GO-terms Semantic Similarity Measures

Guangchuang Yu

February 20, 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.

```
> 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. The mapping from Entrez Gene IDs to GO IDs can be restricted based on evidence codes.

```
> geneSim("241", "2561", ont = "MF", drop = "IEA", 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.

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 similarity 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", measure = "Wang")
> sim

       835  5261   241
835  1.000 0.237 0.599
5261 0.237 1.000 0.339
241  0.599 0.339 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
```

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.