MotifOrganizer 1.0 beta testing
programmer: Steve Qin
Date: 04/15/07
Revised 01/26/08

1. Introduction.

MotifOrganizer is designed to cluster large set of DNA sequence motifs to identify common patterns. Different motif widths are allowed. It uses a two-stage divide-conquer-combine strategy to achieve scalability and efficiency.

2. Files in this package.

executables of C++ programs:

bmcvar - cluster motifs of different widths.

bmcfix - cluster motifs of the same widths.

varview - producing final clustering output files. it can also produce clean clustering results using Bayes ratio and posterior probability thresholds.

Perl script:

batch.pl - automate the whole analysis.

motifstack.pl combined motif sequences from clustering results into input motif file for second stage clustering analysis.

split.pl - used to partition the input motif data files into subsets to run the two stage clustering strategy.

allcombine.pl -combine cluster information (member, shift, width) from first stage clustering results into combined cluster information files.

seq.pl produce FATSA format results file, names, and pure aligned sequence files from clustering outputs.

Others:

sample.txt - sample input file.

readme - short description of this program.

3. Input file format.

We require that the input file is in FASTA format where different motif blocks are separated by an empty line. The first word after ">" is assumed to be the motif ID hence they are required to be the same within each motif block and assumed to be unique in the entire dataset. After the

motif ID, additional information can be added for which there is no restriction. More detailed annotation information can be added here. Note that no empty line is allowed at the end of the file. The following is a sample entry in an input file.

```
>motif0-0
GGACCAATAAGC
>motif0-0
TAACCAACCAGC
>motif0-0
CAGCCAATAAGA
>motif0-0
TGACCTATCACG
>motif0-0
CCGCCAATCGGC
>motif0-0
CAGCCAATCAAA
>motif0-0
CAACCGATCGAA

>motif0-1
TTTCCAATCAGA
>motif0-1
CGTCCAATAAGA
>motif0-1
GGGCCACTGCCT
>motif0-1
CGGCCAATCTGA
```

4. Syntax for executables.

bmcfix [input file name] [output summary file name] [membership summary file name] [number of Markov chains] [number of cycles in each chain] [number of initial cluster sizes]

there are 6 parameters that need to be specified, all are required. Among them,

input file name - string, name of an existing file contains all input motif sequences. See section 3. for input data format.

output summary file name - string, name of a new file to be created to store clustering result summary.

membership summary file name - string, name of a new file to be created to store clustering membership information of the result.

number of Markov chains - integer, number of independent Markov chains to be used in the MCMC procedure of this program. The default value is 5.

number of cycles in each chain - integer, number of times each motifs were reassigned in one chain in the MCMC procedure. The default value is 20.

number of initial cluster sizes - integer, number of motifs in each initial clusters. The total number of motifs divide this number is the number of initial clusters. The default value is 5.

bmcvar [input file name] [output summary file name] [membership summary file name] [width summary file name] [shift summary file name] [number of Markov chians] [number of cycles in each chain] [number of initial cluster sizes] [maximum allowed shift]

There are 9 parameters that need to be specified, all are required.

input file name - string, name of an existing file contains all input motif sequences. Please see section 3 for input data format.

output summary file name - string, name of a new file to be created to store clustering result summary.

membership summary file name - string, name of a new file to be created to store clustering membership information of the result.

width summary file name - string, name of a new file to be created to store widths of the result clusters.

shift summary file name - string, name of a new file to be created to store shifts of the result clusters.

number of Markov chains - integer, number of independent Markov chains to be used in the MCMC procedure of this program. The default value is 5.

number of cycles in each chain - integer, number of times each motifs were reassigned in one chain in the MCMC procedure. The default value is 20.

number of initial cluster sizes - integer, number of motifs in each initial clusters. Total number of motifs divide this number is the number of initial clusters. The default value is 5.

maximum allowed shift - integer, number of maximum base differences allowed for a motif to join a cluster of different width.

varresultview [input file name] [input membership summary file name] [input shift summary file name] [input width summary file name] [final output summary file name] [final cleaned output summary file name] [final cleaned membership summary file name] [posterior probability cut off value] [cluster Bayes ratio cutoff value]

There are 9 parameters that need to be specified, all are required.

input file name - string, name of an existing original motif sequence file contains all input motif sequences.

input summary file name - string, name of a new file to be created to store clustering result summary.

input membership summary file name - string, name of the existing file that contains all cluster membership information, typically is the membership output file obtained from running perl script <allcombine.pl>.

input shift summary file name - string, name of the existing file that contains all motif shift information (relative to the cluster it belongs to), typically is the shift output file obtained from running the perl script <allcombine.pl>.

input width summary file name - string, name of the existing file that contains all motif width information, typically is the width output file obtained from running the perl script <allcombine.pl>.

final output membership summary file name - string, name of a new file to be created to store the final clustering membership information.

final cleaned output summary file name - string, name of a new file to be created to store the final clustering result summary on the subset of motifs that passed the clean up criteria (for clusters, Bayes ratio larger than the pres-set threshold and for motifs, posterior probability larger than the pre-set threshold.

final cleaned output membership summary file name - string, name of a new file to be created to store the final clustering membership information on the subset of motifs that passed the clean up criteria (for clusters, Bayes ratio larger than the pres-set threshold and for motifs, posterior probability larger than the pre-set threshold.

posterior probability cutoff value - double float between 0 and 1, posterior probability cutoff. The default value is 0.5.

cluster Bayes ratio cutoff value - double float. Bayes ratio cutoff. The default value is 2.

5. Useage.

The command for running the perl script is

perl batch.pl [input filename] [number of Markov chains] [number of cycles in each chain] [number of initial clusters] [number of initial cluster sizes (used in combine step)] [maximum allowed shift]

There are 6 parameters that need to be specified.

input file name - string, name of an existing file contains all input motif sequences. Please see section 3 for input data format.

number of Markov chains - integer, number of independent Markov chains to be used in the MCMC procedure of this program. The default value is 5.

number of cycles in each chain - integer, number of times each motifs were reassigned in one chain in the MCMC procedure. The default value is 20.

number of initial clusters - integer, number of initial clusters. The default value is 5 (have to be less than the number of motifs in each input motif subset.

number of initial cluster sizes - integer, maximum number of motifs in each initial clusters. Total number of motifs divide this number is the number of initial clusters. The default value is 10.

maximum allowed shift - integer, number of maximum base differences allowed for a motif to join a cluster of different width. The default value is 0.

Example: to cluster the motifs in the sample data attached, one can issue the following command:

perl batch.pl sample 5 20 5 10 0 &

6. Output file format.

There are six files that will be generated from the batch.pl run.

bmc.out - raw outputs from various scripts and executables.

bmc.log - log file show status of intermediate jobs.

out.final - clustering result summary file.

out.final.clean - clustering result summary file, with lose cluster and poor-fit motifs removed.

member.final - membership summary. Each line corresponds to a cluster, with its members listed by their IDs, separated by white space.

member.final.clean - membership summary as in member.final, but with lose clusters and poor-fit motifs removed.

7. Credit.
This program is developed together with Misha Bilenky, Gordon Robertson, Steven Jones from Canada's BC Cancer Centre and Gang Su at the Bioinformatics Program at University of Michigan.

8. Contact.
For questions, comments, suggestions, please email Steve at qin@umich.edu