User Manual for BayesIso 1.0.0 demo (Matlab)

Introduction

Bayesian identification of differentially expressed isoforms using a novel joint model (BayesIso) is a computational approach for the identification of differentially expressed isoforms through a sampling procedure. As a joint model, BayesIso, with differential state of the isoforms as model parameters, is developed to account for the intrinsic complicate variability of RNA-seq data at isoform level, taking into consideration that different isoforms of the same gene may have different variability. In specific, within-sample variability and between-sample variability of each transcript are modeled by a Poisson-Lognormal model and a Gamma-Gamma model, respectively. Under the Bayesian framework, the differential state of each transcript as well as other model parameters are estimated jointly through a Markov Chain Monte Carlo (MCMC) process.

Citation

Xiao Wang, Xu Shi, Jinghua Gu, Ayesha N Shajahan-Haq, Leena Halakivi-Clarke, Robert Clarke, Jianhua Xuan, "BayesIso: Bayesian identification of differentially expressed isoforms using a novel joint model of RNA-seq data".

Requirement

The Matlab package of BayesIso method was tested under Windows7 64bit Matlab 2012b and Ubuntu 10.04 64bit Matlab 2012b and Matlab 2014a. The package also includes several external Matlab packages and functions for string operation and numerical calculations, which are:

'string' package: includes functions such as strsplit.m;

alogam.m: computes the logarithm of the Gamma function;

Usage

- ***Pre-processing of annotation information***

The processed annotation information for human isoform structure is provided by 'Isof_structure_hg19.mat' in folder 'annotation_info'. The details to generate the annotation information are as follows.

The original annotation information for human isoform structure in 'bed' format was downloaded from UCSC table browser (http://genome.ucsc.edu/cgibin/hgTables). For the implementation of BayesIso, the annotation information was processed by: 1. The overlapped exons were split into non-overlaped 'sub-exons'; 2. Genes were sorted according to the number of involved isoforms. Then, the processed annotation information for human was saved as 'hg19.gene.post.bed' in the folder 'annotation_info', which can be further processed by 'generate_isoform_structure.m' to formulation annotation information in '.mat' format.

- ***Simulation study***

Run demo_BayesIso_sim.m to generate a demo simulation experiment according to the following steps.

1. Simulate over-dispersed RNA-seq counts

With the processed annotation information for isoform structure ('isof_structure_forDemo.mat'), we simulate over-dispersed RNA-seq counts on a randomly selected a set of genes with the parameter settings in Table 1.

Table 1. Parameter setting for simulation data

Parameter	Value	Description
tau	1.78	Within-sample over-dispersion parameter
alpha	1	Gamma shape parameter for β
alpha0	0.5	Gamma shape parameter for λ
v	0.1	Gamma rate parameter for λ
J1	10	Number of samples in phenotype 1
J2	10	Number of samples in phenotype 2
Num_genes	200	Number of genes to be simulated
K	2	Number of isoforms per gene
DEGPCENT	0.5	Percentage of differentially expressed isoforms

2. Implement BayesIso using:

[beta_e1, d_e] = func_BayesIso(Y0, X0, I,G,K,J1,J2,Result_dir)

The inputs to the func_BayesIso function are:

G	Number of genes.
Y0	$G \times 1$ Cell array. Each cell is a matrix of simulated counts for a gene, where each row corresponds to an exon and each column corresponds to a sample.
X0	$G \times 1$ Cell array. Each cell is a matrix indicating isoform structure of a gene, where each row corresponding to an exon; each column corresponds to an isoform; and each element is the length of the exon if involved in the corresponding isoform.
I	$G \times 1$ Vector indicating the number of exons for the selected genes.
K	Number of isoforms per gene.
J1	Number of samples in phenotype 1
J2	Number of samples in phenotype 2
Result_dir	Directory to save the results.

The outputs to the func_BayesIso function are:

beta_e1	Estimated abundance for each isoform
d_e	Estimated probability that the isoform is differentially expressed
result.mat*	Intermediate results from the sampling procedure.

3. Evaluate the performance

The ROC curve of the demo experiment is shown by Fig. 1 with AUC=0.8723. With threshold 'Prob(d_e)>0.75', precision = 0.939, recall = 0.695, F-score=0.800. Fig. 2 shows the histogram of estimated probability for differential expression.

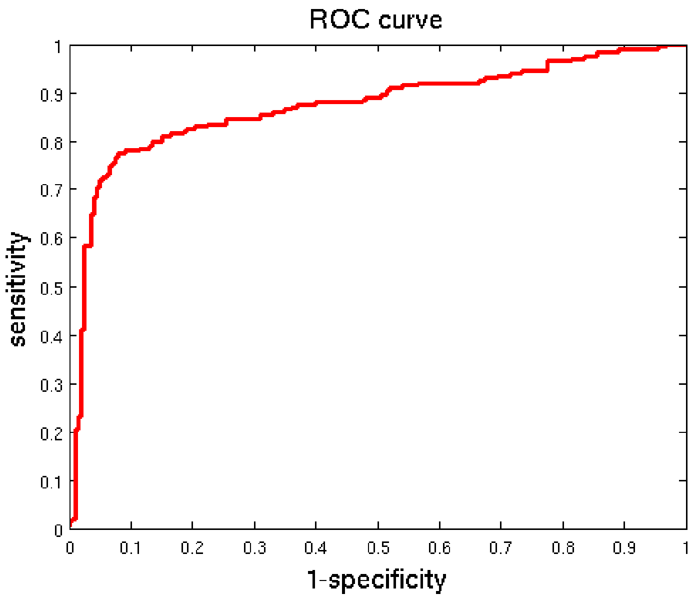

Fig. 1 ROC curve of demo experiment for differential isoform identification

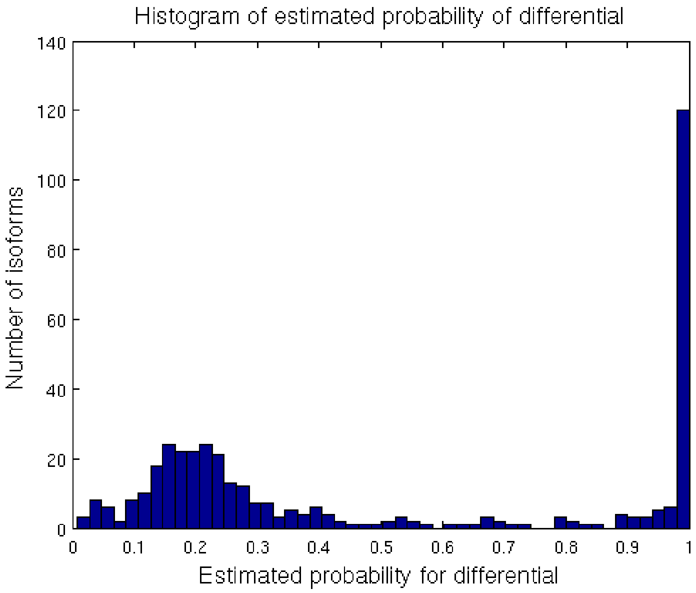

Fig. 2 Histogram of estimated probability for differential expression.