MEME: Discovering and analyzing DNA and protein sequence motifs

Timothy L. Bailey [a,*], Nadya Williams [b],

Chris Misleh [b], and Wilfred W. Li [b]

[a] *Institute of Molecular Bioscience, The University of Queensland, QLD 4072, St Lucia, Australia.*

[b] *SDSC, UCSD, La Jolla, California, USA.*

Abstract

MEME (Multiple EM for Motif Elicitation) is one of the most widely used tools for searching for novel "signals" in sets of biological sequences. Applications include the discovery of new transcription factor binding sites (TFBSs) and protein domains. MEME works by searching for repeated, ungapped sequence patterns that occur in the DNA or protein sequences provided by the user. Users can perform MEME searches via the web server hosted by the National Biomedical Computation Resource (`http://meme.nbcr.net`) and several mirror sites. Via the same web server, users can also access the Motif Alignment and Search Tool (MAST) to search sequence databases for matches to motifs encoded in several popular formats. By clicking on buttons in the MEME output, users can compare the motifs discovered in their input sequences with databases of known motifs, search sequence databases for matches to the motifs, and display the motifs in various formats. This article describes the freely accessible web server and its architecture, and discusses ways to use MEME effectively to find new sequence patterns in biological sequences and analyze their significance.

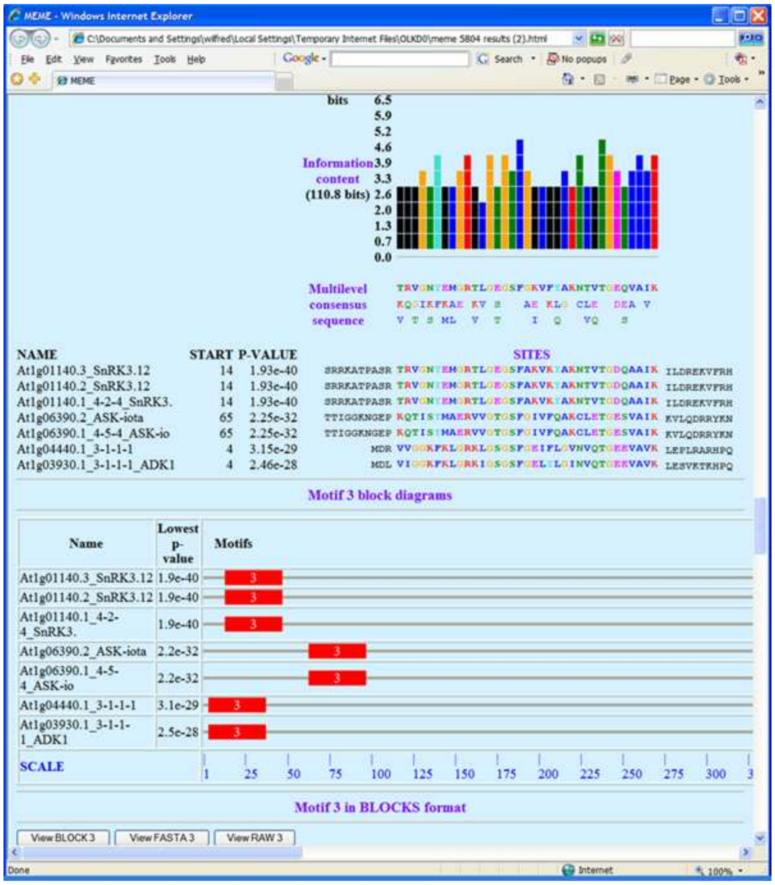
* Corresponding author.
Email address: `t.bailey@imb.uq.edu.au` (Timothy L. Bailey).

Introduction

The purpose of MEME (rhymes with "team") [1,2] is to allow users to discover signals (called "motifs") in DNA or protein sequences. The user of MEME inputs a set of sequences believed to share some (unknown) sequence signal(s). For example, some or all of a set of promoters from co-expressed and/or orthologous genes may contain binding sites (the "signal") for the same transcription factor [3]. Similarly, a set of proteins that interact with a single host protein may do so via similar domains (the "signal") [4]. Both types of sequence signals can often be represented as motifs–ungapped, approximate sequence patterns. Using a process akin to gapless, local, multiple sequence alignment, MEME searches for statistically significant motifs in the input sequence set. In this way, MEME can discover the binding sites for the shared transcription factor in the set of promoters or the common protein-protein binding domains in the set of proteins. MEME can also be used to discover motifs describing many other types of DNA or protein signals besides transcription factor binding sites and protein-protein interaction domains.

To use MEME via the web site, the user provides a set of sequences in FASTA format by either uploading a file or by cut-and-paste. The only other required input is an email address where the results will be sent. (A planned future version will remove this requirement by providing temporary storage of the results on the web server for a preset period of time.) By default, MEME looks for up to three motifs, each of which may be present in some or all of the input sequences. MEME chooses the width and number of occurrences of each motif automatically in order to minimize the "E-value" of the motif–the probability of finding an equally well-conserved pattern in random sequences. By default, only motif widths between 6 and 50 are considered, but the user may change this as well as several other aspects of the search for motifs.

The MEME output is HTML and shows the motifs as local multiple alignments of (subsets of) the input sequences, as well as in several other formats (Fig. 1). “Block diagrams” show the relative positions of the motifs in each of the input sequences. Buttons on the MEME HTML output allow one or all of the motifs to be forwarded for analysis by other web-based programs. Clicking on a button allows all of the motifs to be sent to the MAST web server where various sequence databases (or uploaded sequences) can be searched for sequences matching the motifs. This is useful in cases, for example, where the user would like to find whether the motif of interest is also present in other genes or genomes.



```
Multilevel   TRVGNYEMGRTLGEGSFGKVFYAKNTVTGEQVAIK
consensus    KQSIKFKAE KV S  AE KLG CLE  DEA V
sequence     V T S ML  V  T   I  Q  VQ   S

NAME                     START P-VALUE                         SITES
At1g01140.3_SnRK3.12        14  1.93e-40  SRRKATPASR TRVGNYEMGRTLGEGSFAKVKYAKNTVTGDQAAIK ILDREKVFRH
At1g01140.2_SnRK3.12        14  1.93e-40  SRRKATPASR TRVGNYEMGRTLGEGSFAKVKYAKNTVTGDQAAIK ILDREKVFRH
At1g01140.1_4-2-4_SnRK3.    14  1.93e-40  SRRKATPASR TRVGNYEMGRTLGEGSFAKVKYAKNTVTGDQAAIK ILDREKVFRH
At1g06390.2_ASK-iota        65  2.25e-32  TTIGGKNGEP KQTISYMAERVVGTGSFGIVFQAKCLETGESVAIK KVLQDRRYKN
At1g06390.1_4-5-4_ASK-io    65  2.25e-32  TTIGGKNGEP KQTISYMAERVVGTGSFGIVFQAKCLETGESVAIK KVLQDRRYKN
At1g04440.1_3-1-1-1          4  3.15e-29         MDR VVGGKFKLGRKLGSGSFGEIFLGVNVQTGEEVAVK LEFLRARHPQ
At1g03930.1_3-1-1-1_ADK1     4  2.46e-28         MDL VIGGRFKLGRKIGSGSFGELYLGINVQTGEEVAVK LESVRTKHPQ
```

Motif 3 block diagrams

Name	Lowest p-value	Motifs
At1g01140.3_SnRK3.12	1.9e-40	3
At1g01140.2_SnRK3.12	1.9e-40	3
At1g01140.1_4-2-4_SnRK3.	1.9e-40	3
At1g06390.2_ASK-iota	2.2e-32	3
At1g06390.1_4-5-4_ASK-io	2.2e-32	3
At1g04440.1_3-1-1-1	3.1e-29	3
At1g03930.1_3-1-1-1_ADK1	2.5e-28	3

Motif 3 in BLOCKS format

Fig. 1. **Sample MEME output.** This portion of a MEME HTML output form shows a protein motif that MEME has discovered in the input sequences. The sites identified as belonging to the motif are indicated, and above them is the “consensus” of the motif and a color-coded bar graph showing the conservation of each position in the motif. Some of the hyperlinked buttons that allow the motif to be viewed and analyzed in other ways can be seen at the bottom of the screen shot.

MAST is a web-based tool that can be used to search for sequences that match one or more motifs. It can be used to look for sequences that contain motifs found by MEME, by other motif discovery tools, or that are taken from a motif database. The MAST web site, reached via the same URL as the MEME web site, provides numerous nucleotide and protein databases for searching. MAST queries may contain any number of motifs, and it scores each sequence in the selected database using all of the motifs. In the first example above, MAST can search DNA sequences for matches to the putative transcription factor binding site motifs found by MEME in a set of promoter sequences. MAST can search for matches in protein sequences to the putative protein-protein interaction motifs found in the second MEME example.

Users of MEME via the web site or locally installed versions are asked to cite this article as well as the primary reference for MEME: Timothy L. Bailey and Charles Elkan, “Fitting a mixture model by expectation maximization to discover motifs in biopolymers”, *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28-36, AAAI Press, Menlo Park, California, 1994. Users of MAST are asked to cite this article and: Timothy L. Bailey and Michael Gribskov, “Combining evidence using p-values: application to sequence homology searches”, Bioinformatics, 14(48-54), 1998.

Motif Discovery Strategies

Motif discovery can be viewed as a “needle in a haystack” problem. The motif discovery algorithm is looking for a set of similar short sequences (the needle) in a set of much longer sequences (the haystack). The problem is easier when the motif instances are long and very similar to each other. It gets much harder when the motif instances are short and/or degenerate, or the input sequences are very long.

Discovering transcription factor binding site motifs in a set of DNA sequences (e.g., genomic regions upstream of genes) is a difficult task due to the tendency of binding sites to be short and degenerate, and due to the fact that promoter regions are often difficult to identify precisely. The problem tends to be worse in eukaryotes than in prokaryotes and yeast because eukaryotic TFBS tend to be shorter and more variable [5].

To successfully discover TFBS motifs with MEME, it is necessary to choose and prepare the input sequences carefully. Candidate sequences can be the promoters of genes believed to be co-regulated based on evidence from expression micro-array experiments, or sequences appearing to bind to a transcription factor based on chromatin immunoprecipitation experiments. The sequences should be as short as possible and contain as few "noise" sequences (sequences not containing any motif) as possible. Ideally, the sequences should be less than 1000 base-pairs long [6]. Including more than 40 motif-containing sequences generally does not improve TFBS motif discovery with MEME and similar algorithms [7]. If the sequences contain low-information segments that do not contain motifs of interest, it can be helpful to remove them using the DUST program (R. L. Tatusov and D. J. Lipman, unpublished NCBI/Toolkit), which is available for downloading at `http://blast.wustl.edu/pub/dust/`. Repetitive DNA elements should also be removed from the sequences input to MEME using the RepeatMasker program (A. Smit, R. Hubley and P. Green, unpublished data), which can be accessed via the web at URL `http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker`.

It should be noted that MEME is not suited to whole-genome TFBS motif discovery. Because of their shortness and degeneracy, TFBS motifs become statistically "invisible" in the context a whole genome. The sensitivity of the search for TFBS motifs can be improved by using a "higher-order background sequence model," but this option is only available currently when users download the MEME source code and install it locally. Instructions for doing this are available at the MEME web site (`http://meme.nbcr.net/meme/website/meme-download.html`) by clicking on "View MEME man page", see the documentation for the "-bfile" switch there.

Protein motifs are generally easier to discover due to the length of the protein alphabet and the chemical similarity among groups of amino acids. This allows shorter motifs to be more statistically significant and makes it easier to distinguish functional motifs from statistical artifacts. To use MEME to discover protein motifs, the same basic guidelines apply as with DNA motifs–keep the sequences as short as possible and include as few sequences not likely to contain the motif as possible in the input to MEME. Low complexity regions can be removed from the protein input sequences using the SEG program [8].

Analyzing Motifs Using the MEME Output Hyperlinks

The MEME HTML output contains buttons making it easy to analyze the motifs it discovers. By clicking on the button labeled "Compare PSPM to known motifs in JASPAR database" following each motif, the DNA motif can be compared to each of the motifs in the JASPAR database [9] of known TFBS motifs. Similarly, protein motifs may be compared with protein motifs in the BLOCKS database of protein motifs [10] by clicking on the "submit BLOCK" button following each motif on the MEME form. This takes the user to the "BLOCKS server" where clicking on "LAMA" will compare the motif with those in the BLOCKS database. The BLOCKS server also allows users to display protein motifs in many different ways including as LOGOS [11] or as phylogenetic trees by clicking on the corresponding buttons on the BLOCKS server form. By clicking on one of the file output formats under Logos, one is able to obtain a LOGOS diagram like that shown in Fig. 2.

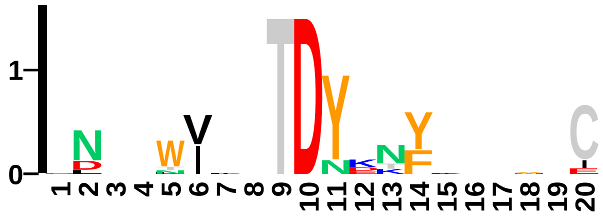


Fig. 2. **LOGO of protein motif.** LOGOS are a visualization tool for motifs. The height of a letter indicates its relative frequency at the given position (X-axis) in the motif.

To search sequences for matches to the motifs found by MEME, users can click on the "MAST" button at the top of the MEME output form. That will take the user to the MAST web site where they can select the database to search. Since MAST is sequence-oriented, TFBS motifs should be only be used to search promoter regions. These are listed in the MAST database pull-down menu as "Upstream Sequence Databases". Currently, only a few organisms are supported. However, users can upload there own database of promoter sequences for searching by MAST. Protein motifs can be used to search any of the sequence databases provided by the MAST web site since MAST can search either protein or nucleotide databases with protein motifs. The MAST databases are updated weekly.

Web Server and User Support

As of MEME version 3.5, the configuration and installation of MEME (including the web server) is significantly simplified by using Autoconf (`http://www.gnu.org/software/autoconf/autoconf.html) and Automake` (`http://www.gnu.org/software/automake/automake.html`) from the GNU Build System. An installation session for MEME and MAST web server may be as simple as follows:

```
cd meme_3.5.2
./configure --prefix=$HOME/meme --with-url=http://www.nbcr.net/meme --enable-web
make
make test
make install
```

Supported platforms now include Linux, Solaris, MacOS X, Cygwin and Irix.

The MEME web server hosted by NBCR is queried by about 800 different users (based on unique email addresses) each month. Usage has been growing steadily since the service was first introduced in 1996. Fig. 3 shows usage growth at the NBCR server since 2000.

To meet the growing user demand and take advantage of the emerging grid computing resources [12], we have made MEME available for installation on Linux clusters using either the RPM package manager or Rocks. The RPM package manager is a tool for managing software installation on computers running many versions of the Linux operating system. Rocks (`http://www.rocksclusters.org`) is a highly customized toolkit for computational biologists and engineers to build and maintain Linux clusters. The current NBCR MEME web server cluster is built with the MEME roll for Rocks and requires minimal maintenance effort.

MEME and MAST can be downloaded and installed free of charge by academic users via the web site: `http://meme.nbcr.net/meme/website/meme-download.html`. Approximately 300 users download the MEME/MAST software each month. The MEME support team offers assistance to the MEME and MAST user community through the forum

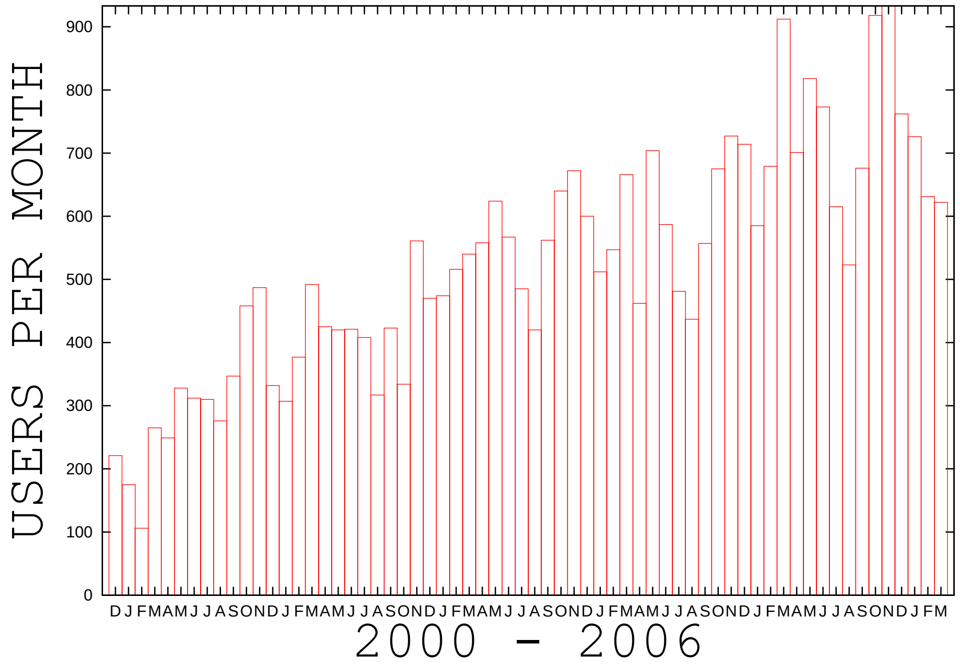


Fig. 3. **Usage of MEME at the NBCR web server.** The plot shows the number of different users submitting jobs to the NBCR MEME web server each month since December, 2000. Usage figures for March, 2006 include only to up to March 20.

(http://nbcr.net/forum/viewforum.php?f=5) or the mailing list (meme@nbcr.net). Institutes interested in setting up MEME mirror sites are encouraged to contact us for any assistance.

Future Directions

To increase the sensitivity of MEME searches, we will add an option to the web server to let the user upload a background sequence model to MEME. We hope to add algorithms for removing low-complexity regions (SEG and DUST) and repeated elements (RepeatMasker) to the MEME web site as a convenience to users. These services will also be exposed as web services and integrated using workflow tools developed by NBCR.

We also plan to add buttons to the MEME output to allow TFBS motifs to be used in searching for cis-regulatory modules via algorithms such as MCAST [13]. MCAST will be configured to be able to search the same DNA databases as MAST. In conjunction with this, we will add databases of upstream sequences for many additional organisms to the MAST/MCAST web sites to facilitate the analysis of TFBS motifs discovered by MEME.

NBCR has developed a set of tools built on top of open source software that allows bioinformatics applications to be deployed as Web Services easily (Sriram Krishnan, Brent Stearn, Karan Bhatia, Wilfred W. Li, and Peter Arzberger (2006) "Opal: Wrapping Scientific Grid Applications as Web Services", ICWS, Chicago, submitted) and leverage the Cyberinfrastructure components transparently [12]. A prototype has been deployed using MEME as a scientific driver [14] that offers a user with a dynamic pool of distributed compute resource, workflow management console and a friendly user interface. This portal will be deployed to the production web server in the future.

Acknowledgements

The authors acknowledge NBCR award from NCRR, NIH P41 RR08605 for support of the MEME and MAST web site. TLB acknowledges support from NIH, R01 RR021692-01 for support of continuing development of MEME and related sequence analysis tools. TLB also acknowledges the ARC Centre for Bioinformatics (ACB) (ARC CE0348221) for infrastructure support for the MEME mirror site at the ACB.

References

[1] Bailey, T. L. and Elkan, C. (1995) Unsupervised learning of multiple motifs in biopolymers using EM. *Machine Learning,* **21**, 51–80.

[2] Bailey, T. L. and Elkan, C. (1995) The value of prior knowledge in discovering motifs with MEME. In Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T., and Wodak, S., (eds.), *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, California: AAAI Press pp. 21–29.

[3] Lyons, T. J., Gasch, A. P., Gaither, L. A., Botstein, D., Brown, P. O., and Eide, D. J. (2000) Genome-wide characterization of the zap1p zinc-responsive regulon in yeast. *Proceedings of the National Academy of Sciences USA,* **97**, 7957–7962.

[4] Fang, J., Haasl, R. J., Dong, Y., and Lushington, G. H. (2005) Discover protein sequence signatures from protein-protein interaction data. *BMC Bioinformatics,* **6**(277), 1–8.

[5] Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Ye, C., and Zhu, Z. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology,* **23**, 137–147.

[6] Pevzner, P. A. and Sze, S. H. (2000) Combinatorial approaches to finding subtle signals in dna sequences. In Bourne, P., Gribskov, M., Altman, R., Jensen, N., Lengauer, D. H. T., Mitchell, J., Scheeff, E., Smith, C., Strande, S., and Weissig, H., (eds.), *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, California: AAAI Press pp. 269–278.

[7] Hu, J., Li, B., and Kihara, D. (2005) Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Research,* **33**, 4899–4913.

[8] Wootton, J. C. and Federhen, S. (1966) Analysis of compositionally biased regions in sequence databases. *Methods in Enzymology,* **266**, 554–571.

[9] Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research,* **32**, D91–D94.

[10] Henikoff, J. G., Pietrokovski, S., and Henikoff, S. (1997) Recent enhancements to the blocks database servers. *Nucleic Acids Research,* **25**, 222–225.

[11] Schneider, T. D. and Stephens, R. M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research,* **18**, 6097–6100.

[12] Foster, I. and Kesselman, C. (2004) The Grid 2: Blueprint for a New Computing Infrastructure, 2 ed., Morgan Kaufmann Publishers, Inc., San Francisco.

[13] Bailey, T. L. and Noble, W. S. (2003) Searching for statistically significant regulatory modules. *Bioinformatics,* **19 Suppl 2**, II16–II25.

[14] Li, W. W., Krishnan, S., Mueller, K., Misleh, C., and Arzberger, P. (2006) Building cyberinfrastructure for bioinformatics using service oriented architecture. In Sung, F. L. B., Abramson, D., Cai, W., Graupner, S., Jin, H., and Sloot, P., (eds.), *2006 IEEE International Symposium on Cluster Computing and the Grid*, USA: IEEE Press (in press).