Functional Enrichment Analysis

Duanchen Sun, Ling-Yun Wu

2018-12-17

Contents

1 NetGen: Network-based Generative Model

NetGen is a network-based generative model for functional enrichment analysis. We first load the required packages

```
library(CopTea)
options(scipen=0)
```

In this example, we use a small subset of GO annotation database, which contains 300 biological processes categories.

```
load("GO_BP_300.RData")
```

The annotation data is stored in a matrix with dimensions:

```
dim(annotation)
```

```
## [1] 6447  300
```

which indicates there are total 6447 genes and 300 GO terms in this annotation dataset.

Next, we load the protein-protein interaction (PPI) dataset:

```
load("PPI.RData")
```

The PPI network is given by its adjacent matrix as follows:

```
dim(adj_matrix)
```

```
## [1] 9453 9453
```

We simulate the gene list of interest:

```
load("active_gene.RData")
```

The list consists of 84 active genes which are derived from the true categories as follows.

```
True_Categories <- c("GO:0019614", "GO:1903249", "GO:2000506", "GO:0015985", "GO:0071962")
```

1.1 Mode1: fixed parameter strategy

In the first part of this example, we try to identify the most enriched categories using a given parameter setting.

```
Enriched_Categories <- netgen(annotation, adj_matrix, active_gene,
                              p1 = 0.8, p2 = 0.1, q = 0.001, alpha = 5, trace=TRUE)
```

```
## [1] "Compute the enriched categories using the given parameter combination."
## [1] "-Inf -Inf -256.938974204227"
## [1] "-256.938974204227 -Inf -255.156043848658"
## [1] "-255.156043848658 -256.938974204227 -253.47143212099"
## [1] "-253.47143212099 -255.156043848658 -252.833422047565"
## [1] "-252.833422047565 -253.47143212099 -256.862566747075"
```

The most enriched categories identified by NetGen are

```
Enriched_Categories
```

```
##        GO ID             p-value
## 1 GO:0019614  1.3282025203837e-07
## 2 GO:0015985 1.82262800994748e-06
## 3 GO:2000506   0.0130293159609119
## 4 GO:1903249   0.0130293159609119
```

and the false negative categories are

```
setdiff(True_Categories, Enriched_Categories[,1])
```

```
## [1] "GO:0071962"
```

1.2 Mode2: mixed parameter strategy

Instead of using a fixed parameter setting, we can run NetGen with several different parameter settings, and then select the result of highest enrichment significance.

```
p1 <- c(0.5, 0.8)
p2 <- c(0.1, 0.3)
q  <- 0.001
Enriched_Categories <- netgen(annotation, adj_matrix, active_gene,
                              p1, p2, q, alpha = 3, trace=FALSE)
```

```
## [1] "Compute the enriched categories using a mixed parameter selection strategy."
## [1] "Computing parameter combination p1=0.5, p2=0.1, q=0.001"
## [1] "Computing parameter combination p1=0.5, p2=0.3, q=0.001"
## [1] "Computing parameter combination p1=0.8, p2=0.1, q=0.001"
## [1] "Computing parameter combination p1=0.8, p2=0.3, q=0.001"
```

The combined p-values of mixed parameter strategy are

```
Enriched_Categories$Term_combined_pvalue
```

```
## [1] 2.037391e-15 8.117211e-14 2.037391e-15 2.830118e-11
```

And the most enriched categories and its corresponding parameter combination are

```
Enriched_Categories$mix_result[which.min(Enriched_Categories$Term_combined_pvalue)]
```

```
## $`p1=0.5 p2=0.1 q=0.001`
##        GO ID             p-value
## 1 GO:0019614  1.3282025203837e-07
## 2 GO:0015985 1.82262800994748e-06
## 3 GO:2000506   0.0130293159609119
## 4 GO:1903249   0.0130293159609119
```

2 CEA: Combination-based Enrichment Analysis model

CEA is a novel combination-based method for gene set functional enrichment analysis. It is based on a multi-objective optimization framework, and the adapted IMPROVED GREEDY algorithm was used to approximatively solve the problem.

We first load the required packages

```
library(CopTea)
options(scipen=0)
```

In this example, we use the same GO annotation database and the active gene list.

```
load("GO_BP_300.RData")
load("active_gene.RData")
```

The list consists of 84 active genes which are derived from the true categories as follows:

```
True_Categories <- c("GO:0019614", "GO:1903249", "GO:2000506", "GO:0015985", "GO:0071962")
```

Note that, not all the active genes are annotated in the annotation matrix.

```
sum(active_gene %in% rownames(annotation))
```

```
## [1] 39
```

We use the CEA function to identify the most enriched catgories.

```
Enrich_result <- CEA(annotation, active_gene, d = 0, times = 5, trace = TRUE)
```

```
## [1] "Number of annotated active genes: 39"
## [1] "Size of categories set in 1 out of 5 repeats : 24"
## [1] "Size of categories set in 2 out of 5 repeats : 24"
## [1] "Size of categories set in 3 out of 5 repeats : 24"
## [1] "Size of categories set in 4 out of 5 repeats : 24"
## [1] "Size of categories set in 5 out of 5 repeats : 24"
## [1] "Number of the identified category sets before unique is 120"
## [1] "Number of the identified category sets after  unique is 25"
```

The result contains the following components:

```
names(Enrich_result)
```

```
## [1] "p.values"   "coverage"   "category"   "annotation"
```

For example, we select the most enriched 5 category sets as the final outputs. These are:

```
Enrich_result$category[1:5]
```

```
## [[1]]
## [1] "GO:1903249" "GO:2000506" "GO:0015985" "GO:0019614" "GO:0006398"
## [6] "GO:0055014" "GO:0006933"
##
## [[2]]
## [1] "GO:1903249" "GO:2000506" "GO:0002763" "GO:0015985" "GO:0019614"
## [6] "GO:0006398" "GO:0055014" "GO:0006933"
##
## [[3]]
## [1] "GO:1903249" "GO:2000506" "GO:0002763" "GO:0015985" "GO:0019614"
## [6] "GO:0006398" "GO:0055014" "GO:0006933" "GO:0006541"
##
## [[4]]
## [1] "GO:1903249" "GO:2000506" "GO:0015985" "GO:0019614" "GO:0055014"
## [6] "GO:0006933"
##
## [[5]]
##  [1] "GO:1903249" "GO:2000506" "GO:0002763" "GO:0046503" "GO:0015985"
##  [6] "GO:0019614" "GO:0006398" "GO:0055014" "GO:0006933" "GO:0006541"
```

The related Fisher's exact test p-values and coverages are:

```
Enrich_result$p.values[1:5]
```

```
## [1] 7.916791e-21 8.951878e-21 1.992085e-20 9.322719e-20 9.323603e-20
```

```
Enrich_result$coverage[1:5]
```

```
## [1] 0.3333333 0.4102564 0.4358974 0.3076923 0.4615385
```

The false negative categories in the first category set are

```
setdiff(True_Categories, Enrich_result$category[[1]])
```

```
## [1] "GO:0071962"
```

We can obtain a more enriched result by setting a larger tolerance parameter d as:

```
Enrich_result <- CEA(annotation, active_gene, d = 1, times = 500, trace = FALSE)
```

```
## [1] "Number of annotated active genes: 39"
## [1] "Number of the identified category sets before unique is 11347"
## [1] "Number of the identified category sets after  unique is 793"
```

The Fisher's exact test p-value of the most enriched category is:

```
Enrich_result$p.values[1]
```

```
## [1] 1.7826e-21
```