Trawler Standalone

User Guide

April 29, 2009

1. Installation instructions

1.1 Linux/Unix and OSX

Trawler_standalone comes as a *.tar.gz* package. To unpack it, type:

```
> gunzip trawler_standalone-<version>.tar.gz
```

```
> tar xf trawler_standalone-<version>.tar
```

This creates a directory callede “trawler_standalone-<version>”, which contains a complete Trawler distribution.

1.1.1 Quick start

The easiest way to find out if you have all the required dependencies installed, to run Trawler_standalone, is to execute the *pre-install_trawler.sh* script provided with the distribution.
Use the following command in a terminal:

```
> ./pre-install_trawler.sh
```

a. If you have the required dependencies already installed, you can now try your Trawler_standalone installation by running the examples (see section 1.1.3. Running the examples).

b. If you are missing some dependencies, please refer to the Dependencies installation section (1.1.2).

1.1.2 Dependencies installation

Perl distribution

Almost every today's Linux/Unix system contains a Perl installation, so we expect that you already have a working Perl setup.

Make sure perl is installed with version 5.6 or higher (http://www.perl.org/)

```
> perl -v
```

If Perl is installed, this command will return its version.

Java Platform

Install Java if needed, Java 5.0 or newer is required to run Trawler standalone.
Java can be downloaded from http://java.sun.com/javase/downloads/index.jsp

If Java is already installed, make sure you are using version 5.0 or higher. In order to check the version of the Java, you can use the following command:

```
> java -version
```

This command will return its version.

Ghostscript

Make sure you have Ghostscript installed (http://pages.cs.wisc.edu/~ghost/)

```
> gs -version
```

If Ghostscript is installed, this command will return its version.

If needed, Ghostscript can be installed your package manager.

Algorithm-Cluster

NOTE: - You will need root privileges, for this installation step.
- see also instructions included with the Algorithm-Cluster release.

```
> sudo perl -MCPAN -e shell
cpan> install Algorithm::Cluster
```

If you do not have root privileges, download the module directly from CPAN (http://search.cpan.org/~mdehoon/)
and use perl Makefile.PL prefix=/some/other/directory
to install the module in /some/other/directory/lib/perl5/.
For Perl, type:

```
> perl Makefile.PL
> make
> make test
> make install
```

For more information, refer to the installation notes included in the Algorithm-Cluster module.

1.1.3 Running the examples

If you have the dependencies installed on your system, it should be possible to immediately run Trawler_standalone using the examples as follows:

```
> cd trawler_standalone-<version>/
```

Run example 1 (running time is less than a minute):

```
> ./examples/run_trawler.sh
```

Run example 2 (running time is about 2 min):

```
> ./examples/creb/run_trawler_creb.sh
```

Run example 3 (running time is about 2 min):

```
> ./examples/creb/run_trawler_creb_orthos.sh
```

Run example 4 (running time is about 1 min):

```
> ./examples/miR106b/run_trawler_miR106b.sh
```

NOTE: - by default results are stored in trawler_standalone-<version>/myResults/

1.1.4 Advanced Configuration

NOTE: - TRAWLER_HOME is the full path to the Trawler installation directory.

Edit the configuration file in TRAWLER_HOME/conf/trawler.cfg

* basepath:
The directory where trawler is installed.

e.g. /home/username/trawler

* resultpath:
e.g. /home/username/Trawler_output
Set this variable, using full path , if you want to change the default output directory.
By default results are stored in TRAWLER_HOME/myResults
It is generally safer to set this variable to keep your results outside your Trawler installation directory.

* logging:
debug and info loggings are disabled by default, if you want to enable the logging replace the value by 1 (instead of 0).

1.2 Windows

1.2.1 Dependencies installation

Perl distribution

ActivePerl is a ready-to-install distribution of Perl you can download from
http://www.activestate.com/Products/activeperl/index.mhtml

Java

You can obtain Java from http://java.sun.com/javase/downloads/index.jsp

Ghostscript

Download and install GPL Ghostscript from http://pages.cs.wisc.edu/~ghost/

Configure weblogo:
Go to trawler_standalone-<>\weblogo
Rename logo.conf.init to logo.conf
Set the gs PATH (by removing the '#' symbol at the beginning of the line)
eg. gs=C:\Program Files\gs\gs8.63\bin\gswin32c.exe

Algorithm-Cluster module

To install Algorithm::Cluster for Perl 5.8 and Perl 5.10 on Windows,
you can use the precompiled package by executing this command in a Shell window:
(Start menu -> Run... -> type cmd then OK)

```
> ppm install http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/Algorithm-Cluster.ppd
```

1.2.2 Running the examples

If you have the dependencies installed on your system, it should be possible to immediately run Trawler using the examples as follows:

Open a shell window:
Start menu -> Run... -> type cmd then OK

```
> cd trawler_standalone-<version>\
```

Run example 1 (running time is less than a minute):

```
> examples\run_trawler.bat
```

Run example 2 (running time is about 2 min):

```
> examples\creb\run_trawler_creb.bat
```

Run example 3 (running time is about 2 min):

```
> examples\creb\run_trawler_creb_orthos.bat
```

Run example 4 (running time is about 1 min):

```
> examples\miR106b\run_trawler_miR106b.bat
```

NOTE: - by default results are stored in trawler_standalone-<version>\myResults\

2. Troubleshooting

- **Out-of-memory errors**

Depending on the size of your data set, trawler can fail when the computer runs out of memory. Trawler will display an error message such as Java OutOfMemoryError. The problem can be addressed by closing other applications that are running on the same computer or increasing Trawler memory setting using the JAVA_OPTS property within the TRAWLER_HOME/conf/trawler.cfg file.

- **Out-of-memory errors**

Depending on the size of your dataset, trawler can fail when the computer runs out of memory. Trawler will display an error message such as Java OutOfMemoryError. The problem can be addressed by closing other applications that are running on the same computer or increasing trawler memory setting using the JAVA_OPTS property within the TRAWLER_HOME/conf/trawler.cfg file.
By default the Java memory allocation is -Xmx256m but can be replaced with a higher value, for example JAVA_OPTS = -Xms500m

- **No motifs found**

If Trawler cannot detect any over-represented motifs, the following error will appear in the shell: "No motif satisfying your criteria has been found". Please re-run Trawler using different parameters. Increase or decrease the parameter "Minimum number of motif occurrence" for instance.

- **Html output: no images**

The images of the motif's PWM cannot be displayed (especially for Windows users). Make sure that you have installed the Ghostscript dependency (please refer to the Installation section details).

- **Html output: no tables**

The web browser does not have Javascript enabled. Modify your browser preferences to enable Javascript (for example Safari : go to preference/ security check enable javascript).

3. Command line arguments

USAGE

```
trawler -sample [file containing the enriched sequences] -background [file containing the background sequences]
```

- **sample** (FASTA format) better to be repeat-masked.

- **background** (FASTA format)

OPTIONAL PARAMETERS

[MOTIF DISCOVERY]

- **occurrence** (optional) minimum occurrence in the sample sequences. [DEFAULT = 10]
- **mlength** (optional) minimum motif length. [DEFAULT = 6]
- **wildcard** (optional) number of wild card in motifs. [DEFAULT = 2]
- **strand** (optional) single or double. [DEFAULT = single]

[CLUSTERING]

- **overlap** (optional) in percentage. [DEFAULT = 70]
- **motif_number** (optional) total number of motifs to be clustered. [DEFAULT = 200]
- **nb_of_cluster** (optional) fixed number of cluster. if this option is used, the k-mean clustering algorithm with fixed k will be used instead of the self organizing map (SOM). [DEFAULT = NULL]

[VISUALIZATION]

- **directory** (optional) output directory. [DEFAULT = "$TRAWLER_HOME/myResults"]
- **dir_id** (optional) gives an id to the results directory. [DEFAULT = NULL]
- **xtralen** (optional) add bases around the motifs for the logo. [DEFAULT = 0]
- **alignments** (optional) file containing the list of files containing the aligned sequences (see README file for more info) [DEFAULT = NULL]
- **ref_species** (optional) name of the reference species [DEFAULT = NULL]
- **clustering** (optional) if 1 the program clusters the instances, if 0 no clustering. [DEFAULT = 1]
- **web** (optional) if 1 the output will be a web page with all the information [DEFAULT = 1]

USAGE EXAMPLES

Trawler-standalone comes in two flavors: [1] with the conservation information or [2] without. For the conservation you need to provide Trawler-standalone with a file containing all the orthologous sequences (not aligned).

[1] with conservation

```
> trawler.pl -sample file_sample -background file_background -alignments file_alignments -ref_species reference_species_name
```

[2] without conservation

```
> trawler.pl -sample file_sample -background file_background
```

By default Trawler-standalone performs the full analysis nevertheless it is possible to

[1] get only the over-represented motifs:

```
> trawler.pl -sample file_sample -background file_bakground -clustering 0
```

[2] get only the over-represented motifs and the cluster (PWM):

```
> trawler.pl -sample file_sample -background file_bakground -web 0
```

OPTIONS DESCRIPTION

[MOTIF DISCOVERY]

- **occurrence** : [DEFAULT = 10] The minimum number of time the motif occurs in the sample sequence. The lower limit defined by the suffix tree is 2 motifs. Nevertheless if your chromatin IP contains more than just a few sequences, you would expect to see the motif occurring quite often. For typical chromatin IP data, a good minimum number of motif occurrence is about 10 to 20. If you do not get anything interesting with this values, try a higher or lower number according to your sample size.
- **mlength** : [DEFAULT = 6] Trawler-standalone does not assess motifs of fixed length. The maximum length has been set to 20 nucleotides but the minimum length is a user defined parrameter. A good minimum length is around 5-6 bp for typical transcription factor binding sites.
- **wildcard** : [DEFAULT = 2] The number of wildcards is the maximum number of positions in the motif that can have a degenerate nucleotide. In IUPAC code, the possible wildcards used in Trawler-standalone are the following :
  M = [AC], R = [AG], Y = [CT], W = [AT], S = [CG], K = [GT], N = [ATCG] (the latest counts as 2 mismatches). A possible motif with -wildcard option set to 2 can be AGCMTWA for example.
- **strand** : [DEFAULT = single] The analysis can be performed either on single-stranded or double-stranded sequences.

[CLUSTERING]

- **overlap** : [DEFAULT = 70] The minimum percentage overlap for two motifs to be considered in the same cluster. The value is by default 70 %; lowering down this value will cluster more distant motifs.
- **motif_number** : [DEFAULT = 200] The maximum number of motifs to be considered for clustering. If the option is set to 200, the best 200 motifs returned by the discovery step will be used for the clustering step.
- **nb_of_cluster** : [DEFAULT = NULL] By default, the number of cluster(s) is defined by the self organizing map algorithm (SOM). By using this option, the user set the number of expected cluster(s). In this case, the SOM algorithm is replaced by the k-mean clustering algorithm.

[VISUALIZATION]

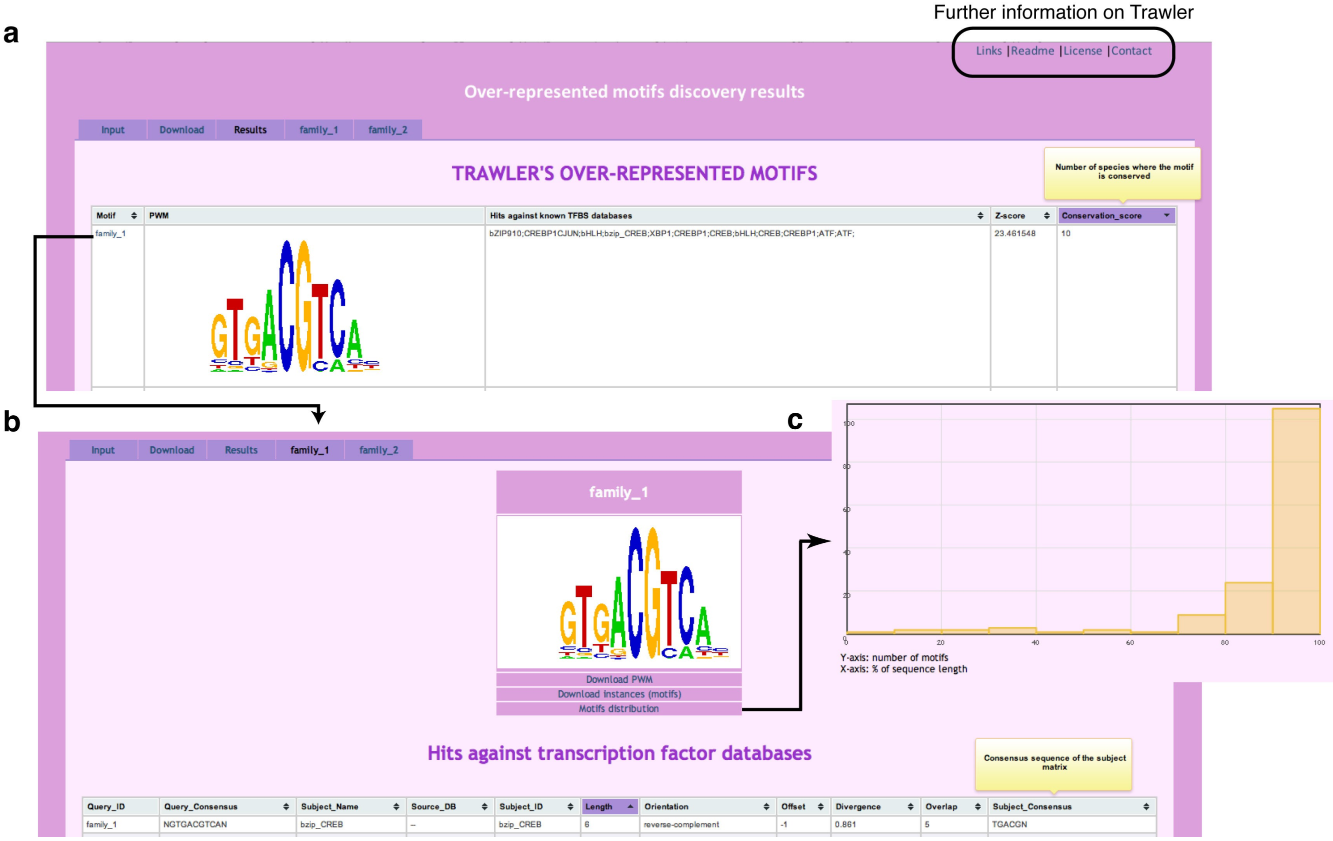
- **directory** : [DEFAULT = "$TRAWLER_HOME/myResults"] name of the directory where all the outputs will be stored. If no directory is specified, then a default directory "myResults" will be created, Individual results will be stored in a separate directory with the following pattern "tmp_yyyy-mm-dd_hh:mm_ss" (ex : tmp_2009-05-20_12h13_20).
- **dir_id** : [DEFAULT = NULL] gives a meaningful ID to the results directory. For instance, if this option is set to "myID", this will create a directory named "myID_yyyy-mm-dd_hh:mm_ss".
- **xtralen** : [DEFAULT = 0] The number of additional nucleotides included in the PWM flanking the core motif.
- **alignments** : [DEFAULT = NULL] Alignment file in the correct format. If alignments are provided, the motifs are visualized in the context of an alignment using Jalview. If the -alignment option is activated, the option -ref_species needs to be set.
- **ref_species** : [DEFAULT = NULL] Name of the reference species. Only use this option if the -alignment option is activated.
- **clustering** : [DEFAULT = 1]. By default, Trawler proceeds to the clustering step. To only obtain the over-represented motifs, set the -clustering option to 0. If the -clustering option is set to 0, no webpage is produced.
- **web** : [DEFAULT = 1] By default, Trawler proceeds to create a web page summarizing the result. If this option is set to 0, Trawler proceeds until the clustering step only. Useful for batch analysis.

4.Additional features

Additional reference matrices can be added to the known matrices used by Treg comparator by adding to the file matrix_set_all your PWM in the following format :

ID#name#text#text#0.3548828125 ...
Matrix entries (float) are separated by a whitespace. The matrix entries start with the frequency of "A", followed by that of "C", "G", and "T" in the first position (5'), then the same for the second position and so on.

5. Web Output


a
Further information on Trawler
Links |Readme |License |Contact
Over-represented motifs discovery results
Input
Download
Results
family_1
family_2
Number of species where the motif is conserved
b
c
family_1
Download PWM
Download instances (motifs)
Motifs distribution
100
80
60
40
20
0
20
40
60
80
100
Y-axis: number of motifs
X-axis: % of sequence length
Consensus sequence of the subject matrix

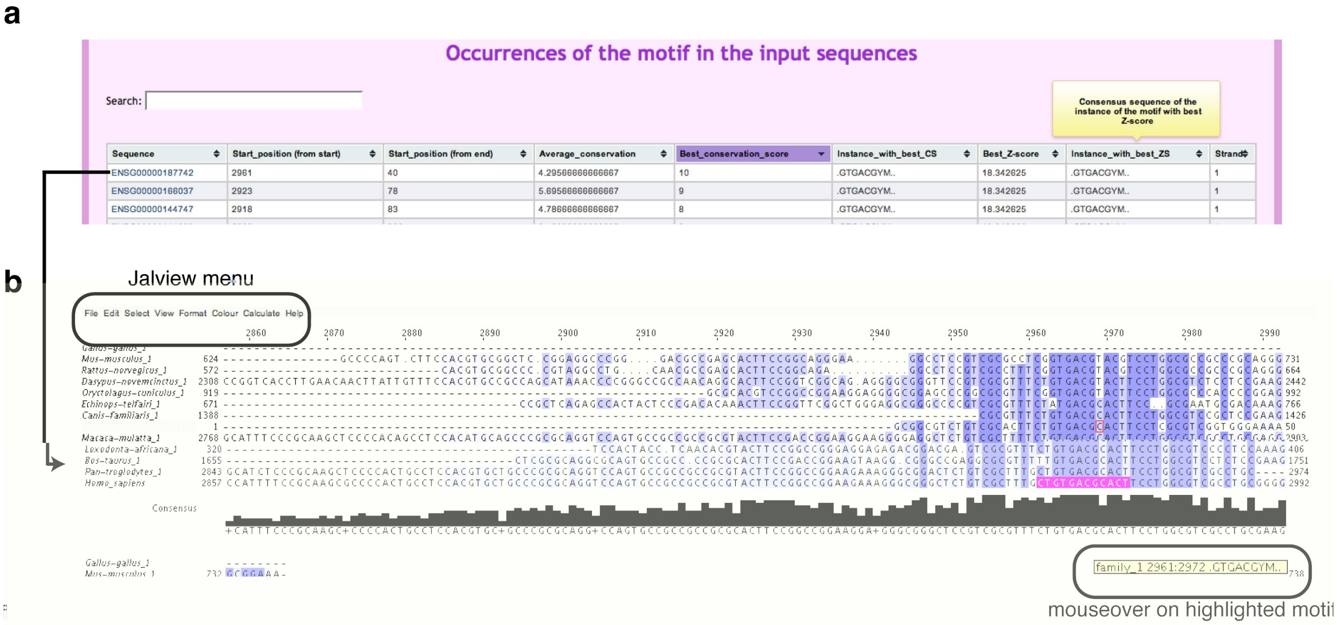
TRAWLER'S OVER-REPRESENTED MOTIFS

Motif	PWM	Hits against known TFBS databases	Z-score	Conservation_score
family_1		bZIP910;CREBP1CJUN;bHLH;bzip_CREB;XBP1;CREBP1;CREB;bHLH;CREB;CREBP1;ATF;ATF;	23.461548	10

Hits against transcription factor databases

Query_ID	Query_Consensus	Subject_Name	Source_DB	Subject_ID	Length	Orientation	Offset	Divergence	Overlap	Subject_Consensus
family_1	NGTGACGTCAN	bzip_CREB	–	bzip_CREB	6	reverse-complement	-1	0.861	5	TGACGN

a - output summary: The list of over-represented motifs are displayed. The conservation column only appears if a multiple alignment of orthologous sequences has been submitted.
b - List of similar motifs identified in TFBS databases.
c - Distribution plot of the motifs positions.


a
b
Jalview menu
File Edit Select View Format Colour Calculate Help
mouseover on highlighted motif

Occurrences of the motif in the input sequences

Search:

Consensus sequence of the instance of the motif with best Z-score

Sequence	Start_position (from start)	Start_position (from end)	Average_conservation	Best_conservation_score	Instance_with_best_CS	Best_Z-score	Instance_with_best_ZS	Strand
ENSG00000187742	2961	40	4.29566666666667	10	.GTGACGYM..	18.342625	.GTGACGYM..	1
ENSG00000166037	2923	78	5.69566666666667	9	.GTGACGYM..	18.342625	.GTGACGYM..	1
ENSG00000144747	2918	83	4.78666666666667	8	.GTGACGYM..	18.342625	.GTGACGYM..	1

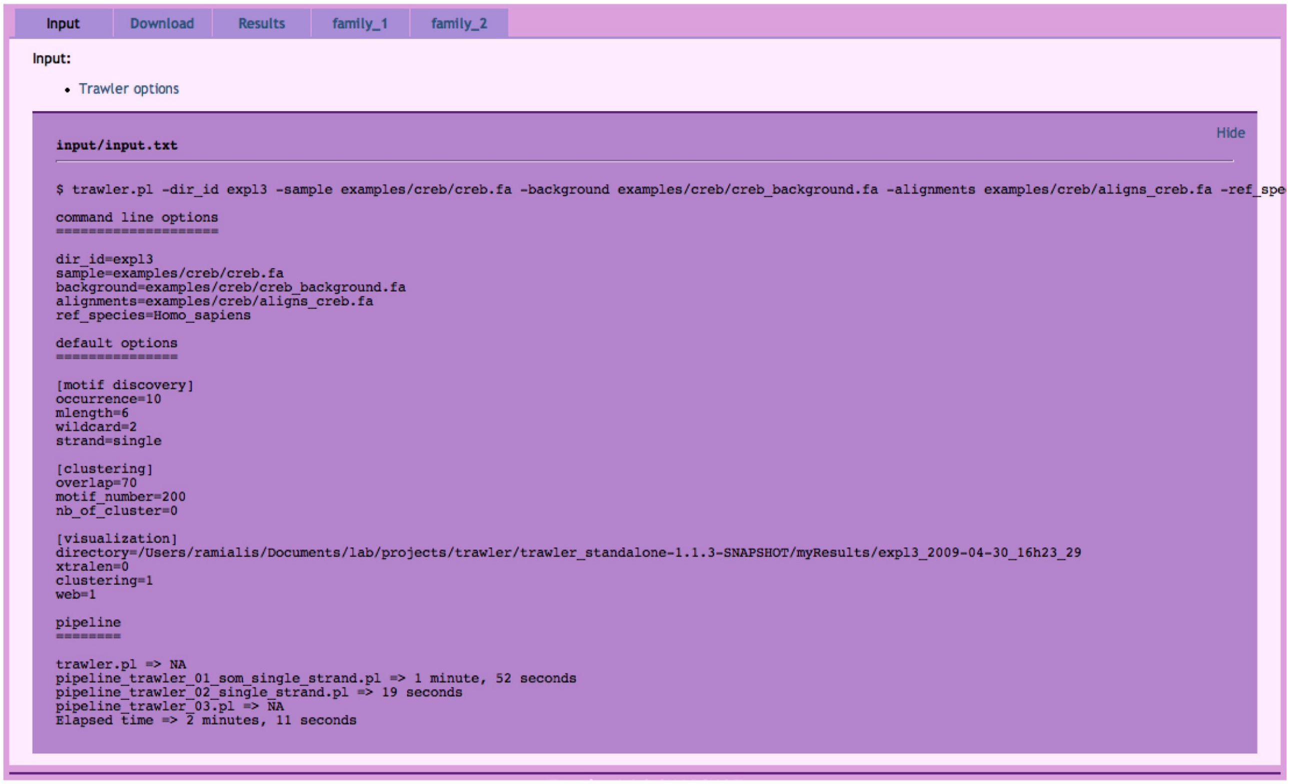
```
                                    2860      2870      2880      2890      2900      2910      2920      2930      2940      2950      2960      2970      2980      2990
Gallus-gallus_1               ----------------------------------------------------------------------------------------------------------------------------------------------
Mus-musculus_1           624  ---------------GCCCCAGT.CTTCCACGTGCGGCTC.CGGAGGCCGG....GACGCCGAGCACTTCCGGCAGGGAA.......GGCCTCCGTCGCGCCTCGGTGACGTACGTCCTGGCGCCGCCCGCAGGG 731
Rattus-norvegicus_1      572  ---------------------------CACGTGCGGCCC.CGTAGGCCTG....CAACGCCGAGCACTTCCGGCAGA..........GGCCTCCGTCGCCTTTCGGTGACGTACGTCCTGGCGCCGCCCGCAGGG 664
Dasypus-novemcinctus_1   2308 CCGGTCACCTTGAACAACTTATTGTTTCCACGTGCCGCCAGCATAAACCCGGGCCGCCAACAGGCACTTCCGGTCGGCAG.AGGGGCGGGTTCCGTCGCGTTTCTGTGACGTACTTCCTGGCGTCTCCTCCGAAG 2442
Oryctolagus-cuniculus_1  919  ------------------------------------------------------------GCGCACGTCCGGCCGGAAGGAGGGGCGGAGCCCGGCGCGTTTCGGTGACGTACTTCCTGGCGCCCACCCGGAG 992
Echinops-telfairi_1      671  ---------------------------------------CCGCTCAGAGCCACTACTCCCGACACAAACTTCCGGTTCGGCTGGGAGGCGGGCCCCGTCGCGTTTCTATGACGCACTTCC..GCGAATGCGACAAAG 766
Canis-familiaris_1       1388 ------------------------------------------------------------------------------------------------------CGCGTTTCTGTGACGCACTTCCTGGCGTCCGCTCCGAAG 1426
                         1    ------------------------------------------------------------------------------------------GCGGCGTCTGTCGCACTTCTGTGACGCACTTCCTCGCGTCGGTGGGAAAA 50
Macaca-mulatta_1         2768 GCATTTCCCGCAAGCTCCCCACAGCCTCCACATGCAGCCCGCGCAGGTCCAGTGCCGCCGCCGCGTACTTCCGACCGGAAGGAAGGGGAGGCTCTGTCGCTTTTCTGTGACGCACTTCCTGGCGTCGCCTGCGAGG 2903
Loxodonta-africana_1     320  ---------------------------------------------TCCACTACC.TCAACACGTACTTCCGGCCGGGAGGAGAGACGGACGA.GTCGCGTTTCTGTGACGCACTTCCTGGCGTCCCCTCCAAAG 406
Bos-taurus_1             1655 -------------------------------------CTCGCGCAGGCGCAGTGCCGCC.CCGCGCACTTCCGACCGGAAGTAAGG.CGGGCCGAGGCGCGTTTTTGTGACGCACTTCCTGGCGTCCTCTCCGAAG 1751
Pan-troglodytes_1        2843 GCATCTCCCGCAAGCTCCCCACTGCCTCCACGTGCTGCCCGCGCAGGTCCAGTGCCGCCGCCGCGTACTTCCGGCCGGAAGAAAGGGCGGACTCTGTCGCTTTGCTGTGACGCACTTCCTGGCGTCGCCTGC---- 2974
Homo_sapiens             2857 CCATTTTCCGCAAGCGCCCCACTGCCTCCACGTGCTGCCCGCGCAGGTCCAGTGCCGCCGCCGCGTACTTCCGGCCGGAAGAAAGGGCGGGCTCTGTCGCTTTGCTGTGACGCACTTCCTGGCGTCGCCTGCGGGG 2992

Consensus
                              +CATTTCCCGCAAGC+CCCCACTGCCTCCACGTGC+GCCCGCGCAGG+CCAGTGCCGCCGCCGCGCACTTCCGGCCGGAAGGA+GGGCGGGCTCCGTCGCGTTTCTGTGACGCACTTCCTGGCGTCGCCTGCGAAG

Gallus-gallus_1               --------
Mus-musculus_1           732  GCGGAAA-
```

family_1 2961:2972 .GTGACGYM.. 738

a - List of sequences containing one instance of the motif.

b - Visualization of one sequence using Jalview. The instance of the motif within this sequence is highlighted in pink.

a

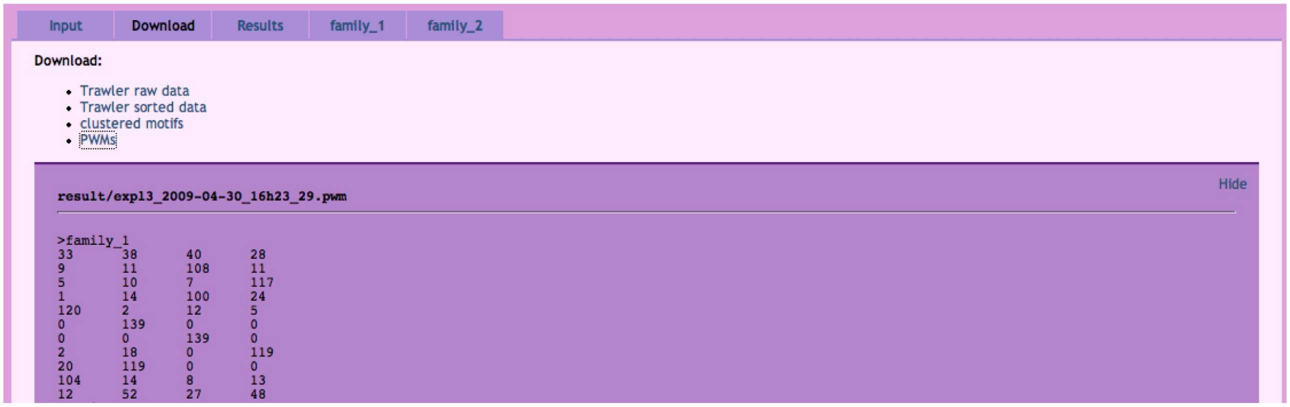


Input:

- Trawler options

```
input/input.txt

$ trawler.pl -dir_id expl3 -sample examples/creb/creb.fa -background examples/creb/creb_background.fa -alignments examples/creb/aligns_creb.fa -ref_spe

command line options
====================

dir_id=expl3
sample=examples/creb/creb.fa
background=examples/creb/creb_background.fa
alignments=examples/creb/aligns_creb.fa
ref_species=Homo_sapiens

default options
===============

[motif discovery]
occurrence=10
mlength=6
wildcard=2
strand=single

[clustering]
overlap=70
motif_number=200
nb_of_cluster=0

[visualization]
directory=/Users/ramialis/Documents/lab/projects/trawler/trawler_standalone-1.1.3-SNAPSHOT/myResults/expl3_2009-04-30_16h23_29
xtralen=0
clustering=1
web=1

pipeline
========

trawler.pl => NA
pipeline_trawler_01_som_single_strand.pl => 1 minute, 52 seconds
pipeline_trawler_02_single_strand.pl => 19 seconds
pipeline_trawler_03.pl => NA
Elapsed time => 2 minutes, 11 seconds
```

b



Download:

- Trawler raw data
- Trawler sorted data
- clustered motifs
- PWMs

```
result/expl3_2009-04-30_16h23_29.pwm

>family_1
33      38      40      28
9       11      108     11
5       10      7       117
1       14      100     24
120     2       12      5
0       139     0       0
0       0       139     0
2       18      0       119
20      119     0       0
104     14      8       13
12      52      27      48
```

a - Summary of the parameters used.
b - List of files available for download.