Method proposed by Wang

Guangchuang Yu

February 20, 2009

The method described here which was 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, we 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 $GO = \{go_1, go_2 \ldots go_k\}$ is defined as:

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

Therefore, given two GO terms sets $GO_1 = \{go_{11}, go_{12} \ldots go_{1m}\}$ and $GO_2 = \{go_{21}, go_{22} \ldots go_{2n}\}$, 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.

References

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.