Last updated on Mar 26, 2013

This documentation file supports software published as part of:

Bo Jiang, Jun S. Liu and Martha L. Bulyk (2013) Bayesian hierarchical model of protein binding microarray *k*-mer data reduces noise and identifies transcription factor subclasses and preferred *k*-mers. *Bioinformatics*, in press.

Bayesian Analysis Suite of Protein Binding Microarray (PBM)

This analysis suite contains five components:

Note that the executables are compiled under Unix system, Macs versions of the software are compiled with a suffix *-mac.

I. Bayesian ANOVA decomposition of PBM *k*-mer data: a Bayesian ANOVA model for decomposing observed E-scores into background noise, TF-common effect and TF-preferred effect was implemented in C and complied as an executable “pbm.anova” (the source code is in the directory “anova_source”). We also provide an R script “anova.result.r” to analyze the output from running “pbm.anova” and input and output examples in the directory “anova_example”.

II. *k*-mer background noise correction: an R script “background.correction.r” for correcting observed 8-mer E-scores based on estimated background noise, with input and output examples in the directory “background_example”.

III. Bayesian partition model for identifying TF subclasses: a Bayesian hierarchical partition model for simultaneously inferring subclasses within a DBD class and identifying subclass-common *k*-mers was implemented in C and complied as an executable “pbm.partition” (the source code is in the directory “partition_source”). We also provide an R script “partition.result.r” to analyze the output from running “pbm.partition” and input and output examples in the directory “partition_example”.

IV. Position weight matrix and motif logo generation based on *k*-mer sequences: A Gibbs sampling algorithm to generate position weight matrix (PWM) based on 8-mer sequences was implemented in C and complied as an executable “pbm.motif” (the source code is in the directory “motif_source”). We also provide an R script “motif.logo.r” to generate PWM motif logos and input and output examples in the directory “motif_example”.

V. Statistical evaluation of TF-preferred *k*-mers: R scripts “preferred.analysis.r”and “preferred.run.r” evaluate the statistical significance of differentially preferred 6-mers (a 6-mer that is preferred by one TF but not the other). The script “preferred.analysis.r” contains the implementation of evaluation and plotting functions. The directory “preferred_example” gives input and output examples for using the R scripts. A python wrapper “PreferredKmerAnalysis.py” for the R script “preferred.cmd.r” (a command line version of “preferred.run.r”) is also provided.

Bayesian ANOVA decomposition of PBM k-mer data

(1) Guide for using “pbm.anova”:

Required arguments:

-i <input directory, containing family.txt, score.txt>

-o <output directory>

Optional arguments:

-m <number of burn-in iterations, default: 100>

-n <number of mcmc sampling iterations, default: 100>

-p <log prior for family specific binding, default: -6.0>

-q <log prior for tf specific binding, default: -6.0>

Example: ./pbm.anova -i anova_example/input/ -o anova_example/output/

Input format: The input directory (see /anova_example/input/) should contain the following files (with corresponding file names): (a) score.txt: a 32,896 by *n* (total number of TFs) matrix of observed E-scores; (b) family.txt: the first two rows indicate total number of TFs and total number of families, followed by the family membership (represented by the index of the TF family) of each TF; (c) kmers.txt: enumeration of all possible 8-mers (corresponding to the row order of observed E-scores in score.txt); (d) names.txt: names and corresponding family (DBD class) of each TF (corresponding to the column order of observed E-scores in score.txt and the indices of TFs in family.txt).

Output format: The output directory (see /anova_example/output/) after running "pbm.anova" will contain the following files: (a) trace.txt: log-likelihood trace from MCMC; (b) param.txt: posterior sample of parameters including sigma and gammas; (c) tau.txt: posterior sample of 8-mer background noises, i.e. τ_k in Eqn.1 of the paper (Jiang *et al.*, 2013); (d) omega1.txt: posterior sample of ω_f^+ in Eqn.1 of the paper (Jiang *et al.*, 2013); (e) omega2.txt: posterior sample of ω_f^- in Eqn.1 of the paper (Jiang *et al.*, 2013); (f) delta1.txt: posterior sample of δ_i^+ in Eqn.1 of the paper (Jiang *et al.*, 2013); (g) delta2.txt: posterior sample of δ_i^- in Eqn.1 of the paper (Jiang *et al.*, 2013); (h) familybindfreq.txt: posterior frequency of being preferred (TF-common) by a TF family; (i) familybindprob.txt: posterior of the probability (given parameters) of being preferred (TF-common) by a TF family; (j) familybindnum.txt: posterior sample of number of TF-common 8-mers preferred by a TF family; (k) familynobindfreq.txt: posterior frequency of being disfavored by a TF family; (l) familynobindprob.txt: posterior of the probability (given parameters) of being disfavored by a TF family; (m) familynobindnum.txt: posterior sample of number of 8-mers disfavored by a TF family; (n) tfbindfreq.txt: posterior frequency of being preferred (TF-preferred) by a TF; (o) tfbindprob.txt: posterior of the probability (given parameters) of being preferred (TF-preferred) by a TF; (p) tfbindnum.txt: posterior sample of number of (TF-preferred) 8-mers preferred by a TF; (q) tfnobindfreq.txt: posterior frequency of being disfavored by a TF; (r) tfnobindprob.txt: posterior of the probability (given parameters) of being disfavored by a TF; (s) tfnobindnum.txt: posterior sample of number of 8-mers disfavored by a TF.

(2) Guide for using "anova.result.r":

The R script "anova.result.r" helps to analyze the output from running "pbm.anova". The R script requires the installation of R package "gplots", which can be downloaded from CRAN (http://cran.r-project.org/web/packages/gplots/index.html). Here are the steps to run "anova.result.r":

Step1: In the R script, set the working directory, which contains the software directory "Bayesian_PBM_Analysis".

Step2: In the R script, set the input directory and the output directory when running "pbm.anova".

Step3: Set the family index and TF index to identify TF-common and TF-preferred *k*-mers (or set two TF indices to identify *k*-mers preferred by the first TF but not the second TF).

Run: R CMD BATCH anova.result.r (or run the script "anova.result.r" in the R console).

Output: (a) Observed E-score heatmap (top_50_sticky_heatmap.pdf) and sequences (sticky_kmers.seq) of top 50 'sticky' 8-mers; (b) Observed E-score heatmap (tf_common_heatmap.pdf) and sequences (tf_common_kmers.seq) of TF-common 8-mers for the TF family specified by the TF family index; (c) Observed E-score heatmap (tf_preferred_heatmap.pdf) of TF-preferred 8-mers for the TF specified by the TF index; (d) Observed E-score heatmap (tf_distinct_heatmap.pdf) of 8-mers by one TF but not the other specified by two TF indices.

k-mer background noise correction

Guide for using "background.correction.r":

The R script "background.correction.r" calculates corrected 8-mer E-scores based on estimated background noise. Here are the steps to run "background.correction.r":

Step1: In the R script, set the working directory, which contains the software directory "Bayesian_PBM_Analysis".

Step2: In the R script, set the input directory which contains estimated *k*-mer background noise file (see "estimated_noise.txt" in "background_example") and observed 8-mer E-score files from PBM experiments, and set the names of these files. **Note that the observed 8-mer E-score file must have the first four columns in the order of 8-mer, 8-mer (reverse complementary), median intensity and observed E-score.**

Step3: Set the output file name to store the corrected E-scores (the output format will be the same as the input observed 8-mer E-score file) and the threshold for selecting specific binding 8-mers (default is 0.35).

Run: R CMD BATCH background.correction.r (or run the script "background.correction.r" in the R console).

Output: (a) specific binding 8-mer sequences with observed E-score greater than the threshold (observed.kmers.seq); (b) specific binding 8-mers sequences (with the same number of sequences as in observed.kmers.seq) after background noise correction (corrected.kmers.seq); (c) Histogram of 8-mer background noises (background_noise_hist.pdf); (d) Scatter plots of observed E-scores and 8-mer background noises (background noise_plot.pdf), in which a vertical line indicates the threshold and red points indicate specific binding 8-mers after correction.

Bayesian partition model for identifying TF subclasses

(1) Guide for using "pbm.partiton":

Required arguments:

-i <input file>

-o <output directory>

-n <total number of TFs>

Optional arguments:

-m <total number of k-mer groups, default: 100>

-t <number of burn-in iterations, default: 100>

-s <number of mcmc sampling iterations, default: 1000>

Example:

```
./pbm.partition -i partition_example/score_homeo.txt -o partition_example/output_homeo/ -n 173
```

Input format: A file (see /partition_example/score_homeo.txt) containing a 32,896 by *n* (total number of TFs) matrix of observed E-scores.

Output format: The output directory (see /partition_example/output_homeo/) after running "pbm.parition" will contain the following files: (a) scale.score.txt: standardized observed E-score; (b) trace.txt: log-likelihood trace from MCMC; (c) resulty.txt: posterior probability of a *k*-mer group being subclass-common, $i.e., I_g$ in Eqn.2 of the paper (Jiang *et al.*, 2013); (d) resultc.txt: posterior sample of subclass memberships of TFs $i.e., C_i$ in Eqn.2 of the paper (Jiang *et al.*, 2013); (e) resultyg.txt: posterior sample of *k*-mer group membership; (d) resultcy.txt: sample of subclass membership of TFs from burn-in period; (f) resultgy.txt: sample of *k*-mer group membership from burn-in period.

(2) Guide for using "partition.result.r":

The R script "partition.result.r" helps to analyze the output from running "pbm.partitoin". The R script requires the installation of R package "gplots", which can be downloaded from CRAN (http://cran.r-project.org/web/packages/gplots/index.html). Here are the steps to run "partition.result.r":

Step1: In the R script, set the working directory, which contains the software directory "Bayesian_PBM_Analysis".

Step2: In the R script, set directory that contains the input file and the output directory when running "pbm.partition".

Step3: Set input files names including a file containing the observed E-score matrix (see "/partition_example/score_homeo.txt"), a file containing enumeration of all possible 8-mers (corresponding to the row order of the observed E-scores matrix; see "/partition_example/kmers.txt") and a file containing names and corresponding family (DBD class) of each TF (corresponding to the column order of the observed E-scores matrix; see "/partition_example/names_homeo.txt").

Step4: Set the output directory when running "pbm.partition"

Run: R CMD BATCH partition.result.r (or run the script "partition.result.r" in the R console).

Output: (a) Observed E-score heatmap (subclass_heatmap.pdf) of subclass-common 8-mers with hierarchical clustering of TFs into subclasses; (b) 8-mer sequences ("kmers.group**X**.seq") from each *k*-mer group X. Note that *k*-groups are in the numerical order (indexed by the number **X** in "kmers.group**X**.seq") from left to right in the heatmap generated in (a).

Position weight matrix and motif logo generation based on k-mer sequences

(1) Guide for using "pbm.motif":

Required arguments:

-i <input file>

Optional arguments:

-m <number of burn-in iterations, default: 100>

-n <number of mcmc sampling iterations, default: 100>

Example: ./pbm.motif -i motif_example/kmers.seq

Input format: A file containing 8-mer sequences in rows (see /motif_example/kmers.seq)

Output format: A text file "seqalign.txt" that contain the aligned sequences and a .pwm file that contains the position weight matrix (PWM) with width 10. To use a different width of PWM, one can change the parameter "nL" in the source code "/motif_source/pbm.motif.c".

(2) Guide for using "motif.logo.r":

The R script "motif.logo.r" plot motif logo based on position weight matrix generated by "pbm.motif". It requires the installation of R package "seqLogo" and "Biostrings". To install these packages, start R and enter: source("http://bioconductor.org/biocLite.R"); biocLite("Biostrings"); biocLite("seqLogo"). Here are the steps to run "motif.logo.r":

Step1: In the R script, set the working directory, which contains the software directory "Bayesian_PBM_Analysis".

Step2: In the R script, set directory that contains the input PWM file and the name of the file

Run: R CMD BATCH motif.logo.r (or run the script "motif.logo.r" in the R console).

Output: motif logo (motif_logo.pdf).

Statistical evaluation of TF-preferred k-mers

(1) Guide for using "preferred.run.r":

The R script "preferred.run.r" evaluate the statistical significance of differentially preferred 6-mers (a 6-mer that is preferred by one TF but not the other). Here are the steps to run "preferred.run.r":

Step 1: Set working director that contains source code (preferred.analysis.r), 6-mer sequences (kmer.6.RData) and files of 8-mer scores to be analyzed, and set input and output directory.

Step 2: Set "file1" and "file2" to be the name of the files that contain observed E-scores to be compared. The first three columns of the files need to be: 8-mer, 8-mer(reverse complementary), and E-score. Note that the first row of files are assumed to be a header line and the file contents are separated by tab. Some common file types that may be used as input are XXXXX_contig8mers.txt and XXXXX_8mers_11111111.txt.

Step 3(Optional): Set "mu1", "mu2",and "p.adjust.method" in preferred.pvalue() function. Note that "mu1" and "mu2" control the practical significant level, "mu2" may need to be adjusted to 0.5 if practical significance is too small. p.adjust.methods should be one of "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none" ("fdr" by default)

Step 4(Optional): Specify "kmers.6", a set of 6-mers to be plotted by preferred.plot() function.

Step 5(Optional): Specify "xlab","ylab","xlim","ylim" and "main" in preferred.plot() function.

Run: R CMD BATCH preferred.run.r (or run the script “preferred.run.r” in the R console).

Output: (a) Scatter plot showing top 3 differentially preferred 6-mers by each TF (preferred_analysis_plot.pdf); (b) Two text files “preferred.tf1.txt” and “preferred.tf2.txt” contain the sequences and p-values of significantly preferred 6-mers TF1 and TF2, respectively (c) R object “result” output from function “preferred.analysis”: result$preferred.1 and result$preferred.2 shows the significant preferred 6-mers by the 1st and 2nd experiments, respectively. The attributes of result$preferred.1 and result$preferred.2 include: (i) six.mer and six.mer.rc: 6-mers with adjusted p-value lower than result$level; (ii) p.value: original p-values; (iii) p.adjust: p-values adjusted by result$p.adjust.method; (iv) mean: standardized mean of the corresponding (1st or 2nd) experiment; (v) diff:standardized difference between 1st and 2nd (1st-2nd or 2nd-1st) experiments.

(2) Guide for using “PreferredKmerAnalysis.py”

PreferredKmerAnalysis.py provides a user friendly access to the R script “preferred.cmd.r” (a command line version of “preferred.run.r”), giving users feedback about which arguments are necessary or missing.

Usage: python PreferredKmerAnalysis.py [-h] [-N N] tf1_8mer_file.txt tf2_8mer_file.txt output_file.png name1 name2 R_script_path

Arguments:

N: plot top N differentially preferred 6-mers by each TF

tf1_8mer_file.txt, tf2_8mer_file.txt: observed E-score file of the first and the second TF, respectively

output_file.png: the name of the output figure containing the scatter plot

name1, name2: the names of the first and the second TF, respectively

R_script_path: absolute of R scripts (the directory containing “preferred.cmd.r”, “preferred.analyssi.r” and “kmer.6.Rdata”)

Questions, comments or suggestions regarding the materials presented here should be directed to Martha L. Bulyk (mlbulyk at receptor.med.harvard.edu).