GO-terms Semantic Similarity Measures

Guangchuang Yu

Jinan University, Guangzhou, China

February 3, 2012

```
> library(GOSemSim)
> library(org.Hs.eg.db)
> library(GO.db)
```

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 [Yu et al., 2010] 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], and 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 many species. We used Bioconductor package org.At.tair.db, org.Ag.eg.db, org.Bt.eg.db, org.Cf.eg.db, org.Gg.eg.db, org.Pt.eg.db, org.Sco.eg.db, org.EcK12.eg.db, org.EcSakai.eg.db, org.Dm.eg.db, org.Hs.eg.db, org.Pf.plasmo.db, org.Mm.eg.db, org.Ss.eg.db, org.Rn.eg.db, org.Mmu.eg.db, org.Ce.eg.db, org.Xl.eg.db, org.Sc.sgd.db and org.Dr.eg.db to calculate the information content of Arabidopsis, Anopheles, Bovine, Canine, Chicken, Chimp, Coelicolor, E coli strain K12 and strain Sakai, Fly, Human, Malaria, Mouse, Pig, Rat, Rhesus, Worm, Xenopus, Yeast

and Zebrafish 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. We implemented four methods which called *max*, *average*, *rcmax*, and *rcmax.avg* to combine semantic similarity scores of multiple GO terms. The similarities among gene products and gene clusters which annotated by multiple GO terms were also calculated by the same combine methods mentioned above.

Given two GO terms sets $GO_1 = \{go_{11}, go_{12} \cdots go_{1m}\}$ and $GO_2 = \{go_{21}, go_{22} \cdots go_{2n}\}$, method *max* calculate the maximum semantic similarity score over all pairs of GO terms between these two sets, method *average* calcuate the average semantic similarity score over all pairs of GO terms.

Similarities between GO terms form a matrix, and method *rcmax* use the maximum of RowScore and ColumnScore as the similarity, where RowScore (or ColumnScore) is the average of maximum similarities on each row (or column).

And method *rcmax.avg* calculate the average of all maximum similarities on each row and column, and defined as:

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

3 Functions

A *Params* stores a set of parameters for measuring semantic similarity. *Params* containing parameters are ontology, organism, method, combine, and dropCodes. Parameter ontology specify which ontology were used in measurement, organism specifiy which GO Map were loaded for mapping Gene IDs to GO terms, dropCodes restrict evident codes when mapping Gene IDs to GO Terms, method specify which method to be used to measure the similarity and combine sepcify which combine method was used to combining semantic similarity scores.

```
> params <- new("Params", ontology="MF", organism="human", method="Wang")
```

A *GOSet* stores two set of GO IDs.

```
> go1 <- c("GO:0004022", "GO:0004024", "GO:0004023")
> go2 <- c("GO:0009055", "GO:0020037")
> gos <- new("GOSet", GOSet1=go1, GOSet2=go2)
```

A *GeneSet* containing two set of Gene IDs.

```
> gs1 <- c("835", "5261","241", "994", "514", "533")
> gs2 <- c("578","582", "400", "409", "411")
> gs <- new("GeneSet", GeneSet1=gs1, GeneSet2=gs2)
```

A *GeneClusterSet* containing a list of gene clusters.

```
> x <- org.Hs.egGO
> hsEG <- mappedkeys(x)
> set.seed <- 123
> clusters <- list(a=sample(hsEG, 20), b=sample(hsEG, 20), c=sample(hsEG, 20))
> geneClusters <- new("GeneClusterSet", GeneClusters=clusters)
```

Function *sim* was designed to measuring semantic similarity among *GOSet*, *GeneSet* and *GeneClusterSet*.

```
> sim(gos,params)
```

```
           GO:0009055 GO:0020037
GO:0004022      0.318      0.151
GO:0004024      0.290      0.136
GO:0004023      0.290      0.136
```

```
> setCombineMethod(params)<-"rcmax.avg"
> sim(gos,params)
```

```
[1] 0.273
```

```
> sim(gs, params)
```

```
[1] "loading GOMap..."
[1] "Done..."
       578   582   400   409   411
835  0.736 1.000 0.471 0.726 0.576
5261 0.591 0.842 0.527 0.606 0.376
241  0.744 1.000 0.533 0.700 0.382
994  0.467 0.647 0.377 0.555 0.665
514  0.292 0.390 0.421 0.277 0.603
533  0.382 0.481 0.354 0.359 0.406
```

```
> sim(geneClusters, params)
```

```
      a     b     c
a 1.000 0.762 0.736
b 0.762 1.000 0.772
c 0.736 0.772 1.000
```

The old function calls which maybe more user friendly were also supported. They are wrapper functions of function *sim* mentioning above.

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

```
[1] 0.252
```

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

```
[1] 0.299
```

```
> geneSim("241", "251", ont="MF", organism="human", measure="Wang", combine="rcmax.avg")
```

```
$geneSim
[1] 0.368

$GO1
[1] "GO:0005515" "GO:0004051" "GO:0008047" "GO:0019899"
[5] "GO:0042803" "GO:0046982" "GO:0047485" "GO:0050544"

$GO2
[1] "GO:0004035" "GO:0046872" "GO:0016787"
```

```
> mgeneSim(genes=c("835", "5261","241", "994"), ont="MF", organism="human", measure="Wang")
```

```
       835  5261   241   994
835  1.000 0.584 0.666 0.666
5261 0.584 1.000 0.562 0.435
241  0.666 0.562 1.000 0.466
994  0.666 0.435 0.466 1.000
```

```
> clusterSim(gs1, gs2, ont="MF", organism="human", measure="Wang", combine="rcmax.avg")

[1] 0.737

> mclusterSim(clusters, ont="MF", organism="human", measure="Wang", combine="rcmax.avg")

      a     b     c
a 1.000 0.762 0.736
b 0.762 1.000 0.772
c 0.736 0.772 1.000
```

Session Information

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

```
R version 2.14.1 (2011-12-22)
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] GO.db_2.6.1           org.Hs.eg.db_2.6.4
[3] GOSemSim_1.12.1       AnnotationDbi_1.16.11
[5] Biobase_2.14.0        DOSE_1.0.0
[7] RSQLite_0.11.1        DBI_0.2-5

loaded via a namespace (and not attached):
 [1] DO.db_2.3.0             IRanges_1.12.5
 [3] org.Ag.eg.db_2.6.4      org.At.tair.db_2.6.4
 [5] org.Bt.eg.db_2.6.4      org.Ce.eg.db_2.6.4
 [7] org.Cf.eg.db_2.6.4      org.Dm.eg.db_2.6.4
 [9] org.Dr.eg.db_2.6.4      org.EcK12.eg.db_2.6.4
[11] org.EcSakai.eg.db_2.6.4 org.Gg.eg.db_2.6.4
[13] org.Mm.eg.db_2.6.4      org.Mmu.eg.db_2.6.4
[15] org.Pf.plasmo.db_2.6.4  org.Pt.eg.db_2.6.4
[17] org.Rn.eg.db_2.6.4      org.Sc.sgd.db_2.6.4
```

```
[19] org.Sco.eg.db_2.4.2     org.Ss.eg.db_2.6.4
[21] org.Xl.eg.db_2.6.4      plyr_1.7.1
[23] qvalue_1.28.0           tcltk_2.14.1
[25] tools_2.14.1
```

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-abCL3.1.11&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.

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. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq064. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/26/7/976. PMID: 20179076.