Contents

[1. loadIntervals: 3](#_Toc497822346)

[2. exportMemeIntervalSequences 4](#_Toc497822347)

[3. runMeme/runRandDreme: 5](#_Toc497822348)

[4. memeSelfMatches/randDremeSelfMatches: 6](#_Toc497822349)

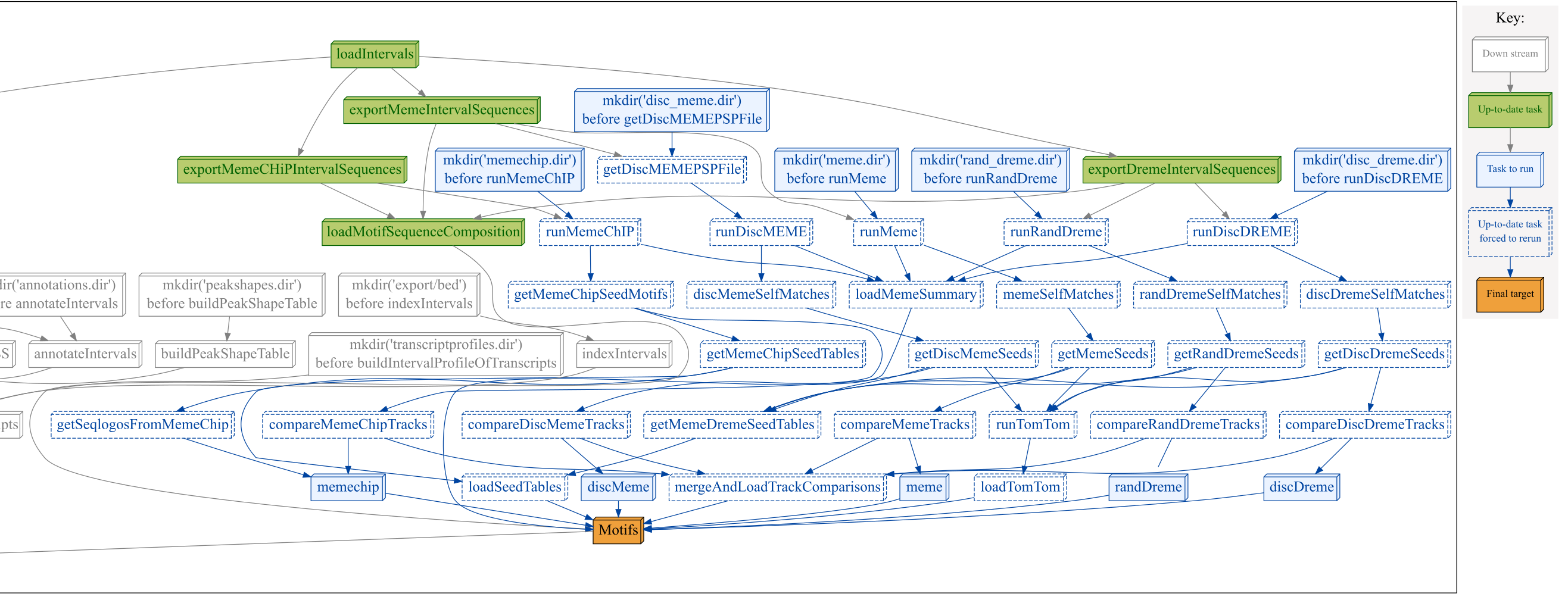
[5. getMemeSeeds/getRandDremeSeeds: 9](#_Toc497822350)

[6. compareRandDremeTracks/compareMemeTracks: 17](#_Toc497822351)

[7. loadMotifSequenceComposition: 18](#_Toc497822352)

[8. runTomTom 19](#_Toc497822353)

[9. Sample run 20](#_Toc497822354)



load intervals from ‘bed.gz’ formatted files into

the database.

# loadIntervals:

Starting from a bed.gz file generates a table in the database of the form:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **avgval** | **contig** | **disttostart** | **end** | **genelist** | **interval\_id** | **length** | **npeaks** | **nprobes** | **peakcenter** | **peakval position** | **score** | **start** | **strand** |
| 1 | chr1 | 0 | 24482400 | 1 | 600 | 1 | 1 | 24482100 | 1 | 0 | 1 | 24481800 | . |
| 1 | chr1 | 0 | 26866800 | 2 | 200 | 1 | 1 | 26866700 | 1 | 0 | 1 | 26866600 | . |
| 1 | chr1 | 0 | 40313200 | 3 | 200 | 1 | 1 | 40313100 | 1 | 0 | 1 | 40313000 | . |
| 1 | chr1 | 0 | 52034600 | 4 | 200 | 1 | 1 | 52034500 | 1 | 0 | 1 | 52034400 | . |
| 1 | chr1 | 0 | 52115800 | 5 | 200 | 1 | 1 | 52115700 | 1 | 0 | 1 | 52115600 | . |
| 1 | chr1 | 0 | 150977000 | 6 | 200 | 1 | 1 | 150976900 | 1 | 0 | 1 | 150976800 | . |

Where:

* contig: The 1st field in the bed.gz file (contig).
* start: The 2nd field in the bed.gz file (start).
* end: The 3st field in the bed.gz file (end).
* genelist: The 4th field in the bed.gz file (In this case, the name of the region).

# exportMemeIntervalSequences

Gets the intervals from the database and generates the genome sequences (fasta files) from them:

>None\_118 chr3:46393800-46394400

caggtggtgtggctcagctaaaacctgagaaccaagaatgagCCTTCTGATATCTCAGGGTTAAATCGGCACAAATGAAACCATTTTATAACCAAACAACATTTTCCCTCCTGAACTGCACTTTTTCTAGAACCCATTTAATCCTTCCTAACATGTTCAATCCAGCCCCGTGATTTTTCAGTCAAATGCAAGATGCGTAGGATTCCTTTTCTGTTCAGTGCTCCCCAGCCAATCTTTTTCCCTTCTCCCACCCCACAGTTACTAATTCCAGCGTTTCAgttctcagcagaggatgacagttccaccaggagacatcagtgttatctggagacattttagttgtcccagctgagggcacagtgctgctggcatctagtgggtgaaggtcaggaatgcacctcaacatcctggaacactcaggacaatccccagcaacaaaaacctacccatccagaacatcaacagggctgtggctgaggagccctgCTCTAGTGCTTCCTCCTTACTGTATCCTCATCAAGGAGGGCTCTGGGGCATGTCAGGGGCTGCTACAGGCTCCTGGCAGATTTCGGCGACTTACTCCAGACTTGTGGAACATTC

>None\_73 chr16:66550600-66551000

tcacgcctgtaatcccagcactttgggaggccaaggaaggcagatcacctgaggccagaagttcgagaccagccttggccaacatggcgaaaccccatcgctactaaaaatgcaaaaattagctgggtgtggtggggggcacccgtattcccagctactcaggaggctgaggcaggaaaatcgcttgaacctgggaggcagagattgcagtgagccgagatcactgcactgcactccagcctgagcaacagagggacacttcgttccccccgccaaaaaaaaaaTTTTTTTTTAATGATCAGAAGCggctgggcgcggtggctcgggcctgtaatcccagcactttcggaggccgaggaaggtagatcacttgaggtcatgagttcaagaccagcctgac

# runMeme/runRandDreme:

Runs Meme/Dreme with specified parameters, outputs the results to a temporary directory and then moves all the results to the directory specified in “params[exportdir]”/meme.dir/track.meme/ (Eg./mnt/fastdata/mbp15ja/PC\_MM\_enhancers\_raw\_data\_RNA\_pipeline\_intervals/export/meme.dir/track.meme/).

It copies the output to the base directory/meme.dir (eg. /mnt/fastdata/mbp15ja/PC\_MM\_enhancers\_raw\_data\_RNA\_pipeline\_intervals/meme.dir), which contains a simplified, text-only version of the output. (See -text, below, for the text output format.) into:

* meme.dir(or rand\_dreme.dir)/track.meme(or .dreme). Eg. meme.dir/enhancers\_intersect\_overexpressed\_genes\_id.meme

# memeSelfMatches/randDremeSelfMatches:

The idea here is to compare sequences to the same sequences, we are trying to see similar sequences which will be grouped later.

Tomtom compares one or more motifs against a database of known motifs (e.g., JASPAR). Tomtom will rank the motifs in the database and produce an alignment for each significant match (sample output for motif and JASPAR CORE 2014 database). Tomtom searches one or more query motifs against one or more databases of target motifs (and their reverse complements when applicable), and reports for each query a list of target motifs, ranked by p-value. The E-value and the q-value of each match is also reported.

The q-value is the minimal false discovery rate at which the observed similarity would be deemed significant, lower q-value, more significant (than by chance). The output contains results for each query, in the order that the queries appear in the input file.

Convert to p-value (significance over all offsets for a particular query – target comparison) and report the minimum p-value (obtained by a specific offset comparison)

Each offset’s score: Add the scores of the overlapping positions for one offset.

Done for multiple offsets between query and target

Each non-overlapping position in the query sequence is given the median score of randomly matching this base

Minimum overlap

Each overlapping position in the query sequence is given a similarity score

Target motif sequence

Query sequence

Compensate for multiple testing (one query to all targets): Convert each query-target p-value to E-value:

E-value = p-value x (2x number of target motifs(=comparisons))

Also q-values for each p-value (q-value is the minimal false discovery rate at which the observed similarity would be deemed significant).

For a given pair of motifs, the program considers all offsets between the motifs, while requiring a minimum number of overlapping positions. For a given offset, each overlapping position is scored using one of seven column similarity functions defined below. Columns in the query motif that don't overlap the target motif are assigned a score equal to the median score of the set of random matches to that column.

In order to compute the scores, Tomtom needs to know the frequencies of the letters of the sequence alphabet in the database being searched (the 'background' letter frequencies), the target sequences. By default, the background letter frequencies included in the query motif file are used.

The scores of columns that overlap for a given offset are summed. This summed score is then converted to a p-value. The reported p-value is the minimal p-value over all possible offsets. To compensate for multiple testing, each reported p-value is converted to an E-value by multiplying it by twice the number of target motifs. As a second type of multiple-testing correction, q-values for each match are computed from the set of p-values and reported.

Inputs:

-Query Motifs: A file containing one or more motifs in MEME format or the HTML (.html) or plain text (.txt) output of MEME or DREME. Each of these motifs will be searched against the target databases. If you only wish to search with a subset of these motifs then look into the -m and -mi options.

-Target Motif Databases: File(s) containing MEME formatted motifs. Outputs from MEME and DREME are supported, as well as Minimal MEME Format. You can convert many other motif formats to MEME format using conversion scripts available with the MEME Suite.

Output:

-Tomtom writes its output to files in a directory named tomtom\_out, which it creates if necessary. (You can also cause the output to be written to a different directory; see -o and -oc, below.)

The main output file is named tomtom.html and can be viewed with a web browser. The tomtom.html file is created from the tomtom.xml file. An additional file, tomtom.txt, contains a simplified, text-only version of the output. (See -text, below, for the text output format.)

For each query-target match, two additional files containing LOGO alignments may also be written -- an encapsulated postscript file (.eps) if the -eps flag is specified and a portable network graphic file (.png) if the -png flag is specified. An install of ghostscript is required to create the png file.

Only matches for which the significance is less than or equal to the threshold set by the -thresh switch will be shown. By default, significance is measured by q-value of the match. The q-value is the estimated false discovery rate if the occurrence is accepted as significant.

For each .meme file runs a command similar to:

tomtom -verbosity 1 -text -thresh 0.05 meme.dir/enhancers\_intersect\_overexpressed\_genes\_id/enhancers\_intersect\_overexpressed\_genes\_id.meme meme.dir/enhancers\_intersect\_overexpressed\_genes\_id/enhancers\_intersect\_overexpressed\_genes\_id.meme 2> meme.dir/enhancers\_intersect\_overexpressed\_genes\_id/enhancers\_intersect\_overexpressed\_genes\_id.meme.self\_matches.log | sed 's/#//' > meme.dir/enhancers\_intersect\_overexpressed\_genes\_id/enhancers\_intersect\_overexpressed\_genes\_id.meme.self\_matches

Therefore compare sequences to the same sequences, output to text mode and use a false discovery rate of 0.05.

# getMemeSeeds/getRandDremeSeeds:

Uses the result from memeSelfMatches (table transposed for easier visualization) for each query, in the order that the queries appear in the input file (most significant binding motifs found first):

|  |  |  |  |
| --- | --- | --- | --- |
| **Query ID** | 1 | 1 | 1 |
| **Target ID** | 1 | 9 | 7 |
| **Optimal offset** | 0 | 4 | -19 |
| **p-value** | 0 | 7.34E-07 | 0.000117362 |
| **E-value** | 0 | 1.47E-05 | 0.00234724 |
| **q-value** | 0 | 1.47E-05 | 0.00156483 |
| **Overlap** | 30 | 18 | 11 |
| **Query consensus** | CTCAGCCTCCCAAAGTGCTGGGATTACAGG | CTCAGCCTCCCAAAGTGCTGGGATTACAGG | CTCAGCCTCCCAAAGTGCTGGGATTACAGG |
| **Target consensus** | CTCAGCCTCCCAAAGTGCTGGGATTACAGG | CCACCTCGGCCTCCCAAAGTGC | GGGATTACAGGCATGA |
| **Orientation** | + | - | - |

Also gets the output from runMeme:

*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*MEME - Motif discovery tool*

*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*MEME version 4.11.2 (Release date: Mon Oct 24 12:27:06 2016 -0700)*

*For further information on how to interpret these results or to get*

*a copy of the MEME software please access http://meme-suite.org .*

*This file may be used as input to the MAST algorithm for searching*

*sequence databases for matches to groups of motifs. MAST is available*

*for interactive use and downloading at http://meme-suite.org .*

*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*REFERENCE*

*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*If you use this program in your research, please cite:*

*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, pp. 28-36,*

*AAAI Press, Menlo Park, California, 1994.*

*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*TRAINING SET*

*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*DATAFILE= ./enhancers\_intersect\_overexpressed\_genes\_id.meme.fasta*

*ALPHABET= ACGT*

*Sequence name Weight Length Sequence name Weight Length*

*------------- ------ ------ ------------- ------ ------*

*None\_163 1.0000 200 None\_165 1.0000 400*

*None\_32 1.0000 800 None\_117 1.0000 200*

*None\_71 1.0000 200 None\_219 1.0000 200*

*None\_183 1.0000 400 None\_78 1.0000 200*

*None\_226 1.0000 400 None\_86 1.0000 200*

*None\_206 1.0000 600 None\_141 1.0000 600*

*None\_15 1.0000 200 None\_229 1.0000 200*

*None\_205 1.0000 200 None\_121 1.0000 400*

*…*

*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*COMMAND LINE SUMMARY*

*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*This information can also be useful in the event you wish to report a*

*problem with the MEME software.*

*command: meme ./enhancers\_intersect\_overexpressed\_genes\_id.meme.fasta -dna -revcomp -mod anr -nmotifs 20 -oc ./ctmpFsUDIr -maxsize 300000 -minw 5 -maxw 30*

*model: mod= anr nmotifs= 20 evt= inf*

*object function= E-value of product of p-values*

*width: minw= 5 maxw= 30*

*width: wg= 11 ws= 1 endgaps= yes*

*nsites: minsites= 2 maxsites= 50 wnsites= 0.8*

*theta: spmap= uni spfuzz= 0.5*

*global: substring= yes branching= no wbranch= no*

*em: prior= dirichlet b= 0.01 maxiter= 50*

*distance= 1e-05*

*data: n= 101600 N= 229 shuffle= -1*

*strands: + -*

*sample: seed= 0 ctfrac= -1 maxwords= -1*

*Letter frequencies in dataset:*

*A 0.285 C 0.215 G 0.215 T 0.285*

*Background letter frequencies (from dataset with add-one prior applied):*

*A 0.285 C 0.215 G 0.215 T 0.285*

*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*MOTIF 1 MEME width = 30 sites = 49 llr = 1337 E-value = 2.7e-307*

*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*--------------------------------------------------------------------------------*

*Motif 1 Description*

*--------------------------------------------------------------------------------*

*Simplified A :::52::::::785:3:22111a::7:8::*

*pos.-specific C 9:71:aa:aa7::11:27::1::112822:*

*probability G :::38::::::3244:8::989::::::87*

*matrix T 1a3::::a::2:::47:18:1::99:2::2*

*bits 2.2 \**

*2.0 \*\* \*\**

*1.8 \*\* \*\*\*\*\* \* \*\**

*1.6 \*\* \*\*\*\*\* \* \*\**

*Relative 1.3 \*\*\* \*\*\*\*\*\* \* \* \*\*\*\* \* \**

*Entropy 1.1 \*\*\* \*\*\*\*\*\*\* \* \* \*\*\*\* \* \*\**

*(39.4 bits) 0.9 \*\*\* \*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*0.7 \*\*\* \*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*0.4 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*0.2 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*0.0 ------------------------------*

*Multilevel CTCAGCCTCCCAAAGTGCTGGGATTACAGG*

*consensus TGA TGGGTA A CT CT*

*sequence*

*--------------------------------------------------------------------------------*

*--------------------------------------------------------------------------------*

*Motif 1 sites sorted by position p-value*

*--------------------------------------------------------------------------------*

*Sequence name Strand Start P-value Site*

*------------- ------ ----- --------- ------------------------------*

*None\_14 + 138 8.52e-19 atccacctgc ctcagcctcccaaagtgctgggattacagg catgagccac*

*None\_59 - 793 8.52e-19 ATCTGTCTTC CTCAGCCTCCCAAAGTGCTGGGATTACAGG TATGAGCCAC*

*None\_93 + 631 1.50e-18 atccgcccgc ctcggcctcccaaagtgctgggattacagg tgtgagccac*

*…*

*--------------------------------------------------------------------------------*

*--------------------------------------------------------------------------------*

*Motif 1 block diagrams*

*--------------------------------------------------------------------------------*

*SEQUENCE NAME POSITION P-VALUE MOTIF DIAGRAM*

*------------- ---------------- -------------*

*None\_3 2.6e-14 175\_[-1]\_3*

*None\_232 2.6e-14 176\_[-1]\_2*

*None\_156 2.6e-14 128\_[+1]\_50*

*None\_55 2.6e-14 140\_[+1]\_38*

*None\_387 2.6e-14 26\_[+1]\_152*

*…*

*--------------------------------------------------------------------------------*

*--------------------------------------------------------------------------------*

*Motif 1 in BLOCKS format*

*--------------------------------------------------------------------------------*

*BL MOTIF 1 width=22 seqs=50*

*None\_3 ( 176) TGGGATTACAGGCGTGAGCCAC 1*

*None\_232 ( 177) TGGGATTACAGGCGTGAGCCAC 1*

*None\_156 ( 129) TGGGATTACAGGCGTGAGCCAC 1*

*None\_55 ( 141) TGGGATTACAGGCGTGAGCCAC 1*

*None\_387 ( 27) TGGGATTACAGGCGTGAGCCAC 1*

*None\_150 ( 79) TGGGATTACAGGCGTGAGCCAC 1*

*None\_34 ( 93) TGGGATTACAGGCGTGAGCCAC 1*

*…*

*--------------------------------------------------------------------------------*

*--------------------------------------------------------------------------------*

*Motif 1 position-specific scoring matrix*

*--------------------------------------------------------------------------------*

*log-odds matrix: alength= 4 w= 22 n= 90932 bayes= 11.3181 E= 4.4e-250*

*-1229 -1229 -1229 181*

*-225 -1229 210 -383*

*-225 -1229 213 -1229*

*-283 -1229 213 -383*

*…*

*--------------------------------------------------------------------------------*

*--------------------------------------------------------------------------------*

*Motif 1 position-specific probability matrix*

*--------------------------------------------------------------------------------*

*letter-probability matrix: alength= 4 w= 22 nsites= 50 E= 4.4e-250*

*0.000000 0.000000 0.000000 1.000000*

*0.060000 0.000000 0.920000 0.020000*

*0.060000 0.000000 0.940000 0.000000*

*0.040000 0.000000 0.940000 0.020000*

*--------------------------------------------------------------------------------*

*--------------------------------------------------------------------------------*

*Motif 1 regular expression*

*--------------------------------------------------------------------------------*

*TGGGATTACAGG[CT][GA]TG[AC]GCCAC*

*--------------------------------------------------------------------------------*

For each seed motif m in the meme/dreme motif file (ordered by decreasing significance), it finds in the selfmatches if any overlap with any other motif with q-value < 0.05 and groups any motifs complying with this together with the motif m. After this pass it performs another pass where it tries to cluster groups with weaker (q-value < 0.10) similarity. It reports the seed motif m from each group (the seed will always be the more significant motif from runMeme/runRandDreme).

meme.dir/enhancers\_intersect\_overexpressed\_genes\_id\_200bp.meme.self\_matches, meme.dir/enhancers\_intersect\_overexpressed\_genes\_id\_200bp.meme -> meme.dir/enhancers\_intersect\_overexpressed\_genes\_id\_200bp.meme.seeds

# compareRandDremeTracks/compareMemeTracks:

Does the same as memeSelfMatches/randDremeSelfMatches comparing all the clustered seeds (after selfMatches and clustering) for all files in a certain method (eg. All files from randDreme) between each other, both ways: file1 with file 2 and file2 with file1.

# loadMotifSequenceComposition:

Gets all the fasta files and appends them to the database, creating a new table for each:

Select \* from enhancers\_intersect\_overexpressed\_genes\_id\_200bp\_meme\_motifseq\_stats LIMIT 5;

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| id | nA | nAT | nC | nG | nGC | nN | nT | nUnk | pA | pAT | pC | pG | pGC | pN | pT |
| None\_201 chr21:43305200-43305400 | 37 | 72 | 66 | 62 | 128 | 0 | 35 | 0 | 0.185 | 0.36 | 0.33 | 0.31 | 0.64 | 0 | 0.175 |
| None\_32 chr1:183843600-183843800 | 71 | 139 | 27 | 34 | 61 | 0 | 68 | 0 | 0.355 | 0.695 | 0.135 | 0.17 | 0.305 | 0 | 0.34 |
| None\_61 chr11:65263200-65263400 | 29 | 101 | 61 | 38 | 99 | 0 | 72 | 0 | 0.145 | 0.505 | 0.305 | 0.19 | 0.495 | 0 | 0.36 |
| None\_46 chr10:21888600-21888800 | 74 | 118 | 28 | 54 | 82 | 0 | 44 | 0 | 0.37 | 0.59 | 0.14 | 0.27 | 0.41 | 0 | 0.22 |
| None\_467 chr7:116225000-116225200 | 63 | 133 | 31 | 36 | 67 | 0 | 70 | 0 | 0.315 | 0.665 | 0.155 | 0.18 | 0.335 | 0 | 0.35 |

nX: Number of X bases (GCTA).

pX: Percentage of X.

# runTomTom

For meme tracks and dreme tracks independently, uses the result from getMemeSeeds/getRandDremeSeeds on a per track basis: for one track gets the most significant seeds for each cluster and then looks for these seeds in the databases of known motifs provided.

Note, this doesn’t use the result from comparing and getting seeds between tracks (result from compareRandDremeTracks/compareMemeTracks).

Differently to the methods memeSelfMatches/randDremeSelfMatches and compareMemeTracks/compareRandDremeTracks, produces the results (see output files in [here](#_memeSelfMatches/randDremeSelfMatche)) in a temporary directory and then moves all the results to the directory specified in “params[exportdir]”**/tomtom** (Eg./mnt/fastdata/mbp15ja/PC\_MM\_enhancers\_raw\_data\_RNA\_pipeline\_intervals/export/tomtom/). It copies the file tomtom.txt, which contains a simplified, text-only version of the output. (See -text, below, for the text output format.) into:

meme.dir(or rand\_dreme.dir)/track.meme(or dreme).tomtom

**Order of the results in Tomtom:**

In the “Matches” section of the HTML file, the results are ordered first alphabetically from the seed motif letters, thenby decreasing significance of enrichment of the meme seed (**seed enrichment in Tomtom used database, not original enrichment of seed in the original regions**).

**To also get the enrichment in the original regions, the task “outputTomTomWithMotifEnrichment” must be used.**

# Sample run

Starting with two bed files: enhancers\_intersect\_overexpressed\_genes\_id\_200bp.bed.gz and enhancers\_intersect\_overexpressed\_genes\_id.bed.gz

enhancers\_intersect\_overexpressed\_genes\_id\_200bp.bed.gz contains the same regions as enhancers\_intersect\_overexpressed\_genes\_id.bed.gz but split into 200bp windows.

**runRandDreme** **(motifs commonly found in the regions) merged:**

# Word RC Word Pos Neg P-value E-value

BEST GTARTCCCA TGGGAYTAC 39 0 3.1e-013 6.4e-007

letter-probability matrix: alength= 4 w= 9 nsites= 43 E= 6.4e-007

# BEST AAAAAHAAA TTTDTTTTT 66 14 4.5e-011 9.3e-005

letter-probability matrix: alength= 4 w= 9 nsites= 108 E= 9.3e-005

# BEST AAATABA TVTATTT 97 41 7.6e-009 1.5e-002

letter-probability matrix: alength= 4 w= 7 nsites= 141 E= 1.5e-002

**runRandDreme (motifs commonly found in the regions) 200bp:**

# Word RC Word Pos Neg P-value E-value

# BEST AAAAAWAMA TKTWTTTTT 87 15 5.0e-015 9.7e-009

letter-probability matrix: alength= 4 w= 9 nsites= 109 E= 9.7e-009

# BEST GTARTCCCA TGGGAYTAC 40 0 4.1e-013 7.8e-007

letter-probability matrix: alength= 4 w= 9 nsites= 42 E= 7.8e-007

# BEST GSCCAGGM KCCTGGSC 61 9 1.6e-011 3.0e-005

letter-probability matrix: alength= 4 w= 8 nsites= 69 E= 3.0e-005

# BEST CCTCCCR YGGGAGG 73 20 3.2e-009 6.0e-003

letter-probability matrix: alength= 4 w= 7 nsites= 88 E= 6.0e-003

# BEST GTGGCWCA TGWGCCAC 37 3 5.5e-009 1.0e-002

letter-probability matrix: alength= 4 w= 8 nsites= 37 E= 1.0e-002

**randDremeSelfMatches** **(motifs commonly found in the regions matching with themselves) merged:**

Query ID Target ID Optimal offset p-value E-value q-value Overlap Query consensus Target consensus Orientation

GTARTCCCA GTARTCCCA 0 2.75188e-08 8.25565e-08 1.65113e-07 9 GTAATCCCA GTAATCCCA +

AAAAAHAAA AAAAAHAAA 0 1.12844e-05 3.38532e-05 6.77065e-05 9 AAAAAAAAA AAAAAAAAA +

AAATABA AAATABA 0 8.70694e-05 0.000261208 0.000522416 7 AAATATA AAATATA +

**randDremeSelfMatches (motifs commonly found in the regions matching with themselves) 200bp:**

Query ID Target ID Optimal offset p-value E-value q-value Overlap Query consensus Target consensus Orientation

AAAAAWAMA AAAAAWAMA 0 2.03866e-09 1.01933e-08 2.03866e-08 9 AAAAAAAAA AAAAAAAAA +

GTARTCCCA GTARTCCCA 0 1.5744e-07 7.87199e-07 1.5744e-06 9 GTAATCCCA GTAATCCCA +

GSCCAGGM GSCCAGGM 0 3.63395e-08 1.81697e-07 3.63395e-07 8 GCCCAGGC GCCCAGGC +

CCTCCCR CCTCCCR 0 5.95966e-06 2.97983e-05 5.95966e-05 7 CCTCCCA CCTCCCA +

GTGGCWCA GTGGCWCA 0 5.45092e-07 2.72546e-06 5.45092e-06 8 GTGGCTCA GTGGCTCA +

**getRandDremeSeeds** **(grouping of motifs commonly found in the regions matching with themselves in terms of similarity, more significant motifs discovered in runRandDreme reported first) merged:**

MOTIF GTARTCCCA DREME

letter-probability matrix: alength= 4 w= 9 nsites= 43 E= 6.4e-007 nClustered= 1 totalHits= 43

MOTIF AAAAAHAAA DREME

letter-probability matrix: alength= 4 w= 9 nsites= 108 E= 9.3e-005 nClustered= 1 totalHits= 108

MOTIF AAATABA DREME

letter-probability matrix: alength= 4 w= 7 nsites= 141 E= 1.5e-002 nClustered= 1 totalHits= 141

**getRandDremeSeeds (grouping of motifs commonly found in the regions matching with themselves in terms of similarity, more significant motifs discovered in runRandDreme reported first) 200bp:**

MOTIF AAAAAWAMA DREME

letter-probability matrix: alength= 4 w= 9 nsites= 109 E= 9.7e-009 nClustered= 1 totalHits= 109

MOTIF GTARTCCCA DREME

letter-probability matrix: alength= 4 w= 9 nsites= 42 E= 7.8e-007 nClustered= 1 totalHits= 42

MOTIF GSCCAGGM DREME

letter-probability matrix: alength= 4 w= 8 nsites= 69 E= 3.0e-005 nClustered= 1 totalHits= 69

MOTIF CCTCCCR DREME

letter-probability matrix: alength= 4 w= 7 nsites= 88 E= 6.0e-003 nClustered= 1 totalHits= 88

MOTIF GTGGCWCA DREME

letter-probability matrix: alength= 4 w= 8 nsites= 37 E= 1.0e-002 nClustered= 1 totalHits= 37

**compareRandDremeTracks:** **for the same method (RandDreme) compare the seeds found in the previous step among the files,** **order of comparison: merged to 200bp:**

Query ID Target ID Optimal offset p-value E-value q-value Overlap Query consensus Target consensus Orientation

GTARTCCCA GTARTCCCA 0 1.5744e-07 7.87199e-07 1.5744e-06 9 GTAATCCCA GTAATCCCA +

AAAAAHAAA AAAAAWAMA 0 8.99048e-07 4.49524e-06 8.99048e-06 9 AAAAAAAAA AAAAAAAAA +

AAATABA AAAAAWAMA 2 0.00115438 0.00577189 0.0115438 7 AAATATA AAAAAAAAA +

**compareRandDremeTracks for the same method (RandDreme) compare the seeds found in the previous step among the files, order of comparison: 200bp to merged:**

Query ID Target ID Optimal offset p-value E-value q-value Overlap Query consensus Target consensus Orientation

AAAAAWAMA AAAAAHAAA 0 1.25382e-05 3.76147e-05 7.52294e-05 9 AAAAAAAAA AAAAAAAAA +

GTARTCCCA GTARTCCCA 0 2.75188e-08 8.25565e-08 1.65113e-07 9 GTAATCCCA GTAATCCCA +

CCTCCCR GTARTCCCA 2 0.00197465 0.00592395 0.0118479 7 CCTCCCA GTAATCCCA +