vahed, Jun-ichi Ishihara, Hiroki Takahashi.

Visualization: Mohammad Vahed, Hiroki Takahashi.

Writing – original draft: Mohammad Vahed, Jun-ichi Ishihara, Hiroki Takahashi.

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. https://doi.org/10.3389/fgene.2016.00024 PMID: 26941778
2. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007; 316(5830):1497–502. https://doi.org/10.1126/science.1141319 PMID: 17540862
3. Shultzaberger RK, Bucheimer RE, Rudd KE, Schneider TD. Anatomy of Escherichia coli ribosome binding sites. J Mol Biol. 2001; 313(1):215–28. https://doi.org/10.1006/jmbi.2001.5040 PMID: 11601857
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. https://doi.org/10.1021/mp7001126 PMID: 18076137
5. Haldenwang WG. The sigma factors of Bacillus subtilis. Microbiol Rev. 1995; 59(1):1–30. PMID: 7708009
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. https://doi.org/10.1007/bf00729452 PMID: 6181373
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. https://doi.org/10.1093/nar/gkh074 PMID: 14681362
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. https://doi.org/10.1128/JB.184.21.5826-5832.2002 PMID: 12374814
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. https://doi.org/10.1093/nar/gkq780 PMID: 20817675
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. https://doi.org/10.1128/jb.177.15.4532-4536.1995 PMID: 7635837
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. https://doi.org/10.1128/jb.178.14.4319-4322.1996 PMID: 8763966
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. https://doi.org/10.1073/pnas.0712188105 PMID: 18272477
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. https://doi.org/10.1038/nature03441 PMID: 15735639
16. Handschin C, Meyer UA. Induction of drug metabolism: the role of nuclear receptors. Pharmacol Rev. 2003; 55(4):649–73. https://doi.org/10.1124/pr.55.4.2 PMID: 14657421

17. GuhaThakurta D, Stormo GD. Identifying target sites for cooperatively binding factors. Bioinformatics. 2001; 17(7):608–21. https://doi.org/10.1093/bioinformatics/17.7.608 PMID: 11448879

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. PMID: 11262934

19. Bi C, Rogan PK. Bipartite pattern discovery by entropy minimization-based multiple local alignment. Nucleic Acids Res. 2004; 32(17):4979–91. https://doi.org/10.1093/nar/gkh825 PMID: 15388800

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. https://doi.org/10.1093/nar/gkw1036 PMID: 27899659

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. https://doi.org/10.1371/journal.pone.0024576 PMID: 21931761

22. Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics. 1999; 15(7–8):563–77. https://doi.org/10.1093/bioinformatics/15.7.563 PMID: 10487864

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. https://doi.org/10.1126/science.8211139 PMID: 8211139

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. PMID: 7584402

25. Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000; 16(1):16–23. https://doi.org/10.1093/bioinformatics/16.1.16 PMID: 10812473

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. https://doi.org/10.1093/nar/29.13.2860 PMID: 11433033

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. https://doi.org/10.1534/genetics.112.138685 PMID: 22505627

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. https://doi.org/10.1038/nbt.2486 PMID: 23354101

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. https://doi.org/10.1093/nar/gkq274 PMID: 20439311

30. Siddharthan R. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. PLoS One. 2010; 5(3):e9722. https://doi.org/10.1371/journal.pone.0009722 PMID: 20339533

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. https://doi.org/10.1038/nbt1053 PMID: 15637633

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. https://doi.org/10.1093/nar/gkv1156 PMID: 26527724

33. Martínez-Antonio A, Collado-Vides J. Identifying global regulators in transcriptional regulatory networks in bacteria. Curr Opin Microbiol. 2003; 6(5):482–9. PMID: 14572541

34. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator. Genome Res. 2004; 14:1188–90. https://doi.org/10.1101/gr.849004 PMID: 15173120

35. Jensen ST, Liu JS. BioOptimizer: a Bayesian scoring function approach to motif discovery. Bioinformatics. 2004; 20(10):1557–64. https://doi.org/10.1093/bioinformatics/bth127 PMID: 14962923