How To Use DOSim

Jiang Li

August 9, 2011

Contents

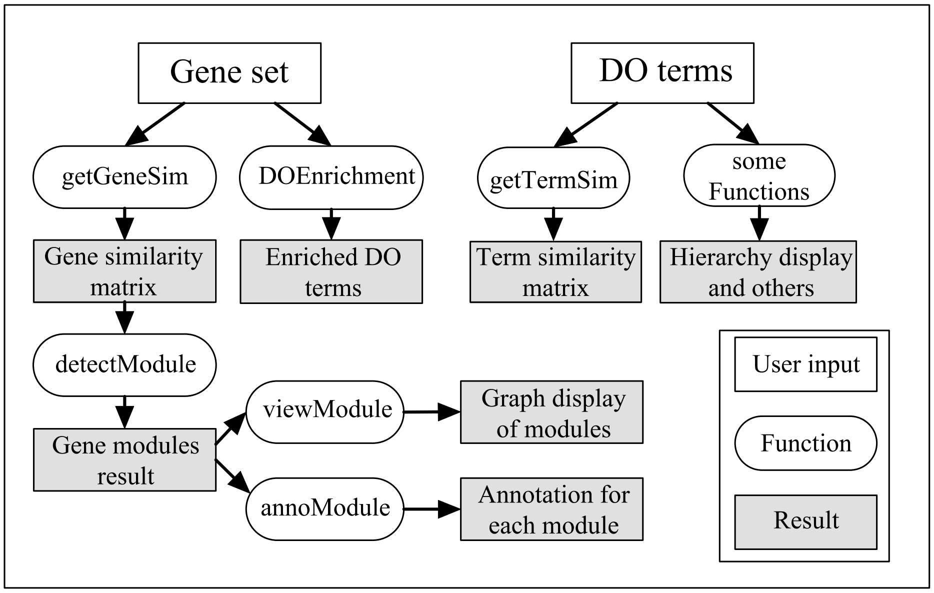
1 Overview

This vignette demonstrates how to use the DOSim package easily. DOSim is developed on DO to measure the similarity between DO terms, measure the similarity between human genes based on DO, detect disease-related gene modules and explore their functional meaning from gene sets, conduct DO enrichment analysis, and visualize hierarchies in DO and extract related terms for the given DO terms. It focuses on the reflection of the modular characteristics of disease related genes and we believe it will promote our understanding of the complex pathogenesis of diseases.

To use DOSim package, type the following codes to get a summary of DOSim and the document for each function:

```
> library(DOSim)
> help(DOSim)
```

In the following text, we will introduce the usage of DOSim mainly into two parts, one uses genes as data source and the other uses DO terms as data source. The flow chart of DOSim is shown as below.



2 Analysis for gene sets

Using gene sets as the data source, users could calculate the gene similarity matrix and further detect the modules on it, or simply conduct a DO enrichment analysis.

2.1 Conducting DO enrichment analysis

In DOSim, DO-based enrichment analysis is implemented to explore the disease feature of the gene sets. Significance of the enrichment analysis is assessed by hypergeometric test and the p value is adjusted by false discovery rate (FDR). DOSim selects the DO terms satisfied two criterions for enrichment analysis. One criterion is that the term should include 'n' genes, the other is that it should be the terms beneath depth 'm' in the DAG of DO, where 'n' and 'm' can be set by users when conducting DO enrichment analysis.

To do it, you can simply invoke the function *DOEnrichment*. Here is an example.

```
> genelist = getDefaultBackground()[1:10]
> DOEnrichment(genelist, filter = 5, cutoff = 0.01, layer = NULL)
```

```
               DOID                                  Term annGeneNumber
DOID:934   DOID:934              viral infectious disease             5
DOID:1117 DOID:1117 respiratory system infectious disease             2
          annBgNumber geneNumber bgNumber      odds       pvalue       qvalue
DOID:934           10         77     4054  26.32468 5.076558e-07 6.091869e-05
DOID:1117          10          8     4054 101.35000 1.521626e-04 9.129758e-03
```

2.2 Measuring the similarity between human genes based on DO

In our package, we calculate the similarity between two genes based on the similarity of their DO term annotation groups (See section 3.1). Five different methods are implemented in DOSim, which are the arithmetic maxima and average of pairwise similarity between two groups of DO terms describing the two genes (max, mean) [1], the arithmetic maxima and average between similarities for two directional comparisons of the similarity matrix S of two genes (funSimMax, funSimAvg)[2], and the best-match average approach (BMA) [3].

Let DO_1 and DO_2 be the groups of annotation terms for two genes g_1 and g_2, and m and n are the number of terms included in DO_1 and DO_2 respectively. A similarity matrix S contains all pairwise similarity scores of mappings from DO_1 to DO_2 and vice verse with size mxn. '*rowScore*' and '*columnScore*' of S are the averages over the row maxima and the column maxima, which give similarity scores for the comparison of DO_1 to DO_2 and the comparison of DO_2 to DO_1, respectively.

$$rowScore = \frac{1}{m}\sum_{i=1}^{m}\max_{1\le j\le n} s_{ij} \tag{1}$$

$$columnscore = \frac{1}{n}\sum_{j=1}^{n}\max_{1\le i\le m} s_{ij} \tag{2}$$

With these definitions, the five similarity methods for the computation of gene similarity between two genes g_1 and g_2 are defined as follows:

$$Sim_{max}\left(g_1, g_2\right) = \max_{1\le i\le m, 1\le j\le n} s_{ij} \tag{3}$$

$$Sim_{mean}\left(g_1, g_2\right) = \frac{1}{m\times n}\sum_{i=1}^{m}\sum_{j=1}^{n} s_{ij} \tag{4}$$

$$Sim_{funSimMax}\left(g_1, g_2\right) = \max\{rowScore, columnScore\} \tag{5}$$

$$Sim_{funSimAvg}\left(g_1, g_2\right) = \frac{rowScore + columnScore}{2} \tag{6}$$

$$Sim_{BMA}(g_1, g_2) = \frac{\sum\limits_{i=1}^{m} \max\limits_{1 \le j \le n} s_{ij} + \sum\limits_{j=1}^{n} \max\limits_{1 \le i \le m} s_{ij}}{m + n} \tag{7}$$

To do it, you can simply invoke the function *getGeneSim*. Here is an example to get five genes pairwise similarities.

```
> genelist <- c("10003", "10008", "10015", "10042", "10036")
> gsim <- getGeneSim(genelist, similarity = "BMA", similarityTerm = "Resnik")

> gsim
```

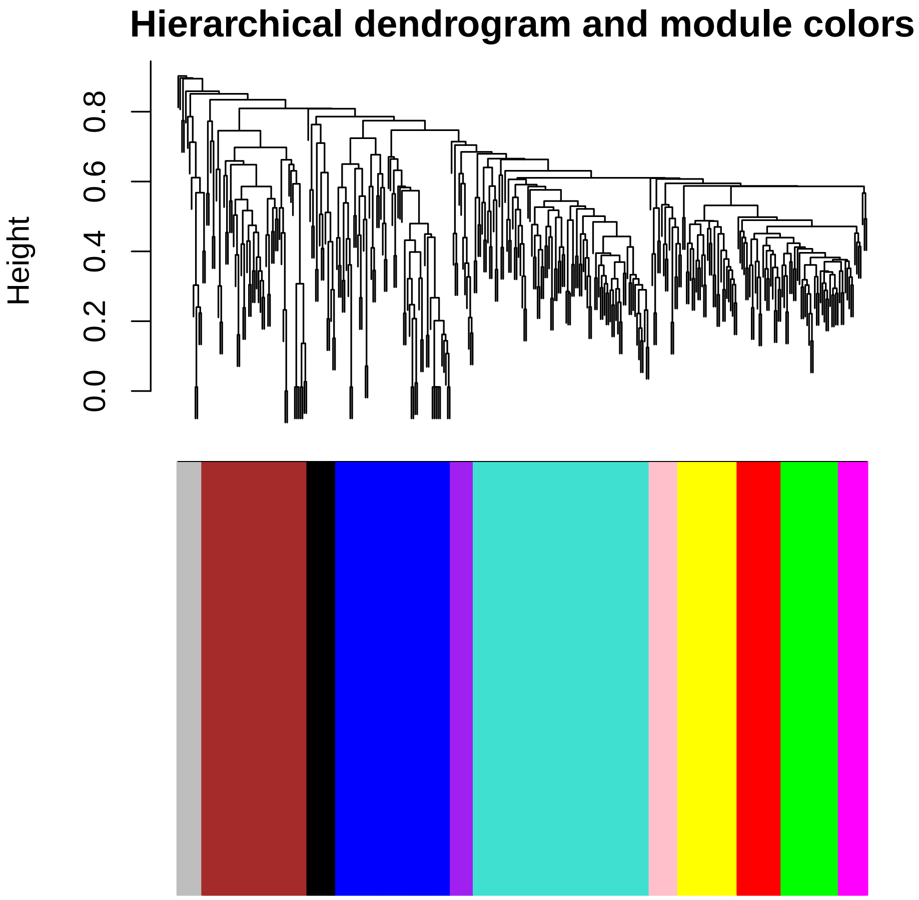
```
          10003 10008      10015 10042      10036
10003 1.0000000     0 0.00000000     0 0.12921344
10008 0.0000000     1 0.00000000     0 0.00000000
10015 0.0000000     0 1.00000000     0 0.03210972
10042 0.0000000     0 0.00000000     1 0.00000000
10036 0.1292134     0 0.03210972     0 1.00000000
```

2.3 Detecting gene modules and multilayer annotation

Gene module is a group of highly correlated genes. In DOSim, for a gene set, once the gene similarity matrix has been constructed, a hierarchical clustering is performed using the standard R function hclust and one of the three branch cutting methods is applied (one constant-height cutting and two dynamic branch cutting methods are embed in our package) [4], then the gene modules can be detected. After the gene modules have been detected, DOSim provides multilayer enrichment analysis (DO, GO and KEGG annotation) to explore the biological meaning implied in the modules, where DO annotations are conducted with DO enrichment analysis (section 2.1) , the GO annotations are conducted with the GOSim [5] and the KEGG annotations are gotten by SubpathwayMiner [6].

Meanwhile, we provide a function to visualize the module result. Here, we demonstrate the module detection and visualization of detected module applied on the obesity genes.

```
> data(obesity)
> module <- detectModule(obesity, method = "tree", minClusterSize = 10)
> viewModule(module)
```



3 Analysis for DO terms

Using DO terms as the data source, users can obtain the term similarity matrix (disease similarity matrix) and other information for DO term, e.g., the hierarchical structure relationship of the given DO terms.

3.1 Measuring similarity between DO terms

Here, we implemented ten semantic similarity measures for DO term pairs in DOSim, which are Resnik measure [7], Lin measure [8], Jiang and Conrath measure (JC) [9], Relevance measure (relevance) [2], Graph Information Content measure (GIC) [10], Information Coefficient similarity measure (simIC) [11], Wang measure [3], modified Resnik measure (CoutoResnik) [12], modified Lin measure (CoutoLin) [12], and modified Jiang and Conrath measure (CoutoJC) [12] respectively. Except that the Wang measure uses a hybrid measure, the other nine measures are based on information content (IC).

The IC of a term t is defined as $IC(t) = -logp(t)$, where $p(t)$ is the number of genes annotated to the term t and its descendants divided by the number of all genes annotated to DO. When characterizing the shared IC between two terms, two concepts, which are most information common ancestor (MICA) and disjunctive common ancestor (DCA), are widely used [12]. The MICA of two terms t_1 and t_2 is the one that possesses the maximum IC among all the common ancestor terms of t_1 and t_2. And the DCAs of two terms t_1 and t_2 are the MICA of disjunctive ancestors of t_1 and t_2, which can be defined as follows:

$$\begin{aligned} &DisjCommonAnc\left(t_1,t_2\right) = \{a_1 \mid \\ &\quad a_1 \in CommonAnc\left(t_1,t_2\right) \wedge \\ &\quad \forall a_2 : \left[\left(a_2 \in CommonAnc\left(t_1,t_2\right)\right) \wedge \left(IC\left(a_1\right) \leq IC\left(a_2\right)\right)\right] \Rightarrow \\ &\quad \left[\left(a_1,a_2\right) \in \left(DisjAnc\left(t_1\right) \cup DisjAnc\left(t_2\right)\right)\right]\} \end{aligned} \tag{8}$$

where disjunctive ancestors of the term t, $DisjAnc(t)$, can be described as that two ancestors a_1 and a_2 are disjunctive ancestors of the term t if there is a path from a_1 to t not passing through a_2 and a path from a_2 to t not passing through a_1. It can be formulated as follows:

$$\begin{aligned} &DisjAnc\left(t\right) = \{\left(a_1,a_2\right) \mid \\ &\quad \left(\exists p : \left(p \in Paths\left(a_1,t\right)\right) \wedge \left(a_2 \notin p\right)\right) \wedge \\ &\quad \left(\exists p : \left(p \in Paths\left(a_2,t\right)\right) \wedge \left(a_1 \notin p\right)\right)\} \end{aligned} \tag{9}$$

Then the shared information of two terms t_1 and t_2, $Share(t1,t2)$, is defined as the average of the IC of the DCAs, which is formulated as follows:

$$Share\left(t_1,t_2\right) = \overline{\{IC\left(a\right) \mid a \in DisjCommonAnc\left(t_1,t_2\right)\}} \tag{10}$$

Let t_{MICA} represents the MICA term of two terms t_1 and t_2, then the nine IC-based similarity measures are calculated as follows:

$$Sim_{Resnik}\left(t_1,t_2\right) = IC\left(t_{MICA}\right) \tag{11}$$

$$Sim_{Lin}\left(t_1,t_2\right) = \frac{2 \times IC\left(t_{MICA}\right)}{IC\left(t_1\right) + IC\left(t_2\right)} \tag{12}$$

$$Sim_{JC}\left(t_1,t_2\right) = 1 - \min\left(1, IC\left(t_1\right) + IC\left(t_2\right) - 2 \times IC\left(t_{MICA}\right)\right) \tag{13}$$

$$Sim_{relevance}\left(t_1,t_2\right) = Sim_{Lin}\left(t_1,t_2\right) \times \left(1 - p\left(t_{MICA}\right)\right) \tag{14}$$

$$Sim_{GIC}\left(t_1,t_2\right)=\frac{\sum\limits_{t\in\left(Ancestor\left(t_1\right)\cap Ancestor\left(t_2\right)\right)} IC\left(t\right)}{\sum\limits_{t\in\left(Ancestor\left(t_1\right)\cup Ancestor\left(t_2\right)\right)} IC\left(t\right)} \tag{15}$$

$$Sim_{simIC}\left(t_1,t_2\right)=Sim_{Lin}\times\left(1-\frac{1}{1+IC\left(t_{MICA}\right)}\right) \tag{16}$$

$$Sim_{CoutoResnik}\left(t_1,t_2\right)=Share\left(t_1,t_2\right) \tag{17}$$

$$Sim_{CoutuLin}\left(t_1,t_2\right)=\frac{2\times Share\left(t_1,t_2\right)}{IC\left(t_1\right)+IC\left(t_2\right)} \tag{18}$$

$$Sim_{CoutoJC}\left(t_1,t_2\right)=1-\min\left(1,IC\left(t_1\right)+IC\left(t_2\right)-2\times Share\left(t_1,t_2\right)\right) \tag{19}$$

In Wang measure, each edge is given a weight according to the types of relationships. For a term A, a sub-DAG comprised of the term A and all its ancestor terms can be represented as $DAG_A=(A,T_A,E_A)$, where T_A is the ancestor term set of the term A (including A itself) and E_A is the set of edges connecting to the terms in DAG_A. For any term t in DAG_A, Wang et al. defined the semantic contribution of t to A, $DA(t)$, as the product of all the edge weights in the "best" path from term t to A, where the "best" path is the one that maximizes the product (the semantic contribution of the term A to itself is set to 1). It could be represented as follow:

$$\begin{cases} S_A\left(A\right)=1 \\ S_A\left(t\right)=\max\left\{w_e\times S_A\left(t'\right)\mid t'\in childrenof\left(t\right)\right\} \quad if\ t\neq A \end{cases} \tag{20}$$

where w_e is the semantic contribution factor of edge e ($e\in E_A$). It is set between 0 and 1 according to the types of relationships, e.g., "is-a" or "part-of". In DO, there is only one type of relationships, defined as "is-a", and we set w_e to 0.7 in DOSim. Then the semantic similarity between two terms A and B is calculated as follows:

$$Sim_{Wang}\left(A,B\right)=\frac{\sum\limits_{t\in T_A\cap T_B}\left(S_A\left(t\right)+S_B\left(t\right)\right)}{SV\left(A\right)+SV\left(B\right)} \tag{21}$$

where $SV\left(A\right)$ (or $SV\left(B\right)$) is the total semantic contribution to term A (or B) in DAG_A (or DAG_B), which could be calculated as follows:

$$SV\left(A\right)=\sum_{t\in T_A}S_A\left(t\right),\quad SV\left(B\right)=\sum_{t\in T_B}S_B\left(t\right) \tag{22}$$

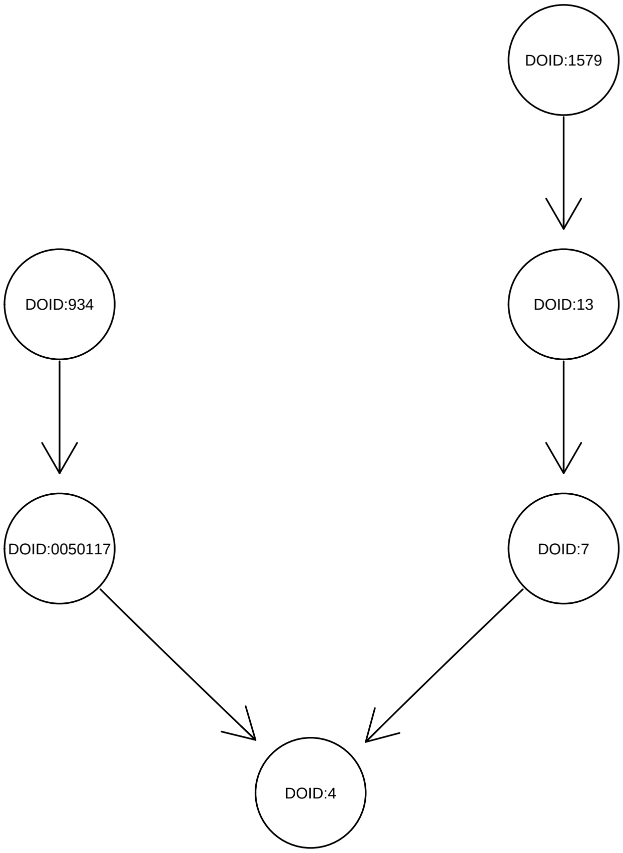
As terms in DO are disease names or disease-related concepts. Exploring the similarity between them can facilitate us to understand the similarity between diseases. Here we take an example to use the relevance measure to calculate four DO terms pairwise similarity. The code and result are below:

```
> termlist = c("DOID:399", "DOID:1117", "DOID:2313", "DOID:2040")
> tsim <- getTermSim(termlist, method = "relevance", verbose = TRUE)

> tsim

           DOID:399 DOID:1117 DOID:2313 DOID:2040
DOID:399  0.9765664 0.3421396 0.9609378         0
DOID:1117 0.3421396 0.9610261 0.3471034         0
DOID:2313 0.9609378 0.3471034 0.9740997         0
DOID:2040 0.0000000 0.0000000 0.0000000         1
```

3.2 Displaying DO hierarchical structures

DO is a collection of terminologies associated with human diseases and the terms in DO are organized in DAG. Hierarchical structures of DO terms can be represented as a *graphNEL* object and function *getDOGraph* in DOSim can be used to fetch the DO graph with specified DO terms at its leave. A demonstration is shown below:

```
> terms <- c("DOID:934", "DOID:1579")
> if (require(graph)) {
+     g <- getDOGraph(terms)
+     if (require(Rgraphviz)) {
+         plot(g)
+     }
+ }
```



3.3 Extracting related terms for the given DO terms

Here, we provide functions for users to extracting related terms for the given DO terms (e.g., get a DO terms parent terms). This includes a series of functions, they are described in the following sub-sections.

3.3.1 getParents

Returns a list of all direct parents associated to each DO term.

```
> terms <- c("DOID:934", "DOID:1579")
> getParents(terms)

[1] "Start to fetch the parents"
$`DOID:934`
[1] "DOID:0050117"
```

```
$`DOID:1579`
[1] "DOID:13"
```

3.3.2 getAncestors

Returns the list of all ancestors associated to each DO term.

```
> terms <- c("DOID:934", "DOID:1579")
> getAncestors(terms)

[1] "Start to fetch the ancestors"
$`DOID:934`
[1] "DOID:0050117" "DOID:4"

$`DOID:1579`
[1] "DOID:4"  "DOID:13" "DOID:7"
```

3.3.3 getOffsprings

Returns the list of all offsprings associated to each DO term.

```
> terms <- c("DOID:10533", "DOID:550")
> getOffsprings(terms)

[1] "Start to fetch the offsprings"
$`DOID:10533`
[1] "DOID:14473" "DOID:14476" "DOID:14475" "DOID:10510" "DOID:14474"
[6] "DOID:14472" "DOID:14477"

$`DOID:550`
[1] NA
```

3.3.4 getChildren

Returns the list of all direct children associated to each DO term.

```
> terms <- c("DOID:934", "DOID:1579")
> getChildren(terms)

[1] "Start to fetch the children"
$`DOID:934`
 [1] "DOID:0050079" "DOID:10533"   "DOID:1301"    "DOID:1329"    "DOID:13801"
 [6] "DOID:1385"    "DOID:1884"    "DOID:2295"    "DOID:2931"    "DOID:2932"
[11] "DOID:2937"    "DOID:2940"    "DOID:2947"    "DOID:2950"    "DOID:3294"
[16] "DOID:4121"    "DOID:623"     "DOID:6297"    "DOID:8568"    "DOID:8672"
```

```
[21] "DOID:8867"    "DOID:937"

$`DOID:1579`
 [1] "DOID:0050161" "DOID:10458"   "DOID:11091"   "DOID:1116"    "DOID:11565"
 [6] "DOID:1273"    "DOID:2945"    "DOID:4298"    "DOID:4493"    "DOID:9395"
[11] "DOID:974"
```

3.3.5 getDoTerm

Returns the list of DO term's name associated to each DO ID.

```
> terms <- c("DOID:934", "DOID:1579")
> getDoTerm(terms)
```

```
$`DOID:934`
[1] "viral infectious disease"

$`DOID:1579`
[1] "respiratory system disease"
```

3.3.6 getDoAnno

Get gene list associated to each DO term

```
> terms <- c("DOID:1579")
> getDoAnno(terms)
```

```
$`DOID:1579`
[1] "1636"
```

References

[1] P~W Lord~* AB R D~Stevens, Goble CA: **Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation**. *Bioinformatics* 2003, **19**(10):1275–1283.

[2] A~Schlicker FD: **A new measure for functional similarity of gene products based on Gene Ontology**. *BMC Bioinformatics* 2006.

[3] James~ZWang ZD: **A new method to measure the semantic similarity of GO terms**. *Bioinformatics* 2007, :1274–1281.

[4] Langfelder P, Zhang B, Horvath S: **Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R**. *Bioinformatics* 2008, **24**(5):719–720, [[http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/5/719]].

[5] Frohlich H, Speer N, Poustka A, BeiSZbarth T: **GOSim - an R-package for computation of information theoretic GO similarities between terms and gene products**. *BMC Bioinformatics* 2007, **8**:166.

[6] Li C, Li X, Miao Y, Wang Q, Jiang W, Xu C, Li J, Han J, Zhang F, Gong B, Xu L: **SubpathwayMiner: a software package for flexible identification of pathways**. *Nucl. Acids Res.* 2009, **37**(19):e131–, [[http://nar.oxfordjournals.org/cgi/content/abstract/37/19/e131]].

[7] Resnik P: **Using Information Content to Evaluate Semantic Similarity in a Taxonomy**. *Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal* 1995, **1**:448–453.

[8] Lin D: **An Information-Theoretic Definition of Similarity** 1998, :296–304.

[9] Jiang J, Conrath D: **Semantic similarity based on corpus statistics and lexical taxonomy**. *Proceedings of the International Conference on Research in Computational Linguistics, Taiwan* 1998.

[10] C˜Pesquita DF: **Evaluating GO-based Semantic Similarity Measures**. *In: Proc. 10th Annual Bio-Ontologies Meeting* 2007, :37–40.

[11] B˜Li AF J˜Wang: **Effectively Integrating Information Content and Structural Relationship to Improve the GO-based Similarity Measure Between Proteins**. *BMC Bioinformatics* 2009.

[12] Couto F, Silva M, Coutinho P: **Semantic Similarity over the Gene Ontology: Family Correlation and Selecting Disjunctive Ancestors**. *Conference in Information and Knowledge Management* 2005.