```
*-------------------------------------------------------------------------
|                                                                        |
|      GADEM: a motif discovery tool for large scale sequence data       |
|                                v1.3                                    |
|                                                                        |
|         Multiple runs are recommended for 'unseeded' analysis.         |
|      Each unseeded run automatically uses a different random seed.     |
|        'Seeded' runs are deterministic; no repeat runs are needed.         |
*-------------------------------------------------------------------------
```

Usage: gadem -fseq seqFile optional arguments

Optional arguments that need attention:

-posWt	0,1,2,or 3	Weight profile for positions on the sequence (see documentation). 0 - no weight (uniform spatial prior, default), 1 - small or zero weights for the ends and large weights for the center (e.g. the center 50 bp) If you expect strong central enrichment (as in ChIP-seq) and your sequences are long (e.g. >200 bp), choose type 1.
-widthWt	integer	For -posWt 1 or 3, width of central sequence region with large EM weights for PWM optimization (default: 50). This argument is ignored when -posWt is 0 (uniform prior) or 2 (Gaussian prior).
-ev	decimal	ln(E-value) cutoff for selecting MOTIFS (default: 0.0). If a seeded analysis fails to identify the expected motif, run GADEM with -verbose 1 to show motif ln(E-value)s on screen, then rerun with a larger ln(E-value) cutoff. This can help in identifying short and/or low abundance motifs, for which the default E-value threshold may be too low. The subroutine for E-value calculation is adapted from the MEME package.
-pv	decimal	P-value cutoff for declaring BINDING SITES (default: 0.0002). Depending on data size and the motif, you might want to assess more than one value. For ChIP-seq data (e.g., 10 thousand +/-200-bp max-center peak 'cores'), p=0.0002 often seems appropriate. However, short motifs may require a less stringent setting. Given a subsequence s of length w, GADEM computes the log likelihood (llr) score, log{p(s|M)/p(s|B)}, where M is the EM-derived motif model, B is the

b-th order Markov background model and w is the motif length. The subsequence is declared a binding site if its llr score is at or above the llr score corresponding to the p-value cutoff. This requires knowing the distribution of the llr score (under the null), and GADEM implements two approaches for approximating the null distribution: Staden probability generating function (pgf) method (Comput. Appl. Biosci., 5,89,1989) and an empirical approximation method through generating many background sequences. Both approaches are described briefly below.

-minN integer Minimal number of sites required for a motif to be reported (default: numSeq/20).

-fpwm0 string File name for the seed PWM, when a 'seeded' approach is used. A PWM (format below) can be used as the starting PWM for the EM algorithm. This will help find an 'expected' motif and is much faster than 'unseeded' de novo discovery. Also, when a seed PWM is specified, the run results are deterministic, so only a single run is needed (repeat runs with the same settings will give identical results). In contrast, unseeded runs are stochastic, and we recommend comparing results from several repeat runs.

Format: number of rows & columns followed by integer counts OR decimal frequencies.
Example: PWM (CREB, JASPAR MA0018) in two acceptable representations:

```
4     12
0     3     0     2     5     0     0    16     0     0     1     5
7     5     3     3     1     0     0     0    16     0     5     6
5     4     6    11     7     0    15     0     0    16     0     3
4     4     7     0     3    16     1     0     0     0    10     2

4     12
0.000 0.188 0.000 0.125 0.312 0.000 0.000 1.000 0.000 0.000 0.062 0.312
0.438 0.312 0.188 0.188 0.062 0.000 0.000 0.000 1.000 0.000 0.312 0.375
0.312 0.250 0.375 0.688 0.438 0.000 0.938 0.000 0.000 1.000 0.000 0.188
0.250 0.250 0.438 0.000 0.188 1.000 0.062 0.000 0.000 0.000 0.625 0.125
```

-pgf 1 or 0 By default, GADEM uses the Staden probability generating function (pgf) method to approximate the exact llr score null distribution.

The pgf method assumes that the background model is independent and identically-distributed (iid). When this method is used, GADEM takes the frequencies of [a,c,g,t] in the input data as estimates of the parameters of this iid model (0th-order), and a user-specified background model is not

required. In other words, when -pgf is set to 1 (default) or is unspecified, both the -fbm and -bOrder flags are ignored by GADEM.

Alternatively (when -pgf is set 0), the GADEM approximates the null using the llr scores of many background subsequences of length w, where w is the motif length. It generates the background subsequences using the [a,c,g,t] frequencies in the input data.

-bOrder integer The order of the background Markov model for computing llr scores:
0 - 0th
1 - 1st
2 - 2nd
...
8 - 8th

-fbm string Name of the file containing the user-specified background model. The GADEM package includes pre-computed genome-wide frequencies data for human, mouse and Drosophila, and source code for generating such files.

The background Markov model can be estimated from the input data by GADEM or read from a file using the -fbm argument (see format below). We recommend -bOrder 0 when -fbm is not used. Otherwise, a higher order (e.g., -bOrder 3 or 4) may be reasonable.

Up to 8th order (octamer + 1nt = nonamer) is allowed.

```
#monomer frequency
a       0.20850000001660
c       0.29149999998340
g       0.29149999998340
t       0.20850000001660
#dimer frequency
aa      0.04800960194357
ac      0.05151030207800
ag      0.08171634323790
at      0.02720544114470
             ....
ta      0.03460692142891
tc      0.05361072215865
tg      0.07231446287687
tt      0.04800960194357
#trimer frequency
aaa     0.01200480194395
```

```
                    aac     0.01550620248175
                    aag     0.01420568228200
                    aat     0.00630252106809
                                 ....
                    tta     0.01190476192858
                    ttc     0.01180472191322
                    ttg     0.01230492199004
                    ttt     0.01200480194395
                                 ....
```

If the file containing the background model is not specified and -pgf is set to 0, GADEM estimates the model from the input sequences. Note that when GADEM estimates a background model from input data that consists of short sequences (e.g. ChIP-seq), a higher order model is not recommended, as the sequences generated by the resulting background model may be too similar to the input sequences. For such cases we suggest setting -bOrder to 0.

Other optional arguments:

-gen	integer	Number of genetic algorithm (GA) generations (default: 5).
-pop	integer	GA population size (default: 100). Both default settings should work well for most datasets (ChIP-chip and ChIP-seq). The above two arguments are ignored in a seeded analysis, because spaced dyads and GA are no longer needed (-gen is set to 1 and -pop is set to 10 internally, corresponding to the 10 maxp choices).
-fullScan	0 or 1	GADEM keeps two copies of the input sequences internally: one (D) for discovering PWMs and one (S) for scanning for binding sites using the PWMs. Once a motif is identified, its instances in set D are always masked by Ns. However, masking motif instances in set S is optional, and scanning unmasked sequences allows sites of discovered motifs to overlap. 0 (default) - scan masked sequences in S (disallow motif site overlap). 1 - scan unmasked sequences in S (allow motif sites to overlap) (was default in v1.2).
-em	integer	Number of EM steps (default: 40). One might want to set it to a larger value (e.g. 80) in a seeded run, because such runs are fast.
-fEM	decimal	Fraction of sequences used in EM to obtain PWMs in an unseeded analysis (default: 0.5). For unseeded motif discovery in a large dataset (e.g. >10 million nt), one might want to set -fEM to a smaller value (e.g., 0.3 or 0.4)

```
                    to reduce run time.

                    When only partial input data are used in EM and ?verbose is set to 1, the
                    number of binding sites printed on screen is the number of sites found only
                    in the sequences that are used in EM optimization.

                    This argument is ignored in a seeded analysis, which uses all sequences EM to
                    obtain the PWMs.

-extTrim  1 or 0    Base extension and trimming (1 -yes, 0 -no) (default: 1).

-maxw3    integer   Number of top-ranked trimers for spaced dyads (default: 20).
-maxw4    integer   Number of top-ranked tetramers for spaced dyads (default: 40).
-maxw5    integer   Number of top-ranked pentamers for spaced dyads (default: 60).

-mingap   integer   Minimal number of unspecified nucleotides in spaced dyads (default: 0).
-maxgap   integer   Maximal number of unspecified nucleotides in spaced dyads (default: 10).
                    -mingap and -maxgap control the lengths of spaced dyads, and, with -extrim,
                    control motif lengths. Longer motifs can be discovered by setting -maxgap to
                    larger values (e.g. 50).

-useScore 0 or 1    Use top-scoring sequences for deriving PWMs. Sequence (quality) scores are
                    stored in sequence header (see documentation).
                    0 - no (default, randomly select sequences), 1 - yes.

-fpwm     string    Name of output PWM file in STAMP format (http://www.benoslab.pitt.edu/stamp).
                    (default: observedPWMs.txt). This file can be loaded into STAMP to compare
                    each PWM with PWMs in databases for similarity.

-fout     string    Name of main GADEM output file (see documentation for description) (default:
                    gadem.txt).

-nbs      integer   Number of sets of background sequences (default: 10). The background sequences
                    are simulated using the [a,c,g,t] frequencies in the input sequences, with
                    length matched between the two sets. The background sequences are used as the
                    random sequences for assessing motif enrichment in the input data. Another set
                    (same default: 10) of background sequences is independently generated to
                    approximate the empirical llr score distribution when -pgf is set to 0.

-verbose  1 or 0    Print immediate results on screen [1-yes (default), 0-no]. These results
                    include the motif consensus sequence, number of sites (in sequences subjected
                    to EM optimization, see -fEM, above), and ln(E-value).
```

Examples:

1. Unseeded analysis for ChIP data with expected central enrichment (llr distr.: pgf method - default)

```
gadem -fseq input.seq -minN 1000 -posWt 1 –verbose 1
```

2. Unseeded analysis for ChIP data with expected central enrichment (llr distr.: empirical approx.)

```
gadem -fseq input.seq -minN 1000 -posWt 1 -pgf 0 -fbackg freq.txt -bOrder 4 –verbose 1
```

3. Seeded analysis for ChIP data with expected central enrichment (llr distr.: pgf method - default)

```
gadem -fseq input.seq -minN 1000 -posWt 1 -fpwm0 startPWM.mx –verbose 1
```

4. Seeded analysis for ChIP data with expected central enrichment (llr distr.: empirical approx.)

```
gadem -fseq input.seq -minN 1000 -posWt 1 -fpwm0 startPWM.mx -pgf 0 -fbackg freq.txt -bOrder 4 –verbose 1
```

Sequence format

All sequences should be in FASTA format (below). 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
```

Outputs

`GADEM` outputs the following files (all in ASCII text).

1 `info.done.txt` (initially as `info.txt` and renamed after the job is completed).
This file contains the summary information on the run including the command line options and all parameter settings used in the analysis.

2. File containing the main `GADEM` result (file name specified by `-fout` option, default: `gadem.txt`).
This file contains not only the individual motifs identified but also the locations (seqID and position) of the sites in the original sequence data. 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 in both input and background/random sequences.

The first column contains the sequence header. The second column reports the sequence of a predicted site in upper case with 10-bp flanking sequences in lower case. The third column indicates the strand orientation of the site in the original data. The fourth 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 fifth 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). Here is an example:

```
Cycle[  1] motif[1]:
spaced dyad:                nGknCAAAGkyCAn
motif consensus:            rGknCAAAGkyCAn
m=ac r=ag w=at s=cg y=ct k=gt b=cgt d=agt h=act v=acg
motif length(w):            14
maxpfactor:                 0.3000
number of sites:            10810
ln(E-value):                -11693.48
pwm p-value cutoff:         2.000000e-04

Seqs with 0,1,2,>2 sites:    3751,5924,1915,336
8175(68.55%) of 11926 seqs have >=1 site

Sequence header                   10bp flanking--MOTIF--10bp flanking  strand  seqID  pos  p-value
>ht28_chr4:136561558-136561957_+  gctgctttaaAGCGCAAAGTCCACtttcagcctg     +     2988   149  0.000000e+00
>ht37_chr13:37824125-37824524_+   tcataacctgAGCGCAAAGTCCACccggagcttg     -     8912   218  0.000000e+00
>ht17_chr2:21144091-21144490_+    atggtttgctGGGCCAAAGTCCAAgcgtagccct     +     902    228  0.000000e+00
>ht16_chr13:97424144-97424543_+   ctcttcctgaGGGCCAAAGTCCAAttatcaacac     +     9179   192  0.000000e+00
>ht16_chr11:116433053-116433452_+ gatggcaaagAGGTCAAAGTCCAAgaggacctcc     +     8229   160  0.000000e+00
>ht23_chr18:80646417-80646816_+   gtcgctggacAGGTCAAAGTCCAAatcctgggtg     +     11475  232  0.000000e+00
>ht45_chr7:51720573-51720972_+    agagagtcagGGGTCAAAGTCCAAagttcattca     +     4845   177  0.000000e+00
>ht20_chr8:85921560-85921959_+    gagcttactgGGGTCAAAGTCCAAccatggtcta     -     5644   234  0.000000e+00
>ht21_chr10:126414108-126414507_+ ctggagcacaGGGTCAAAGTCCAAcaaggtccct     +     7402   212  0.000000e+00
>ht21_chr11:105019360-105019759_+ tgatggatcaGGGTCAAAGTCCAAactcaggagc     +     8114   196  0.000000e+00
                                                  ⋮
```

```
background set[ 1] Seqs with 0,1,2,>2 sites 0,1,2,>2: 10148 1665 110 3
background set[ 2] Seqs with 0,1,2,>2 sites 0,1,2,>2: 10183 1613 124 6
background set[ 3] Seqs with 0,1,2,>2 sites 0,1,2,>2: 10182 1632 106 6
background set[ 4] Seqs with 0,1,2,>2 sites 0,1,2,>2: 10175 1633 113 5
background set[ 5] Seqs with 0,1,2,>2 sites 0,1,2,>2: 10185 1600 135 6
background set[ 6] Seqs with 0,1,2,>2 sites 0,1,2,>2: 10170 1621 127 8
background set[ 7] Seqs with 0,1,2,>2 sites 0,1,2,>2: 10295 1514 110 7
background set[ 8] Seqs with 0,1,2,>2 sites 0,1,2,>2: 10204 1590 128 4
background set[ 9] Seqs with 0,1,2,>2 sites 0,1,2,>2: 10211 1602 111 2
background set[10] Seqs with 0,1,2,>2 sites 0,1,2,>2: 10201 1611 106 8
average number of sites in background sequences: 1858, fold enrichment: 5.818.
average number of background sequences that contain at least one site: 1730, fold enrichment: 4.739.
```

3. File containing all observed PWMs corresponding to the identified motifs (file name specified by `-fpwm`, default: `observedPWMs.txt`)
This file can be loaded directly to STAMP (http://www.benoslab.pitt.edu/stamp/) to check for similarity between each of the identified motifs and the known motifs in databases such as TRANSFAC, JASPER, FLYREG, etc.

4 Individual files containing the sequences of the predicted sites
Each file is numbered according to the order in which the motif is identified. Those files can be used to create motif logos using `Weblogo` (http://weblogo.berkeley.edu/). This software can be run at the server or downloaded and run locally as:

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

This will generate a png logo file (01.png) using sequence file 1.seq.

Additional examples of usages:
For the genetic algorithm (GA), the default number of generations is 10 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 5
```

The default *p*-value cutoff for this declaring binding site is 0.0002 ($2.0 \cdot 10^{-4}$). The following command line resets this threshold to a less stringent $5 \cdot 10^{-4}$.

```
gadem -fseq OCT4_ChIP_chip.seq -pv 0.0005
```

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 0.0. For short and/or low abundance motifs, if GADEM fails to identify it, set the ln(*E*-value) cutoff large:

```
gadem –fseq input.seq –pv 0.0005 –ev 5000
```

To change the default number (40) of EM steps:

```
gadem –fseq input.seq –em 80
```

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 sequences divided by 20 as the minimum number. Setting a non-default value for the `–minN` option is recommended.

```
gadem –fseq OCT4_ChIP_chip.seq –pv 0.0005 –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.` This may be useful for a seeded analysis for which you do not wish to change the motif length.

```
gadem –fseq OCT4_ChIP_chip.seq –pv 0.0005 –extTrim 0 –fpwm0 Oct.mx
```

For an unseeded analysis, `GADEM` obtains its initial PWM models from the spaced dyads that are constructed from over-represented 3-mer, 4-mer, and 5-mer words in all input sequences. Up to two of the three *k*-mer ($k = 3,4,5$) lengths can be switched off by setting their parameters to 0. For example, if you wish to search for short motifs, you might set `-maxw5` 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 –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 5 bp (a pentamer), the default minimal and maximal initial motif lengths are (3+0+3=6) and (5+10+5=20), 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 20+10+10=40 bp, respectively.

To search for very long motifs (>40 bp), you might set, for example, `–mingap` and `–maxgap` to 40 and 70 bp and use a more stringent PWM score *p*-value cutoff (e.g., 0.000001) using the `–pv` argument. Typically, longer motifs require longer search times.

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 –fEM argument so that the EM algorithm uses a smaller fraction of the sequences, say, 20% or 25%.

```
gadem –fseq CTCF_ChIP_chip.seq –fEM 0.25
```

The –verbose argument prints out immediate results on screen. It does not affect the output file. Setting –verbose to 1 is particularly useful when GADEM fails to identify any motifs. This will allow GADEM to print on screen the number of predicted sites and the ln(*E*-value). One may adjust the settings for –minN and –ev, accordingly.

```
gadem –fseq OCT4_ChIP_chip.seq –pv 0.0005 –extTrim 0 –fpwm0 Oct.mx –verbose 1
```

In ChIP-chip or ChIP-seq datasets, enriched regions may be assigned a score that is related to enrichment or significance. 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	ER**1**

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 set `-useScore` to 1, 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.

GADEM is reasonably robust to errors in setting score values in sequence headers. If you provide no quality scores in sequence headers but set `-useScore` to 1, the `-useScore` 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 `-fEM` 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.

The follow command line allows GADEM to choose the top-scoring `-fEM` sequences (e.g., 25% highest scoring) instead of a randomly selected `-fEM` sequences to derive PWMs:

```
gadem -fseq CTCF_ChIP_chip.seq -fEM 0.25 -useScore 1
```

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.

Change global settings

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).

ACKNOWLEDGEMENT

Thanks to Gordon Robertson at the BC Cancer Agency Genome Sciences Centre for providing the opportunity to work with the MORGEN project and with whom I developed the seeded algorithm, David Umbach and Grace Kissling at NIEHS for helpful suggestions and Robert Bass for creating and maintaining the GADEM web site.

REFERENCE

Li, L. (2009) GADEM: A genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. *J. Comput. Biol.*, **16**, 317-329.

Li, L., Robertson, G., Hoffman, B.G., Marra, M.A., Hoodless, P.A., and Jones, S.J.M. Identifying motifs using GADEM with a starting PWM. *Submitted*.

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

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

Last modification: September 15, 2009