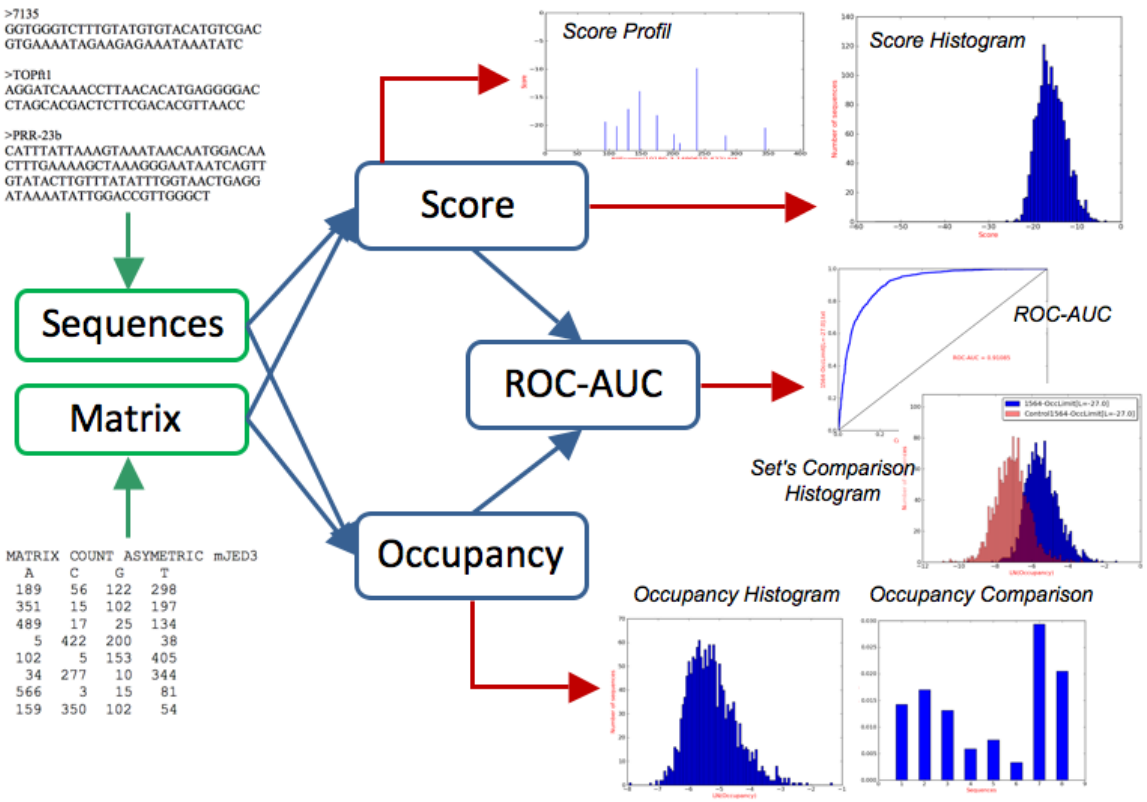
MORPHEUS

http://biodev.cea.fr/morpheus/

Prediction of Transcription Factors Binding Sites based on Position Weight Matrix.

Reference: *MORPHEUS, a Webtool for Transcripton Factor Binding Analysis Using Position Weight Matrices with Dependency* (Minguet EG, Charavay C, Segard S, Parcy F) PLos One. 2015 Aug 18 ; 10(8):e0135586. doi: 10.1371/journal.pone.0135586. eCollection 2015

Morpheus is a suite of tools to analyse transcription factor binding sites (TFBS) on DNA sequences. As input, the program uses a set of sequences in FASTA format and a matrix (Count Matrix, a Frequency Matrix or Position Weight Matrix [PWM]).



SEQUENCES

Sequences must be in standard ***FASTA format***. To facilitate data visualization and identification, we suggest to use short sequence names and to avoid duplicated names. A number between parentheses will be added if a duplicated name is found.

Characters not allowed in sequence name:

1 .- Avoid the use of "-" in sequence name as is used as mark for genomic information (see below). Their presence could result in incorrect name truncations and ambiguity in the results.

2.- Characters "\(:)/?*|<>" cannot be used either as several programs use it for creating individual profiles and data results files. By default, those characters are erased from sequence names. White spaces in sequence's name are also substituted by "_".

Example of Fasta Format:

```
>7135
GGTGGGTCTTTGTATGTGTACATGTCGACGTGA
AAATAGAAGAGAAATAAATATCAGCAAAA

>TOPft1
GGGCTTTGGACCGGATCAATAGCCACCTAACC
AAATCTGTGAACGGCGTTACTTTCAAATAAC

>PRR_23b
AAATACTAACAGAAATATCATTGTAGATAAGAC
TTTAGGGTCTTCGATCGACATATAATAAAGATA
AGGTTTGGGACCCTTTCAAGC
```

Unambiguous sequences are needed for score calculation of Transcription Factor Binding Sites (TFBS) because a single position can have a big influence over TFBS score. Although only "A,C,G,T" characters are allowed for score calculation, Morpheus gives the worst possible score if the site contains any ambiguous character (IUPAC DNA code: A, C, G, T, U, R, Y, S, W, K, M, B, D, H, V, N, . , -). Log file contains a list with sequences containing such characters to prevent incorrect interpretation of results.

Genomic information: If genomic information are available (chromosome number, start position and sequence size) as in the case of ChIP-CHIP or ChIP-seq data, it can be indicated using "-" in sequence name as follow:

Name-chromosome number-start position-size

For instance :

```
>7135-3-236478-62
GGTGGGTCTTTGTATGTGTACATGTCGACGTGA
AAATAGAAGAGAAATAAATATCAGCAAAA

>TOPft1-1-35642-63
GGGCTTTGGACCGGATCAATAGCCACCTAACC
AAATCTGTGAACGGCGTTACTTTCAAATAAC

>PRR_23b-5-1997005-87
AAATACTAACAGAAATATCATTGTAGATAAGAC
TTTAGGGTCTTCGATCGACATATAATAAAGATA
AGGTTTGGGACCCTTTCAAGC
```

Occupancy result files have independent columns for genomic information (empty if not available) to help further results analysis.

LogInfo file : You can find in the LogInfo file information about the process as the number of sequences that have been identified, matrix data used or any problem in sequences/matrix recognition.

MATRIX

Matrix information can be found in transcription factor databases such as JASPAR or generated from a binding sites alignment using tools such as MEME. Here we describe the adequate way to write that information so that it can be correctly interpreted by Morpheus. In order to facilitate conversion, the mPWM tool is available to generate Morpheus matrix format directly from a binding sites sequence alignment.

Morpheus matrix Conversion tool

This tool takes as input a file with binding sites sequence alignment either in fasta format or only aligned sequences. All sequences must have the same size. If not, the smaller size will be used. All IUPAC characters for DNA are allowed although only ACGT characters will be used for counting.

Optionally, if dependencies must be taken into account another file in txt format with dependencies positions will be necessary. In this file, positions for each dependency must be indicated between squares brackets, one dependency per line. As for example:

```
[2,3,5]
[6,9]
```

User must select if matrix is symmetric or asymmetric. When symmetric is selected, the corresponding symmetric dependencies will be generated even if they have not been indicated in the dependency information file. For example, if the dependencies indicated in the example describes dependencies in a symmetric binding site of a total of 15 positions, program will create 4 dependencies:

```
[2,3,5]
[6,9]
[7,10]
[11,13,14]
```

File generated will contain matrix information in Morpheus format ready to use with any analysis tool in Morpheus web.

Morpheus Matrix Format

The simplest format for a Matrix file is as follow:

```
MATRIX COUNT ASYMMETRIC mJED3
 A    C    G    T
189   56  122  298
351   15  102  197
489   17   25  134
  5  422  200   38
102    5  153  405
 34  277   10  344
566    3   15   81
159  350  102   54
```

The first line specifies the characteristics of the matrix separated by whitespaces, followed by information of nucleotide occurrence for each position (one position by line; in the example, a binding site of size 8 positions).

MATRIX --> Indicate that next data is a matrix data. ALL matrix files must start with this word.

Type of Data --> **COUNT** or **FREQUENCY** or **SCORE**

- **COUNT** --> Data is the occurrence number of each nucleotide (ACGT)
- **FREQUENCY** --> Data is the occurrence frequency of each nucleotide (ACGT)
- **SCORE** --> Data is the position weight matrix.

If the data is in "COUNT" or "FREQUENCY" formats, transformation to weights data is done as follows:

For each position --> Score $_{nt(ACGT)} = LN\left(\frac{freq_{nt}}{freq_{max}}\right)$

PSEUDOCOUNTS

To avoid null values for $freq_{nt}$, that are incompatible with score calculation, Morpheus uses Pseudocounts. When a COUNT matrix is provided, null values will be replaced by 1 before the matrix generation. The user can instead replace zeros by a value of its choice as there are different ways to deal with pseudocount values. When a FREQUENCY matrix is provided, a error message is written in LogInfo file if null values are present and no score matrix is generated. The user has to decide by which value the zeros have to be replaced before providing the frequency matrix.

SYMMETRY --> **ASYMMETRIC** or **SYMMETRIC**

In some cases, due to structure of transcription factor or because it binds as homodimer, the matrix is palindromic, so the scores of a sequence and of its reverse complement are the same. In these cases matrix has a SYMMETRIC structure.

If SYMMETRIC option is chosen but the matrix is asymmetric, scores will be calculated on a single DNA strand!

NAME --> The name of your matrix.

Default values: Only the term MATRIX and the matrix type are required for running the program. If one or both of the two others descriptive data (symmetry or name) are missing, the program will use default parameters.

- Symmetry --> ASYMMETRIC
- Name --> Unknown

LogInfo file: You can find in the LogInfo file information about matrix importation.

DEPENDENCY DATA

Position weight matrices assume that each base of the binding site contributes independently to the Transcription Factor/DNA affinity. However, there is evidence that interdependencies between positions exist and should be taken into account. Morpheus format allows using independent positions together with dependency in the needed positions.

For dependency data, a line with the word "DEPENDENCY" followed by the positions implicated, between brackets, precedes dependency data. This data is organized by alphabetical order for duplets (AA, AC, AG, … TG, TT) or triplets (AAA, AAC, AAG, … , TTG, TTT). In order to facilitate data visualization the use of [A,C,G,T] characters are allowed, for example:

```
DEPENDENCY (6 7 8)
       A       C       G       T
AA -6.0225 -4.0576 -6.0225 -5.8493
AC -6.0225 -5.1562 -6.0225 -6.0225
AG -3.3644 -5.8493 -5.8493 -5.1562
AT -6.0225 -6.0225 -6.0225 -6.0225
CA -6.0225 -6.0225 -6.0225 -5.8493
CC -6.0225 -6.0225 -6.0225 -6.0225
CG -4.2399 -6.0225 -6.0225 -6.0225
```

```
CT -5.8493 -6.0225 -6.0225 -6.0225
GA -5.1562 -4.0576 -6.0225 -5.8493
GC -6.0225 -6.0225 -6.0225 -6.0225
GG -0.7194 -4.7507 -5.8493 -3.5467
GT -4.2399 -6.0225 -6.0225 -6.0225
TA -1.8063  0.0000 -6.0225 -1.1954
TC -6.0225 -6.0225 -6.0225 -6.0225
TG -2.4820 -2.8048 -5.8493 -1.9992
TT -4.4630 -5.8493 -6.0225 -6.0225
```

Data in independent matrix for positions implicated in dependencies are not taken into account for score calculation.

INPUT FILES

Score & Occupancy: These programs need a file with matrix information (Morpheus format) and a file with sequences (Fasta format).

ROC-AUC: This program needs two result files from Score (option best only) or Occupancy, one with the positive data set and the other with the negative data set.

***m*PWM**: This program needs aligned motifs either in fasta format or raw format.

Negative Set Generator Tool: This program needs a file with sequences (fasta format); and optionally a text file with a list of chromosome sizes.

OUTPUT FILES

Two types of files are generated from each program with TFBS information:

1) File(s) in txt format with data organized in columns (score, position, TFBS sequence, etc.)

2) Image file(s) with TFBS information (score profile, score distribution, ROC-AUC, etc.), using general parameters.

If other graphic representations are needed, txt files contains all required data to generate them.

Sometimes txt output files are not correctly visualized with some simple text processor. If you are not able to see information in organized columns try to use other text processor.

PROGRAMS

***m*PWM Format Conversion Tool (see also MATRIX format)**

This tools allows generation of a matrix file in Morpheus format from an alignment of binding sites. Alignment can be either in fasta format or raw sequence alignment (a list of the motifs sequences).

This tool DO NOT do sequences alignment, it use motifs alignment (from tools like MEME) to generate matrix information in mPWM format. All sequences must have the same size so each position has the same information, because of that when several sizes are detected sequences are trimmed in the 3' end to match the smaller motif.

If the model must contain some dependency information, a list of the involved positions for each dependency must be provided in a text file: one dependency by line between square brackets, with positions separated by commas.

```
[4,6]
[9,10,11]
[2,17]
```

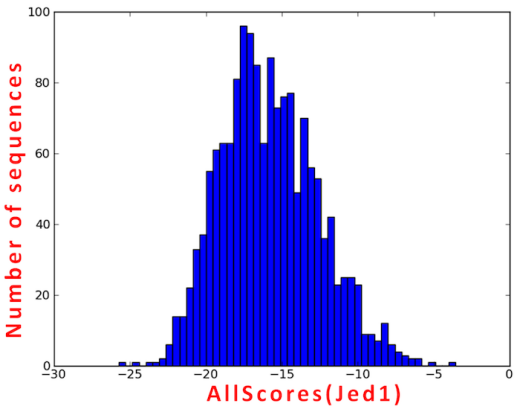
Score

This tool calculates the scores of TFBS present in a list of DNA sequences. Program needs a file with matrix information in Morpheus format (*m*PWM) and a file with sequences in fasta format. The available options are:

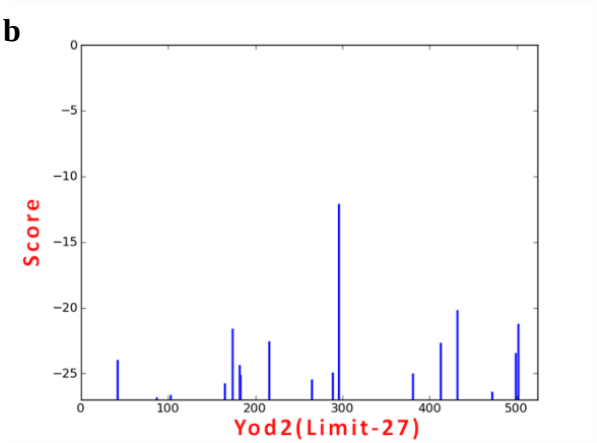
- Option "**All**" determines the score of all sites in each DNA sequence and generate quantitative profiles.
- Option "**Limit**" gives similar results than "All" but only site with a score higher than a given limit will be shown. User must indicate the desired limit.
- Option "**Best**" gives the site with the best score in each DNA sequence. Result of this option can be used as input for the ROC-AUC tool.

This tool generates graphic outputs what allows a quick overview of results. In all options a histogram is generated to reflect distribution of scores (a), that allows a quick general evaluation. "All" and "Limit" give an additional graphic output showing the position of the best sites in each sequence (b). A score profile is generated for each sequence in the input fasta format.

a



b



Occupancy

This tool calculates Transcription Factor Occupancy as the relative expected number of bound transcription factor molecules of each DNA sequence based on previous formalism (Roider et al., 2007). Affinity score correlates with the relative equilibrium dissociation constant following a linear curve:

$$\text{score}_S = -\ln(K_{D,S})\,a + b \rightarrow K_{D,S} = e^{\frac{(b - \text{score}_S)}{a}}$$

The linear correlation between i*n vitro* binding measurements and computed scores allows an experimental determination of "a" and "b" values

The occupancy is calculated as:

$$POcc = \sum_{s=1}^{L-W} p_s = \sum_{s=1}^{L-W} \frac{K_{A,s} \cdot [\text{TF}]}{1 + K_{A,s} \cdot [\text{TF}]}$$

Where $K_{A,S}$ is the relative equilibrium association constant for site S (the inverse of the relative equilibrium dissociation constant):

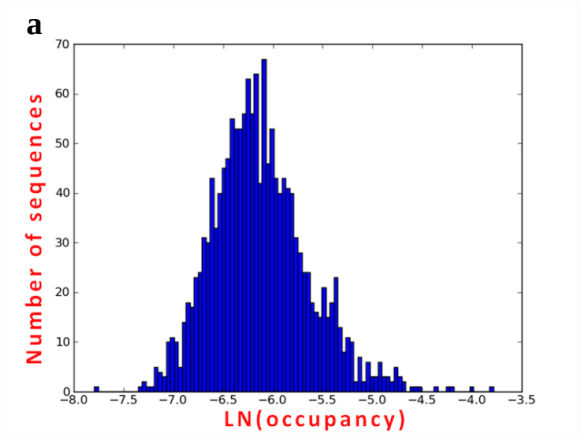
$$K_{A,S} = 1/K_{D,S}$$

Finally, the protein concentration of the transcription factor is needed and can be provided by the user. However, since protein concentration *in vivo* is rarely determined, a default value will be used so that the probability of binding to the best site (maximal score) is equal to 0.5, and the corresponding value will be indicated in the output files.

Program has two options:

- Option "**All**" use all sites in each sequence for Occupancy calculation.
- Option "**Limit**" gives similar results than "All" but a score limit can be specified so all sites with score worse than the limit will not be used. If this option is selected, limit value will have to be given as input.

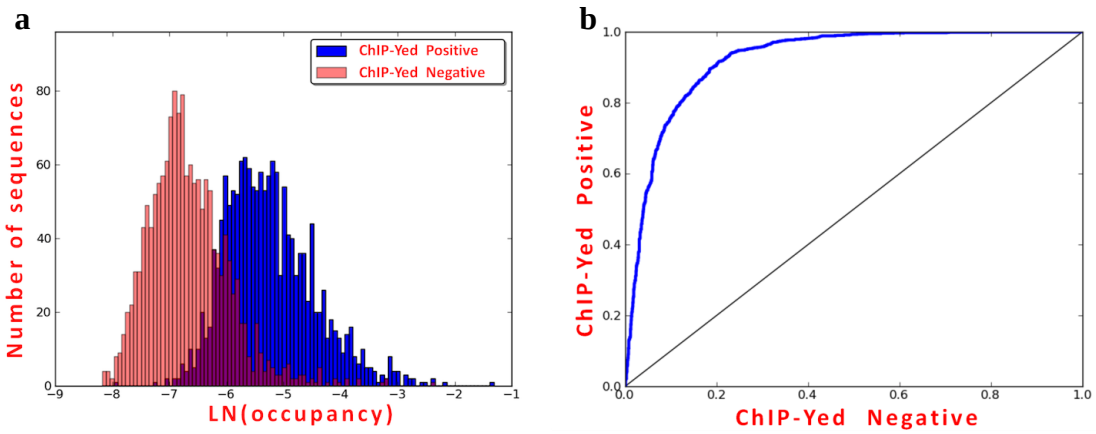
As graphic output, a histogram is generated to show the occupancy distribution (a). This tool also gives a graphic output showing occupancy of each sequence (b). The Result file can be used as input for ROC-AUC tool.



ROC-AUC

This tool calculates Receiver Operating Characteristic Area Under the Curve (**ROC-AUC**) from two data sets (such as bound and not bound regions, from a ChIP experiment for instance) from the Score tool (option "Best") or the Occupancy tool outputs.

A double histogram (with values for each dataset) is generated to reflect values distribution and as a visual representation of dataset's overlap (a). It also gives a graphical output with ROC curve (b).



Negative Set Generator Tool

This tools allows generation of a negative set that can be used in comparison with an experimental positive set of sequences.

In all cases, the negative set has the same number of sequences of the original positive set and with the same nucleotide size. This is particularly important when using occupancy prediction. We offer two methods to generate a negative set :

- "**SUFFLE**" This method randomly shuffle each sequence in the positive set of sequences so the new sequence has exactly the same proportion of the four nucleotides than the original positive sequence.
- "**GENOMIC**" This method randomly select genomic regions that are not present in the positive set of sequences. This method requieres genomic information (chrosomome, position and size of each sequence in the positive set; see "Sequences" point of this user guide for sequence's name) and the nucleotide size of each chromosome as a list in a text file (only numbers are allowed! Avoid using any other character; neither points or commas). In this file the order of numbers must correspond wih the natural order of chromosome number: first number is chromosome 1, the next one the chromome 2, and so on. If some chromosomes are not named as numbers, just do the assotiation by yourself and remember it.

This option randomly generates a negative set of sequences that do not overlap with any sequence in the positive set and produce a file with their genomic coordinates (NOT the sequences). Maybe some users will find it as a little inconvenient but we have designed this option so it can be used for any species now and in the future, and for any species with a complete genome version it must be really easy to get a set of sequences from their genomic coordinates.

Each method can have its advantages depending of the origin of the positive set and the porpose of the comparison. However we can indicate some comments and suggestions:

- If no complete genome is available only shuffle option can be used, obviously.

- With Shuffle option the negative set preserves the nucleotide proportion of the original positive set of sequences, what can be an advantage for some binding site evaluation methods that uses nucleotides background to calculate a p-value for each motif. Shuffle option reflects the probability of random generation of binding sites.

- With Genomic option the negative set corresponds with random real sequences present in the experiment in competition with those in the positive set. Although binding *in vivo* it is the result of several factors (as chromatine structure) and not just the sequence affinity, it is assumed that the positive set must be enriched with the motifs bound by the protein. A model that do not reflect this situation will indicate any bad assumption in the binding model.