Disease Ontology Semantic and Enrichment analysis

Guangchuang Yu, Li-Gen Wang

Jinan University, Guangzhou, China

April 4, 2013

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 DO 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:

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

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:

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

  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 *BMA*, were implmented to combine semantic similarity scores of multiple DO terms.

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

[1] 0.217
```

- Calculate Gene products Similarity

```
> geneid <- list(a=c("5320", "338"),
+                b= c("341", "581", "885"))
> params <- new("DOParams",
+               IDs=geneid,
+               type="GeneID",
+               method="Wang",
+               combine="BMA")
> x <- sim(params)
> x
```

```
       341   581   885
5320 0.886 0.687 0.729
338  0.847 0.643 0.610
```

DOSE implement `simplot` to visualize the semantic similarity matrix.

```
> simplot(x)
```

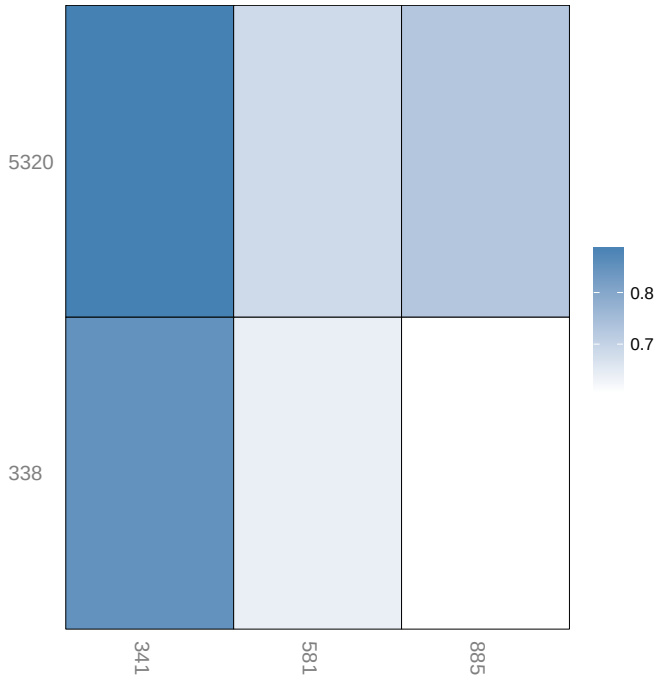

Figure 1: Heatmap plot for semantic similarity matrix

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

```
> data(AL1)
> x <- enrichDO(AL1, pvalueCutoff=0.05)
> head(summary(x))
```

```
                   ID       Description GeneRatio BgRatio
DOLite:214 DOLite:214         Gigantism      1/62  1/4051
DOLite:310 DOLite:310 Lewy body disease      1/62  1/4051
DOLite:258 DOLite:258     Hyperglycemia      3/62 39/4051
DOLite:168 DOLite:168  Encephalopathies      3/62 51/4051
               pvalue    qvalue          geneID Count
DOLite:214 0.01530486 0.5978242            2597     1
DOLite:310 0.01530486 0.5978242           10376     1
DOLite:258 0.02109968 0.5978242 10135/9588/2739     3
DOLite:168 0.04227576 0.7876214  2184/1967/2023     3
```

User can use the following command for mapping gene IDs to their corresponding gene symbol.

```
> setReadable(x) <- TRUE
> head(summary(x))
```

```
                   ID       Description GeneRatio BgRatio
DOLite:214 DOLite:214         Gigantism      1/62  1/4051
DOLite:310 DOLite:310 Lewy body disease      1/62  1/4051
DOLite:258 DOLite:258     Hyperglycemia      3/62 39/4051
DOLite:168 DOLite:168  Encephalopathies      3/62 51/4051
               pvalue    qvalue           geneID Count
DOLite:214 0.01530486 0.5978242            GAPDH     1
DOLite:310 0.01530486 0.5978242           TUBA1B     1
DOLite:258 0.02109968 0.5978242 NAMPT/PRDX6/GLO1     3
DOLite:168 0.04227576 0.7876214  FAH/EIF2B1/ENO1     3
```

DOSE package implement bar plot and gene-category network plot for visualization.

In the category-network plot, if expression values is provided, the `plot` function will use them to label the gene nodes. Red indicates up-regulated and green indicates down-regulated.

```
> AL1expr
```

```
  396  8349  7334 10539  7178 11332  5034  1891  1072  5094
-4.12 -2.14 -2.07 -3.03  2.40  2.34  2.39  2.07  2.22  2.56
 8668  7384  6282   708  3735  4436 10135  5713   811   811
-2.34 -3.72  2.53  2.13 -2.06 -2.82  2.03  4.04 -3.06  2.01
```

```
> plot(x, type="bar")
```

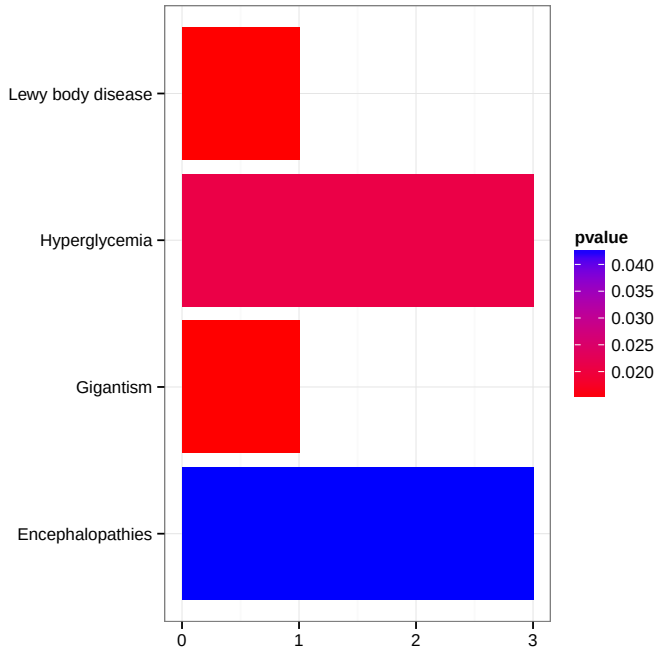

Figure 2: Bar Plot of Enrichment Result

```
  811   811 10102   839 10935  3315  8533  3417  7001  7001
-3.06  2.01  2.13  2.62 -2.57 -2.04  3.52  2.48 -7.47 -2.45
 7001  7001  5690 80273 25796  2184  9342  9380 28970  4678
-7.47 -2.45  2.51 -2.05  2.36  2.50 -2.17 -2.69  2.37  2.32
 3050  2288  2597 10327  9588  6636  2739  7520  1967 10015
 2.59 -2.06  2.53  3.04  2.25  3.34 -2.91  2.41  3.55  2.01
 8574 30968 23193   689  5478  2023  3925  7203  5721 10694
 2.51  2.24  2.33  2.38  2.77  3.06  2.17  2.23 -2.61 -3.23
10376  7296
-2.01  2.01
```

The plot was re-generate by using this log fold change expression values as follows:

5 Session Information

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

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

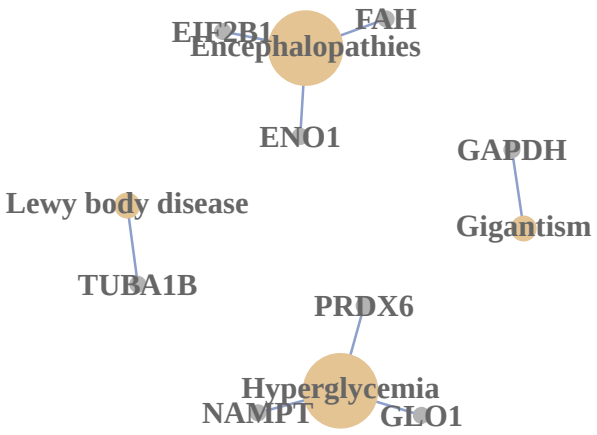

Figure 3: Category-Network Plot of Enrichment Result

```
R version 3.0.0 (2013-04-03)
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] parallel  stats     graphics  grDevices utils
[6] datasets  methods   base

other attached packages:
[1] org.Hs.eg.db_2.9.0   DO.db_2.6.0
```

```
> plot(x,showCategory=5, logFC=AL1expr, categorySize="geneNum",output="fixed")
```

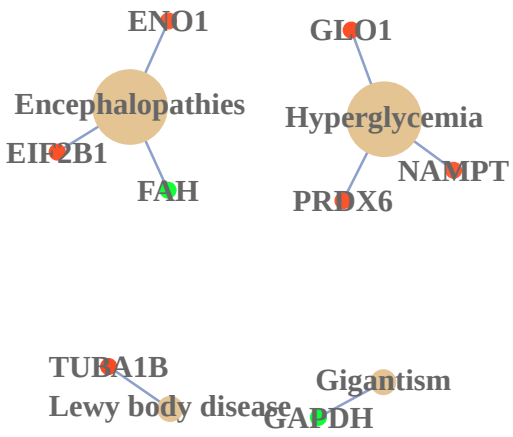

Figure 4: Category-Network Plot of Enrichment Result

```
[3] DOSE_1.6.0           AnnotationDbi_1.22.0
[5] Biobase_2.20.0       BiocGenerics_0.6.0
[7] RSQLite_0.11.2       DBI_0.2-5
[9] ggplot2_0.9.3.1

loaded via a namespace (and not attached):
 [1] GO.db_2.9.0         GOSemSim_1.18.0
 [3] IRanges_1.18.0      MASS_7.3-26
 [5] RColorBrewer_1.0-5 colorspace_1.2-1
 [7] dichromat_2.0-0     digest_0.6.3
 [9] grid_3.0.0          gtable_0.1.2
[11] igraph_0.6.5-1      labeling_0.1
[13] munsell_0.4         plyr_1.8
[15] proto_0.3-10        qvalue_1.34.0
[17] reshape2_1.2.2      scales_0.2.3
[19] stats4_3.0.0        stringr_0.6.2
[21] tcltk_3.0.0         tools_3.0.0
```

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.