

Using *DOSE* for Disease Ontology Semantic and Enrichment analysis

Guangchuang Yu

Jinan University, Guangzhou, China

November 1, 2011

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. The calculation details can be referred to the vignette of R package *GOSemSim* [Yu et al., 2010]. Hypergeometric test was implemented for enrichment analysis.

This document presents an introduction to the use of *DOSE*.

To start with *DOSE* package, type following code below:

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

2 Quick start

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:4001" "DOID:12328" "DOID:9563" "DOID:5583"
[5] "DOID:10587"

$b
[1] "DOID:0050127" "DOID:4772" "DOID:3674"
[4] "DOID:2917" "DOID:106" "DOID:450"

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

```

	DOID:0050127	DOID:4772	DOID:3674	DOID:2917
DOID:4001	0.025	0.149	0.111	0.034
DOID:12328	0.038	0.031	0.025	0.048
DOID:9563	0.172	0.031	0.025	0.116
DOID:5583	0.025	0.149	0.111	0.034
DOID:10587	0.064	0.024	0.020	0.080

	DOID:106	DOID:450
DOID:4001	0.025	0.025
DOID:12328	0.038	0.038
DOID:9563	0.038	0.093
DOID:5583	0.025	0.025
DOID:10587	0.029	0.064

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

```

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

[1] 0.116

```

- Calculate Gene products Similarity

```

> data(EG2DO)
> set.seed(123)
> geneid <- list(a=sample(names(EG2DO), 5), b= sample(names(EG2DO), 6))
> geneid

```

```

$a
[1] "2069" "6642" "1892" "11036" "3664"

```

```

$b
[1] "4772" "9436" "362" "613203" "6425" "6557"

```

```
> params <- new("DOParams", IDs=geneid, type="GeneID", method="Wang", combine="rcmax.av
> sim(params)
```

```
      4772 9436    362 613203 6425 6557
2069   -Inf -Inf  -Inf    NA -Inf    NA
6642  0.845   NA 0.018  0.028   NA 0.028
1892  0.213   NA 0.028  0.038   NA 0.036
11036 0.227   NA 0.170  0.200   NA 0.167
3664  0.071   NA 0.050  0.061   NA 0.053
```

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

```
> genes <- as.character(1:100)
> x <- enrichDO(genes, pvalueCutoff=0.05)
> summary(x)
```

	D0ID	Description
D0ID:3191	D0ID:3191	nemaline myopathy
D0ID:13068	D0ID:13068	renal osteodystrophy
D0ID:13336	D0ID:13336	congenital toxoplasmosis
D0ID:11758	D0ID:11758	iron deficiency anemia
D0ID:9965	D0ID:9965	toxoplasmosis
D0ID:2796	D0ID:2796	desquamative interstitial pneumonia
D0ID:2596	D0ID:2596	larynx cancer

	GeneRatio	BgRatio	pvalue	qvalue	geneID
D0ID:3191	2/100	6/3930	0.008994975	1	58/70
D0ID:13068	1/100	1/3930	0.025445293	1	54
D0ID:13336	1/100	1/3930	0.025445293	1	24
D0ID:11758	1/100	1/3930	0.025445293	1	48
D0ID:9965	1/100	1/3930	0.025445293	1	48
D0ID:2796	1/100	1/3930	0.025445293	1	21
D0ID:2596	2/100	14/3930	0.047812156	1	9/10

	Count
D0ID:3191	2
D0ID:13068	1
D0ID:13336	1
D0ID:11758	1
D0ID:9965	1
D0ID:2796	1
D0ID:2596	2

3 Session Information

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

R Under development (unstable) (2011-10-25 r57433)
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] DO.db_2.3.0	AnnotationDbi_1.17.0
[3] Biobase_2.15.0	DOSE_1.1.0
[5] RSQLite_0.10.0	DBI_0.2-5

loaded via a namespace (and not attached):

[1] IRanges_1.13.0	plyr_1.6	qvalue_1.29.0
[4] tcltk_2.15.0	tools_2.15.0	

References

- 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.
- 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, J  rg Rahnenf  jhrer, 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.