Sequence analysis

MotifVoter: a novel ensemble method for fine-grained integration of generic motif finders
Edward Wijaya\textsuperscript{1,2}, Siu-Ming Yiu\textsuperscript{3}, Ngo Thanh Son\textsuperscript{1}, Rajaraman Kanagasabai\textsuperscript{2} and Wing-Kin Sung\textsuperscript{1,4,*}

\textsuperscript{1}School of Computing, National University of Singapore, Singapore 119260, \textsuperscript{2}Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, \textsuperscript{3}Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong and \textsuperscript{4}Genome Institute of Singapore, 60 Biopolis Street, #02-01 Genome, Singapore 138672

Received on May 9, 2008; revised on August 3, 2008; accepted on August 7, 2008
Advance Access publication August 12, 2008
Associate Editor: Alex Bateman

\textbf{ABSTRACT}

\textbf{Motivation:} Locating transcription factor binding sites (motifs) is a key step in understanding gene regulation. Based on Tompa’s benchmark study, the performance of current \textit{de novo} motif finders is far from satisfactory (with sensitivity $\leq 0.222$ and precision $\leq 0.307$). The same study also shows that no motif finder performs consistently well over all datasets. Hence, it is not clear which finder one should use for a given dataset. To address this issue, a class of algorithms called ensemble methods have been proposed. Though the existing ensemble methods overall perform better than stand-alone motif finders, the improvement gained is not substantial. Our study reveals that these methods do not fully exploit the information obtained from the results of individual finders, resulting in minor improvement in sensitivity and poor precision.

\textbf{Results:} In this article, we identify several key observations on how to utilize the results from individual finders and design a novel ensemble method, MotifVoter, to predict the motifs and binding sites. Evaluations on 186 datasets show that MotifVoter can locate more than 95% of the binding sites found by its component motif finders. In terms of sensitivity and precision, MotifVoter outperforms stand-alone motif finders and ensemble methods significantly on Tompa’s benchmark, \textit{Escherichia coli}, and ChIP-Chip datasets. MotifVoter is available online via a web server with several biologist-friendly features.

\textbf{Availability:} http://www.comp.nus.edu.sg/\textquotesingle bioinfo/MotifVoter
\textbf{Contact:} ksung@comp.nus.edu.sg
\textbf{supplementary information:} Supplementary data are available at Bioinformatics online.

1 \textbf{INTRODUCTION}

Understanding the regulatory mechanism of genes is one of the major challenges faced by biologists today. Central to this problem is the identification of recurring patterns (motifs) in regulatory sequences, which represent binding sites of a transcription factor. In general, one would like to identify transcription factors whose binding sites are statistically over-represented in the promoter of a set of co-regulated genes. Many motif finders have been proposed to undertake this problem using different approaches, such as profile-based methods (Nimwegen, 2007) and consensus-based methods (Eskin and Pevzner, 2002; Pavesi \textit{et al.}, 2001; Wijaya \textit{et al.}, 2007). However the existing tools are still not effective for discovering motifs (Das and Dai, 2007; Tompa \textit{et al.}, 2005). For example, as shown in Tompa \textit{et al.}’s evaluation, even the best performing algorithm has sensitivity $<13\%$ and precision $<35\%$ (sensitivity is percentage of true nucleotides that are predicted and precision is the percentage of predicted nucleotides that are true).

To deal with this issue, some literatures (Harbison \textit{et al.}, 2004; Hu and Kihara, 2005; MacIsaac and Fraenkel, 2006) hinted that assembling the results of multiple motif finders can provide better results for motif finding. For example, Harbison \textit{et al.} (2004) observed different motif finders have different strengths. They successfully identified more binding sites by combining results of six motif finders compared to using only single finder. In fact, the benchmark datasets from Tompa \textit{et al.} (2005) also support this. By simply taking the union of all binding sites predicted by 10 selected motif finders, the sensitivity can be increased by more than double over each selected motif finder. However, the union of all predicted sites could contain a lot of noise. It is not trivial to distinguish the real binding sites from the noise.

Six ensemble methods have been developed. SCOPE (Chakravarty \textit{et al.}, 2007) and BEST (Dongsheng \textit{et al.}, 2005; Jensen and Liu, 2006) rerank all motifs predicted by individual finders based on a certain scoring function and report the top motif. WebMotifs (Gordon \textit{et al.}, 2005; Romer \textit{et al.}, 2007), MultiFinder (Huber and Bulyk, 2006) and RGSMiner (Huang \textit{et al.}, 2004) assume motifs given by the consensus of several motif finders are likely to be the real motifs. They cluster the motifs and report one motif from the best cluster. All above methods select one representative motif among the predicted motifs from the individual finders. If none of the motifs from the finders can capture the binding sites accurately, the performance of these methods will suffer. EMD (Hu \textit{et al.}, 2006) goes a step further and considers the binding sites predicted by the motifs. Motifs of the same rank from different component motif finders are grouped together. For each group, based on the binding sites predicted by all the motifs...
in the group, new motifs are derived. Though these ensemble methods indeed help to improve the performance of motif finding, the improvement is not significant. For example, in Tompa’s benchmark and Escherichia coli datasets, the average sensitivity is only improved by 62% but the average precision is reduced by 15%. We believe that some important aspects have been overlooked by existing ensemble methods:

1. Existing ensemble methods make use of the observation that true motifs are shared by multiple motif finders, i.e. the coincidence of the predictions from multiple finders. However, the (negative) information from the false positive (spurious) motifs is not exploited in existing approaches. These spurious motifs, which are usually only reported by a few or even one motif finder, provide additional information to guide the identification of good motifs.

2. Although existing ensemble methods assume that true motifs will be shared by multiple motif finders, they do not explicitly use the fact that the motifs (binding sites) predicted by more motif finders have a higher chance to be the correct ones.

These observations are also supported by Hu and Kihara (2005) and MacIsaac and Fraenkel (2006). In this article, we show how to formalize these observations and derive two criteria, namely discriminative and consensus criteria, to combine results from multiple motif finders.

Extensive evaluation shows that our proposed ensemble method, MotifVoter, can extract almost all information from multiple motif finders. On 186 datasets for Tompa’s benchmark (Tompa et al., 2005) and E.coli (Salgado et al., 2004), MotifVoter retains more than 95% of true binding sites predicted by at least one individual motif finder. This implies that MotifVoter improves the sensitivity by 120% and precision by 77% when compared with the best individual motif finder and it also shows an improvement of 90% in sensitivity and 135% in precision when compared to the existing ensemble methods. Applying MotifVoter on 140 yeast and mammalian ChIP-Chip datasets, we demonstrate that MotifVoter is scalable. Furthermore, in these ChIP-Chip datasets, we notice that the motifs reported by MotifVoter are more similar to true motifs when compared with those reported by existing stand-alone motif finders and ensemble methods.

2 APPROACH

Figure 1 shows the overall framework of MotifVoter. Given a set of sequences, MotifVoter executes \( m \) different motif finders, each reporting \( n \) motifs. Note that each motif also defines a set of predicted binding sites. MotifVoter comprises two stages. (1) Motif filtering: in the first stage, MotifVoter processes the candidate motifs predicted by the \( m \) motif finders and attempts to remove the spurious motifs (Fig. 1a). The main idea is to find a cluster of motifs with high conformity based on a motif similarity measure (defined below). (2) Sites extraction: based on the candidate motifs retained in Stage 1, MotifVoter then identifies a set of sites with high confidence that they are real (Fig. 1b). At the end, for visualization purposes, we generate the PWM and the Weblogo of the motifs (Fig. 1c).

![Figure 1](https://academic.oup.com/bioinformatics/article-abstract/24/20/2288/258515)

**Fig. 1.** (a) We apply \( m \) motif finders on the input sequences. Each motif finder reports a set of motifs and their respective sets of binding sites. The pairwise similarities among predicted motifs can be visualized in a graph where each node represents a motif and the similarity between two motifs is represented by a weighted edge. We expect the clustered motifs to approximate the real motif while the rest of the motifs are spurious. Stage 1 aims to find a cluster such that (1) the motifs in the cluster are given by as many motif finders as possible and (2) the motifs inside the cluster are close (similar) to one another, and those outside the cluster are distant (different) from one another. (b) The figure shows the ideas behind the second stage of MotifVoter. Real binding sites are represented by blue colors. The remaining colors represent the binding sites predicted by four other motif finders. As shown by the framed binding sites, MotifVoter can discover more true binding sites compared to individual motif finders. Furthermore, using a confidence measure, MotifVoter aims to detect the true binding site which can only be discovered by one motif finder (represented by purple sites). (c) After extracting the binding sites, we align them using MUSCLE and generate a PWM to model the motif. The diagram shows the resulting weblogo of the example.
It may be noted that, unlike other ensemble methods that only work on the predicted motifs, MotifVoter and EMD work on both the predicted motifs and binding sites. MotifVoter differs from EMD in the discriminative and consensus criteria used for grouping (clustering) the motifs. First, the discriminative criterion requires the selected cluster to of motif share as many binding sites as possible, while requiring that motifs outside the cluster share none or few binding sites. Second, the consensus criterion requires the motifs in the selected cluster to be contributed by as many motif finders as possible. We propose a scoring function that captures the two criteria and derive an efficient algorithm to select a cluster that optimizes this scoring function. In addition to these two criteria, MotifVoter selects only those binding sites of high confidence to construct the new motif, rather than using all binding sites predicted by motifs in the same group (cluster).

Compared to existing ensemble schemes, MotifVoter has the following advantages:

1. Existing ensemble methods infer the final motif from sites of either one motif or a set of motifs of the same ranking. MotifVoter, on the other hand, infers the final motif using sites of high confidence gathered from all predicted motifs of multiple motif finders. Hence, MotifVoter has potential to extract more true sites than existing ensemble methods.

2. MotifVoter can potentially yield higher precision than existing ensemble methods by the following reasoning. The discriminative criterion helps to select a cluster of motif that are not only similar (i.e. share as many binding sites as possible), but also have the property that motifs outside the cluster are distant from one another. At the same time, the consensus criterion ensures that the selected cluster of motifs is predicted by as many motif finders as possible. These two criteria help to remove almost all spurious motifs. MotifVoter further improves the precision by removing spurious sites in the selected cluster.

3 METHODS

MotifVoter uses 10 basic motif finders. The first group consists of three motif finders based on (l, d) model, namely MITRA (Eskin and Pevzner, 2002), Weeder (Pavesi et al., 2001), and SPACE (Wijaya et al., 2007). The second group consists of seven motif finders based on PWM model, namely AlignACE (Hughes et al., 2000), ANN-Spec (Workman and Stormo, 2000), BioProspector (Liu et al., 2001), Improbizer (Ao et al., 2004), MDScan (Liu et al., 2002), MEME (Bailey and Elkan, 1995) and MotifSampler (Thijs et al., 2001). The details of these component motif finders and parameter descriptions can be found in Supplementary Material 10.

Section 3.1 formally describes the discriminative and consensus criteria in the motif filtering step. Then, in Section 3.2, we present the heuristic algorithm for finding a cluster of motifs that satisfy the two criteria. The details of the sites extraction step are given in Section 3.3. Finally, Section 3.4 describes the performance evaluation measures we used for benchmarking MotifVoter against other existing methods.

3.1 The discriminative and consensus criteria

Given two motifs x and y, we first describe how to measure their similarity based on the binding sites defined by these motifs. Let I(x) be the set of binding sites defined by motif x. Let I(x) ∩ I(y) be the set of regions covered by at least one site in x and one site in y. Let I(x) ∪ I(y) be the set of regions covered by any site of x or y. We denote the total number of nucleotides of all the regions in I(x) ∩ I(y) and I(x) ∪ I(y), as |I(x) ∩ I(y)| and |I(x) ∪ I(y)|, respectively. The similarity of x and y, denoted sim(x, y), is expressed as |I(x) ∩ I(y)|/|I(x) ∪ I(y)|. Note that 0 ≤ sim(x, y) ≤ 1 and sim(x, x) = 1.

Now, we define the scoring function to capture the discriminative criterion. Given m motif finders and each motif finder reports its top-n candidate motifs, there will be a set P of mn candidate motifs. Among all the candidate motifs in P, some of them will approximate the real motif while the other will not. We would like to identify the subset X of P such that the candidate motifs in X are likely to approximate the real motif. The principle idea is that if the candidate motifs in X can model the real motif, motifs in X should be highly similar while motifs outside X should be distant from one another (discriminative criteria).

Let X be some subset of candidate motifs of P. The mean similarity among the candidate motifs in X, denoted as sim(X), is defined as:

\[ \sum_{i,j \in X} \text{sim}(x_i, y_j) / |X|^2 \]

The \( w \) score of X, denoted by \( \text{w}(X) \), is defined as:

\[ \sqrt{\sum_{i,j \in X} (\text{sim}(x_i, y_j) - \text{sim}(X))^2} \]

The function \( w(X) \) measures the degree of similarity among the candidate motifs in X. If many of the candidate motifs in X approximate the real motif, we should expect to have a high \( w(X) \). On the other hand, we expect the complement of X, that is \( X - X \), to have a low \( w(P - X) \). Thus, \( w(P - X) \) constitutes the discriminative criterion in the clustering procedure. In other words, if X is the set of candidate motifs which approximates the real motif, we expect to have a high \( A(X) \) score, where:

\[ A(X) = \frac{\text{w}(X)}{\text{w}(P - X)} \]

Note that there may be multiple sets of X with the same \( A(X) \) score. Among those X’s, we would select one that contains maximum number of motif finders. The latter constitutes the consensus criterion.

In summary, this stage aims to find \( X \subseteq P \) which maximizes \( A(X) \) while \( X \) contains the candidate motifs predicted by maximum number of motif finders. The naive method to identify X is to enumerate all possible X and check if they satisfy the above two criteria. However, this approach is computationally infeasible. In the next section, we describe our proposed heuristics to identify X to overcome this difficulty.

3.2 Heuristics for motif filtering

In this subsection, we describe the heuristic algorithm for identifying X, the set of similar candidate motifs in the first stage.

Let P be all motifs found by m motif finders, where each motif finder returns n motifs. Steps 1–3 compute the pairwise similarity scores for all pairs of motifs. Based on these similarity scores, we apply the following heuristics approach to find X (Steps 4–9).

Instead of enumerating all possible subsets in P which takes exponential time, we only consider subsets \( X_{i,j} = \{z \in P_i \} \) for every \( z \in P \) and for every \( 1 \leq i \leq |P| - 1 \) where \( \text{sim}(z, P_i) > \text{sim}(z, P_{i+1}) \) and \( P \subseteq P \). The rationale behind examining only these subsets is as follows. For each \( z \in P \), let a and b be two other motifs such that \( \text{sim}(a, z) > \text{sim}(b, z) \). That is, a is more similar to z than b. Then, it is likely that \( A(z, a, b) > A(z, b) \). By including \( P_i \) in the subset, we also include all \( P_j \) with \( i < j \).

For each \( X_{i,j} \), we compute the value of \( A(X_{i,j}) \) and the number of motif finders contributing to \( X_{i,j} \), denoted as \( Q(X_{i,j}) \). Finally, we set X to \( X_{i,j} \) with maximum \( A(X_{i,j}) \) value. If there are more than one such subset, pick the one with the largest \( Q(X_{i,j}) \). In case of a tie again, we pick one randomly.

The time complexity of heuristics can be shown to be \( O(m^2 n^2) \) as follows. There are \( |P| = mn \) motifs. For each motif z, we consider mn different
Algorithm 1 MotifVoter

Input: $P, |P| = k$
Output: PWM of the aligned sites
1: for each $z, y \in P$ do
2: compute $\text{sim}(z, y)$
3: end for
4: for each $z \in P$ do
5: sort $p_1, p_2, \ldots, p_k$ such that $\text{sim}(z, p_i) \geq \text{sim}(z, p_{i+1}); \forall i = 1, \ldots, k$
6: consider sets $X_{z,j} = \{z, p_1, \ldots, p_j\}; j = 1, \ldots, k$
7: compute $A(X_{z,j})$ for all such $X_{z,j}$
8: set $Q(X_{z,j}) = \text{number of motif finders contributing to } X_{z,j}$
9: end for
10: set $X = X_{z,j}$ with the maximum $A(X_{z,j})$ score, if there are two such $X_{z,j}$’s, pick the one with the largest $Q(X_{z,j})$
11: extract and align sites of motifs in $X$
12: construct PWM

$X_{z,j}$ subsets. For each subset, we need to compute the value of $A(X_{z,j})$ for all $j$. Note that we can obtain the value of $A(X_{z,j})$ from the value of $A(X_{z,j-1})$ in constant time. So, for each motif $z$, we compute all values of $A(X_{z,j})$, it takes $O(mn^2)$ time. Therefore, the overall time complexity is $O(mn^2n^2)$. For the actual running time of the heuristics, please refer to Supplementary Material 5.

3.3 Sites extraction

From $X$, we obtain the list of sites using two requirements. First, we accept all regions which are covered by sites defined by at least two motifs $x$ and $y$ in $X$, where $x$ and $y$ are predicted by two different motif finders. The reason behind is that it is unlikely that several motif finders predict the same spurious binding sites.

Second, we accept all the sites of the motif $x$ in $X$ with the highest confidence in approximating real motif. The confidence score is defined as follows. Let $B(x)$ be the total number of nucleotides covered by the sites defined by motif $x$. Let $O(x)$ be the total number of nucleotides covered by the sites defined by motif $x$ and all other motifs in $X$. The confident score of motif $x$ is defined as $O(x)/B(x)$.

For the selected sites that are covered by more than one motif finder, we further apply a post-processing procedure to refine each site by removing the nucleotides that are only covered by a single finder to increase the precision of our prediction as these nucleotides are likely to be noise.

Given all the sites predicted by MotifVoter, we generate a PWM motif to model the motif. To achieve this, we first align these sites using MUSCLE (Edgar, 2004), then a PWM is generated from this alignment to model the motif. To achieve this, we first align these sites using MUSCLE (Edgar, 2004), then a PWM is generated from this alignment to model the motif.

For each motif $z$, we compute all values of $A(X_{z,j})$, it takes $O(mn^2)$ time. Therefore, the overall time complexity is $O(mn^2n^2)$. For the actual running time of the heuristics, please refer to Supplementary Material 5.

3.4 Performance evaluation measures

For performance evaluation, we use the same four measures proposed in (Tompa et al., 2005) namely, sensitivity (nSN), positive predictive value$^1$ (nPPV), performance coefficient (nPC) and correlation coefficient (nCC). Index $n$ is used to denote that the assessment is done at the nucleotide level instead of site level. The definitions of these performance measures can be found in Supplementary Material 4.

For ChIP-Chip datasets which do not have sites position information, we use sum of squared distance (SSD) to evaluate the performance of motif finders. It is one of the most widely used methods for comparing PWMs (Mahony et al., 2007). Let $L$ be length of two aligned motifs $X$ and $Y$, SSD is defined as:

$$SSD(X, Y) = \frac{1}{T} \sum_{i=1}^{L} \sum_{b=h}^{T} (p_i(b) - p_y(b))^2$$

$^1$Throughout the article, we use the term precision instead of positive predictive value.

where $p_i(b)$ and $p_y(b)$ are the probabilities of base $b$ occurring at position $j$ in motif $X$ and $Y$, respectively. Note that $0 \leq SSD(X, Y) \leq 2$ and $SSD(X, X) = 2$. If the motifs are of different length, we use the shorter motif to align different regions of the longer motif and obtain the maximum SSD value.

4 EXPERIMENTAL RESULTS

There is no universal benchmark dataset for evaluating a motif finder. We thus performed an extensive evaluation on MotifVoter over 326 datasets from Tompa’s benchmark, E.coli and ChIP-Chip datasets. In our experiments, we consider the top-30 (i.e. $n = 30$) motifs reported by each individual motif finder. For the effect of different values of $n$, please refer to Supplementary Material 6.

4.1 Comparison using Tompa’s benchmark dataset

This section compares MotifVoter with existing methods using Tompa’s benchmark datasets. The datasets are constructed based on real transcription factor binding sites drawn from four different organisms (human, fruitfly, mouse and yeast). The background sequences are based on three categories: (1) real upstream sequences (real), (2) randomly chosen upstream sequences with binding sites inserted (generic) and (3) sequences randomly generated by a Markov chain with binding sites inserted (markov). It consists of 56 datasets. The number of sequences per dataset ranges from 1 to 35 and the sequence lengths are up to 3000 bp (Tompa et al., 2005). Such diverse characteristics of the benchmark make it a good candidate for evaluating the robustness of a motif finder.

The graph in Figure 2a shows the average performance of all stand-alone and ensemble motif finders based on four performance evaluation measures [see also Table 1 for the actual figures of sensitivity (nSN) and precision (nPPV)] using Tompa’s benchmark dataset. Among the stand-alone motif finders, Weeder and SPACE perform somewhat similar and outperform the other 15 stand-alone motif finders by all four measures. However, integration of motif finders enables MotifVoter to further enhance the performance. The improvement gained over SPACE is 215% in sensitivity and 45% in precision. We believe that MotifVoter is able to piece together different subregions of sites identified by the stand-alone motif finders, thus yielding significant improvement in sensitivity. Furthermore, since MotifVoter imposes strict discriminative and consensus criteria on the clustering procedure, we get significant improvement in precision also.

For ensemble methods, MotifVoter outperforms the sensitivity of BEST and precision of SCOPE by 54.1% and 226.2%, respectively. We cannot directly evaluate RGMSMiner and MultiFinder because of their limited applicability. For WebMotifs, it only takes probe names as input, so we only evaluate WebMotifs on ChIP-Chip dataset (see Section 4.3). In principle, these three ensemble methods select one motif out of all motifs reported by the component motif finders. Although we are not able to run these finders on the dataset, to predict the upper bound of the performance of these finders, we can select the most sensitive and precise motif for each dataset. We found that even if we do so, the sensitivity is at least 48.6% lower than MotifVoter and the precision is at least 41.2% lower.

For EMD, a direct evaluation is done in the next subsection on E.coli dataset. However, in this benchmark if we cluster the motifs...
of the same rank given by 10 motif finders and pick the most sensitive cluster, the highest possible sensitivity is 21.3% lower than MotifVoter. The main reason why MotifVoter yields better performance is because apart from the ability to capture true sites from various motif finders, it also includes true sites that may come from different ranks.

4.2 Comparison using E.coli dataset

Despite its comprehensive assortments of datasets, Tompa’s benchmark does not fully represent real biological settings since it still contains synthetic sequences (markov) and mixture of real and synthetic (generic) sequences. Here we evaluate MotifVoter using real E.coli datasets. This species is not included in Tompa’s benchmark. The datasets are taken from RegulonDB (Salgado et al., 2004) (http://regulondb.ccg.unam.mx/), and the sequences are generated from the intergenic regions of E.coli genome. In total, there are 62 datasets. The average number of sequences per dataset is 12 and the average sequence length is 300 bp.

The graph in Figure 2b shows the average performance of all stand-alone and ensemble motif finders based on four performance evaluation measures [see also Table 1 for the actual figures of sensitivity (nSN) and precision (nPPV)] using the E.coli dataset. MotifVoter consistently outperforms the stand-alone motif finders by about 168% in terms of sensitivity. For precision, although the improvement is not so much as for sensitivity, MotifVoter still yields a significant improvement over stand-alone motif finders exceeding the highest precision by 80.9%.

Three ensemble methods (SCOPE, BEST, EMD) evaluated on this dataset outperform the stand-alone algorithms in terms of sensitivity. However, even for the most sensitive ensemble method (EMD) among the three, the improvement is not significant. On the other hand, MotifVoter outperforms the EMD ensemble method by 130.2% in sensitivity. As for precision, MotifVoter also consistently performs better with a 45.9% improvement over the best ensemble method (SCOPE).

4.3 Application on ChIP-Chip datasets

In this section, we evaluate MotifVoter on yeast and mammalian ChIP-Chip datasets. These datasets provide an ideal platform for assessing the scalability and applicability of MotifVoter to the entire genome.

| Table 1. The table lists the average sensitivity (nSN) and precision (nPPV) of each motif finder for Tompa’s Benchmark and E.coli datasets |
| | Tompa | E.coli |
| | nSN | nPPV | nSN | nPPV |
| QuickScore | 0.017 | 0.030 | – | – |
| Consensus | 0.021 | 0.113 | – | – |
| GLAM | 0.026 | 0.038 | – | – |
| MITRA | 0.031 | 0.062 | 0.206 | 0.193 |
| OligoDyad | 0.040 | 0.154 | – | – |
| MDScan | 0.053 | 0.063 | 0.217 | 0.289 |
| AlignAce | 0.055 | 0.112 | 0.171 | 0.243 |
| MotifSampler | 0.060 | 0.107 | 0.110 | 0.339 |
| SSeiCMC | 0.061 | 0.037 | – | – |
| YMF | 0.064 | 0.137 | – | – |
| MEME | 0.067 | 0.107 | 0.224 | 0.267 |
| Improbizer | 0.069 | 0.070 | 0.225 | 0.264 |
| MEME3 | 0.078 | 0.091 | – | – |
| WEEDER | 0.086 | 0.300 | 0.173 | 0.3 |
| ANN-Spec | 0.087 | 0.088 | 0.079 | 0.346 |
| BioProspector | 0.111 | 0.215 | 0.158 | 0.304 |
| SPACE | 0.126 | 0.334 | 0.225 | 0.341 |
| SCOPE* | 0.098 | 0.149 | 0.229 | 0.423 |
| BEST* | 0.266 | 0.044 | 0.244 | 0.217 |
| RGSMiner* | ≤0.282 | ≤0.347 | ≤0.441 | ≤0.446 |
| WebMotifs* | ≤0.282 | ≤0.347 | ≤0.441 | ≤0.446 |
| EMD* | ≤0.338 | ≤0.071 | 0.262 | 0.296 |
| MotifVoter* | 0.410 | 0.486 | 0.603 | 0.617 |

The nSN and nPPV for stand-alone motif finders in Tompa’s benchmark are taken directly from Tompa et al. (2005). For E.coli dataset, we are unable to obtain the result of a few stand-alone motif finders (marked with dash), because either these programs are not available or the output does not give binding sites for us to evaluate the results (e.g YMF), during the preparation of this article. Motif finders marked with asterisks (*) are ensemble methods. Note that for some ensemble methods, we are not able to execute on the given datasets, we only estimate their upper bound (see Section 4.1 for details on how we obtain the upper bound).
Fig. 3. Scatter plot shows the SSD values of each stand-alone and ensemble motif finders in (a) yeast and (b) mammalian ChIP datasets. Every point in the scatter plot is taken from the best motif from each stand-alone and ensemble motif finders. For MotifVoter we only take top motif.

The yeast ChIP-Chip datasets and TF profiles are obtained from Harbison et al. (2004) (http://fraenkel.mit.edu/Harbison/). For mammalian datasets, we evaluate nine transcription factors: CREB (Zhang et al., 2005), E2F (Ren et al., 2002), HNF4/HNF6 (Odom et al., 2004), MYOD/MYOG (Cao et al., 2006), NFKB (Schrieber et al., 2006), NOTCH (Palomero et al., 2006), SOX (Boyer et al., 2005). Further details about the yeast and mammalian datasets sources can be found in Supplementary Material 1.

Out of 65 cases, MotifVoter can identify 56 motifs. As for the missing ones, we found that the individual finders do not perform well. Figure 3 shows the evaluation using SSD values of the stand-alone and ensemble motif finders based on the 56 successful cases. Note that SSD values can give an estimation of how close the prediction of one motif finder is to the real PWM in each dataset. Table 2 shows the average SSD performance of each motif finder.

On yeast ChIP-Chip datasets, among all stand-alone and ensemble motif finders, Improbizer has the highest average SSD value (1.639). MotifVoter outperforms Improbizer with an average SSD value of 1.919. Weblogos of the actual motifs predicted by MotifVoter on these datasets can be found in Supplementary Material 3.

Figure 4 shows the nine PWMs found by MotifVoter in agreement with canonical motifs in mammalian datasets. Evaluation on MotifVoter shows that it has an average SSD value of 1.824, while the best performing existing motif finder achieves an average SSD value of 1.530. SSD of MotifVoter on mammalian datasets is lower than that on yeast since the datasets of mammalian genome are much more complex than yeast.

| Table 2. The table lists the average SSD between the true motif and the predicted motif for each motif finders, on yeast and mammalian ChIP datasets |
|-----------------------------------|----------------|----------------|
|                                   | Yeast SSD     | Mammal SSD     |
| AlignAce                          | 1.206         | 1.530          |
| MotifSampler                      | 1.278         | 1.132          |
| MEME                              | 1.292         | 0.750          |
| Weeder                            | 1.314         | 1.360          |
| ANNSpec                           | 1.344         | 1.137          |
| MDSscan                           | 1.361         | 1.036          |
| BioProspector                     | 1.370         | 1.246          |
| MITRA                             | 1.515         | 1.470          |
| SPACE                             | 1.539         | 1.121          |
| Improbizer                        | 1.639         | 1.475          |
| BEST*                             | 1.159         | 1.323          |
| WebMotifs*                        | 1.344         | –              |
| SCOPE*                            | 1.492         | 1.247          |
| MotifVoter*                       | 1.919         | 1.824          |

Note that the SSD score is at most 2.

5 DISCUSSION

This section further investigates the effectiveness of MotifVoter in terms of the following aspects: (1) the optimality of MotifVoter; (2) the effects of the discriminative and consensus criteria used in MotifVoter and (3) the robustness of MotifVoter and the running time of MotifVoter.

5.1 Optimality of MotifVoter

In general, the performance of any ensemble method is bounded by its component motif finders. The total number of true binding sites that are covered by any of its component motif finders can be regarded as a rough upper bound on the performance of an ensemble method. We say an ensemble motif finder is \( \epsilon \)-optimal if it can find \( \epsilon \) fraction of the true sites covered by at least one of its component motif finders. Evaluation on Tompa’s benchmark shows that the highest possible sensitivity that can be achieved by all the component motif finders is 0.440. The sensitivity of MotifVoter is 0.419 which
The discriminative criterion of MotifVoter makes use of the study shows that MotifVoter is a near optimal ensemble method. This empirical sensitivity is 0.643 and MotifVoter achieves 95.7%. This empirical improvement was made on yeast datasets (84%).

MotifVoter relies on its component motif finders. Thus, a natural question is whether the performance of MotifVoter will degrade if we include some motif finders with poor performance. The performance of MotifVoter does degrade as more random motif finders (representing motif finders with poor performance) are included. However, even if we include five random motif finders (that is half of the real motif finders we used), the performance of MotifVoter is still greater than that of the best individual motif finder by 183% in sensitivity and by 10.4% in precision. These results imply that MotifVoter is robust even if some of the component motif finders perform badly. We remark that it is possible that MotifVoter may predict wrong motifs if almost every individual finder reports similar wrong motifs. However, MotifVoter can reduce the chance of finding such wrong motifs significantly in comparison to individual finders because the chance of having many individual finders to report the same (similar) spurious motif is low since every finder has its own scheme of predicting the motifs.

Running all 10 component motif finders for MotifVoter is not always practical. We show that the sensitivity and precision of MotifVoter are still significantly better than the best motif finder when we run only the five fastest motif finders. The details of the above robustness evaluations are shown in Supplementary Material 5. In Supplementary Material 8, we evaluate and discuss the characteristics of binding sites missed by MotifVoter. We also show that MotifVoter can identify motifs in motif modules that contain multiple motifs work in groups, provided that these motifs have signals strong enough to be detected by the individual motif finders used in MotifVoter. Details can be found in Supplementary Material 9.

5.3 Robustness and time complexity MotifVoter

We also evaluate the performance of MotifVoter on various background sequences and various species in Tompa’s benchmark. MotifVoter achieved the highest nSN and nPPV on all three backgrounds. The major sensitivity improvement is on real datasets (275%), followed by generic datasets (128%). As for evaluation on three species, MotifVoter made major sensitivity improvement on human dataset (314%) followed by fruitfly (263%), while the least improvement was made on yeast datasets (84%).

MotifVoter relies on its component motif finders. Thus, a natural question is whether the performance of MotifVoter will degrade if we include some motif finders with poor performance. The performance of MotifVoter does degrade as more random motif finders (representing motif finders with poor performance) are included. However, even if we include five random motif finders (that is half of the real motif finders we used), the performance of MotifVoter is still greater than that of the best individual motif finder by 183% in sensitivity and by 10.4% in precision. These results imply that MotifVoter is robust even if some of the component motif finders perform badly. We remark that it is possible that MotifVoter may predict wrong motifs if almost every individual finder reports similar wrong motifs. However, MotifVoter can reduce the chance of finding such wrong motifs significantly in comparison to individual finders because the chance of having many individual finders to report the same (similar) spurious motif is low since every finder has its own scheme of predicting the motifs.

Running all 10 component motif finders for MotifVoter is not always practical. We show that the sensitivity and precision of MotifVoter are still significantly better than the best motif finder when we run only the five fastest motif finders. The details of the above robustness evaluations are shown in Supplementary Material 5. In Supplementary Material 8, we evaluate and discuss the characteristics of binding sites missed by MotifVoter. We also show that MotifVoter can identify motifs in motif modules that contain multiple motifs work in groups, provided that these motifs have signals strong enough to be detected by the individual motif finders used in MotifVoter. Details can be found in Supplementary Material 9.

6 CONCLUSIONS

We have presented MotifVoter, a simple yet effective ensemble method that utilizes both positive and negative information in the motifs returned by the basic motif finders and achieves near-optimal sensitivity. In extensive evaluations on many ChIP-Chip and benchmark datasets, we observed that MotifVoter outperforms all stand-alone and existing ensemble motif finders significantly. It is also robust with respect to the inclusion of random motif finders and number of component motif finders.

A key advantage of MotifVoter is its flexibility in incorporating new algorithms. That is, if a novel superior motif finding algorithm is made available, it can be readily incorporated into the MotifVoter platform. In the website version, we provide an option for users to provide results from the component motif finders that are not included in our main list. Furthermore, MotifVoter, as a web application, has several biologist friendly features such as a parameter-free interface, ability for users to choose their preferred...
component motif finders and option for speedier return of results using the fastest motif finders.

ACKNOWLEDGEMENTS

We are grateful to the authors of the component motif finders used by MotifVoter for making their tools available for public use. We thank C. Dongsheng for providing command line version of BEST. Finally, the authors would like to thank the reviewers for all their useful and constructive comments.

Funding: National University of Singapore (grant R-252-000-326-112); Research Output Prize (Faculty of Engineering) of the University of Hong Kong to S.M.Y.

Conflict of Interest: none declared.

REFERENCES


