

THE UNIVERSITY OF WARWICK

Original citation:

Brown, P., Baxter, L., Hickman, R. D. G. (Richard D. G.), Beynon, J., Moore, J. D. and Ott, S. (2013) MEME-LaB : motif analysis in clusters. *Bioinformatics* . ISSN 1367-4803

Permanent WRAP url:

<http://wrap.warwick.ac.uk/54629>

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes the work of researchers of the University of Warwick available open access under the following conditions.

This article is made available under the Attribution 3.0 Unported (CC BY 3.0) license and may be reused according to the conditions of the license. For more details see:

<http://creativecommons.org/licenses/by/3.0/>

A note on versions:

The version presented in WRAP is the published version, or, version of record, and may be cited as it appears here.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

warwick**publications**wrap

highlight your research

<http://go.warwick.ac.uk/lib-publications>

Sequence Analysis

MEME-LaB: motif analysis in clustersPaul Brown^{1a*}, Laura Baxter^{1a}, Richard Hickman^{1b}, Jim Beynon², Jonathan D. Moore¹ and Sascha Ott¹¹Warwick Systems Biology Centre, University of Warwick, Coventry, CV4 7AL, United Kingdom²School of Life Sciences, University of Warwick, Coventry, CV4 7AL, United Kingdom^a These authors contributed equally to this work and should be considered joint first authors^b Current address: Department of Biology, Faculty of Science, Utrecht University, PO Box 800.56, 3508 TB Utrecht, The Netherlands

Associate Editor: Prof. Martin Bishop

ABSTRACT

Summary: Genome-wide expression analysis can result in large numbers of clusters of co-expressed genes. While there are tools for *ab initio* discovery of transcription factor binding sites, most do not provide a quick and easy way to study large numbers of clusters. To address this, we introduce a web-tool called MEME-LaB. The tool wraps MEME (an *ab initio* motif finder), providing an interface for users to input multiple gene clusters, retrieve promoter sequences, run motif finding, and then easily browse and condense the results, facilitating better interpretation of the results from large-scale datasets.

Availability: MEME-LaB is freely accessible at: <http://wsbc.warwick.ac.uk/wsbcToolsWebpage/>

Contact: p.e.brown@warwick.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online

1 INTRODUCTION

Analyzing microarray expression data using cluster analysis is a common and frequently performed task in functional genomics. Typically, a large number of clusters are produced, each containing a large number of genes (e.g. 50 clusters of 200 genes). Each cluster is predicted to contain a set of genes that are co-expressed, and as such would be expected to share common regulatory features, such as transcription factor binding sites (TFBSs). There are established methods and tools for predicting known TFBSs (such as Athena, O'Connor *et al.*, 2005), but *ab initio* motif discovery remains an important aspect to consider. Several tools exist to perform this task and MEME in particular is a well-recognized suite for motif discovery (Bailey *et al.*, 2006), but the MEME web-suite offers limited usage on large numbers of clusters and subsequent navigation and post-processing of the results. Machanick and Bailey (2012) provide a web-tool (MEME-ChIP) that is specifically tailored towards ChIP-seq data, providing a useful expansion in MEME's functionality.

Here we describe a web-tool called MEME-LaB (MEME Launcher and Browser), which wraps the MEME tool in ways that are ideally suited to the task of *ab initio* motif finding in co-expressed gene clusters: (i) Users can input multiple gene clusters at once. (ii)

Promoter sequences are automatically retrieved from a local database or from a user-specified file. (iii) MEME is run on all clusters simultaneously and the results are presented in a condensed and navigable format. (iv) Identified motifs are compared for similarity to known TFBS motifs.

2 IMPLEMENTATION**2.1 General workflow**

The MEME-LaB web service is designed to simplify the task of identifying putative TFBSs in the promoters of co-expressed gene clusters, and provides an easy way to navigate through and filter the results.

2.2 Input

Users are not required to register for the service, and are automatically logged in as a guest user. Gene clusters are uploaded as a simple tab-separated file consisting of two columns: the first for numbers identifying the clusters, and the second for gene ids. For the Arabidopsis genome, sequences will be automatically retrieved from a local database; for all other genomes, a second file containing promoter sequences in FASTA format is uploaded by the user.

2.3 Processing

Uploaded files are verified for validity, and the user warned of any detected errors in the input. Users specify the minimum and maximum length of promoter regions to search, between 50 and 1000 nucleotides in length, and optionally to stop at a neighbouring gene if there is one present in this region. Users select the number of motifs to find per cluster, within a minimum and maximum motif length (6-20 nucleotides). Details of comparisons to known motifs from JASPAR (Bryne *et al.*, 2008) and PLACE (Higo *et al.*, 1999) are provided in the output (users with a valid login for TRANSFAC can also compare motifs against TRANSFAC motifs on our server). An email is sent to the address provided notifying when the job is complete and results ready for retrieval.

2.4 Output

Results are provided as interactive html pages, and can also be downloaded. For each cluster, the specified number of motifs are identified using the MEME algorithm, and displayed as motif logos. Additional information for each motif is displayed, including

*To whom correspondence should be addressed.

its distribution amongst the input set, positional bias, strand bias, and similarity to known motifs.

3 EXAMPLE RESULTS

We demonstrate the functionality offered by MEME-LaB, using co-expression clusters derived from a time-course microarray experiment of Arabidopsis responses to infection with *Botrytis cinerea* (Windram *et al.*, 2012). (Input files and the complete results are available as Supplementary Data 1). The usefulness of the tool is demonstrated by being able to easily browse motifs for all clusters on a single web-page and to reduce a large set of motifs to the most informative results based on motif properties (Figure 1A and B). Additional information on each motif's positional distribution amongst the input set is provided (Figure 1C). For each *ab initio*

motif predicted, up to 5 of the most similar known motifs are listed, with a distance measure, and additional information and links are provided (Figure 1D). In the example result view (Figure 1), filtering to show only motifs that occur in 25% or more of sequences in a cluster, occurring in 20 or more sites, and have an information content greater than 10 has resulted in 21 of the possible 220 motifs being displayed. MEME identified motifs similar to I-box (Figure 1A, top) and G-box (Figure 1A, bottom), which are consistent with previous findings, but also a third motif (Figure 1A, middle) that is not closely similar to any known motif but is present in all the sequences in cluster 4. The MEME-LaB service makes existing functionality more widely and more easily applicable, enabling the identification of significant motifs in large co-expression cluster datasets.

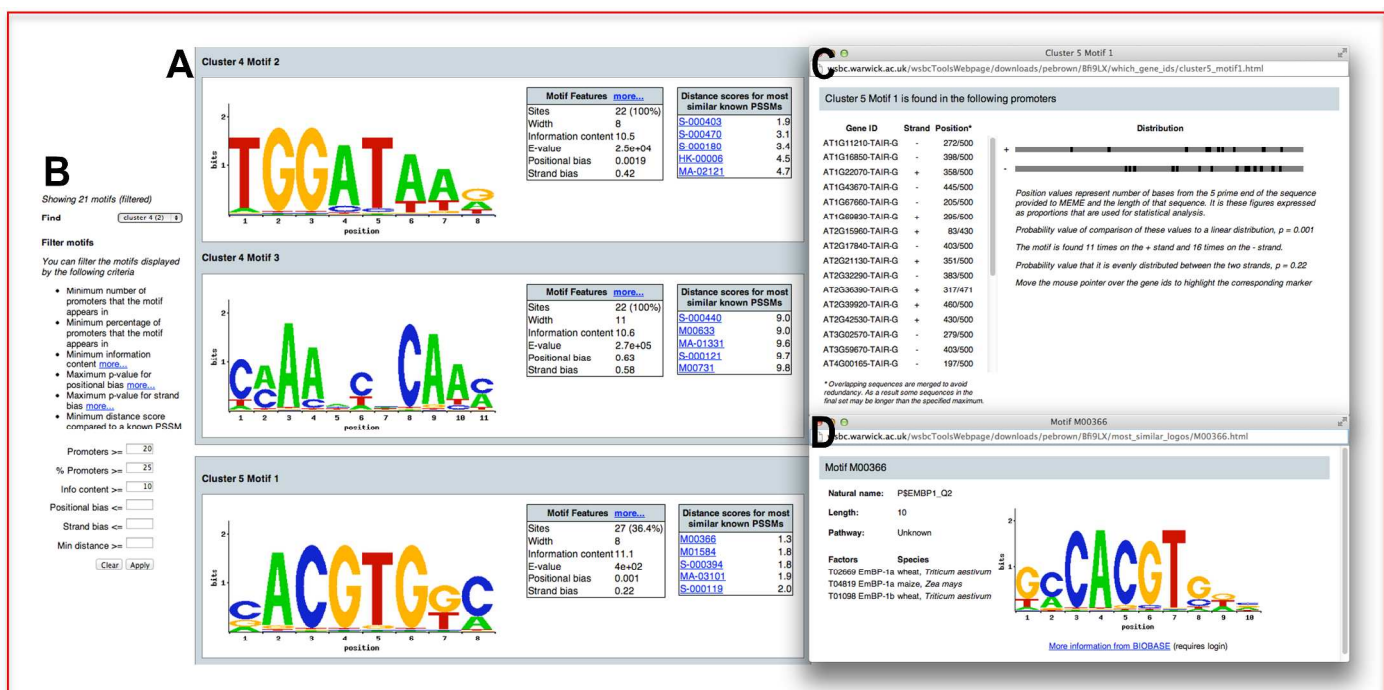


Figure 1. Screenshot of a typical MEME-LaB result. Motifs are displayed as motif logos, and information about each motif is shown (A). Results can be filtered on motif properties (B). Additional information on positional distribution (C) and properties of similar motifs (D) are accessed in pop-up windows.

ACKNOWLEDGEMENTS

Funding: This work was supported by the Biotechnology and Biological Sciences Research Council (BBSRC), [BB/F005806/1 to L.B., J.D.M., J.B., and S.O.]; and the Engineering and Physical Sciences Research Council/BBSRC-funded Warwick Systems Biology Doctoral Training Centre [RH].

Conflict of Interest: none declared.

REFERENCES

Bailey, T.L., *et al.* (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**: W369-W373.

Bryne, J.C., *et al.* (2008). JASPAR, the open access database of transcription factor-binding profiles: New content and tools in the 2008 update. *Nucleic Acids Res.*, **36** (Database issue): D102–D106.

Higo, K., *et al.* (1999). Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.* **27**: 297–300.

Machanic, P. and Bailey, T.L. (2012). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27** (12): 1696–1697.

O'Connor, T.R., *et al.* (2005). Athena: a resource for rapid visualization and systematic analysis of Arabidopsis promoter sequences. *Bioinformatics* **21** (24): 4411–4413.

Windram, O., *et al.* (2012). Arabidopsis defense against *Botrytis cinerea*: chronology and regulation deciphered by high-resolution temporal transcriptomic analysis. *Plant Cell*. DOI: <http://dx.doi.org/10.1105/tpc.112.102046>.