ScanWM-PL

The program for site search in DNA sequences by score matrices.

The program's brief description.

ScanWM-PL is a program that search for motifs in "+" and "-" strands of DNA using score matrices. The program takes DNA sequences one by one from FASTA file, takes matrices from the score matrices file and annotates DNA sequences by finding motifs (potential sites for binding of transcription factors) in accordance to score matrices. Nucleotide sequences are referred to as motifs (potential sites for binding of transcription factors) if their score is more or equal to "cut-off value" of score matrix; at that the score of sequence is calculated as sum of its nucleotides' score, and the score of a nucleotide in appropriate position is defined in accordance to score matrix. Since ScanWM works with score matrices, elements of which are "log likelihood ratios", the summation is used at sequence score detection.

Algorithm.
In the current version of the program there is no checking for overlapping motifs. Checking for overlapping motifs could be of importance for motifs of those sites, sequences of which can be read similarly (or almost similarly) in both forward and backward orientations.
Definition of the data volumes.

Initially, the program does not know the approximate number of motifs, that can be found in a single sequence using a single score matrix.
For storing motifs the dynamic container is used. If, at a certain step, the number of motifs becomes greater than the current volume of container, then its volume increases by the number of elements, defined by the "increment"-value of the container's volume.
In the current version of the program, the initial and "increment-" volumes of container for motifs are set equal to 100 and 100.
FASTA file.

In the current version of program, the maximal number of symbols in a line of FASTA file = 999.

Format of a file with score matrices

Score matrices in a score matrices file have the following record format:

```
2. AC: RSP00002//OS: Brassica napus /GENE: Oleosin/RE: ABRE-3 /BF: ...

     1430     9.29   10.28   12.76    6.79    1.49

       1       2       3       4       5       6       7       8       9
A     0.96   -2.46    1.12   -2.57   -2.76   -3.49   -3.24   -2.12   -1.15
C    -0.44    1.63   -4.85    1.65   -3.60   -3.47   -3.47   -2.12    1.53
G    -2.55   -2.02   -3.47   -2.72    1.67  -10.16    1.69    1.38   -1.91
T    -2.34   -2.36   -3.29   -2.66   -2.91    1.12   -3.49   -0.37   -2.06
```

Each score matrix takes 10 lines in a file.
The first line - ID-line of a score matrix;
The third line - "line of values" (see below);
The fifth line - score matrix's positions;
The sixth to ninth lines - the score matrix itself (in a format, shown above).
The empty lines: second, fourth and tenth ones.
Format and table-description of "values' lines".

```
 1430     9.29   10.28   12.76    6.79    1.49
```

value (example)	Description
1430	Number of sequences, used to build the score matrix.
9.29	Site's IC
10.28	Average score (*)
12.76	Maximal score (*)
6.79	Minimal score (*)
1.49	Standard deviation (*)

(*) Using the matrix, the scores for sequences, used to build the matrix, are calculated, and average, maximal and minimal scores as well as standard deviation are revealed.
In the current version of ScanWM, if -t: parameter is set to 1, i.e. -t:1, then of all "values' line" numbers the average score and standard deviation (see table) only are used. Other "values' line" numbers are not used, and at preparation of user-defined files with score matrices can be set, for example, to zero.

Format of a file with results of searching for motifs using score matrices

Format of a file with results of searching for motifs using score matrices has a following structure.
In the header, the data on a program version and parameters used for program launch are shown:

```
Program  ScanWM  (Softberry Inc.)

 Search for motifs by Weight Matrixes of Regulatory Elements
 Version 1.2004

 SET of WMs: derived from subsection of REGSITE DB (Plants; version IV)
_____________________________________________________________

 File with QUERY Sequences: TEST_SEQ.seq

 Search PARAMETERS:
     Threshold type                        : 2
     Threshold value                       : 0.90
     Search for motifs on "+" strand       : yes
     Search for motifs on "-" strand       : yes

 NOTE: WM - Weight Matrix of Regulatory Element
       AC - Accession No  of Regulatory Element in a given DB
       OS - Organism/Species
       BF - Binding Factors or One of them

============================================================
```

Further, for each DNA sequence (from designated set), there are located its ID-string and length followed by results of searching for motifs using score matrices: for each of the score matrices, the ID-string and motifs found on "+" and/or "-" strands of DNA are shown;
For each of found motifs, there are shown its sequence, coordinates in "QUERY sequence" and a score, obtained using a score matrix;
Motifs, found on "-" strand, are shown in 5'-3' orientation, and thus, since coordinates are shown relatively to "+" strand (which corresponds to "QUERY sequence"), the first coordinate should be greater then the second one (see example below);
In the end, the total number of motifs, found in a sequence, and the total number of score matrices, used for search, are shown.
Below there is an example of output for a single sequence and a single score matrix (ID-string of a sequence and ID-string of a score matrix are shown incompletely):

```
------------------------------------------------------------

 QUERY: >At4g00860 stress-related ozone-induced protein (OZI1)...
 Length of Query Sequence:        350
```

```
.............................................................
 WM:     228. AC: RSP00231//OS: Arabidopsis thaliana /GENE: AGAMAOUS (AG)...

 Motifs on "+" strand (in DIR orientation):  Found   1

     121  CCAATCT      127       7.73

 Motifs on "-" strand (in INV orientation):  Found   1

     192  CCCATCT      186       6.65

.............................................................
 Totally       2 motifs  of     1 different WMs have been found
-------------------------------------------------------------

If no motifs were found in a sequence, then output for this sequence is displayed
as following:
-------------------------------------------------------------

 QUERY: >At1g04660 68414.t00411 glycine-rich protein
 Length of Query Sequence:        350

.............................................................
 Any Motif not found
-------------------------------------------------------------
```

OUTPUT EXAMPLE

The whole output of ScanWM-PL for some test sequence is shown below.

```
Program  ScanWM  (Softberry Inc.)

 Search for motifs by Weight Matrixes of Regulatory Elements
 Version 1.2004

 SET of WMs: derived from subsection of REGSITE DB (Plants; version IV)
_____________________________________________________________

 File with QUERY Sequences: TEST_SEQ.seq

 Search PARAMETERS:
     Threshold type                        : 2
     Threshold value                       : 0.90
     Search for motifs on "+" strand       : yes
     Search for motifs on "-" strand       : yes

 NOTE: WM - Weight Matrix of Regulatory Element
       AC - Accession No  of Regulatory Element in a given DB
       OS - Organism/Species
       BF - Binding Factors or One of them

=============================================================

 QUERY: >At4g00160 [-300,+50] region of F-box family protein
 Length of Query Sequence:        350

.............................................................
 WM:     >151. AC: RSP00151//OS: tomato, Lycopersicon esculentum /GENE: Lhcb1*1,
Lhcb1*2, Lhca3, Lhca4/RE: CRE, consensus /BF:unknown
```

```
 Motifs on "+" strand (in DIR orientation):  Found   1

      79  CAAGTACATC       88       7.76

.............................................................
 WM:     >174. AC: RSP00174//OS: Phaseolus vulgaris /GENE: beta-phaseolin, or
phas/RE: ATCATC motif /BF:unknown

 Motifs on "+" strand (in DIR orientation):  Found   2

      21  ATCATC       26       7.98
     102  ATCATC      107       7.98

.............................................................
 WM:     >359. AC: RSP00359//OS: barley, Hordeum vulgare /GENE: GCCGAC motif/RE:
HVA1s /BF: HvCBF1

 Motifs on "-" strand (in INV orientation):  Found   1

     103  ATCGAC       98       4.73

.............................................................
 WM:     >707. AC: RSP00707//OS: /GENE: /RE: W-box (consensus 1) /BF: transcription
factors of WRKY family

 Motifs on "-" strand (in INV orientation):  Found   3

     120  AATGACC      114       4.56
     137  AATGACC      131       4.56
     286  AATGACT      280       4.42

.............................................................
 WM:     >722. AC: RSP00722//OS: Nicotiana plumbaginifolia /GENE: rbcS 8B/RE: I-box
/BF: unknown transcription factor

 Motifs on "-" strand (in INV orientation):  Found   1

     251  GATAAGA      245       9.12

.............................................................
 Totally       8 motifs  of     5 different WMs have been found
```

Parameters:

Input	
Sequences	File with fasta sequences. In the current version of program, the maximal number of symbols in a line of FASTA file = 999.
Options	
Threshold type	threshold type, formula to calculate weight matrix cut-off value: **Based on weights of training motifs** - formula is: *Cut-off = Average + THR_VALUE * Std_dev* *"Average"* and *"Std_dev"* (standard deviation) are calculated for weights of motifs from which a weight matrix has been built. *THR_VALUE* is a real number (including 0). *THR_VALUE* is specified by "Threshold value" option. **Based on similarity to weight matrix** - formula is: *Cut-off = WM_Min_Value + THR_VALUE * (WM_Max_Value - WM_Min_Value)*

	"WM_Min_Value" and *"WM_Max_Value"* are minimal and maximal values that can be obtained with a corresponding weight matrix. *THR_VALUE* must belong to interval [0;1] (with default value = 0.9). *THR_VALUE* is specified by "Threshold value" option.
Threshold value	threshold value
DNA chain	DNA chain: **Direct** **Reverse** **Both**