



A dynamic programming approach to integrate gene expression data and network information for pathway model generation

Yuexu Jiang, Yanchun Liang, Duolin Wang, Dong Xu, and Trupti Joshi
Bioinformatics 2020

Students: Mahdi Rostami ,Mahdi Jamshid por

Pathways

Why active pathways?

- ✓ Disease
- ✓ Drug design
- ✓ Lung cancer
- ✓

Why dynamic programming?

- ✓ Step by step
- ✓ Trace back
- ✓ Economical system

Previous works

Paper	Explanation
Tuncbag et al. (2013)	Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. J. Comput. Biol., 20, 124–136.
Ideker,T. et al. (2001)	Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science, 292, 929–934.
Kim et al., (2011)	Identifying causal genes and dysregulated pathways in complex diseases. PLoS Comput. Biol., 7, e1001095.
Lan et al. (2011)	ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. Nucleic Acids Res., 39, W424–W429.
Min et al. (2018)	Edge-group sparse PCA for network-guided high dimensional data analysis. Bioinformatics, 34, 3479–3487.

Limitations

Previous works problems

- Unclear pathway
- Hard to trace back
- Time-consuming
- information loss
- Related pathways

Input data

- **KEGG**

Kyoto Encyclopedia of Genes and Genomes

- **PPI**

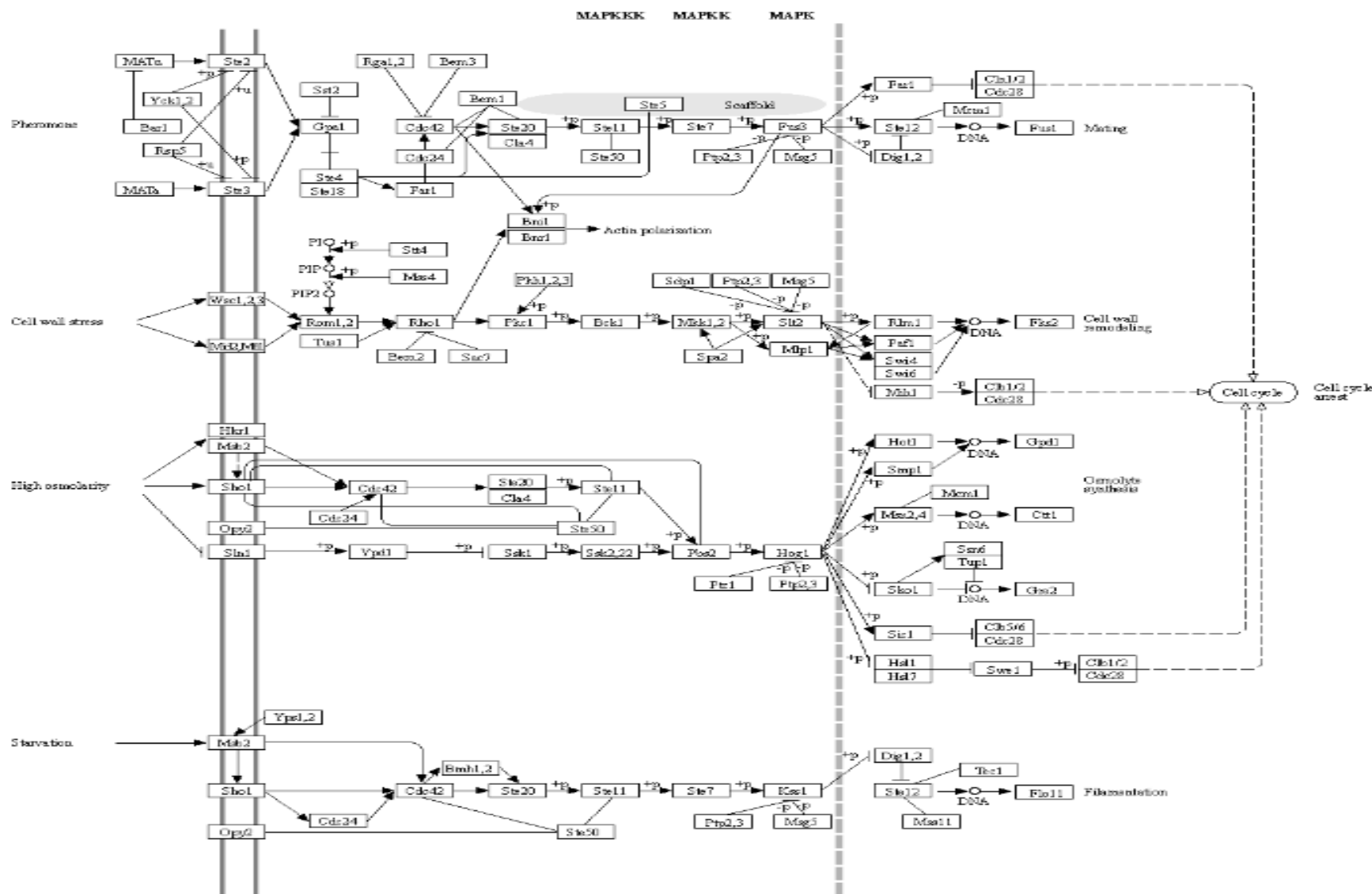
protein-protein interaction

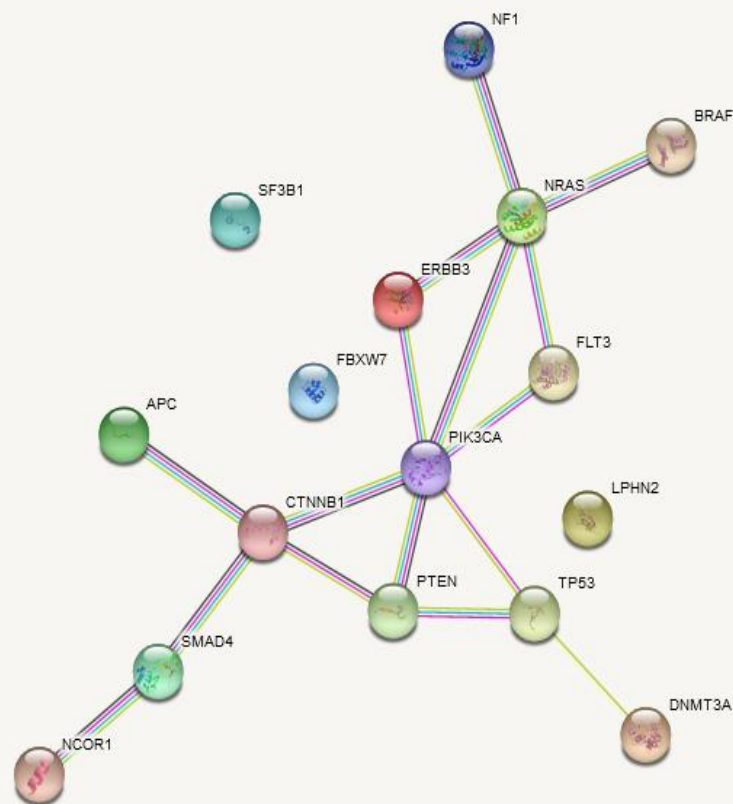
- **GEO**

Gene Expression Omnibus

- **Omics data**

MAPK SIGNALING PATHWAY - YEAST





Network

currently showing

Summary view: shows current interactions. Nodes can be moved; popups provide information on nodes & edges.

Cooccurrence

Gene families whose occurrence patterns across genomes show similarities.

Experiments

Co-purification, co-crystallization, Yeast2Hybrid, Genetic Interactions, etc ... as imported from primary sources.

Databases

Known metabolic pathways, protein complexes, signal transduction pathways, etc ... from curated databases.

Coexpression

Proteins whose genes are observed to be correlated in expression, across a large number of experiments.

Neighborhood

Groups of genes that are frequently observed in each other's genomic neighborhood.

NF1

Information

Neurofibromin; Stimulates the GTPase activity of Ras. NF1 shows greater affinity for Ras GAP, but lower specific activity. May be a regulator of Ras activity; Armadillo-like helical domain containing Identifier: ENSP00000351015, NF1 Organism: Homo sapiens

e!

UniProt

RefSeq

NCBI

KEGG

Ensembl

Actions

re-center network on this node

remove this node from input nodes

show protein sequence

homologs among STRING organisms

Pathways, Functions, Resources (GeneCards)

highlight enriched terms in the analysis table

1 of 12

AlphaFold model (P21359)

identity: 100%

NCOR1

DNMT3A

Network

currently showing

Summary view: shows current interactions. Nodes can be moved; popups provide information on nodes & edges.

Cooccurrence

Gene families whose occurrence patterns across genomes show similarities.

Experiments

Co-purification, co-crystallization, Yeast2Hybrid, Genetic Interactions, etc ... as imported from primary sources.

Databases

Known metabolic pathways, protein complexes, signal transduction pathways, etc ... from curated databases.

Coexpression

Proteins whose genes are observed to be correlated in expression, across a large number of experiments.

Neighborhood

Groups of genes that are frequently observed in each other's genomic neighborhood.

GEO

GEO	Title
GSE19804	Genome-wide screening of transcriptional modulation in non-smoking female lung cancer in Taiwan
GSE31176	Expression data from yeast (wild type, rlm1 and swi3 mutants) exposed to Congo Red
GSE71433	Expression data from yeast (wild type and gcn5 mutants) exposed to Congo Red (CR)
GSE13097	mRNA amount analysis of wild type strain subjected to osmotic stress

NCBI > GEO > [Accession Display](#) [?](#) Not logged in | [Login](#) [?](#)

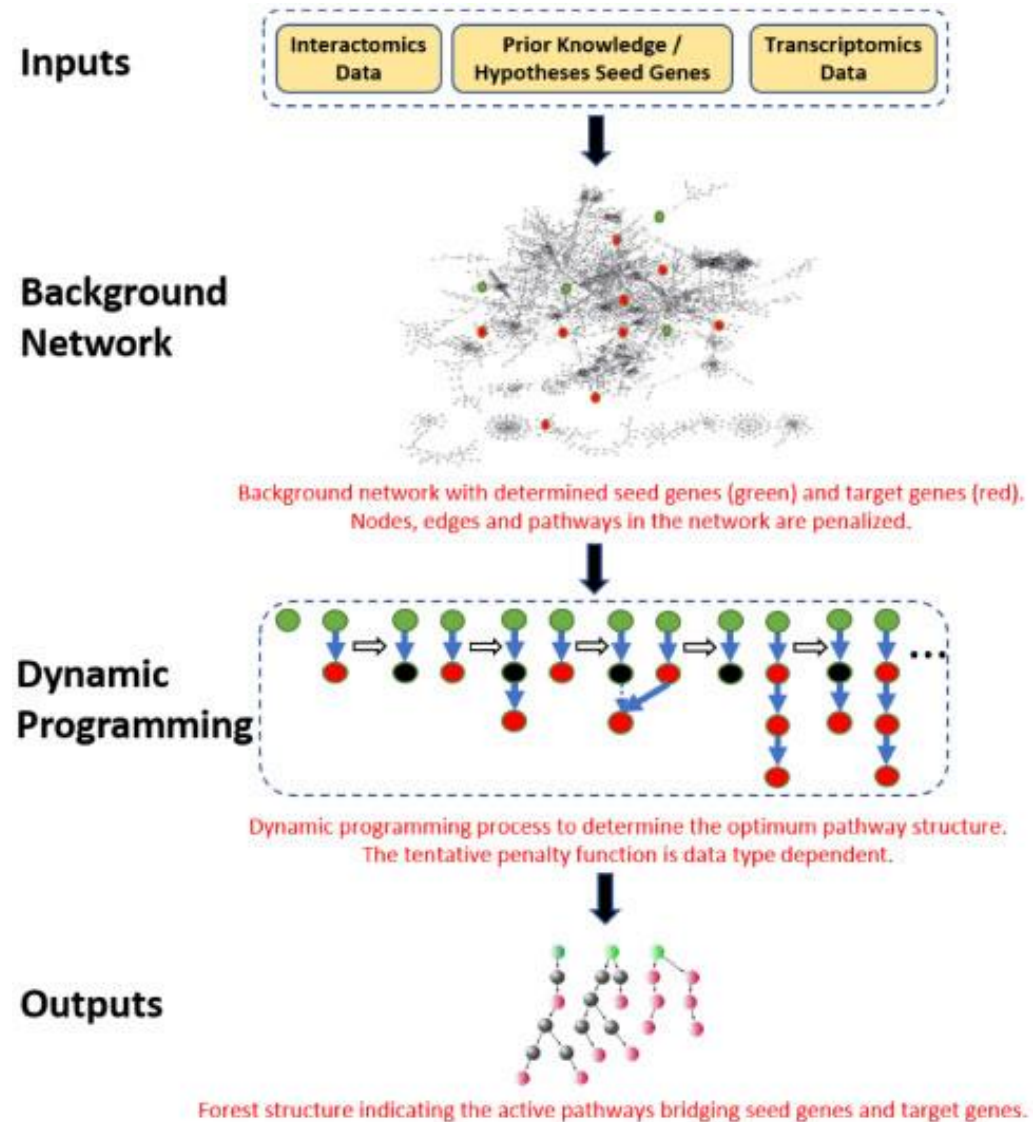
Scope: Format: Amount: GEO accession:

Series GSE19804 [Query DataSets for GSE19804](#)

Status	Public on Jan 30, 2011
Title	Genome-wide screening of transcriptional modulation in non-smoking female lung cancer in Taiwan
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	Although smoking is the major risk factor for lung cancer, only 7% of female lung cancer patients in Taiwan have a history of cigarette smoking, extremely lower than those in Caucasian females. This report is a comprehensive analysis of the molecular signature of non-smoking female lung cancer in Taiwan.
Overall design	RNA was extracted from paired tumor and normal tissues for gene expression analysis.
Contributor(s)	Lu T, Lai L, Chuang EY
Citation(s)	Lu TP, Tsai MH, Lee JM, Hsu CP et al. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. <i>Cancer Epidemiol Biomarkers Prev</i> 2010 Oct;19(10):2590-7. PMID: 20802022 Lu TP, Hsiao CK, Lai LC, Tsai MH et al. Identification of regulatory SNPs associated with genetic modifications in lung adenocarcinoma. <i>BMC Res Notes</i> 2015 Mar 24;8:92. PMID: 25889623
Submission date	Jan 08, 2010
Last update date	Jan 15, 2020
Contact name	Tzu-Pin Lu
E-mail(s)	tplu@ntu.edu.tw
Organization name	National Taiwan University, Taiwan
Department	Department of Public Health, Institute of Epidemiology and Preventive Medicine
Street address	No. 1, Sec. 4, Roosevelt Road
City	Taipei
ZIP/Postal code	10617
Country	Taiwan
Platforms (1)	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
Samples (120)	GSM494556 Lung Cancer 2T GSM494557 Lung Cancer 3T GSM494558 Lung Cancer 6T
This SubSeries is part of SuperSeries: GSE33356 Genome-wide screening of genomic alterations and transcriptional modulation in non-smoking female lung cancer in Taiwan	
Relations	
BioProject	PRJNA153899
<input type="button" value="Analyze with GEO2R"/>	
Download family	Format
SOFT formatted family file(s)	SOFT ?
MINiML formatted family file(s)	MINiML ?

Method

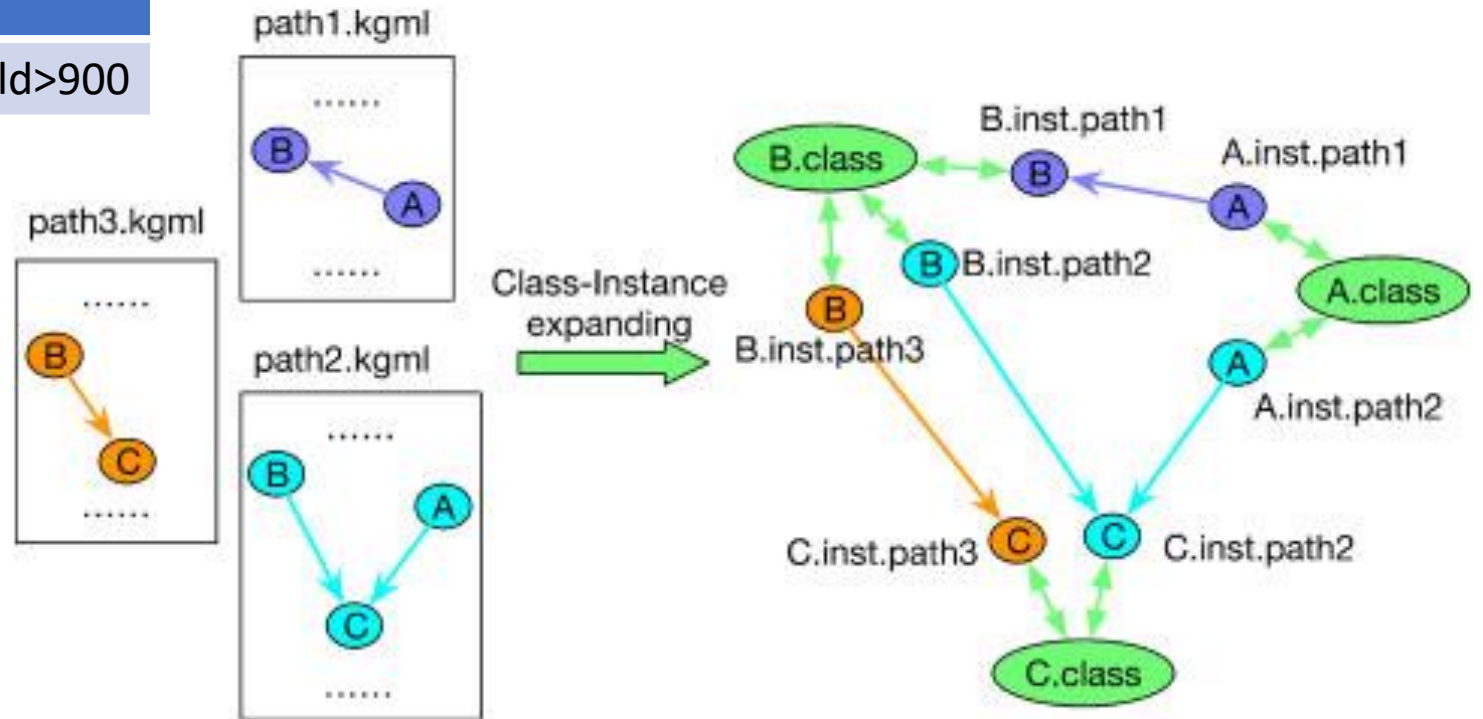
1. Constructing a merged pathway network
2. Define seed genes and target genes
3. Assign penalties based on omics data
4. Calculating minimum involvement score using a shortest path algorithm
5. Detecting final active pathways by truncating and backtracking



Method

Constructing a merged pathway network

1	KEGG		
2	PPI	STRING v10	Threshold>900



Method

Define seed genes and target genes

Seed genes	Perspective
Root nodes	Algorithm
Mutated	Biological
Receptor genes	Biological
Effective	Natural potential

Target genes	
Case-control	q-value threshold of 0.01
time-series	maSigPro

Method

Assign penalties based on omics data

Static penalty:

	Case control	Time series	Equation
Node penalty	+		1
Edge penalty		+	2
Pathway penalty	+	+	3

1. t-test -> P-value

$$\text{penalty}(v) = \frac{1}{-\log_2(P_{value}(v))} \quad (1)$$

2. v1, v2 = vectors

$$\text{penalty}(e) = 1 - |\text{correlation}(v1, v2)| \quad (2)$$

3. Fisher test -> P-value

$$\text{penalty}(p) = \frac{1}{-\log_2(P_{value}(p))} \quad (3)$$

4. Unknown pathway = 1

Method

Calculating minimum involvement score using the shortest path algorithm (Dijkstra)

Dynamic penalty:

$$\text{tentative pentalty}(s, v) = \begin{cases} \text{penalty}(s) + \text{Sigmoid}\left(fc(s) \cdot fc(v)\right) \cdot \left(\text{penalty}(v) + \text{penalty}(p_v)\right), & \text{negative relation} \\ \text{penalty}(s) + \text{Sigmoid}\left(-fc(s) \cdot fc(v)\right) \cdot \left(\text{penalty}(v) + \text{penalty}(p_v)\right), & \text{positive relation} \end{cases}$$

Case control	
fc	fold
Pv	pathway
relation	KEGG

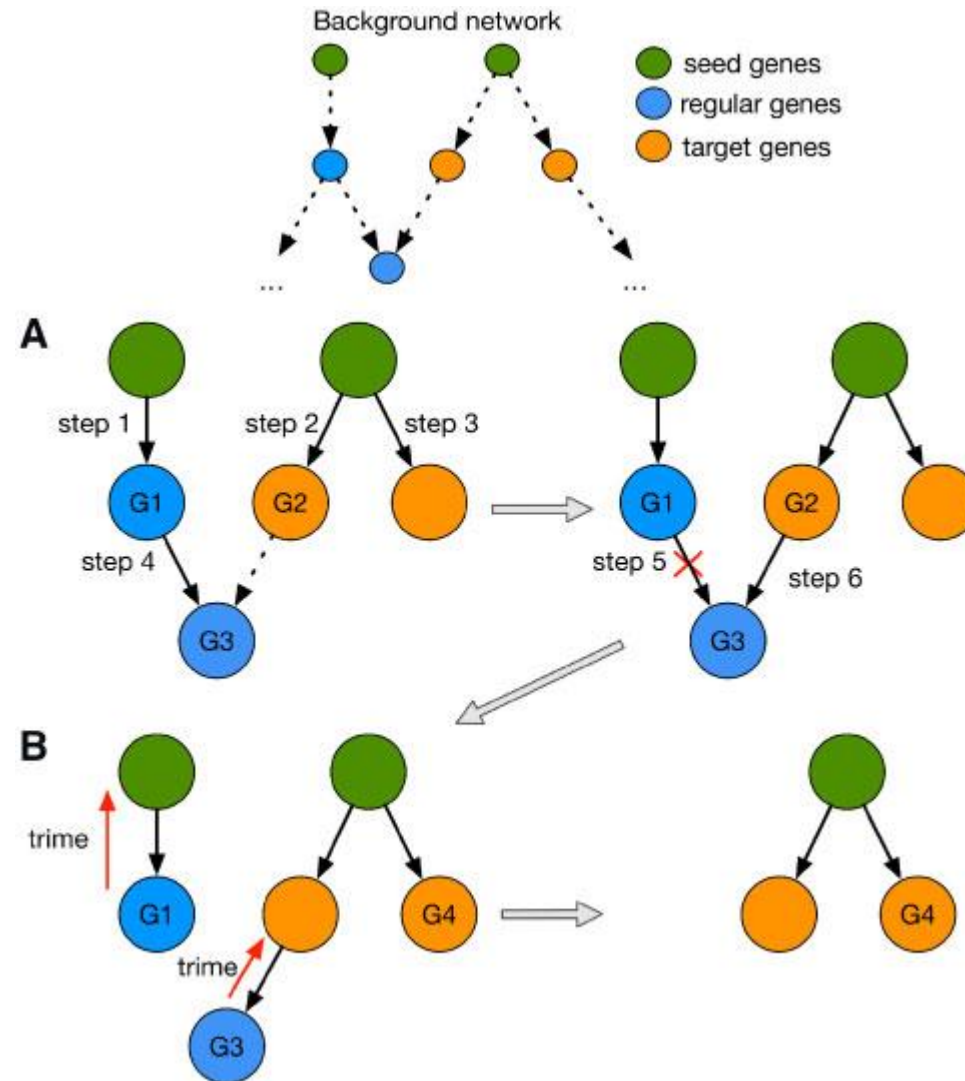
Time series	
e	Interaction (KEGG ,PPI)
Pv	pathway

$$\text{tentative penalty}(s, v) = \text{penalty}(s) + \text{penalty}(e) \cdot \text{penalty}(p_v)$$

Method

Detecting final active pathways by truncating and backtracking

A	Penalty update
B	Truncating rules



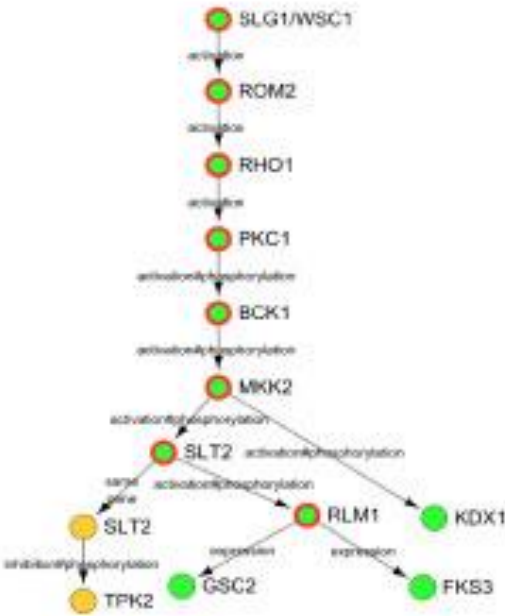
Evaluation

- Application on yeast cell wall damage stress dataset

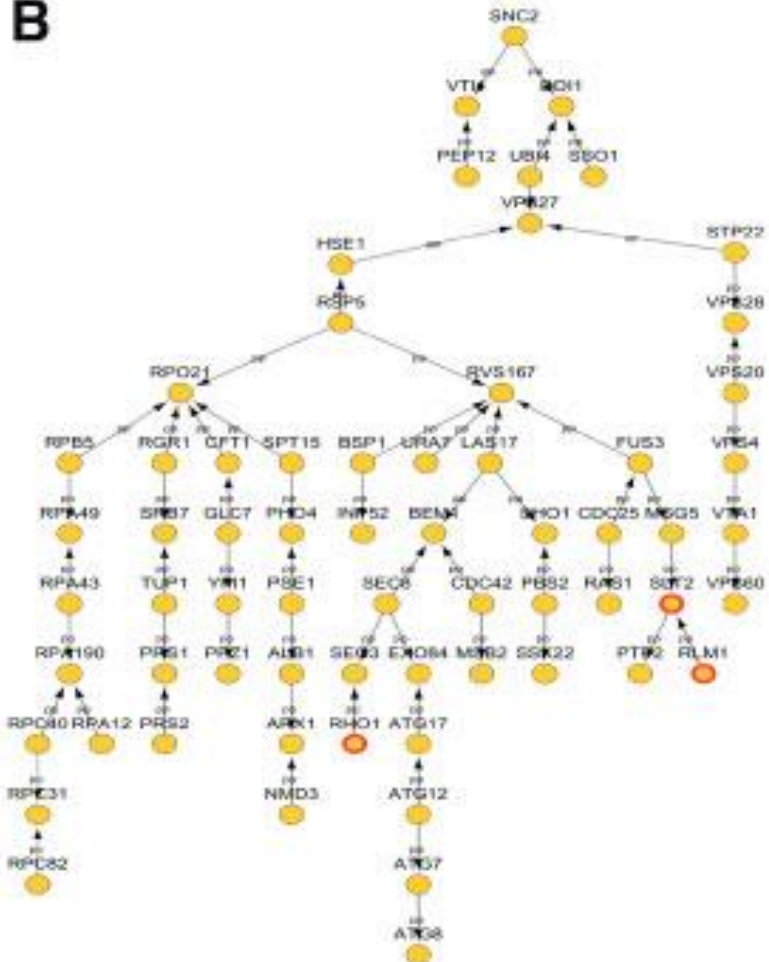
Method	Shape
Steiner tree	B
IMPres	A

GEO	GSE31176 GSE71433
-----	----------------------

A

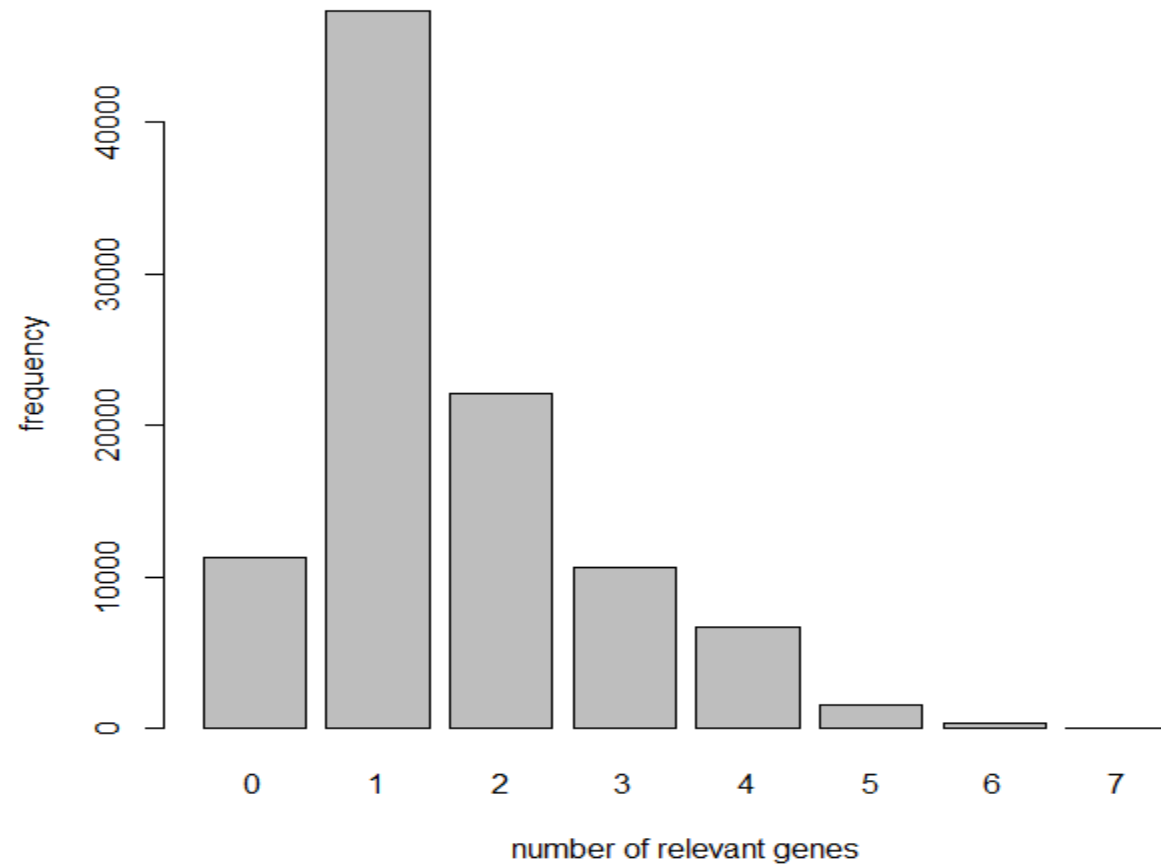


B



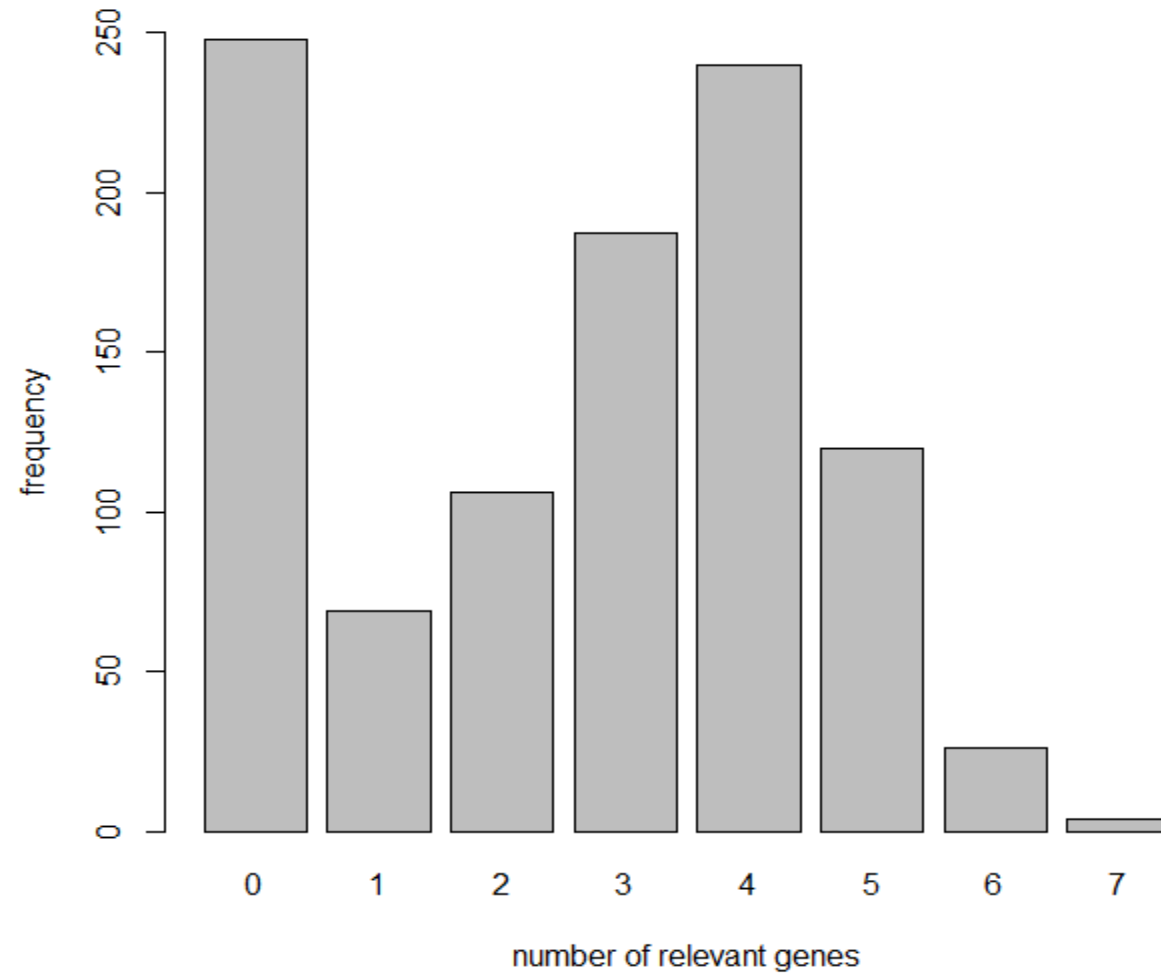
Evaluation

Is it by chance?

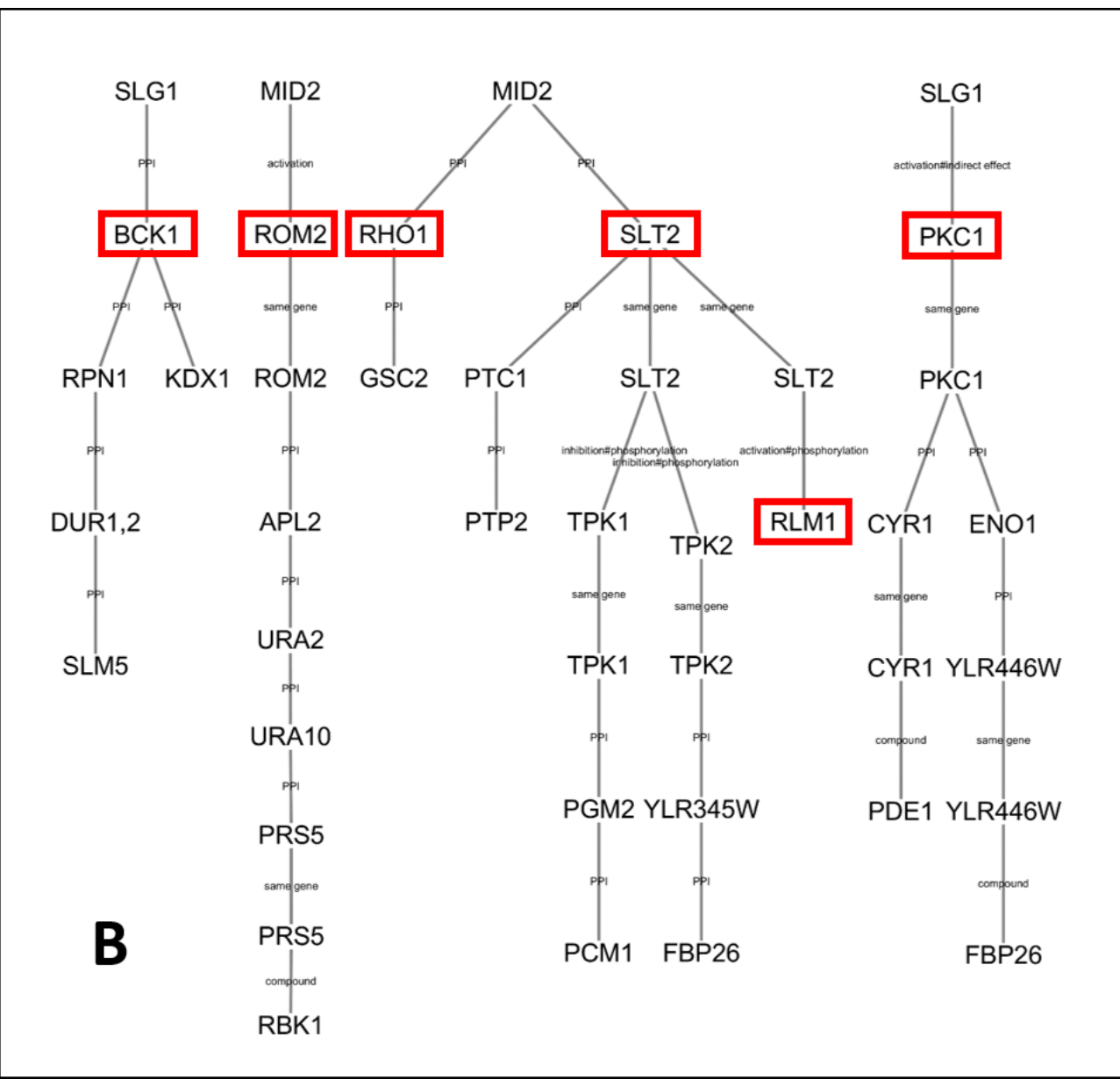
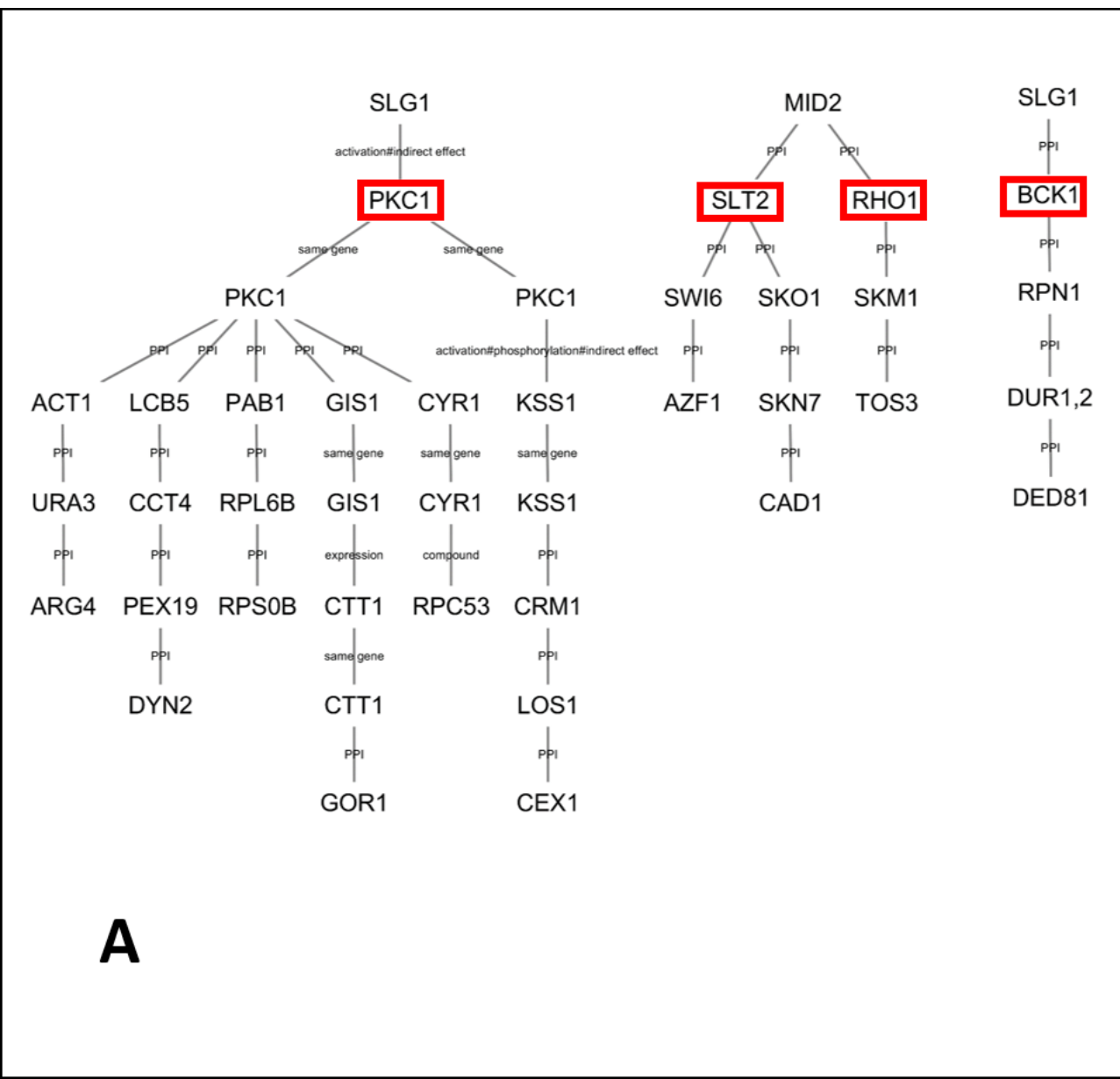


Evaluation

Does expression data have an impact?



Evaluation




Evaluation

- Application on yeast perturbation dataset

Method	Small network	Large network	Network PPI
Ideker et al.(2002)	43 nodes	340 genes	362
IMPres	10 786 nodes		54 249

Evaluation

Log2(fold change)
-5.42  7.35

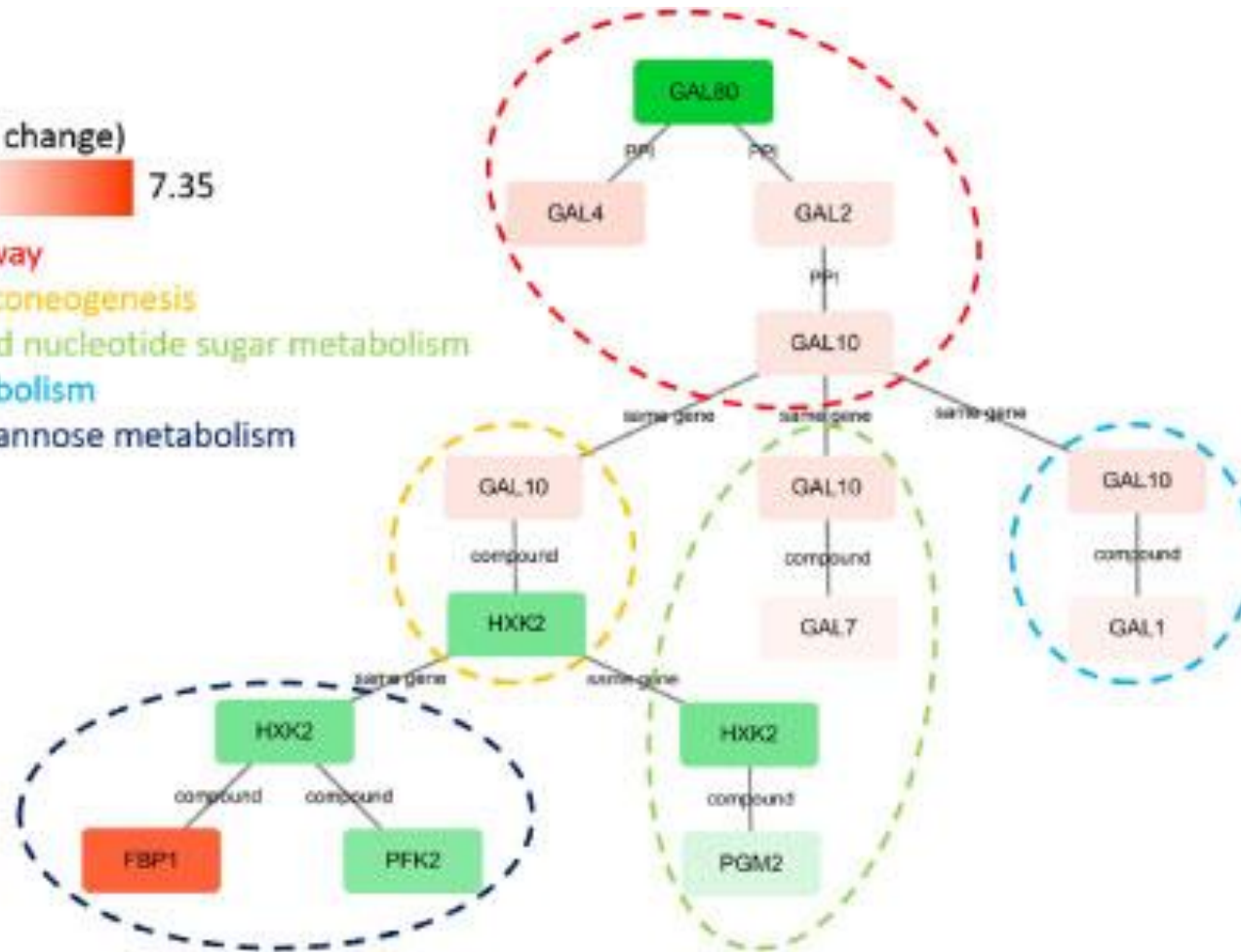
Unknown pathway

Glycolysis / Gluconeogenesis

Amino sugar and nucleotide sugar metabolism

Galactose metabolism

Fructose and mannose metabolism



Evaluation

- Application on yeast high osmolality stress dataset

GEO

GSE13097

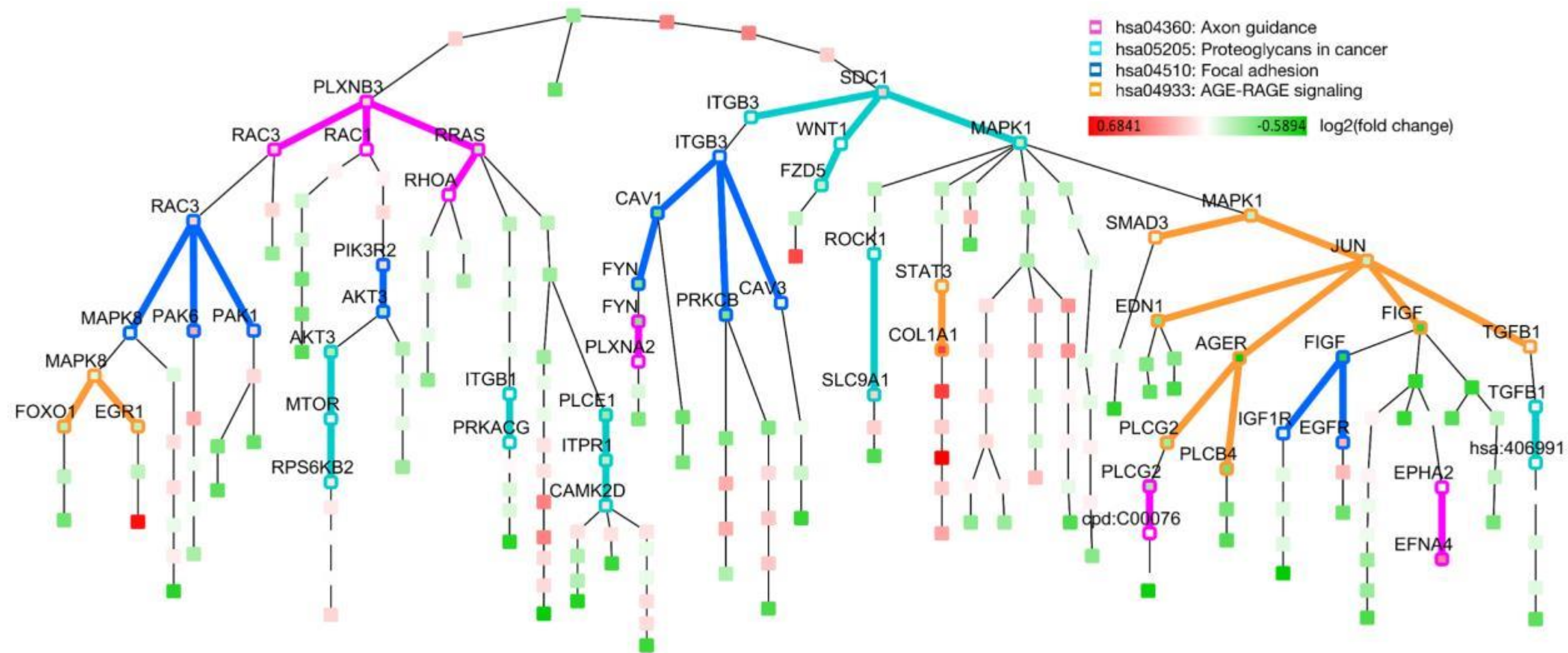
Algorithm	KEGG_ top10	KEGG_ top30	KEGG_ top40	KEGG_ top50	K+PPI_ top10	K+PPI_ top20	K+PPI_ top30	SDREM	PNM	ResponseNet
Total predictions	5	9	12	23	28	44	66	58	445	61
Predicts Hog1	Y	Y	Y	Y	Y	Y	Y	Y	Y	N
Predicted internal	4	6	9	20	25	41	62	30	374	4
Gold standard internal	21	21	21	21	29	29	29	30	30	30
Internal overlap	4	5	5	5	3	7	7	6	9	1
Internal significance	4.5E-8	3.4E-9	6.9E-8	7.7E-6	5.1E-4	4.2E-9	8.5E-8	1.11E-8	1.61E-4	0.0227
Internal precision	100%	83%	55.6%	25%	12%	17%	11%	20%	2%	25%
Predicted TFs	1	3	3	3	3	3	4	28	71	57
Gold standard TFs	6	6	6	6	6	6	6	7	7	7
TF overlap	1	3	3	3	3	3	3	4	2	2
TF significance	0.24	8.7E-3	8.7E-3	8.7E-3	1.4E-4	1.4E-4	5.6E-4	7.7E-3	0.77	0.632
TF precision	100%	100%	100%	100%	100%	100%	75%	14%	2.8%	3%

Result

- Human lung cancer

Pathways	P-value	Importance
AGE-RAGE	3.99E-9	Rap1
proteoglycans	5.43E-7	5 downregulated
focal adhesion	2.77E-5	PAK1, PAK6, EGFR, AKT3

Result



Web server

[HOME](#) [Info](#) [New Job](#) [Tutorial](#) [Tools](#)

IMPRes

Integrative MultiOmics Pathway Resolution

Introduction

Learn what is IMPRes and what you can do with IMPRes.

[Learn more +](#)

New Job

Submit a new job of your own dataset or try our web server by using our sample dataset, you can use our [tools](#) to convert your data to our acceptable format.


[Start a new job +](#)

Tutorial

New to IMPRes? Our detailed tutorial will guide you through the process of using IMPRes.

[Learn more +](#)

© Digital Biological Laboratory 2017



IMPRE Usage Tutorial

How to submit a job

To submit a job, first thing is to specify the organism you are working on.

Organism

human

Then you should upload the seed genes file. The seed genes are the root nodes where the pathways detection process begins. The seed genes file format should be like below...

```

file Edit Format View Help
chr1:844,012..VL:811216
chr1:844,012..VL:811216
chr1:844,012..VL:811216
chr1:844,012..VL:811216

```

It contains the ID from KEGG database and String PPI database. If you do not integrate PPI into reference network, the String ID is not required. You can use our [gene file converter tool](#) to convert your gene list to our acceptable format.

Next, you need to provide some information about the target genes. They are the genes that become active during the process and show changes in their expression patterns. You can either upload your own target gene list as the format as seed genes, or you can let us determine target genes automatically based on your target number.

Then you need to upload your expression data. You have to specify the data type of your expression data. If it is a case control type, the format should be a tab separated list file like below...

chr1:844,012..VL:811216	0.0	0.000001	0	0.200000	0.010000	0	0.7	0.700000	0	0.797001	0	0.160002
chr1:844,012..VL:811216	0.422007	0.200714	1	0.690007	0.704540	0	0.275429	0.5	0.260044	0	0.198727	
chr1:844,012..VL:811216	0.06	0	1.044	0.275714	0.552796	0.620009	0.38	0.8	0	0.842791	1	0
chr1:844,012..VL:811216	0.440002	0.391409	0	0.510004	0.420140	0	0.880001	0	0.390027	0	0.550701	0.220000
chr1:844,012..VL:811216	0.870708	0.260028	0.243002	0.290514	0	0.000000	0.710714	0	0.081003	0	0.550001	
chr1:844,012..VL:811216	0.170414	0.710001	0.201178	0.527048	0.201707	0	0.103444	0	0.733004	0	0.630046	
chr1:844,012..VL:811216	0.434703	0.347004	0.260007	0.720000	0.837907	0	0.116010	0	0.440207	0	0.100000	
chr1:844,012..VL:811216	0	0.00020	0.00000	0	0.000200	0.010001	0	0.8070	0.75	0.391007	0.600107	1
chr1:844,012..VL:811216	0.680270	0	0.000000	0.700704	0.010000	0	0.300107	0.020000	0.00010	0	0.700200	
chr1:844,012..VL:811216	0.780000	0.000010	0	0.120707	0.470000	0.000010	0	0.100000	0.820007	0	0.000010	
chr1:844,012..VL:811216	0	0.000000	0	0.000000	0	0.000000	0.000000	0.000000	0.000000	0	0.000000	

Control

HOME

Info

New Job

Tutorial

Tools

Job Submission

Organism

human

Seed genes

Upload Seed gene file:

Choose File

No file chosen

Target genes

☒ Upload my own Target genes

Choose File

No file chosen

☐ Derive Target genes automatically

#Target genes:

Expression data

Upload expression data file:

Choose File

No file chosen

Data type

☒ Case-control

#Case samples:

#Control samples:

☐ Time-series

#Time points:

#Replicates:

☐ Integrate PPS into reference network

Submit

The screenshot shows the 'Architecture Tree' interface. On the left, a sidebar lists components categorized by type:

- Activation:** 'tanh', 'sigmoid', 'relu', 'maxout'.
- Normalization:** 'batchnorm', 'layernorm', 'groupnorm'.
- Pooling:** 'maxpool', 'avgpool', 'lstm'.
- Other:** 'dropout', 'attention', 'resnet'.

The main area displays a hierarchical tree of architectures. The root node is 'ResNet-50'. It branches into 'ResNet-50 (v1)' and 'ResNet-50 (v2)'. 'ResNet-50 (v1)' further branches into 'ResNet-50 (v1a)' and 'ResNet-50 (v1b)'. 'ResNet-50 (v2)' branches into 'ResNet-50 (v2a)' and 'ResNet-50 (v2b)'. The tree continues to branch out, showing various architectural variations and their associated performance metrics.

<https://impres.missouri.edu/impres>