Published online 12 April 2011

Nucleic Acids Research, 2011, Vol. 39, No. 11 ***e76***
doi:10.1093/nar/gkr205

Finding subtypes of transcription factor motif pairs with distinct regulatory roles

Abha Singh Bais[1,*], Naftali Kaminski[2] and Panayiotis V. Benos[1,3,*]

[1]Department of Computational and Systems Biology, [2]Dorothy P. and Richard P. Simmons Center for Interstitial Lung Disease, Division of Pulmonary, Allergy and Critical Care Medicine and [3]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15260, USA

Received November 12, 2010; Revised March 1, 2011; Accepted March 22, 2011

ABSTRACT

DNA sequences bound by a transcription factor (TF) are presumed to contain sequence elements that reflect its DNA binding preferences and its downstream-regulatory effects. Experimentally identified TF binding sites (TFBSs) are usually similar enough to be summarized by a 'consensus' motif, representative of the TF DNA binding specificity. Studies have shown that groups of nucleotide TFBS variants (subtypes) can contribute to distinct modes of downstream regulation by the TF via differential recruitment of cofactors. A TF_A may bind to TFBS subtypes a_1 or a_2 depending on whether it associates with cofactors TF_B or TF_C, respectively. While some approaches can discover motif pairs (dyads), none address the problem of identifying 'variants' of dyads. TFs are key components of multiple regulatory pathways targeting different sets of genes perhaps with different binding preferences. Identifying the discriminating TF–DNA associations that lead to the differential downstream regulation is thus essential. We present DiSCo (Discovery of Subtypes and Cofactors), a novel approach for identifying variants of dyad motifs (and their respective target sequence sets) that are instrumental for differential downstream regulation. Using both simulated and experimental datasets, we demonstrate how current motif discovery can be successfully leveraged to address this question.

INTRODUCTION

Transcription factors (TFs) are DNA binding proteins that recognize and bind a small set of similar DNA binding sites with high specificity to regulate the expression of multiple target genes. The binding sites of a TF can be determined using *in vitro* or *in vivo* techniques. The former include methods like SELEX and its variants (1,2) and protein binding microarrays (PBMs) (3), whereas the latter routinely involve chromatin immunoprecipitation (ChIP) coupled with either microarrays (ChIP-chip) or deep sequencing (ChIP-seq) [for reviews see Ref. (4,5)]. In most cases, experimentally found binding sites of a TF are similar enough to be summed up by a 'consensus' motif or a 'position weight matrix' [for an excellent review, see Ref. (6)]. In some cases, however, a TF may show different binding preferences *in vitro* and *in vivo* (7,8). Furthermore, even *in vivo*, a TF may bind variants of its main (or 'canonical') motif to target distinct downstream genes [(9–11); reviewed in (7)]. Such variant sites are referred to as 'non-canonical' sites and may contribute to distinct modes of downstream regulation (12–16).

The precise nucleotide sequence of a TFBS plays an important role, not only in attracting the corresponding TF, but also in the recruitment of its cofactors and hence in the mode of regulation of its targets. A TF A may bind to TFBSs of subtype a_1 or a_2 and target different genes depending on whether it associates with cofactor B or C, respectively (Figure 1). For example, in *Escherichia coli*, the cyclic AMP receptor protein (CRP) binds a 22-bp consensus motif CRP-N to regulate roughly 100 genes involved in the response to sugar starvation in the cell (17). In *Haemophilus influenzae*, CRP recognizes the typical CRP-N sites, but it also recognizes and binds to a CRP-N variant, the non-canonical CRP-S motif (18,19). The CRP-S sites are found in the promoters of genes involved in 'competence', a process by which cells can take up DNA from the environment, and which require the presence of both CRP and another protein, Sxy, for transcriptional activation. A similar functional CRP-S regulon has also been identified in *E. coli* (13,20), suggesting possibly a similar mechanism. Hence, the protein CRP recognizes two highly similar but distinct motifs, subtypes CRP-N and CRP-S, in the presence or absence of Sxy thus

*To whom correspondence should be addressed. Email: bais@pitt.edu
Correspondence may also be addressed to Panayiotis V. Benos. Tel: +1 412 6483315; Fax: +1 412 5483163; Email: benos@pitt.edu

regulating different sets of downstream genes. Single nucleotide changes too have been associated with the choice of cofactors and the set of target genes as in the case of the glucocorticoid receptor (16,21), NF-κB (22), Pit-1 (23), Foxa2 (24), etc. In the following, we use the term 'subtypes' of TFBSs to refer to groups of highly similar but distinct nucleotide variants of the canonical binding motif of a TF.

There is a plethora of computational TFBS motif discovery methods [reviewed in Ref. (25–27)]. Typically these methods ignore differences between canonical and non-canonical motifs. Almost all tend to average slight variations of the enriched motifs into 'consensi', thereby not identifying TFBS subtypes. An initial attempt to discover subtypes of single TFBS motifs was a kernel estimations-based method presented by Kel *et al.* (28). The authors tested their method on similar TFBSs of two distinct TFs, AP-1 and CREB. Hannenhalli *et al.* (29) and Georgi *et al.* (30) used mixture-model representations of existing position-specific probabilistic models for TFBSs to account for subtypes of TFBSs. Computational studies analyzing ChIP-chip/seq data have also uncovered non-canonical variants of the main TFBS motifs using information about distance from the TSS or presence or absence of cofactor TFBSs (9,12,21,24). In some cases, the discovered subtypes of TFBSs had similar sequence consensus overall but varying strengths of the sites (24). In other cases, the two motifs were drastically different in sequence composition as well as in length (9). As useful as these approaches are, they do not systematically address the fact that TF binding preferences may change depending on different cofactor associations.

Few motif discovery approaches focus on dyad motif discovery. A dyad motif consists of two motifs separated by a certain number of bases. The predictions from dyad-based methods can be used further for building *cis*-regulatory modules comprised of more than two components. Guhathakurta *et al.* (31) and Liu *et al.* (32) proposed Gibbs-sampling-based approaches to identify dyads, defined as cooperatively acting TFs, two-block motifs or gapped motifs. The method BIPAD (33) predicts a dyad motif pattern, or a 'two-block structured motif' with a given maximum inter-site distance by optimizing over the total information content (IC) of the two half-sites. Another approach GADEM (34) uses a genetic algorithm followed by expectation-maximization to discover 'spaced dyads', defined therein as two words that are separated by a spacer. While most of these dyad-discovery methods are PWM based, the approaches of van Helden *et al.* (35) and Eskin *et al.* (36) are consensus sequence based. The former uses word counting followed by estimation of statistical significance to discover spaced dyad elements, defined as a pair of short conserved words separated by a region of fixed size and variable content. The latter uses a single motif discovery step to identify putative instances of composite signals followed by an exhaustive search. Here, composite patterns are referred to as groups of 'monads', or single motifs, that occur near each other, with dyads being the special case of a pair of monads that appear approximately within a given distance. In another work, Smith *et al.* (37) use a set of candidate motifs that best predict localization data and search the sequence neighborhood for putative cofactor motifs.

None of the aforementioned approaches addresses the question of identifying variants of the same motif of a TF **A**, depending on whether it binds to DNA in the presence of cofactors **B** or **C** (Figure 1). In the case of single motif discovery, an average motif may be calculated for **A**; whereas in the case of dyad motif discovery, a method may fail to detect any dyad motif if the TFBSs of **B** and **C** are substantially different. Since TFs usually participate in multiple regulatory pathways, sometimes targeting different sets of genes with perhaps different binding preferences, it is essential to identify the discriminating TF–DNA associations that lead to different downstream target genes.

We present DiSCo (Discovery of Subtypes and Cofactors), an integrated approach to identify variants

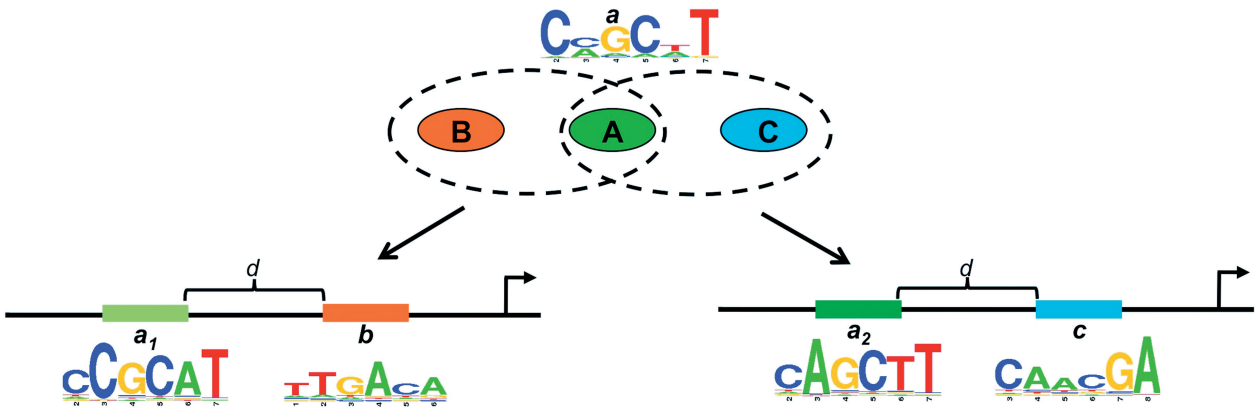

Figure 1. Dyad-dependent modes of regulation. A TF A with DNA binding specificity *a* may bind subtypes, a_1 and a_2, of binding sites depending on whether they co-occur with binding sites *b* and *c* of cofactors B and C, respectively. Alternatively, a TF that binds as a dimer may bind dyads a_1:*b* or a_2:*c* depending on the sets of downstream targets and/or vice versa. Finally, two distinct TFs with highly similar but distinct binding sites a_1 and a_2, may recognize and bind their respective TFBSs by associating with corresponding cofactors B and C, respectively. Logos represent the consensus motifs of the TFs and *d* is the maximum inter-site distance.

of TFBS dyad motifs and the subsets of sequences that reflect the different modes of regulation adopted by the TF. We refer to motifs of same length that differ marginally as different subtypes of the main TFBS. While this definition precludes examples where motifs of different lengths have been discovered as being bound by the same TF, it enables us to distinguish TFBS subtypes based on variation purely in sequence composition and not length differences. For simplicity, we do not take into account position dependencies within a motif. As mentioned earlier, we refer to pairs of binding sites that occur within a maximum distance, say d, as dyads. Slight variations from the consensus in either one or both components of the dyad are termed as *dyad subtypes*. Unless otherwise stated, the order or the orientation of the motifs inside a dyad is not considered. Our assumption is that a TF may adopt different binding preferences depending on the constitution of enriched dyads in a set of target sequences (Figure 1), which could lead to different modes of downstream regulation. Although it is possible that a TF adopts more than two distinct modes of DNA binding and regulation along with associated cofactors, to the best of our knowledge, no such examples have been observed yet. Note that while our main goal is to study subtype-cofactor dependent modes of TF binding, TFs that bind as dimers with varying dimer composition also fall into the same technical framework. Finally, DiSCo can be used to differentiate between highly similar DNA motifs of two different TFs.

To highlight the applicability of our approach, we utilize an existing PWM-based tool, BioProspector (32), which performs both single and dyad motif discovery, and is available for download and use. However, the approach we present here is general and allows the use of other dyad-discovery tools. We test DiSCo on simulated and experimental data, and demonstrate how existing motif discovery algorithms can be leveraged to yield subtypes of pairs of TFBS motifs as well as the sequence subsets that are enriched with them. The DiSCo software consists of Perl scripts and R code for clustering of TFBSs found using the freely available tool BioProspector, and is available from the authors on request.

MATERIALS AND METHODS

Let A be a TF with a DNA binding specificity represented by the PWM P_A. Let P_{A1} and P_{A2} be the PWMs representing two subtypes of the consensus binding site of A that are associated with motifs P_B and P_C, respectively, within a maximum distance d. Let $S = S_{A1:B} \cup S_{A2:C}$ be the set of sequences bound by A and containing instances of motifs P_{A1} and P_{A2}, respectively. In this setting, a standard dyad discovery (SDD) method may yield motif pairs that cluster together instances of both kinds of subtype-cofactor pairs and hence have poorer discriminating performance. We would like to have a method that better characterizes the probabilistic models P_{A1}, P_{A2}, P_B and P_C, and partitions S into subsets $S_{A1:B}$ and $S_{A2:C}$.

Approach

We propose a novel approach for dyad motif subtype discovery, based on dyad enrichment that distinguishes target sequence sets. The general idea is as follows (Supplementary Figure S1). Sequences are partitioned into two clusters based on initial predictions of a dyad motif discovery algorithm. Each of the two clusters is then subject to a second round of dyad motif discovery. Focusing on subsets of sequences with highly similar binding sites increases the signal-to-noise ratio, which in turn enables higher quality prediction of binding site subtypes. At the same time, the dyad motif discovery yields TFBS subtype-cofactor relations and—if the query dataset is a set of promoters—the respective downstream targets. Let $S_1, \ldots, S_N$ be the N sequences containing the targets of TF A. Let W and w be the widths of the two components of the dyad which are enriched within a maximum inter-site distance of d.

Step 1: de novo dyad discovery. In this article, motifs are discovered using BioProspector (32), although, as mentioned earlier, any appropriate dyad-discovery tool can be used. BioProspector is a Gibbs-sampling-based method adapted for efficient gapped motif discovery and consideration of higher order background models. In the predictive-update step, it initializes two motif models by randomly choosing two positions, a_i, b_i on the same strand in each sequence S_i such that $d = b_i - (a_i + W - 1)$. The models are constructed by using the substrings of lengths W and w, respectively, starting at a_i, b_i. In the sampling step, the randomly chosen pair (x_1, x_2) in each sequence is scored as $F(x_1, x_2) = m_1(x_1) \cdot m_2(x_2)/\pi(x_1) \cdot \pi(x_2)$ where $m_1(x_1), m_2(x_2)$ are the respective probabilities of generating x_1, x_2 under the two corresponding motif models and π is the background distribution. For every sequence, new positions for a_i, b_i are sampled with a probability proportional to $F(x_1, x_2)$ and the corresponding substrings are used to formulate two new motif models. The position for a substring x_1 for the first motif is sampled using its marginal distribution $F(x_1, *) = \sum_{x_2} F(x_1, x_2)$, where the sum is over all substrings of width w within the gap range d downstream from x_1. Then the position for segment x_2 is chosen with probability $F(x_1, x_2)/F(x_1, *)$ conditioned on x_1.

In the default dyad-discovery mode, BioProspector searches both strands of the input sequences for motif pairs with widths W and w, respectively, within a gap range $[g_{\min}, g_{\max}]$ ($d = g_{\max} - g_{\min}$) such that not all sequences need contain a copy of the motif pair. After multiple initializations (default = 40) to avoid local optima, it outputs the top five scoring motif pairs with the corresponding sites, locations, strands and scores in the relevant sequences. The scores correspond to the average IC of the two halves of a dyad. Hence for a motif pair, there can be multiple instances of site pairs predicted in a sequence. If no background model file is given, the input sequences are used to calculate the background.

For our approach the following settings of BioProspector are used: (i) each sequence is searched for at least one occurrence of the motif pair with a maximum gap of $g_{max} = d$, hence $g_{min} = 0$, and (ii) only the locations of the highest scoring site pair in every sequence for the top ranking motif pair are stored in each search. This implies that for every sequence S_i, Step 1 yields position pairs (a_i, b_i) corresponding to the best scoring site pair under the top ranking motif pair. For the purposes of the present study, the results of Step 1 correspond to those of a SDD method.

Step 2: Site-based clustering. Let $(s_i^{(1)}, s_i^{(2)})$ be the substrings of lengths W and w, respectively, starting at positions (a_i, b_i) in S_i. Our aim is to cluster together sequences that have highly similar binding sites of one or both components in the dyad. Hence, for all sequences with a predicted site pair, we calculate the pairwise proximity matrix $D(i, j)$ by comparing the respective highest scoring site pairs $(s_i^{(1)}, s_i^{(2)}), (s_j^{(1)}, s_j^{(2)})$. The comparisons can be between the sites of the first motif (i.e. with respect to the first motif, parameter $wrt = 1$), sites of the second motif ($wrt = 2$) or sites of both motifs ($wrt = 3$). Unless otherwise stated, we calculate the Hamming distance (HD) between the sites of both motifs ($wrt = 3$) in a pair of sequences S_i, S_j, i.e. $D(i,j) = \mathrm{HD}(s_i^{(1)}, s_j^{(1)}) + \mathrm{HD}(s_i^{(2)}, s_j^{(2)})$. We note that other distance metrics can be used too. Once the pairwise proximity matrix is formulated, we use the iterative clustering method of Dubnov *et al.* (38) for clustering. This is a non-parametric approach predicted to converge to two broad clusters in most cases via a transformation on the pairwise proximity matrix followed by hierarchical clustering. The pairwise proximity matrix $D(i, j)$ constructed from the HD is iteratively transformed according to the Jensen–Shannon divergence (39) and two broad clusters extracted.

Step 3: Motif discovery on clusters. In a sequence, the location of the site corresponding to the first motif in the best scoring site pair is used as seed for the next round of motif discovery in the respective cluster the sequence belongs to. This means that in Gibbs sampling the first motif is initialized at the locations of the corresponding sites predicted in the previous round. Motif discovery is performed on both clusters yielding four newly discovered PWMs, two from each cluster. The results of this motif discovery after clustering are referred to as those of our algorithm DiSCo, i.e. Steps 1–3 constitute a single run of DiSCo. While the algorithm can be iterated until convergence, we observed that a single iteration has sufficient discriminating power and there is no practical improvement in performance after the first iteration (data not shown). Hence, in the following tests, we present results where we only run DiSCo (Steps 1–3) once.

Post-processing of stochastic dyad motif finder results

Our approach uses the initial predictions of a dyad motif-finding program like BioProspector (32). In general, since most dyad-discovery methods are stochastic, their outputs may vary in each run. Thus, for the analysis of biological examples, our algorithm includes a post-processing step, in which the pipeline runs multiple times and the final output is decided by majority polling. Consider the motif-discovery step of either before or after clustering. On running motif discovery 'Nrun' times, for each sequence, the pair of sites predicted most often across all runs is considered a 'robust' prediction. Similarly, the run in which the output PWMs are composed of robust pairs of sites for the majority of sequences is likely to be the one with the most robust output. That is, the PWMs are formulated from sites that are predicted most often across 'Nrun' runs. For both SDD and DiSCo, we select the run that yields the most robust site pairs over all runs for the majority of sequences as the final output. Hence, both SDD and DiSCo end up predicting only one dyad occurrence per sequence.

Suppose that after SDD, the highest scoring site pair in sequence S_i in the j-th run is located at $(a_i^{(j)}, b_i^{(j)})$. Hence, over 'Nrun' runs, we have the locations of the highest scoring site pair for sequence $S_i : \{(a_i^{(1)}, b_i^{(1)}), (a_i^{(2)}, b_i^{(2)}), \ldots, (a_i^{(\mathrm{Nrun})}, b_i^{(\mathrm{Nrun})})\}$. Let $(a_i^{(*)}, b_i^{(*)})$ be the robust prediction for sequence S_i. That is, it is predicted most often to be the location of the highest scoring site pair in Nrun runs, say it is predicted in $R_i : \{r_i^{(1)}, r_i^{(2)}, \ldots, r_i^{(N_i)}\}$ runs. Hence, for each sequence S_i, we have R_i as the set of runs with the robust location as the final prediction. We choose the output of the run $R^* = \cap_{i=1}^{N} R_i$ that predicts robust locations for all sequences. Clearly, it is possible that no single run predicts the robust locations of all sequences, in which case we pick a run that predicts it for the maximum number of sequences. In case all Nrun runs yield different predictions, we choose the prediction of a randomly selected run as (a_i^*, b_i^*). If more than one location pair is predicted often and same number of times, we randomly choose one to be the robust location and so on. The same procedure is followed for DiSCo.

The dyad(s) discovered after the majority polled run is taken as the final output of the method (SDD or DiSCo). To provide an estimate of the significance of the motif found, a P-value for each site of each of the motifs constituting the final dyad(s) is also calculated. The P-value is calculated by simulating the background motif score distribution by sampling $N = 100\,000$ instances from the input nucleotide distribution. The P-value of a site corresponds to the probability of such a random instance having a motif log-odds score at least as high as that of the site. Additionally, Monte Carlo simulations can be performed by running DiSCo with the same parameters on randomly generated datasets with the same sequence number, length and nucleotide distribution as the original data and comparing the mean motif scores of the final dyads found.

Evaluation datasets

Simulated data. First, we compare the performance of DiSCo (i.e. results after clustering in Step 3) versus SDD (i.e. results after Step 1) in a controlled setting. Simulated datasets are constructed by implanting binding sites sampled from artificial PWMs (10 bp long), of varying

IC into 20 (200-bp long) random sequences sampled from the uniform background distribution. The artificial PWMs were generated from randomly selected matrix columns in the TRANSFAC (40) database (version 11.2). The TRANSFAC matrix columns were categorized as being of *low* (*L;* $1 \leq IC \leq 1.5$) or *high* (*H*; $1.5 \leq IC \leq 2$) IC. To create a PWM of length W of high (or low) IC, we randomly choose W number of high (or low) IC columns. For generating the subtype a_2 of motif a_1, a proportion δ ($\delta = 0\%$, 30% or 40%) of the columns of a_1 are randomly selected and the consensus nucleotide (A, C, G or T) is permutated. In each sequence, a site sampled from one matrix (say a_1 or a_2) is implanted at a random location followed by one sampled from another matrix (either b or c) within a randomly selected distance $d' \leq d = 20$. For each dataset, two kinds of sequences with different site pairs are constructed and the known locations of the implanted sites stored. One complete dataset S is made up of two subsets of sequences: S_1 with sites sampled from motif models of subtype a_1 and motif b (a_1:b), and S_2 with sites from subtype a_2 and motif c (a_2:c). Without loss of generality, sites are implanted in the forward strand only. Since, for each sequence, only the best scoring site pairs are reported, we calculate the true positive (TP) and false positive (FP) sites by comparing the predictions after SDD and DiSCo with the known locations. Steps 1–3 are run multiple times (Nrun = 20) to calculate average sensitivity. For *de novo* dyad discovery, both in SDD at Step 1 and DiSCo in Step 3, we search for pairs of motifs of widths $W = w = 10$ with a maximum gap of $d = 20$. Only the forward strand is searched for at least one instance of the motif pair. BioProspector is run for the default number of initializations and the sites of both motifs are used for calculating the HD at the clustering step (i.e. $wrt = 3$).

CRP dataset. The *E. coli* CRP protein is a well-studied TF known to regulate hundreds of genes involved in carbon-energy starvation response amongst other functions [for a review see Ref. (41)]. In *E. coli*, CRP typically recognizes and binds a 22-bp symmetric motif (CRP-N) *AAA***TGTGA**(N_6)**TCACA***TTT* (42). In *H. influenzae*, CRP binds not only the typical CRP-N sites, but also a variant CRP-S, which is enriched in the promoters of competence genes (18,19). This variant differs from CRP-N sites at one position in the core of each of the half-sites, **TGCGA**(N_6)**TCGCA** in CRP-S instead of **TGTGA**(N_6)**TCACA** in CRP-N (Figure 3), and requires CRP and another protein, Sxy, for transcriptional activation (19,43). Recently, Sinha *et al.* (20) identified the equivalent CRP-S regulon in *E. coli* and demonstrated its requirement of both CRP as well as Sxy for transcriptional activation. The fact that regulation of the competence genes requires the CRP-S and not the CRP-N motif shows that differences in the binding motifs (and associated cofactors) can be directly associated with differential regulation of certain categories of genes.

We focus on 300 bases upstream of the CRP-N and CRP-S target genes in both *E. coli* and *H. influenzae*. Using the available CRP matrix (44), we retrieved the list of the *E. coli* CRP-N genes with experimentally found sites within 300-bp upstream of the TSS. For the *E. coli* CRP-S sequences, we used the 34 transcriptional units found to be differentially regulated by Sxy and considered putative CRP-S genes by Sinha *et al.* (20). For both sets, we extracted the corresponding upstream sequences from the *E. coli* K12 genome using RSATools (35), allowing for overlaps and not admitting imprecise positions. For completeness, we manually obtained sequences for genes whose identifiers did not yield matches. This resulted in a total of 25 CRP-N sequences and 37 CRP-S sequences in *E. coli* with 48 and 43 sites, respectively. The *H. influenzae* CRP-N and CRP-S genes were taken from the Supplementary Data of Cameron *et al.* (13). Using the complete genome sequence and coordinates information of the *H. influenzae* Rd KW 20 genome as downloaded from *The Comprehensive Microbial Resource* (45), we extracted the upstream regions for the respective transcriptional units for the two sets that contained 41 and 13 sequences respectively.

AP-1 and CREB dataset. We also considered two distinct TFs, AP-1 and CREB, that have binding sites that differ slightly at the 3′ end of the core regions (see below). A mixture of TFBSs retrieved from the TRANSFAC database (40) for these two TFs was used by Kel *et al.* (28) in a study aimed to distinguish between subtypes of single patterns. Their method successfully discovered two motifs of lengths 7 and 8 that were identical to the TRANSFAC motifs of AP-1 and CREB, respectively. Our focus is not just extracting subtypes of TFBSs but also the motifs from the surrounding regions associated with potential cofactors. Hence, for the present study, we retrieved 50 bp around the sites contributing to the matrices for AP-1 and CREB in TRANSFAC (40) (version 11.2) (IDS V$AP1_Q2_01 and V$CREB_Q4_01, respectively). Although there are numerous matrices for AP-1 and CREB in TRANSFAC, we choose to use these since they are formulated from experimentally derived sites from vertebrate species. We retrieved sequences from EBI using 'dbFetch' for the sites with EMBL IDs in the 'site.dat' file of the TRANSFAC database. The sites were then mapped onto the EBI sequences and ±50-bp regions extracted for analysis. We ran DiSCo for Nrun = 20 times and compared the resulting PWMs with the TRANSFAC database using STAMP (46), with no trimming for weak IC columns and default settings otherwise.

NF-κB dataset. Using a computational approach, Busse *et al.* (14) studied the NF-κB TFBSs in the regulatory regions of genes activated in the Toll and Imd pathways in Drosophila. They identified pathway-specific characteristics of the κB sites: nearly two-thirds of the Imd-specific promoter set had a GGGGA 5′ half-site. This site is absent in the Toll-specific set which instead had either GGGA or GGAA 5′ half-site in the upstream regions of the corresponding genes (14). We retrieved the 16 and 11 Toll- and Imd-specific sequences with a total of 17 and 21 sites, respectively, from the Supplementary Data of Busse *et al.* (14).

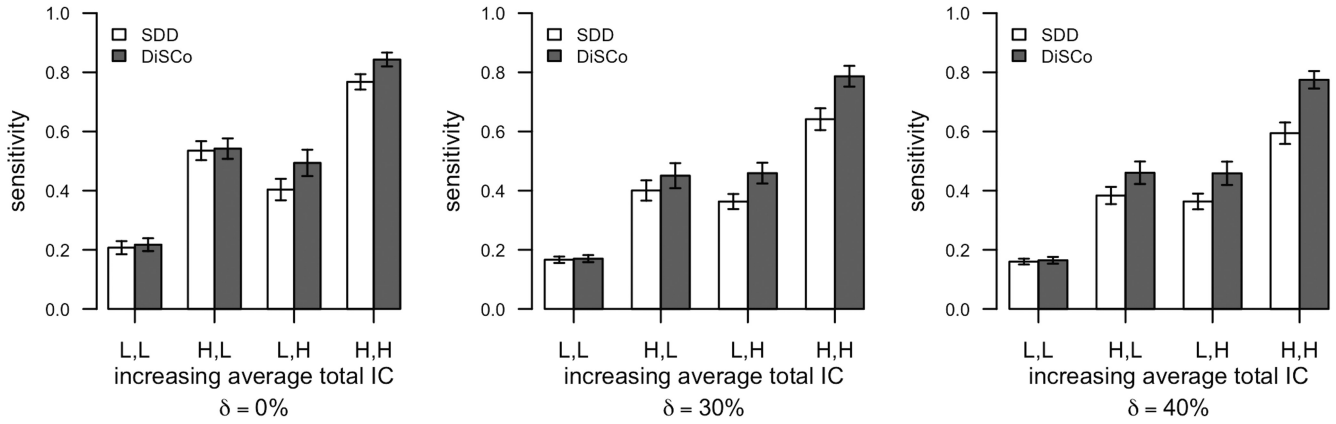

Figure 2. DiSCo outperforms SDD on simulated datasets. The average sensitivity after Standard dyad discovery (SDD; white) and DiSCo (grey) as the proportion of dissimilarity between the subtypes a_1 and a_2 increases from $\delta = 0\%$, 30% and 40%, is shown. The bars are plotted for datasets with increasing mean total IC of the constituting matrices with the labels denoting the individual IC levels for the first (a_1 or a_2) and second motifs (b or c) used to construct a dataset. For example, 'L, H' denotes the datasets where motif models a_1 and a_2 are of low IC levels and b and c of high IC levels, respectively. L: low IC and H: high IC category, respectively. The error bars indicate standard errors. The proportion of dissimilarity is measured as the percentage of dissimilar positions between two matrices.

RESULTS

Evaluation of performance on simulated data

We study the problem of identifying dyad subtypes through simulated datasets that emulate sequences with binding site pairs that vary in signal strengths. The aim is to compare the performance of SDD (standard *de novo* dyad discovery before clustering) with DiSCo (dyad discovery after clustering), in a controlled setting. To this end, we generate sequence datasets with pairs of sites implanted at a maximum inter-site distance d. Sites are sampled from artificial matrices a_1, b, a_2 and c of varying IC. Each dataset, S, consists of two kinds of sequence subsets, S_1 and S_2, such that S_1 has implanted sites sampled from motif models a_1 and b and S_2 has those sampled from a_2 and c, respectively. Here, a_1, a_2 represent subtypes of a motif with varying amounts of dissimilarities between each other (see 'Materials and Methods' section). The predicted locations after SDD and DiSCo are compared with the implanted locations to calculate the TP and FP predictions. Since the number of known and predicted sites per sequence are equal, we use the average sensitivity (TP/TP + FN) over multiple runs for performance evaluation. The results are presented in Figure 2.

In general, irrespective of the average total IC of the matrices used, the performance of *de novo* discovery both before and after clustering deteriorates as the similarity between a_1 and a_2 decreases (i.e. from $\delta = 0\%$ to 40%, Figure 2). Clearly, the presence of sites that are sampled from the same or very similar motifs throughout the sequence set enables the SDD to yield more correct predictions, which also helps DiSCo. As the dissimilarity between the two subtypes increases, SDD deteriorates in performance yielding more false predictions. Since the output of SDD is used to seed the DiSCo motif discovery, incorrect predictions after SDD negatively affect DiSCo. However, the decrease in average sensitivity for DiSCo is less than that of SDD as the similarity between a_1 and a_2 decreases. For most matrix combinations, DiSCo yields a significantly higher average sensitivity than SDD, a direct consequence of the increased signal-to-noise ratio after clustering. In the case where both matrices are of low IC, both methods perform poorly, with no significant difference in their average sensitivity values. Again, at low IC, the predicted sites for both motifs after SDD are poor, which in turn influences the behavior of DiSCo. Since all motifs are poorly informative, putative predicted sites of a motif after SDD are inclined to be quite dissimilar to each other making them more unlikely to fall in the same cluster, leading to a detrimental effect on the performance of DiSCo. Finally, as expected, the increase in the average total IC of the constituting matrices improves the performance of both methods. The case of the (L, H) dataset is interesting. This is where the first and second motifs have low and high IC, respectively. A drop in the average sensitivity values is observed from the (L, H) to (H, L) dataset even though there is an increase in the total average IC. This is a consequence of using BioProspector, which is designed to find the strongest motif first and then search only downstream for the second motif. Hence, here BioProspector discovers the stronger motif (in this case, the second motif) first, which leads to increased false predictions for the weaker motif [located upstream in the (L, H) dataset] yielding a dip in sensitivity values. Again, although this affects negatively both SDD and DiSCo, DiSCo performs better (Figure 2). We performed similar simulations with search parameters changed to motif widths of $W = w = 8$ and got similar results. Additionally, we generated and analyzed simulated datasets with increased inter-site distance ($d = 50$ bp) and the results were practically the same (data not shown). In summary, as the difference between a_1 and a_2 increases, there is an improvement in DiSCo performance due to clustering.

Evaluation on the CRP dataset

CRP motifs provide an appropriate dataset for testing our algorithm for the following reasons. Although both CRP-N half-sites are highly similar between *E. coli* and

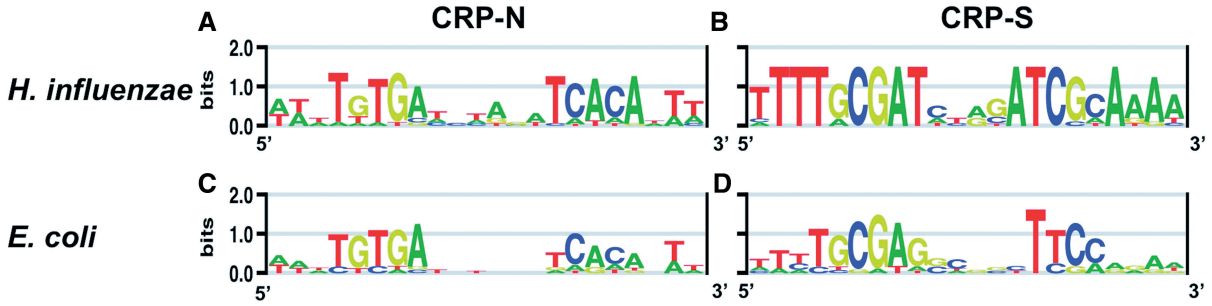

Figure 3. Motif logos of the CRP-N and CRP-S sites of *H. influenzae* and *E. coli*. *Haemophilus influenzae* CRP-N (**A**) and CRP-S (**B**) motifs created from sites retrieved from the Supplementary Tables of Cameron *et al*. (13); *E. coli* CRP-N (**C**) and CRP-S (**D**) motifs created from sites retrieved from (44) (CRP-N) and (19) (CRP-S). Logos generated using enoLOGOS (49).

H. influenzae (TGTGA), the CRP-S sites differ between the two species, especially in the second half-site. In particular, both species have similar first half-sites in the CRP-S motifs, except in the last base (Figure 3B and D); whereas the second half-site, that lies 6-bp away, appears to be species specific. Thus, we can treat the two half-sites as two separate motifs: one that is shared amongst all sequences and another that differs between the two subsets (in this case, the two subsets come from two different organisms). Focusing on a single species, the complete CRP datasets can also be used to test DiSCo. In *H. influenzae* the two CRP motifs vary in both half-sites (Figure 3A and B). Here, one type of dyad is enriched in CRP-S related promoters and the other in CRP-N related ones. In this case, both components of the dyads are specific to a (different) regulatory mode—competence or non-competence related. Lastly, in both species, the complete 22-bp CRP sites have variants CRP-S and CRP-N, depending on whether they are derived from promoters of competence or non-competence genes, respectively. In the former case, CRP binding requires the presence of the protein Sxy. Here, one TF (CRP) binds to subtypes of TFBSs, CRP-N and CRP-S, respectively, depending on the presence of cofactors. This provides an opportunity to explore novel motifs that might be specifically associated with the regulation of competence genes by CRP.

The different aspects of the CRP dataset fit well into our technical framework and provide an excellent test case scenario. We use the *E. coli* and *H. influenzae* CRP datasets to demonstrate the applicability of DiSCo in a biological setting through three main analyses. First, as a proof of principle we run DiSCo on the pooled set of CRP-S sequences from these two species to automatically partition them into two species-specific clusters and identify the dyads enriched in each. The two species-specific datasets have similar first half-sites but different second half-sites. Second, we run DiSCo on the set of pooled CRP-N and CRP-S sequences from *H. influenzae* in order to predict the two distinct dyad types and their corresponding targets. In this case, both first and second half-sites are different between the two datasets. Finally, we use DiSCo to study putative motifs that might be associated with each complete CRP dyad subtype in *H. influenzae*. In each case, site *P*-values are calculated as described in 'Materials and Methods' section. The maximum *P*-values for each motif output are presented in Supplementary Table S1.

Similar first half-sites, different second half-sites. We studied the pooled *E. coli* and *H. influenzae* CRP-S-associated promoter sequences. The motifs of the best scoring sites in the majority polled runs of SDD and DiSCo, when searching both strands for motif pairs of widths 5 each at a distance of 6 bp, are presented in Figure 4. Both methods discover the first half-site of the two species-specific TFBSs, though in one case the predicted motif is slightly shifted. For the second half-site, the motif discovered in SDD resembles more closely the second half-site of the *E. coli* CRP-S motif. Clearly, the greater number of *E. coli* CRP-S related sequences (37 *E. coli* versus 13 *H. influenzae* CRP-S) dominated the motif discovery in SDD, resulting in a motif that seems to average the individual dyads. In contrast, DiSCo identifies two sequence clusters with enriched dyads that correctly match the CRP-S motifs of *H. influenzae* (C_1) and *E. coli* (C_2). Since the first half-sites are very similar between the two species and present in almost all sequences, it is easy for SDD to correctly predict the locations of the first motif in the sequences. Using these predicted locations to initialize the motif discovery after clustering enables correct predictions for this motif in DiSCo. This ultimately leads to better identification of species-specific sequences as well as dyad subtypes. All *H. influenzae* CRP-S sequences correctly fall in the same cluster and majority of *E. coli* CRP-S sequences are clustered together, with only ~24% of *E. coli* CRP-S sequences being misclassified (i.e. clustered with *H. influenzae* CRP-S). It should be stressed here that the pooled and weak second half-site discovered using the SDD is not necessarily the shortcoming of the underlying method used (in this case, BioProspector). Given the limitations of experimental knowledge, most motif discovery methods are inclined to do the same. However, as demonstrated, it is feasible to use the current state-of-the-art methods to further address questions of binding site subtypes.

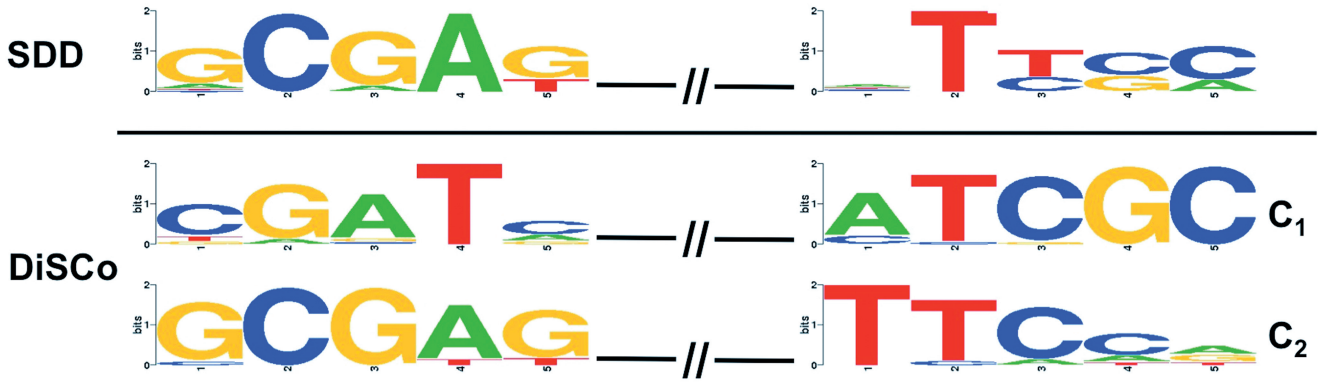

Figure 4. Motifs found on CRP-S sequences of *E. coli* and *H. influenzae*. The dyads discovered in the majority polled run after SDD (top row) and DiSCo (two bottom rows) are shown. SDD yields a dyad whose components predominantly resemble the two half-sites of the *E. coli* CRP-S motif. In contrast, DiSCo identifies clusters C_1 and C_2; C_1 is enriched with a dyad similar to the *H. influenzae* CRP-S motif (Figure 3B) and C_2 is enriched with one that is similar to the *E. coli* CRP-S motif (Figure 3D). Average misclassification error = 0.2045.

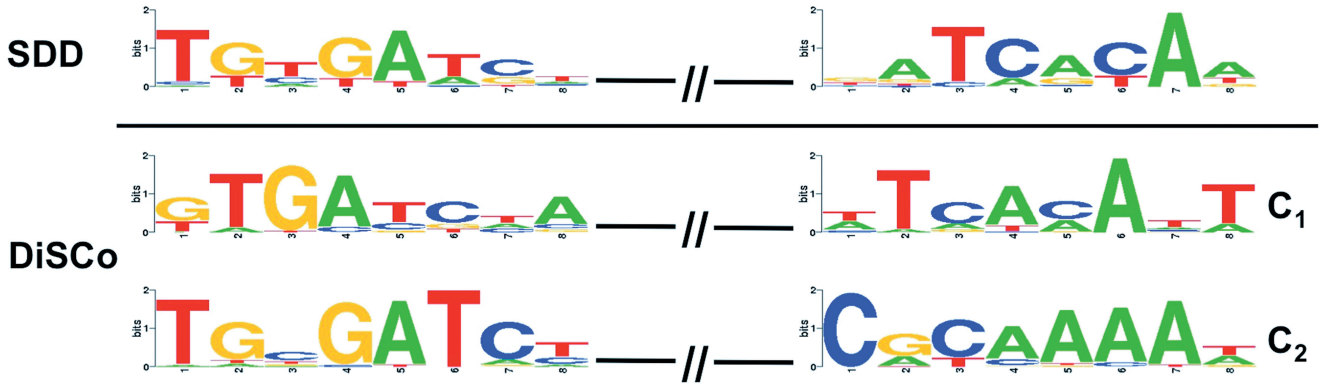

Figure 5. Motifs found on CRP-N and CRP-S sequences of *H. influenzae*. The dyads discovered in majority polled run after SDD (top row) and DiSCo (bottom two rows) are shown. SDD yields a dyad whose components have clustered together the half-sites of both types of *H. influenzae* CRP motifs (Figure 3A and B). DiSCo however is able to successfully identify two clusters C_1 and C_2 whereby C_1 is enriched with a dyad similar to the *H. influenzae* CRP-N motif (Figure 3A) and C_2 with one that resembles the *H. influenzae* CRP-S motif (Figure 3B). Average misclassification error = 0.23.

Different first and second half-sites. Next, we analyzed the pooled CRP-N and CRP-S related sequences in *H. influenzae* with resulting motifs shown in Figure 5 (both strands searched for motif pairs of widths 8 each at a distance of 6 bp). Here, besides the difference in the cores, the differences in the surrounding regions of the two half-sites in CRP-S and CRP-N are also prominent. The dyad motif discovered after SDD matches the CRP-N motif of *H. influenzae*, which might be due to the greater number of CRP-N related sequences (41) as compared to CRP-S related sequences (13). The two dyads DiSCo yields clearly resembling the two motifs. The CRP-S sequences again are clearly clustered together yielding a strong CRP-S like motif. Hence, despite the two half-sites differing in both sets, DiSCo successfully partitions the sequences into those enriched with the specific subtype of CRP hetero-dimers.

Sequence properties of putative cofactors. To study if there exist additional sequence motifs that are needed for regulation of CRP-S sequences, we use DiSCo to analyze the target sequences of both CRP variants to search for dyads where one motif is of length 22 bp. The aim here is 2-fold. One, to study if DiSCo is able to identify the two CRP variants along with their corresponding sequence subsets in each species; and two, to investigate additional sequence signals that might co-occur with each CRP binding site variant, and might aid in the decision of the specific mode of regulation employed by CRP. Since Sxy itself lacks a DNA binding domain and no Sxy binding sites are known, we run DiSCo multiple times with varying parameter values, like motif widths for the second component of the dyad and multiple gap values. This is a typical procedure for many biological problems, when no additional information is known about the potential cofactors. For one of the components, we search for 22-bp long motif. In both species, irrespective of the width of the second component, DiSCo successfully identified the CRP motifs and their target sequence subsets (clusters). On searching for motif pairs of widths $W = 8$ and $w = 22$, within a maximum gap of 10 bp on both strands, we found both CRP-N- and CRP-S-like sites

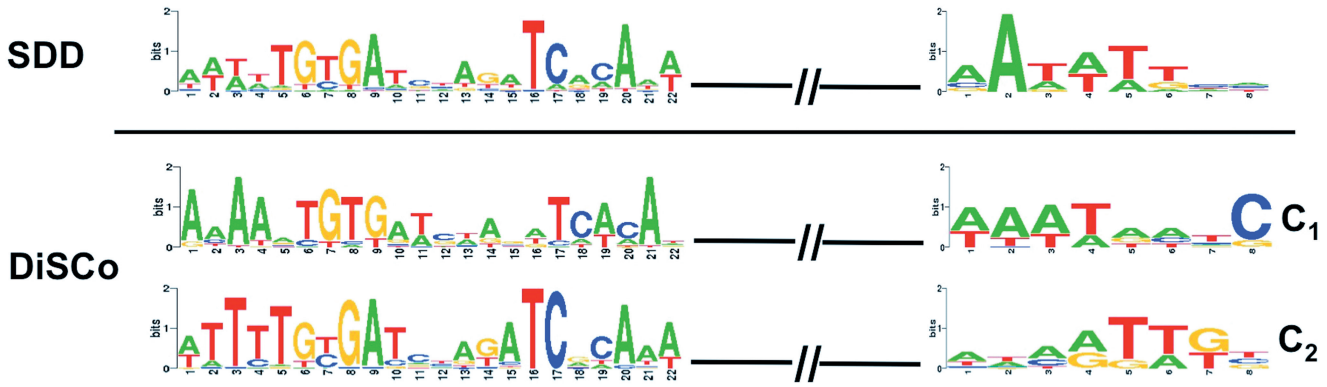

Figure 6. Motifs found on CRP-N and CRP-S sequences of *H. influenzae* with one motif of complete CRP length. The dyads discovered in majority polled run after SDD (top row) and DiSCo (two bottom rows) when the search is performed for the complete CRP motif as the main motif, are shown. SDD yields a pair of motifs, one of which has grouped together the complete CRP-N and CRP-S motifs of *H. influenzae* (Figure 3A and B), and the other is an AT-rich motif. In contrast, DiSCo successfully identifies two clusters C_1 and C_2 where C_1 is enriched with the *H. influenzae* CRP-N-like motif (Figure 3A) and C_2 with the *H. influenzae* CRP-S-like motif (Figure 3B). Additionally, the second motif discovered in C_2 closely resembles the first half of the *E. coli* σ70 motif, found previously also by Cameron *et al.* (43). Average misclassification error = 0.24.

co-occurring with AT-rich motifs (data not shown). Previously, Cameron *et al.* (43) had observed A + T runs upstream of the CRP-S sites in *H. influenzae* that were required for promoter activation. From our analysis, such motifs seem to be present close to CRP-N sites also. On reversing the order of motif widths ($W = 22$ and $w = 8$ bp), again yields AT-rich motifs (data not shown). It seemed that there are only slight differences between the associated motifs of each CRP variant. However, for their motif analysis, Cameron *et al.* (43) aligned sequences that were 200-bp upstream of genes and found *E. coli* σ70-like sites downstream of the CRP-S sites. Following that study, we also restricted our search space to 200-bp regions. For this search, we used the same motif length parameters $W = 22$ and $w = 8$ bp, and a maximum gap of 20 bp. We searched both strands and the forward strand only. In both cases, we identified two clusters, each enriched with one type of CRP motif. In the latter case, though, for the cluster enriched with the CRP-S like motif, the second component matches the first half of the *E. coli* σ70 motif (Figure 6) while the cluster enriched with the CRP-N like motif had a second motif which is AT-rich, but dissimilar to the TTG stretch of *E. coli* σ70. Hence, while the CRP-S sites seem to have at least part of an *E. coli* σ70-like motif downstream, the CRP-N sites do not. In general, by using DiSCo to analyze this biological dataset thoroughly, we were able to identify the two CRP motif subtypes separately along with the distinguishing possible cofactor motif. This shows the direct applicability and usefulness of DiSCo in addressing biological problems.

Evaluation on the AP-1 and CREB dataset

AP-1 and CREB are two TFs with similar but not identical binding sites, which mainly differ at the 3′ end of the core regions (Figure 7A). We tested whether DiSCo is able to identify the two types of binding sites, automatically partition their target sequence sets and in the process identify the surrounding sequence motifs associated with them (if any). To this end, we pooled together and analyzed the complete set of sequences containing both kinds of TFBSs. We searched both strands of all sequences for motif pairs of widths 7 bp each and a maximum gap of 5 bp using sites of both motifs to calculate the clustering measure in DiSCo (i.e. *wrt* = 3). The motifs found in the majority polled runs of SDD and DiSCo are presented in (Figure 7B, *P*-values are reported in Supplementary Table S1). SDD identified a dyad where one motif is similar to both AP-1 and CREB. In other words, the two kinds of TFBSs are pooled together (Figure 7B, top row). However, DiSCo segregates the two sequence sets into two clusters and identifies the individual TFBS motifs (Figure 7B, two bottom rows). The cluster of sequences enriched with the CREB-like motif, contains another co-occurring motif with high-ranking matches to CAC binding motif and Pax-4 in TRANSFAC. On comparing with JASPAR (47) (version 2010), this motif matches Krueppel-like factor 4 (KLF-4). The protein KLF-4 has been shown to be involved in the regulation of mouse B2R promoter by the formation of a higher-order complex with CREB and p53 in conjunction with the co-activator p300/CBP (CREB binding protein) (48). It is likely that CREB and KLF-4 are together involved in the regulation of other genes too, using possibly a similar mechanism. However, we could not find any evidence for the second motif we identified in the sequence set enriched with the AP-1–like motif. In summary, for this set of TFBSs, surrounding sequence lengths and search parameters, the TFBSs of CREB tend to co-occur with those of KLF-4.

Evaluation on the NF-κB dataset

NF-κB has two highly similar sites enriched in pathway-specific promoters (14) (Figure 8). However, there are a couple of points that need to be noted here. First, the spacer lengths in the two sites is different; four or five bases for Toll-specific sites, and two or three bases for

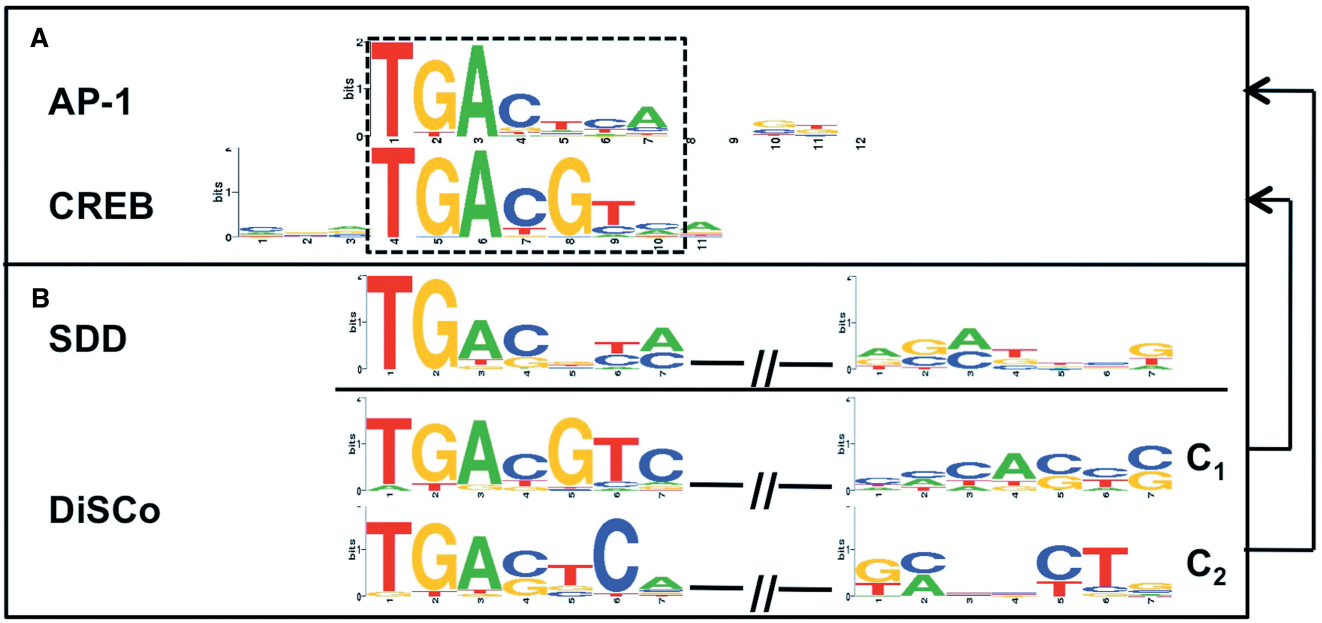

Figure 7. Analysis on the AP-1 and CREB dataset. (**A**) Logos of AP-1 (V$AP1_Q2_01) and CREB (V$CREB_Q4_01) matrices from TRANSFAC are shown. (**B**) Logos formulated from the best scoring sites of the dyads discovered in the majority polled run after SDD (top row) and after clustering (DiSCo; two bottom rows) on the complete set of AP-1 and CREB sequences are shown. SDD yields a dyad composed of a CREB-like motif (STAMP best match to ATF4, *E*-value ~10*e*-09 and to CREB with *E*-value ~9*e*-08) and a motif that matches DEAF1 (*E*-value ~2*e*-04). The clusters resulting from DiSCo are enriched with AP-1 and CREB motifs, respectively. The cluster C_1 [second row in (B)] is enriched with a dyad whose components match CREB and CAC-binding motif (*E*-values of ~2*e*-11 and ~10*e*-04, respectively). On comparing with JASPAR, the second motif best matches KLF-4 (*E*-value = 4.4*e*-03). The components of the dyad discovered in cluster C_2 [third row in (B)] match AP-1 and Adf-1 (*E*-values of ~6*e*-11 and 4*e*-03, respectively). Average misclassification error rate = 0.22.

the Imd-specific sites. Since we do not distinguish between variants based on intra-site spacer lengths, this information is not taken into consideration by DiSCo. Second, the Toll and Imd sets also differ in the number of sites they have per promoter. While most of the promoter sequences in the Toll dataset contain only single κB site instances, most of the promoters in the lmd dataset contain multiple κB sites. This implies that while the total number of Toll and Imd promoters was 16 and 11, respectively, the total number of sites was 17 and 21, respectively. Finally, in approximately half of the Imd sites the 3′ half-site diverged from the canonical site also, which makes it difficult to use dyad discovery to distinguish between the Toll and Imd sets, since differences already exist among the promoters in the Imd set. Keeping these issues into consideration, we used DiSCo to analyze the complete set of Toll- and Imd-specific sequences. We searched for motifs of widths 4 and 5, respectively, on both strands and we calculated the clustering measure based on the sites of the second motif only (i.e. *wrt* = 2). The logos of the best scoring sites in the dyads found in the majority polled run by SDD and DiSCo are shown in Figure 8 (*P*-values are reported in Supplementary Table S1). Since the two subtypes are highly similar, SDD predicts a dyad that summarizes the slight dissimilarities by clustering the two pathway-specific NF-κB motifs into one. On the other hand, DiSCo successfully separates the sequences into two clusters that are enriched with one kind of dyad each. Since here the dyad is comprised of the whole NF-κB motif in each set, the two pathway-specific motifs are hence identified.

DISCUSSION

Advanced technologies that enable the collection of high-throughput *in vivo* and *in vitro* TF–DNA interaction data have shed new light on the changes in the DNA binding preferences of a TF *in vivo*. For eukaryotic organisms with complex mechanisms of gene regulation, the *in vivo* binding preferences of a TF frequently depend on the presence or absence of cofactors. Research has unearthed examples where subtypes of TFBSs contribute to the regulation of different pathway- or function-specific target genes. Typically, the standard approach for analyzing TF-bound sequences is to perform simple *de novo* single or dyad motif discovery. In this work, we presented DiSCo, a first attempt to computationally address the following question: given a set of unaligned DNA sequences bound by a TF A, identify putative subtypes a_1 and a_2 of TFBSs that depend on the presence of TFBSs for possible cofactors B and C, respectively. The main steps of DiSCo consist of dyad motif discovery on the complete sequence set, followed by clustering of sequences based on the similarity of the found dyads, and finally performing dyad motif discovery in each cluster separately. The clustering step helps in reducing the search space, making the second dyad motif discovery more efficient. Through analysis of both

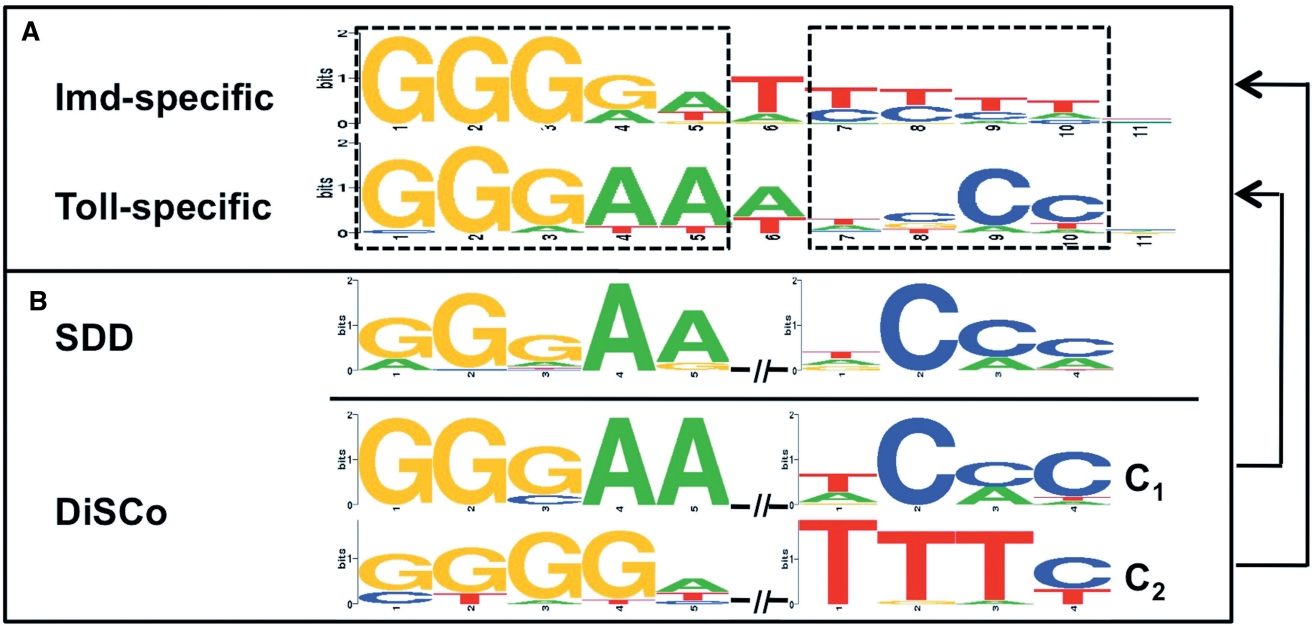

Figure 8. Analysis on NF-κB dataset. (**A**) Logos of Imd- and Toll-specific NF-κB sites. (**B**) Dyads discovered in the majority polled run after SDD (top row) and DiSCo (two bottom rows) for the complete set of Imd- and Toll-specific NF-κB sequences. SDD yields a dyad that has pooled the two pathway-specific subtypes of sites. The clusters resulting from the Step 3 of DiSCo are enriched with Toll specific [C_1, second row in (B)] and Imd specific [C_2, third row in (B)] κB sites, respectively. Average misclassification error rate = 0.28.

artificial and biological datasets, we demonstrated that the approach of DiSCo improves on TFBS detection as compared to standard dyad motif discovery. This was so, even when we considered the top two dyads of the standard motif discovery method or when we masked the occurrences of the first dyad and then searched for a second one (see Supplementary Data).

In its current form, DiSCo uses BioProspector (32) for dyad discovery, although the underlying approach is adaptable to any appropriate dyad-discovery algorithm. Naturally, DiSCo inherits the same problem of signal-to-noise ratio that all motif finders have. In most practical applications where no prior knowledge is available, the user typically performs multiple searches with varying parameter values (motif length, promoter length, etc.) in order to identify motifs that may have biological significance. Also, DiSCo is flexible in the choice of the clustering method and the pair-wise similarity measure it uses. The current clustering method used to extract two broad clusters can be used to recursively extract more clusters, in effect yielding multiple instances of dyad subtypes. Clearly, in this case, larger input datasets may be needed. By default, DiSCo performs the clustering based on the weaker of the two motifs to enable better discriminating power. However, when this motif has very poor IC (close to the background), then using both halves of the dyad for clustering might be more efficient. Similar to most other SDD algorithms, DiSCo reports the average IC of the dyads discovered after running multiple initializations, which can be used to compare the quality of dyads predicted across multiple parameter settings.

The DiSCo approach can also be used to analyze a stringently selected set of bound sequences to identify dyads with relatively short inter-motif distance *d*. But even stringently selected peaks from ChIP-chip or ChIP-seq data tend to be noisy. When we analyzed noisy artificial datasets where 20% of the sequences did not contain any motif, we found that DiSCo performance was fairly stable and superior to SDD (see Supplementary Data for results with additional noise levels). Large-scale datasets (e.g. from ChIP-chip and ChIP-seq experiments) are paving the way for in-depth analysis of *in vivo* binding preferences of a TF, and an approach such as DiSCo has the potential to help elucidate the preferential mode of TF–DNA interaction under different conditions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Dennis Kostka for helpful discussions. The constructive criticism of two reviewers is gratefully acknowledged.

FUNDING

Funding for open access charge: National Institutes of Health (grants R01LM007994 and R01LM009657).

Conflict of interest statement. None declared.

REFERENCES

1. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
2. Zhao,Y., Granas,D. and Stormo,G.D. (2009) Inferring binding energies from selected binding sites. *PLoS Comput. Biol.*, **5**, e1000590.
3. Bulyk,M.L., Huang,X., Choo,Y. and Church,G.M. (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl Acad. Sci. USA*, **98**, 7158–7163.
4. Buck,M.J. and Lieb,J.D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**, 349–360.
5. Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
6. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
7. Farnham,P.J. (2009) Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, **10**, 605–616.
8. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
9. Rada-Iglesias,A., Wallerman,O., Koch,C., Ameur,A., Enroth,S., Clelland,G., Wester,K., Wilcox,S., Dovey,O.M., Ellis,P.D. *et al.* (2005) Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays. *Human Mol. Genet.*, **14**, 3435–3447.
10. Rabinovich,A., Jin,V.X., Rabinovich,R., Xu,X. and Farnham,P.J. (2008) E2F in vivo binding specificity: comparison of consensus versus nonconsensus binding sites. *Genome Res.*, **18**, 1763–1777.
11. Wei,G.H., Badis,G., Berger,M.F., Kivioja,T., Palin,K., Enge,M., Bonke,M., Jolma,A., Varjosalo,M., Gehrke,A.R. *et al.* Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J.*, **29**, 2147–2160.
12. Liu,R., McEachin,R.C. and States,D.J. (2003) Computationally identifying novel NF-kappa B-regulated immune genes in the human genome. *Genome Res.*, **13**, 654–661.
13. Cameron,A.D. and Redfield,R.J. (2006) Non-canonical CRP sites control competence regulons in Escherichia coli and many other gamma-proteobacteria. *Nucleic Acids Res.*, **34**, 6001–6014.
14. Busse,M.S., Arnold,C.P., Towb,P., Katrivesis,J. and Wasserman,S.A. (2007) A kappaB sequence code for pathway-specific innate immune responses. *EMBO J.*, **26**, 3826–3835.
15. Hollenhorst,P.C., Chandler,K.J., Poulsen,R.L., Johnson,W.E., Speck,N.A. and Graves,B.J. (2009) DNA specificity determinants associate with distinct transcription factor functions. *PLoS Genet.*, **5**, e1000778.
16. Meijsing,S.H., Pufall,M.A., So,A.Y., Bates,D.L., Chen,L. and Yamamoto,K.R. (2009) DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science*, **324**, 407–410.
17. Kolb,A., Busby,S., Buc,H., Garges,S. and Adhya,S. (1993) Transcriptional regulation by cAMP and its receptor protein. *Annu. Rev. Biochem.*, **62**, 749–795.
18. Macfadyen,L.P. (2000) Regulation of competence development in Haemophilus influenzae. *J. Theor. Biol.*, **207**, 349–359.
19. Redfield,R.J., Cameron,A.D., Qian,Q., Hinds,J., Ali,T.R., Kroll,J.S. and Langford,P.R. (2005) A novel CRP-dependent regulon controls expression of competence genes in *Haemophilus influenzae*. *J. Mol. Biol.*, **347**, 735–747.
20. Sinha,S., Cameron,A.D. and Redfield,R.J. (2009) Sxy induces a CRP-S regulon in Escherichia coli. *J. Bacteriol.*, **191**, 5180–5195.
21. Morin,B., Nichols,L.A. and Holland,L.J. (2006) Flanking sequence composition differentially affects the binding and functional characteristics of glucocorticoid receptor homo- and heterodimers. *Biochemistry*, **45**, 7299–7306.
22. Leung,T.H., Hoffmann,A. and Baltimore,D. (2004) One nucleotide in a kappaB site can determine cofactor specificity for NF-kappaB dimers. *Cell*, **118**, 453–464.
23. Shewchuk,B.M., Ho,Y., Liebhaber,S.A. and Cooke,N.E. (2006) A single base difference between Pit-1 binding sites at the hGH promoter and locus control region specifies distinct Pit-1 conformations and functions. *Mol. Cell Biol.*, **26**, 6535–6546.
24. Tuteja,G., Jensen,S.T., White,P. and Kaestner,K.H. (2008) Cis-regulatory modules in the mammalian liver: composition depends on strength of Foxa2 consensus site. *Nucleic Acids Res.*, **36**, 4149–4157.
25. Bailey,T.L. (2008) Discovering sequence motifs. *Methods Mol. Biol.*, **452**, 231–251.
26. Hannenhalli,S. (2008) Eukaryotic transcription factor binding sites–modeling and integrative search methods. *Bioinformatics*, **24**, 1325–1331.
27. MacIsaac,K.D. and Fraenkel,E. (2006) Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput. Biol.*, **2**, e36.
28. Kel,A., Tikunov,Y., Voss,N., Borlak,J. and Wingender,E. (2005) Application of Kernel Method to Reveal Subtypes of TF Binding Motifs. *Lect. Notes Comput. Sci.*, **3318**, 42–51.
29. Hannenhalli,S. and Wang,L.S. (2005) Enhanced position weight matrices using mixture models. *Bioinformatics*, **21(Suppl 1)**, i204–i212.
30. Georgi,B. and Schliep,A. (2006) Context-specific independence mixture modeling for positional weight matrices. *Bioinformatics*, **22**, e166–e173.
31. GuhaThakurta,D. and Stormo,G.D. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics*, **17**, 608–621.
32. Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, **6**, 127–138.
33. Bi,C. and Rogan,P.K. (2004) Bipartite pattern discovery by entropy minimization-based multiple local alignment. *Nucleic Acids Res.*, **32**, 4979–4991.
34. Li,L. (2009) GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. *J. Comput. Biol.*, **16**, 317–329.
35. van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res.*, **31**, 3593–3596.
36. Eskin,E. and Pevzner,P.A. (2002) Finding composite regulatory patterns in DNA sequences. *Bioinformatics*, **18(Suppl 1)**, S354–S363.
37. Smith,A.D., Sumazin,P., Das,D. and Zhang,M.Q. (2005) Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics*, **21(Suppl 1)**, i403–i412.
38. Dubnov,S., El-Yaniv,R., Gdalyahu,Y., Schneidman,E., Tishby,N. and Yona,G. (2002) A new nonparametric pairwise clustering algorithm based on iterative estimation of distance profiles. *Mach. Learn.*, **47**, 35–61.
39. Lin,J. (1991) Divergence measures based on the Shannon entropy. *IEEE Trans Inform. Theor.*, **37**, 145–151.
40. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
41. Lawson,C.L., Swigon,D., Murakami,K.S., Darst,S.A., Berman,H.M. and Ebright,R.H. (2004) Catabolite activator protein: DNA binding and transcription activation. *Curr. Opin. Struct. Biol.*, **14**, 10–20.
42. Ebright,R.H., Ebright,Y.W. and Gunasekera,A. (1989) Consensus DNA site for the Escherichia coli catabolite gene activator protein (CAP): CAP exhibits a 450-fold higher affinity for the consensus DNA site than for the E. coli lac DNA site. *Nucleic Acids Res.*, **17**, 10295–10305.
43. Cameron,A.D. and Redfield,R.J. (2008) CRP binding and transcription activation at CRP-S sites. *J. Mol. Biol.*, **383**, 313–323.
44. Robison,K., McGuire,A.M. and Church,G.M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. *J. Mol. Biol.*, **284**, 241–254.
45. Peterson,J.D., Umayam,L.A., Dickinson,T., Hickey,E.K. and White,O. (2001) The Comprehensive Microbial Resource. *Nucleic Acids Res.*, **29**, 123–125.

46. Mahony,S. and Benos,P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, **35**, W253–W258.
47. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
48. Saifudeen,Z., Dipp,S., Fan,H. and El-Dahr,S.S. (2005) Combinatorial control of the bradykinin B2 receptor promoter by p53, CREB, KLF-4, and CBP: implications for terminal nephron differentiation. *Am. J. Physiol. Renal Physiol.*, **288**, F899–F909.
49. Workman,C.T., Yin,Y., Corcoran,D.L., Ideker,T., Stormo,G.D. and Benos,P.V. (2005) enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res.*, **33**, W389–W392.