MOTIFSIM Command-line Manual

Version 2.2

Ngoc Tam L. Tran and Chun-Hsi Huang

Department of Computer Science and Engineering, University of Connecticut

Storrs, CT 06269, USA

Introduction

MOTIFSIM command-line is a software tool for detecting similarity in a single or multiple DNA motif datasets. It accepts nine different input formats from several motif finders and generates multiple results. 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 the global significant motifs, the global and local significant motifs, as well as best matches for each motif in a combined list or in a single dataset.

Version 2.2 allows combining similar motifs discovered in the results. It also allows comparing the global significant motifs as well as every motif in the combined list with motif database. In addition, the relationship between motifs can be visualized via motif tree.

Version 2.2 for Linux platform provides more options for generating output file format. The results can be generated in HTML and PDF formats.

MOTIFSIM is written in C++ and OpenMP for multithreaded utilization. It can be downloaded at http://motifsim.org for Windows and Linux environments.

How to Use MOTIFSIM?

Motif Input Format

MOTIFSIM accepts nine different motif input formats 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```
		One empty line must be present to separate two motifs. Space or tab can be used to separate matrix's elements.

| | ```
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
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
``` | 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. |

| | ```
1 1 0 0 0
2 0 0 0 1
3 1 0 0 0
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
 0.197183  0.000000  0.000000  0.802817
 0.000000  0.084507  0.690141  0.225352
 0.091549  0.605634  0.169014  0.133803
``` | One empty line must be present to separate two motifs. Space or tab can be used to separate matrix's elements. |

| | ```
----------------------------------------
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
TWWWAAAAAWWAAAAWWAAAAAAAAAA

>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
CCAAGTTTAAGGGGTTTTGGA
AAACCGGTTAAAACCTTGCGC

>Test2
CCCAATAGCTTTT-TTTTTTAAACCCCC-CC
GGGTGTGCGCGACCACCAAAATTTTAAAAAA
AAACCCTTTGGGCCCGGGTTAAACCCCGGGG
TTTTTTCCCAAACCCAAAGGGTTTTTCGCCC
CACAAAAACCGGTTTTTTGCCGCGCCCCAAA
CCAAAAAACCCTT-TTTTCCCAAAAGGGGGG
``` | One empty line must be present to separate two motifs. |

	CACAAACACCCCCCCCCCAAAATTT-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

MOTIFSIM can be run on Windows and Linux environments. Below are the examples for running MOTIFSIM 2.2. The examples show comparing two motif datasets, which are included with the tool.

Running MOTIFSIM 2.2 on Windows

```
H:\06-18-18>motifsim_v2-2-wins64.exe
Please, enter number of files to read (must be > 0):
2
Please, enter number of top significant motifs (must be > 0 and <= 50):
10
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):
H:\06-18-18\dataset\

Enter input file names and formats (for example: 1). See the 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)
(10) Unspecified

Please, enter file name (in text format .txt, name without spaces):
PScanChIP_DM05.txt
Please, enter file format:
10
Please, enter file name (in text format .txt, name without spaces):
W-ChIPMotifs_DM05.txt
Please, enter file format:
10


Enter database name by selecting a number in the list below:

(1) Jaspar 2016 (All)
(2) Jaspar 2016 (Fungi)
(3) Jaspar 2016 (Insects)
(4) Jaspar 2016 (Nematodes)
(5) Jaspar 2016 (Plants)
(6) Jaspar 2016 (Urochordates)
(7) Jaspar 2016 (Vertebrates)
(8) Transfac (Free version)
(9) UniPROBE (Human)
(10) UniPROBE (Mouse)
(11) UniPROBE (Parasite)
(12) UniPROBE (Worm)
(13) UniPROBE (Yeast)
(14) None


10
Would you like to generate motif tree? (Y or N):
Y
Would you like to combine similar motifs? (Y or N):
Y
Please, enter an output file type (Global-Only, All):
All
please, enter output file's location (full path, for example, C:\MyDocuments\ for Windows and
/home/MyFolder/ for Linux):
H:\06-18-18\result\

Reading input ...

Your input files, types, and counts are:

File Name                   Count of Motifs     Dataset #
PScanChIP_DM05.txt          16                  1
W-ChIPMotifs_DM05.txt       11                  2

Processing input ...


Your output files have been saved in H:\06-18-18\result\

Runtime: 298.873 seconds.
```

Running MOTIFSIM 2.2 on Linux

```
user@user-Inspiron-5535:~/motifsim_commandline_v2.2$ ./motifsim_v2-2-linux64
Please, enter number of files to read (must be > 0):
2
Please, enter number of top significant motifs (must be > 0 and <= 50):
```

```
10
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 4.
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):
./dataset/

Enter input file names and formats (for example: 1). See the 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)
(10) Unspecified

Please, enter file name (in text format .txt, name without spaces):
PScanChIP_DM05.txt
Please, enter file format:
10
Please, enter file name (in text format .txt, name without spaces):
W-ChIPMotifs_DM05.txt
Please, enter file format:
10

Enter database name by selecting a number in the list below:

(1)  Jaspar 2016 (All)
(2)  Jaspar 2016 (Fungi)
(3)  Jaspar 2016 (Insects)
(4)  Jaspar 2016 (Nematodes)
(5)  Jaspar 2016 (Plants)
(6)  Jaspar 2016 (Urochordates)
(7)  Jaspar 2016 (Vertebrates)
(8)  Transfac (Free version)
(9)  UniPROBE (Human)
(10) UniPROBE (Mouse)
(11) UniPROBE (Parasite)
(12) UniPROBE (Worm)
(13) UniPROBE (Yeast)
(14) None

10
Would you like to generate motif tree? (Y or N):
Y
Would you like to combine similar motifs? (Y or N):
Y
Please, enter an output file type (Global-Only, All):
All
Please, enter an output file format (Text, HTML, PDF, All):
All
please, enter output file's location (full path, for example, C:\MyDocuments\ for Windows and
/home/MyFolder/ for Linux):
./result/

Reading input ...

Your input files, types, and counts are:

File Name                       Count of Motifs     Dataset #
```

```
PScanChIP_DM05.txt              16                    1
W-ChIPMotifs_DM05.txt           11                    2

Processing input ...


Your output files have been saved in ./result/

Runtime: 668.135 seconds.
```

Input Parameters

The input parameters and description are listed in the table below.

Input parameter	Description
Number of files	The number of motif datasets for comparison. It must be ≥ 1. MOTIFSIM can also compare motifs in a single dataset.
Number of top significant motifs	The tool currently supports a maximum of 50 top significant motifs. This is a cutoff for the number of top global significant motifs as well as the number of top global and local significant motifs generated in the results.
Number of best matches	The tool currently supports a maximum of 50 best matches. 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 value is used for selecting the number of most similar motifs to motif *i* and report them in the results. These best matched motifs are listed in order of similarity with the most similar one on the top of the list.
Similarity cutoff	Currently, the tool supports cutoff values **≥ 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 using a cutoff **≥ 0.75** as this value shows a good threshold in our case studies. If a higher cutoff value is used, fewer similar motifs are generated in the results. However, these motifs are much more similar to the motif being compared.
Number of threads	The number of threads for running the tool. It must be between 1 and the maximum number of threads available on the user's machine.
Input file's location	Full path to input file location (for example, **C:\MyDocuments** for Windows and **/home/MyFolder/** for Linux).
Input file name	File name *without space including extension*. File format must be in **text** (**.txt**).

Input file format	From version 2.0 and above, the tool can automatically detect the input format for each motif. Motifs in different formats can be mixed in one input file. Users can select the input format using an option from 1 through 9 below or by letting the tool to detect the input format for each motif by choosing option 10 in the list.
	
	`(1) TRANSFAC`
	`(2) TRANSFAC-like`
	`(3) PSSM`
	`(4) Jaspar`
	`(5) MEME output`
	`(6) Consensus sequence`
	`(7) Sequence Alignment`
	`(8) Matrices (Horizonal)`
	`(9) Matrices (Vertical)`
	`(10) Unspecified`

Database name	Users can select a motif database for comparing with the global significant motifs, the global and local significant motifs, as well as every motif in the combined list. Currently, the tool supports Jaspar 2016 [1], Transfac [2] (free version), and UniPROBE [3] databases. The list of database names is below. If users do not want to compare motifs with a motif database, they can choose the last option in the list.
	
	`(1) Jaspar 2016 (All)`
	`(2) Jaspar 2016 (Fungi)`
	`(3) Jaspar 2016 (Insects)`
	`(4) Jaspar 2016 (Nematodes)`
	`(5) Jaspar 2016 (Plants)`
	`(6) Jaspar 2016 (Urochordates)`
	`(7) Jaspar 2016 (Vertebrates)`
	`(8) Transfac (Free version)`
	`(9) UniPROBE (Human)`
	`(10) UniPROBE (Mouse)`
	`(11) UniPROBE (Parasite)`
	`(12) UniPROBE (Worm)`
	`(13) UniPROBE (Yeast)`
	`(14) None`

Motif tree	The motif tree can be generated for the global significant motifs as well as for the entire combined motif list. The tree is built using *hclust* function in R [4]. This function implements the hierarchical clustering algorithm. The tool generates the distance matrix for building the tree. This matrix contains best similarity scores between motifs. Users can choose option `Y` for generating the motif tree and `N` if not.
Combine similar motifs	Similar motifs discovered in the results can be combined into new motifs. Two similar motifs can be combined into a new motif if the new motif satisfies the similarity threshold with both of its parents. Users can choose option `Y` for combining similar motifs and `N` if not.
Output file type	Users can choose between `Global-Only` and `All`. The first option generates only the global significant motifs in the results. The second option generates everything and it requires longer time for processing large datasets.

9

Output file format	The tool provides four options for output file format: `Text`, `HTML`, `PDF`, and `All`. The option `All` generates all formats. `HTML`, `PDF`, and `All` require longer time for processing large datasets. For Windows environment, the tool supports `Text` format only.
Output file's location	Full path (for example, **C:\MyDocuments** for Windows and **/home/MyFolder/** for Linux).

Output Files

MOTIFSIM generates multiple results, which can be in different formats for Linux. The result files are described in the table below.

	Output file name (without Job ID prefix)	Description	Supported Platform
1.	input_motifs.txt	This file contains motifs from all datasets. The motifs are in position specific probability matrices. They are in the order of the datasets are entered by the user.	Windows and Linux
2.	HTML_Results.html	The results in HTML format. Motif's logos are included.	Linux
3.	HTML_Results.pdf	The results in PDF format. Motif's logos are included.	Linux
4.	Results.txt	The results in Text format. Motifs are in position specific probability matrices.	Windows and Linux
5.	Results_Without_Motif_Details.txt	The results in Text format. Motif's detail is not included.	Windows and Linux
6.	HTML_Results_Database_Matching.html	The results match with motif database in HTML format. Motif's logos are included.	Linux
7.	HTML_Results_Database_Matching.pdf	The results match with motif database in PDF format. Motif's logos are included.	Linux
8.	Results_Database_Matching.txt	The results match with motif database in Text format. Motifs are in position specific probability matrices.	Windows and Linux
9.	Results_Database_Matching_Without_Motif_Details.txt	The results match with motif database in Text format. Motif's detail is not included.	Windows and Linux
10.	HTML_Global_Matching_Results.html	This file contains only the global significant motifs in HTML format.	Linux
11.	HTML_Global_Matching_Results.pdf	Same as in (10) but the file is in PDF format.	Linux
12.	Results_Global_Matching.txt	Same as in (10) but the file is in Text format.	Windows and Linux

13.	HTML_Results_Database_Matching_Global_Matching.html	This file contains only the global significant motifs that match with a motif database. The results are in HTML format. Motif's logos are included.	Linux
14.	HTML_Results_Database_Matching_Global_Matching.pdf	Same as in (13) but the file is in PDF format.	Linux
15.	Results_Database_Matching_Global_Matching.txt	This file contains only the global significant motifs that match with a motif database. The results are in Text format. Motifs are in position specific probability matrices.	Windows and Linux
16.	Results_Combined_Motifs_Global_Matching.txt	This file contains combined motifs for the global significant motifs and their best matches. The motifs are combined in a pair-wise fashion. The motifs are in IUPAC format. Pair-wise matching information is included.	Windows and Linux
17.	Results_Combined_Motifs_Global_and_Local_Matching.txt	This file contains combined motifs for the global and local significant motifs and their best matches. The motifs are combined in a pair-wise fashion. The motifs are in IUPAC format. Pair-wise matching information is included.	Windows and Linux
18.	Results_Combined_Motifs_Best_Matches.txt	This file contains combined motifs for every motif and its best matches. The motifs are combined in a pair-wise fashion. The motifs are in IUPAC format. Pair-wise matching information is included.	Windows and Linux
19.	Motif_Tree.html	The motif tree demonstrates the relationship between motifs in the entire combined list. The file is in HTML format.	Linux
20.	Global_Matching_Motif_Tree.html	The motif tree demonstrates the relationship between motifs for the global significant motifs and their best matches. The file is in HTML format.	Linux
21.	Motif_Tree.png	Same as in (19) but the file is in PNG image format.	Windows and Linux
22.	Global_Matching_Motif_Tree.png	Same as in (20) but the file is in PNG image format.	Windows and Linux

Each result file has two sections: Input and Results. The Input section lists input parameters entered by the user.

The result files (numbers 2-5 in the table above) have three subsections for: (1) global significant motifs, (2) global and local significant motifs, and (3) best matches for each motif.

The results match with motif database (numbers 6-9 in the table above) show best matches in a motif database for each motif in the combined list.

The global matching result files (numbers 10-12 in the table above) contain only the global significant motifs and their best matches.

The result files (numbers 13-15 in the table above) contain only the global significant motifs and their best matches in a motif database.

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 by the user.
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 a 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 [5].
Alignment	Alignment of two motifs in IUPAC format.

Dependencies

The dependencies are required only for Linux platform. The tool requires installing Prince [6] software package for converting HTML to PDF format. Users can use a free version of Prince for this conversion. To obtain this version and instructions for installing the software, please go to https://www.princexml.com/download/ . The tool also requires WebLogo [7] software package version 3.4, which is available at https://github.com/WebLogo/weblogo/tree/3.4 , for creating motif's logo. Lastly, the tool requires installing `R` and its four packages: `ggplot2`, `ggdendro`, `dendextend`, and `ape`, for drawing the motif tree.

Source Code

The source code in C++ is available on the tool's website for downloading.

Databases

The tool supports Jaspar 2016, Transfac (free version), and UniPROBE databases. These database files are included with the source code zip file for downloading. The `databases` folder must reside in the same location with the executable file.

WebLogo

`weblogo-3.4` is also included with the source code zip file for downloading. The folder `weblogo-3.4` must reside in the same location with the executable file.

Memory Usage

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

References

1. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles**. *Nucleic Acids Research* 2004, **32**:D91-D94.

2. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC®: transcriptional regulation, from patterns to profiles**. *Nucleic Acids Research* 2003, **31(1)**:374-8.

3. Newburger N and Bulyk M. **UniPROBE: an online database of protein binding microarray data on protein–DNA interactions**. *Nucleic Acids Research* 2009, **37**:D77-D82.

4. R Core Team (2016). **R: A language and environment for statistical computing. R Foundation for Statistical Computing,** Vienna, Austria. URL https://www.R-project.org/.

5. Tran NTL, Huang C-H: **MOTIFSIM: A Web Tool for Detecting Similarity in Multiple DNA Motif Datasets**. *BioTechniques* 2015, **59**(1):26-33.

6. **Prince**. http://www.princexml.com/ (2002). Accessed 19 Mar 2016.

7. Crooks GE, Hon G, Chandonia JM, Brenner SE. **WebLogo: A sequence logo generator**. *Genome Research* 2004, **14**:1188-1190.