`fdrMotif`: Identifying *cis*-elements by an EM Algorithm Coupled with False Discovery Rate Control

http://www.niehs.nih.gov/research/resources/software/fdrMotif/index.cfm

Leping Li[1], Robert L Bass[2], and Yu Liang[3]

[1]Biostatistics Branch and [2]Computational Biology Facility, National Institute of Environmental
Health Sciences, NIH, DHHS, Research Triangle Park, North Carolina 27709, USA
li3@niehs.nih.gov
bass1@niehs.nih.gov

[3]SAS Institute, 100 SAS Campus Drive, Cary, North Carolina 27513, USA
Yu.Liang@sas.com

Usage: `fdrMotif -fm inputSeqFile –fpwm initialPWMFile(s) [optional arguments]`

`-fm`	`inputSeqFile`	File containing sequences to be searched in FASTA format, e.g., ChIP file
`-fpwm`	`startPWMFile`	File(s) containing initial PWM

optional arguments:

`-fdrb`	`FDRbound`	Upper bound of False Discovery Rate (default: 0.1)
`-g`	`maxIteration`	Maximum number of iteration (default: 100)
`-nbsets`	`numBackgSets`	Number of sets of background sequences (default: 10)
`-Cmax`	`maxNumSites`	Maximum number of occurrences of a motif in any given sequence in input data (default: 10)
`-fb`	`backgSeqFile`	File containing sequences other than the input/motif sequences. Sequences in this file are either used directly as the background sequences or used to estimate a Markov model from which background sequences are generated (default: generate Markov background sequences from sequences in inputSeqFile with –fm argument)
`-Imb`	`indicator`	Generate Markov background sequences or not (1 – yes, 0 – no) (default: 1)
`-fo`	`outputFIle`	Name of the output file (default: fdrMotif.txt)
`-fM`	`motifSummary`	Name of output file receiving motif scoring summary (default: fdrMotifSummary.txt)

`fdrMotif` is a tool that identifies the transcription factor binding sites in a set of sequences when zero or multiple occurrences of a motif exist in a sequence.

Two parameters are required to run `fdrMotif`: 1) file containing sequences to be searched; 2) file (s) containing initial PWM(s). One may run fdrMotif with a single starting PWM or all PWMs in a database such as Transfac or Jasper at once (see an example below). When many PWMs are provided, fdrMotif takes one at a time as the starting PWM. The motifs identified are sorted according to E-values computed using subroutines from MEME (Bailey and Elkan, 1994).

Note, if a file is not in the same directory where `fdrMotif` is run, the path for the file must be specified. An example of command line would be:

Example 1

```
fdrMotif -fm Wei_p53_ChIP.seq -fpwm p53_consensus.mx
```

This command line allows `fdrMotif` to run motif analysis on sequences in Wei_p53_ChIP.seq with p53_consensus.mx as the starting PWM with all default settings. The background sequences are generated from a Markov model estimated from sequences in Wei_p53_ChIP.seq.

Example 2

```
fdrMotif -fm Wei_p53_ChIP.seq -fpwm p53_consensus.mx -Imb 1
-fb cds.seq
```

This command line allows `fdrMotif` to run the same analysis as in **example 1** except that the background sequences are generated from a Markov model estimated from sequences in cds.seq

Example 3

```
fdrMotif -fm Wei_p53_ChIP.seq -fpwm p53_consensus.mx -Imb 0
-fb cds.seq
```

This command line allows `fdrMotif` to use sequences in cds.seq as the background sequences.

Example 4

```
fdrMotif -fm Wei_p53_ChIP.seq -fpwm *.mx
```

This command line allows `fdrMotif` to use all PWMs with `.mx` extension as the starting PWMs.

Example 5

```
fdrMotif -fm Wei_p53_ChIP.seq -fpwm p53_consensus.mx -g 100 -fdrb 0.10
-nbsets 10 -Cmax 10 -fb Wei_p53_ChIP.seq -Imb 1 -fo output.txt
```

Run `fdrMotif` on sequences in Wei_p53_ChIP.seq using p53_consensus.mx as the starting PWM. The maximal number of iterations is 100. Upper bound of FDR is controlled at 10%. Background sequences are generated from a Markov model estimated from sequences in Wei_p53_ChIP.seq. Ten sets of such background sequences are generated.

Sequences

All sequences should be in FASTA format. The maximal numbers of sequences allowed for motif set and background set are 2000 (MAX_NUM_SEQ) and 20000 (MAX_NUM_BSEQ). The maximal sequence length (MAX_SEQ_LENGTH) allowed is 8000. These settings can be changed in “`fdr_defines.h`”. An example of the FASTA format is as follows:

```
>chr1:12610241-12610625 5'pad=0 3'pad=0 revComp=FALSE strand=
ggaagagttaatcggatcggctttggctgatagttcaggctccaaagttc
agtcccagtcagagccaccccggaggaattgtaaatctcagggcagtatt
taacaaaacaaaagcaacctggaattacatgcaggtttggttttctacag
tacatatttacttaatccccaaggtatgcggctccatgtcagatcagctg
gctttgcggccctttcacccccctagttcacaacagtttaagtttcaaac
taattccctgttttcgctcttcctcttcacagggctggctggagacagcc
```

```
tggcctgcctccctctcctgatggctctggtcaccgcgtgagtcagcctg
gcctgggctgggagttgggtgacagcctgcccact
>chr1:18703077-18703668 5'pad=0 3'pad=0 revComp=FALSE strand=
cgcatcagcccgcacaacttctggccgaggccagccggcagaggcggact
tggggttggagtgtttgtttgtttgaacttcctcgtcgtcgccaccttcc
ctccccccaacctccaccccacctcacccccctccccagcttctggacgc
gtttgactgcagccaggggtggggggtgggggtagggagtgtgtgtggag
gggagggagaagaggttaaaaaaaagaagacgaagaagacggaaagaaag
agatcgcagcaggggtgaagggagcggacgggaagcgattttgccgact
ttggattcgtccccggcgtgcgcaagaatggcggcccttcccggcacggt
accgagaatgatgcggccggctccggggcagaactacccccgcacgggat
tccctttggaaggtaagaacgcccaggctggcctcgccgcgactccgccg
cccggaactcggggtccttggagaggctgcggtctccaggggacggtggc
ggcgccggcgatagcagagggatcccgttctcttctgggtcccagtccgg
gcgcggaacccagggagtttctgggacccatacttgtccgct
>chr1:52581595-52581889 5'pad=0 3'pad=0 revComp=FALSE strand=
caaagaacgaaacaagtagagtgctttacaaatgcagatggagggaaagt
catcactgagcatcagggtgcggagggcaggaatgctcctgcttctaggc
tgttggcttccgccttccccccctgcaaactcagttccctgcagcgcggga
agccttttaggaatcggagtgtggaacagaggaacgctcttaacagttcn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnngactgagtctacagagaacgtcctcagtttcagaggttcc
```

`fdrMotif` allows a user to choose one of the following three options for background sequences:

1) to simulate background sequences from a Markov model estimated from sequences in file with –fm argument

2) to simulate background sequences from a Markov model estimated from sequences in file with `–fb` argument

3) to directly use sequences in file with `–fb` argument as background sequences

Option 1 or 2 is preferred. `fdrMotif` automatically determines the order (from 0th up to 5th) of the Markov model according to the size of the data (Liu et al., 2001). When option 1 or 2 is used, `fdrMotif` generates `nbsets` (default 10) of background sequences of the same length as the average length of the input/motif sequences. *However, when option 3 is used, make sure the file specified with the* `–fb` *argument contains* `nbsets` *times* (×) *the number of sequences that are at least as long as the average length of the input sequences*. For instance, if one wishes to use a set of randomly selected cds regions as the background sequences for 542 p53 ChIP dataset, she/he would need $nbsets \times N = 10 \times 542$ background sequences, each of which is at least 1183bp (average length of the 542 ChIP sequences) long.

Initial position weight matrix (PWM)

`fdrMotif` reads in the initial PWM from an input file with the argument `–fpwm` followed by the filename. The first line must contain two (only two) integers that specify the dimension of the PWM. The first integer should always be 4 whereas the second integer should match the length of the motif/PWM. For example:

```
4  17
0  7  0  4  4  0  0 11  0  4 11  0 11 11  4  0  1
```

```
7  0  2  0  0  0 11  0  0  2  0 11  0  0  0  0  1
2  1  2  0  0 10  0  0  0  1  0  0  0  0  0 11  1
2  3  7  7  7  1  0  0 11  4  0  0  0  0  7  0  1
```

The value in each cell can be an integer or a float number, e.g., frequency. Standardization will be carried out by `fdrMotif`.

Upper bound of False Discovery Rate

`fdrMotif` selects as many binding sites as possible while controlling a user specified False Discovery Rate (FDR) bound. FDR is defined as the expected proportion of non-motif subsequences that are falsely declared as binding sites. The FDR bound can be any float number greater than 0 and less than 1. Since fdrMotif controls the bound of the FDR, FDR=0.15 may be reasonable. The default can be changed by using the argument `–fdrb`.

Maximal number of iterations

`fdrMotif` stops when the PWM is converged or the maximal number of iterations has been reached. The default is 100. The default can be changed by specifying the argument `–g` followed by a number, e.g., `–g` 100.

Number of sets of background sequences

At each step of PWM estimation, statistical tests are used to decide the number of binding sites in each sequence in the input file. In order to control the FDR, `fdrMotif` needs to generate many sets of background sequences. The FDR is controlled by monitoring the proportion of background subsequences that are (falsely) declared as binding sites. The default is 10; this means that 10 times the number of input sequences are used as the background sequences. This parameter can be changed by using the `-nb` argument.

We found that 5-10 sets should work well. The result might not be stable when less than 5 are used. On the other hand, the larger the number, the more the memory is needed.

Maximum number of binding sites in a single sequence ($C_{\max}$)

`Cmax` is the maximal number of binding sites in any of the sequences. The default is 10. If one believes that more than 10 binding sites may be present in at least one of the sequences, this should be changed by using the `–nbsets` argument.

Output

`fdrMotif` outputs the PWM, specified FDR bound, the computed FDR bound, and the number of binding sites identified in the input and background sequences at each iteration. When the PWM converges or the number of iterations exceeds a maximum, `fdrMotif` stops and prints out the binding sites with 10 bp flanking each side, their locations and strand orientations.

`fdrMotif` also outputs the proportions of the binding sites in both input and background sequences. If the proportion in background sequences is small (e.g., <0.05), one might consider increasing the FDR bound and re-running `fdrMotif`. With a larger FDR bound, more binding sites in the input sequences should be found. Other factors that may affect the results include 1) the quality of the starting PWM; and 2) the background model that is estimated from either the input sequences (`default`) or user-specified sequences other than the input sequences (e.g., negative ChIP sequences). We recommend that users try different starting PWMs.

When multiple starting PWMs are provided to `fdrMotif`, multiple motifs may be identified. `fdrMotif` outputs each motif into a separate file. For each motif, fdrMotif also computes log likelihood ratio (LLR) score and the corresponding E-value using subroutines from the MEME package (Bailey and Elkan, 1994). The motifs are sorted according E-values and are summarized in “fdrMotifSummary.txt”.

References

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*. *Bol*,, **2**, 28-36.

Liu,X. et al. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac. Symp. Biocomput., 6, 127-138.

Li, L., Bass, R.L., and Liang, Y. (*tbd*) fdrMotif: Identifying cis-elements by an EM Algorithm Coupled with False Discovery Rate Control. Submitted.

Please address comments, questions and suggestions to Leping Li at li3@niehs.nih.gov