

CGI: a new approach for prioritizing genes by Combining Gene expression and protein-protein Interaction data

Xiaotu Ma, Hyunju Lee, Li Wang, and Fengzhu Sun

Running environment: the programs provided in this web site can be run in MS windows.

Part A For yeast data

A.1 Configuration file of CGI.exe

The program CGI.exe for the yeast data is executable. It first read a configuration file CGI_onfig.txt to determine which task it should perform. A sample configuration file is given as follows,

```
DataPath=C:/CGI/data/  
OutPutPath=C:/CGI/out/  
GeneExpressionFile=Cycle.txt  
KernelFile=MIPS_Kernel_L070.txt  
GeneSetFile=GO_data.txt  
GeneNames=cycle_A_I_G.txt  
InformativeGONodes=informativeGO.txt  
OutputFileName=cellcycleoutput.txt
```

A configuration file is composed of statements, in the format of parameter=value, where the parameter is a named variable in the program and the value is the value that user wants to pass to the corresponding variable. All available parameters are given in the included example for yeast Cell cycle data, user can simply copy and modify the example file to get their own configuration file.

A.2 Running steps

After put the configuration file config.txt in the same folder as CGI.exe, the user can run the program CGI.exe by double click it or type its name in the command line and hit return. The program CGI.exe will output an R script file (cellcycleoutput.txt.R in the above example) in the designated OutPutPath (C:/CGI/out/ in the above example) which is directly run-able in R environment. The user can use the File->Source code menu to run it.

A.3 Input file

Five input file are required as input file. In the above example,

(1) Gene expression data file = Cycle.txt

This is the gene expression data file. It is a data matrix. The first row is the condition names. All other rows represent one gene each and the gene name is indicated in the first column. Expression values are separated by space (“ ”). Missing value is indicated by “NA”. The following is an example:

	0	10	20	30	40	50	60	70	80						
1	cln3.1	cln3.2	clb2.2	clb2.1	alpha0	alpha7	alpha14	alpha21	alpha28	alpha35	alpha42				
2	YAL001C	0.15	NA	-0.22	0.07	-0.15	-0.15	-0.21	0.17	-0.42	-0.44	-0.15	0.24	-0.10	NA
3	YAL002W	-0.07	-0.76	-0.12	-0.25	-0.11	0.10	0.01	0.06	0.04	-0.26	0.04	0.19	-0.22	-C
4	YAL003W	-1.22	-0.27	-0.10	0.23	-0.14	-0.71	0.10	-0.32	-0.40	-0.58	0.11	0.21	0.09	C
5	YAL004W	-0.09	1.20	0.16	-0.14	-0.02	-0.48	-0.11	0.12	-0.03	0.19	0.13	0.76	0.07	C
6	YAL005C	-0.60	1.01	0.24	0.65	-0.05	-0.53	-0.47	-0.06	0.11	-0.07	0.25	0.46	0.12	0.4
7	YAL007C	0.65	1.39	-0.29	-0.54	-0.60	-0.45	-0.13	0.35	-0.01	0.49	0.18	0.43	-0.23	-C
8	YAL008W	-0.36	-0.22	-0.20	0.10	-0.28	-0.22	-0.06	0.22	0.25	0.13	0.34	0.44	-0.32	0.
9	YAL009W	0.25	-0.79	-0.22	-0.54	-0.03	-0.27	0.17	-0.12	-0.27	0.06	0.23	0.11	0.03	-C
10	YAL010C	-0.30	-0.60	-0.18	0.01	-0.05	0.13	0.13	-0.21	-0.45	-0.21	0.06	0.32	0.00	0.
11	YAL011W	-0.15	-0.71	-0.15	-0.25	-0.31	-0.43	-0.30	-0.23	-0.13	-0.07	0.08	0.12	-0.0	C
12	YAL012W	-1.22	0.66	-0.64	-0.17	0.02	-0.33	-0.49	-0.30	-0.15	-0.24	0.40	0.53	0.25	C
13	YAL013W	-0.34	-1.06	-0.45	-0.29	-0.36	-0.19	0.00	-0.32	-0.27	-0.12	0.04	0.17	0.06	

(2) Protein interaction data file = MIPS_Kernel_L070.txt (the diffusion kernel in section 2.3 of the paper)

This is the protein interaction data file. It is a data matrix and its format is the same as the gene expression data file.

(3) Gene set file = GO_data.txt

This is a two column data matrix. The second column is the gene names and the first column is the corresponding functional annotations (see the following example).

```
GO:0005743 YMR056C
GO:0005471 YMR056C
GO:0006839 YMR056C
GO:0009060 YMR056C
GO:0005743 YBR085W
GO:0005471 YBR085W
GO:0009061 YBR085W
GO:0008372 YJR155W
GO:0018456 YJR155W
```

(4) Gene name file = cycle_A_I_G.txt

This is the file with all the gene names you are going to prioritize. It in general is the same gene list as that in the gene expression data file. Sometimes you may want a smaller list (e.g., all genes included in the expression data and protein interaction data and having functional annotation)

(5) Informative GO Nodes file = informativeGO.txt

This is the informative GO nodes. Refer to section 2.1 in the paper.

A.4 Output file

Two files will be generated after the R script file is submitted to the R environment. In the above example, they are cellcycleoutput.txt.R_rlt.txt and cellcycleoutput.txt.R_rllist.txt. The following is an example of the output file cellcycleoutput.txt.R_rlt.txt

```
MCC 0.740625976630315
CGI 0.838586224404736
```

The gene prioritizing method by using expression data only have performance index 0.74, and our CGI method have performance index 0.839.

The following is an example of the output file cellcycleoutput.txt.R_rftlist.txt

GOCLASS	Gene	p-Value by exp	p-Value by exp + ppi	total members
GO:0006414	YAL003W	1.03772410496861e-05	1.19124036019880e-05	12
GO:0006457	YAL005C	3.38108781861024e-07	1.07086019607294e-07	35
GO:0030437	YAL009W	0.501463961061275	0.154888571582294	30
GO:0007015	YAL016W	5.31147694942113e-05	1.57212970031129e-06	41
GO:0006468	YAL017W	0.141072879139435	0.0322477296279593	66
GO:0030490	YAL026C	0.633328050712797	0.973051169671187	23
GO:0000082	YAL040C	0.343506936452639	0.0305257652519236	23
GO:0007124	YAL041W	0.74573033901567	0.252520741753467	31
GO:0007264	YAL041W	0.490087248917976	0.0364636028671746	17
GO:0006888	YAL042W	0.00116816432105682	0.000204119801954783	43

The gene names in the second column are referred as target gene in the paper (section 2.4).

The first column is its corresponding functional annotation from Gene Ontology.

The third column is the one-sided Wilcoxon rank sum test P value of the prioritizing method by using expression data only.

The fourth column is the one-sided Wilcoxon rank sum test P value of the prioritizing method by our CGI method.

The fifth column is the number of genes in the corresponding functional annotation (indicated in column 1).

Part B For real phenotype data (human Alzheimer's disease data)

B.1 Configuration file of realdata_CGI.exe

The program realdata_CGI.exe for real phenotype data (here the human Alzheimer's disease data) is executable. Similar to the CGI.exe in part A, It first read a configuration file real_config.txt to determine which task it should perform. A sample configuration file is given as follows,

```
DataPath=C:/CGI/realdata/  
OutPutPath=C:/CGI/out/  
GeneExpressionFile=Alzheimerdata.txt  
KernelFile= HumanPPI_DF070.txt  
GeneNames= allgenes.txt  
OutputFileName=MMSE_tau070.txt  
PhenoNameFile=pheno_names.txt
```

B.2 Running steps

After put the configuration file real_config.txt in the same folder as realdata_CGI.exe, the user can run the program realdata_CGI.exe by double click it or type its name in the command line and hit return. The program realdata_CGI.exe will output the result (MMSE_tau070.txt in the above example) in the designated OutPutPath (C:/CGI/out/ in the above example).

B.3 Input file

Five input file are required as input file. In the above example,

(1) Gene expression data file = Alzheimerdata.txt

Same as section A.4.

(2) Protein interaction data file = HumanPPI_DF070.txt (the diffusion kernel in section 2.3 of the paper)

Same as section A.4.

(3) Gene name file = allgenes.txt

Same as section A.4.

(4) Phenotype name file = pheno_names.txt

This is the phenotype names file.

B.4 Output file MMSE_tau070.txt

The following is the content in the output file:

100	MCCScore	-0.1516	CGIScore	0.2296	MCCrank	2899	CGIrank	2089
1000	MCCScore	-0.2775	CGIScore	0.2336	MCCrank	3687	CGIrank	1475
10000	MCCScore	0.1254	CGIScore	0.2244	MCCrank	1408	CGIrank	2656
10001	MCCScore	-0.3590	CGIScore	0.2318	MCCrank	4124	CGIrank	1791
10006	MCCScore	0.3028	CGIScore	0.2383	MCCrank	661	CGIrank	810
10007	MCCScore	0.0745	CGIScore	0.2153	MCCrank	1660	CGIrank	3182
10010	MCCScore	0.1347	CGIScore	0.2346	MCCrank	1361	CGIrank	1296
10013	MCCScore	-0.0862	CGIScore	0.2289	MCCrank	2474	CGIrank	2175
			.					
			.					
			.					
MMSE	MCCScore	1.0000	CGIScore	1.0000	MCCrank	1	CGIrank	1

The first column is the Entrez gene ID.

The third column is its corresponding association score with the phenotype (here MMSE value).

The fifth column is its CGI score.

The seventh column is rank of this gene by MCC score.

The ninth column is the rank of this gene by CGI score.

All ranks are “smaller means better.” (note that the last row is the rank for the phenotype MMSE itself.)

Reference:

Xiaotu Ma, Hyunju Lee, Li Wang, and Fengzhu Sun. (2007) CGI: a new approach for prioritizing genes by Combining Gene expression and protein-protein Interaction data. *Bioinformatics*. (in press)