User Manual for DM-BLD (Matlab Demo V2.0)

Introduction

Differential **M**ethylation detection using a hierarchical **B**ayesian model exploiting **L**ocal **D**ependency, namely DM-BLD, is a computational approach for the identification of differentially methylated genes based on a Bayesian framework. DM-BLD detects DNA methylation changes in functional regions closely associated with genes (from promoter regions to 3’UTR) under the hypothesis that genes involving a sequence of CpG sites with methylation change are more likely to exhibit abnormal methylation activity. Normalization and other required preprocessing steps should be carried out before the detection of differentially methylated genes.

Three major steps of the proposed method, DM-BLD, are summarized as follows: (1) estimating the true methylation level of CpG sites by modeling the local spatial correlation of methylation level and the dependency of methylation change among neighboring CpG sites; (2) calculating the differential methylation score of genes from the estimated methylation change of CpG sites; (3) performing permutation-based significance tests on the genes. Specifically in the first step, the DM-BLD approach features a joint model to capture both the local dependency of measured loci and the variability of methylation in samples. Specifically, the local dependency is modeled by Leroux conditional autoregressive (CAR) structure; the dependency of methylation changes is modeled by a discrete Markov random field for differential analysis. A hierarchical Bayesian model is developed to fully take into account the local dependency for differential analysis, in which differential states are embedded as hidden variables. A Gibbs sampling procedure, based on conditional distributions, is designed to estimate the true methylation level and other model parameters. In the second step, the differential methylation score of a gene is calculated from the estimated methylation change of involved CpG sites. Genes can be prioritized according to the differential methylation score for further biological validation. Finally in the third step, permutation-based hypothesis tests are implemented and performed to assess the significance of the identified differentially methylated genes for real data analysis.

Citation

Xiao Wang, Jinghua Gu, Leena Halakivi-Clarke, Robert Clarke, Jianhua Xuan, “DM-BLD: Differential methylation detection using a hierarchical Bayesian model exploiting local dependency”.

Requirement

The Matlab package of DM-BLD method was tested under Windows7 64bit Matlab R2012b, Matlab R2014a and Ubuntu 10.04 64bit Matlab R2012b, Matlab R2014a. Rscript "BMIQ.R" is used for real data preprocessing.

Usage

I. Pre-processing of location information

The association and location information between CpG sites and genes is first generated from the annotation of the profiling technique. Specifically, CpG sites are mapped to the corresponding genes. For each gene, the information of the involved CpG sites is saved by the data format in Table 1.

Table 1. Data format for mapping between CpG sites and genes

Data field	Description
gene_symbol	Official gene symbol
probe_id	ID of the measured CpG site
CpG_perGene_sorted	ID of the involved CpG sites sorted by genomic location
idx_site_perGene_sorted	Index of the CpG site in probe_id
W	A binary matrix indicating the neighborhood information of the CpG sites
num_site_perGene	Number of CpG sites involved in the gene
funcLoc_perGene_sorted	Functional location of the CpG sites (1: before TSS; 2. 5'UTR; 3. first Exon; 4. Body; 5. 3'UTR)

For Illumina Infinium HumanMethylation450 BeadChip Kit (Illumina 450k), the processed association and location information between CpG sites and genes is provided by 'CpG_gene.mat' in folder named 'Location_info'.

II. Simulation study

Step 1. Simulate methylation data

Run 'Generate_simData.m' in the folder named 'codes' to generate simulation data. The simulated methylation data is saved in 'input_demo.mat'. In specific, with the processed association and location information between CpG sites and genes ('./Location_info/CpG_gene.mat'), we simulate methylation data on a set of randomly selected genes in the following steps:

a) randomly select a subset of gene to be differentially methylated;

b) for each differentially methylated gene, randomly select a neighorhood of CpG sites to be differentially methylated;

c) generate methylation data following the Leroux model with the parameter settings in Table 2.

In the demo, more than 10 CpG sites are involved in each of the simulated genes.

Table 2. Parameter setting for simulation data

Parameter	**Value**	**Description**
G	500	Number of genes to be simulated
DEGPCENT	0.7	Percentage of differentially methylated genes
J1	10	Number of samples in phenotype 1
J2	10	Number of samples in phenotype 2
τ_e	1	Precision parameter of normal distribution for methylation data **Y**
τ	1	Parameter of precision of the Leroux model for basal methylation **θ**
ρ	0.3	Parameter of dependency for the Leroux model for basal methylation **θ**
μ_0	0.7	Methylation change of gene

As a result, the methylation level of the CpG sites in each gene of samples in two phenotypes will be generated and saved in cell matrix **Y**.

Step 2. Run DM-BLD on the simulation data

[gene_score, gene_site, H] = ***func_DM_BLD_all****(Y, W_slt, gene_symbol_slt, ite, J1,J2, num_run);*

The inputs to the ***func_DM_BLD_all*** function are:

Y	A $G \times 1$ cell matrix; each cell contains a $M_n \times J$ matrix which is the methylation value of the M_n involved CpG sites in J samples of the two phenotypes. In specific, each row corresponds to a CpG site and each column corresponds to a sample. Samples of the same phenotype are grouped together.
W	A $G \times 1$ cell matrix; each cell contains a $M_n \times M_n$ binary matrix indicating the neighborhood information of the CpG sites. $w_{i,j} = 1$, if CpG site i and CpG site j locate within a neighborhood (defined as 1000 base in this demo); $w_{i,j} = 0$, otherwise.
Ite	Number of iterations of the Gibbs sampling in each run.
J1	Number of samples in phenotype 1
J2	Number of samples in phenotype 2

num_run Number of runs of the Gibbs sampling

In the ***func_DM_BLD_all*** function, genes are divided into two groups: the first group consists of the genes with neighbored CpG sites, i.e. sum(sum(**W**{i_gene}))>0; the second group consists of the genes with isolated CpG sites which does not form any neighborhood. The first set of genes is estimated by **func_DMBLD_withNeighbor** with Gibbs sampling designed for the Leroux model; the second set of genes is estimated by **func_DMBLD_noNeighbor** with Gibbs sampling designed for a simplified Bayesian model for independent CpG sites.

For Gibbs sampling, the software provides two types of implementations: a single run and multiple random runs, specified by argument 'num_run'.

- If num_run = 1, a single run of Gibbs sampling with 'ite' samples will be conducted. Every other sample after the burn-in period is recorded for the estimation of the parameters.
- If num_run > 1 (typically num_run = 5), 'num_run' runs of Gibbs sampling with random initiation of the parameters as well as different random seeds will be conducted. At the end of the multiple runs, the distributions generated from the independent runs will be checked to see whether a specific number of different runs (typically 3 times) produce samples from the same distribution. If so, all of the samples from all runs will be used for the estimation of the parameters. If not, another set of fixed number of runs will be conducted continuously.

If num_run is not provided, num_run = 1.

The outputs to the ***func_DM_BLD_all*** function are:

gene_site Index of detected reprehensive CpG sites for the gene
gene_score Differential methylation score of the gene
H Direction of methylation change: H = 1, if phenotype 2 > phenotype 1; H = -1, otherwise.

Finally, the genes are prioritized according to the estimated differential methylation score.

Step 3. Evaluate the performance

The genes are sorted by the estimated differential methylation scores, and the ROC curves of the demo experiment using one single run and 5 independent runs are shown by Fig. 1.

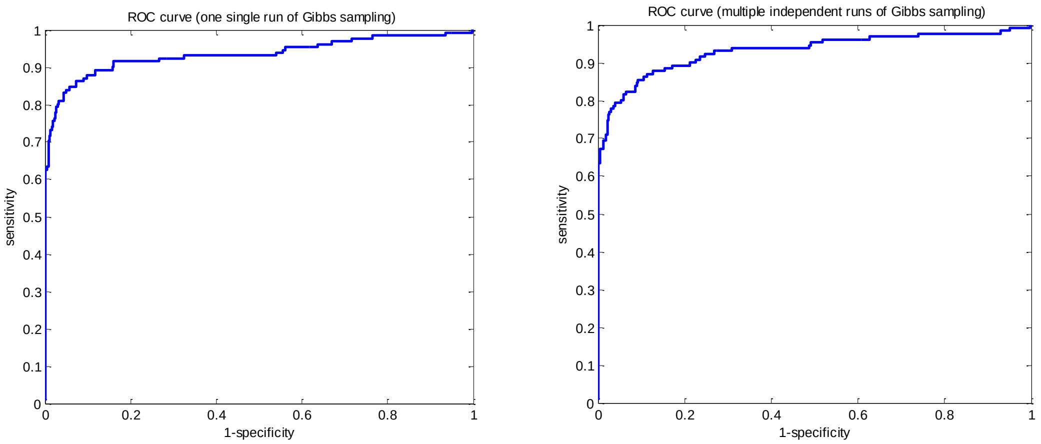

Fig. 1 ROC curves of the demo experiment for DM-BLD using one single run (left) and multiple independent runs (right) of Gibbs sampling.

III. Pipeline for real data analysis

To run DM-BLD on real data profiled by Illumina 450K, the methylation data measured by beta values should first be corrected for two types of probes. The pipeline for real data analysis includes the following steps.

Step 1. Preprocessing of methylation data

In this study, we used BMIQ (a R package) for the preprocessing of data. R script “BiasCorrection.R” is used to correct the group of samples by calling “BMIQ.R”.

The inputs to “BiasCorrection.R” are:

- “methylaiton_data.txt”: the matrix of raw methylation data (beta value), where each raw corresponds to a CpG site and each column corresponds to a sample.
- “CpGsiteType.txt”: the information of probe type of the CpG sites subtracted from the annotation file.

The output of “BiasCorrection.R” is:

- “methylation_data_corrected.txt”: the matrix of processed methylation data after bias correction. CpG ID is saved in the first column; sampleID is saved in the first row, named following the pattern “cond*****_sample*****”.

Step 2. Run DM-BLD for differential methylation analysis

We apply DM-BLD to the processed methylation data by running “DM_BLD_RealData.m”.

The inputs of “DM_BLD_RealData.m” are:

- “CpG_gene.mat”: the association and location information between CpG sites and genes.
- “methylation_data_corrected.txt”: the matrix of corrected methylation data.

The results are saved in the variable “result” with five columns specified as follows:

Gene Symbol	Differential direction	CpG sites	adjusted_Pval_local	adjusted_Pval_global

The significance of the differential level of the genes is derived from permutation tests. Specifically, the sample labels as well as the location of the CpG sites were rearranged for 100 times, and then DM-BLD is applied onto the perturbed methylation data. The ‘global’ and ‘local’ null distributions of the differential methylation score of genes are generated from the random trails, from which the p-values of differentially methylated genes can be calculated. The ‘global’ null distribution is generated from all the genes in consideration, while the ‘local’ null distribution is generated for each individual gene. Benjamini-Hochberg correction is further used for multiple testing correction to obtain the adjusted p-values (adjusted_Pval_local and adjusted_Pval_global).