Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W217–W220***
doi:10.1093/nar/gkm376

WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches

Katherine A. Romer[1], Guy-Richard Kayombya[1] and Ernest Fraenkel[2,3,*]

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA, [2]MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Cambridge, MA 02139, USA and [3]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received January 16, 2007; Revised April 27, 2007; Accepted April 30, 2007

ABSTRACT

WebMOTIFS provides a web interface that facilitates the discovery and analysis of DNA-sequence motifs. Several studies have shown that the accuracy of motif discovery can be significantly improved by using multiple *de novo* motif discovery programs and using randomized control calculations to identify the most significant motifs or by using Bayesian approaches. WebMOTIFS makes it easy to apply these strategies. Using a single submission form, users can run several motif discovery programs and score, cluster and visualize the results. In addition, the Bayesian motif discovery program THEME can be used to determine the class of transcription factors that is most likely to regulate a set of sequences. Input can be provided as a list of gene or probe identifiers. Used with the default settings, WebMOTIFS accurately identifies biologically relevant motifs from diverse data in several species. WebMOTIFS is freely available at http://fraenkel.mit.edu/webmotifs.

INTRODUCTION

One of the principal challenges in analyzing genomic sequences is to identify patterns, or 'motifs,' that represent functional elements (1). An important use of sequence motifs is to represent sites where transcriptional regulatory proteins bind and modulate expression of genes. There are many algorithms for finding such motifs. Given the same input data, these algorithms often discover different motifs, with no one algorithm consistently recovering all biologically significant patterns. Several studies have demonstrated that it is possible to achieve higher accuracy and sensitivity by combining the results from multiple motif discovery programs (2–4). However, this approach requires considerable computational overhead for managing data in a variety of formats and for clustering and scoring the large number of discovered motifs.

WebMOTIFS is a user-friendly web-based program that makes it easy to follow the current 'best practice' in motif discovery. A single web-based form facilitates data entry. WebMOTIFS automatically performs motif discovery on these data with several programs. The results from these programs are scored and integrated, and the most significant motifs are provided in an easily interpreted graphic output. WebMOTIFS also offers users the opportunity to analyze their data with THEME (5), a Bayesian motif discovery program that incorporates prior knowledge about the biochemical properties of many DNA-binding domains. The THEME approach is much more powerful than *de novo* motif discovery programs in mammalian species, and reveal both the motif and the class of DNA-binding protein that regulates a set of sequences.

PROGRAM DESCRIPTION AND ORGANIZATION

WebMOTIFS is designed to automate the identification of regulatory sequence motifs using multiple motif discovery algorithms. Users may provide gene names (RefSeq or yeast ORF names) or probe identifiers from one of the several microarray platforms for *Saccharomyces cerevisiae*, *Mus musculus* and *Homo sapiens*. WebMOTIFS automates all the remaining steps and sends the user an email with a link to the results, which are kept on our server for 30 days. In addition to graphical output, the data can be downloaded in text-based formats. An overview of the processing is shown in Figure 1.

Run with the default options, WebMOTIFS reports integrated results from four motif discovery programs: MEME (6), AlignACE (7), MDscan (8) and Weeder (9,10). WebMOTIFS retrieves sequences corresponding to

*To whom correspondence should be addressed. Tel: 617 258 8702; Fax: 617 395 2652; Email: fraenkel-admin@mit.edu

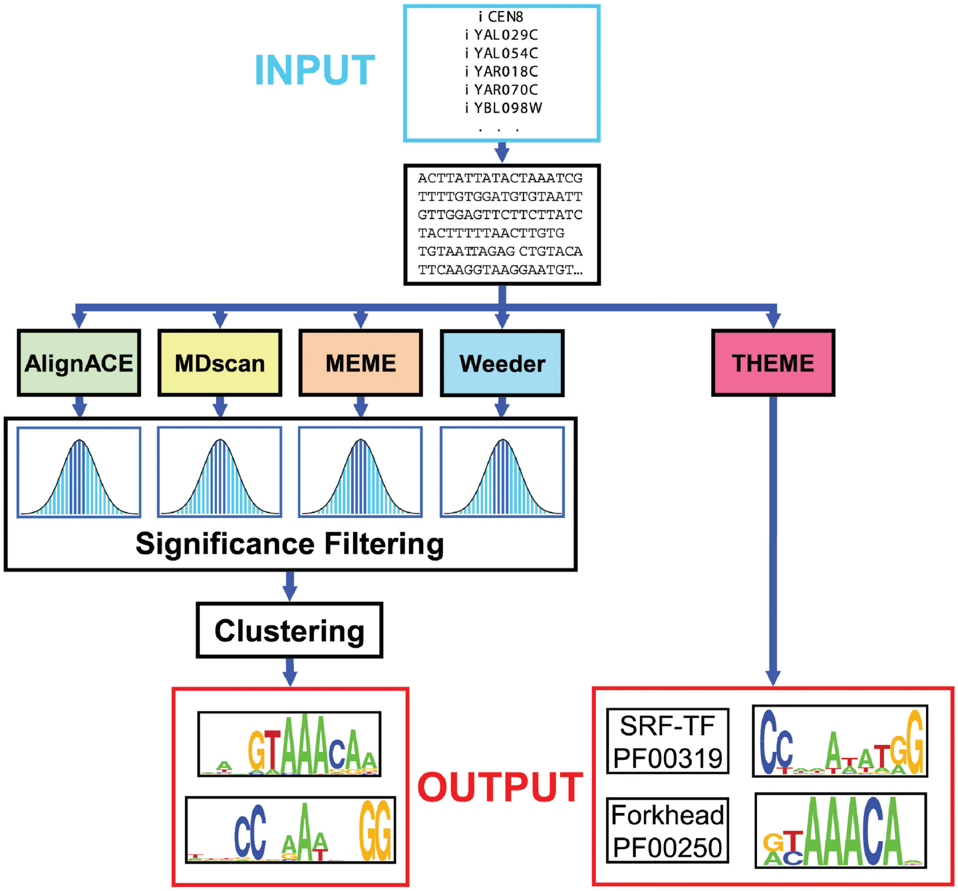


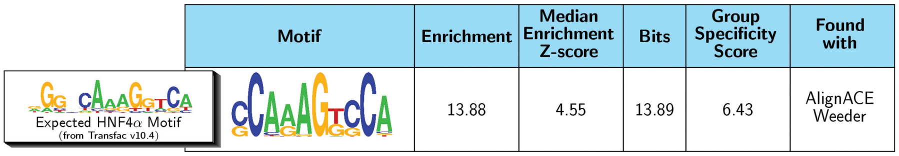
Figure 1. Overview of the WebMOTIFS analysis package. The user provides a set of gene or probe identifiers that WebMOTIFS converts to sequences. In the default mode, the sequences are analyzed by four motif discovery programs. The outputs of each program are tested for statistical significance and clustered to reveal a small set of likely motifs. In the advanced mode, the Bayesian motif discovery program THEME is also used to find motifs consistent with particular DNA-binding domain families. THEME includes its own principled scoring algorithms, eliminating the need for post-processing. The motifs discovered using ChIP-chip data for Fkh2 in high-H_2O_2 conditions are shown. The motifs discovered using the default settings match the known specificity of Fkh2 and of the interacting protein Mcm1. THEME also finds the Fkh2 and Mcm1 motifs, and in both cases correctly identifies the DNA-binding domain family. Note that many motif discovery algorithms are non-deterministic. Therefore, results may very among repeated runs of WebMOTIFS using the same input.

each input gene or probe name. If a gene name is provided or the probe is from a microarray designed for transcriptional profiling, the sequences are chosen based on the corresponding transcriptional start site. The sequence surrounding a probe is used for arrays designed for ChIP-chip. These sequences are automatically passed to the requested motif discovery programs without masking.

To combine the results from different motif discovery programs, WebMOTIFS objectively evaluates the significance of each motif. It compares the hypergeometric enrichment score for each motif to the distribution of scores for motifs found by the same program in sets of randomly selected promoters (2). Next, since motif discovery programs may discover very similar motifs, the significant motifs are clustered and a single representative motif for each cluster is computed. Currently, WebMOTIFS uses the clustering algorithm reported in Harbison *et al.* (2), although more sophisticated algorithms have been developed (11).

WebMOTIFS also provides the option of Bayesian motif discovery with THEME. The THEME algorithm searches for motifs consistent with proteins from specified DNA-binding domain families. The significance of each discovered motif is determined using cross-validation. THEME is particularly powerful in revealing motifs in mammalian promoters that are often missed by other methods (5). The user can specify which DNA-binding domains are expected to be involved in the regulation of the input sequences or test all the available DNA-binding domain families.

WebMOTIFS offers a unique combination of features. First, WebMOTIFS is completely web-based, with all jobs running on our server, so it can work on any operating system. Although many motif discovery programs have web interfaces, it is difficult to merge the results of these programs. Other available tools for running multiple motif discovery programs, such as BEST (the Binding-site Estimation Suite of Tools) (12) and TAMO (Tools for Analysis of Motifs, the software package on which WebMOTIFS is based) (13), are downloadable software packages. Second, WebMOTIFS analyzes the results of motif discovery automatically with default values that


Expected HNF4α Motif
(from Transfac v10.4)

Motif	Enrichment	Median Enrichment Z-score	Bits	Group Specificity Score	Found with
	13.88	4.55	13.89	6.43	AlignACE Weeder

Figure 2. Sample output from WebMOTIFS applied to sequences from human ChIP-chip experiments. The expected motif for the Hnf4α protein is discovered.

typically produce useful results. WebMOTIFS has few adjustable parameters, which makes it less flexible than TAMO and BEST, but also makes it easier to use and less vulnerable to user error. Third, WebMOTIFS facilitates both input and output. It automatically extracts sequences corresponding to gene and probe names, clusters the discovered motifs and produces sequence logos representing the results. Clustering and visualizing the results of motif discovery with multiple programs helps make sense of the large number of discovered motifs, providing a quick summary of the results.

Example

We evaluated WebMOTIFS by analyzing previously reported genome-wide chromatin-immunoprecipitation experiments in *S. cerevisiae* (2), taking the list of bound genes for each transcription factor in each condition as input to WebMOTIFS. We compared the results from WebMOTIFS with the motifs previously reported by MacIsaac *et al.* (14). Run with the default settings (without Bayesian motif discovery), WebMOTIFS discovers the correct motif in 51 out of the 64 transcription factors. These results are particularly striking, because, in contrast to MacIsaac *et al.* (14), the programs currently incorporated in WebMOTIFS do not take advantage of information from evolutionary conservation.

The significance filtering and clustering steps provided by WebMOTIFS reveal the most statistically significant motifs, which are frequently also the most biologically relevant. For example, applying WebMOTIFS to the genes bound by the transcription factor Fkh2 in high-H_2O_2 conditions produced 163 motifs. Significance filtering eliminated most of these results, and clustering grouped the remaining 32 motifs into two clusters. The highest-ranked cluster is a good match to the known specificity of Fkh2, and the second cluster is a good match to the known specificity of Mcm1. The Mcm1 motif is the most significant motif identified by THEME, which correctly attributes it to the SRF-TF family (15). The next most significant motif matches the known specificity and the DNA-binding domain family of Fkh2 (Figure 1). Fkh2 and Mcm1 have previously been reported to bind cooperatively to a number of promoters in *S. cerevisiae* (16,17).

We also tested WebMOTIFS on chromatin-immunoprecipitation data from mouse and human (5), taking up to 200 bound probes for each transcription factor as input. In general, motif discovery is more difficult on human and mouse data than on yeast sequences (4). Nevertheless, WebMOTIFS is often able to identify the correct motif as significant when run with the default settings. For instance, applying WebMOTIFS to binding data for Hnf4α in human hepatocytes produced 640 motifs. After significance filtering and clustering, the only remaining motif matches the known specificity of this protein, as reported in TRANSFAC (18) (Figure 2). Running WebMOTIFS with the THEME option also reveals the Hnf4α motif, which is correctly attributed to the nuclear hormone receptor family (19).

CONCLUSION

WebMOTIFS is an easy-to-use motif discovery tool that provides an interface to integrated motif discovery and automatically does clustering, scoring, and visualization of the results. WebMOTIFS automatically selects significant motifs, taking the best results from different motif discovery programs and combining them in an intuitive output format. It also offers the opportunity to incorporate prior knowledge of transcription factor structure using Bayesian motif discovery. Run with the default settings, it provides accurate results on a variety of data.

ACKNOWLEDGEMENTS

We thank Kenzie MacIsaac and other members of the Fraenkel laboratory for helpful discussions. We are grateful to the authors of AlignACE, MDscan, MEME and Weeder for permission to include their programs. Supported by funding from the Whitaker Foundation, the John S. Reed (1961) Fund and the MIT Undergraduate Research Opportunities Program. Funding to pay the Open Access publication charges for this article was provided by the Whitaker Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. D'Haeseleer,P. (2006) What are DNA sequence motifs? *Nat. Biotechnol.*, **24**, 423–425.
2. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
3. MacIsaac,K.D. and Fraenkel,E. (2006) Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput. Biol.*, **2**, e36.
4. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
5. Macisaac,K.D., Gordon,D.B., Nekludova,L., Odom,D.T., Schreiber,J., Gifford,D.K., Young,R.A. and Fraenkel,E. (2006) A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data. *Bioinformatics*, **22**, 423–429.

6. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
7. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.*, **296**, 1205–1214.
8. Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.
9. Pavesi,G., Mereghetti,P., Zambelli,F., Stefani,M., Mauri,G. and Pesole,G. (2006) MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. *Nucleic Acids Res.*, **34**, W566–W570.
10. Pavesi,G., Mereghetti,P., Mauri,G. and Pesole,G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.*, **32**, W199–W203.
11. Jensen,S.T., Shen,L. and Liu,J.S. (2005) Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes. *Bioinformatics*, **21**, 3832–3839.
12. Che,D., Jensen,S., Cai,L. and Liu,J.S. (2005) BEST: binding-site estimation suite of tools. *Bioinformatics*, **21**, 2909–2911.
13. Gordon,D.B., Nekludova,L., McCallum,S. and Fraenkel,E. (2005) TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. *Bioinformatics*, **21**, 3164–3165.
14. MacIsaac,K.D., Wang,T., Gordon,D.B., Gifford,D.K., Stormo,G.D. and Fraenkel,E. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 113.
15. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
16. Boros,J., Lim,F.L., Darieva,Z., Pic-Taylor,A., Harman,R., Morgan,B.A. and Sharrocks,A.D. (2003) Molecular determinants of the cell-cycle regulated Mcm1p-Fkh2p transcription factor complex. *Nucleic Acids Res.*, **31**, 2279–2288.
17. Hollenhorst,P.C., Pietz,G. and Fox,C.A. (2001) Mechanisms controlling differential promoter-occupancy by the yeast forkhead proteins Fkh1p and Fkh2p: implications for regulating the cell cycle and differentiation. *Genes Dev*, **15**, 2445–2456.
18. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
19. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.