

Improving detection of differentially expressed gene sets by applying cluster enrichment analysis to Gene Ontology

Xutao

January 3, 2009

1. Overview

The CeaGO is a test for the detection of differentially expressed gene sets by applying cluster enrichment analysis to Gene Ontology (GO) (Ashburner, Ball et al. 2000). The clustered GO classes are obtained from a semantic similarity clustering process hampered by a pre-defined cutoff. Then the most significant GO subset is obtained by applying the PAGE (Kim and Volsky 2005) algorithm to test all subsets derived from the GO class.

Two kinds of test analysis are implemented in CeaGO: 1) **getSigClusters** function to enrich GO clusters. 2) **getSigTerms** function to enrich individual GO terms. In the example, a microarray data set (Chiaretti, Li et al. 2005) from a clinical trial in acute lymphoblastic leukemia (ALL) is used. Here, we focus on B-cell lines derived from ALL, and in particular, on comparing expression changes between BCR/ABL and NEG samples (those with no observed cytogenetic abnormalities).

2. Preprocessing and Inputs

First, we load package CeaGO and the ALL expression data set from (Chiaretti, Li et al. 2005). The data is collected for characterizing the relationship between gene expression signatures in ALL associated cells and genotypic abnormalities in adult patients. It consists of 12,625 genes and 128 samples. From several phenotype variables, we use BCR/ABL and NEG groups from B-cell lines for enrichment. There are 37 samples for the BCR/ABL group and 42 samples for the NEG group.

```
> library(CeaGO)
> library(ALL)
> data(ALL)
```

Normalization on these samples is carried out using the variance stabilizing method VSN (Huber, von Heydebreck et al. 2002).

```
> library(vsn)
> ALLadjust <- justvsn(ALL)
```

The data set is rearranged into the allowed pattern for CeaGO (such as 00001111, pattern like 011001 is forbidden).

```
> bcrabl <- exprs(ALLadjust)[, ALLadjust$mol.biol[in%"BCR/ABL"]
> neg <- exprs(ALLadjust)[, c((ALLadjust$mol.biol[in%"NEG"])[1:95], rep(FALSE, 33))]
> ex <- cbind(bcrabl, neg)
```

```
> ph <- c(rep(0, 37), rep(1, 42))
```

Note that, *ex* is the expression data and *ph* is a vector that defines the clinical diagnosis for the BCR/ABL and NEG group.

Next, we need to load the annotation data. The chip is HGU95aV2 Affymetrix.

```
> affyLib <- paste(annotation(ALL))
```

Next, a **CeaGOdata** object is build. The first input *ontology* is the ontology of interest (BP, MF or CC). The second input *allGenes* is the names of the genes/probe IDs. The third input *expressionMatrix* and *phenotype* is the expression matrix and phenotype vector for BCR/ABL and NEG, respectively. The *ann.FUN* function is an annotation function which is contained in the CeaGO package. The last input *affyLib* is the annotation dataset.

```
> CeaBPdata <- new("CeaGOdata", ontology = "BP", allGenes = rownames(ex),  
expressionMatrix = ex, phenotype = ph, annot = ann.FUN, affyLib = affyLib)
```

```
Building mounted GOs ..... ( 3066 GO terms found. )
```

```
Build GO DAG topology ..... ( 4913 GO terms and 8791 relations. )
```

```
Annotating nodes ..... ( 10503 genes annotated to the GO terms. )
```

User can type **CeaBPdata** to gain insight into the result of **CeaGOdata** object.

```
----- CeaGOdata object -----
```

Description:

-

Ontology:

- BP

12625 genes from the array:

- symbol: 1000_at 1001_at 1002_f_at 1003_s_at 1004_at ...

10503 number of genes mounted to GO:

- symbol: 32822_at 34988_at 36879_at 37181_at 38950_r_at ...

GO graph:

- a graph with directed edges

- number of nodes = 4913

- number of edges = 8791

```
----- CeaGOdata object -----
```

3. Enrichment of GO clusters

Given the **CeaGOdata** object, we are now ready to start the GO cluster analysis. Before the

enrichment, a list of parameters must be specified. The list is incorporated into a **CeaScore** object.

```
> para_c <- new("CeaScore", name = "Cluster test", height = 2, simmeasure = "Resnic")
```

Where *height* is the cutoff of the dendrogram of GO graph which is clustered by the hierarchical clustering algorithm and *simmeasure* is the semantic similarity algorithm to calculate the similarity between two GO terms. Note that, there are two semantic similarity algorithm implemented in the package: Resnic and Lin. The maximum of the height of dendrogram for Resnic is the maximum distance of two GO terms of the ontology.

We apply the `getSigClusters` function to enrich clustered GO terms and compute *p*-values. This process is performed in three steps: 1) Building dissimilarity matrix. 2) Building cluster environment. 3) Extracting the significant cluster from the cluster environment.

```
> c <- getSigClusters(CeaBPdata, para_c)
```

```
Building BP dissimilarity matrix...OK
```

```
Building BP cluster enviroment...OK
```

```
Extracting the significant clusters from the BP cluster enviroment...OK
```

This `getSigClusters` function can take a while, depending on the size of the GO graph under dissimilarity matrix calculating and clustering. Here, the gene set analysis algorithm implemented in the package is based on Z-statistic named PAGE.

Afterwards, a suitable correction for multiple testing has to be applied. In this example, the raw *p*-values are adjusted using the false discovery rate method from Benjamini and Yekutieli (Benjamini and Yekutieli 2001).

```
> p_c <- correctedPvalues(c, correction = "BY")
```

Next, we apply the `PrintTable` function to the *p_c*, the result with well-formatted table is provided to investigate the significant GO clusters.

```
> gotable_c <- PrintTable(CeaBPdata, "p-value" = p_c)
```

```
> print(gotable_c, right = FALSE)
```

NO.	GO ID	Term	<i>p</i> -value
1	GO:0043122	regulation of I-kappaB kinase/NF...	3.7e-06
	GO:0043123	positive regulation of I-kappaB kinase/...	
	GO:0043124	negative regulation of I-kappaB kinase...	
2	GO:0000084	S phase of mitotic cell cycle	2.9e-05
	GO:0000115	S-phase-specific transcription in mitoti...	
3	GO:0032715	negative regulation of interleukin-6 pro...	0.012
	GO:0032755	positive regulation of interleukin-6 pro...	

	GO:0042226	interleukin-6 biosynthetic process	
	GO:0045408	regulation of interleukin-6 biosynthetic...	
	GO:0045410	positive regulation of interleukin-6 bio...	
4	GO:0032088	inhibition of NF-kappaB transcription fa...	0.027
	GO:0043392	negative regulation of DNA binding	
	GO:0043433	negative regulation of transcription fac...	
5	GO:0007257	activation of JNK activity	0.045
	GO:0043507	positive regulation of JNK activity	

4. Enrichment of individual GO terms

Individual term enrichment is also implemented in the package. Given the **CeaGOdata** object, we specify a list of parameters into a **StScore** object.

```
> para_i <- new("stScore", name = "Individual test")
> i <- getSigTerms(CeaBPdata, para_i)
> p_i <- correctedPvalues(i, correction = "BY")
> gotable_i <- PrintTable(CeaBPdata, "p_values" = p_i, topNodes = 10)
> print(gotable_i, right = FALSE)
```

NO.	GO ID	Term	p-value
1	GO:0008037	cell recognition	5.6e-12
2	GO:0006955	immune response	7.2e-06
3	GO:0043123	positive regulation of I-kappaB kinase/N...	1.5e-05
4	GO:0007260	tyrosine phosphorylation of STAT protein	2.9e-05
5	GO:0007262	STAT protein nuclear translocation	0.00027
6	GO:0006928	cell motility	0.00027
7	GO:0006954	inflammatory response	0.00055
8	GO:0030036	actin cytoskeleton organization and biog...	0.00080
9	GO:0007067	mitosis	0.00084
10	GO:0007249	I-kappaB kinase/NF-kappaB cascade	0.00204

Another insight way of looking at the result is to investigate the list of GO cluster enrichment that is not presented in the top list of individual GO term analysis.

```
> gotable <- PrintTable(CeaBPdata, "p-value" = p_c, "Individual" = p_i)
> print(gotable_i, right = FALSE)
```

NO.	GO ID	Term	p-value	Rank
1	GO:0043122	regulation of I-kappaB kinase/NF...	3.7e-06	3
	GO:0043123	positive regulation of I-kappaB kinase/...		
	GO:0043124	negative regulation of I-kappaB kinase...		
2	GO:0000084	S phase of mitotic cell cycle	2.9e-05	

	GO:0000115	S-phase-specific transcription in mitoti...	
3	GO:0032715	negative regulation of interleukin-6 pro...	0.012
	GO:0032755	positive regulation of interleukin-6 pro...	
	GO:0042226	interleukin-6 biosynthetic process	
	GO:0045408	regulation of interleukin-6 biosynthetic...	
	GO:0045410	positive regulation of interleukin-6 bio...	
4	GO:0032088	inhibition of NF-kappaB transcription f...	0.027
	GO:0043392	negative regulation of DNA binding	
	GO:0043433	negative regulation of transcription fac...	
5	GO:0007257	activation of JNK activity	0.045
	GO:0043507	positive regulation of JNK activity	

5. Session Information

R version 2.7.1 (2008-06-23) i386-pc-mingw32

- locale:
LC_COLLATE=Chinese_People's Republic of China.936;
LC_CTYPE=Chinese_People's Republic of China.936;
LC_MONETARY=Chinese_People's Republic of China.936; LC_NUMERIC=C;
LC_TIME=Chinese_People's Republic of China.936
- attached base packages: splines, tools, stats, graphics, grDevices, utils, datasets, methods, base
- other attached packages: multtest_1.20.0, survival_2.34-1, hgu95av2.db_2.2.0, vsn_3.6.0, limma_2.14.5, affy_1.18.2, preprocessCore_1.2.0, affyio_1.8.0, lattice_0.17-8, ALL_1.4.4, CeaGO_1.0.0, SparseM_0.78, GO.db_2.2.0, AnnotationDbi_1.2.2, RSQLite_0.6-9, DBI_0.2-4, Biobase_2.0.1, graph_1.18.1
- loaded via a namespace (and not attached): cluster_1.11.11, grid_2.7.1

References

- Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nat Genet* **25**(1): 25-9.
- Benjamini, Y. and D. Yekutieli (2001). "The control of the false discovery rate in multiple testing under dependency." *Ann. Statist* **29**(4): 1165-1188.
- Chiaretti, S., X. Li, et al. (2005). "Gene expression profiles of B-lineage adult acute lymphocytic leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of transformation." *Clin Cancer Res* **11**(20): 7209-19.
- Huber, W., A. von Heydebreck, et al. (2002). "Variance stabilization applied to microarray data calibration and to the quantification of differential expression." *Bioinformatics* **18 Suppl 1**: S96-104.
- Kim, S. Y. and D. J. Volsky (2005). "PAGE: parametric analysis of gene set enrichment." *BMC Bioinformatics* **6**: 144.