MProfiler: A Profile-Based Method for DNA Motif Discovery

Doaa Altarawy, Mohamed A. Ismail, and Sahar M. Ghanem

Computer and Systems Engineering Dept.
Faculty of Engineering, Alexandria University
Alexandria 21544, Egypt
{doaa.altarawy,maismail,sghanem}@alex.edu.eg

Abstract. Motif Finding is one of the most important tasks in gene regulation which is essential in understanding biological cell functions. Based on recent studies, the performance of current motif finders is not satisfactory. A number of ensemble methods have been proposed to enhance the accuracy of the results. Existing ensemble methods overall performance is better than stand-alone motif finders. A recent ensemble method, MotifVoter, significantly outperforms all existing stand-alone and ensemble methods. In this paper, we propose a method, MProfiler, to increase the accuracy of MotifVoter without increasing the run time by introducing an idea called center profiling. Our experiments show improvement in the quality of generated clusters over MotifVoter in both accuracy and cluster compactness. Using 56 datasets, the accuracy of the final results using our method achieves 80% improvement in correlation coefficient nCC, and 93% improvement in performance coefficient nPC over MotifVoter.

Keywords: Bioinformatics, DNA Motif Finding, Clustering.

1 Introduction

Computational identification of overrepresented patterns (motifs) in DNA sequences is a long-standing problem in Bioinformatics. Identification of those patterns is one of the most important tasks in gene regulation which is essential in understanding biological cell functions. Over the last few years, the sequencing of the complete genome of large variety of species (including human) has accelerated the advance in the filed of Bioinformatics [1].

The problem of DNA motif finding is to locate common short patterns in a set of co-regulated gene promoters (DNA sequences). Those patterns are conserved but still tend to vary slightly [2]. Normally the patterns (motifs) are fairly short (5 to 20 base pair long) [3]. Those motifs are the locations where transcription factors (TF) bind to in order to control protein production in cells. DNA motifs are also called transcription factor binding sites (TFBS). Many computational methods are being proposed to solve this problem. Their strategies can be divided into two main classes: exhaustive enumeration and probabilistic methods [4].

V. Kadirkamanathan et al. (Eds.): PRIB 2009, LNBI 5780, pp. 13–23, 2009.

Review of the field and description of some motif finding methods can be found in [1,2,3,4,5,6].

Several studies show that current motif finding methods are unsatisfactory [3,5,7]. In Tompa et al.'s [7] assessment, 13 motif finding methods were examined. Their study shows that the accuracy of those methods in terms of sensitivity and specificity is low. Despite the large number of methods being proposed for motif finding, it is still a challenging problem.

Motifs found by different methods are not always the same, meaning that their results can be complementary [7,8]. Although the accuracy of a single motif finder method is low, ensemble methods are promising. Ensemble methods are compound algorithms that combine the results of multiple predictions from multiple algorithms. Thus, combing more than one stand-alone method can increase the sensitivity (more true positives), but without a good filtering method it will reduce the specificity (more false positives) [7]. In the last few years, some ensemble methods have been proposed such as SCOPE [9], BEST [10], EMD [11], and more recently, MotifVoter [8]. MotifVoter significantly outperforms all existing stand-alone and ensemble methods. For example, in Tompa's benchmark MotifVoter increased the accuracy of the results (correlation coefficient nCC) over the best stand-alone method by more than 100%.

MotifVoter formulates the ensemble motif finding as an optimization search problem and uses a heuristic to generate a search space consisting of clusters of motifs. It uses a variance-based objective function to select the best cluster among the generated search space. In this paper, we propose a method called MProfiler to increase the accuracy of MotifVoter by using a new heuristic to generate the search space. Enhancing the search space in both accuracy and quality improves the final results and reduces the chances of falling in local maximums. A more accurate search space is the one that has higher percentage of motifs with higher collaboration coefficient with the true motifs. The quality of the search space is the compactness of its higher accuracy clusters, since the selection function is variance-based. The proposed technique for search space generation has more than 200% improvement over MotifVoter's in terms of percentage of generated sets having nCC greater than 0.5. In addition, the generated sets are having higher mean and lower variance (i.e more compact) when compared to the sets generated by MotifVoter's approach. Having compact sets is a desirable feature for the objective function because it is variance-based. In our experiments, we compare the proposed MProfiler technique with MotifVoter on 56 different datasets that are proposed on Tompa's benchmark [7]. The correlation coefficient nCC and performance coefficient nPC are used as measures of accuracy for Motif finding methods.

Our experimental results show that MProfiler increases correlation coefficient by 80% over MotifVoter on the same benchmark. In addition, MProfiler increases the performance coefficient by 93%.

The rest of the paper is organized as follows: Section 2 provides an overview of related word and the motivation of our proposed algorithm. Section 3 introduces

the MProfiler algorithm. Section 4 presents experimental results and discussions. Section 5 concludes the paper along with future work.

2 Motivation

Many ensemble motif finding methods make use of the observation that true motifs are shared by multiple motif finders. In addition, MotifVoter in addition proposed the use of the negative information from the false positive motifs that are usually predicted by a few or even one motif finder. MotifVoter performs two selective criteria to find an optimal cluster [8]:

1. Discriminative criterion: select a cluster of motifs that are not only similar, but also have the property that motifs outside the cluster are distant from each other. This is done with a variance-based objective function (see equation (6) in methods section).
2. Consensus criterion: the selected cluster must be predicted by as many motif finders as possible.

After the cluster is chosen, one representative motif is extracted from the cluster (i.e. a cluster center) using a process called site extraction.

The enumeration technique is unfeasible since it takes exponential time. Instead MotifVoter uses a simple heuristic to generate the search space. Let P be the set of all input motifs. MotifVoter only considers subsets $X_{z,j} = \{z, p_1, ..., p_j\}$ for every $z \in P$ and for every $1 < j < |P| - 1$, where p_i's are sorted descending according to its similarity to z, i.e. $sim(z, p_i) > sim(z, p_{i+1})$ and $p_i \in P$.

The heuristic used by MotifVoter to generate the search space produces good search space for the motif finding problem. MotiVoter outperform all stand-alone and ensemble motif finding methods in terms of accuracy [8].

Because the objective function is variance-based, it favors compact clusters even if they are not the optimal ones. In addition, using a variance-based function with different size sets can mislead the selection to smaller clusters, since smaller clusters appear more compact. Therefore, the capability of the objective function to select a more accurate cluster can be improved by making the clustered sets of nearly equal size. In this paper, an ensemble motif finding method, MProfiler, is proposed that improves MotifVoter search space in three desirable features:

1. Increase the percentage of higher accuracy sets.
2. The generated sets are more compact, i.e. having higher mean and lower variance.
3. Clusters examined by the objective function are nearly of equal size.

The proposed MProfiler technique constructs profiles of similar motifs predicted by different finders. Then, the profiles are used to generate the search space. The constructed profiles increase the similarities between motifs if they exist, thus giving them higher score from the variance-based function. Details of how to generate and use the profiles are described in the following section.

3 MProfiler Methods

3.1 Definitions

Problem Statement: Ensemble DNA Motif finding problem can be formalized as follows: given a set of DNA sequences, and the output of different motif finding methods, each output is a set of motifs (i.e $n * m$ motifs in total), it is required to construct a representative motif that is the best approximate of the real motif which is shared in the input sequences.

Motif: A motif is a set of sites where each site is a continuous range of positions representing a subsequence from a DNA sequence.

Motif Similarity: In [8], the similarity between two motifs is defined by (1), where $cov(x_i)$ is all positions covered by motif x. From that motif similarity definition $0 \leq sim(x_i, x_j) \leq 1$, and $sim(x_i, x_i) = 1$.

$$sim(x_i, x_j) = \frac{cov(x_i) \cap cov(x_j)}{cov(x_i) \cup cov(x_j)} \tag{1}$$

Cluster Similarity: The similarity among a cluster X of motifs is defined as the mean pairwise similarity among its members given by (2), where $|X|$ is the number of motifs in the set X.

$$sim(X) = \frac{\sum\limits_{x_i, x_j \in X} sim(x_i, x_j)}{|X|^2} \tag{2}$$

Cluster Center: We define the center of a cluster as the motif that consists of all positions covered by two or more motifs in the cluster, i.e. it is the pairwise intersection of its members, and can be calculated using (3).

$$center(X) = \bigcup_{\substack{x_i, x_j \in X \\ x_i \neq x_j}} [cov(x_i) \cap cov(x_j)] \tag{3}$$

Consensus Cluster Center (Profile): We define consensus center of a cluster as the motif that consists of all positions covered by at least two motifs, such that the intersecting motifs are predicted by two different motif finders and can be calculated using (4). An extra refinement is added by removing sites (continuous positions) that has only two contributing finding methods.

$$consCenter(X) = \bigcup_{\substack{x_i, x_j \in X \\ finder(x_i) \neq \\ finder(x_j)}} [cov(x_i) \cap cov(x_j)] \tag{4}$$

Cluster Weight: There are several weighing functions that can be used to give a score to a set of motifs. In this paper, we compare our technique to MotifVoter [8], and apply the same weight used by MotifVoter as defined by (5).

$$weight(X) = \frac{sim(X)}{\sqrt{\sum_{x_i,x_j \in X} (sim(x_i, x_j) - sim(X))^2}} \quad (5)$$

Objective Function: The objective function is defined in [8] as the ratio between the weight of a chosen set X, and the weight of remaining motifs not belonging to X (i.e. $\bar{X}$) as shown in (6).

$$A(X) = \frac{weight(X)}{weight(\bar{X})} \quad (6)$$

Accuracy Measures: Following Tompa et al. [7] and others, the following accuracy measures are considered. *Sensitivity* is the percentage of known sites that the algorithm was able to find correctly. *Specificity* is the percentage of the predicted sites that are correct.

- ***Nucleotide Correlation Coefficient*** *(nCC)**:*** Nucleotide Correlation Coefficient combines both sensitivity and specificity (Positive predictive value). As nCC calculated by (7), if the predicted motif perfectly coincide with the known motif, then the value of nCC is 1. If they are independent, then the value of nCC is 0. TP, FP, TN and FN are nucleotide true positive, false positive, true negative and false negative respectively [7].

$$nCC = \frac{TP.TN - FN.FP}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \quad (7)$$

- ***Performance Coefficient*** *(nPC)**:*** Performance coefficient captures both specificity and sensitivity in a single accuracy measure using a simple equation. It is the ratio between true positives (true motifs) and all regions that is marked as motifs either correctly or incorrectly. Nucleotide level performance coefficient (nPC) is defined in 8. It ranges from 0 (worst) to 1 (best).

$$nPC = \frac{TP}{TP + FN + FP} \quad (8)$$

3.2 MProfiler Algorithm

Given the output of m stand-alone motif finding methods, it is desirable to produce a motif that best approximates the real motif.

MotifVoter algorithm has three steps. First, a search space consisting of sets of motifs is generated using the heuristic described in section 2. Second, a set is chosen from the generated search space the maximizes the variance-based objective function in equation (6) with consensus criterion satisfied. Finally, the final motif is extracted from the chosen set as described in MotifVoter [8].

Instead of using $n * m$ motifs given by the stand-alone finders in MotifVoter, the proposed MProfiler technique uses a set of generated motif profiles called consensus cluster centers as defined in section 3.1. Using those profiles helps

increase similarities between motifs in the same cluster, if it exists, thus giving them higher score from the variance-based function in equation (6). A profile has at least 3 intersecting motifs predicted by 3 different motif finders. The generation of the profiles is described in Algorithm (1).

Input : set P contains $n * m$ Motifs.
Output: one Motif and PWM for its aligned sites.

1 **foreach** $x_i, x_j \in P$ **do** *compute sim*(x_i, x_j);
2 profiles $\leftarrow \emptyset$;
3 **foreach** *motif* $z \in P$ **do**
4 X $\leftarrow \emptyset$;
5 sortedP $\leftarrow$ sort P according to sim(z,p_i);
6 **for** *top n* $p_i \in$ *sortedP* **do**
7 X $\leftarrow X + p_i$;
8 **if** *(sim(profiles.lastElement,consCenter(X))*$< \epsilon$*)* **then** profiles $\leftarrow$ profiles + consCenter(X);
9 **end**
10 **end**
11 *acceptedCluster* $\leftarrow$ *MotifVoter(profiles)*;
12 *extractSites and generate PWM*;

Algorithm 1. MProfiler pseudo code

The condition in line 8 avoids obtaining very similar profiles from the same group which actually represent the same profile. A new profile is generated only if it differs by at least ϵ within its group where ϵ is any similarity value between 0 and 1. Small ϵ values generate larger number of profiles, which will be merged in line 11. In line 11, MotifVoter algorithm is used to find the cluster X using the objective function in equation (6). Consensus criterion is not needed in this step because it is already applied in generating the profiles.

3.3 Site Extraction

Final sites are extracted from the selected cluster of motifs as in equation (3). Accepted positions are the positions covered by more than one motif in the cluster. The sites are then aligned using MUSCLE [12] and a Position Weight Matrix (PWM) is generated. PWM is a common representation of motifs. A position weight matrix is a matrix of score values that gives a weighted match to any given substring of fixed length. It has one row for each symbol of the alphabet (A, C, G, T), and one column for each position in the motif.

3.4 Time Complexity

Given m motif finders, each with n predicted motifs, the time complexity of our method is $O(m^2n^2)$, which is the same for MotifVoter. First, at most mn^2

profiles are generated. Then, for each profile, the objective function is calculated for m subsets. As in MotifVoter, since motifs are added one by one, the objective function can be calculated in a constant time from the previous value. Unlike MotifVoter, for each profile MProfiler algorithm did not need to add all other profiles to the growing clusters of motifs because sets are more compact. Instead the first most similar m profiles are examined. Thus the final running time is $O(m^2n^2)$.

4 Results and Discussion

4.1 Stand-Alone Motif Finders

We used the same 10 finders used by MotifVoter with the same parameters described in [8]. The stand-alone motif finders are: MEME [13], Weeder [14], Bioprospector [15], SPACE [16], MDScan [17], ANN-Spec [18], MotifSampler [19], MITRA [20], AlignACE [21], and Improbizer [22]. Any other DNA motif finder can be used. For each finder, the first 30 predicted motifs are taken. The top 30 motifs achieve maximum sensitivity (nSn) on Tompa's benchmark [8]. Since Tompa's benchmark is a good representative of real motifs, using top 30 motifs for other datasets is a quite reasonable approximation.

4.2 Datasets

Datasets used in the comparison are the Tompa et al. [7] benchmark consisting of 56 different datasets, which cover 4 different species (Mouse, Fruit fly, Human and Yeast). The datasets are constructed based on real transcription factor binding sites (TFBS).

4.3 Improvement in Search Space

Accuracy: Using performance coefficient nPC as a measure of accuracy, MProfiler has 380% improvement over MotifVoter in percent of generated sets having accuracy $nPC > 0.5$. Fig. 1 shows the total improvement in nPC for all 56 datasets. Improvement of nPC over MotifVoter is calculated using (9).

$$Improvement(nPC) = \frac{nPC_{MProfiler} - nPC_{MotifVoter}}{nPC_{MotifVoter}} \tag{9}$$

Also MProfiler's search space (generated sets) has more than 200% improvement over MotifVoter in percentage of generated sets having higher correlation coefficient, i.e $nCC > 0.5$ as shown in Fig. 2. The figure shows the combined nCC for all 56 datasets. Improvement of nCC over MotifVoter is calculated using (10).

$$Improvement(nCC) = \frac{nCC_{MProfiler} - nCC_{MotifVoter}}{nCC_{MotifVoter}} \tag{10}$$

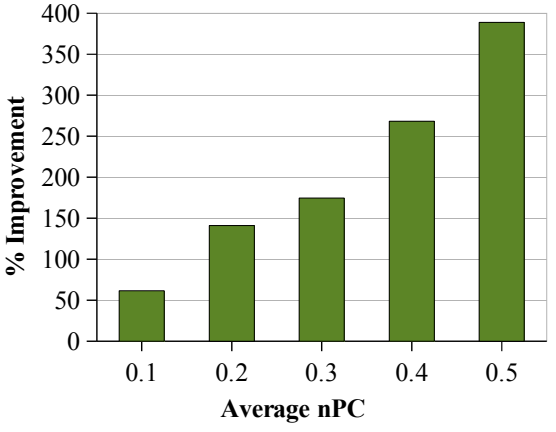

Fig. 1. The overall improvement in accuracy (nPC) of MProfiler over MotifVoter. Y-axis: percent improvement of number of generated clusters having nPC greater than or equal to x. MProfiler's search space (generated sets) has 380% improvement over MotifVoter in percentage of generated sets having higher accuracy (i.e with $nPC > 0.5$).

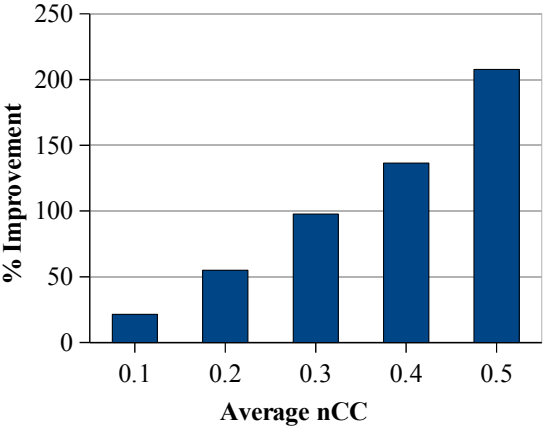

Fig. 2. The overall improvement in accuracy (nCC) of MProfiler over MotifVoter. Y-axis: percent improvement of number of generated clusters having nCC greater than or equal to x. MProfiler's search space (generated sets) has more than 200% improvement over MotifVoter in percentage of generated sets having higher accuracy (i.e with $nCC > 0.5$).

More accurate clusters mean a higher probability to find the correct set, given that their quality are better. Notice that MProfiler has more improvement in higher nCC and nPC values than lower ones which is a desirable feature (i.e. it increases the percentage of higher quality sets more than lower quality sets).

Average Mean and Variance: Since the objective function is based on the mean and the variance of cluster similarity (see equation (6)), it is desirable

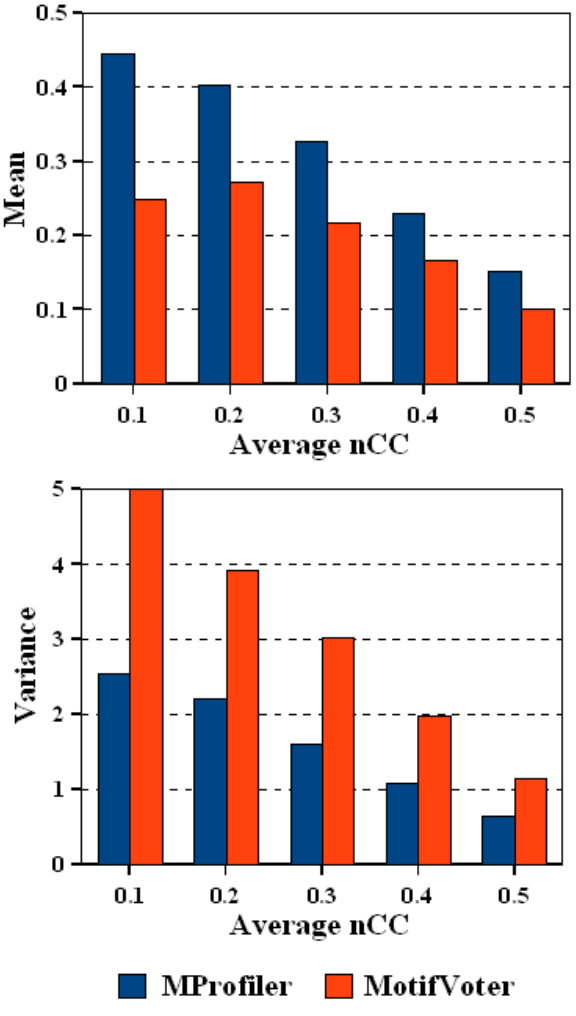

Fig. 3. Average mean and variance of similarity in generated clusters. y-axis is the average mean/variance of similarity for sets having nCC greater than or equal to x.

to make higher accuracy clusters more compact, i.e. with higher mean and lower variance. MProfiler improves both mean (higher value) and variance (lower value) over MotifVoter which led to the improvement in the optimal cluster selected. Fig. 3 shows the improvement of mean and variance of MProfiler generated sets over MotifVoter for all 56 datasets.

4.4 Comparison of Final Results

On 56 different datasets, MProfiler has 80% improvement in accuracy (nucleotide correlation coefficient nCC) over MotifVoter results using the same input and

the same objective function implemented as described by Wijaya et al. [8]. Also MProfiler has 93% improvement in accuracy using performance coefficient nPC as a measure of accuracy. Comparison with the results stated in [8] was not possible because the exact implementation of the objective function is not described in their paper and the source code is not available.

5 Conclusion

Ensemble methods provide improvement in motif finding accuracy without the need to use additional data (such as phylogenetic information or characterization of the domain structure of the transcription factor), which are not always available. Our proposed method, MProfiler, improves the best existing motif finding ensemble method, MotifVoter, in terms of accuracy without increasing time complexity.

On the widely used Tompa's benchmark with 56 different datasets, MProfiler's search space has 200% improvement over MotifVoter in percentage of generated sets having higher accuracy (i.e with $nCC > 0.5$), and 380% improvement for sets having performance coefficient $nPC > 0.5$. For final motif results, our method achieves 80% improvement in final accuracy using correlation coefficient, and 93% improvement using performance coefficient over MotifVoter.

6 Future Work

The problem of computational motif finding is still standing in bioinformatics. Even with ensemble methods the accuracy is low. The upper-bound of ensemble methods is limited by the underlying stand-alone finders. Thus, using better stand-alone finders will raise the maximum possible senstivity for ensemble methods. Moreover, other objective functions can be suggested to enhance the accuracy. The idea of generating the profiles can also be used with other stand-alone or ensemble methods.

References

1. Qiu, P.: Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. Biochemical and Biophysical Research Communications 309(3), 495–501 (2003)
2. Wei, W., Yu, X.D.: Comparative analysis of regulatory motif discovery tools for transcription factor binding sites. Genomics Proteomics Bioinformatics 5(2), 131–142 (2007)
3. Das, M., Dai, H.K.: A survey of DNA motif finding algorithms. BMC Bioinformatics 8(suppl. 7) (2007)
4. Li, N., Tompa, M.: Analysis of computational approaches for motif discovery. Algorithms for Molecular Biology 1(1), 8–15 (2006)
5. Hu, J., Li, B., Kihara, D.: Limitations and potentials of current motif discovery algorithms. Nucleic Acids Res. 33(15), 4899–4913 (2005)

6. Stormo, G.D.: DNA binding sites: representation and discovery. Bioinformatics 16(1), 16–23 (2000)
7. Tompa, M., et al.: Assessing computational tools for the discovery of transcription factor binding sites. Nature Biotechnology 23, 137–144 (2005)
8. Wijaya, E., Yiu, S., Son, N.T., Kanagasabai, R., Sung, W.: Motifvoter: a novel ensemble method for fine-grained integration of generic motif finders. Bioinformatics 24, 2288–2295 (2008)
9. Chakravarty, A., Carlson, J.M., Khetani, R.S., Gross, R.H.: A novel ensemble learning method for de novo computational identification of DNA binding sites. BMC Bioinformatics 8, 249–263 (2007)
10. Che, D., Jensen, S., Cai, L., Liu, J.S.: BEST: Binding-site estimation suite of tools. Bioinformatics 21(12), 2909–2911 (2005)
11. Hu, J., Yang, Y.D., Kihara, D.: EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences. BMC Bioinformatics 7, 342–454 (2006)
12. Edgar, R.C.: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5), 1792–1797 (2004)
13. Bailey, T.L., Elkan, C.: Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning 21, 51–80 (1995)
14. Pavesi, G., Mereghetti, P., Mauri, G., Pesole, G.: Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res. 32(Web Server issue) (July 2004)
15. Liu, X., Brutlag, D.L., Liu, J.S.: Bioprospector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In: Pac. Symp. Biocomput., pp. 127–138 (2001)
16. Wijaya, E., Kanagasabai, R., Yiu, S.-M.M., Sung, W.-K.K.: Detection of generic spaced motifs using submotif pattern mining. Bioinformatics 23(12), 1476–1485 (2007)
17. Liu, X.S., Brutlag, D.L., Liu, J.S.: An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat. Biotechnol. 20(8), 835–839 (2002)
18. Workman, C.T., Stormo, G.D.: ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. In: Pac. Symp. Biocomput., pp. 467–478 (2000)
19. Thijs, G., et al.: A higher-order background model improves the detection of promoter regulatory elements by gibbs sampling. Bioinformatics 17(12), 1113–1122 (2001)
20. Eskin, E., Pevzner, P.A.: Finding composite regulatory patterns in DNA sequences. Bioinformatics 18(suppl. 1) (2002)
21. Huang, H.-D., Horng, J.-T., Sun1, Y.-M., Tsou, A.-P., Huang, S.-L.: Identifying transcriptional regulatory sites in the human genome using an integrated system. Nucleic Acids Res. 32(6), 1948–1956 (2004)
22. Ao, W., Gaudet, J., Kent, W.J., Muttumu, S., Mango, S.E.: Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. Science 305, 1743–1746 (2004)