MOTIFSIM Manual

Version 1.0

Ngoc Tam L. Tran and Chun-Hsi Huang

Department of Computer Science and Engineering, University of Connecticut

Storrs, CT 06269, USA

Introduction

MOTIFSIM is a software tool for detecting similarity in multiple motif datasets. It accepts nine different motif input formats and outputs the results in two text files. The tool combines all input datasets into one list and performs pair-wise comparisons on the entire list. MOTIFSIM converts all input motifs into position specific probability matrices for comparisons. The tool reports global significant motifs, global and local significant motifs, as well as best matches for each motif in the combined list or in a single dataset. MOTIFSIM is written in C++ and OpenMP for multithreaded utilization. It can be downloaded at http://biogrid-head.engr.uconn.edu/motifsim/ for Windows and Linux environments.

How to Use MOTIFSIM?

Motif Input Format

MOTIFSIM accepts nine different motif input formats, which are listed in the table below.

Input Format	**Example**	**Restriction**

TRANSFAC	`NA Test1`
	`XX`
	`DE Test1`
	`XX`
	`P0 A C G T`
	`01 4 36 5 5 C`
	`02 39 0 9 2 A`
	`03 10 30 0 10 C`
	`04 2 1 38 9 G`
	`05 4 3 5 38 T`
	`06 9 0 31 10 G`
		One empty line must be present to separate two motifs. Space or tab can be used to separate matrix's elements.

	07 4 6 21 10 G 08 1 9 10 30 T XX NA Test2 XX DE Test2 XX P0 A C G T 01 0 40 10 0 C 02 38 0 10 2 A 03 0 30 10 10 C 04 2 11 28 9 G 05 9 3 10 28 T XX	
TRANSFAC-like	DE Test1 01 4 31 5 5 C 02 29 0 9 2 A 03 0 30 0 10 C 04 2 1 28 9 G 05 4 3 5 28 T 06 9 0 31 0 G XX DE Test2 01 0 40 10 0 C 02 38 0 10 2 A 03 0 30 10 10 C 04 2 11 28 9 G 05 9 3 10 28 T 06 50 0 0 0 A 07 0 50 0 0 C 08 0 0 50 0 G 09 0 0 0 50 T XX	Columns 2, 3, 4, and 5 in the matrix represent A, C, G, and T values respectively. One empty line must be present to separate two motifs. Space or tab can be used to separate matrix’s elements.
	DE sscCCCGCGcs 1 5 15 9 5 2 4 18 10 2 3 0 23 8 3 4 1 29 4 0 5 0 28 6 0 6 0 27 7 0 7 0 0 34 0 8 0 34 0 0 9 0 2 32 0 10 4 16 8 6 11 0 11 18 5 XX DE atactttggc 1 1 0 0 0 2 0 0 0 1 3 1 0 0 0	Columns 2, 3, 4, and 5 in the matrix represent A, C, G, and T values respectively. One empty line must be present to separate two motifs. Space or tab can be used to separate matrix’s elements.

	4 0 1 0 0 5 0 0 0 1 6 0 0 0 1 7 0 0 0 1 8 0 0 1 0 9 0 0 1 0 10 0 1 0 0 XX	
PSSM	>TFW3 73 81 407 61 44 578 0 0 485 65 0 72 0 570 52 0 79 0 0 543 0 0 622 0 >TFW1 0 0 1 39 0 0 0 40 4 1 33 2 6 25 2 7 7 2 25 6 2 33 1 4 40 0 0 0 39 1 0 0	Columns 1, 2, 3, and 4 in the matrix represent A, C, G, and T values respectively. One empty line must be present to separate two motifs. Space or tab can be used to separate matrix's elements.
Jaspar	>NR4A2 A [8 13 0 3 2 0 14 3] C [1 0 0 0 2 13 0 8] G [3 1 13 11 0 0 0 2] T [2 0 1 0 10 1 0 1] >RORA_1 A [15 9 6 11 21 0 0 0 0 25] C [1 1 12 2 0 0 0 0 25 0] G [2 0 4 5 4 25 25 0 0 0] T [7 15 3 7 0 0 0 25 0 0]	One empty line must be present to separate two motifs. Space or tab can be used to separate matrix's elements.
MEME's output	-- Motif 1 position-specific probability matrix -- letter-probability matrix: alength= 4 w= 11 nsites= 142 E= 6.0e-015 0.000000 0.598592 0.176056 0.225352 0.000000 0.626761 0.000000 0.373239 0.000000 0.000000 0.000000 1.000000 0.000000 0.408451 0.514085 0.077465 0.091549 0.823944 0.028169 0.056338 0.133803 0.690141 0.000000 0.176056 0.042254 0.281690 0.000000 0.676056 0.007042 0.683099 0.197183 0.112676	One empty line must be present to separate two motifs. Space or tab can be used to separate matrix's elements.

	0.197183 0.000000 0.000000 0.802817 0.000000 0.084507 0.690141 0.225352 0.091549 0.605634 0.169014 0.133803 -- -- Motif 2 position-specific probability matrix -- -- letter-probability matrix: alength= 4 w= 14 nsites= 24 E= 6.3e-010 0.833333 0.000000 0.166667 0.000000 0.000000 1.000000 0.000000 0.000000 0.875000 0.000000 0.000000 0.125000 0.083333 0.875000 0.041667 0.000000 0.958333 0.000000 0.000000 0.041667 0.125000 0.875000 0.000000 0.000000 0.833333 0.166667 0.000000 0.000000 0.000000 0.750000 0.000000 0.250000 0.666667 0.000000 0.208333 0.125000 0.000000 0.833333 0.041667 0.125000 0.791667 0.125000 0.083333 0.000000 0.000000 1.000000 0.000000 0.000000 0.708333 0.000000 0.083333 0.208333 0.000000 0.958333 0.000000 0.041667	
Consensus sequence	>C001 CYCYYSHGGCCASMAGAGGGCRCYAGATCCCCT >C002 WWWWWWWWWWWWAAAAAAAAAWWAAWWWWW >C003 VTGYRYRYACACACACAYRCAYRYR >C004 SNSVCCCSBCCCCCSCCCCCSSY >C005 SCSCSSSSSCSSCSCCSSSSCCSSSSSSC >C006 TWWWAAAAAAWWAAAAWWAAAAAAAAAA >C007 MYVGAGGCCAGAAGAGGGCAYCAGATYCCHT	Motif is in IUPAC format. One empty line must be present to separate two motifs.
Sequence Alignment	>Test1 GATACGTGGCAAAACCCTGGG GCCACGT-CCGGGAACCTGGG CGCATGTGCACCAATTACACC ACGACGTGTTCCCAAATTTTT CACACGTGCCCCCCAAGTTTG GGGGGTTACACCCTTTTAAAA	One empty line must be present to separate two motifs.

	CCAAGTTTAAGGGGTTTTGGA AAACCGGTTAAAACCTTGCGC >Test2 CCCAATAGCTTTT-TTTTTTAAACCCCC-CC GGGTGTGCGCGACCACCAAAATTTTAAAAAA AAACCCTTTGGGCCCGGGTTAAACCCCGGGG TTTTTTCCCAAACCCAAAGGGTTTTTCGCCC CACAAAAACCGGTTTTTTGCCGCGCCCCAAA CCAAAAAACCCTT-TTTTCCCAAAAGGGGGG CACAAACACCCCCCCCCAAAATTT-TGGGCG	

Matrix (Horizontal)	7 10 6 13 4 21 0 22
	1 4 3 4 10 0 2 1
	4 2 6 4 2 2 0 0
	11 7 8 2 7 0 21 0
	
	27 0 1 27 27 20
	0 0 9 0 0 0
	0 0 0 0 0 1
	0 27 17 0 0 6
		Rows 1, 2, 3, and 4 represent A, C, G, and T values respectively. One empty line must be present to separate two motifs. Space or tab can be used to separate matrix's elements.

Matrix (Vertical)	0.450000 0.250000 0.000000 0.300000
	0.800000 0.000000 0.000000 0.200000
	0.850000 0.000000 0.150000 0.000000
	1.000000 0.000000 0.000000 0.000000
	0.750000 0.000000 0.250000 0.000000
	0.400000 0.300000 0.050000 0.250000
	0.850000 0.100000 0.000000 0.050000
	0.850000 0.150000 0.000000 0.000000
	1.000000 0.000000 0.000000 0.000000
	0.500000 0.100000 0.400000 0.000000
	
	0.000000 0.812500 0.125000 0.062500
	0.375000 0.000000 0.000000 0.625000
	0.062500 0.000000 0.937500 0.000000
	0.562500 0.125000 0.187500 0.125000
	0.000000 0.000000 1.000000 0.000000
	0.062500 0.937500 0.000000 0.000000
		Columns 1, 2, 3, and 4 represent A, C, G, and T values respectively.
		One empty line must be present to separate two motifs. Space or tab can be used to separate matrix's elements.

Running MOTIFSIM

The tool can be run by command line on Windows and Linux. An example for running MOTIFSIM on Windows for comparing two motif datasets, which are included with the tool, is below.

```
C:\Path> motifsim-v1-0-wins64
Please, enter number of files to read (must be > 0):
2
Please, enter number of best matches (must be > 0 and <= 50):
5
```

```
Please, select a cutoff for similarity (>= 0.5, >= 0.6, >= 0.7, >= 0.75, >= 0.8, >=
0.85, >= 0.9):
0.75
Please, enter number of threads (must be >= 1):
1
Maximum number of threads available on your machine is 1.
This is the maximum number of threads can be allocated to run this program.

Please, enter input file's location (full path, for example, C:\MyDocuments\ for
Windows and /home/MyFolder/ for Linux):

C:\Enter\Location\of\Input\Files\

Enter input file names and formats (for example: 1). See user manual for each format:

(1) TRANSFAC
(2) TRANSFAC-like
(3) PSSM
(4) Jaspar
(5) MEME output
(6) Consensus sequence
(7) Sequence Alignment
(8) Matrices (Horizonal)
(9) Matrices (Vertical)

Please, enter file name (in text format .txt, name without spaces):
PScanChIP_DM05.txt
Please, enter file format:
4
Please, enter file name (in text format .txt, name without spaces):
W-ChIPMotifs_DM05.txt
Please, enter file format:
3
Please, enter output file's location (full path, for example, C:\MyDocuments\ for
Windows and /home/MyFolder/ for Linux):

C:\Location\To\Save\Output\Files\

Your input files, types, and counts are:
File Name                      File Type      Count of Motifs     Dataset #
PScanChIP_DM05.txt             4              16                  1
W-ChIPMotifs_DM05.txt          3              11                  2

Your output files have been saved in C:\Location\To\Save\Output\Files\
```

Input Parameters

The required input parameters are listed in the table below.

Parameter	Description
Number of files	Number of motif datasets that need to be compared. It must be > 0. MOTIFSIM can also compare motifs in a single dataset.
Number of best matches	Users are required to select the number of best matched motifs. This value is currently limited to ≤ 50. The number of

	best matches is the number of motifs that are most similar to motif *i* (*i* from 1 to *m*) in a combined motif list *M*. This threshold is used for selecting the numbers of most similar motifs to motif *i* and report them in the result files. These best matched motifs are listed in order of similarity with the most similar one on the top of the list.
Similarity cutoff	Cutoff values are >= 0.5, >= 0.6, >= 0.7, >= 0.75, >= 0.8, >= 0.85, and >= 0.9. A value >= 0.75 indicates a match of 75 % or greater between two motifs. We suggest to set a cutoff >= 0.75 as this value showed a good start for threshold in our case studies. If a higher cutoff value is set, fewer similar motifs will be returned in the results. However, these motifs are much more similar to the motif being compared.
Number of threads	Number of threads to run the tool. It must be between 1 and the maximum number of threads available on the machine.
Input file's location	Full path to input file location (for example, for example, C:\MyDocuments\ for Windows and /home/MyFolder/ for Linux).
Enter input file name	File name in text format without space including extension .txt
Enter file format	Select a number represents a format listed in the menu below. (1) TRANSFAC (2) TRANSFAC-like (3) PSSM (4) Jaspar (5) MEME output (6) Consensus sequence (7) Sequence Alignment (8) Matrices (Horizonal) (9) Matrices (Vertical)
Enter output file's location	Full path (for example, C:\MyDocuments\ for Windows and /home/MyFolder/ for Linux).

Output files

MOTIFSIM outputs the results in two text files namely `Results.txt` and `Results_Without_Motif_Details.txt.` The former includes motif's detail in position specific probability matrices. The latter does not. Each result file includes two sections: Input and Results. The Input section contains input parameters entered. The Results section includes three subsections: (1) global significant motifs, (2) global and local significant motifs, and (3) best matches for each motif. The number of significant motifs as well as the number of best matches returned by the tool are selected by the users when entering the cutoff for best matches. Other output information can be found in the table below.

Output Information	**Description**
Dataset #	Dataset is numbered from 1, 2, 3, … , *n* in the order they are entered.
Motif ID	Each motif in the combined list is assigned a unique ID, which is an integer from 1, 2, 3, … *n*, in the order of the dataset enters.
Motif Name	Motif name in the input file if available.
Matching Format of First Motif	Matching format of the *first* motif in the comparison. The format can be the original motif or its reverse complement.
Matching Format of Second Motif	Matching format of the *second* motif in the comparison. The format can be the original motif or its reverse complement.
Direction	Matching can be in forward or backward direction.
Position #	Matching position number. Starting at position 1 on the top if it is in forward direction or at the bottom if it is in a backward direction.
# of Overlap	The number of overlapping columns when matching two motifs.
Similarity Score	This score is described in our algorithm.

Memory Use

MOTIFSIM requires over 2G of RAM for comparing more than 250 motifs.