GibbsOS User's Manual

Introduction

GibbsOS is a Matlab package to infer Transcriptional Regulatory Networks (TRN) by integrating gene expression data with ChIP-on-chip or motif binding information. The main purpose of developing GibbsOS is to identify the true target genes for the given transcription factors under certain condition, which are supported by evidences from both expression profiles and binding information. We adopted a novel statistic, namely Outlier Sum of regression t-statistic to evaluate the quality of the target genes' expressions to represent the activity of its regulator. To tackle complicated cases that one target gene may be regulated by multiple transcription factors, GibbsOS employed a sampling scheme to estimate the marginal distribution of the proposed Outlier Sum statistic from its conditional distribution. The software package is available at http://www.cbil.ece.vt.edu/software.htm.

Reference

Jinghua Gu, Jianhua Xuan, Li Chen, Rebecca B. Riggins, Robert Clarke and Yue Wang, "Robust Identification of Transcription Regulatory Networks Using a Gibbs Sampler on Outlier Sum Statistic", *Bioinformatics*, (doi: 10.1093/bioinformatics/bts296; first published online: May 17, 2012), 2012.

Jinghua Gu, Jianhua Xuan, Yue Wang, Riggins R. B. and Clarke, R., "Identification of Transcriptional Regulatory Networks by Learning the Marginal Function of Outlier Sum Statistic", *Proc. 9th International Conference on Machine Learning and Applications*, 2010, page: 281-286.

System Requirement

GibbsOS has been developed on Microsoft Windows XP with 2GB of RAM. The required Matlab version is v7.0.4 or higher.

USAGE

Function: [seed, os]=gibbs_sampling(E, A, ITE); // Gibbs sampling

Argument:

Table 1

Name	Description
E	Expression matrix whose N rows correspond to N genes and M columns correspond to M samples. **It is required by the algorithm to take logarithm of raw expression data and standardize gene expression row by row (for each row, subtract the mean and divide by the standard deviation).**
A	Connectivity matrix whose N rows correspond to N genes and L columns correspond to L transcription factors. If there is evidence showing that gene i is regulated by TF j, A_{ij}=1; otherwise A_{ij}=0.
ITE	Number of iterations. Default value: 1000.
seed	Sampled seed gene index in every iteration for each transcription factor. See more description in Table 5.
os	Outlier sum estimated in each iteration using the Gibbs sampler.

Function: [E, A, pid, tf]=cell_line_up_early(TF)*; // preprocessing BC cell line data

Argument:

Table 2

Name	Description
E	Expression matrix whose N rows correspond to N genes and M columns correspond to M samples.
A	Connectivity matrix whose N rows correspond to N genes and L columns correspond to L transcription factors. If there is evidence showing that gene i is regulated by TF j, A_{ij}=1; otherwise A_{ij}=0.
pid	Affymetrix probe id of the target genes.
TF	Input transcription factor name.
tf	Output transcription factor name. TFs with no target genes will be removed.

*: Several .mat files are required in order to run the preprocessing for breast cancer cell line data, which are:

Table 3

cell_line_all_up_early.mat	Expression profile of breast cancer cell line under estrogen induced, early up-regulated condition (Creighton *et al*). The genes are identified by Affymetrix probe ids.
prob2bnm133A.mat	Mapping Affymetrix U133A probe ids to RefSeq ids.
hg2_MM.mat	Motif matrix for homo sapiens within 2K bases from the promoter region. The entries are identified by RefSeq ids.

Function: p=null_distribution(E, A, os, alpha); // Significant test of TRNs

Argument:

Table 4

Name	Description
p	p-value for each transcription factor (or TRN).
alpha	alpha=1.96 corresponds to 95% confidence level in two-tailed t-test.

Function: convergence_check (seed, bp1, bp2, A); // Check convergence and plot CDFs

Argument:

Table 5

Name	Description
seed	Sampled seed gene index in every iteration for each transcription factor.
bp1	Break point 1. Default bp1=0.
bp2	Break point 2. Default bp2=ITE/2.
A	Connectivity matrix whose N rows correspond to N genes and L columns correspond to L transcription factors.

Function: [sens, spec]=ROC (true_tar, false_tar, gene_rank); // Calculate sensitivity and specificity

Argument:

Table 6

Name	Description
true_tar	True target genes
false_tar	False target genes
gene_rank	Ranked gene list according to certain criterion, e.g., sampling frequency.
sens	sensitivity
spec	specificity

Function: [E, A, At, tLabel, plabel]=synthetic_data_gen(N, M, L, SNR, p,th); // Generate synthetic data

Argument:

Table 7

Name	Description
N	Number of foreground genes
M	Sample size
L	Number of TFs
SNR	Signal-to-noise ratio.
p	Background gene ratio. Number of background gene is p*N. Due to the randomness, the actual ratio may vary.
th	Degree parameter. The larger th is, the sparser the network is.
E	Synthetic gene expression.
A	Synthetic connectivity matrix of all genes.
At	Connectivity matrix of true target genes.
tLabel	Gene label. 1: foreground gene; 0: background gene.
pLabel	Permuted gene label.

Examples

Demo file 1: main_CL_early.m will call the following functions: cell_line_up_early(TF), gibbs_sampling(E, A, ITE) and null_distribution(E, A, os, alpha) in succession. A sample set of 26 transcription factors are defined at the beginning of the m-file. Users can replace this set by their own transcription factors of interest.

Demo file 2: main_synthetic.m will call the following functions:

synthetic_data_gen(N,M,L,SNR,p,th);

gibbs_sampling(E, A, ITE)

ROC(true_tar,false_tar,gene_rank);

convergence_check(seed1,0,ITE/2,A);

Results

Running GibbsOS demo program will yield the following results:

Graphical results:

1. A hierarchical clustering of the expression data E will be conducted while running the preprocessing function cell_line_up_early(TF), as is shown in Fig. 1.

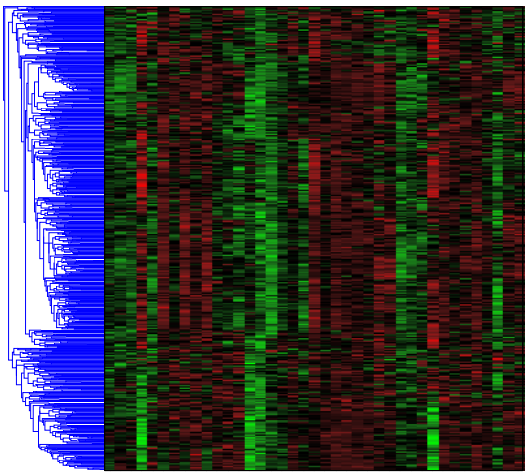

Fig. 1 Heatmap of breast cancer cell line data

2. The function: convergence_check will plot the CDFs using half earlier samples and half newer samples. Fig. 2 is an example of synthetic data for ITE=1000.

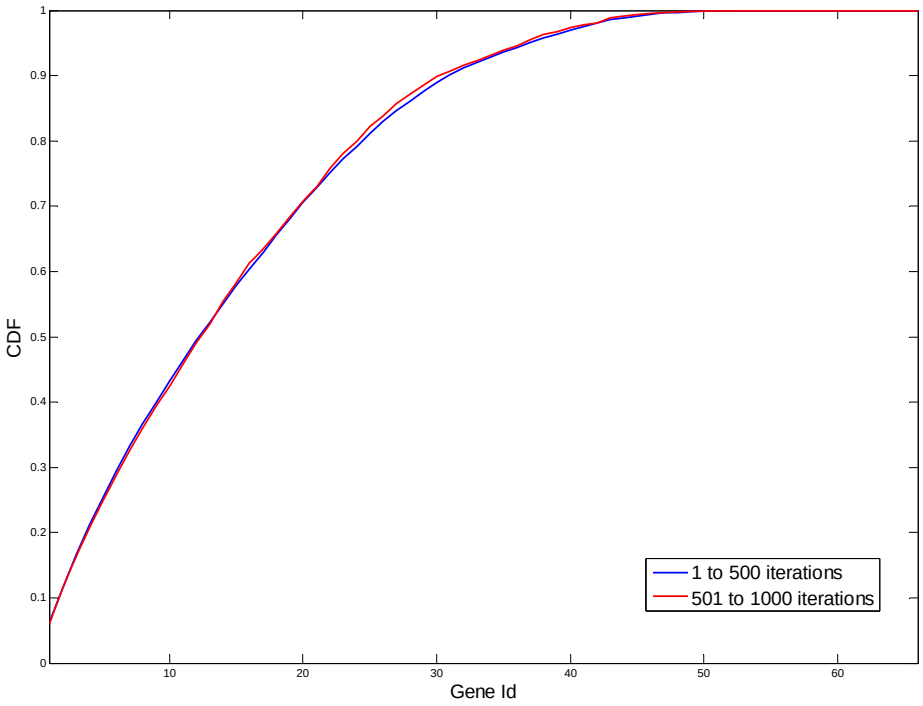

Fig. 2 CDF plot for convergence check

3. We also plotted the ROC figure for synthetic data generated using N=50, M=30, L=10, p=2, th=0.9 and SNR=2.

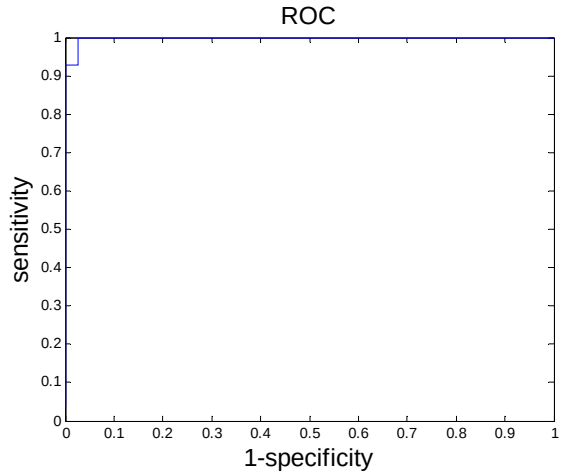

Fig. 3 ROC for synthetic data

Numerical results:

The output of function gibbs_sampling() and null_distribution() provide metrics that evaluates the significance of the transcription regulatory networks and prioritize the target genes for each TRN.

Table 8

seed	L by n matrix where L is the number of TFs and n is the number of iterations. To determine the most confident target gene set for a given transcription factor, the user need to rank all the corresponding seed genes that sampled in n iterations and the genes with higher sampling frequencies are more likely to be true target genes for this TF.
p	The p-value for each transcription factor. It characterizes the consistency between the expression profile and binding information within one TRN.