Disease Ontology Semantic and Enrichment analysis

Guangchuang Yu, Li-Gen Wang

Jinan University, Guangzhou, China

April 21, 2012

1 Introduction

Disease Ontology (DO) provides an open source ontology for the integration of biomedical data that is associated with human disease. DO analysis can lead to interesting discoveries that deserve further clinical investigation.

DOSE was designed for semantic similarity measure and enrichment analysis.

Four information content (IC)-based methods, proposed by Resnik [Philip, 1999], Jiang [Jiang and Conrath, 1997], Lin [Lin, 1998] and Schlicker [Schlicker et al., 2006], and one graph structure-based method, proposed by Wang [Wang et al., 2007], were implemented. These methods were also implemented in our *GOSemSim* [Yu et al., 2010] package for measuring GO-term semantic similarities. Hypergeometric test [Boyle et al., 2004] was implemented for enrichment analysis.
To start with *DOSE* package, type following code below:

```
> library(DOSE)
> help(DOSE)
```

2 Semantic Similarity Measurement

The *DOSE*package contains functions to estimate semantic similarity of GO terms based on Resnik's, Lin's, Jiang and Conrath's, Rel's and Wang's method. Details about Resnik's, Lin's, and Jiang and Conrath's methods can be seen in [Lord et al., 2003], details about Rel's method can be seen in [Schlicker et al., 2006], and details about Wang's method can be seen in [Wang et al., 2007].

IC-based method depend on the frequencies of two DO terms involved and that of their closest common ancestor term in a specific corpus of DO annotations. Information content is defined as frequency of each term occurs in the corpus.

As DO allow multiple parents for each concept, two terms can share parents by multiple paths. We take the minimum p(t), where there is more than one shared parents. The p_{ms} is defined as:

$$p_{ms}(t1, t2) = \min_{t \in S(t1,t2)} \{p(t)\})$$

Where S(t1,t2) is the set of parent terms shared by t1 and t2.

- Resnik's method is defined as:

$$sim(t1, t2) = -\ln p_{ms}(t1, t2)$$

- Lin's method is defined as:

$$sim(t1, t2) = \frac{2 \times \ln(p_{ms}(t1, t2))}{\ln p(t1) + \ln p(c2)}$$

- Schlicker's method, which combine Resnik's and Lin's method, is defined as:

$$sim(t1, t2) = \frac{2 \times \ln p_{ms}(t1, t2)}{\ln p(t1) + \ln p(p2)} \times (1 - p_{ms}(t1, t2))$$

- Jiang and Conrath's method is defined as:

$$sim(t1, t2) = 1 - \min(1, d(t1, t2))$$

where

$$d(t1, t2) = \ln p(t1) + \ln p(p2) - 2 \times \ln p_{ms}(t1, t2)$$

Graph-based methods using the topology of DO graph structure to compute semantic similariy. Formally, a DO term A can be represented as $DAG_A = (A, T_A, E_A)$ where T_A is the set of DO terms in DAG_A, including term A and all of its ancestor terms in the DO graph, and E_A is the set of edges connecting the DO terms in DAG_A.

- Wang's method

  To encode the semantic of a DO term in a measurable format to enable a quantitative comparison, Wang firstly defined the semantic value of term A as the aggregate contribution of all terms in DAG_A to the semantics of term A, terms closer to term A in DAG_A contribute more to its semantics. Thus, defined the contribution of a DO term t to the semantics of DO term A as the S-value of DO term t related to term A. For any of term t in DAG_A, its S-value related to term A. $S_A(t)$ is defined as:

$$\begin{cases} S_A(A) = 1 \\ S_A(t) = \max\{w_e \times S_A(t') | t' \in childrenof(t)\} \text{ if } t \neq A \end{cases}$$

where w_e is the semantic contribution factor for edge $e \in E_A$ linking term t with its child term t'. Wang defined term A contributes to its own as one. After obtaining the S-values for all terms in DAG_A, the semantic value of GO term A, SV(A), is calculated as:

$$SV(A) = \sum_{t \in T_A} S_A(t)$$

Thus, given two DO terms A and B, the semantic similarity between these two terms, $DO_{A,B}$, is defined as:

$$S_{GO}(A, B) = \sum_{t \in T_A \cap T_B} \frac{S_A(t) + S_B(t)}{SV(A) + SV(B)}$$

where $S_A(t)$ is the S-value of DO term t related to term A and $S_B(t)$ is the S-value of DO term t related to term B.

This method proposed by Wang [Wang et al., 2007] determines the semantic similarity of two DO terms based on both the locations of these terms in the DO graph and their relations with their ancestor terms.

3 Enrichment Analysis

Enrichment analysis is a widely used approach to identify biological themes. Here we implement hypergeometric model to assess whether the number of selected genes associated with disease is larger than expected. We also implement a bar plot and gene-category-network for visualization.

- Calculation of Statistical Significance

  To determine whether any DO terms annotate a specified list of genes at frequency greater than that would be expected by chance, *DOSE*calculates a p-value using the hypergeometric distribution:

  $$p = 1 - \sum_{i=0}^{k-1} \frac{\begin{pmatrix} M \\ i \end{pmatrix} \begin{pmatrix} N-M \\ n-i \end{pmatrix}}{\begin{pmatrix} N \\ i \end{pmatrix}}$$

  In this equation, N is the total number of genes in the background distribution, M is the number of genes within that distribution that are annotated (either directly or indirectly) to the node of interest, n is the size of the list of genes of interest and k is the number of genes within that list which are annotated to the node. The background distribution by default is all the genes that have DO annotation.

4 Example

The following lines provide a quick and simple example on the use of *DOSE*.

- Calculate DO terms Similarity

```
> data(DO2EG)
> set.seed(123)
> terms <- list(a=sample(names(DO2EG), 5),b= sample(names(DO2EG), 6))
> terms

$a
[1] "DOID:1474" "DOID:6432" "DOID:2571" "DOID:8622"
[5] "DOID:9206"

$b
[1] "DOID:10591" "DOID:332"   "DOID:8689"  "DOID:3458"
[5] "DOID:2893"  "DOID:9346"

> ## Setting Parameters...
> params <- new("DOParams", IDs=terms, type="DOID", method="Wang")
> ## Calculating Semantic Similarities...
> sim(params)

          DOID:10591 DOID:332 DOID:8689 DOID:3458 DOID:2893
DOID:1474      0.100    0.078     0.041     0.026     0.026
DOID:6432      0.660    0.116     0.057     0.041     0.041
DOID:2571      0.139    0.116     0.057     0.041     0.041
DOID:8622      0.093    0.082     0.093     0.075     0.075
DOID:9206      0.173    0.149     0.071     0.055     0.055
          DOID:9346
DOID:1474     0.100
DOID:6432     0.139
DOID:2571     0.139
DOID:8622     0.093
DOID:9206     0.173
```

Four combine methods which called *max*, *average*, *rcmax* and *rcmax.avg*, were implmented to combine semantic similarity scores of multiple DO terms.

```
> params <- new("DOParams",
+               IDs=terms,
+               type="DOID",
+               method="Wang",
+               combine="rcmax.avg")
> sim(params)
```

```
[1] NaN
```

- Calculate Gene products Similarity

```
> geneid <- list(a=c("920", "100"),
+                b= c("919", "4221", "3458"))
> params <- new("DOParams",
+               IDs=geneid,
+               type="GeneID",
+               method="Wang",
+               combine="rcmax.avg")
> sim(params)

     919 4221  3458
920 -Inf -Inf 0.754
100 -Inf -Inf   NaN
```

- Enrichment analysis of a list of genes can also be performed as shown in the following examples.

```
> data(DO2ALLEG)
> genes = DO2ALLEG[[1]]
> genes

 [1] "10062" "335"   "338"   "339"   "341"   "345"   "348"
 [8] "367"   "3952"  "3953"  "4043"  "4276"  "4295"  "4586"
[15] "5320"  "581"   "58191" "64240" "64241" "7434"  "885"

> x <- enrichDO(genes, pvalueCutoff=0.05)
> head(summary(x))

                       ID
DOID:0000000 DOID:0000000
DOID:10211     DOID:10211
DOID:77           DOID:77
DOID:1936       DOID:1936
DOID:2348       DOID:2348
DOID:2349       DOID:2349
                                        Description
DOID:0000000                      gallbladder disease
DOID:10211                             cholelithiasis
DOID:77              gastrointestinal system disease
DOID:1936                           atherosclerosis
DOID:2348    arteriosclerotic cardiovascular disease
DOID:2349                          arteriosclerosis
             GeneRatio  BgRatio       pvalue       qvalue
DOID:0000000     21/21  21/2690 5.218390e-53 1.054664e-50
DOID:10211       19/21  19/2690 1.860766e-46 1.880353e-44
```

```
DOID:77          21/21 266/2690 3.799087e-22 2.559385e-20
DOID:1936        10/21 132/2690 1.286159e-08 5.198791e-07
DOID:2348        10/21 132/2690 1.286159e-08 5.198791e-07
DOID:2349        10/21 138/2690 1.990542e-08 6.617278e-07

DOID:0000000 10062/335/338/339/341/345/348/367/3952/3953/4043/4276/4295/4586/5320/581/5
DOID:10211             10062/335/338/339/341/345/348/367/3952/3953/4043/4276/4295/4586/
DOID:77      10062/335/338/339/341/345/348/367/3952/3953/4043/4276/4295/4586/5320/581/5
DOID:1936                                                                  10062/335/338/341/34
DOID:2348                                                                  10062/335/338/341/34
DOID:2349                                                                  10062/335/338/341/34
             Count
DOID:0000000    21
DOID:10211      19
DOID:77         21
DOID:1936       10
DOID:2348       10
DOID:2349       10

> setReadable(x) <- TRUE
```

5 Session Information

The version number of R and packages loaded for generating the vignette were:

```
R version 2.15.0 (2012-03-30)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=C                 LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets
[6] methods   base

other attached packages:
[1] org.Hs.eg.db_2.7.1   DO.db_2.4
[3] AnnotationDbi_1.18.0 Biobase_2.16.0
[5] BiocGenerics_0.2.0   DOSE_1.2.1
[7] RSQLite_0.11.1       DBI_0.2-5
```

```
> plot(x,showCategory=5, categorySize="geneNum",output="fixed")
```

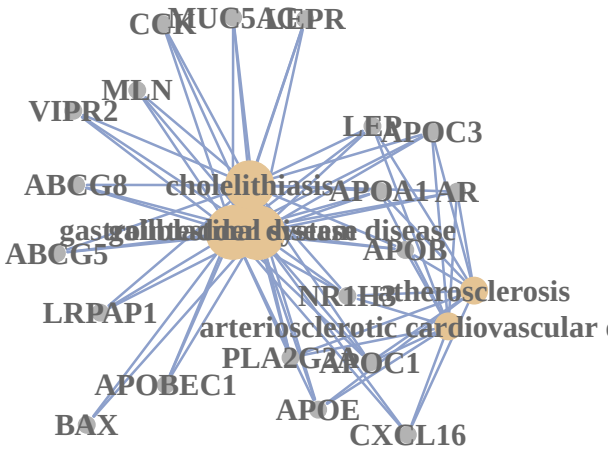

Figure 1: Category-Network Plot of Enrichment Result

```
loaded via a namespace (and not attached):
 [1] IRanges_1.14.2     MASS_7.3-17
 [3] RColorBrewer_1.0-5 colorspace_1.1-1
 [5] dichromat_1.2-4    digest_0.5.2
 [7] ggplot2_0.9.0      grid_2.15.0
 [9] igraph_0.5.5-4     memoise_0.1
[11] munsell_0.3        plyr_1.7.1
[13] proto_0.3-9.2      qvalue_1.30.0
[15] reshape2_1.2.1     scales_0.2.0
[17] stats4_2.15.0      stringr_0.6
[19] tcltk_2.15.0       tools_2.15.0
```

References

Elizabeth I Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J Michael Cherry, and Gavin Sherlock. GO::TermFinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics (Oxford, England)*, 20(18):3710–3715, December 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth456. URL `http://www.ncbi.nlm.nih.gov/pubmed/15297299`. PMID: 15297299.

Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of 10th International Conference on Research In Computational Linguistics*, 1997.

Dekang Lin. An Information-Theoretic definition of similarity. *In Proceedings of the 15th International Conference on Machine Learning*, pages 296—304, 1998.

P W Lord, R D Stevens, A Brass, and C A Goble. Semantic similarity measures as tools for exploring the gene ontology. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 601–12, 2003. ISSN 1793-5091. doi: 12603061. URL `http://www.ncbi.nlm.nih.gov/pubmed/12603061`. PMID: 12603061.

Resnik Philip. Semantic similarity in a taxonomy: An Information-Based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

Andreas Schlicker, Francisco S Domingues, JÃűrg RahnenfÃijhrer, and Thomas Lengauer. A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, 7:302, 2006. PMID: 16776819.

James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip S Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics (Oxford, England)*, 23:1274–81, May 2007. PMID: 17344234.

Guangchuang Yu, Fei Li, Yide Qin, Xiaochen Bo, Yibo Wu, and Shengqi Wang. Gosemsim: an r package for measuring semantic similarity among go terms and gene products. *Bioinformatics*, 26:976–978, 2010. PMID: 20179076.