SIOMICS V1.1 Manual

1.Prerequisites

In order to use this software, you should get the following things ready:

(1) You need to have python installed (Python 2 or Python 3).
You can download Python from here .
Besides,you need Tkinter (Python 2) or tkinter (Python 3) module to enable the GUI.
For Windows users, the module was already included in the windows Python installer.(link)
For Linux users, please see here for installation instructions.

(2)You would also have to install Jave Runtime Environment. (**Optional but highly recommended, it will improve the time efficiency**)
You can check here for installation instructions.

(3) You may also need to configure the Environment Variable for running java and python (**Optional for SIOMICS GUI version users.**)
A. For configuring Java Environment Variable, you can see here for instructions.
B. For configuring Python paths, you might see here for help.

2.Paramters

(1) Description of parameters

python SIOMICS.py
-i <input_peak_sequences> Input TF ChIP-seq peak sequences should be in FASTA format, see here.
-o <ouput directory> This parameter used to specify the directory for output, all results will be put there.
-w <length_of_motifs> This parameter specifies the length of motifs,(Motif length range: 6-14).
-m <maximal_number_of_output_motifs> The maximal number of output motifs.
-s <support_value_of_motif_combination> The minimal number of sequences a motif module needs to occur in order to be considered as significant or frequent.
-c < corrected_pvalue cutoff> The multiple comparison corrected p-value cutoff for motif module prediction.
-r < number_of_iterations> The number of iterations.

(2) Recommended parameters for SIOMICS

- -w: 8, motif length =8
- -m: 100, in every iteration,100 top motif candidates will be used to predict motif modules.
- -s: 1% of total number of sequences (Please keep in mind: s should be Integer, for example, the total input sequence 10,000 , s=1%*10,000=100)
- -c: 0.01, corrected p-value cutoff 0.01
- -r: 20, 20 iterations at the maximal.

---

3. Software Usage

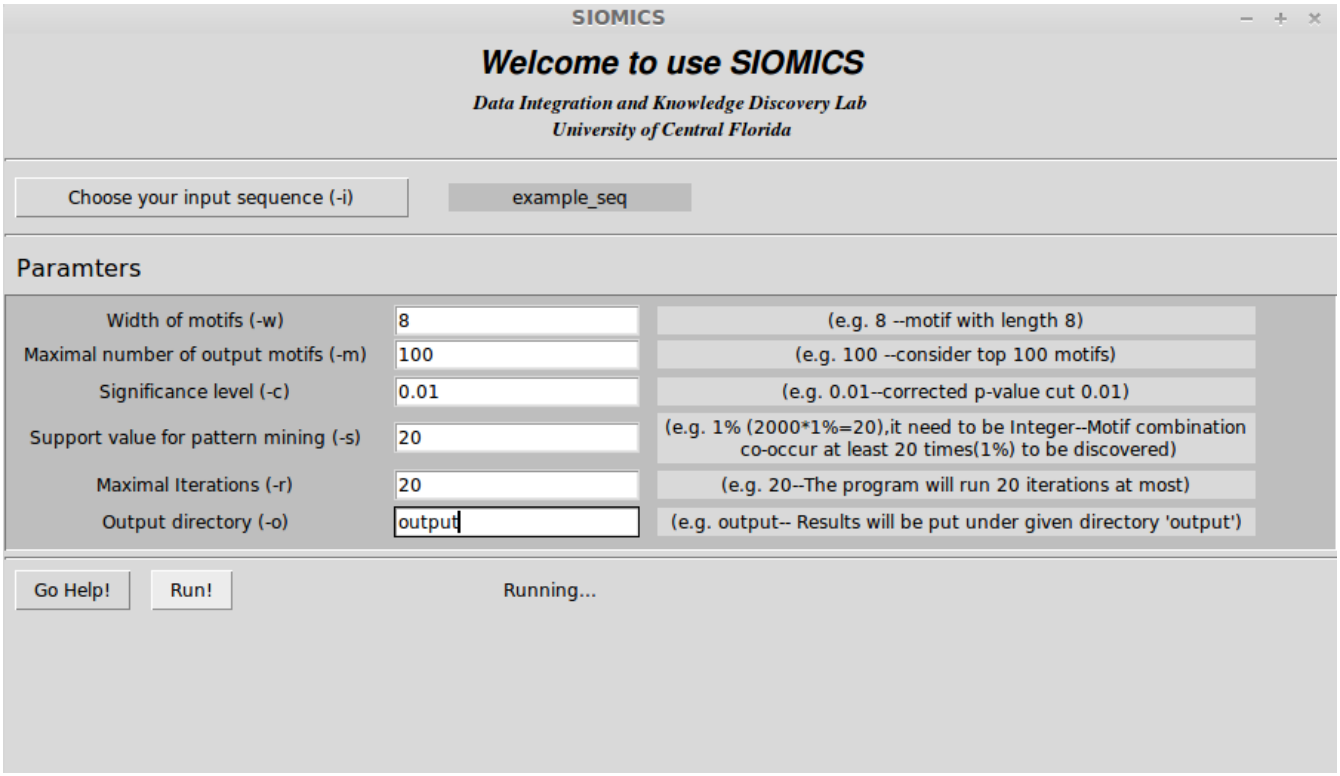
The following example shows you how to use the SIOMICS (both command line and GUI).
(1) Command line example
For example, if we want to identify motifs with length=8 from the provided "example_seq" dataset under the "example" directory . We can use the following command:
python SIOMICS.py -i example/example_seq -o example_output -w 8 -s 20 -c 0.01 -r 20 -m 100.

The meaning of the above parameters:
Try to identify motifs with length =8, corrected p-value < 0.01. The motifs need to co-occur at least 20 times to be claimed as modules. The maximal number of predicted motifs =100. The maximal iterations =20.
Note:The format of input sequence is the FASTA format.

If you do not want to specify every parameters by yourself, you can use the "batch_siomics.py" scipt we provided. This script can be used to run SIOMICS on a batch of peak sequences with default parameters.
Take the "example" folder included in the software as an example:
We can get the predictions for all datasets under the "example" folder by using the following command:
python batch_siomics.py example

SIOMICS will be run on each of sequence file under "example" folder sequentially.
The output files will be put into directory names as <DatasetName_out>

(2) GUI example
In order run GUI version of SIOMICS, just double click "SIOMICS_GUI.py". See the following GUI example:



---

4. SIOMICS Results

When the software is running, you will see "Running..." shown in the bottom of the GUI. It might show "Not Responding" when SIOMICS is running on Windows, but it's OK. It will show "done" on the bottom of GUI once the results were obtained. SIOMICS will provide two files as the results of prediction in the output directory provided.
(1) X.motifs
(2) X.mc
X.motifs is the result file of the predicted motifs in the format of frequency matrix
X.mc is the result file of motif modules predicted.
See the following example to see the meaning of X.motifs:

```
>M0     8.810747221152823       3.359771834161662E-5    8.0935338229
0.8764 0.0299 0.0518 0.0418
0.9081 0.0304 0.04 0.0215
0.8847 0.041 0.0416 0.0327
0.8378 0.064 0.0532 0.045
0.839 0.0654 0.0515 0.0441
0.8756 0.0456 0.0423 0.0365
0.9033 0.0333 0.0349 0.0286
0.8752 0.0337 0.0463 0.0449
```

The first line represents:
ID of motifs : M0
Scan_cutoff: 8.810747221152823 Used to define a putative TFBS of the motif.
lambda: 3.359771834161662E-5 Represent the probability of this motif occurring in random sequences (per nucleotide).
MDScan_score: 8.0935338229 MDScan score used to represent the statistical significance of predicted motifs
The remaining lines represent the frequencies of"A,C,G,T" in each position.

See the following example to see the meaning of X.mc:

```
M66 M21 (58)    (6.7476668697e-10)
```

This denotes M66 and M21 were regarded as a motif module (co-occur in 58 sequences). The corrected pvalue is 6.7476668697e-10.

---

5. Contact Info

If you have any question regarding to the SIOMICS software or you have found any bugs, please feel free to contact us via xiaoman@mail.ucf.edu. For any non-academic use of this software, please also contact xiaoman@mail.ucf.edu.