Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Performance evaluation of DNA motif discovery programs

Chandra Prakash Singh[1*], Feroz Khan[2], Bhartendu Nath Mishra[3], Durg Singh Chauhan[4]

[1]Department of Computer Sciences, R.S.M.T., U.P. College, Varanasi, India; [2]Bioinformatics Division, C.I.M.A.P. (C.S.I.R.), Lucknow, India; [3]Department of Biotechnology, I.E.T., U.P. Technical University, Lucknow, India; [4]Uttarakhand Technical University, Dehradun, India; Chandra Prakash Singh* - Email: chand_11_2000@yahoo.com; Phone: 91 542 2319081; * Corresponding author

received November 11, 2008; accepted December 01, 2008; published December 31, 2008

Abstract:

Methods for the identification of transcription factor binding sites have proved to be useful for deciphering genetic regulatory networks. The strengths and weaknesses for a number of available web tools are not fully understood. Here, we designed a comprehensive set of performance measures and benchmarked sequence-based motif discovery tools using large scale datasets (derived from *Escherichia coli* genome and RegulonDB database). The benchmark study showed that nucleotide based and binding site based prediction accuracy is often low and activator binding site based prediction accuracy is high.

Keywords: regulatory proteins; DNA binding site; motif discovery; accuracy; evaluation

Background:

Computational identification of transcription factor (TF) binding sites from the upstream regions of genes has proved to be extremely valuable in functional genomics studies **[1]**. There are a large number of prediction tools for identifying regulatory elements from DNA sequences with or without additional information. It has also been realized that current activator binding site discovery tools are far from perfect. Evolutionary relationship profile and microarray data has also been incorporated to improve prediction accuracy **[2]**.

The availability of different web tools for regulatory binding site discovery is large in number. Yet, there are few systematic comparative benchmark studies to independently evaluate the prediction performance of existing TF binding site discovery tools **[3, 4, 5, 6]**. These measures are defined to guide users to choose from consensus methods. However, no widely used web tools are evaluated for activators and repressors TF binding sites. The effect of factors such as sequence number and scalability is also not yet characterized. Two other web tools using synthetic data and real datasets from yeast are compared with YMF program **[7, 8]**. About 13 motif discovery tools have also been evaluated using a well defined datasets **[4, 9, 10]**.

Here, we used the activators binding sites for the evaluation of available tools. We defined a set of prediction performance indexes for activator binding site discovery tools and performed a comparative evaluation of existing activator binding site discovery tools using prediction accuracy, scalability and reliability of significance scores for activators transcription factors from RegulonDB (binding site information for *E. coli* is available).

We investigated the width of a target activator binding site, the number of input sequences and the information content of target activator binding sites for prediction accuracy. We then describe rules for using and improving activator binding site discovery tools. Tools such as AlignACE **[11]**, MEME **[12]**, Consensus **[5]**, MotifSampler **[13]** and PoSSuMsearch **[14]** are evaluated in this study. The tools are selected based on wide usage, scalability and availability. The benchmark experiments show that the performance of popular activator binding site tools based only on DNA sequence information is low with ~14–25% accuracy for nucleotide sequences and 36–46% for binding sites.

Methodology:

Datasets

We used activators binding sites information of *Escherichia coli* K-12 stored in RegulonDB **[16]** to generate various types of input sequence sets and also to compare the difference in prediction performance **[6]**. RegulonDB is selected based for the following reasons: (1) It is used by many groups for benchmark study **[13]**; (2) It complements the latest benchmark study in which all the known *E. coli* TF binding sites from RegulonDB were used **[6]**; (3) It has also been used for a comparative study of TF binding site representation and motif search algorithms **[4, 10]**; and (4) for evaluation of a motif discovery and a motif search algorithm tools **[9]**.

A dataset of 45 test sequences are generated from RegulonDB activators transcription factors. The raw data for generating input sequences include three files: (1) Ecoli_RegulonDB containing experimentally determined activators binding sites information including TFs, start and end positions on the genome and location on the forward or reverse sequence **[16]**; (2) Ecoli_Gene containing start and end positions of genes in the genome; (3) Ecoli_Genome which is the whole *E. coli* genome sequence taken from KEGG database **[17]**.

Binding site records in RegulonDB are organized in groups which bind to the same TF. The following binding sites records are discarded from RegulonDB: (a) any record that does not have positional information on the genome, (b) any duplicated record, (c) any record that differs with other binding site records by a < 5 nt shift, (d) finally, we remove binding site groups with only one sequence. We refer to this curated dataset as ECOLI_ACTIVATORS_RDB_45. It should be noted that datasets ECOLI_ACTIVATORS_RDB_45 are the source for a variety of input datasets.

We generated two types of datasets (Type A and B) from ECOLI_ACTIVATORS_RDB_45. Type-A datasets are generated from the intergenic regions of *E. coli* genome. It is generated as follows: for each known activator binding site of an activator binding site group, we align it to the *E. coli* genome, locating the adjacent genes to the binding site and extracting the intergenic region to generate one input sequence. If all the binding sites in an activator binding site group are located in the same intergenic region, then only one intergenic sequence is extracted and the corresponding activator binding site group was discarded. The final dataset has 40 activator binding site groups and is termed as ECOLI_ACTIVATORS_RDB_40A. It has the following characteristics: the average number of sequences per activator binding site group: 10; the average number of sites per sequence: 2; the average sequence length: 400 nt; and the average site width: 15.

Type-B datasets include sequences with symmetric margins on both sides of known activators binding sites. For each binding site of an activator binding site group, we align it with the *E. coli* genome and extend the binding site in both directions by adding symmetric margins of a given length along the genome. In this manner, we can define a series of datasets with increasing margin sizes to test the scalability of activator binding site discovery algorithms. In some activator binding site groups, multiple binding sites appear in a single sequence when the margin size is large. Thus, when the margin size is >200, all the binding sites appear in each of the input sequences despite their differences. We kept these exceptional cases in the dataset because this case also happens in a real situation. Each Type B dataset namely, ECOLI_ACTIVATORS_RDB_45B-X with margin size X has the following characteristics: there are 45 binding site groups, each with at least two sequences; the average number of sequences per activator binding site group is 10; the average number of sites per sequence is 2; the average site width is 15. The high values of the standard deviations reflect the diversity and variation among input sequence sets. For type B datasets, we observe that when the margin sizes are larger (e.g. >400 nt), some part of the sequences are located in the coding regions. However no significant influence has been observed of these variations on the prediction accuracy. Type A dataset is suitable for analyzing activator binding site discovery for co-expressed genes while type B provides a good model for analyzing data from ChIP–chip experiments.

Algorithms tested

The characteristics of algorithms used in this study are briefly described.

AlignACE

AlignACE **[11]** is a stochastic activator binding site discovery algorithm based on widely adopted Gibbs Sampling method **[18]**. Compared with the original Gibbs Sampling method, it adds the following major features: both strands of sequences are searched; near-optimum sampling is improved; an iterative masking approach is used to search multiple activator binding sites. Running parameters for AlignACE are set as default except that the gc_back (the background GC content) is set as 0.5 and the expected activator binding site width is set to 15 unless otherwise specified. We have investigated the effect of the activator binding site width setting as summarized in Table 3 (supplementary material). The major statistical score used by AlignACE, the MAP score, measures the degree to which an activator binding site is over-represented relative to the expected random occurrence of such an activator binding site in the sequence.

MotifSampler

MotifSampler **[13]** is another motif discovery algorithm based on Gibbs sampling. It extends the original Gibbs Sampling approach in two ways. First, it introduces a higher-order Markov background model. Second, it incorporates a Bayesian mechanism to estimate the number of activator binding sites occurring in each sequence. We made the following adjustments to the default parameter values. We only search input sequences without including its reverse complements because all known sites are aligned on the forward direction of the input sequences. We search five different activator binding sites with width 10. The number of repeating runs is set to five. The background frequency model is generated using the intergenic region sequences of *E. coli* genome and the third-order Markov model is used unless otherwise specified.

Consensus

Consensus **[5]** is an enumerative deterministic greedy algorithm. It selects several top activators binding site candidates according to the chip-array enhancement score to build activator binding site models and then employs a greedy strategy to improve the models. We used the default parameter set except for the activator binding site width, which is set to 10.

MEME

MEME (Multiple Expectation Maximization Estimation) **[12]** is based on the expectation maximization (EM) technique. With a given activator binding site of width *w*, MEME first decomposes original sequences into *w*-mers. Each *w*-mer could be an activator binding site or a background subsequence to be determined by the activator binding site and background model components. The search space increases significantly with increasing number of sequences and sequence lengths. It is the only algorithm in this evaluation that does not require an activator binding site width parameter. We set the maximum dataset size in characters as 1 million, maximum running time as 3600 CPU seconds, maximum number of activator binding sites to find as five and minimum number of sites for each activator binding site as one. The rest of the parameters are used as default. The background frequency model is generated using the whole *E. coli* genome and the third-order Markov model is used unless otherwise specified.

PoSSuMsearch

PoSSuMsearch **[14]** includes fast and sensitive matching of position specific scoring matrices using enhanced suffix arrays. It has a new non-heuristic algorithm called ESAsearch, to efficiently find matches of such matrices in large databases. It preprocesses the search space e.g. a complete genome or a set of protein sequences and builds an enhanced suffix array which is stored in a file. The enhanced suffix array only requires 9 bytes per unit symbol and allows searching a database with a PSSM in sub linear expected time. It also addresses the problem of non-comparable PSSM-scores by developing a method which allows for computing a matrix similarity threshold for a PSSM, given an E-value or a P-value. It is based on dynamic programming. In contrast to other methods it employs lazy evaluation of the dynamic programming matrix. It only evaluates those matrix entries that are necessary to derive the sought similarity threshold.

Measures of prediction accuracy

There are several prediction accuracy measures for evaluating motif discovery algorithms **[3, 5]**. Here, we use three levels of performance criteria: nucleotide, binding site and motif levels (see supplementary material).

Results:

Performance on ECOLI_ACTIVATORS_RDB_40A set

Table 1 (supplementary material) shows the prediction performance at the nucleotide, binding site and motif levels for the five motif discovery algorithms especially used only for known activators of *E. coli*. The accuracy scores of AlignACE, MotifSampler and Consensus are averaged over 100 runs. First, we found that at the nucleotide level, the prediction accuracy of all algorithms is relatively low. The maximum sensitivity, specificity and performance coefficient are only 0.365, 0.401 and 0.32, respectively. The accuracy levels are higher than the performance scores reported previously **[4]**. This is due to their longer sequences ranging from 400 to 2000 nt, while the sequence lengths in ECOLI_ACTIVATORS_RDB_40A vary from 100 to 800 nt (average: 400 nt). PossumSearch achieved the highest performance coefficient and specificity while MEME has the best sensitivity, partly due to its capability to estimate activator binding site lengths.

The prediction performance at the binding site level is better than the nucleotide level. The maximum specificity reaches 0.584 for PossumSearch and the maximum performance coefficient reaches 0.469 for PossumSearch. These accuracy scores are higher than what was reported before because we regard overlaps with one or more nucleotides as sufficient to qualify as a correct prediction, while at least 4 nt overlaps were needed in the previous work **[4]**. The justification is that when a predicted binding site overlaps with the true site with at least 1 nt, it is not difficult for experimental biologists to locate the true binding site position around the predicted anchor position since the activator binding site width is only 10–20 nt on average. This higher prediction accuracy at binding sites implies that at least these algorithms can locate rough positions of binding sites. At the binding site level, MEME is comparable with Consensus in terms of performance coefficient scores, both are better than MotifSampler and AlignACE. This means that MEME indeed improves the prediction performance of the simple Gibbs Sampling method. We also found that MEME is the best in terms of sensitivity and PossumSearch is best in terms of sequence success rate (sSr) while MEME is the second.

We found that the activator binding site success rates of all five algorithms are > 0.90. This comparison demonstrates that the algorithms are able to reliably predict at least one correct binding site from all activator binding site groups. This fact could be potentially exploited to improve existing algorithms. We also found the activator binding site level success rate, *mSr* of PossumSearch is the highest among the five algorithms, showing that PossumSearch can handle more diverse input sequences.

Another interesting observation is that the prediction accuracy of stochastic algorithms, such as AlignACE and MotifSampler, are very stable over multiple runs. For the mean nPC scores of AlignACE and MotifSampler, the standard deviation is 0.08 for 100 runs. In this study, we evaluated the accuracy of the best prediction out of top five scoring predictions. This is because in practice biologists can test candidate activator binding sites by experiments if they know the correct sites are included in the top five predictions with a reasonably high probability.

First, it is evident that on average the top-scoring activator binding site is not the best prediction. For example, in the case of MotifSampler the top-scoring activator binding site corresponds to the best prediction in only 40% of the cases. Second, the discrepancy of the accuracy between the best and the worst prediction is relatively larger for AlignACE, MEME and MotifSampler, and the mean accuracy of them are lower than the other two algorithms. We found that this results from the way these three algorithms find the next best-scoring activator binding sites: once the top-scoring activator binding site is found, its positions are masked out so that no subsequent sites are overlapped with them. Therefore, averaging the accuracy of the multiple top-scoring activator binding sites is disadvantageous for the three algorithms.

Scalability

The scalability is about the performance changes with the increase in the number of sequences, the activator binding site width and the sequence length. We generated six types of datasets with different margin sizes (extending on both sides of target activator binding sites) of 50, 100, 200, 300, 400 and 500. Hence, the total sequence length is the target activator binding site width plus twice the margin size. Each type has 45 activator binding site groups with at least two sequences in a dataset. We ran the five algorithms with the same parameter settings as in the previous section.

First at the nucleotide level, the performance of all the algorithms decreases significantly as the sequence length increases. When the margin size is <200 nt, all algorithms except for AlignACE showed a similar performance. What is interesting is that when the margin size becomes larger than 400 nt, PossumSearch, Consensus and MEME become the best algorithms, while MotifSampler and AlignACE become quite ineffective. It should be noted that AlignACE and MotifSampler are based on Gibbs sampling strategy while MEME and Consensus have an enumerative component in their search strategy. This performance discrepancy shows that for long input sequences, Gibbs sampling strategy tends to become too inefficient to identify the binding sites correctly.

At the binding site level, PossumSearch, Consensus and MEME are the best algorithms, especially when the sequence length (double margin size) becomes >300 nt. Here, PossumSearch is the best with its capability to locate at least one correct binding site for a given dataset. In this test, PossumSearch also has a high success rate. To examine why PossumSearch has the highest activator binding site level success rate, we compare the sensitivity (nSn) and the specificity (nSp) of these algorithms. It is illuminating that PossumSearch has a dominant sensitivity, contributing to its high success rate (mSr). One possible explanation of PossumSearch's high sensitivity is due to its enhanced lazy algorithm unlike others having a local search component.

Effect of different parameters on expected width and number of motifs

Motif discovery algorithms have several parameters to tune its prediction performance. Here, we examined the effect of two of the most critical parameters of the algorithms. One is the expected motif width W_e, and the other is the expected number N_{bs} of binding sites for a dataset. To evaluate how the parameter W_e affects the performance, we run the algorithms on the ECOLI_ACTIVATORS_RDB_40 B-200 dataset using different W_e ranging from 5 to 20. For stochastic algorithms, such as AlignACE and MotifSampler, the experiments are repeated for 10 times and the average scores are reported. Since MEME can adaptively estimate the best motif width, we only conducted a single run using the parameter setting specified in the Methodology.

Table 3 (supplementary material) shows how the nucleotide level accuracy varies with the different parameter of estimated motif width, W_e. Generally speaking, if W_e is too small, the algorithms will be affected in sensitivity. If W_e is too large, they will be affected in specificity. We found that for both PossumSearch and Consensus, the best performance is achieved at W_e = 15, while PossumSearch, Consensus and MEME work best with W_e =10.

We have chosen 15 as the expected activator binding site width, which is approximately the average between the default values of the algorithms (which is 10 expect for MEME) and the average size of the binding sites in the benchmark set. The results in Table 3 (supplementary material) shows that the optimal value for the activator binding site width differs from algorithm to algorithm even for the same benchmark dataset. The value 15 used in this analysis performed better or equal to MEME for five of the algorithms than using the default value of 10. These results illustrate the difficulty for biologists to tune good parameters when they use these algorithms.

Effect of the number of input sequences

In this section, we investigate how the number of sequences in a given input sequence set affects the prediction accuracy because it is a dominant factor that determines the time complexity of activator binding site discovery algorithms. For this study, input sets with K (= 5, 10, 20, 30, 40) number of sequences is generated as follows: first, we select those activator binding site groups which has at least 40 sequences. For each activator binding site group, we extend each binding site with 200 nt on both sides to create raw input sequences. We randomly select K sequences without duplicates from each such set of raw input sequences. Ten such sequence sets are created for each K. We then run the activator binding site discovery algorithms on all datasets. The prediction accuracy scores are then averaged for all the input sequence sets with the same number of (K) sequences. It is observed that when the number of sequences becomes >10, the performance coefficient at nucleotide level becomes stable. More input sequences do not improve the prediction accuracy. Results show that the binding site level accuracy is almost independent of the number of sequences except that PossumSearch seems to benefit from more input sequences. Therefore, for a large input sequence set, it is recommended to use an algorithm which has a good scalability to the number of sequences, such as PossumSearch. If a user insists to use a computationally demanding algorithm, such as MEME, this observation suggests a novel approach: i.e., one can input only partial input sequences to an activator binding site discovery algorithm to obtain an activator binding site model and then use this model to find activator binding sites in the remaining sequences. In this manner, a significant reduction in the running time can be achieved without sacrificing the prediction accuracy.

Discussion:

We have developed a comprehensive set of performance measures at the nucleotide, binding site and activator binding site levels and systematically evaluated five activator binding site discovery algorithms using a *E. coli* activators motif dataset, ECOLI_ACTIVATORS_RDB_45. We selected algorithms which solely use input sequences for finding motifs, because this is the baseline for any of the recent algorithms which will also incorporate additional information. Special attention is paid to carefully examine factors that affect the prediction accuracy, which have not been carried out in the previous studies. We found that the prediction accuracy at the nucleotide and binding site levels is relatively low while the activator binding site level prediction accuracy is surprisingly high. These conclusions complement the evaluation work reported for eukaryotic datasets **[4]**. We compared the scalability of these algorithms and found that Gibbs Sampling based algorithms tend to fail for long sequences. Other algorithms also show significant degradation when the sequence lengths increase. These results suggest a need for improving scalability of activator binding site discovery algorithms, which is particularly important when activator binding sites are sought from an increasing number of complete genome sequences. We also found that the capability of adapting activator binding site length is important which partially contributes to the dominance of

MEME's high prediction sensitivity. Interestingly, it is observed that increasing the number of input sequences does not always improve the prediction accuracy once it reaches a threshold level, which can be exploited to reduce computational complexity of some algorithms. Another observation is that for noisy real datasets, no strong correlation between significance scores and prediction accuracy is observed for all activator binding site discovery algorithms across all datasets. It implies that the high degree of consensus among multiple predictions of one or more algorithms may indicate their correctness.

Limitations of current DNA motif discovery algorithms
Despite the long-time effort for the motif discovery problem, our benchmarking results show that current sequence-based activator binding site discovery algorithms have several fundamental limitations. First, the nucleotide level and binding site level prediction accuracy are still very low (i.e. nPC and sPC) even on the prokaryotic activator binding sites, which are supposed to be easier to be captured than eukaryotic ones. Therefore, in the current situation users should be aware of the limitations and be extremely careful in interpreting computational predictions. It should be also noticed that the significance score of algorithms do not necessarily correspond directly to the accuracy of found activator binding sites. The lack of scalability is another problem for all the evaluated algorithms. The important technical difficulty which causes limitations is the inherent low signal/noise ratio in only-sequence-based binding site discovery problems. As shown in results, prediction performance decreases significantly as the length of sequences increases for all five algorithms. Several strategies have been proposed to increase the signal-to-noise ratio.

The limitation also comes from the pattern model used to capture the regularity among the binding sites for transcription factors. The PSSM model is used for all five algorithms, with a slight variation. This model, however, has difficulty in modeling gapped motifs and assumes that the nucleotide positions are independent of each other, which is not true in reality. The syntactic deterministic activator binding site models, such as consensus sequence models, suffer from application for short and highly conserved sequences **[21]**. Several methods have been proposed to incorporate position-dependence information, including a novel hidden Markov model method **[21]**, which tries to capture dependency between non-adjacent positions using a position re-ordering method. Besides, per-position information content was also reported as well as local pairwise nucleotide dependencies to improve the activator binding site search performance **[10]**. However, such more advanced activator binding site models have not been incorporated into current motif discovery algorithms.

The local optima phenomena in optimization algorithms should also be mentioned here. Many popular motif discovery algorithms are based on heuristic search algorithms such as Greedy search, Gibbs sampling and Expectation Maximization. The performances of these methods are subject to potential suboptimal solutions in the search space. While usually 10–20 starting points are evaluated to find the most potential search direction, the effectiveness of this simple approach is usually limited for large multi-modal search spaces found in datasets with long sequences. Extensive experiments are needed to evaluate how severe the local optima issue could limit the performance of existing heuristic-based methods and whether stronger global optimization techniques, such as genetic algorithms and others **[22]**, could be used to improve it.

Potentials of DNA motif discovery algorithms
Although the low prediction performance has been revealed on the nucleotide and the binding site level accuracy, we believe that sequence-based activator binding site discovery still has room for improvement. First, we could take advantage of the high motif-level success rate ($mSr \geq 0.9$), a capability to identify at least one binding site correctly for an activator binding site group in ECOLI_ACTIVATORS_RDB_45 (i.e. mSr). A remarkable characteristic of the motif level success rate is the better tolerance to a longer input sequence size, i.e. a better scalability. Based on this observation, one natural idea of searching activator binding sites in a set of long sequences is to perform the motif search in two steps, namely, to perform the second search just in the vicinity of motifs identified in the initial search. In this manner, the search space could be greatly reduced. A trend of recent activator binding site discovery algorithms is to incorporate additional information, such as phylogenetic trees or family sequences, to improve the predication accuracy **[7]**. This strategy can effectively increase the signal/noise ratio, thus improving both specificity and sensitivity. In contrast, in our study, we have showed limitations and potentials of current sequence-based algorithms, and indicated ways to take advantage of the potentials for improvement. Since sequence-based approach is the baseline of any modern algorithms, our finding will surely benefit to improve almost all the algorithms.

References:

[01] J. J. Wyrick and R. A. Young, *Curr. Opin. Genet. Dev.,* 12: 130 (2002)
[02] A. Prakash *et al.*, *Pac. Symp. Biocomput.,* 9: 348 (2004)
[03] S. Sinha and M. Tompa, *Nucleic Acids Res.,* 31:3586 (2003)
[04] M. Tompa *et al.*, *Nat. Biotechnol.,* 23:137 (2005)
[05] X. S. Liu *et al.*, *Nat. Biotechnol.,* 20: 835 (2002)
[06] J. Hu *et al.*, *Nucleic Acids Research,* 33: 4899 (2005)
[07] S. Sinha *et al.*, *BMC Bioinformatics,* (2004) 5:170
[08] S. Sinha and M. Tompa, *Third IEEE Symposium on BioInformatics and BioEngineering (BIBE'03), Bethesda, Maryland*, 214 (2003)
[09] E. Benitez-Bellon *et al.*, *Genome Biol.,* (2002) 3:research0013.1.
[10] R. Osada *et al.*, *Bioinformatics,* 20: 3516 (2004)
[11] F. P. Roth *et al.*, *Nat. Biotechnol.,* 16: 939 (1998)
[12] T. L. Bailey and C. Elkan, *Machine Learning,* 21: 51 (1995)
[13] G. Thijs *et al.*, *Comput. Biol.,* 9:447 (2002)
[14] M. Beckstette *et al.*, *BMC Bioinformatics,* 7: 389 (2006)
[15] J Buhler and M Tompa., *J. Comput. Biol.,* 9: 225 (2002)

[16] H. Salgado *et al., Nucleic Acids Res.,* 32: D303 (2004)

[17] M. Kanehisa *et al., Nucleic Acids Res.*, 30: 42 (2002)

[18] C. E. Lawrence *et al., Science,* 262: 208 (1993)

[19] M. Burset and R. Guigo, *Genomics,* 34: 353 (1996)

[20] S. Rogic *et al., Genome Res.,* 11:817 (2001)

[21] K. Ellrott *et al., Bioinformatics,* 18: S100 (2002)

[22] C. G. Moles *et al., Genome Res.,* 13: 2467 (2003)

Edited by P. Kangueane

Citation: Singh *et al.*, Bioinformation 3(5): 205-212 (2008)

Supplementary material

Tools	Nucleotide level			Binding site level			Motif level	
	nPC	nSn	nSp	sPC	sSn	sSp	mS_r	sSr
AlignACE	0.17	0.24	0.198	0.276	0.399	0.379	0.91	0.54
MotifSampler	0.19	0.225	0.29	0.31	0.442	0.392	0.91	0.61
Consensus	0.25	0.282	0.335	0.341	0.431	0.458	0.93	0.61
MEME	0.26	0.365	0.305	0.397	0.565	0.54	0.94	0.63
PossumSearch	0.32	0.348	0.401	0.469	0.557	0.584	0.95	0.65
Average	0.238	0.292	0.3058	0.3586	0.4788	0.4706	0.928	0.608

Table 1: Prediction accuracy on the *E. coli* intergenic region dataset is given.

Algorithm	Best	Worst	Mean	Standard deviation
AlignACE	0.17	0.02	0.10	0.07
MotifSampler	0.19	0.03	0.11	0.08
Consensus	0.25	0.07	0.15	0.09
MEME	0.26	0.05	0.15	0.10
PossumSearch	0.32	0.06	0.16	0.13

Table 2: The statistics of the top five predictions in terms of nPC on ECOLI_ACTIVATORS_RDB_40A set.

Algorithm	Motif width			
	5	10	15	20
AlignACE	0.068	0.241	0.194	0.131
MotifSampler	0.031	0.240	0.196	0.136
Consensus	0.162	0.248	0.282	0.148
MEME	0.149	0.246	0.254	0.136
PossumSearch	0.177	0.251	0.292	0.158

Table 3: Influence of estimated motif width on the nucleotide level prediction accuracy (nPC)

Nucleotide level accuracy

First, for each target binding site with overlapping predicted binding sites in an input sequence, we define the following values for calculating accuracy metrics at the nucleotide level: *n*TP (true positive), the number of target binding site positions predicted as binding site positions; *n*TN (true negative), the number of non-target binding site positions predicted as non-binding site positions; *n*FP (false positive), the number of non-target binding site positions predicted as binding site positions; *n*FN (false negative), the number of target binding site positions predicted as non-binding site positions. The sensitivity, specificity and performance coefficient over a pair of target/predicted binding sites is defined as:

Sensitivity **Specificity** **Performance coefficient [4]**

$$nS_n = \frac{nTP}{nTP + nFN} \qquad nS_p = \frac{nTP}{nTP + nFP}. \qquad nPC = \frac{nTP}{nTP + nFP + nFN}$$

According to this definition, the nPC value ranges over (0, 1) with the perfect prediction being the value of 1. Compared with the correlation coefficient (CC) **[19, 20]**, nPC has several benefits: it is straightforward to interpret and practically, it also tells the experimental biologists the probable ranges that the true binding sites are located around the predicted positions.

Binding site level accuracy

The binding site level accuracy indicates whether predicted binding sites overlap with true binding sites by one or more nucleotide position. We define, sTP, sFP and sFN as follows: sTP, the number of predicted binding sites which overlaps with the true binding sites by at least 1 nt; sFP, the number of predicted binding sites which have no overlaps with the true binding sites; sFN, the number of true binding sites that have no overlaps with any predicted binding sites. For each input sequence, we define the following accuracy metrics at the binding site level:

Performance coefficient **Sensitivity** **Specificity**

$$\mathrm{sPC} = \frac{\mathrm{sTP}}{\mathrm{sTP} + \mathrm{sFP} + \mathrm{sFN}} \quad \mathrm{sSn} = \frac{\mathrm{sTP}}{\mathrm{sTP} + \mathrm{sFN}}. \quad \mathrm{sSp} = \frac{\mathrm{sTP}}{\mathrm{sTP} + \mathrm{sFP}}.$$

The binding site level accuracy score of an input sequence set (e.g. ArcA) is the average of the scores over all its sequences. The binding site level accuracy score of the entire benchmark dataset is the average of the scores for all input sequence sets.

Sequence motif level accuracy

To evaluate the capability to find at least one binding site in an input sequence, we define the sequence level success rate as the number of sequences N_s that have at least one correctly predicted motif divided by the total number N of sequences in an input sequence set:

$$\mathrm{sSr} = \frac{N_s}{N} \qquad \mathrm{mS_r} = \frac{Np}{M}.$$

The overall sequence success rate of an algorithm is thus the average of sSr over all the input sequence sets. We introduce the motif level success rate score *mSr*, a sensitivity measure, to evaluate the adaptability of an algorithm to different types of motifs, is defined as the number of target motif groups N_p, which have at least one correctly predicted binding site divided by the total number of target motifs (M = 45). A prediction is regarded as correct when the predicted motif overlaps with the target motif by at least 1 nt.