Gene Ontology-based Semantic Similarity Measures

Xiang Guo

April 30, 2008

1 Introduction

Quantitative measure of functional similarity between gene products is important for post-genomics study. The similarity measures may be used to validate high-throughput protein interaction data, help the development of new pathway modelling tools and clustering methods, and enable the identification of functionally related gene products independent of homology [Guo et al., 2006, Schlicker et al., 2006].

Functional relationship of gene products is usually estimated by their shared annotation in a controlled vocabulary system, such as Gene Ontology (GO). GO comprises of three orthogonal ontologies, molecular function (MF), biological process (BP), and cellular component (CC). GO terms and their relationships are represented in the form of directed acyclic graphs (DAG). GO-based semantic similarity measures can be classified into two categories. The first category defines semantic similarity based on the graph structure of GO. For each gene product, we can obtain an induced graph which includes the specific set of GO annotations for the gene product and all parents of those GO terms. Union-intersection (UI) method estimates semantic similarity of two gene products by the number of nodes two induced graphs share divided by the total number of nodes in two graphs. Another method, longest shared path approach (LP), uses the depth of the longest path shared by two induced graphs as the similarity score. These two methods have been implemented in a Bioconductor package GOstats.

The other category of semantic similarity measures is based on the concept of information content that is defined as the frequency of each GO term, or any of its children, occurring in an annotated data set. Semantic similarity of gene products is estimated by the information content of specific GO annotations and their shared parents. The assumption is that the more information two terms share, the more similar they are. The shared information of two terms is indicated by the information content of the terms

that subsume them in DAG. Given the information content of each term, there are several ways to calculate similarity scores between annotated gene products.

2 Semantic Similarity Measures

The *SemSim* package contains functions to estimate information content-based similarity scores of GO terms and gene products. Four semantic similarity measures have been implemented with different strategies for the combination of multiple similarity scores in the case of multiple annotations for a gene product. Details about *Resnik*, *Lin*, and *Jiang*'s methods can be seen in Lord et al. [2003], while the *Relevance* method is described in Schlicker et al. [2006]. The information content of each GO term has been precomputed based on annotations available in GO Consortium. Species specific estimation may be chosen if the study is limited in one species. If gene products are annotated with multiple GO terms, maximum or average GO term similarity may be taken as the gene similarity. Alternatively, different scores may be calculated for two directional comparisons of gene products A and B. Given a matrix containing pairwise similarity values of GO terms for gene product A and B, the maximum values in the rows and columns represent the best hits for the comparison of A with B and the comparison of B with A respectively. The averages over the row maxima and the column maxima give scores for two directions, and they are combined to generate one gene similarity score. This approach provides a robust measure for the gene products with incomplete annotation [Schlicker et al., 2006].

```
> library(SemSim)
> geneSim("852695", "5261", ont = "BP")

$Sim
[1] 0.307

$GO1
[1] "GO:0006914" "GO:0016050"

$GO2
[1] "GO:0006091" "GO:0006468" "GO:0008150"
```

Each GO category generates one score for a pair of gene products, and they may be combined to form a single functional similarity score. Taking

the sum or average of two or three scores does not distinguish the cases where two gene products have average scores in all ontologies or they have one high score in one ontology and low scores in the other. Squaring individual score gives higher similarity for the latter case than the former, which is desirable. Sophisticated machine learning and statistical methods may also be used to integrate GO similarity scores as well as other genomic features for the study of systems biology.

Besides entrezgene identifiers, GO annotation may be retrieved for other identifiers such as affy ids, RefSeq and Ensembl ids. Then, function *mtermSim* can be called to get semantic similarity scores.

```
> library(hgu133plus2)
> go1 <- sapply(hgu133plus2GO[["203140_at"]], function(x) x$GOID)
> go2 <- sapply(hgu133plus2GO[["208368_s_at"]], function(x) x$GOID)
> mtermSim(go1, go2, ont = "BP")
```

```
[1] 7.026
```

3 Functional Clustering and Validation

Given GO-based similarity scores, gene products may be clustered by their function to identify annotation patterns in the data set. Wolting et al. [2006] have demonstrated the feasibility of this methodology using two protein array data sets. One data set includes proteins that bind small molecule inhibitors of rapamyclin. Clustering by annotation reveals subsets of proteins that may help to elucidate how rapamycin affects cell growth. *SemSim* package provides a function, *mgeneSim*, that returns pairwise similarity scores for a list of genes with GO annotations available. It can be used with other functions to perform functional clustering analysis.

```
> library(cluster)
> data(Schreiber)
> sim <- mgeneSim(Schreiber, ont = "CC", measure = "Lin", multiple = "rcmax")
> sim[1:5, 1:5]
```

```
       852215 852291 852369 852514 851334
852215  1.000  0.539  0.539      0  0.447
852291  0.539  1.000  1.000      0  0.444
852369  0.539  1.000  1.000      0  0.313
852514  0.000  0.000  0.000     NA  0.000
851334  0.447  0.444  0.313      0  1.000
```

```
> pamClust <- pam(as.dist(1 - sim[complete.cases(sim), complete.cases(sim)]),
+     9)
> pamClust$clustering

852215 852291 852369 851334 851567 851746 856926 850502 852817 852919 852972
     1      2      2      3      4      2      4      2      1      1      4
856445 856453 856471 856529 854692 854855 853462 850675 850825 854987 855240
     4      5      6      7      1      7      8      2      4      4      3
855669 855659 855625 855544 855797 854065 854058 854052 854042 856210
     1      1      9      1      9      3      1      9      4      1

> pamClust$silinfo$avg.width

[1] 0.6293889
```

The high average silhouette width validates the quality of GO-based clustering results. Moreover, validity indices calculated by GO similarity scores may be applied to assess data-driven clustering results [Bolshakova et al., 2005]. Suppose a set of genes have been clustered by their gene expression measurements. We can then use *mgeneSim* in *SemSim* and *silhouette* in *cluster* package to estimate the GO-based silhouette index. This knowledge-driven approach facilitates biological assessment of data mining results.

References

N. Bolshakova, F. Azuaje, and P. Cunningham. A knowledge-driven approach to cluster validity assessment. *Bioinformatics*, 21:2546–2547, 2005.

X. Guo, R. Liu, C.D. Shriver, H. Hu, and M.N. Liebman. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, 22:967–973, 2006.

P. Lord, R. Stevens, A. Brass, and C. Goble. Semantic similarity measures as tools for exploring the Gene Ontology. In *Pacific Symposium on Biocomputing*, volume 8, pages 601–612, 2003.

A. Schlicker, F. Domingues, J. Rahnenfuhrer, and T. Lengauer. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7:302, 2006.

C. Wolting, C.J. McGlade, and D. Tritchler. Cluster analysis of protein array results via similarity of gene ontology annotation. *BMC Bioinformatics*, 7:338, 2006.