GO-terms Semantic Similarity Measures

Guangchuang Yu

January 23, 2010

1 Introduction

Functional similarity of gene products can be estimated by controlled biological vocabularies, such as Gene Ontology (GO). GO comprises of three orthogonal ontologies, i.e. molecular function (MF), biological process (BP), and cellular component (CC).

Four methods have been presented to determine the semantic similarity of two GO terms based on the annotation statistics of their common ancestor terms (Resnik[Philip, 1999], Jiang[Jiang and Conrath, 1997], Lin[Lin, 1998] and Schlicker[Schlicker et al., 2006]). Wang [Wang et al., 2007] proposed a new method to measure the similarity based on the graph structure of GO. Each of these methods has its own advantages and weaknesses. The *GOSemSim* package is developed to compute semantic similarity among GO terms, sets of GO terms, gene products, and gene clusters, providing both five methods mentioned above.

2 Semantic Similarity Measurement Based on GO

The *GOSemSim* 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], details about Wang's method can be seen in [Wang et al., 2007].

Formally, a GO term A can be represented as $DAG_A = (A, T_A, E_A)$ where T_A is the set of GO terms in DAG_A, including term A and all of its ancestor terms in the GO graph, and E_A is the set of edges connecting the GO terms in DAG_A.

To encode the semantics of a GO term in a measurable format to enable a quantitative comparison between two term's semantics, we firstly define 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, define the contribution of a GO term t to the semantics of GO term A as the S-value of GO term t related to term A. For any of term t in $DAG_A = (A, T_A, E_A)$, 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'. We 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:

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

Given two GO terms A and B, the semantic similarity between these two terms, $GO_{A,B}$, 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 GO term t related to term A and $S_B(t)$ is the S-value of GO term t related to term B.

The method descriped above is proposed by Wang[Wang et al., 2007]. The Wang's method determines the semantic similarity of two GO terms based on both the locations of these terms in the GO graph and their relations with their ancestor terms.

The other four methods proposed by Resnik[Philip, 1999], Jiang[Jiang and Conrath, 1997], Lin[Lin, 1998] and Schlicker[Schlicker et al., 2006] are information content (IC) based, which depend on the frequencies of two GO terms involved and that of their closest common ancestor term in a specific corpus of GO annotations. Information content is defined as frequency of each term occurs in the corpus. At present, *GOSemSim* supports analysis on nine species, which are human, fly, mouse, rat, yeast, zebrafish, worm, E coli strain 12 and arabidopsis. We used Bioconductor package org.Hs.eg.db, org.Dm.eg.db, org.Mm.eg.db, org.Rn.eg.db, org.Sc.sgd.db, org.Dr.eg.db, org.Ce.eg.db, org.EcK12.eg.db, org.At.tair.db to calculate the information content of human, fly, mouse, rat, yeast, zebrafish, worm, E coli strain 12 and arabidopsis species respectively. The information content will update regularly.

As GO allow multiple parents for each concept, two terms can share parents by multiple paths. We take the mininmum p(t), where there is more than on 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. The similarity of Resnik's method is defined as:

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

The similarity of Lin's is defined as:

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

The similarity of Schlicker's method combine Resnik's and Lin's method is defined as:

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

The Jiang and Conrath's method define a semantic similarity as:

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

where

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

In *GOSemSim*, on the basis of semantic similarity between GO terms, we can also compute semantic similarity among sets of GO terms, gene products, and gene clusters. The semantic similarity of one GO term go and a GO terms set $GO = \{go_1, go_2 \cdots go_k\}$ is defined as:

$$Sim(go, GO) = \max_{1 \le i \le k} (S_{GO}(go, GO_i))$$

Therefore, given two GO terms sets $GO_1 = \{go_{11}, go_{12} \cdots go_{1m}\}$ and $GO_2 = \{go_{21}, go_{22} \cdots go_{2n}\}$, the semantic similarity between these two sets is defined as:

$$Sim(GO1, GO2) = \frac{\sum\limits_{1 \le i \le m} Sim((go_{1i}), (GO_2)) + \sum\limits_{1 \le j \le n} Sim((go_{2j}), (GO_1))}{m+n}$$

3 Functions

Six functions are provided by *GOSemSim* package. The function *goSim*, *mgoSim*, *geneSim*, *clusterSim* can compute the semantic similarity among GO terms, sets of GO terms, GO descriptions of gene products and GO descriptions of gene clusters respectively. The function *mgeneSim* and *mclusterSim* are designed to calculate the similarity scores matrix of a set of genes and gene clusters.

For all these six functions, the ontology of GO used in measurement can be restricted by assigning the corresponding parameter to ”BP” (biological process), ”MF” (molecular function) and ”CC” (cellular component).

The function *goSim* gives the semantic similarity score for a pair of GO terms. The output value of goSim is between 0 and 1. The higher the value obtained the more similar between them. For example:

```
> library(GOSemSim)
> goSim("GO:0004022", "GO:0005515", ont = "MF", measure = "Wang")
```

```
[1] 0.252
```

The function *mgoSim* generates the similarity score of two GO terms lists. For example:

```
> go1 = c("GO:0004022", "GO:0004024", "GO:0004174")
> go2 = c("GO:0009055", "GO:0005515")
> mgoSim(go1, go2, ont = "MF", measure = "Wang")
```

```
[1] 0.299
```

The function *geneSim* estimate semantic similarity of two genes. The mapping from Gene IDs to GO IDs can be restricted based on evidence codes. Gene IDs and species are needed for the function. For human, rat, mouse, zebrafish, fly, worm and e coli strain 12, the Gene IDs is refer to Entrez Gene IDs, while for yeast, the Gene IDs is refer to ORF identifiers from Saccharomyces Genome Database (SGD), and for arabidopsis, the Gene IDs is refer to TAIR gene identifiers from The Arabidopsis Information Resource (TAIR). For example:

```
> geneSim("241", "2561", ont = "MF", organism = "human", measure = "Wang")
```

```
$geneSim
[1] 0.185

$GO1
[1] "GO:0047485" "GO:0050544"

$GO2
[1] "GO:0004890"
```

For reduce the dependences of *GOSemSim*, it only depends on species specific annotation package org.Hs.eg.db for human. When calculated semantic similarities of other species, *GOSemSim* will auto load the corresponding annotation package if it was installed or download and install the package automatically if it was not installed.

Given GO based similarity scores, gene products may be clustered by their function. GOSemSim package provides *mgeneSim* function to compute pairwise similarity scores for a list of genes. The *mgeneSim* function will automatically remove the genes without annotations. Gene IDs and species are needed for the function. For human, rat, mouse, zebrafish, fly, worm and e coli strain 12, the Gene IDs is refer to Entrez Gene IDs, while for yeast, the Gene IDs is refer to ORF identifiers from Saccharomyces Genome Database (SGD), and for arabidopsis, the Gene IDs is refer to TAIR gene identifiers from The Arabidopsis Information Resource (TAIR). The mgeneSim function can be used for large-scale analysis, such as gene clustering. For example:

```
> sim <- mgeneSim(c("835", "5261", "241", "934"), ont = "MF", organism = "human",
+     measure = "Wang")
> sim
```

```
       835  5261   241
835  1.000 0.217 0.620
5261 0.217 1.000 0.191
241  0.620 0.191 1.000
```

```
> library(cluster)
> pamCluster <- pam(as.dist(1 - sim[complete.cases(sim), complete.cases(sim)]),
+     2)
> pamCluster$clustering
```

```
 835 5261  241
   1    2    1
```

Two functions have also been implemented to estimate similarities among gene clusters. The *clusterSim* function is developed for calculating semantic similarity between two gene clusters. The *clusterSim* function first calculate all pair-wise similarities between gene products from different clusters, and then give the average of all these similarities as the similarity of two gene clusters. The *mclusterSim* function is developed for calculating pair-wise similarities of a set of gene clusters, and its kernel algorithm is just the same as the *clusterSim* function.

Gene IDs and species are needed for these functions. For human, rat, mouse, zebrafish, fly, worm and e coli strain 12, the Gene IDs is refer to Entrez Gene IDs, while for yeast, the Gene IDs is refer to ORF identifiers from Saccharomyces Genome Database (SGD), and for arabidopsis, the Gene IDs is refer to TAIR gene identifiers from The Arabidopsis Information Resource (TAIR). For example:

```
> cluster1 <- c("835", "5261", "241", "994", "514", "517", "533")
> cluster2 <- c("578", "582", "583", "400", "409", "411")
> cluster3 <- c("307", "308", "317", "321", "506", "540", "378",
+     "388", "396")
> clusters <- list(a = cluster1, b = cluster2, c = cluster3)
> clusterSim(cluster1, cluster2, ont = "MF", organism = "yeast",
+     measure = "Wang")
```

```
[1] NA
```

```
> clusters <- list(a = cluster1, b = cluster2, c = cluster3)
> mclusterSim(clusters, ont = "MF", organism = "yeast", measure = "Wang")
```

```
[1] NA
```

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. URL http://www.citebase.org/abstract?id=oai:arXiv.org:cmp-lg/9709008.

Dekang Lin. An Information-Theoretic definition of similarity. *In Proceedings of the 15th International Conference on Machine Learning*, pages 296—304, 1998. doi: 10.1.1.55.1832. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.55.1832.

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. URL `http://nzdl.sadl.uleth.ca/cgi-bin/library?e=d-00000-00--off-0jair-00-0-0-10-0--0--0prompt-10--4-----0-1l-11-en-50--20-about--00-0-1-00-0-0-11-1-0utfZz-8-00&cl=CL3.1.11&d=jair-514&x=1`.

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. ISSN 1471-2105. doi: 1471-2105-7-302. 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. ISSN 1460-2059. doi: btm087. URL `http://www.ncbi.nlm.nih.gov/pubmed/17344234`. PMID: 17344234.