GADEM: A Genetic Algorithm Guided Formation of Spaced Dyads Coupled with an EM Algorithm for Motif Discovery

Leping Li
Biostatistics Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, North Carolina 27709, USA
Li3@niehs.nih.gov

GADEM combines spaced dyads and an expectation-maximization (EM) algorithm. Candidate words (4-6 nucleotides) for constructing spaced dyads are prioritized by their degree of overrepresentation in the input sequence data. Spaced dyads are converted into starting position weight matrices (PWMs). GADEM then employs a genetic algorithm (GA), with an embedded EM algorithm to improve starting PWMs, to guide the evolution of a population spaced dyads toward one whose entropy scores are more statistically significant. Spaced dyads whose entropy scores reach a pre-specified significance threshold are declared motifs.

Usage: gadem -fseq seqFile [optional arguments]

Optional arguments (any order):

```
  -gen      maxGeneration  number of GA generations (default: 5)
  -pop      populationSize GA population size (default: 100)
  -maxw3    numTop3mer     number of top-ranked trimers   for dyads (default: 20)
  -maxw4    numTop4mer     number of top-ranked tetramers for dyads (default: 40)
  -maxw5    numTop5mer     number of top-ranked pentamers for dyads (default: 60)
  -maxw6    numTop6mer     number of top-ranked hexamers  for dyads (default: 100)
  -mingap   minSpacerWidth minimal number unspecified nucleotides in spaced dyads
                           (default: 0)
  -maxgap   maxSpacerWidth maximal number of unspecified nucleotides in spaced dyads
                           (default: 10)
  -em       numEM          number of EM steps (default: 20)
  -fracEM   percentSeqEM   fraction of sequences for EM algorithm (default: 0.5)
 -score     scoreInfo      indicator for using top-scoring sequences for EM algorithm
                           ChIP-seq score may be stored in sequence header (see below)
                           [0 – do not use scores (default); 1 – use scores]
  -pv       pvalue         pvalue cutoff for declaring binding site (default: 0.000025)
  -ev       E_valueCutoff  E-value cutoff for declaring motif (default: 0)
  -minN     minSites       minimal number of sites in a motif (default: #bases/5000
  -nbps     numBasePerSite minimal number of bases per binding site (default: 5000)
  -extTrim  extTrim        base extension and trimming [default: 1 (1 -yes,0 -no)]
  -window   window         width of sliding window for comparing motif similarity
                           (default: 6)
  -fout    output          output file name
  -verbose verbose         print on screen [1-yes, 0-no (default)]
```

GADEM requires only the name of the sequence file. The default parameters should work well for most datasets. An example of a simple command line would be:

```
GADEM –fseq p53_ChIP_PET.seq
```

Additional arguments provide options and flexibility.

Note that if the sequences in an input dataset total more than 3-5 Mb, you should use the –fracEM argument to reduce the run time (see below).

For the genetic algorithm (GA), the default number of generations is 5 and population size is 100. These parameters can be changed using the `-gen` and `-pop` arguments, respectively. Using more generations and a larger population sizes will make run times longer and will not guarantee better results.

```
GADEM -fseq p53_ChIP_PET.seq -pop 150 -gen 8
```

GADEM obtains its initial PWM models from the spaced dyads that are constructed from over-represented 3-mer, 4-mer, 5-mer and 6-mer words in all input sequences. Because the top-ranked *k*-mers are re-estimated in every GADEM cycle with motifs after motifs identified in the previous cycle have been masked, the default *k*-mer parameter settings should be large enough. Although using larger values for these parameters may help find low abundance motifs, this will also increase the number of possible spaced dyads, and so the search time.

```
GADEM -fseq p53_ChIP_PET.seq -maxw3 10
```

Up to three of the four *k*-mer lengths can be switched off by setting their parameters to 0. For example, if you wish to search for short motifs, you might set both `-maxw5` and `-maxw6` to 0, the maximal number of unspecified nucleotides in the spaced dyads to 0 (see below), and possibly `-extTrim` to 0. You will be warned if all four parameters are set to 0.

```
GADEM -fseq p53_ChIP_PET.seq -maxw5 0 -maxw6 0 -maxgap 0
```

The minimal and maximal numbers of unspecified nucleotides between the two words in spaced dyads control the lengths of the motifs (defaults: 0 and 10 bp). Setting a larger spacer value permits finding longer motifs. Since the minimal word length in a spaced dyad is 3 bp (a trimer) and the maximal length is 6 bp (a hexamer), the default minimal and maximal initial motif lengths are (3+0+3=6) and (6+10+6=22), respectively. The final motif lengths are determined at the post-processing step through base extension and trimming. A motif can be extended by up to 10 bp on each side, but this can be changed in `defines.h`. Thus, the default minimal and maximal motif lengths could be 6+0=6 bp and 22+10+10=42 bp, respectively.

To search for very long motifs (>40 bp), you might set, for example, `-mingap` and `-maxgap` to 40 and 70 bp. You might also extend the sliding window length for comparing motif similarity to 10 bp (`-window` 10) and use a lower PWM score *p*-value cutoff (e.g., 0.000001) using the `-pv` argument. Typically, longer motifs require longer search times.

GADEM carries out 20 steps of EM on all or a subset (see below) of selected sequences to derive an EM-optimized PWM or until the PWM converges. This number of steps has worked well for all datasets tested. The larger the number of EM steps, the longer the search times.

```
GADEM -fseq p53_ChIP_PET.seq -em 40
```

By default, GADEM randomly selects 50% of the sequences (without replacement) for the EM algorithm. For genome-wide data sets consisting of thousands to tens of thousands of sequences, a 25% to 50% sample should be adequate for obtaining a good estimate of the PWM. For sequence inputs larger than 3-5 Mb, you might want to use the `-fracEM` argument so that the EM algorithm uses a smaller fraction of the sequences, say, 20% or 25%.

In ChIP-chip or ChIP-seq datasets, enriched regions are typically assigned a score that is related to enrichment or significance. By adding a 'score' to the header of a each input sequence (see Sequence format below) and using the option `-score` 1 (default 0), you can specify that the EM algorithm should use the top-scoring `-fracEM` sequences (e.g., 25% highest scoring) instead of a randomly selected `-fracEM` sequences. For example, the following command line allows GADEM's EM algorithm to derive PWMs from sequences with the top 25% of scores:

```
GADEM -fseq CTCF_ChIP_chip.seq -fracEM 0.25 -score 1
```

GADEM filters motifs using *E*-value threshold and by the minimal number of binding sites in a motif. For each EM-optimized PWM, GADEM first computes its exact score distribution using the probability generating functions method of Staden (1989). A subsequence of length *w* (motif length) is declared a binding site when the *p*-value of its PWM score is equal to or less than a pre-specified value. The default value for this threshold is 0.000025 ($2.5 \cdot 10^{-5}$). The following command line resets this threshold to a less stringent $5 \cdot 10^{-5}$.

```
GADEM –fseq OCT4_ChIP_chip.seq –pv 0.00005
```

GADEM uses a subroutine from MEME (Bailey and Elkan, 1994) to compute the *E*-value of a motif (i.e. of a set of aligned binding sites). Details can be found in MEME documentation (http://meme.sdsc.edu/meme/intro.html) and in Bailey and Gribskov (1998). The default threshold for the natural log of the *E*-value is 0.

You can adjust the minimal number of binding sites in a motif by using the –minN argument. This argument applies to all motifs identified in the data. If you do not set –minN in the command line, by default GADEM uses the total number of nucleotides in the input sequence data divided by 5,000 as the minimum number. For data containing short sequences, the GADEM determined minimum can be small (e.g., 1000×200/5000=40 for 1000 sequences of 200bp in length, each). Setting a non-default value for the –minN option is recommended.

```
GADEM –fseq OCT4_ChIP_chip.seq –pv 0.00005 –minN 150
```

GADEM automatically adjusts the widths of the motifs that it finds using information content profiles through base extension and trimming at the post-processing step. To turn this off, set -extTrim to 0.

After each GA generation, GADEM identifies unique motifs in the population by comparing motifs using a sliding window. Two motifs are considered similar when the similarity measure (see supplementary material) between the two motifs in any sliding window is less than or equal to a threshold value. The default window size is 6 bp. For very long motifs (>40 bp), a longer window is recommended (e.g. 10 bp).

The –verbose argument prints out immediate results on screen. It does not affect the output file.

The maximal number of sequences is set to 20,000 (MAX_NUM_SEQ), and the maximal sequence length (MAX_SEQ_LENGTH) allowed is 15,000. You can work with datasets larger than these limits by changing the values in defines.h, which is located in the src directory, then rebuilding the executable by going to the directory above src, typing 'make clean' and then 'make install' (see installation instruction on GADEM web site).

Sequence format

All sequences should be in FASTA format. An example of the FASTA format is follows. Each sequence consists of a header in a single line starting with the '>" character. The nucleotides in a sequence can be in a single line [maximal length=MAX_BUFFER_LENGTH (15,000) defined in defines.h in the src directory] or in multiple lines. Note that GADEM will report sequences by an integer ID number that it assigns to represent each input sequence in the file specified by –fout argument, and does not pass any information from a sequence header through into its report, so you are free to include any combination of text and whitespace in the header.

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

```
gcctgggctgggagttgggtgacagcctgcccact
```

As noted above, if you specify a 'score' in a sequence header, GADEM can use the score as a guide when sequences are selected that the EM algorithm will use to derive optimized motif models (PWMs). If you generate sequence sets by exporting from the UCSC genome browser, one way to include the score in the sequence header is to add a fifth column in the UCSC BED file (http://genome.ucsc.edu/FAQ/FAQformat#format1). For instance,

chr1	10000000	10000200	Name1	Score=15.0
chr1	10000000	10000200	Name1	score=15.0
chr1	10000000	10000200	Name1	15.0
chr1	10000000	10000200	Name1	[any number or char]_15.0

```
>hg18_ct_test_name1_Score=15.0 range=chr1:1000001-1000200 5'pad=0 3'pad=0 strand=+
ACGTGGCTGCTCTCACACATGGGCCATGTGTTCACACGCTCTATGCCCCC
GTGTCCACAGGCTCTCACACACGTGCCGTGTCCGGAAGCTCACATATGCC
ATGTCCACACTCACACACGCCGTGTCCACACTCACACGCCGTGTCCACAC
TCTCACACACATGCCATGTCCACATGCTCTCACACACGTGCCCTGTGTCC

>hg18_ct_test_name1_score=15.0 range=chr1:1000001-1000200 5'pad=0 3'pad=0 strand=+
ACGTGGCTGCTCTCACACATGGGCCATGTGTTCACACGCTCTATGCCCCC
GTGTCCACAGGCTCTCACACACGTGCCGTGTCCGGAAGCTCACATATGCC
ATGTCCACACTCACACACGCCGTGTCCACACTCACACGCCGTGTCCACAC
TCTCACACACATGCCATGTCCACATGCTCTCACACACGTGCCCTGTGTCC

>hg18_ct_test_name1_15.0 range=chr1:1000001-1000200 5'pad=0 3'pad=0 strand=+
ACGTGGCTGCTCTCACACATGGGCCATGTGTTCACACGCTCTATGCCCCC
GTGTCCACAGGCTCTCACACACGTGCCGTGTCCGGAAGCTCACATATGCC
ATGTCCACACTCACACACGCCGTGTCCACACTCACACGCCGTGTCCACAC
TCTCACACACATGCCATGTCCACATGCTCTCACACACGTGCCCTGTGTCC

>hg18_ct_test_2830_15.0 range=chr1:1000001-1000200 5'pad=0 3'pad=0 strand=+
ACGTGGCTGCTCTCACACATGGGCCATGTGTTCACACGCTCTATGCCCCC
GTGTCCACAGGCTCTCACACACGTGCCGTGTCCGGAAGCTCACATATGCC
ATGTCCACACTCACACACGCCGTGTCCACACTCACACGCCGTGTCCACAC
TCTCACACACATGCCATGTCCACATGCTCTCACACACGTGCCCTGTGTCC
```

GADEM automatically recognizes all four types of headers. GADEM looks for the key word 'score=' (case insensitive) in the first string (first character string before a space) following '>' in the header of each sequence and takes the number following the key word as the quality score for the sequence. If no such key word is found, GADEM takes the number following the last '_' in the first string in the header as the quality score. However, this flexibility can misinterpret the wrong field as the score. For instance, if you use a '_' in the fourth column (name field) in a BED file, e.g., ER_1, ER_2, etc, the number following the '_' (1, 2, etc, in this example) will be interpreted as the sequence quality score.

Consider not using '_' in the name field (the 4th column) when you do not provide a score column (the 5th column). Alternatively, one might want to add the 5th column (score column) in the UCSC BED file:

Option1:

chr1	10000000	10000200	ER1

Remove "_" in the 4th column if no score is provided

Option2:

chr1	10000000	10000200	ER_1	**15.0**

Add a 5th column (score).

Option3:

chr1	10000000	10000200	ER_1_**15.0**

Append score to the 4th column with a '_'.

When you use scores, you might check the `info.txt` file that is written to the output folder in order to verify that GADEM correctly identified the number of sequences containing scores. If the `-score` is set to 1, this file should also report the quality score cutoff used in selecting the sequences for EM.

GADEM is reasonably robust to errors in setting score values in sequence headers. If you provide no quality scores in sequence headers but set `-score` to 1, the `-score` option is ignored. However, if only a subset of the sequences (n1) have quality scores and the number of sequences (n2) specified by the option `-fracEM` exceeds the number of sequences having quality scores, then GADEM will choose n1 + the first (n2-n1) sequences that do not have scores for the EM algorithm.

Output

Individual motifs are numbered according to the order in which they are identified. Those files contain only the sequences of the binding sites which can be used to create motif logos using `Weblogo` (this can be downloaded at http://weblogo.berkeley.edu/). For example, the following command will generate a PNG logo for motif '1':

```
Weblogo –F PNG –w 18 –b –h 5 –a –c –p –Y –f 1 –o 01
```

The report file whose path and name can be specified by the `-fout` argument contains not only the individual motifs but also the locations (seq. and position) of the sites in the original sequence data. For each motif, the report gives the as the observed PWM from all identified binding sites. It also includes the spaced dyad from which the motif is derived, PWM score *p*-value cutoff for the run, the natural log of the motif's *E*-value, and the numbers of sequences containing 0 (no predicted sites), 1, 2, and >2 predicted sites. Below is an example:

```
Cycle[  1] motif[4]:
   spaced dyad:                aaannnnnnnnngccctg
   motif consensus:            rggtcasnstgacct
   motif length (w):           15
   number of sites:            1554
   log(E-value):               -1163.73
   log(E-value)/w:             -77.58
   pwm p-value cutoff:         2.500000e-05

   #seq_0_site  #seq_1_site  #seq_2_site  #seq_>2_site
        2331      1150       155        29

tgcttgacccAGGTCACTGTGACCTgcccatttttt   +   3442   645    4.384301e-10
ctgatgttctGGGTCAGGCTGACCTggtttaacta   +   1013   407    1.805042e-09
cctagctaccAGGTCACACTGCCCTggcaaaggca   +   2574   132    3.593016e-09
agcattggtaAGGTCAGGCTGACCTgacttgcagc   +   3235   688    3.773176e-09
tactgtttccGGGTCAGGGTGACCTctggggtgag   +   1998   183    4.098112e-09
cagctggcctGGGTCACAGTGACCTgacctcaaac   +    173   542    4.746667e-09
ctcatcataaAGGTCACAGTGACCTggaaaatgag   +    639   431    5.230866e-09
gccatgcagaAGGTCACCCTGACCTccgttcttgt   +   1645   325    5.914815e-09
agaaataacaAGGGCAGAGTGACCTggaacacaaa   +   2019   287    6.517660e-09
acagcttgacAGGTCAGCCTGACCTcacatctaga   +    498   659    7.202862e-09
```

The first column reports the sequence of a predicted site in upper case with 10-bp flanking sequences in lower case. The second column indicates the strand orientation of the site in the original data. The third column specifies the position of the site (not counting the flanking regions) relative the start of the sequence (the first base of the sequence being 1). When a site is found in the reverse complementary strand to the input sequence, the last position of the site in the original orientation will be listed as the start of the site. The fourth column lists the ID assigned to the sequence in which the site is located; IDs are integers that give the position in which the sequences occur in the input file, starting with 1 for the first sequence. Finally, the last column lists the *p*-value of the site (see the manuscript for *p*-value computation).

GADEM employs two algorithms (genetic algorithm and expectation-maximization), neither of which guarantees a globally optimum solution. Because of this, we recommend that you carry out several independent runs for each dataset. For each run, GADEM automatically uses a different unique random seed number. Since each run may produce several motifs, you should examine each motif separately and compare it with the same motifs from the other runs. One might consider the motif with the lowest log(E-value) among the same motifs from all runs as the "best" motif. In the following example, one may consider the motif from run 2 as better than that from run 1.

```
Run 1: Cycle[  2] motif[12]:
  spaced dyad:                  ccaggnnnnnnnnnctt
  motif consensus:              aggtcasnstgaccy
  motif length (w):             15
  number of sites:              1608
  log(E-value):                 -1194.81
  pwm p-value cutoff:           2.500000e-05

  #seq_0_site  #seq_1_site  #seq_2_site  #seq_>2_site
  2288         1182         169          26

Run2:  Cycle[  1] motif[5]:
  spaced dyad:                  aggnnnnnnnncca
  motif consensus:              aggkcasnstgacc
  motif length (w):             14
  number of sites:              1633
  log(E-value):                 -1622.02
  pwm p-value cutoff:           2.500000e-05

#seq_0_site  #seq_1_site  #seq_2_site  #seq_>2_site
2298         1145         187          35
```

ACKNOWLEDGEMENT

Special thanks to Gordon Robertson at the BC Cancer Agency Genome Sciences Centre for testing the GADEM software, insightful comments and suggestions and careful reading of this documentation.

REFERENCE

Li, L. GADEM: A genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery, *in review*.

Send questions and comments to li3@niehs.nih.gov

Package download: http://www.niehs.nih.gov/research/resources/software/gadem/