Minimal Parameter Tuning Guideline for Motif Discovery Program Evaluation & Parameter setting for the benchmark experiments

A supplementary material for
"Limitations and Potentials of Current Motif Discovery algorithms", Hu, Li, Kihara, Nucleic Acid Research, 2005

In most cases, the motivation of computational prediction of regulatory binding sites is to help experimental biologists to make informed decisions on experiment planning. There are more than 30 such motif discovery programs available, which makes it hard for them to make a choice. In addition, most of these programs have as many as up to 20 parameters that can be tuned to improve the prediction accuracy if more information about the input data can be fed into the program. However, it is not easy or realistic for the biologists to spend a lot of time to tune the unfamiliar program parameters or to collect other statistics/distribution data to set a good parameter.

Previous motif discovery program benchmarking study (Tompa et. al., 2005) allowed algorithm experts (in many case the authors of the algorithms) to tune their algorithms and do whatever pre- and post-processing to generate the predictions. This best scenario performance evaluation is good for understanding the algorithms themselves but is less relevant to the real-world prediction performance of these programs with ordinary program users. It is from this aspect that we propose to use the minimal-parameter-tuning guideline for setting the parameters of the benchmark study in this paper. So the reported performance should be looked at as the Lazy User's expected performance which can be potentially improved if more tuning is conducted.

Specifically, we did the following to set the parameters:

- Most of the parameters are set as default setting of the original algorithms.
- Some simple statistics is calculated to set the parameter values: for example, the –gc parameter of AlignACE can be easily computed by running the auxiliary program on the e.coli genome; also the Markov models can be easily computed also using the whole genome sequence
- Some basic biology knowledge is used to set the parameters. For example, the DNA regulatory binding site ranges from 5 to 20. We set it as 15.
- We don't specifically to calculate some parameters. For example, before one run the motif algorithm, he/she has no idea how many sites it covers in the sequence and whether each sequence has one or more or what is the estimated length of the motif. Although one can design an algorithms to scan all RegulonDB to get such a distribution of number of binding sites, for other organisms, these information may be not available at all.

In the following sections, we listed all the command line options of the five evaluated programs and all the parameter settings that may affect the prediction accuracy.

AlignACE 4.0 05/13/04

```
Usage: AlignACE -i seqfile (options)
 Seqfile must be in FASTA format.
Options:
 -numcols       number of columns to align (10)
 -expect        number of sites expected in model (10)
 -gcback        background fractional GC content of input sequence (0.38)
 -minpass       minimum number of non-improved passes in phase 1 (200)
 -seed          set seed for random number generator (time)
 -undersample   possible sites / (expect * numcols * seedings) (1)
 -oversample    1/undersample (1)
```

	Parameter	Value	Meaning with default values
1	-numcols	15	number of columns to align (10)
2	-expect	10 (default)	number of sites expected in model (10)
3	-gcback	0.5 calculated from e.coli whole genome	background fractional GC content of input sequence (0.38)
4	-minpass	200 (default)	minimum number of non-improved passes in phase 1 (200)
5	-seed	time(default)	set seed for random number generator (time)
6	-undersample	1 (default)	possible sites / (expect * numcols * seedings) (1)
7	-oversample	1 (default)	1/undersample (1)

BioProspector 4/15/04.

```
Usage: ./BioProspector -i seqfile (options)
 Seqfile must be in restricted FASTA format.
Options:
 -W    <motif width (default 10)>
 -o    <output file (default stdout)>
 -f    <background distribution file (default seqfile)>
 -b    <background sequence file (default input sequences)>
 -k    <force x order Markov in background (default 3)>
 -n    <number of times trying to find motif (default 40)>
 -r    <number of top motifs to report (default 5)>
 -w    <second motif block width for two-block motif (default 0)>
 -p 1  [if two-block motif is palindrome (default 0)]
 -G    <max gap between two motif blocks (default 0)>
 -g    <min gap between two motif blocks (default 0)>
 -d 1  [if only need to examine forward (default 2)]
 -a 1  [if every sequence contains the motif (default 0)]
 -h 1  [if want more degenerate sites (default fewer sites)]
 -e    <expected bases per motif site in the sequences
         (will use Bayes motif scoring, don't specify if unknown)>
```

	Parameter	Value	Meaning with default values
1	-W	15	motif width (default 10)
2	-f	3rd Order model built from whole E. coli genome	background distribution file (default seqfile)
3	-b	(default)	background sequence file (default input sequences)
4	-k	3 (default)	force x order Markov in background (default 3)
5	-n	40 (default)	number of times trying to find motif (default 40)
6	-r	5 (default)	number of top motifs to report (default 5)
7	-w	0 (default)	second motif block width for two-block motif (default 0)
8	-p	0 (default)	1 if two-block motif is palindrome (default 0)
9	-G	0 (default)	max gap between two motif blocks (default 0)
10	-g	Unset (default)	min gap between two motif blocks
11	-d	2 (default)	1 if only need to examine forward (default 2)
12	-a	0 (default)	1 if every sequence contains the motif (default 0)
13	-h	0 (default fewer sites)	if want more degenerate sites (default fewer sites)
14	-e	Unset (default)	expected bases per motif site in the sequences (will use Bayes motif scoring, don't specify if unknown)

MDScan

module usage:
-i <input sequences>
-w <motif width (default 10)>
-t <number of top sequences to look for candidate motifs (default 5)>
-c <number of top sequences to confirm candidate motifs (default whole dataset)>
-e <expectd bases per motif site in the top sequences (don't specify if unknown)>
-f <background frequency file (default yeast genome intergenic region)>
-b <background sequence file (default input sequences)>
-s <number of candidate motifs to scan and refine (default 30)>
-r <number of top motifs to report at the end (default 5)>
-n <number of refinement iterations (default 10)>
-o <output file (default stdout)>
-g 1 <if you don't want to see messages during the run>

	Parameter	Value	Meaning with default values
1	-w	15	motif width (default 10)
2	-t	5 (default)	number of top sequences to look for candidate motifs (default 5)
3	-c	Whole dataset(default)	number of top sequences to confirm candidate motifs (default whole dataset)
4	-e	unset (default)	expected bases per motif site in the top sequences (don't specify if unknown)
5	-f	3rd order Markov model generated from whole e. coli genome	background frequency file (default yeast genome intergenic region)
6	-b	Unset	background sequence file (default input sequences)
7	-s	30 (default)	number of candidate motifs to scan and refine (default 30)
8	-r	5 (default)	number of top motifs to report at the end (default 5)
9	-n	10 (default)	of refinement iterations (default 10)

MEME 3.0.4

```
USAGE:
      meme    <dataset> [optional arguments]
      <dataset>              file containing sequences in FASTA format
      [-h]                   print this message
      [-dna]                 sequences use DNA alphabet
      [-protein]             sequences use protein alphabet
      [-mod oops|zoops|tcm]   distribution of motifs
      [-nmotifs <nmotifs>]    maximum number of motifs to find
      [-evt <ev>]            stop if motif E-value greater than <evt>
      [-nsites <sites>]       number of sites for each motif
      [-minsites <minsites>]  minimum number of sites for each motif
      [-maxsites <maxsites>]  maximum number of sites for each motif
      [-wnsites <wnsites>]    weight on expected number of sites
      [-w <w>]               motif width
      [-minw <minw>]         minumum motif width
      [-maxw <maxw>]         maximum motif width
      [-nomatrim]            do not adjust motif width using multiple
                             alignment
      [-wg <wg>]             gap opening cost for multiple alignments
      [-ws <ws>]             gap extension cost for multiple alignments
      [-noendgaps]           do not count end gaps in multiple alignments
      [-bfile <bfile>]       name of background Markov model file
      [-revcomp]             allow sites on + or - DNA strands
      [-pal]                 force palindromes (requires -dna)
      [-maxiter <maxiter>]    maximum EM iterations to run
      [-distance <distance>]  EM convergence criterion
      [-prior dirichlet|dmix|mega|megap|addone]
                             type of prior to use
      [-b <b>]               strength of the prior
      [-plib <plib>]         name of Dirichlet prior file
      [-spfuzz <spfuzz>]      fuzziness of sequence to theta mapping
      [-spmap uni|pam]        starting point seq to theta mapping type
      [-cons <cons>]         consensus sequence to start EM from
      [-text]                output in text format (default is HTML)
      [-print_fasta]         print sites in FASTA format (default BLOCKS)
      [-maxsize <maxsize>]    maximum dataset size in characters
      [-nostatus]            do not print progress reports to terminal
      [-p <np>]              use parallel version with <np> processors
      [-time <t>]            quit before <t> CPU seconds consumed
      [-sf <sf>]             print <sf> as name of sequence file
also see http://www.psc.edu/general/software/packages/meme/meme_usage.html
```

	Parameter	Value	Meaning with default values
1	-dna	set as -dna	sequences use DNA alphabet
2	-protein	unset	sequences use protein alphabet
3	-mod oops|zoops|tcm	tcm	distribution of motifs
4	-nmotifs<nmotifs>	5	maximum number of motifs to find
5	-evt<ev>	Unset	stop if motif E-value greater than <evt>
6	-nsites<sites>	Unset	number of sites for each motif
7	-minsites<minsites>	sqrt(no. of sequences) (default)	minimum number of sites for each motif
8	-maxsites<maxsites>	Unset	maximum number of sites for each motif
9	-wnsites<wnsites>	Unset	weight on expected number of sites
10	-w<w> motifwidth	Unset	motif width
11	-minw<minw>	8 (default)	minumum motif width
12	-maxw<maxw>	50(default)	maximum motif width
13	-nomatrim	Unset	do not adjust motif width using multiple alignment
14	-wg<wg>	Unset	gap opening cost for multiple alignments
15	-ws<ws>	Unset	gap extension cost for multiple alignments
16	-noendgaps	Unset	do not count end gaps in multiple alignments
17	-bfile<bfile>	3rd order Markov model built from whole e. coli genome	name of background Markov model file
18	-revcomp	Unset	allow sites on + or - DNA strands
19	-pal	Unset	force palindromes (requires -dna)
20	-maxiter<maxiter>	50 (default)	maximum EM iterations to run
21	-distance<distance>	0.001 (default)	EM convergence criterion
22	-prior dirichlet|dmix|mega| megap|addone	Unset	type of prior to use
23	-b<b>	Unset	strength of the prior
24	-plib<plib>	Unset	name of Dirichlet prior file
25	-spfuzz<spfuzz>	Unset	fuzziness of sequence to theta mapping
26	-spmap uni|pam	Unset	starting point seq to theta mapping type
27	-cons<cons>	Unset	consensus sequence to start EM from
28	-maxsize<maxsize>	100000000	maximum dataset size in characters
29	-time<t>	3600	quit before <t> CPU seconds consumed

MotifSampler 3.0

```
Required Arguments
 -f <fastaFile>      Sequences in FASTA format
 -b <bgFile>         File containing the background model description
Optional Arguments
 -s <0|1>            Select strand. (default both)
                  0 is only input sequences, 1 include reverse complement.
 -p <value>          Sets prior probability of 1 motif copy. (default 0.2).
 -M <value>          Maximal number of motif instances per sequence. (default unset = 0)
 -n <value>          Sets number of different motifs to search for (default 4).
 -w <value>          Sets length of the motif (default 8).
 -x <value>          Sets allowed overlap between different motifs. (default 1)
 -r <runs>           Set number of times the MotifSampler should be repeated
                  (default = 1).
Output formatting Arguments
 -o <outFile>        Output file to write results (default stdout).
 -m <matrixFile>     Output file to write retrieved motif models.
Version 3.0
Questions and Remarks: <gert.thijs@esat.kuleuven.ac.be>
```

	Parameter	Value	Meaning with default values
1	-s	0 (only input sequence)	Select strand. (default both) 0 is only input sequences, 1 include reverse complement
2	-p	0.2	prior probability of 1 motif copy . (default 0.2).
3	-M	Unset 0(default)	Maximal number of motif instances per sequence. (default unset = 0)
4	-n	5	Sets number of different motifs to search for (default 4).
5	-w	15	Sets length of the motif (default 8).
6	-x	1 (default)	Sets allowed overlap between different motifs. (default 1)
7	-r	5	Set number of times the MotifSampler should be repeated (default = 1).