Bioinformatics Advance Access published July 10, 2013

Original Papers

An integrated toolkit for accurate prediction and analysis of *cis* regulatory motifs at a genome scale

Qin Ma[1,a], Bingqiang Liu[2,a], Chuan Zhou[1,2], Yanbin Yin[3], Guojun Li[1,2], Ying Xu[1,4,5, *]

[1]Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA

[2]School of Mathematics, Shandong University, Jinan 250100, China

[3]Department of Biological Sciences, Northern Illinois University, DeKalb, IL 60115-2857, USA

[4]BioEnergy Science Center (http://bioenergycenter.org/), USA, and

[5]College of Computer Science and Technology, Jilin University, Changchun, Jilin, China

[a]The first two authors contributed equally to this paper

[*] Corresponding author: Ying Xu

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Motivation: We present an integrated toolkit, BoBro2.0, for prediction and analysis of *cis* regulatory motifs. This toolkit can (i) reliably identify statistically significant *cis* regulatory motifs at a genome scale; ii) accurately scan for all motif instances of a query motif in specified genomic regions using a novel method for *p*-value estimation; (iii) provide highly reliable comparisons and clustering of identified motifs, which takes into consideration the weak signals from the flanking regions of the motifs; and (iv) analyze co-occurring motifs in the regulatory regions.

Results: We have carried out systematic comparisons between motif predictions by BoBro2.0 and by the MEME package. The comparison results on *E. coli* K12 genome and the Human genome show that BoBro2.0 can identify the statistically significant motifs at a genome scale more efficiently, identify motif instances more accurately and get more reliable motif clusters than MEME. In addition BoBro2.0 provides correlational analyses among the identified motifs to facilitate the inference of joint regulation relationships of transcription factors.

Availability: The source code of the program is freely available for noncommercial uses at http://code.google.com/p/bobro/.

Contact: Email: xyn@bmb.uga.edu

1 INTRODUCTION

Computational identification of conserved *cis* regulatory motifs represents an important problem in computational genomics, and it can provide a key piece of information for inference of gene regulation networks (Brohee, et al., 2011; Davidson and Levin, 2005). In the last three decades, numerous tools have been developed to find *cis* regulatory motifs in the promoter regions of given genes (Chen, et al., 2008; Das and Dai, 2007; Li, et al., 2011; Sinha, 2007), and have been successfully applied to several organisms to generate large-scale regulatory networks (Baumbach, 2010; Brohee, et al., 2011). In addition, two related problems are also of great interest: (i) *motif scanning* for additional motif instances across a genome based on known or predicted motifs, which needs a reliable measurement for the statistical significance of the scanned motif instances; and high false-positive prediction rates are the main issue with the existing prediction tools; and (ii) *reliability assessment* of the predicted motifs against annotated motifs in motif databases, which requires an effective way to compare the predicted motifs with the documented *cis* regulatory motifs in the published literature and databases. A number of software packages have been developed to deal with such issues (Bailey, et al., 2009; Thomas-Chollier, et al., 2008). For example, the MEME package (Bailey, et al., 2009) was originally developed to identify conserved motifs (Bailey and Elkan, 1994), and now consists of a number of analysis capabilities, such as FIMO (Tanaka, et al., 2011) and MAST (Bailey and Gribskov, 1998) for motif scanning, and TOMTOM (Tanaka, et al., 2011) for motif comparison. These additional capabilities have substantially extended the utility of the MEME program.

While substantial efforts have been invested to study these motif related problems since the mid-80's, they are still largely unsolved, especially for genome-scale applications (Das and Dai, 2007; Stormo, 2000; Tompa, et al., 2005). A number of challenging issues persist and await for better solutions, including (i) more effective ways for reliably assessing the statistical significance of the predicted motifs (Tompa, et al., 2005), hence possibly overcoming the high false-positive rate issue; (ii) improved capabilities for evaluating predicted motif instances to decrease the false positive rates in motif scanning (da Fonseca, et al., 2008; Medina-Rivera, et al., 2011); and (iii) improved similarity measures between two motifs, which currently suffer from the inability to effectively deal with sequence variations in motifs, hence leading to low prediction sensitivities (Tanaka, et al., 2011).

We have recently developed an improved version of our previous tool BoBro (Li, et al., 2011), BoBro2.0, to address some of these issues (see Fig. 1 for the flowchart of BoBro2.0). Compared to BoBro, the new toolkit has a number of novel capabilities: (i) motif refinement and evaluation based on information extracted from the entire genome and a phylogenetic footprinting method, (ii) motif scanning based on a global *p*-value estimation method, (iii) motif comparison and clustering using a novel and effective technique, and (iv) analysis of motifs' co-occurrences in the regulatory regions. The capability of (i) can make the predicted motifs not only statistically significant but also biologically meaningful; the one in (ii) can improve motif-scanning performance in both the prediction precision and recall. The motif comparison and clustering function in (iii) can identify more reliable motif clusters for a given transcription factor, and the co-occurring motif analysis in (iv) can provide useful information about joint regulations among transcription factors. We have assessed the performance of BoBro2.0 in comparison with MEME and associated analysis tools on large test sets spanning genomic sequences of *E. coli* K12 and Human, and found that our program consistently performed better than those programs.

2 METHODS

BoBro2.0 represents an integrated toolkit for motif identification and analysis, including capabilities for motif refining (BBR), motif scanning (BBS), motif comparison and clustering (BBC), and annotation of co-occurring motifs (BBA). Table 1 summarizes the key features of each of these components, and details about applications and references are available in the supplementary material. Generally, a regulatory motif can be represented by different models, e.g. Consensus (Schneider, 2002), Position Weight Matrix (PWM) (Ben-Gal, et al., 2005), or hidden Markov model (HMM) (Baum, 1970), which are all based on aligned motif binding sites. Hence, in the following, we use *motif* to represent a set of aligned similar binding sites (documented or predicted), and use *instance* or *motif instance* to represent each individual site of the motif.

2.1 BBR: a method for filtering out noises among predicted motifs at a genome scale

Consider a genome-scale motif prediction problem: denote all the motifs predicted by a *de-novo* motif finding tool as Ω; *R* and *C* represent the given set of regulatory sequences for motif identification and a control sequence set, respectively. For any motif $m \in \Omega$, it is considered as a motif if it satisfies the following three criteria: (i) the *p*-value of *m* with respect to a hypothesis that it appears in *R* by chance is below a specified cutoff value; (ii) *R* is more enriched of the instances of *m* than *C*, as defined in formula (1); and (iii) *m* is well-conserved across a diverse set of species, as defined in formula (2).

Criterion (i) is measured using the *p*-value defined in our previous work (Li, et al., 2011). Specifically, let *x* be a random variable denoting the number of instances of a motif in a given set of regulatory sequences, and its probability distribution, *p*(*x*), can be approximated using a Poisson distribution. Hence, the *p*-value of a motif can be calculated by summing up the probability of *p*(*x*) over $x \geq k$, denoting that the motif has at least *k* instances. An enrichment score is defined to evaluate the statistical significance of the ratio between the number of *m*'s instances in *R* and that in *C*, as given in the following,

$$Z = \frac{N_R - (|R|*N_C)/|C|}{\sqrt{(|R|*N_C)/|C|}} \qquad (1)$$

where N_R and N_C are the numbers of instances of *m* in *R* and *C*, respectively; and |*R*| and |*C*| are the sequence lengths of *R* and *C*, respectively. Criterion (iii) is defined in terms of the average enrichment score defined in formula (2), with each Z term being defined in (1) for each organism over a set of diverse species and the original genome.

$$\hat{Z} = \frac{1}{|G_r|}\sum_{i \in Gr} Z_i \qquad (2)$$

where G_r represents a set of species and Z_i is *m*'s enrichment score in species *i*, $i \in G_r$. We consider a motif as statistically significant if its *p*-value < 3.3e-5 (the *p*-value threshold has been corrected for multiple testing based on the estimated number, 300, of TFs in *E. coli*). Criteria (ii) and (iii) are designed to ensure that predicted motifs will be as biologically meaningful as possible (Bailey, 2011).

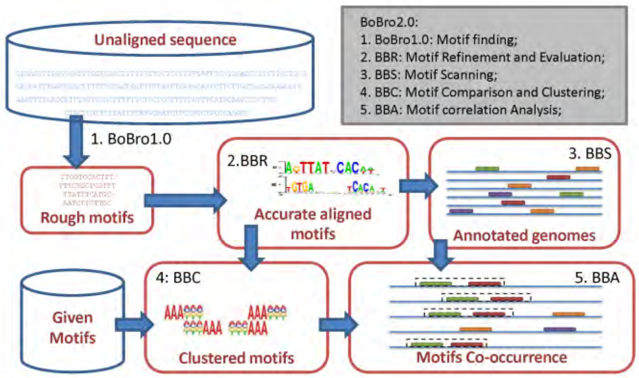

Fig. 1. An outline of the BoBro2.0 toolkit; the blue arrows indicate data flows within BoBro2.0; the red rectangles represent different data status across the whole analysis procedure; and the blue cylinders denote imported information from the other databases.

2.2 BBS: Scanning and ranking new instances of a query motif based on *p*-values

A key to reliable motif-scanning at a genome scale is an ability to effectively evaluate the similarity between a motif instance and a query motif (Das and Dai, 2007; Haverty and Weng, 2004; Medina-Rivera, et al., 2011; Thomas-Chollier, et al., 2008). Obviously different similarity cut-offs may result in different scanning results. BBS provides a global *p*-value for the

Table 1: A comparison of functionalities between BoBro2.0 and MEME; the unique features of BoBro2.0 are listed in the last column.

Functions	MEME	BoBro2.0	Unique features
Motif Refining	N/A	BBR	Strong ability in filtering out noises at a genome scale
Motif Scanning	FIMO	BBS	*p*-value assessment for all the scanned candidate motifs
Motif Comparison	TOMTOM	BBC	1. Utilization of weak conserved signals of motifs' flanking regions when comparing motifs; 2. A motif clustering algorithm
Motif Annotation	N/A	BBA	Motifs' co-occurrence annotation

entire motif instances for each motif scan. We first introduce a few definitions. Let M be an aligned query motif of L nucleotides long and its position weight matrix (PWM) W_M is defined as a 4-by-L matrix, given in (3):

$$W_M = (\log \frac{p_{ij}}{q_i})_{4\times L} \qquad (3)$$

where p_{ij} is the probability of nucleotide $i \in \{A,C,G,T\}$ appearing at position j in M; and q_i is the probability of i appearing in the background sequences, e.g., all the promoter sequences in the entire genome. Comparing with the traditional PWM model that assumes independence among different sequence positions, we assumed 1st order Markov-chain property among consecutive sequence positions in our model. We generated a transition matrix $W_M{}'$, with $W_M{}'(i,i',j)$ representing the probability of a specific nucleotide type i followed by a specific nucleotide type i' in consecutive positions j and j+1 of the query motif (see Supplementary material for details). The similarity between a motif instance $b = \{i_1 i_2 \ldots i_L\}$, $i_j \in \{A,C,G,T\}$ and a query motif M is measured using

$$S_M(b; W_M, W_M^{'}) = \sum_{j=1}^{L} W_M(i,j) + \sum_{j=1}^{L-1} W_M^{'}(i_j, i_{j+1}, j) \qquad (4)$$

Consider a motif M' with t instances $\{a_1, a_2 \ldots., a_t\}$, the average similarity AS(M',M) between M' and M is measured using the following:

$$AS(M', M) = \frac{1}{t}\sum_{i=0}^{t} S_M(b; W_M, W_M^{'}) \qquad (5)$$

A λ-*closure* of M, denoted as $\Omega(M, \lambda)$, is a set of sequence segments in the input regulatory sequences, each having a similarity score no less than λ*AS(M',M). Our previous experience has been that the documented *cis* regulatory motifs tend to have significantly more instances with high similarities among them than the accidental ones; and the size of a λ-*closure* provides a good measure for this (Li, et al., 2011). The p-value $p(M, \lambda)$ of $\Omega(M, \lambda)$ can be approximated using a Poisson distribution based on our previous work (Li, et al., 2011). We can select a λ value λ_M so that the λ-*closure* of M can give the best motif prediction performance measured in terms of prediction sensitivity and specificity. One way to accomplish this is through finding a λ_M that minimizes the following function:

$$p(M, \lambda_M) = \min_{0<\lambda<1} p(M, \lambda) \qquad (6)$$

This capability can be used to derive an optimal similarity cut-off for motif scanning on a statistically sound basis.

2.3 BBC: Motif comparison and clustering

Utilization of weak conserved signals of motifs' flanking regions in motif comparison: We have observed that the flanking regions of *cis* regulatory motifs tend to have some level of sequence conservation. And we have developed the following procedure to take advantage of this information in motif comparison. Define a deformation of *information content* (Schneider, et al., 1986) for a motif M of length L as follows,

$$DIC(M) = \sum_{j=1}^{L}(\sum_{i\in\{A,C,G,T\}} F_M(i,j) * P_M(i,j))^2 \qquad (7)$$

where $F_M = (p_{ij})_{4\times L}$ and the other items are the same as in formula (3). Consider two motifs M_1 and M_2 with lengths L_1 and L_2, respectively, and L= min $\{L_1, L_2\}$. Let $M_1{}'$ and $M_2{}'$ be the two extended motifs formed by concatenating the $\lfloor L_1/2 \rfloor$ and $\lfloor L_2/2 \rfloor$ nucleotides on each side of each motif instance sequence of M_1 and M_2 respectively[1]; hence their lengths are $2L_1$ and $2L_2$. The similarity between the extended instances of M_1 and M_2 is defined as,

$$S(M_1, M_2) = \max_{0\le p\le L_1, 0\le q\le L_2} \frac{DIC_{p,q}(M_1) + DIC_{p,q}(M_2)}{DIC(M_1) + DIC(M_2)} \qquad (8)$$

where

$$DIC_{p,q}(M_1^{'}) = \sum_{j=1}^{L}(\sum_{i\in\{A,C,G,T\}} F_{M_2^{'}}(i,q+j).P_{M_1^{'}}(i,p+j) \times \sum_{i\in\{A,C,G,T\}} F_{M_1^{'}}(i,q+j).P_{M_1^{'}}(i,p+j))$$

and

$$DIC_{p,q}(M_2^{'}) = \sum_{j=1}^{L}(\sum_{i\in\{A,C,G,T\}} F_{M_1^{'}}(i,q+j).P_{M_2^{'}}(i,p+j) \times \sum_{i\in\{A,C,G,T\}} F_{M_2^{'}}(i,q+j).P_{M_2^{'}}(i,p+j))$$

Supplementary Fig. S1 shows an example of motif comparison using this measure, which illustrates the idea of this measure using the information from the motif flanking regions.

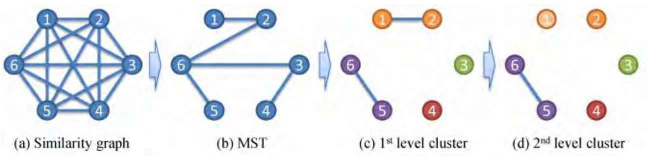

Fig. 2. An example of a two-level clustering of motifs using a minimal spanning tree, consisting of 6 motifs: (a) a complete similarity graph is constructed with the weight of each edge representing the two corresponding motifs' similarity; (b) an MST {(1,2), (2,6), (3,6), (3,4), (5,6)} is constructed using Kruskal's algorithm; (c) four connected components of the MST created using the first-level threshold T_1 , i.e. {1, 2}, {5, 6}, {3} and {4}, reflecting that motifs 1 and 2, 5 and 6 are similar compared to the other motif pairs; (d) the motif cluster {1, 2} is split into two dependent motif clusters {1} and {2} using the threshold T_2, reflecting that the similarity between motifs 1 and 2 are lower than that between motifs 5 and 6.

Motif clustering using the new similarity measure: A group of motifs can be clustered into sub-groups of similar motifs using the following algorithm, which is based on a maximum spanning tree (MST) representation of the candidate motifs. First consider a complete graph defined over a list of candidate motifs, each represented as a node and each pair of motifs connected by an edge; the weight of an edge is the similarity between the two corresponding motifs (see Fig. 2a). An MST of the graph is constructed using Kruskal's algorithm (Thomas H. Cormen, 2001). We have clustered the predicted motifs based on two different similarity thresholds, T_1 and T_2, giving rise to two classes of motif clusters, namely, highly reliable and relatively reliable motif clusters, respectively. We have compared each pair of documented motifs in the RegulonDB database (Salgado, et al., 2013) and assigned the median and the upper quartile of all the similarities to T_1 and T_2, respectively. Each of the two thresholds is used to remove edges with similarities lower than the threshold, giving rise to the final list of motif clusters (see Fig. 2) represented as a connected sub-tree of the MST after application of the threshold. Then, all instances of each motif cluster are mapped back to the original regulatory sequences, facilitating further analysis and interpretation of the motif-prediction results (Supplementary Fig. S2).

2.4 BBA: Motif co-occurrence analysis

We have implemented a function BBA to evaluate the co-occurrences among the identified motifs in a given set of regulatory sequences, which can reveal joint regulation relationships by multiple transcription factors. For a given motif pair, a and b, and the entire set of promoter sequences P, let A and B be the subsets of P that contain motif instances of a and b, respectively (we assume, without loss of generality, |A| ≤ |B|). Let k = |A∩B|; then the probability of A and B sharing k promoter sequences can be calculated using the following hyper-geometric function,

[1] If the location information of given motifs in their original genome is available, we can use the flanking region of each motif to generate the extended motif sequence.

$$\Pr_P(k \mid P, A, B) = \frac{\binom{|B|}{k}\binom{|P \setminus B|}{|A|-k}}{\binom{|P|}{|A|}} \qquad (9)$$

and the *p*-value of *a* and *b* co-occurring in the same regulatory regions is calculated as the probability of *A* and *B* sharing at least *k* regulatory sequences. For a pair of motifs, a significant *p*-value means their instances tend to occur in same regulatory sequences, hence indicating that their corresponding transcription factors may co-regulate the same genes with high probability.

2.5 Data preparation

To test the motif-finding performance of BoBro2.0, we have collected 2,462 promoter sequences (also referred to as regulatory sequences), each being 300 bps long, covering all the predicted operons in *E. coli* K12, which were retrieved from the DOOR database (Dam, et al., 2007; Mao, et al., 2009). 216 bacterial genomes within the same phylum but in different genre of *E. coli* were collected from the NCBI (2011-11-01). In each genus, we selected the largest genome to avoid potential selection bias in comparative genomics studies (Che, et al., 2006). In addition for the motif scanning assessment, we have collected all the known *cis* regulatory motifs of *E. coli* K12 from the RegulonDB database (Salgado, et al., 2013), which has the PWM matrices for 52 transcription factors (TFs) of *E. coli* K12. Out of these TFs, we removed 17 whose *cis* regulatory motifs are known to be not conserved according to a study by Medina-Rivera *et al.* (Medina-Rivera, et al., 2011) and four additional ones that have been reported as nucleoid associated proteins whose *cis* motifs are known to be not conserved, which leaves 31 TFs. In addition to *E. coli* K12, we also collected 1,460 human *cis* regulatory motifs from (Xie, et al., 2005). Further, we retrieved the detailed information of 8 global TFs from RegulonDB, representing the eight largest regulons in the database, namely, CRP, FNR, Fur, LexA, IHF, GntR, PhoP and UlaR, to assess the performance of motif comparison methods.

3 RESULTS

Here we compare BoBro2.0 with the latest version of the MEME suite, a most popular motif-finding and analysis package, in terms of their performance on both prokaryotic and eukaryotic genomes. We found that (i) the predicted motifs by BoBro2.0 have better motif-matching scores and regulon coverage scores than those by MEME; (ii) the average F-score of BBS (0.32) on 31 *E. coli* motifs is significantly higher than that of FIMO (0.14); similar performance results are found in 1,460 human *cis* regulatory motifs (0.34 *versus* 0.15); (iii) BBC can identify more accurate motif clusters than TOMTOM in a constructed motif database; and (iv) BBA can identify jointly regulating TFs which are supported by the published literature. The computational complexity, the actual computing time and selected parameters of each used program can be found in Supplementary Table S1.

3.1 BoBro2.0 can identify *cis* regulatory motifs at a genome scale reliably and efficiently

To assess the motif-finding performance of our toolkit, we have systematically compared BoBro2.0 with MEME on the entire *E. coli* K12 genome. For each program, we take the top 100 predicted motifs as the predictions (the parameters of each program can be found in Table S1). First we note that BoBro is much faster than MEME as it took 2,181 minutes in comparison with 4,492 minutes by MEME to generate the top 100 motifs for the promoter sequences in *E. coli* K12 (both BoBro and MEME are implemented on a computer with 264GB memory and CPU E5-2630 0 @ 2.3 GHz). To highlight the performance of our motif refinement tool, BBR, we have applied it to the motif predictions by both BoBro and MEME, denoted as BoBro+BBR and MEME+BBR, respectively (see Supplementary Appendix 1 for details).

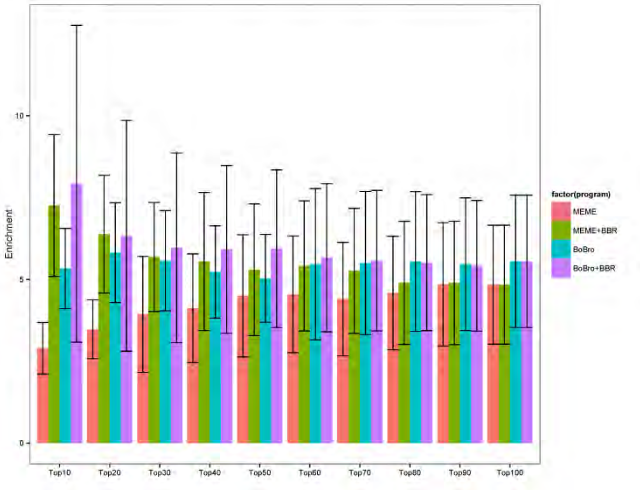

Fig. 3: The *MMSs* comparison between MEME, MEME+BBR, BoBro and BoBro+BBR, where X+BBR means the motif finding tool X combined with our motif refinement tool BBR.

We then compared the motif prediction performance of the four programs, MEME, MEME+BBR, BoBro and BoBro+BBR, in terms of the motif-matching score (*MMS*), defined as follows,

$$MMS = \max_{r \in R}\{\frac{|M_g \cap r_g| / |M_g|}{|r_g| / N}\} \qquad (10)$$

where M_g represents the set of genes in the immediate downstream operon of motif *M*; r_g represents the set of genes in regulon *r*; the set of all the regulons encoded in *E. coli* K12 is denoted as *R*; and *N* is the number of genes in the *E. coli* K12 genome. The *MMS* can be used to infer whether a predicted motif is involved in the regulation of a specific regulon. From the comparison results in Fig. 3, we noted that through top10 to top100, (i) the *MMSs* of the predicted motif by BoBro are significantly higher than that by MEME; (ii) the *MMSs* of the refined motifs by BoBro+BBR and MEME+BBR are higher than predicted motifs by BoBro and MEME, in most cases, respectively. The consensus and enrichment scores for each predicted motif by BoBro and by MEME are shown in Supplementary Table S2.

In addition, we define a regulon coverage score (*RCS*) for each regulon *r* as $(\bigcup_{i=1}^{T} |M_g^i \cap r_g|)/|r_g|$ to measure the coverage of individual operons of a regulon predicted by a prediction program *versus* the known component operons of the regulon, where $M_g^i (1 \le i \le T)$ denotes the predicted gene sets by a program (here *T*=10, 20,…, 100). Note that the larger the *RCS* is, the more component genes of the corresponding regulon are correctly covered by the prediction program. Fig. 4 shows a comparison among the *RCS* values by the four programs on the 12 largest regulons: CRP, Fur, FNR, IHF, Fis, Lrp, CpxR, LexA, NsrR, NarL, Cra and ArcA, each containing at least 20 operons. It is clear that (i) the prediction coverage by BoBro is considerably higher than that by MEME; and (ii) our refinement tool, BBR, can improve the *RCS*s of predicted motifs by both of BoBro and MEME (see details in Supplementary Table S3).

One of the issues that have troubled the motif-finding programs is how to effectively distinguish between *cis* regulatory motifs and the so-called *bacterial-interspersed mosaic elements* (Bachellier, et al., 1999), also known as Repetitive Extra-genic Palindrome (REP) elements (Bachellier, et al., 1999; Keseler, et al.), which are conserved palindromic sequences with various sequence lengths, mostly in the intergenic regions. For example, “CTTATCCGGCCTACAAA” is a key REP pattern in *E. coli* K12. BoBro2.0 can effectively identify such elements when searching for *cis* regulatory motifs through the designed criteria embedded in BBR (see details in METHODS), and filter some of them out although overall the problem remains an unsolved one.

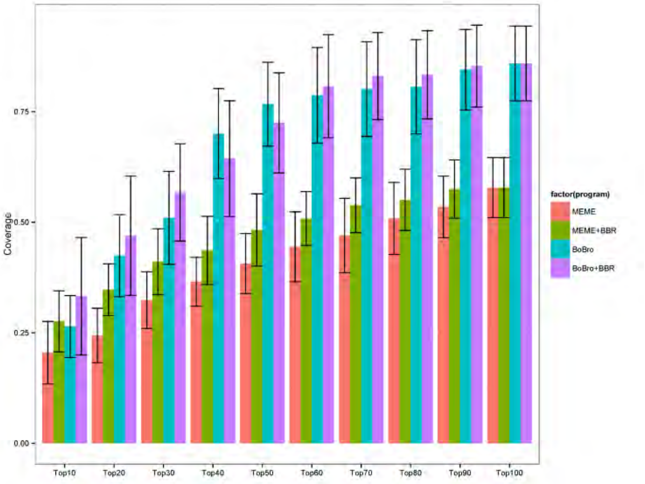

Fig. 4: The *RCSs* comparison between MEME, MEME+BBR, BoBro and BoBro+BBR, where X+BBR mean the a motif finding tool X combined with our motif refinement tool BBR

3.2 BBS can identify motif instances more accurately than FIMO

The BBS provides a global *p*-value for the entire motif prediction when scanning for motif instances at a genome scale, which provides a reliable way for automatically selecting an optimal sequence-similarity threshold for global motif scanning. Our results show that BBS can significantly improve the motif-scanning performance in both *E. coli* K12 and human genomes compared to the FIMO program in MEME. The test set consists of 31 conserved motifs from RegulonDB (see Section 2.5 for details). We used an *F-score* to measure the prediction accuracy (van Rijsbergen, 1979), which is the harmonic mean of *precision* and *recall*,

$$F = \frac{2 * precision * recall}{precision + recall}$$

where *precision* represents the fraction of the predicted motif instances that are documented TF binding sites (Salgado, et al., 2013) (Xie, et al., 2005), and *recall* is the fraction of the documented TF binding sites that are predicted. Fig. 5a shows that the *F-score*s on the 31 TFs by BBS are significantly higher than those by FIMO (see Supplementary Table S4 for details).

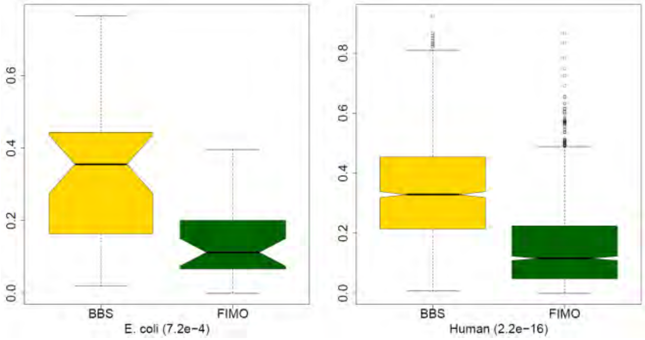

Fig. 5: (a) Performance comparison between BBS and FIMO measured using the *F-score* on 31 TFs of *E. coli* K12. The *p*-value of the Wilcoxon test is 7.2e-4; (b) Performance comparison on a dataset of the human genome. The *p*-value of the Wilcoxon test is 2.2e-16.

In addition, we have also applied BBS to 1,460 regulatory motifs of the human genome, extracted from (Xie, et al., 2005), which are identified in promoters and 3' UTRs by comparative analyses of the human, mouse, rat and dog genomes. Fig. 5b shows a performance comparison on this dataset between BBS and FIMO (see Supplementary Table S5 for details). It is worth noting that the decreased performance by the two programs in comparison with that on *E. coli* K12 is probably due to the higher noise level in the human genome than a bacterial genome. Also the documented motifs representing only a very small fraction of all the encoded *cis* regulatory motifs in the human genome also accounts for the declining in the performance statistics. As shown in Fig. 5, some motifs, scanned by BBS and FIMO, have F-scores close to zero in both human genome and *E. coli* K12, for which the real motifs are known, This is caused by the relatively high false positive rates that all motif scanning programs have to overcome. From the performance statistics, we can see that there is clearly a large room for improvement in human or other complex eukaryotic genomes, possibly by using additional information and techniques.

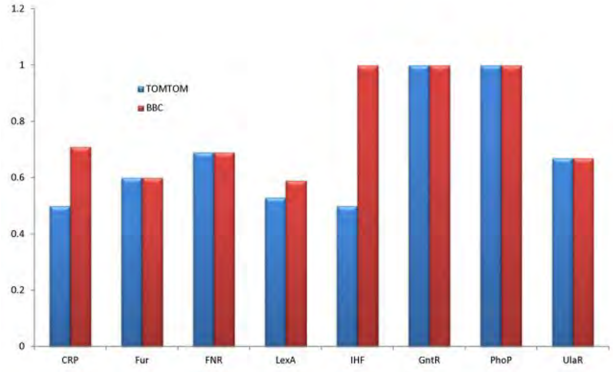

Fig. 6: Comparisons of regulon *BLAST* sensitivity between TOMTOM (blue) and BBC (red).

3.3 BBC can identify more accurate motif clusters than TOMTOM

Motif BLAST is a process for identifying statistically significant motifs in a known motif database, which match given query motifs. We have compared the performance in this area by BoBro2.0 and MEME. We have built a motif dataset using the footprinting technique (Blanchette and Tompa, 2002; Kudla, et al., 2006; Sosinsky, et al., 2007) for the assessment purpose (see Supplementary Material for details). The dataset contains 561 motifs, covering 216 bacterial genomes, for eight global TFs; namely, CRP, FNR, Fur, LexA, IHF, GntR, PhoP and UlaR (see Data preparation). The motifs of these eight TFs in *E. coli* K12 are used as the queries. For any query motif *m*, we define the motif prediction sensitivity as $|B_m \cap I_m|/|I_m|$, where B_m represents all the significant hits of *m* in the collected dataset using a motif comparison program and I_m represents all the implanted motifs of *m* when building this dataset. We compared our method BBC to a recently published program, TOMTOM (Tanaka, et al., 2011), a program of MEME. Fig. 6 shows the comparison results of the identification sensitivity on the eight regulons. BBC has at least as high sensitivity as TOMTOM.

3.4 BBA can identify TFs that jointly regulate genes

It is known that some genes are jointly regulated by multiple transcription factors, and these genes should have the *cis* regulatory motifs of the corresponding TFs, generally arranged in tandem in their promoter sequences (Madan Babu and Teichmann, 2003). We have done statistical analysis on each pair of TFs in *E. coli* K12 to identify such joint regulations. Specifically, we infer that a pair of TFs jointly regulates genes using the motif co-occurrence analysis on all the documented 159 motifs of *E. coli* K12 in RegulonDB (see METHODS). To calibrate the *p*-value distribution, we have run BBA on both documented motifs and randomly simulated motifs (see Supplementary Material for details). Fig. 7 shows the distribution of the *p*-values on both the documented and simulated motifs. We can see that 164 (represented by red bars in Fig. 7) out of all 12,561 pairs of documented motifs have significant *p*-values (less than 0.01) to co-occur in the same promoters. And the comparison with the simulated data (represented by green bars) shows that the *p*-value threshold is significant enough and the chosen pairs are not random noise. Hence, we predict these 164 pairs of TFs jointly regulate gene transcription in *E. coli* K12 and 42 of them have full or partial supporting evidence in the published literature. Table 2 shows the most significant 10 TF pairs and the full list is given in Supplementary Table S6.

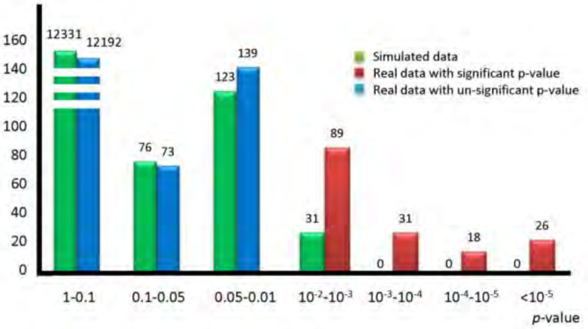

Fig. 7: Distributions of *p*-value for the entire motif pairs among 159 documented *E. coli* K12 motifs and simulated motifs. The number shown on the top of each bar represents the number of motif pairs with corresponding *p*-values on x-axis. The green bars represent the distribution of *p*-value for simulated motif pairs; the red and blue bars are for real data. Motif pairs with *p*-values less than 0.01 (red bars) are considered as significantly co-occurred.

To assess the quality of our predicted TF pairs, we have done an extensive literature search aiming to find published data that may support our predictions. Among the 164 TF pairs, 42 pairs have full or partial supporting evidence with the detailed information given in Supplementary Table S6. We highlight a few examples here. (i) GalS and GalR are known to jointly regulate the *gal* regulon (Geanacopoulos and Adhya, 1997; Weickert and Adhya, 1992); (ii) GadX and GadW are members of the AraC/XylS family of TFs and they collaboratively regulate glutamate-dependent acid resistance in *E. coli* K12 (Gallegos, et al., 1997; Ma, et al., 2002; Martin and Rosner, 2001; Tramonti, et al., 2008); (iii) FNR and CRP belong to the CRP/FNR superfamily of TFs whose members are widely distributed in bacteria; and they have been reported to co-activate the genes involved in reductive dehalogenation of chlorinated aromatic compounds (Gabor, et al., 2006; Korner, et al., 2003); (iv) IHF and Fis are two nucleoid-associated proteins in gram-negative bacteria (Dillon and Dorman) which comprise the prereplication complexes to unwind the origin of replication in *E. coli* K12 (Ryan, et al., 2004); (v) the CytR protein cannot act alone and the synergistic DNA binding is increased by direct interaction with CRP (Sogaard-Andersen, et al., 1990;

Table 2: Top 10 motif pairs with the most significant co-occurrence *p*-values among the total of 12,561 motif pairs of *E. coli* K12; the two motifs in each pair are represented by *a* and *b*, and A and B are the corresponding promoter sets. The *p*-value is for motif co-occurrence.

Motif a	|*A*|	*Motif b*	|*B*|	|*A* ∩ *B*|	*p-value*	*Supporting Literatures*
GalS	5	GalR	5	5	1.33E-15	(Geanacopoulos and Adhya, 1997; Weickert and Adhya, 1992)
FNR	63	ArcA	39	15	3.55E-15	(Cotter and Gunsalus, 1992)
GadX	7	GadW	5	5	2.80E-14	(Gallegos, et al., 1997; Ma, et al., 2002; Tramonti, et al., 2008)
FNR	63	CRP	182	24	2.05E-12	(Gabor, et al., 2006; Korner, et al., 2003)
IHF	57	FNR	63	15	2.27E-12	N/A
IHF	57	Fis	53	13	5.51E-11	(Dillon and Dorman; Ryan, et al., 2004)
CRP	182	ArcA	39	17	3.44E-10	N/A
CytR	8	CRP	182	8	7.72E-10	(Sogaard-Andersen, et al., 1990; Sogaard-Andersen, et al., 1991)
IHF	57	CRP	182	20	9.30E-10	(Bai and Somerville, 1998; Paul, et al., 2007)
FadR	10	ArcA	39	6	2.13E-09	(Cho, et al., 2006)

Sogaard-Andersen, et al., 1990; Sogaard-Andersen, et al., 1991); and (vi) IHF and CRP are known to collaborate to regulate the expression of the tpl promoter and gltBDF operon (Bai and Somerville, 1998; Paul, et al., 2007).

4. CONCLUSION AND DISCUSSION

Compared to the most popular motif analysis software MEME, BoBro2.0 has the following unique and strong features, which all improve the state of the art: it (i) can reliably identify statistically significant *cis* regulatory motifs at a genome scale (ii) provides a reliable way for optimizing the sequence-similarity cut-off in genome-scale motif scanning; (iii) has a reliable capability to compare and cluster motifs, and (iv) can identify TFs that may jointly regulate genes through identification of the co-occurrences of their *cis* regulatory motifs. With these features, we expect that BoBro 2.0 provides a useful tool for motif identification and analysis complementary to the existing tools.

It is worth noting that, based on the above performance on the human genome; BoBro2.0 can be realistically applied to eukaryotic genomes for reliable identification of conserved *cis* regulatory motifs. Specifically BBR is designed to improve the applications of BoBro in eukaryotes. In addition BBA, which is more genome-independent, can clearly be applied to eukaryotes, and we expect that its performance should be about the same as on prokaryotic genomes.

ACKNOWLEDGEMENTS

Funding: This research was supported in part by National Science Foundation (NSF/MCB-0958172, NSF/DEB–0830024 and USG Inter-Institutional Collaborative Grant) and by the U.S. Department of Energy’s BioEnergy Science Center (BESC) grant through the Office of Biological and Environmental Research. GJ Li's work was also supported in part by grants 61070095, 60873207 and 61272016 from NSFC, and ZR2010FM014 from NSF of Shandong Province, China. BQ Liu’s work was supported by grant ZR2011FQ010 and ZR2010AQ018 from NSF of Shandong Province, China.

REFERENCES

Bachellier, S., Clement, J.M. and Hofnung, M. (1999) Short palindromic repetitive DNA elements in enterobacteria: a survey, *Res Microbiol*, **150**, 627-639.

Bai, Q. and Somerville, R.L. (1998) Integration host factor and cyclic AMP receptor protein are required for TyrR-mediated activation of tpl in Citrobacter freundii, *J Bacteriol*, **180**, 6173-6186.

Bailey, T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data, *Bioinformatics*, **27**, 1653-1659.

Bailey, T.L., *et al.* (2009) MEME SUITE: tools for motif discovery and searching, *Nucleic Acids Res*, **37**, W202-208.

Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proc Int Conf Intell Syst Mol Biol*, **2**, 28-36.

Bailey, T.L. and Gribskov, M. (1998) Combining evidence using p-values: application to sequence homology searches, *Bioinformatics*, **14**, 48-54.

Baum, L.E.P., T.; Soules, G.; Weiss, N. (1970) A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains, *The Annals of Mathematical Statistics*, **41**.

Baumbach, J. (2010) On the power and limits of evolutionary conservation--unraveling bacterial gene regulatory networks, *Nucleic Acids Res*, **38**, 7877-7884.

Ben-Gal, I., *et al.* (2005) Identification of transcription factor binding sites with variable-order Bayesian networks, *Bioinformatics*, **21**, 2657-2666.

Blanchette, M. and Tompa, M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting, *Genome Res*, **12**, 739-748.

Brohee, S., *et al.* (2011) Unraveling networks of co-regulated genes on the sole basis of genome sequences, *Nucleic Acids Res*, **39**, 6340-6358.

Che, D., *et al.* (2006) Detecting uber-operons in prokaryotic genomes, *Nucleic Acids Res*, **34**, 2418-2427.

Chen, X., *et al.* (2008) W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data, *Bioinformatics*, **24**, 1121-1128.

Cho, B.K., Knight, E.M. and Palsson, B.O. (2006) Transcriptional regulation of the fad regulon genes of Escherichia coli by ArcA, *Microbiology*, **152**, 2207-2219.

Cotter, P.A. and Gunsalus, R.P. (1992) Contribution of the fnr and arcA gene products in coordinate regulation of cytochrome o and d oxidase (cyoABCDE and cydAB) genes in Escherichia coli, *FEMS microbiology letters*, **70**, 31-36.

da Fonseca, P.G., Guimaraes, K.S. and Sagot, M.F. (2008) Efficient representation and P-value computation for high-order Markov motifs, *Bioinformatics*, **24**, i160-166.

Dam, P., *et al.* (2007) Operon prediction using both genome-specific and general genomic information, *Nucleic Acids Res*, **35**, 288-298.

Das, M.K. and Dai, H.K. (2007) A survey of DNA motif finding algorithms, *BMC Bioinformatics*, **8 Suppl 7**, S21.

Davidson, E. and Levin, M. (2005) Gene regulatory networks, *Proc Natl Acad Sci U S A*, **102**, 4935.

Dillon, S.C. and Dorman, C.J. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression, *Nat Rev Microbiol*, **8**, 185-195.

Gabor, K., *et al.* (2006) Characterization of CprK1, a CRP/FNR-type transcriptional regulator of halorespiration from Desulfitobacterium hafniense, *J Bacteriol*, **188**, 2604-2613.

Gallegos, M.T., *et al.* (1997) Arac/XylS family of transcriptional regulators, *Microbiol Mol Biol Rev*, **61**, 393-410.

Geanacopoulos, M. and Adhya, S. (1997) Functional characterization of roles of GalR and GalS as regulators of the gal regulon, *J Bacteriol*, **179**, 228-234.

Haverty, P.M. and Weng, Z. (2004) CisML: an XML-based format for sequence motif detection software, *Bioinformatics*, **20**, 1815-1817.

Keseler, I.M., *et al.* EcoCyc: a comprehensive database of Escherichia coli biology, *Nucleic Acids Res*, **39**, D583-590.

Korner, H., Sofia, H.J. and Zumft, W.G. (2003) Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs, *FEMS Microbiol Rev*, **27**, 559-592.

Kudla, G., *et al.* (2006) High guanine and cytosine content increases mRNA levels in mammalian cells, *PLoS Biol*, **4**, e180.

Li, G., *et al.* (2011) A new framework for identifying cis-regulatory motifs in prokaryotes, *Nucleic Acids Res*, **39**, e42.

Ma, Z., *et al.* (2002) Collaborative regulation of Escherichia coli glutamate-dependent acid resistance by two AraC-like regulators, GadX and GadW (YhiW), *J Bacteriol*, **184**, 7001-7012.

Madan Babu, M. and Teichmann, S.A. (2003) Evolution of transcription factors and the gene regulatory network in Escherichia coli, *Nucleic Acids Res*, **31**, 1234-1244.

Mao, F., *et al.* (2009) DOOR: a database for prokaryotic operons, *Nucleic Acids Res*, **37**, D459-463.

Martin, R.G. and Rosner, J.L. (2001) The AraC transcriptional activators, *Curr Opin Microbiol*, **4**, 132-137.

Medina-Rivera, A., *et al.* (2011) Theoretical and empirical quality assessment of transcription factor-binding motifs, *Nucleic Acids Res*, **39**, 808-824.

Paul, L., *et al.* (2007) Integration of regulatory signals through involvement of multiple global regulators: control of the Escherichia coli gltBDF operon by Lrp, IHF, Crp, and ArgR, *BMC Microbiol*, **7**, 2.

Ryan, V.T., *et al.* (2004) Escherichia coli prereplication complex assembly is regulated by dynamic interplay among Fis, IHF and DnaA, *Mol Microbiol*, **51**, 1347-1359.

Salgado, H., *et al.* (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more, *Nucleic Acids Res*, **41**, D203-213.

Schneider, T.D. (2002) Consensus sequence Zen, *Applied bioinformatics*, **1**, 111-119.

Schneider, T.D., *et al.* (1986) Information content of binding sites on nucleotide sequences, *J Mol Biol*, **188**, 415-431.

Sinha, S. (2007) PhyME: a software tool for finding motifs in sets of orthologous sequences, *Methods Mol Biol*, **395**, 309-318.

Sogaard-Andersen, L., *et al.* (1990) The CytR repressor antagonizes cyclic AMP-cyclic AMP receptor protein activation of the deoCp2 promoter of Escherichia coli K-12, *J Bacteriol*, **172**, 5706-5713.

Sogaard-Andersen, L., *et al.* (1990) Tandem DNA-bound cAMP-CRP complexes are required for transcriptional repression of the deoP2 promoter by the CytR repressor in Escherichia coli, *Mol Microbiol*, **4**, 1595-1601.

Sogaard-Andersen, L., *et al.* (1991) A novel function of the cAMP-CRP complex in Escherichia coli: cAMP-CRP functions as an adaptor for the CytR repressor in the deo operon, *Mol Microbiol*, **5**, 969-975.

Sosinsky, A., *et al.* (2007) Discovering transcriptional regulatory regions in Drosophila by a nonalignment method for phylogenetic footprinting, *Proc Natl Acad Sci U S A*, **104**, 6305-6310.

Stormo, G.D. (2000) DNA binding sites: representation and discovery, *Bioinformatics*, **16**, 16-23.

Tanaka, E., *et al.* (2011) Improved similarity scores for comparing motifs, *Bioinformatics*, **27**, 1603-1609.

Thomas-Chollier, M., *et al.* (2008) RSAT: regulatory sequence analysis tools, *Nucleic Acids Res*, **36**, W119-127.

Thomas H. Cormen, C.E.L., Ronald L. Rivest, and Clifford Stein (2001) *Introduction to Algorithms*. MIT Press and McGraw-Hill.

Tompa, M., *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites, *Nat Biotechnol*, **23**, 137-144.

Tramonti, A., De Canio, M. and De Biase, D. (2008) GadX/GadW-dependent regulation of the Escherichia coli acid fitness island: transcriptional control at the gadY-gadW divergent promoters and identification of four novel 42 bp GadX/GadW-specific binding sites, *Mol Microbiol*, **70**, 965-982.

van Rijsbergen, C.J. (1979) *Information Retrieval (2nd ed.)*. Butterworth.

Weickert, M.J. and Adhya, S. (1992) Isorepressor of the gal regulon in Escherichia coli, *J Mol Biol*, **226**, 69-83.

Xie, X., *et al.* (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals, *Nature*, **434**, 338-345.