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

Qin Ma1,†, Bingqiang Liu2,†, Chuan Zhou1,2, Yanbin Yin3, Guojun Li1,2 and Ying Xu1,4,5,*
1Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology, Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA, 2School of Mathematics, Shandong University, Jinan 250100, China, 3Department of Biological Sciences, Northern Illinois University, DeKalb, IL 60115-2857, USA, 4BioEnergy Science Center, Oak Ridge, TN, USA and 5College of Computer Science and Technology, Jilin University, Changchun, Jilin, China

Associate Editor: 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 using BoBro2.0 and the MEME package. The comparison results on Escherichia coli K12 genome and the human genome show that BoBro2.0 can identify the statistically significant motifs at a genome scale more efficiently, identify motif instances more accurately and get more reliable motif clusters than MEME. In addition, BoBro2.0 provides correlational analyses among the identified motifs to facilitate the inference of joint regulation relationships of transcription factors.

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

Contact: xyn@bmb.uga.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on January 6, 2013; revised on June 5, 2013; accepted on July 5, 2013

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 past 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.

Although substantial efforts have been invested to study these motif-related problems since the mid-80s, 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) an effective way 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 with 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...
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 other databases.

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 (TF), and the co-occurring motif analysis in (iv) can provide useful information about joint regulations among TFs. 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 BoBro-based motif Refinement tool (BBR), BoBro-based motif Scanning tool (BBS), BoBro-based motif Comparison and Clustering tool (BBC) and BoBro-based motif correlation Analysis tool (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 (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 \( \Omega \), \( R \) 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,

\[
Z = \frac{N_R - (|R|^2N_C)/|C|}{\sqrt{(|R|^2N_C)/|C|}} \tag{1}
\]

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 measured 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.

\[
Z = \frac{1}{|G|} \sum_{i \in G} Z_i \tag{2}
\]

where \( G \) represents a set of species and \( Z_i \) is \( m \)'s enrichment score in species \( i \), \( i \in G \). We consider a motif as statistically significant if its \( P < 3.3 \times 10^{-5} \) (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).

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 cutoffs may result in different scanning results. BBS provides a global \( P \)-value for the 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 PWM \( W_M \) is defined as a 4-by-\( L \) matrix, given in (3):

\[
W_M = \left( \frac{\log P_{\text{ij}}}{q_{\text{ij}}} \right)_{4 \times L} \tag{3}
\]

where \( P_{\text{ij}} \) is the probability of nucleotide \( i \in \{A,C,G,T\} \) appearing at position \( j \) in \( M \); and \( q_{\text{ij}} \) 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 first-order Markov-chain property among consecutive sequence positions in our model. We generated a transition matrix \( W_M' \), with \( W_M'(i,j,f) \) representing the probability of a specific nucleotide type \( f \) 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 \( b = i_1, i_2, ..., i_l \) \( i_j \) \( i_j \in \{A,C,G,T\} \) and a query motif \( M \) is measured using:

\[
S_M(b; W_M, W_M') = \sum_{j=1}^{l-1} W_M(i_j, i_{j+1}) + \sum_{j=1}^{l} W_M'(i_j, f_{j+1}) \tag{4}
\]

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

\[
AS(M', M) = \frac{1}{t} \sum_{i=0}^{t} S_M(a_i; W_M, W_M') \tag{5}
\]

A \( \lambda \)-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
This capability can be used to derive an optimal similarity cut-off for FM where available, we can use the flanking region of each motif to generate the based on our previous work (Li et al., 2011). We can select a \( \lambda \) value \( \lambda_{MC} \) so that the \( \lambda \) 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 \( \lambda_{MC} \) that minimizes the following function:

\[
p(M, \lambda_{MC}) = \min_{0 < \lambda < L} p(M, \lambda)
\]

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:

\[
\text{DIC}(M) = \frac{L}{2} \left( \sum_{i=1}^{L} \sum_{j=1}^{L} F_{M}(i,j) \right)^{2}
\]

where \( F_M = (P_{ij})_{i \leq j} \) 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 \( L_1/2 \) and \( L_2/2 \) nucleotides on each side of each motif instance sequence of \( M_1 \) and \( M_2 \), respectively (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); hence, their lengths are \( 2L_1 \) and \( 2L_2 \). The similarity between the extended instances of \( M_1 \) and \( M_2 \) is defined as follows:

\[
S(M_1, M_2) = \max_{0 \leq L_1' \leq L, 0 \leq L_2'} \frac{\text{DIC}_{p,q}(M_1^{'}) + \text{DIC}_{p,q}(M_2^{'})}{\text{DIC}(M_1) + \text{DIC}(M_2)}
\]

where

\[
\text{DIC}_{p,q}(M_1^{'}) = \sum_{i=1}^{L} \left( \sum_{j=1}^{L} F_{M_1}(i,j) \right) \frac{F_{M_1}(i,j)}{F_{M_1}(i,p)} \left( \sum_{j=1}^{L} F_{M_1}(i,p) \right)
\]

and

\[
\text{DIC}_{p,q}(M_2^{'}) = \sum_{i=1}^{L} \left( \sum_{j=1}^{L} F_{M_2}(i,j) \right) \frac{F_{M_2}(i,j)}{F_{M_2}(i,p)} \left( \sum_{j=1}^{L} F_{M_2}(i,p) \right)
\]

Supplementary Figure 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.

**Motif clustering using the new similarity measure:** A group of motifs can be clustered into subgroups 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, 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 BBCA: motif co-occurrence analysis

We have implemented a function BBCA to evaluate the co-occurrences among the identified motifs in a given set of regulatory sequences, which can reveal joint regulation relationships by multiple TFs. 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, \(|A| \geq |B|\)). Let \( k = \min |A \cap B| \), then the probability of \( A \) and \( B \) sharing \( k \) promoter sequences can be calculated by using the following hyper-geometric function:

\[
Pr_{k}(P, A, B) = \frac{[B]! [P]!}{[P]! [A-k]!}
\]

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 TFs 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 2462 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 (Dan et al., 2007; Mao et al., 2009). Two hundred sixteen 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 genomic 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.

**Table 1. A comparison of functionalities between BoBro2.0 and MEME**

<table>
<thead>
<tr>
<th>Functions</th>
<th>MEME</th>
<th>BoBro2.0</th>
<th>Unique features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motif refining</td>
<td>N/A</td>
<td>BBR</td>
<td>Strong ability in filtering out noises at a genome scale</td>
</tr>
<tr>
<td>Motif scanning</td>
<td>FIMO</td>
<td>BBS</td>
<td>( P )-value assessment for all the scanned candidate motifs</td>
</tr>
<tr>
<td>Motif comparison</td>
<td>TOMTOHM</td>
<td>BBC</td>
<td>(i) Utilization of weak conserved signals of motifs’ flanking regions when comparing motifs; (ii) A motif clustering algorithm</td>
</tr>
<tr>
<td>Motif annotation</td>
<td>N/A</td>
<td>BBA</td>
<td>Motifs’ co-occurrence annotation</td>
</tr>
</tbody>
</table>

**Note:** The unique features of BoBro2.0 are listed in the last column.
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 Supplementary Table S1). First, we note that BoBro is much faster than MEME, as it took 2181 min in comparison with 4492 min by MEME to generate the top 100 motifs for the promoter sequences in E.coli K12. 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. (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 1460 human cis-regulatory motifs from Xie et al. (2005). Further, we retrieved the detailed information of eight 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 (RCSs) 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 1460 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 Supplementary Table S1). First, we note that BoBro is much faster than MEME, as it took 2181 min in comparison with 4492 min 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).

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:

$$MMS = \max_{r \in K} \frac{\left| M_g \cap r \right|}{r_g / N}$$

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 Figure 3, we noted that through top 10 to top 100, (i) the MMSs of the predicted motif by BoBro are significantly higher than that by MEME, and (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 RCS for each regulon $r$ as $\frac{\left(\bigcup_{i=1}^{T} M_g^{i} \cap r \right)}{r_g}$ to measure the coverage of individual operons of a regulon predicted by a prediction program versus the known component operons of the regulon, where $M_g^{i} (1 \leq i \leq T)$ denotes the predicted gene sets by a program $(T=10, 20, \ldots, 100)$. Note that the larger the RCS, the more component genes of the corresponding regulon are correctly covered by the prediction program. Figure 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 RCSs of predicted motifs by both 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 palindromic elements (Bachellier et al., 1999; Keseler et al., 2011), which are conserved palindromic sequences with various sequence lengths, mostly in the intergenic regions. For example, ‘CTTATCCGGCCTACAAA’ is a key repetitive extra-genic palindromic 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 section), and filter some of them out, although overall the problem remains an unsolved one.

### 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 with 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,

\[
F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

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

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. Figure 5b shows a performance comparison on this dataset between BBS and FIMO (see Supplementary Table S5 for details). It is noteworthy that the decreased performance by the two programs in comparison with that on E.coli K12 is probably owing 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 account for the decline in the performance statistics. As shown in Figure 5, some motifs, scanned by BBS and FIMO, have F-scores close to 0 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.

### 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 the 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 8 global TFs, namely, CRP, FNR, Fur, LexA, IHF, GntR, PhoP and UlaR (see Section 2.5). 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 with a recently published program, TOMTOM (Tanaka et al., 2011), a program of MEME. Figure 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 TFs, 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 section). To calibrate the P-value distribution, we have run BBA on both documented motifs and randomly simulated motifs (see Supplementary Material for details). Figure 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) of all 12,561 pairs of documented motifs have significant P-values (<0.01) to co-occur in the same promoters. 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.

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 Dormann, 2010), which comprise the pre-replication 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., 1990a,b, 1991); and (vi) IHF and CRP are known to collaborate to regulate the expression of the tpi promoter and gltBDF operon (Bai and Somerville, 1998; Paul et al., 2007).

4 CONCLUSION AND DISCUSSION

Compared with the most popular motif analysis software MEME, BoBro2.0 has the following unique and strong features, which all improve the state-of-the-art: (i) can reliably identify statistically significant cis-regulatory motifs at a genome scale; (ii) provides a reliable way for optimizing the sequence-similarity cutoff 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 noteworthy 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.

Funding: This research was supported in part by National Science Foundation (NSF/MCB-0958172, NSF/DEB-0830024...
Table 2. Top 10 motif pairs with the most significant co-occurrence P-values among the 12 561 motif pairs of E.coli K12

| Motif a | | Motif b | | P-value | | Supporting literatures |
|---------|-----|--------|-----|---------|-------------------------|
| GalS | 5 | GalR | 5 | 1.33E-15 | Geanacopoulos and Adhya, 1997; Weickert and Adhya, 1992 |
| FNR | 63 | ArcA | 39 | 3.55E-15 | (Cotter and Gunsalus, 1992) |
| GradX | 7 | GradW | 5 | 2.80E-14 | (Gagliogli et al., 1997; Ma et al., 2002; Tramonti et al., 2008) |
| FNR | 63 | CRP | 182 | 2.05E-12 | (Gabor et al., 2006; Kornert et al., 2003) |
| IHF | 57 | Fis | 53 | 5.51E-11 | N/A |
| IHF | 57 | CRP | 182 | 3.44E-10 | N/A |
| CRP | 182 | ArcA | 39 | 3.44E-10 | N/A |
| CytR | 8 | CRP | 182 | 7.72E-10 | N/A |
| IHF | 57 | CRP | 182 | 9.30E-10 | N/A |
| FadR | 10 | ArcA | 39 | 2.13E-09 | N/A |

Note: 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.

REFERENCES


