GO-terms Semantic Similarity Measures

Guangchuang Yu

March 11, 2009

1 Introduction

Functional similarity of gene products can estimated by controlled biological vocabularies, such as Gene Ontology (GO). GO comprises of three orthogonal ontologies, molecular function (MF), biological process (BP), and cellular component (CC).
Four methods proposed by Resnik[Philip, 1999], Jiang[Jiang and Conrath, 1997], Lin[Lin, 1998] and Schlicker[Schlicker et al., 2006] respectively have presented to determine the semantic similarity of two GO terms based on the annotation statistics of their common ancestor terms. 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 strengths and weaknesses. The *GOSemSim* package implemented all these five methods.

2 Semantic Similarity Measures

The *GOSemSim* package contains functions to estimate graph structure based similarity scores of GO terms. Details about Wang's method can be seen in [Wang et al., 2007], details about Rel method can be seen in [Schlicker et al., 2006] and the details about Resnik, Lin, and Jiang's methods can be seen in [Lord et al., 2003]. Resnik, Lin, Rel, and Jiang's methods based on the information content of the GO terms while 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 method proposed by [Wang et al., 2007] is based on the graph structure of each term.

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 of two term's semantics, we 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, 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 contribute to its own as one. After obtaining the S-values for all terms in DAG_A, we calculate the semantic value of GO term A, SV(A), 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 semantic similarity of one GO term go and a GO terms set $GO = \{go_1, go_2 \dots go_k\}$ is defined as:

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

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

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

The *GOSemSim* package contains functions to estimate graph structure based similarity scores of GO terms. Details about this method can be seen in [Wang et al., 2007]. This 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 GOSemSim package implemented four other methods which are based on information content were proposed by Resnik[Philip, 1999], Jiang[Jiang and Conrath, 1997], Lin[Lin, 1998] and Schlicker[Schlicker et al., 2006] respectively.

Information content is defined as frequency of each term occurs in the corpus. We used Bioconductor package org.Hs.eg.db, org.Dm.eg.db, org.Mm.eg.db, org.Rn.eg.db, org.Sc.sgd.db to calculate the information content of human, fly, mouse, rat and yeast species respectively. The information content will update regularly.

Given the information content, we applied the four measures to estimate the semantic similarity between terms.

As GO allows 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. We defined p_{ms} 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 first method Resnik[Philip, 1999] is defined as:

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

The second method Lin[Lin, 1998] is defined as:

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

The third method Rel[Schlicker et al., 2006] combine Resnik's and Lin's method is defined as:

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

The last method Jiang[Jiang and Conrath, 1997] define a semantic distance as:

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

and the corresponding similarity measure for d(t1, t2) is given by:

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

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

[1] 0.252
```

The function goSim generates one score for a pair of GO terms.

```
> 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 mgoSim generates the similarity score of two GO terms lists.

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

$geneSim
[1] 0.29

$GO1
[1] "GO:0005488" "GO:0008047"

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

The function geneSim estimate two genes's semantic similarity. The mapping from Gene IDs to GO IDs can be restricted based on evidence codes. It supports five species, which are "human", "rat", "mouse", "fly", and "yeast".

3 Functional Clustering

Given GO based similarity scores, gene products may be clustered by their function. *GOSemSim* package provides a function, mgeneSim, that returns pairwise similarities scores for a list of genes. It can be used by other functions to perform clustering.

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

       835  5261   241
835  1.000 0.227 0.599
5261 0.227 1.000 0.309
241  0.599 0.309 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
```

We also implemented two functions for estimating similarities among gene clusters. *clusterSim* for calculating semantic similarity between two gene clusters and *mclusterSim* for calculating pairwise similarities of a set of gene clusters. For caculate two gene clusters similarities, we first calculate pairwise similarities among genes, and the maximum similarity score was taken. This strategy is based on WordNet[Fellbaum, 1998].

```
> cluster1 <- c("snR67", "snR40", "snR48", "snR17a", "snR8")
> cluster2 <- c("YOR251C", "YPR137C-B", "YPR010C", "YPR072W")
> cluster3 <- c("YNL133C", "YOL041C", "YOL018C", "YOR236W", "YOR179C",
+     "YOR230W")
> clusterSim(cluster1, cluster2, ont = "MF", organism = "yeast",
+     measure = "Wang")

[1] 0.505

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

```
      a     b c
a 1.000 0.505 1
b 0.505 1.000 1
c 1.000 1.000 1
```

References

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. The MIT Press, illustrated edition edition, May 1998. ISBN 026206197X.

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 seman-

tic 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.