Discovery of transcription factor binding sites through integration of generic motif finders

Edward Wijaya[1,2], Siu-Ming Yiu[3], Ngo Thanh Son[1], Kanagasabai Rajaraman[2], and Wing-Kin Sung[1,4]

Locating transcription factor binding sites is a key step in understanding gene regulation. Due to its importance, many *de novo* motif finding methods (e.g. MEME, MotifSampler, Mitra and Weeder) have been proposed. Individually, these motif finders perform unimpressively overall based on Tompa's benchmark datasets. Moreover, these motif finders vary in their definitions of what constitute a motif, and in their methods for finding statistically overrepresented motifs. This makes different motif finders perform well for identifying binding sites of certain types of datasets only. There is no clear way for biologists to choose the motif finder that is most suitable for their task. The purpose of this work is to describe a method called MotifVoter to identify transcription factor binding sites by integrating the results found by motif finders of different models. Validation of our method on Tompa's benchmark, real *metazoan* and *E. Coli* datasets (186 datasets in total) show that it can improve the sensitivity significantly without sacrificing the precision. MotifVoter can locate almost all the binding sites found by the individual motif finders used and is able to distinguish the real binding sites from noise effectively. Our approach offers a practical alternative for biologists to study novel transcription factors.

Availability: The software is available for public use at:
`http://www.comp.nus.edu.sg/~bioinfo/MotifVoter`

Detection of transcription factor binding sites is one of the major challenges faced by biologists since it is the critical step to understand the regulatory mechanism of genes. Hence, the problem of *de novo* identification of transcription factor binding motifs/sites has been widely studied. Many motif finders have been proposed under different categories such as profile-based methods e.g. Gibbs sampler[1], MotifSampler[2], SeSiMCMC[3], GAME[4], Improbizer[5] and consensus-based methods e.g. Weeder[6], MITRA[7], and SPACE[8].

Though a lot of tools have been developed, little knowledge is known on which motif finder should be used for a particular dataset. This turns out to be a difficult issue for practitioners and biological researchers because of three reasons. Firstly, the performance of individual motif finders is unimpressive overall. For example, in the study of Tompa[9], a well known assessment study of 13 popular motif discovery algorithms over 56 datasets drawn from *H. Sapiens*, *M. Musculus*, *D. melanogaster*, and *S. Cerevisiae* genomes, it is found that even the best motif finder performs very badly. The sensitivity and the precision are ≤ 0.13 and ≤ 0.35, respectively. Secondly, the performance of individual motif finders has been found to vary depending on the input datasets. In the study of the Tompa's benchmark dataset, we cannot find any motif finder which is consistently good for all datasets. This implies that different motif finders are suitable for different datasets. Thirdly, even if we can fix the motif finder, it is not straightforward to decide how many motifs in the output list we should consider. Motifs of lower rank may be useful to reveal real binding sites.

The above discussion implies that there is no single universal motif finder which can predict correctly all types of motifs in the existing motif finders. So it is natural to ask if we can consult

[1] School of Computing, National University of Singapore, Singapore 119260.
[2] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613.
[3] Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong.
[4] Genome Institute of Singapore, 60 Biopolis Street, #02-01 Genome, Singapore 138672.
Correspondence should be addressed to Wing-Kin Sung (ksung@comp.nus.edu.sg)

diverse motif finders to identify the correct motif and the corresponding binding sites. A few methods (e.g. SCOPE[10] and EMD[11]) have showed that this direction is promising. Our experiments also reveal that the combined output from different finders indeed contains a lot more real binding sites than that from individual finders. However, the improvement gained by these methods is not substantial. While the combined output from different finders increases the chance of identifying real binding sites, the amount of noise generated also increases tremendously. The main difficulty lies in how to differentiate the real binding sites from the false positives. This paper would like to answer this question affirmatively by showing how to effectively integrating the output of different motif finders. We hope that this can help the biologists and the practitioners to find the correct motif and the corresponding binding sites as automatically as possible without worrying about which motif finder to use.

We propose a novel method called MotifVoter to integrate the results from multiple motif finders. The principle behind MotifVoter is to remove the noise from the real binding sites based on the consensus of the motif finders in two stages. First, we remove those incorrect candidate motifs according to a variance based statistical measure[12,13]. Second, from the remaining candidate motifs, we filter away the noisy binding sites and retain the real binding sites. Unlike the previous ensemble methods that simply select the best motif from the motif list reported by multiple finders, our solution carefully selects a set of good candidate motifs and goes deeper by selecting the best binding sites of these motifs. **Figure 1** depicts the stages used by MotifVoter.

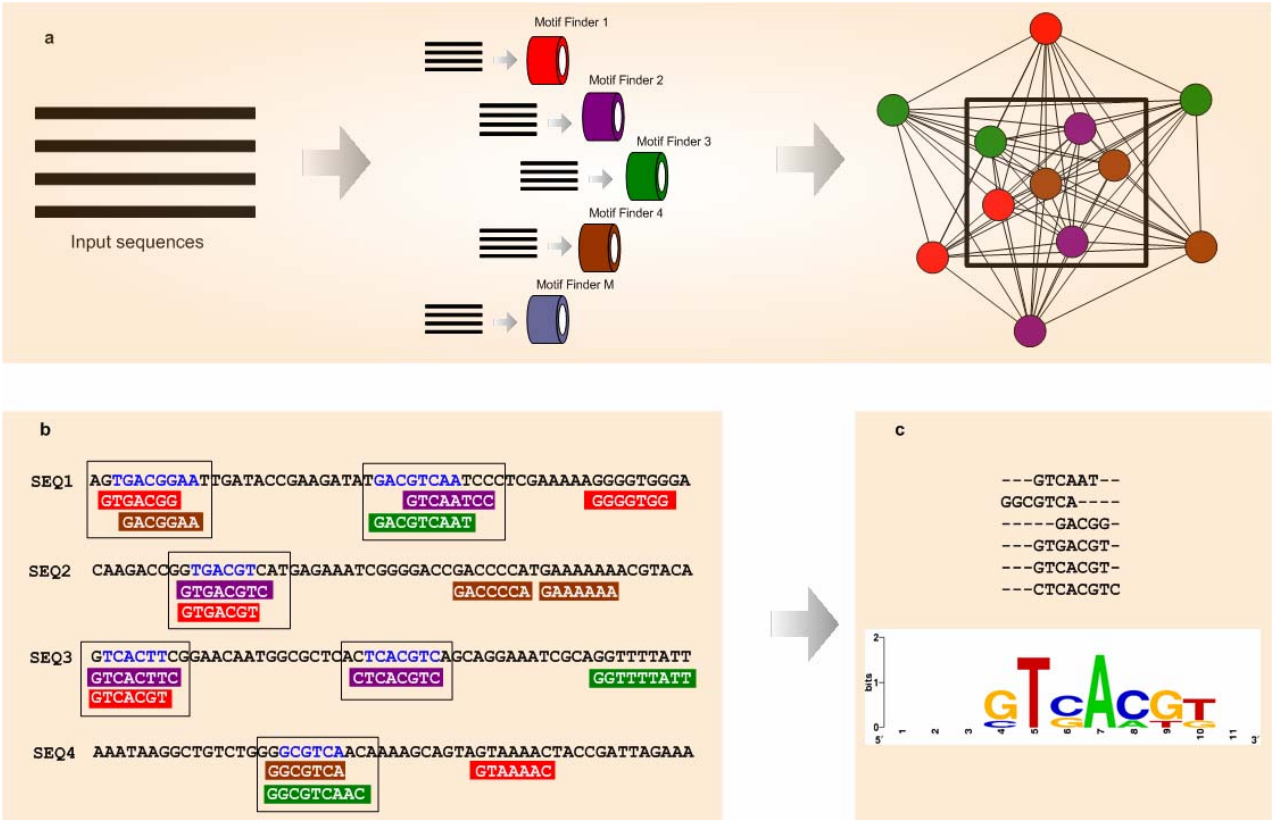

Figure 1 (a) We apply *M* motif finders on the input sequences. From each motif finder we obtain n_i motifs with their respective instances. We compute the pairwise similarity measure between all these motifs. Then, we try to obtain a set of motifs that are highly similar to one another as illustrated in the graph. The similarity between two motifs is represented by a weighted edge (in the diagram, the nodes represent motifs and we use a shorter edge to represent a pair of motifs that are more similar) and the aim is to find a cluster such that (1) motifs predicted by multiple motif finders are involved and (2) the motifs are close (similar) to one another, but far away (dissimilar) from other. We employ a variance-based statistical

approach to achieve this effect. (b) The diagram shows the ideas behind the second stage of MotifVoter. Binding sites in blue are real binding sites. The remaining colors are used to illustrate the binding sites predicted by 4 other motif finders. If we consider the motif finders individually, at most 4 out of 6 true binding sites are found. However, if we consider the overlapping binding sites from different motif finders, more correct binding sites will be discovered. Moreover, the chance of a single motif finder returning false binding sites is higher than that of MotifVoter. For example, the true binding site TCACGTC can still be identified, because among all motif finders, the binding sites of the purple motif finder are highly overlapping with the binding sites found by other motif finders. Thus, we give a higher confidence to it. (c) We compute multiple sequence alignment of these instances using MUSCLE[14], and from the alignment we arrive at the final motif using a PWM model.

We have evaluated MotifVoter and compared it with other 17 motif finders and two most recent ensemble methods. The results show that MotifVoter significantly outperforms all of them in term of both sensitivity and precision. For example, on Tompa's benchmark datasets, MotifVoter improves the sensitivity by 215% and the precision by 45.5%. More importantly, MotifVoter can locate almost all binding sites that are found by its basic motif finders. It can distinguish the real binding sites from the false positives in the aggregation of outputs from the multiple motif finders. We also show that MotifVoter works well across different species and different types of background sequences. In particular, MotifVoter gives the biggest improvement in real background sequences (see description on Tompa's benchmark dataset in the next section) and higher organisms (*H. Sapiens* and *M. Musculus*). Finally, we show that as long as some good motif finders are included in MotifVoter, then even if there are a few motif finders with poor performance, the performance of MotifVoter is still substantially better.

In practice we might not always be able to run a lot of motif finders. Hence, we have studied the performance of MotifVoter by only including the fastest N (N = 3, 4, 5) motif finders. The results show that MotifVoter is stable. Even if we include only 3 finders in the list, the performance, though degrades a little bit, is much better than individual motif finders.

RESULTS

Including motifs of lower rank does not improve the sensitivity of individual motif finders significantly

Tompa et al.'s study[9] assessed rank 1 motifs predict by various motif finders. However, this assessment did not address whether using motifs of lower rank will improve the overall performance of individual motif finders. This section shows that even by including motifs of higher ranks, the performance of individual motif finders cannot be improved substantially. **Figure 2a** shows the sensitivity of the predicted binding sites by the top-*n* motifs of each motif finder. The best individual motif finder has sensitivity 0.130 if we just consider the predicted motifs of rank 1. When we consider the sites predicted by top-30 motifs of the best individual motif finder, the sensitivity is improved to 0.175. This suggests that, even if we consider motifs of rank 2 or above, the sensitivity of individual motif finder is improved by at most 25%. Moreover, the precision decreases significantly since a lot of noise exists in the motif list of rank 2 or above (**Figure 2b**).

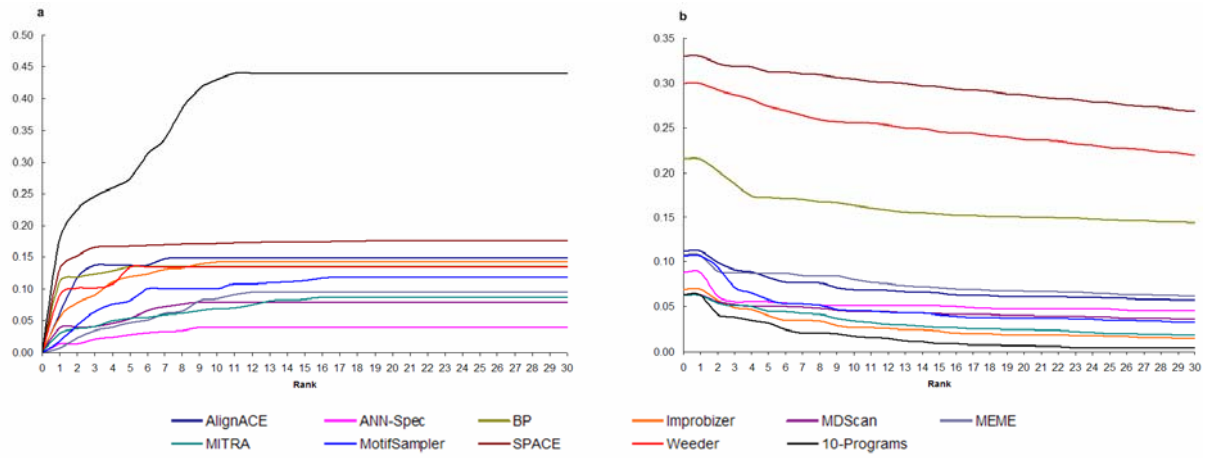

Figure 2. This figure shows the performance of 10 individual motif finders (color curves) and the combined result of all 10 motif finders (black curve). Figures (a) and (b) show the cumulative sensitivity (*nSN*) and precision (*nPPV*), respectively, of these 11 motif finders when we include more motifs with lower rank. The figure shows that the combined result of all 10 motif finders has a much higher sensitivity than any individual motif finder. However, it also reduces the precision a lot.

The black curve in **Figure 2a** shows the sensitivity of the predicted sites by all 10 motif finders. If we just consider the predicted rank 1 motifs of the 10 motif finders, the sensitivity is 0.177. The sensitivity is improved to 0.439 when we consider the top-30 motifs of all 10 motif finders. This suggests an improvement of 148% in sensitivity. This observation suggests that rank 2 or above binding sites predicted by all 10 motif finders are useful.

Though rank 2 or above motifs predicted by various motif finders may help to improve sensitivity, majority of them may be noise. For instance, in Tompa's dataset, among all sites predicted by the rank 2-30 motifs of the 10 motif finders, only 0.47% of them are real binding sites. On the other hand, 6.27% of the sites predicted by the rank 1 motifs of the 10 motif finders are real binding sites (see **Figure 2b**). Hence, there is more noise in rank 2 or above motifs. This suggests that inclusion of motifs from lower rank can only be effective if we consider ensemble methods.

Different motif finders discover different binding sites

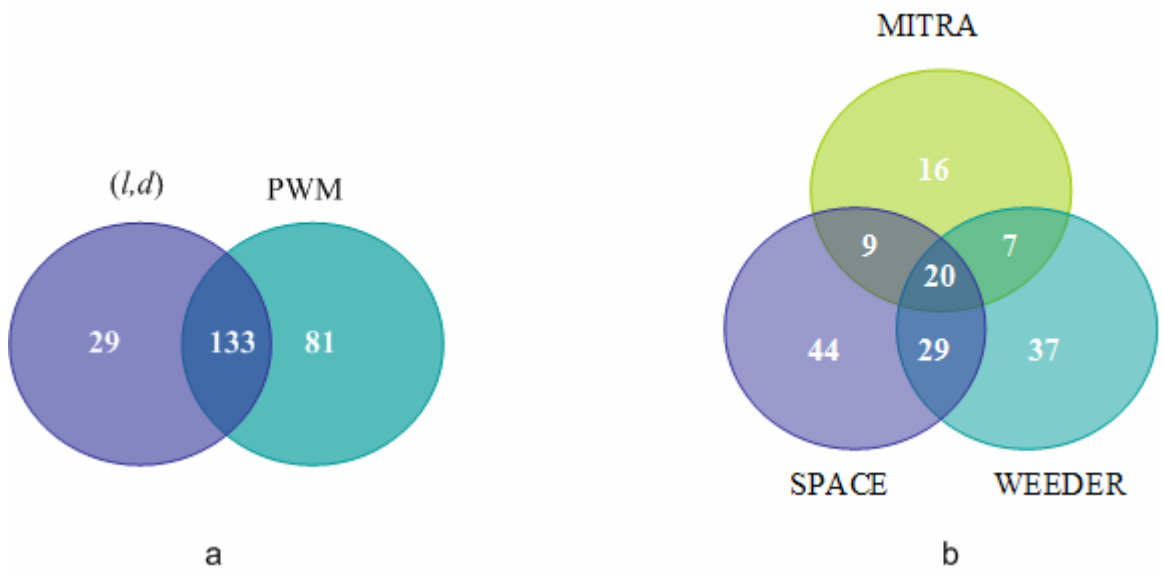

Figure 3 Our study has 3 motif finders based on (*l,d*) model and 7 motif finders based on PWM model. Using their top-30 motifs, the 10 motif finders can discover 243 binding sites in Tompa's benchmark dataset. (a) shows the numbers of sites that can be found by (i) both groups, (ii) (*l,d*) model group only, and (iii) PWM model group only. (b) focuses on the three (*l,d*) motif finders and shows the number of sites that can be found by various combination of 3 (*l,d*) motif finders.

In general, motif finders can be divided into two major types, namely PWM model (profile based) and (*l,d*) model (consensus based). There is no general agreement on which model is better. **Figure 3a** gives a comparison of the binding sites predicted by the two types of motif finders. We divide the motif finders into two groups depending on the model they are based on. The first group consists of 3 motif finders based on (*l,d*) model, which are MITRA[7], Weeder[6], and SPACE[8]. The second group consists of 7 motif finders based on PWM model, which include AlignACE[15], ANN-Spec[16], BioProspector[17], Improbizer[5], MDScan[18], MEME[19], and MotifSampler[2]. It shows the number of sites correctly predicted by (i) both groups, (ii) (*l,d*) model group only, and (iii) PWM model group only. The figure showed that 45.3% of the correctly predicted sites are predicted by either (*l,d*) model or PWM model. This implies that (*l,d*) model and PWM model may be suitable for discovering motifs for different types of datasets.

Even for motif finders of the same type, the individual motif finders may be based on different heuristics and use a different set of parameters, and so may be suitable for discovering motifs from different types of datasets. For instance, consider the three motif finders SPACE, Weeder, and MITRA which are based on (*l,d*) model. **Figure 3b** shows the correctly predicted sites by them. We observe that, even by using the same (*l,d*) model, different motif finders are suitable for finding different types of motifs. And it also provides evidence that combining results from motif finders of the same model may still provide a better motif.

MotifVoter - a method that utilizes the sites predicted by multiple motif finders

Combining results from multiple motif finders and considering motifs of lower rank, e.g. top 30, will obviously include more binding sites, but it will also include more false positives. To develop a robust ensemble method, we need an effective way to distinguish real binding sites from noise based on the outputs from the various motif finders.

Most existing methods (e.g. SCOPE[10] and EMD[11]) are based on integration at the motif level rather at the binding site level. The issue of how to distinguish a real binding site from false binding sites is not adequately addressed in the previous ensemble methods. A naïve approach is to report the binding sites that are covered by more than 2 motifs. However, our experiments show that the improvement is only limited. (For instance, though this naïve approach improves sensitivity (*nSN*) by 68% over the current best motif finder (SPACE), this method looses in precision (*nPPV*) as much as 17.3% over SPACE in Tompa's benchmark dataset). More importantly, it is not trivial to define whether a binding site reported by multiple finders is real or noise.

We developed a novel ensemble method MotifVoter, which integrates the results of 10 motif finders that performed reasonably well on Tompa's benchmark and were easily obtainable from public domain: AlignACE[14], ANN-Spec[15], BioProspector[16], Improbizer[5], MDScan[17], MEME[18], MITRA[7], MotifSampler[2], SPACE[8], and Weeder[6]. It may be noted they are also some of the widely used motif finders in the community of biologists. **Supplementary 7** describes the characteristics and parameters used in each of these motif finders. In the evaluation, we have used three datasets (Tompa's benchmark dataset, the *metazoan* dataset, and the *E. coli* dataset).

The performance of MotifVoter versus individual motif finders

We compare the performance of MotifVoter with 17 individual motif finders on Tompa's benchmark datasets. Tompa's benchmark has been constructed based on real transcription factor binding sites drawn from four different organisms *yeast*, *fruitfly*, *human* and *mouse*[9]. It consists of 56 datasets in total. Each dataset consists of 1-35 sequences and each sequence is of length up to 3000 bp. The datasets are constructed from three different types of background sequences. They are (i) *real* promoter sequences, (ii) randomly chosen promoter sequences from the same genome (called *generic*), and (iii) sequences generated by Markov chain of order 3 (called *markov*).

Figure 3a shows the results. MotifVoter improves the sensitivity (*nSN*) by 215% (from 0.13 to 0.41) when compared with the best performing stand-alone motif finder while the precision (*nPPV*) is improved by 45.5%. (Description about the statistics used can be found at the end of Method section.)

More importantly, MotifVoter can locate almost all binding sites that are found by any existing finders (see **Figure 3b)**. As MotifVoter uses 10 basic motif finders as its components, if the basic motif finders cannot find a particular real binding site, MotifVoter cannot find it too. Thus the highest possible sensitivity that can be achieved by MotifVoter (or any ensemble method) is the fraction of real binding sites that can be found by at least one basic motif finder. Evaluation in Tompa's benchmark datasets shows that the highest possible sensitivity that can be achieved is 0.44. MotifVoter, on the other hand, can achieve a sensitivity of 0.419.

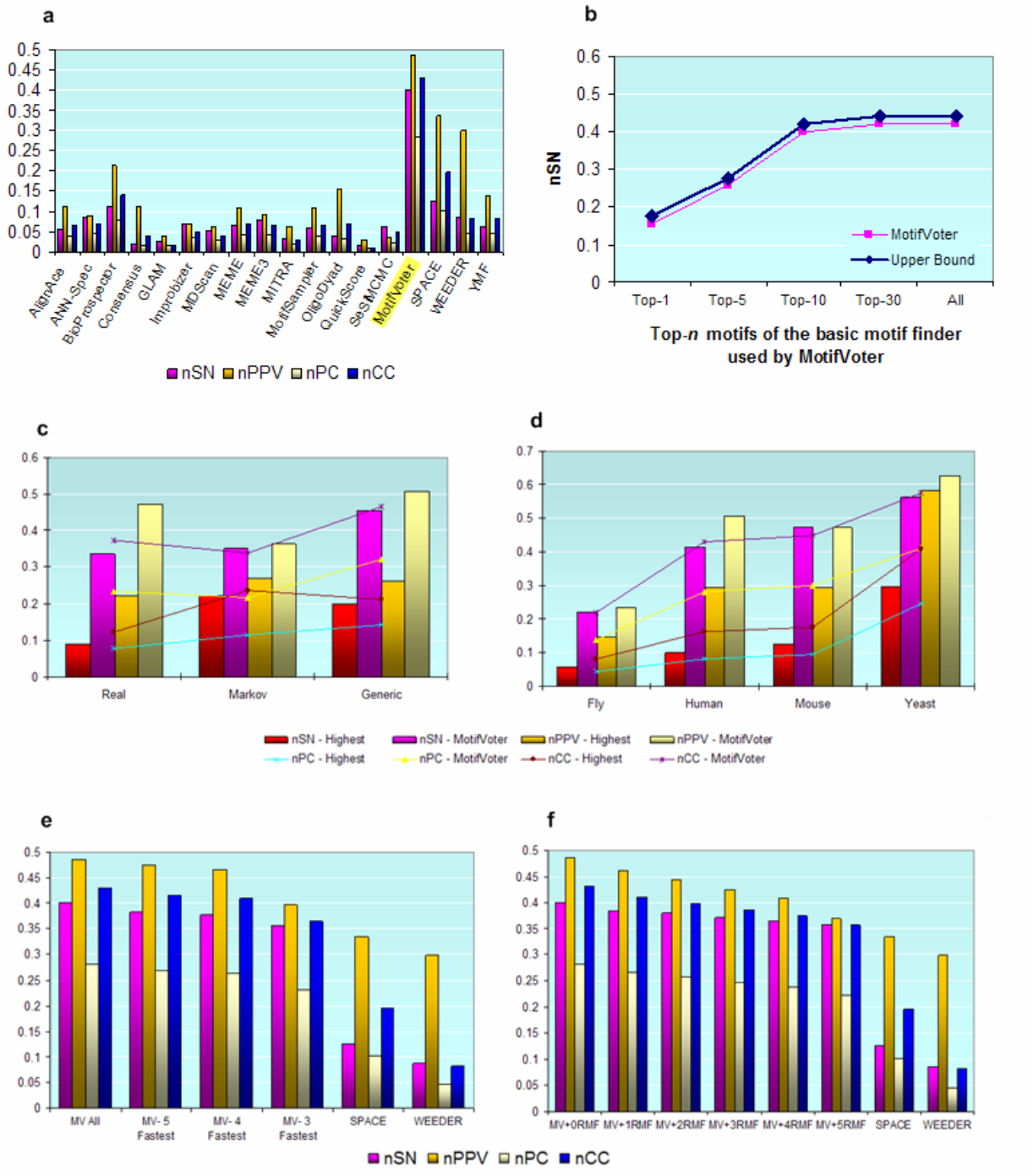

Figure 3 Experimental results on Tompa's benchmark dataset. (a) Comparison of MotifVoter and individual motif finders. (b) The sensitivity of MotifVoter versus the maximum possible sensitivity (using 10 selected motif finders). The blue curve shows the fraction of nucleotides that are found by at least 1 motif finder. The pink curve shows the corresponding nucleotide sensitivity of MotifVoter. Note that the *x*-axis refers to the top-N number of motifs we use from each *basic* motif finder in MotifVoter. For example, top-10 means we use the top 10 motifs from each finder. It is not the number of motifs returned by MotifVoter per se. MotifVoter only returns rank-1 result. (c) Performance of MotifVoter on various types of background sequences when compared with the best individual motif finder. (d) The performance of MotifVoter on various species when compared with the best individual motif finder. (e) The performance of MotifVoter when we use 10 motif finders together with 1-5 random motif finders (as described in the Robustness of MotifVoter section). (f) shows the performance of MotifVoter based on all 10 motif finders (MV), the fastest 5 motif finders (MV-5), the fastest 4 motif finders (MV-4), and the fastest 3 motif finders (MV-3). The fastest 5 motif finders we considered are BP, MDScan, Weeder, ANN-Spec, and Improbizer. (Note that the total running time of these 5 motif finders is faster than MEME.) The detail of execution time is shown in **Supplementary 1**.

Performance of MotifVoter on different background sequences and species

This section discusses the performance of MotifVoter on different species and background sequences. **Figure 3c** shows the performance of MotifVoter on various background sequences in Tompa's benchmark datasets. In this evalution, the major improvement is on *real* datasets (275%), followed by *generic* dataset (128%). Since modeling the background sequences of *real* type is more difficult, individual motif finders usually perform worse in *real* datasets when compared with *markov* and *generic* datasets. On the other hand, MotifVoter combines both PWM and (*l*,*d*) models from different motif finders, and hence it is able to recover more binding sites in *real* datasets.

We obtain similar results in the evaluation based on species also (**Figure 3d**). MotifVoter achieves the highest *nSN* and *nPPV* in datasets on all four species namely human, mouse, fruitfly and *yeast*. But MotifVoter made major improvement on human dataset (314%) followed by fruitfly (263%) while the least improvement is made on yeast dataset (84%). One possible explanation is that the binding sites in *human*, *mouse*, and *fruitfly* are much less conserved than *yeast*. By making use of various modeling capability of different basic motif finders, MotifVoter has a higher chance of capturing more diversed binding sites model on *human*, *mouse*, and *fruitfly*.

Time complexity of MotifVoter

The time complexity is an important issue for MotifVoter. Running all 10 motif finders for MotifVoter is not always practical. We investigated whether MotifVoter can improve the sensitivity and precision compared to the best individual motif finder, if we only execute the fastest N (N=3, 4, 5) motif finders in MotifVoter. (Note that the total running time to execute the fastest 5 motif finders is still smaller than the running time of MEME. **Supplementary 1** shows the detailed running time of the 10 motif finders). **Figure 3e** shows the performance of MotifVoter if we only run the fastest N finders (where N = 3, 4, 5). The results show that the performance of MotifVoter is still significantly better than the best motif finder in terms of sensitivity and precision.

Robustness of MotifVoter

MotifVoter relies on individual motif finders. So, a natural question is whether the performance of MotifVoter will degrade a lot if we include some motif finders that do not perform very well. To study this aspect, we included 1-5 motif finders that predict motifs randomly (to represent motif finders with poor performance) in addition to the 10 motif finders. Each random motif finder picks a random length-l string in the input sequences as a motif. The corresponding motif instances are generated using the (l,d) motif model (that is, length-l substring with at most d mutations from the motif), where the parameter used for (l, d) are: (8,1), (10,2), (10,3), (15,2), (15,3).

Figure 3f shows the evaluation results on this experiment. The performance of MotifVoter does degrade as more random motif finders (representing motif finders with poor performance) are included. However, even if we include 5 random motif finders (that is half of the real motif finders we used), the sensitivity (0.357) of MotifVoter is still significantly greater than that of the best individual motif finder (0.126). A similar observation is obtained for precision. In other words, MotifVoter is quite robust even if some of the component motif finders perform badly.

Performance Validation on Metazoan Datasets

We also examine the performance of MotifVoter on the *metazoan* datasets that have been drawn from real genomic sequences. The *metazoan* datasets are taken from ABS database[20] (`http://genome.imim.es/datasets/abs2005/index.html`), and consist of 68 datasets. The number of sequences ranges from 3-39 and the sequence lengths are up to 500 bp. The binding sites are gathered from the literature where they have been experimentally verified. The sites and the promoter sequences have been manually curated to ensure data consistency. They come from three different organisms: *human, rat* and *mouse.*

When we repeated the same experiments on *metazoan* datasets, we observed similar results. MotifVoter outperforms the best motif finder in this dataset by 103% and 35% in *nSN* and *nPPV* respectively (**Figure 4a**). We also validate the performance of MotifVoter on individual species of the metazoan dataset. MotifVoter also performs better in each case (**Figure 4b**). The highest possible sensitivity for this dataset is 0.650, and the sensitivity of MotifVoter is 0.632 which is again close to the upper bound. Please refer to **Supplementary 2** for the detailed evaluation of MotifVoter on the upper bound analysis. **Figure 5** shows several example binding sites from *metazoan* datasets. It illustrates that MotifVoter finds more binding sites than stand-alone motif finders. Also, in general the predicted motif models are similar to the actual motifs.

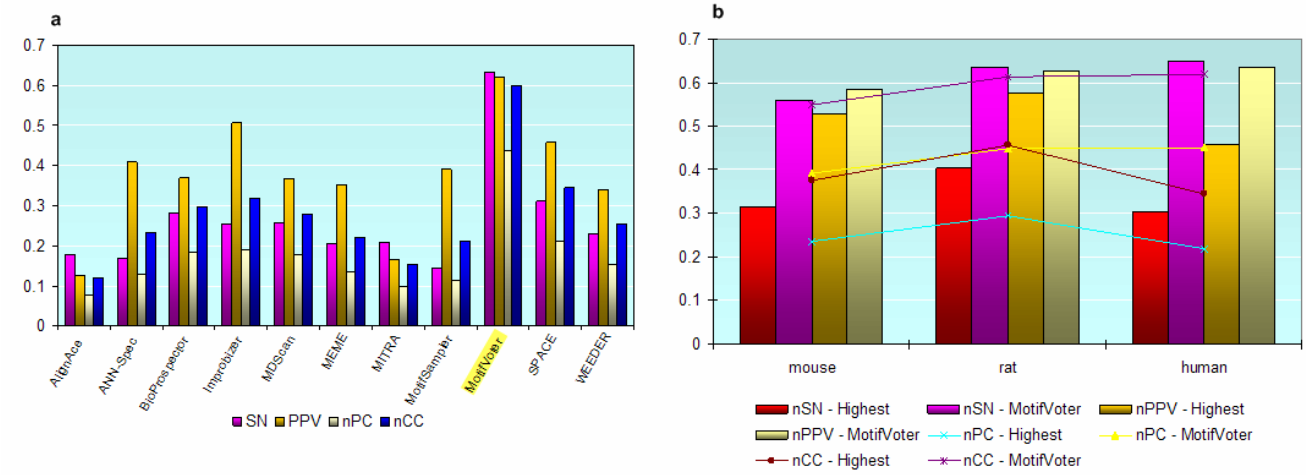

Figure 4 Results on *metazoan* datasets (a) Comparison of MotifVoter and individual motif finders. (b) Performance of MotifVoter on various species compared to the best performing individual motif finders.

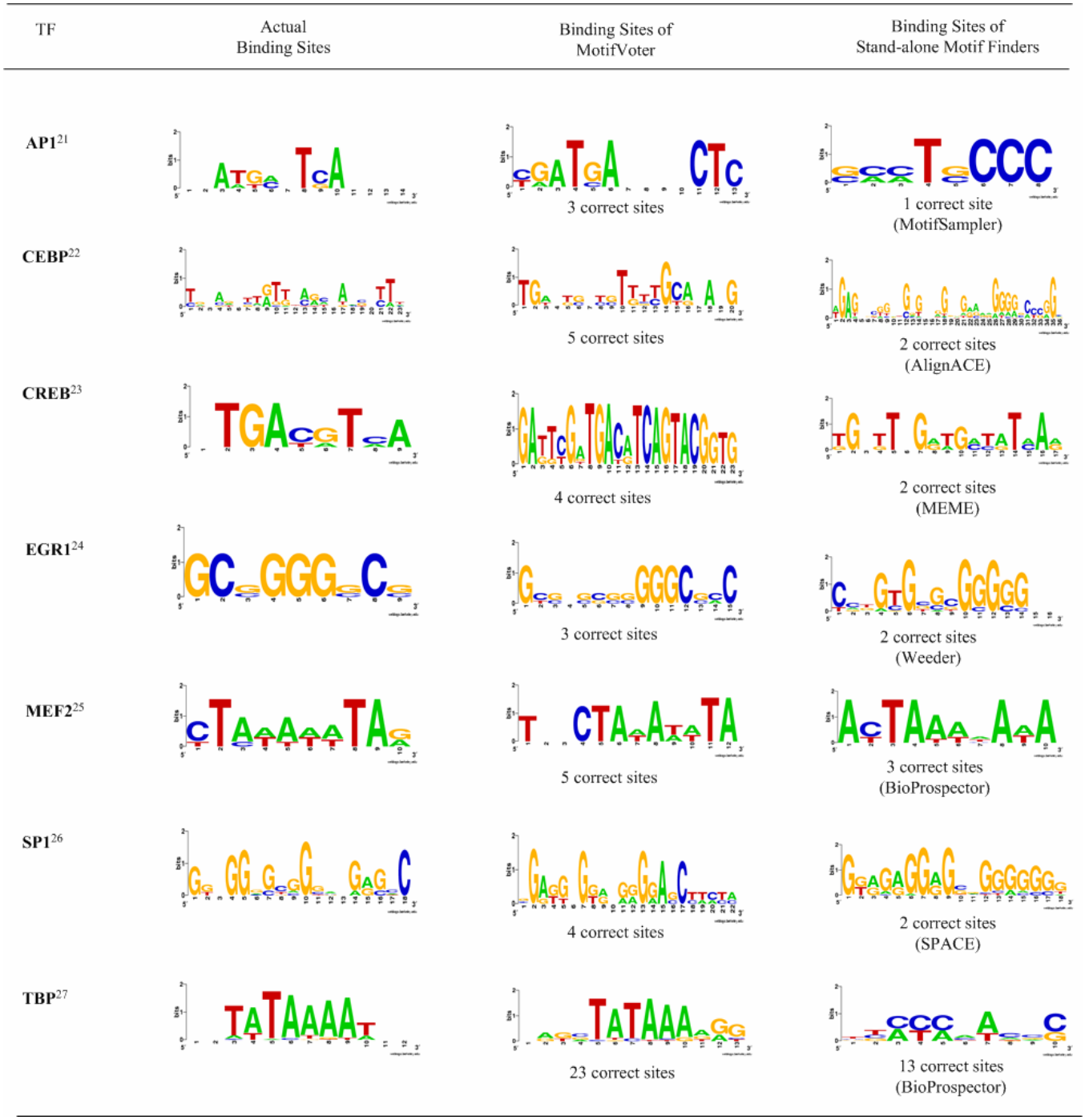

Figure 5 Examples of the binding sites found by MotifVoter and stand-alone motif finders on real *metazoan* datasets. For each of these datasets, we report the result from the best performing stand-alone motif finder.

Comparison of MotifVoter with other ensemble methods

We also compare MotifVoter with the two most recent ensemble methods SCOPE and EMD. We perform experiments on *E.Coli* datasets taken from RegulonDB[28] (`http://regulondb.ccg.unam.mx/`). They are generated from the intergenic regions of *E.Coli* genome. In total they contain 62 datasets. The average number of sequences is 12 and the average sequence length is 300bp. We are unable to perform the evaluation on Tompa's benchmark and the metazoan datasets since EMD is not available for public use. Hence we make the comparison using *E.Coli* datasets alone, the results for which are obtained from EMD's publication.

SCOPE is a motif finder which integrates the motifs predicted by BEAM, PRISM and SPACER while EMD is a motif finder which uses the motifs predicted by AlignACE, BioProsPector, and MDScan. To make a fair comparison, we run a version of the MotifVoter that uses the same three motif finders used by EMD[‡]. **Figure 6** shows the evaluation results.

In this dataset, SCOPE is better than EMD in terms of *nPPV* but has a slightly lower *nSN*. We believe that this is because SCOPE only reports instances from 1 motif, unlike EMD which also considers instances from other motifs of the same rank. Nevertheless, even with 3 motif finders, MotifVoter can improve the *nSN* to 0.448 and *nPPV* to 0.509. For further analysis of SCOPE and EMD, please refer to the discussion section. In **Figure 6**, we also include the performance of the best two individual motif finders (SPACE and Weeder) for reference. Please refer to **Supplementary 3** for the detailed evaluation of MotifVoter and other stand-alone motif finders.

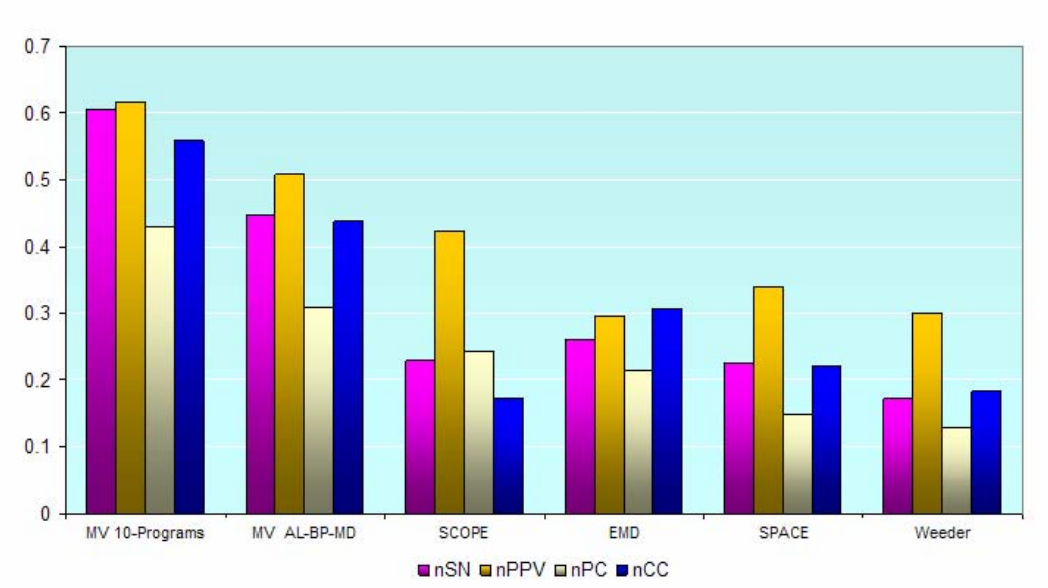

Figure 6. Comparison of MotifVoter with SCOPE and EMD. MotifVoter performs consistently better in both *nSN* and *nPPV*. We also include the performance of the best two individual motif finders (SPACE and Weeder) for reference. It shows that both SCOPE and EMD improve the performance. However, the improvement is not as significant as MotifVoter. In particular, SCOPE is better than SPACE in terms of *nPPV* only. EMD, on the other hand, can only improve the *nSN* of SPACE and Weeder marginally.

[‡]We cannot create a MotifVoter which uses BEAM, PRISM, and SPACER since these three stand-alone motif finders are not available. Also, note that the motif finder SPACER used by SCOPE is different from SPACE used by MotifVoter.

DISCUSSIONS

Integration principle used by MotifVoter

The integration principle used by MotifVoter consists of two stages. The first stage tries to identify the set of candidate motifs that are similar to one another and filter out those that are dissimilar, in the list of candidate motifs predicted by each motif finder (**Figure 1a**). It is based on the belief that the incorrect motifs from different motif finders are less likely to be similar to one another. However, after identifying the set of motifs, they may still contain some false binding sites. The second stage tries to distinguish the false binding sites from the real binding sites (**Figure 1b**). A real binding site is likely to be covered by more than one motif predicted by different motif finders. Also some motif finders are more sensitive to a particular dataset, thus their predicted motifs are more accurate. For this reason, we develop a confidence measure for the motifs based on the amount of real binding sites they can cover and accept all binding sites reported by the highest confidence motif, even if some of its binding sites are covered by a single motif only.

Limitations of existing ensemble methods

In the literature, there are two existing directions for performing ensembles. However, both have their own limitations. The first approach is simply done by grouping all the motifs output by the respective motif finders (excluding the binding sites) and then re-ranking them using a new scoring function. This approach is taken by SCOPE[10]. The advantage of this method is that it can select the best motif out of all the motif finders. However, this approach only selects correct binding sites of one motif predicted by one individual motif finder. It will fail to discover correct binding sites found by more than one motif finders.

The second approach makes use of a clustering technique. EMD[11] follows this direction. It begins by considering motifs of *m* motif finders. The motifs of the same rank reported by *m* motif finders are clustered together. Using a voting scheme it selects the final binding sites in each cluster and finally reports the cluster of the highest ranking. The benefit of this approach is that it can find more binding sites from multiple motif finders. However, it misses the true binding sites that come from motifs of different ranking since true binding sites most likely come from different motifs of different rank.

Observations on the binding sites missed by MotifVoter

In Tompa's benchmark dataset, out of 56 datasets there are 22 datasets in which less than 50% binding sites can be found. Having analyzed those 22 datasets, we suspect that most of these binding sites are highly unconserved. Precisely, out of 22 datasets, 15 datasets are unconserved (70%). For the remaining 7 datasets out of 22 datasets (30%), the density of binding sites (that is the ratio of total length of binding sites over the total size of dataset) is relatively low. Under the low signal to noise ratio, it is harder to discover the binding sites. Please refer to the **Supplementary 4** for the actual examples that illustrate the characteristics of these datasets.

Conclusion and future direction

This paper argues that all current motif models can only approximate the correct motif. To maximize the sensitivity, we should integrate the outputs discovered by multiple motif finders. We proposed Motifvoter, which can effectively retain almost all the correct binding sites discovered by the given individual motif finders while removing significant amount of false binding sites. It also works well across different species and different types of background sequences. We hope MotifVoter can offer a practical alternative for biologist to study novel transcription factors.

Despite of its effectiveness, MotifVoter is still unable to fully model the true binding sites. Since the underlying biology of regulatory mechanism is very incompletely misunderstood, exploitation of additional information such as microarray data[29] or phylogenetic footprinting[30] may help us to recover more binding sites which cannot be found with *de novo* method.

METHODS

Pairwise similarity between motifs

We measure the similarity of two motifs *x* and *y* based on their instances. Let *I(x)* be the set of instances (or the regions covered by the instances) of *x*. Let $I(x) \cap I(y)$ be the set of regions covered by at least one instance in *x* and one instance in *y*. Let $I(x) \cup I(y)$ be the set of regions covered by any instance of *x* or *y*. We denote the total number of nucleotides of all the regions in $I(x) \cap I(y)$ and $I(x) \cup I(y)$, by $|I(x) \cap I(y)|$ and $|I(x) \cup I(y)|$ respectively. The similarity of *x* and *y*, denoted *sim(x, y)*, is expressed as $|I(x) \cap I(y)| \,/\, |I(x) \cup I(y)|$. Note that $0 \leq sim(x, y) \leq 1$ and $sim(x, x) = 1$.

MotifVoter - a method that utilizes the sites predicted by multiple motif finders

Consider *m* basic motif finders, each reporting *n* motifs. Each motif corresponds to its list of predicted binding sites. MotifVoter aims to integrate the information and to give an accurate prediction of the binding sites. The main assumption behind the method is that the true binding sites have a higher chance to be predicted by more than one motif finders.

There are three stages in MotifVoter: (1) Motif filtering: this stage filters away the spurious motifs from all the candidate motifs predicted by the *m* motif finders (see **Figure 1a**). (2) Instance refinement: based on the candidate motifs retained in Stage 1, we identify a set of instances with high confidence that they are real binding sites (see **Figure 1b**). (3) PWM generation: from the instances computed in Stage 2, we generate the PWM of the motif (see **Figure 1c**).

Stage 1: Motif Filtering

This stage uses a variance-based statistical measure[16,17] to distinguish candidate motifs that look similar to the true motif from those that are different. Suppose that we run *m* motif finders and each motif finder reports its top-*n* candidate motifs. We will get a set *P* of *mn* candidate motifs. Among all the candidate motifs in *P*, some of them can approximate the real motif while the other cannot. We would like to identify the subset *X* of *P* such that the candidate motifs in *X* are likely to approximate the real motif. Our basic idea is that if the candidate motifs in *X* can model the real motif, they should have high similarity. Below, we define a score function which allows us to identify *X*.

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:

$$\frac{\sum_{x,y \in X} sim(x, y)}{|X^2|}$$

The *w* score of *X*, denoted by *w(X)*, is defined as:

$$\frac{|X|^2 \, sim(X)}{\sqrt{\sum_{x,y \in X} (sim(x, y) - sim(X))^2}}$$

Note that *w(X)* measures how similar 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 *P-X*, should have a low *w(P-X)*. In other word, if *X* is the set of candidate motifs which approximate the real motif, we expect to have a high *A(X)* score, where:

$$A(X) = \frac{w(X)}{w(P - X)}$$

In addition, we also assume that most of the motif finders are effective. In other word, for each motif finder, if we select its top *n* candidate motifs for some big enough *n*, we expect at least one of these top *n* candidate motifs approximates the real motif. Based on this assumption, we have an additional criterion that *X* must contain candidate motifs predicted by at least *t* motif finders for some pre-defined threshold *t*. In our experiments, we set *n=30* and *t=m*.

In summary, this stage aims to find $X \subseteq P$ which (1) maximizes *A(X)* and (2) *X* contains the candidate motifs predicted by at least *t* motif finders. Please refer to **Supplementary 5** for our proposed heuristics to identify *X*.

Stage 2: Instance Refinement

Given *X*, we obtain the list of instances using two criteria. First, we accept all regions which are covered by instances of 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 instances of the motif in *X* that have the highest confidence to approximate the real motif the best. To rank the candidate motifs *x* in *X*, we use a confident score defined as follows. Let *B(x)* be the total number of nucleotides covered by the instance of *x*. Let *O(x)* be the total number of nucleotides covered by the instances of *x* and also the instances of the motif *y* where *y* is a motif in *X* predicted by some other motif finder. The confident score of *x* is defined as *O(x)/B(x)*.

For the selected instances that are covered by more than one motif finder, we further apply a post-processing procedure to refine each instance 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. **Supplementary 6** gives the example of our post-processing procedure.

Stage 3: PWM generation

Given all the instances predicted by MotifVoter, Stage 3 generates a PWM motif to model the instances. This stage has two steps: First, a multiple sequence alignment of those instances are computed using MUSCLE[14]. Second, a PWM is generated from the alignment to model the motif. **Figure 1c** provides an illustration of Stage 3.

Performance Measure

The performance measures used in the paper are the same as the ones used in Tompa et al.'s study, namely *sensitivity* (*nSn*), *positive predictive value* (*nPPV*), *performance coefficient* (*nPC*), and *correlation coefficient* (*nCC*). Index *n* is used to denote that the fact the assessment is done

at the nucleotide level instead of site level. Note that there is no consensus on what measures are the most appropriate to evaluate all different motif finders. The selected measures focus on the accuracy of predicting the locations of actual binding sites. The definitions of these performance measures are as follows:

- $nSn = \frac{nTP}{nTP + nFN}$
- $nPPV = \frac{nTP}{nTP + nFP}$
- $nPC = \frac{nTP}{nTP + nFN + nFP}$
- $nCC = \frac{nTP\ nTN - nFN\ nFP}{\sqrt{(nTP + nFN)(nTN + nFP)(nTP + nFP)(nTN + nFN)}}$

References

1. Lawrence, C. et.al. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*. **262**, 208-14 (1993).
2. Thijs, G. et.al. A Gibbs sampling method to detect over-represented motifs in the upstream regions of co-expressed genes. *Journal of Computational Biology*. **23,** 447-464 (2002).
3. Favorov, A.V. et.al. A Gibbs sampler for identification of symmetrically structured spaced DNA motifs with improved estimation of the signal length. *Bioinformatics*. **21,** 2240-5 (2005).
4. Wei, Z. and Jensen, S.T. GAME: detecting cis-regulatory elements using a genetic algorithm. *Bioinformatics*. **22,** 1577-1584.
5. Ao, W. et.al. Environmentally induced forgut remodelling by PHA-4/FoxA and DAF-12/NHR. *Science*. **305,** 1743-1746 (2001).
6. Pavesi, G. et.al. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*. **16,** S207-S214 (2001).
7. Eskin, E. and Pevzner, P. Finding composite regulatory patterns in DNA sequences. *Bioinformatics*. **18,** S354-S363 (2002).
8. Wijaya,E., Rajaraman, K., Yiu,S.M., Sung,W.K. Detection of Generic Spaced Motifs Using Submotif Pattern Mining. *Bioinformatics*. **23**, 1476-1485 (2007).
9. Tompa, M. et.al. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*. **23**, 137-144 (2005).
10. Chakravarty, A., Carlson, J.M., Khetani, R.S., and Gross, R.H., A parameter-free algorithm for improved *de novo* identification of transcription factor binding sites. *BMC Bioinformatics* **8**, 249 (2007).
11. Hu, J., Yang, Y.D. and Kihara, D. EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences. *BMC Bioinformatics*. **7,** 342 (2006).
12. Savage, L. *The Foundations of Statistics*, 2nd ed. New York: Dover Publications (1972).
13. Borgonovo E. Measuring uncertainty importance: investigation and comparison of alternative approaches. *Risk Anal*. **26**:1349-61 (2006).

14. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Research* **32**(5), 1792-97 (2004).
15. Hughes, J.D., Estep, P.W., Tavazoeie, S. and Church, G.M. Computational identification of cis-regulatory elements associated with functionally coherent groups of genes in Saccharomyces cerevisiae. *J. Mol. Biol*. **296,** 1205-1214 (2000)**.**
16. Workman, C.T. and Stormo, G.D. ANN-Spec: a method for discovering transcriptor factor binding sites with improved specificity. *Proceedings of 7th Pacific Symposium of BioComputing (PSB)*, 467-478 (2000).
17. Liu, X., Brutlag, D.L., and Liu, J.S. BioProspector: discovering DNA motifs in upstream regulatory regions of co-expressed genes. *Proceedings of 7th Pacific Symposium of BioComputing (PSB)*, 127-138 (2000).
18. Liu, X., et.al. An algorithm for finding protein-DNA binding sites with application to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology*, **20**, 835-839 (2002).
19. Bailey, T. and Elkan, C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, **21**, 51-80 (1995).
20. Blanco, E. et.al ABS: a database of annotated regulatory binding sites from orthologous promoters. *Nucleic Acids Research*. **34,** D63-67 (2006).
21. U. Benbow, U. and Brinckerhoff, C. The AP-1 site and MMP gene regulation: what is all the fuss about?, *Matrix Biology* **15,** 519-526 (1997).
22. Lenhard,B., Sandelin,A, Mendoza,L., Engström, P., Jareborg, N. and Wasserman, W.W., Identification of conserved regulatory elements by comparative genome analysis. *Journal of Biology* **2**,13 (2003).
23. Krivan, W. and Wasserman, W.W. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Research* **11**, 1559-1566 (2001).
24. Wasserman, W.W. and Fickett, J.W. Identification of regulatory regions which confer muscle-specific gene expression. *Journal of Molecular Biology* **278**,167-81 (1998).
25. Rozek, D. and Pfeiffer, G.P. In vivo protein-DNA interactions at the c-jun promoter: preformed complexes mediate the UV response. *Mol. Cell. Biol*. **13**, 5490-5499 (1993).
26. Blanchette, M. and Tompa, M. Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting. *Genome Research* **12**, 739-748 (2002).
27. Dermitzakis, E.T. and Clark, A.G., Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Molecular Biology and Evolution.* **19**, 1114-1121. (2002).
28. Salgado, H. et.al. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherechia coli K-12. *Nucleic Acids Research*. **32,** D303-306 (2003).
29. Bockhorst J, Qiu Y, Glasner J, Liu M, Blattner F, Craven M. Predicting bacterial transcription units using sequence and expression data. *Bioinformatics*. 19, S34-43 (2003).
30. Sinha, S. Blanchette, M. and Tompa, M. PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* 5, 170 (2004).