GO-based Semantic Similarity Measure

The GO terms are catagorised to represent three most general biological concepts : biological processes, molecular functions, and cellular components. The GO database provides annotations for each GO terms. A set of GO terms are used to describe properly the functionality of a protein [1]. As a given gene or protein can perform different biological processes or functions in different environments so each gene or protein can be associated with, or annotated to, one or more GO-term(s). Measuring semantic similarity between concepts in a taxonomy is a common practice in natural language processing and is characterised as structure of the taxonomy or information contents of the concepts. These techniques can be extended to measure the degree of similarity between terms in the GO structure also. The semantic similarity measured between two GO terms can be directly converted to a measurement of the similarity between two proteins.

There are mainly two approaches for measuring semantic similarity in a reference gene set : Graph Structure based (GS) and Information Content based (IC). Graph structure based method consider hierarchical structure of GO in computing semantic similarity whereas information content based method prioritize the *a priory* probabilities or information content in a given gene set.

Czekanowski-Dice similarity [2] is a GS-based method in computing semantic similarity and is defined as : $1 - d(G_1, G_2)$, where distance of genes G_1 and G_2 is defined as :

$$d(G_1, G_2) = \frac{\#(GO(G_1) \bigtriangleup (G_2))}{\#(GO(G_1) \bigcup GO(G_2)) + \#(GO(G_1) \bigcap GO(G_2))}, \tag{1}$$

where $\bigtriangleup$ is the symmetric set difference, $\#$ is the number of elements in a set and $GO(G_i)$ is the set of GO annotations for gene G_i.

The information content of a GO term is computed by the frequency of the term occurring in annotations; a rarely used term contains a greater amount of information. Probability for observing a term t is defined as $p(t) = \frac{freq(t)}{MaxFreq(t)}$, where MaxFreq is the maximum frequency of all terms [3]. The information content for a term t is given as $IC(t) = -log_2 p(t)$. [4] introduced several related similarity metrics that are based on the most informative common ancestor (MICA) of two GO terms. Resnik proposed a semantic similarity measure between two terms $t1$ and $t2$ and is defined as

$$Sim_{resnik}(t1, t2) = IC(A), \tag{2}$$

where A is the most informative common ancestor of $t1$ and $t2$, i.e., A is a term that is an ancestor of both $t1$ and $t2$ and has the maximum IC among common ancestors $CA(t1, t2)$ of the terms. According to Lin [5] the semantic similarity between terms $t1$ and $t2$, is defined as :

$$Sim_{Lin}(t1, t2) = \frac{2IC(A)}{IC(t1) + IC(t2)}, \tag{3}$$

Jiang and Conrath [6] defined a similarity metric as :

$$Sim_{jc}(t1, t2) = \frac{1}{1 + d_{jc}(t1, t2)}, \tag{4}$$

where the semantic distance metric is $d_{jc}(t1, t2) = IC(t1) + IC(t2) - 2IC(A)$.

The Relevance measure [7] that combines Lin's and Resnik's measures is defined as :

$$Sim_{Rel}(t1, t2) = \max_{t \epsilon CA(t1,t)} \frac{2logp(t)(1 - p(t))}{logp(t1) + logp(t)} = \frac{IC(A)(1 - p(a))}{IC(t1) + IC(t2)}. \tag{5}$$

Kappa statistics, a chance-corrected measure of co-occurrence between two sets of categorized data, can be adopted to statistically measure the annotation co-occurrence of any given gene pairs [8, 9]. In Kappa statistics [10] each gene is represented as a binary vector $(g_1 \quad g_2 \quad g_3 \quad \ldots \quad g_N)$, where g_i is 1 if the gene is annotated with the GO term g_i and 0 otherwise. N is the total number of GO terms under consideration. Similarity of genes $G1$ and $G2$ is defined as

$$K_{G1,G2} = \frac{O_{G1,G2} - A_{G1,G2}}{1 - AG1, G2}, \tag{6}$$

where $O_{G1,G2}$ represents observed co-occurrence of GO terms and $A_{G1,G2}$ represents random co-occurrence and $K_{G1,G2}$ is the kappa value representing the degree of annotation co-occurrence between genes $G1$ and $G2$.

The MICA-based measures can be modified by computing the disjunctive ancestor terms [3]. Two ancestors $ansc_1$ and $ansc_2$ of a term t are called disjunctive if there exists independent paths from $ansc_1$ to t and from $ansc_2$ to t. In the GraSM enhancement, when computing the similarity between two terms $t1$ and $t2$ all common disjunctive ancestors of terms $t1$ and $t2$ are considered [3]. GraSM modifies the computation of IC(A) and can be applied to the Resnik, Lin and Jiang-Conrath measures.

References

[1] A. Mukhopadhyay, M. De, and U. Maulik, “Selection of go-based semantic similarity measures through amde for predicting protein-protein interactions,” in *In Proc. Int Conf. SEMCCO*, (Visakhapatnam, India,), pp. 55–62,, December 2011.

[2] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guenoche, and B. Jacq, “Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network.,” *GENOME BIOLOGY*, vol. 5, 2004,.

[3] Couto.FM, Silva.MJ, and Coutinho.PM, “Measuring semantic similarity between gene ontology terms,” *Data Knowl Eng*, vol. 61, pp. 137–152, 2007.

[4] P. Lord, R. Stevens, A. Brass, and C. Goble, “Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation,” *Bioinformatics*, vol. 19, no. 10, pp. 1275–1283, 2003.

[5] D. Lin, “An information-theoretic definition of similarity,” in *Proc. 15th International Conference on Machine Learning*, vol. San Francisco, CA, pp. 296–304, 1998.

[6] J. J. Jiang and D. W. Conrath, “Semantic similarity based on corpus statistics and lexical taxonomy,” in *Proc. International Conference Research on Computational Linguistics*, vol. Taiwan, 1997.

[7] A. Schlicker, F. Domingues, J. Rahnenfhrer, and T. Lengauer, “A new measure for functional similarity of gene products based on gene ontology.,” *BMC Bioinformatics*, vol. 7, p. 302, 2006.

[8] J. Cohen, “A coefficient of agreement for nominal scales,” *Educ Psychol Meas*, vol. 20, pp. 37–46, 1960.

[9] T. Byrt, J. Bishop, and J. Carlin, “Bias, prevalence and kappa.,” vol. 46, pp. 423–429, 1993.

[10] D. Huang, B. Sherman, Q. Tan, J. Collins, W. Alvord, J. Roayaei, R. Stephens, M. Baseler, H.Lane, and R. Lempicki, “The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists,” *Genome Biology*, vol. 8, no. 9, p. R183, 2007.