IMPALA User Manual

Table of Contents

I. Introduction

The IMPALA (Inferring Modularization of PAthway LAndscape) MATLAB package contains demo codes for two pathway identification methods: GIST and SOUL. GIST (Gibbs sampler Infers Signal Transduction) is a distribution learning method that builds pathways among given source and target proteins. By sampling pathway states according to pathway potential distributions, GIST will extract signal transduction pathways that are associated with phenotype and consent with knowledge. SOUL (Structural Organization Uncovers pathway Landscape) is a post-processing method that utilizes pathway samples from GIST to further investigate pathway modularization and landscape. IMPALA offers a novel perspective to identify aberrant signal transduction in cancer cells by emphasizing on locally coherent modules as emerged from pathway landscape.

II. Systems Requirements

The software was developed on MATLAB 7.14.0 and is compatible with both Windows and Linux environments. It may also support lower version of MATLAB as long as:

1. newer version of MATLAB build-in cluster function (clustergram.m) is available, which returns cluster labels;

2. moving average filtering (smooth.m) is available.

III. GIST (Gibbs sampler Infers Signal Transduction)

Input Data Structure

Input data: demo_data.mat*
Augment: see Table1

Table 1. GIST input data and description

Name	Description	Example	Is mandatory?
G0	Sparse binary network (undirected)	0 or 1	Yes
G1	Sparse weighted network (undirected)	1.1353	Yes
G_X	Log 2 gene expression	8.1311	No
G_corr	Node correlation	0.266	No
G_edgez	Edge z-score	1.1353	No
G_entrez	Entrez gene id	7157	NO
G_fld	Log2 fold change	0.2013	Yes**
G_loc	Subcellular location	Nucleus	Yes**
G_locn	Subcellular location id	4	Yes
G_nodesz	Node z-score	1.1384	Yes
G_p	Node p-value	0.0071	Yes**
G_probes	Probeset id	201746_at	No
G_symbol	Official gene symbol	TP53	Yes
t_dmfs	Survival days	760	No

*: preprocessed data from PPI subnetwork identification methods. The data include key information such as network structure, node scores and edge scores.

**: data not directly required for the algorithm, but only needed for complete summarization (annotation) of output results (e.g., fold change of genes).

Key Functions and Usage

Function: F0=bldFlowNet(G0,source,sink,L); %% Build flow network
Arguments: see Table 2

Table 2. Arguments of bldFlowNet()

Name	Description
G0	Sparse binary network (undirected)
source	Source/start proteins
sink	Target/end proteins
L	Length of pathway
F0	Flow network (unweighted)

Function: G=bldWeightMatrix(F0,G0,G1,delta); %% Build directed weighted matrix for given graph
Arguments: see Table 3

Table 3. Arguments of bldWeightMatrix()

Name	Description
F0	Flow network (unweighted)
G0	Sparse binary network (undirected)
G1	Sparse weighted network (undirected)
delta	Baseline score for pseudo-edges
G	Sparse weighted network (directed)

Function: [S V W valid_path valid_edge]=rndInitial(G,G0,F0,H,L); %% Random Initialization, pseudo-edges are introduced
Arguments: see Table 4

Table 4. Arguments of rndInitial()

Name	Description
G	Sparse weighted network (directed)
G0	Sparse binary network (undirected)
F0	Flow network (unweighted)
H	Node z-score
L	Pathway length
S	Initial pathway
V	Pathway node potential
W	Pathway edge potential
valid_path	If pathway contains pseudo-edges (binary indicator)
valid_edge	(L-1)×1 vector indicating if an edge is a pseud-edge

Function:[sampledPaths,pathFreq,pathScore]=gist(G,G0,G_locn,L,F0,H,V,W,S,valid_path,valid_edge,ite,T,rho1,rho2,VBITE); %% GIST algorithm
Arguments: see Table 5

Table 5. Arguments of gist()

Name	Description
G	Sparse weighted network (directed)
G0	Sparse binary network (undirected)
G_locn	1:extracellular space; 2: plasma membrane; 3: cytoplasm; 4: nucleus
L	Pathway length
F0	Flow network (unweighted)

H	Node z-score
V	Pathway node potential
W	Pathway edge potential
S	Initial pathway
valid_path	If pathway contains pseudo-edges (binary indicator)
valid_edge	(L-1)×1 vector indicating if an edge is a pseud-edge
ite	Number of sampling iterations
T	Temperature
rho1	Flow parameter
rho2	Subcellular balance parameter
VBITE	Verbosity parameter
sampledPaths	Sampled pathways
pathFreq	Frequency of sampled pathway
pathScore	Likelihood score of sampled pathway

Function: [Eg1 Eg2 Eg3 gist_slist gist_wlist gist_locn gist_flow]=est_edge(rankedSampledPaths1,rankedPathScore1,G,G_locn,PATHNUM)
%% estimate edge score and direction
Arguments: see Table 6

Table 6. Arguments of est_edge()

Name	**Description**
rankedSampledPaths1	Pathway ID ranked according to pathway potential score
rankedPathScore1	Pathway potential score
G	Sparse weighted network (directed)
G_locn	1:extracellular space; 2: plasma membrane; 3: cytoplasm; 4: nucleus
PATHNUM	Number of top pathways used to estimate edge attributes
Eg1	Directed edge matrix
Eg2	Edge direction matrix
Eg3	Bi-direction edge matrix (normalized score)
gist_slist	Pathway index list
gist_wlist	Pathway potential score list
gist_locn	Pathway location
gist_flow	Pathway flow information

Output data structure

Output to .mat: demo_results.mat
Save all output results to .mat file, which can be used as input for the SOUL method.

Output edge attributes to file: demo_network.xlsx (see Table 7 for the format)

Table 7. Format of demo_network.xlsx

Column	Description	Example
1	Protein 1*	HSP90AA1
2	Protein 2*	BIRC5
3	Edge direction probability	1
4	Normalized edge score	0.0097445

*: in the context of directed network, an edge always starts from protein1 to protein 2.

Output node attributes to file: demo_node_attr.xlsx (see Table 8 for the format)

Table 8. Format of demo_node_attr.xlsx

Column	Description	Example
1	Official gene symbol	BIRC5
2	Log2 fold change	0.407996299
3	p-value	0.002291066
4	Subcellular location	Cytoplasm
5	Node score	40.30067329

IV. SOUL (Structural Organization Uncovers pathway Landscape)

Input Data Structure

Input data: ER_signaling.mat, apoptosis.mat and cell_cycle.mat
Arguments: see Table 9

Table 9. SOUL input data and description

Name	Description
G_symbol	Official gene symbol
G_entrez	Entrez gene id
G_fld	Log2 fold change
G_p	Node p-value
G_loc	Subcellular location
G_locn	Subcellular location id
G_nscore	Node score
Eg2	Edge direction matrix
Eg3	Bi-direction edge matrix (normalized score)
rankedPathScore1	Pathway potential score
rankedPathSymb1	Pathway (gene symbol) ranked according to pathway potential score
rankedSampledPath1	Pathway index ranked according to pathway potential score

Key Procedures

Key procedures of SOUL are summarized in Table 10.

Table 10. Key procedures of SOUL script

Procedure	Description
1. Calculate structural profile	d(i,j) is the overlap between pathway i and j
2. Clustering	Hierarchical clustering
3. Re-organize pathway samples and potential distribution	Structural heatmap generated
4. Smooth potential distribution	Smooth using moving average filtering

5. Select clusters of interest	Four clusters

Output data structure

Output network to file: SOUL_network_bidir.xlsx (see Table 11 for the format)

Table 11. Format of SOUL_network_bidir_xlsx

Column	Description	Example
1	Protein 1*	IRS1
2	Protein 2*	BIRC5
3	Edge direction probability	1
4	Normalized edge score	0.04941

*: in the context of directed network, an edge always starts from protein1 to protein 2.

Output node attributes to file: SOUL_nodes_bidir.xlsx (see Table 12 for the format)

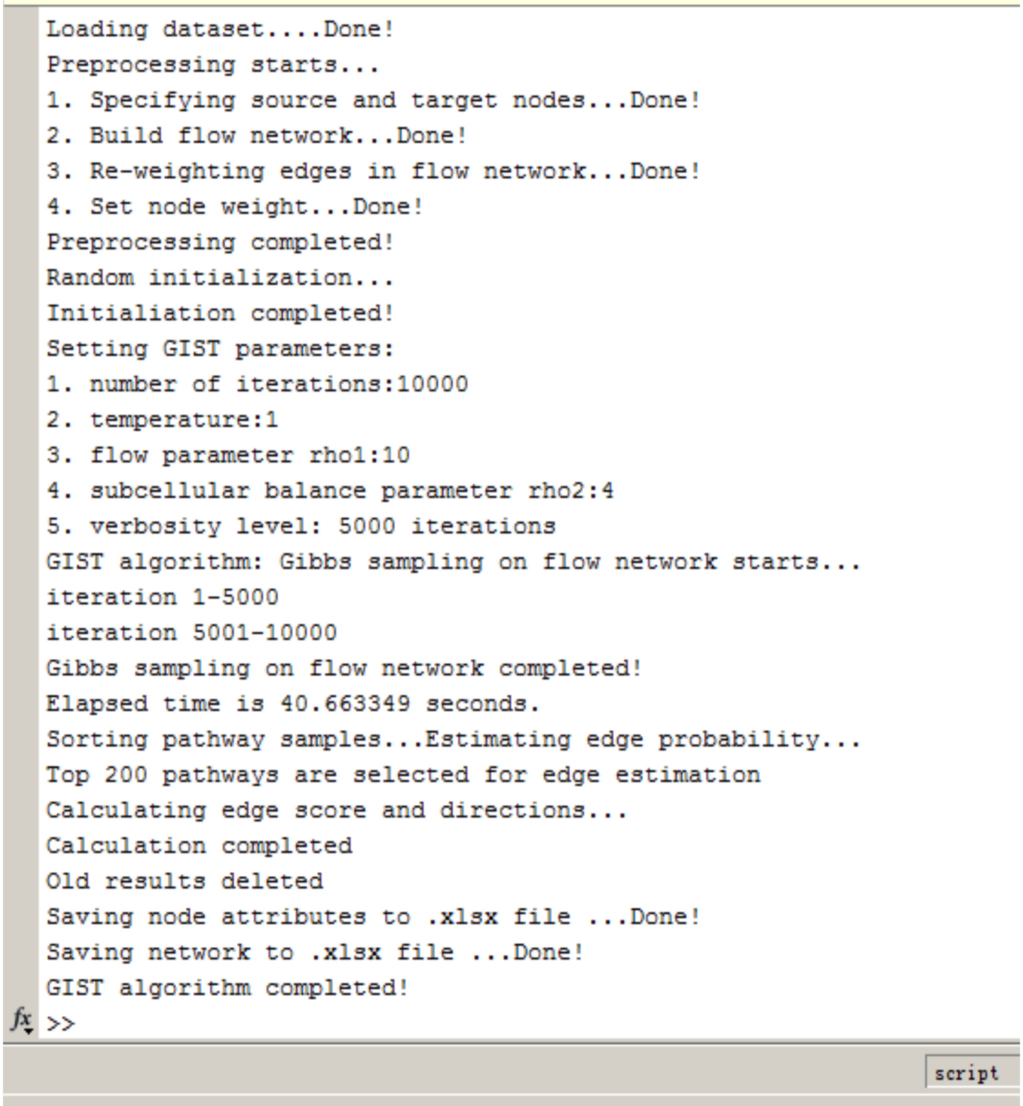
Table 12. Format of SOUL_nodes_bidir.xlsx

Column	Description	Example
1	Official gene symbol	BRCA1
2	Log2 fold change	0.126355513
3	p-value	0.020686546
4	Subcellular location	Nucleus
5	Node score	1864.070147

V. Examples

Running GIST

Run demo_GIST.m and the program will show the running status in MATLAB command window as shown in Fig. 1:



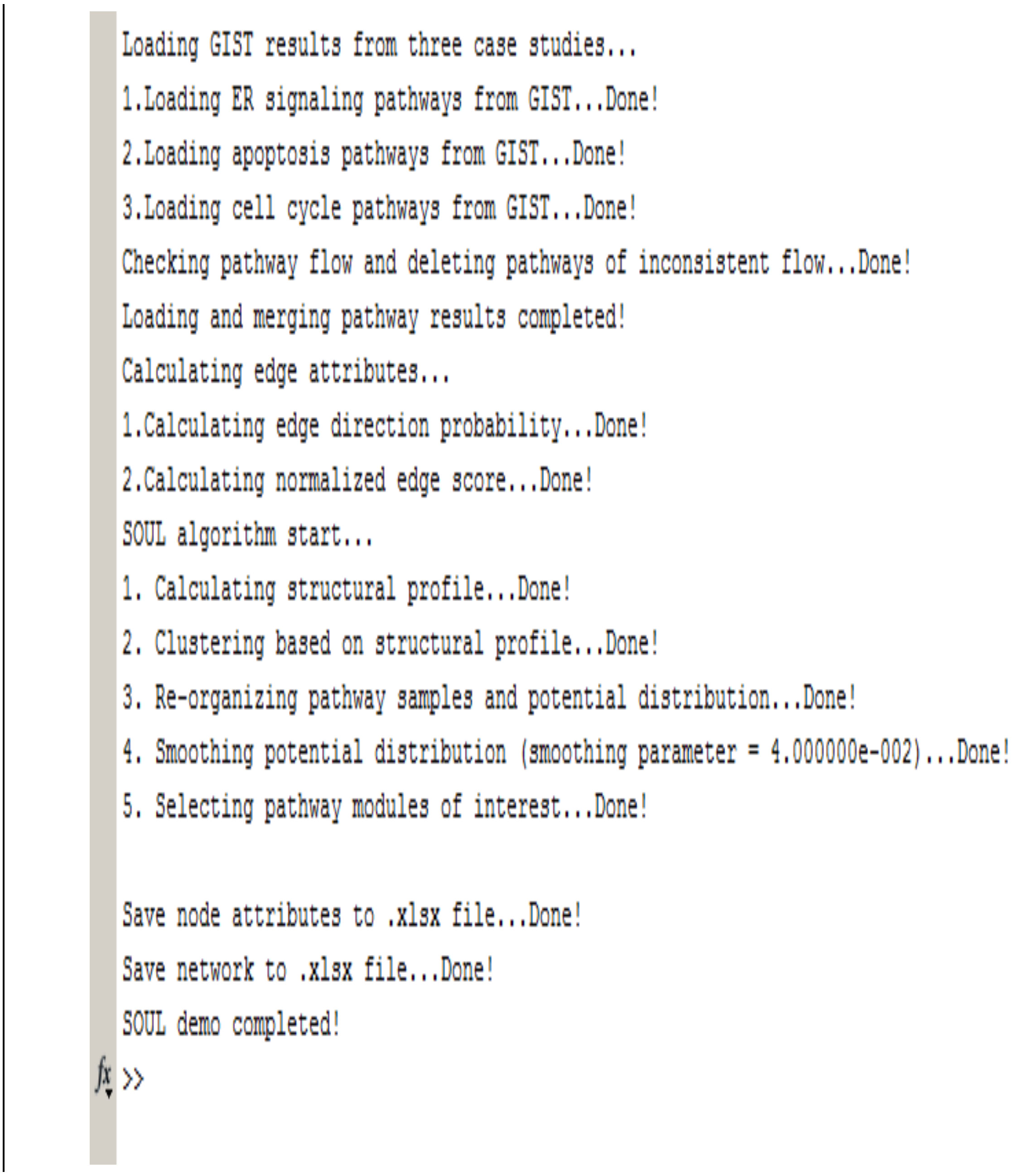
```
Loading dataset....Done!
Preprocessing starts...
1. Specifying source and target nodes...Done!
2. Build flow network...Done!
3. Re-weighting edges in flow network...Done!
4. Set node weight...Done!
Preprocessing completed!
Random initialization...
Initialiation completed!
Setting GIST parameters:
1. number of iterations:10000
2. temperature:1
3. flow parameter rho1:10
4. subcellular balance parameter rho2:4
5. verbosity level: 5000 iterations
GIST algorithm: Gibbs sampling on flow network starts...
iteration 1-5000
iteration 5001-10000
Gibbs sampling on flow network completed!
Elapsed time is 40.663349 seconds.
Sorting pathway samples...Estimating edge probability...
Top 200 pathways are selected for edge estimation
Calculating edge score and directions...
Calculation completed
Old results deleted
Saving node attributes to .xlsx file ...Done!
Saving network to .xlsx file ...Done!
GIST algorithm completed!
fx >>
```

Fig. 1. Running window of GIST

After GIST algorithm completes, three output files (demo_results.mat, demo_network.xlsx and demo_node_attr.xlsx) will be generated.

Running SOUL

Run demo_SOUL.m in MATLAB and the current status of the program will be displayed in command line window (Fig. 2):



```
Loading GIST results from three case studies...
1.Loading ER signaling pathways from GIST...Done!
2.Loading apoptosis pathways from GIST...Done!
3.Loading cell cycle pathways from GIST...Done!
Checking pathway flow and deleting pathways of inconsistent flow...Done!
Loading and merging pathway results completed!
Calculating edge attributes...
1.Calculating edge direction probability...Done!
2.Calculating normalized edge score...Done!
SOUL algorithm start...
1. Calculating structural profile...Done!
2. Clustering based on structural profile...Done!
3. Re-organizing pathway samples and potential distribution...Done!
4. Smoothing potential distribution (smoothing parameter = 4.000000e-002)...Done!
5. Selecting pathway modules of interest...Done!

Save node attributes to .xlsx file...Done!
Save network to .xlsx file...Done!
SOUL demo completed!
>>
```

Fig. 2. Running window of SOUL

After SOUL program completes, two files will be generated to save network information and node attributes. In addition, the algorithm will also generate figures of pathway landscape: structural heatmap and re-organized potential distribution (see Fig. 3 for an example).

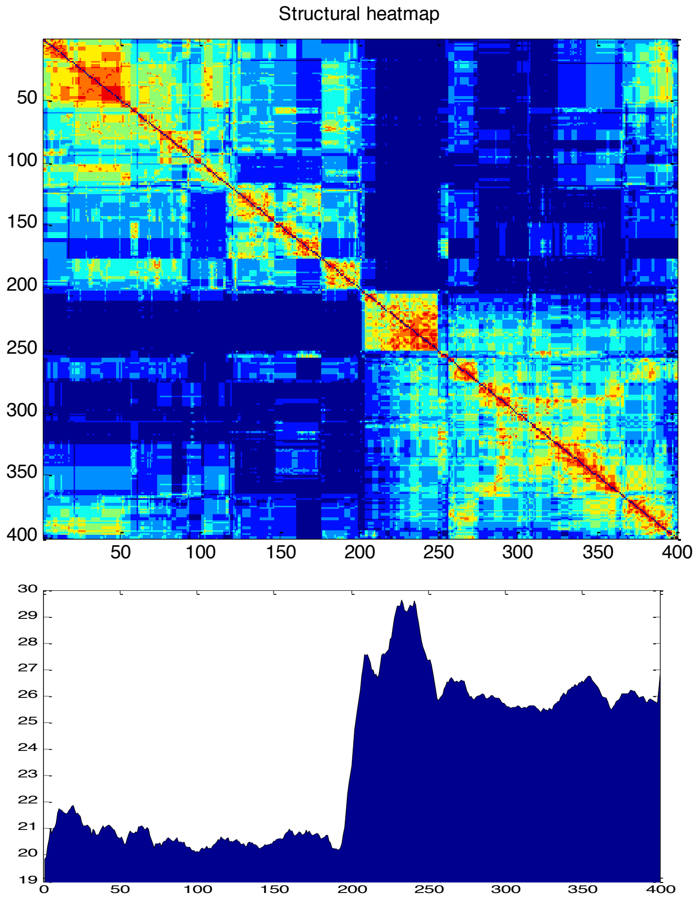


Fig. 3. An example of pathway landscape