User Guide

Honglei Liu

liuhonglei@gmail.com

Overview

This package includes the implementation of the anchor based clustering algorithm (ASC) and an integrated fast motif discovery tool ASC+MEME. ASC+MEME provides similar interfaces and outputs as MEME but is five orders of magnitude faster.

Usage agreement

Downloading is for internal research purpose only. Redistribution and commercial usage are not permitted. For other interests, contact the authors (liuhonglei@gmail.com).

Prerequisite

To use ASC+MEME, you need to have MEME Suite installed.

http://meme-suite.org/doc/download.html

The executable files of MEME Suite should be added to the PATH environment variable, or the following executable files should be copied to the the installation directory of ASC+MEME:

```
meme
meme2images
fimo
```

You will also need to have `Python 2.7` installed.

Installation:

You can download and decompress the package by running

```
wget http://www.cs.ucsb.edu/~honglei/abp/package/ASC_MEME.tar.gz
tar -xvzf ASC_MEME.tar.gz
cd ASC_MEME
```

This package includes the executable file for ASC that could run in a Linux environment. If you need the source code, please contact us.

Note: after the package is downloaded and decompressed, remember to copy the MEME executables to the package directory.

Step by Step

- ### Input file

  ASC+MEME takes a sequence file in fasta format as input. For example, the file should look like

  ```
  >1|356|350|1563|358
  ESGVIWYNEVMHGKS
  >2|80|663|1533|216
  VWERLGPATSWKTEA
  >3|614|193|1231|441
  VDVWYSESVHAKPSV
  >4|344|341|1457|331
  VRGMLPNWYDEMMFS
  >5|167|659|1026|547
  AANPVEMGLLTMSRL
  ```

  A sample file is provided in http://www.cs.ucsb.edu/~honglei/abp/dataset/dataset.tar.gz. You can download and decompress the file by running

  ```
  wget http://www.cs.ucsb.edu/~honglei/abp/dataset/dataset.tar.gz
  tar -xvzf dataset.tar.gz
  ```

- ### Output files

  The outputs of ASC+MEME mainly contain three parts:

  i. pwms.txt

     This file contains the position weight matrix (PWM) of every motif found. For example,

     ```
     MOTIF 1
     letter-probability matrix: alength= 20 w= 8 nsites= 16 E= 8.60e-07
     0.0  0.0  0.0625  0.0  0.0625  0.0  0.0  0.0  0.0  0.0  0.0625  0.0
     0.0625  0.0  0.0  0.25  0.0  0.0  0.0625  0.4375
     ```

```
0.0  0.0  0.0  0.8125  0.0625  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
0.0  0.0  0.125  0.0  0.0  0.0  0.0
0.0  0.0  0.0  0.0  0.0  0.125  0.0  0.0625  0.0  0.0625  0.0625  0.125
0.0  0.0  0.125  0.125  0.0  0.0  0.25  0.0625
0.125  0.0  0.0  0.0625  0.0  0.0625  0.0  0.125  0.0  0.0625  0.125
0.0625  0.0  0.0  0.0  0.1875  0.0  0.1875  0.0  0.0
0.0  0.0  0.0  0.0  0.0  0.125  0.0625  0.0  0.0  0.0625  0.0  0.0625
0.0  0.0  0.0  0.0  0.125  0.0  0.5625  0.0
0.0625  0.0  0.0625  0.0  0.125  0.0625  0.0  0.0  0.0625  0.0625  0.0
0.0  0.0  0.0  0.0  0.3125  0.0  0.0  0.125  0.125
0.0  0.0625  0.125  0.1875  0.125  0.0625  0.125  0.0  0.0  0.0  0.0
0.125  0.0  0.0  0.0  0.125  0.0625  0.0  0.0  0.0
0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.9375  0.0  0.0625  0.0  0.0
0.0  0.0  0.0  0.0  0.0  0.0  0.0
```

ii. pwm_matched_sequences.txt

This file contains the occurrences of each motif. For example,

```
#pattern name   sequence name   start   stop    strand  score   p-value
q-value matched sequence
1       44|407|218|588|333      8       15      +       7.62025 9.91e-05
YSYVLLNM
1       55|257|322|777|104      1       8       +       12.7848 1.71e-05
WSGVWYDK
1       88|497|549|201|70       8       15      +       16.8861 1.4e-06
YESVHLHK
1       691|0|122|235|388       4       11      +       13.4557 1.25e-05
YFYEWDSK
```

iii. A directory named `logos`

This directory includes the logos of all the motifs.

- ## Examples

Assuming you are currently in the package directory `ASC+MEME` and there is an input fasta file `dataset/input.fa`, here are some examples of running the program.

- Running with default settings

  If you are just testing out the program, you can simply run the program with

```
python asc_meme.py dataset/input.fa
```

- Specify an output directory

```
python asc_meme.py dataset/input.fa -o your_output_dir
```

- Specify the number of clusters

  By default, ASC+MEME will run recursively to find the right number of clusters (similar to the number of motifs you want to find), but this could slow down the running process. So, it's better to specify this number. For example,

  ```
  python asc_meme.py dataset/input.fa -npar 50
  ```

- Working with DNA sequences

  By default, ASC+MEME uses protein alphabet. If you are working on DNA sequences, do the following

  ```
  python asc_meme.py dataset/input.fa -dna
  ```

- Change the default settings

  If you want to change the default settings of the program so that you don't need to provide the customized parameters every time, just change the content of the file `settings.txt`.

Options

For more options, check the following.

```
usage: asc_meme.py [-h] [-v] [-o <output dir>] [-text] [-dna] [-evt <ev>]
                   [-minsites <minsites>] [-minw <minw>] [-maxw <maxw>]
                   [-bfile  <bfile>] [-npar <npar>] [-seql <seql>] [-d <d>]
                   [-maxi maxi] [-ct <ct>] [-mt <mt>] [-cs <cs>] [-pa <pa>]
                   [-nsample <nsample>] [-nseq <nseq>] [-nsmotif <nsmotif>]
                   [-occ] [-re] [-klt <klt>] [-njobs <njobs>]
                   <dataset>

The pipeline of ASC+MEME.

positional arguments:
  <dataset>             file containing sequences in FASTA format

optional arguments:
  -h, --help            show this help message and exit
  -v, --verbose         verbose mode

common options:
  -o <output dir>       name of directory for output files will replace
                        existing directory
  -text                 output in text format without generating pwm logos
  -dna                  sequences use DNA alphabet (default is protein
```

```
                        alphabet)
  -evt <ev>             the threshold of E-value for a motif to be significant
                        (default: 0.01)
  -minsites <minsites>  minimum number of sites for each motif (default: 10)
  -minw <minw>          minumum motif width (default: 6)
  -maxw <maxw>          maximum motif width (default is equal to seql)
  -bfile  <bfile>       name of background Markov model file

asc related options:
  -npar <npar>          number of partitions (if not specified, the algorithm
                        will run recursively until it reaches termination
                        condition)
  -seql <seql>          length of sequences (if not specified, the algorithm
                        will set seql equal to the length of the shortest
                        sequence)
  -d <d>                # of anchors in each partition center (default: 5)
  -maxi maxi            maximum # of iterations (default: 1000)
  -ct <ct>              threshold of convergence criteria (default: 0.01)
  -mt <mt>              the threshold of the number of common anchors to
                        combine two partitions (default: 1)
  -cs <cs>              the convergence threshold will be added with this step
                        after each iteration (default: 0.00)
  -pa <pa>              penalty number for small partitions (default: 50)

postprocessing options:
  -nsample <nsample>    the number of samples for each partition (default: 2)
  -nseq <nseq>          the number of sequences per sample (default: 500)
  -nsmotif <nsmotif>    number of motifs to be discovered per sample (default:
                        3)
  -occ                  find occurrences for each motif
  -re                   re-calculate E-value for each motif
  -klt <klt>            the threshold of KL value to merge two motifs
                        (default: 1.5)
  -njobs <njobs>        number of MEME jobs that are run simultaneously
                        (default: 25)
```

Cite

If you use this software in your research, please cite the following paper:

Fast Motif Discovery in Short Sequences
Liu, Honglei, Fangqiu Han, Hongjun Zhou, Xifeng Yan, and Kenneth S. Kosik
Proc. of Int. Conf. on Data Engineering (ICDE 2016)