

UFO: a tool for unifying biomedical ontology-based semantic similarity calculation, enrichment analysis and visualization

Duc-Hau Le^{1,2,*}

¹Department of Computational Biomedicine, Vingroup Big Data Institute, Hanoi, Vietnam.

²School of Computer Science and Engineering, Thuyloi University, Hanoi, Vietnam.

* To whom correspondence should be addressed. Tel: +84 912 324564; Email: hauldhut@gmail.com

Supplementary File 1

Semantic Similarity Measures

I.	Information content	2
II.	Between-term measures.....	3
A.	Node-based measures.....	3
B.	Edge-based measures.....	5
C.	Hybrid-based measures.....	6
III.	Between-entity measures	7
A.	Pairwise.....	7
B.	Groupwise	8
IV.	References	10

Biomedical ontologies are represented in a directed acyclic graph (DAG). In this section, we introduce eleven implemented between-term semantic similarity measures (including eight node-based, two edge-based and one hybrid-based) and eleven between-entity measures (including four pairwise-based and seven group-wise-based ones). First, we introduce information content (IC) of a term. Then, we define semantic similarity measures between terms and between entities.

I. INFORMATION CONTENT

The IC of a term is calculated based on a corpus, (*i.e.*, an annotation database of HPO (Köhler, et al., 2014)) as following:

$$IC_t = -\log(p_t)$$

where p_t is the probability of a term occurred in a given corpus (Lord, et al., 2003):

$$p_t = \frac{f_t}{t_{root}}$$

where

$$f_t = Annot_t + \sum_{c \in Children_t} f_c$$

where $Annot_t$ is number of proteins annotated with term t in a corpus and $Children_t$ is the set of children of term t in given ontology graph.

Figure 1 illustrates Information Content (IC) calculation for Gene Ontology (GO) term using Gene Ontology and Annotation databases.

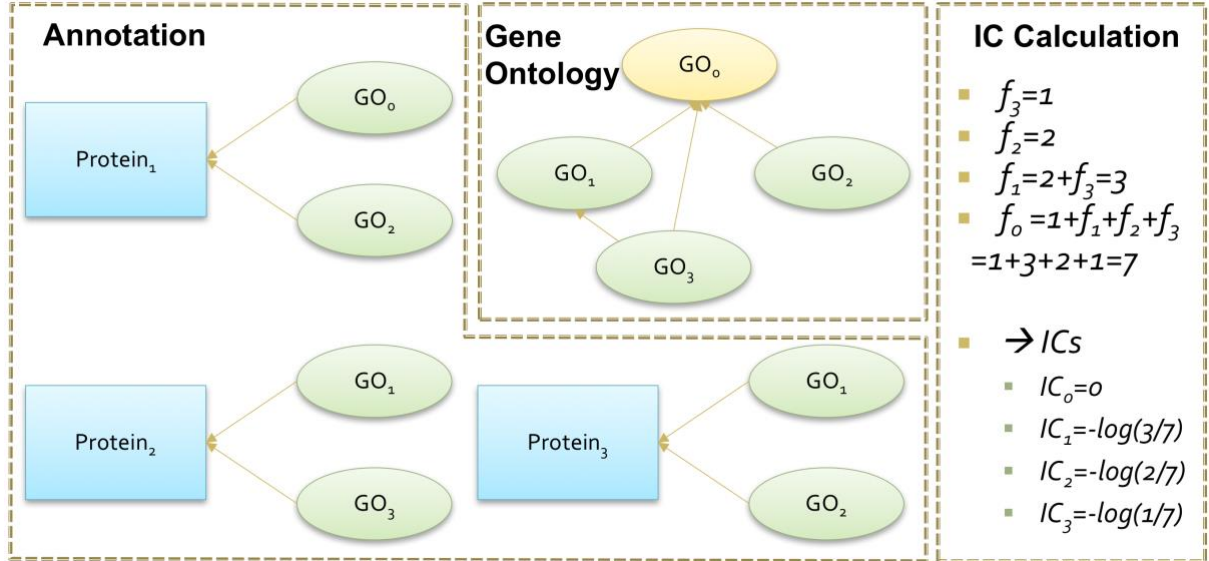


Figure 1: Illustration of Information Content (IC) Calculation for Gene Ontology term using Gene Ontology and Annotation databases.

In the UFO tool, we visualize IC of terms by color. Figure 2 visualizes ICs of some HPO terms by their color (the higher IC of the term is the redder of node is)

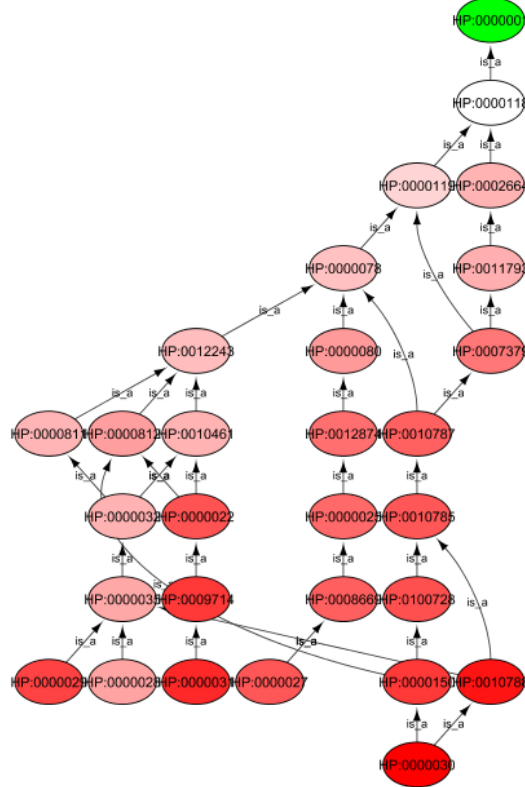


Figure 2: ICs of HPO terms are visualized by red color (blue node is root).

II. BETWEEN-TERM MEASURES

The ontology data is represented in directed acyclic graph, in which each term is located in a node and the relations between terms are represented by edge connecting nodes. Therefore, calculation for semantic similarity between terms are categorized into two main methods: i) node-based and ii) edge-based.

A. Node-based measures

In this section, we firstly introduce four node-based measures for similarity between terms. They are all based on most informative common ancestors (MICA) of the terms, which is defined as following:

$$IC_{MICA} = \max_{c \in SP(t_1, t_2)} (IC_c)$$

where $SP(t_1, t_2)$ are shared ancestors of terms t_1 and t_2 .

Figure 3 shows shared ancestors and MICA term of two HPO terms (HP:0000008 and HP:0000010)

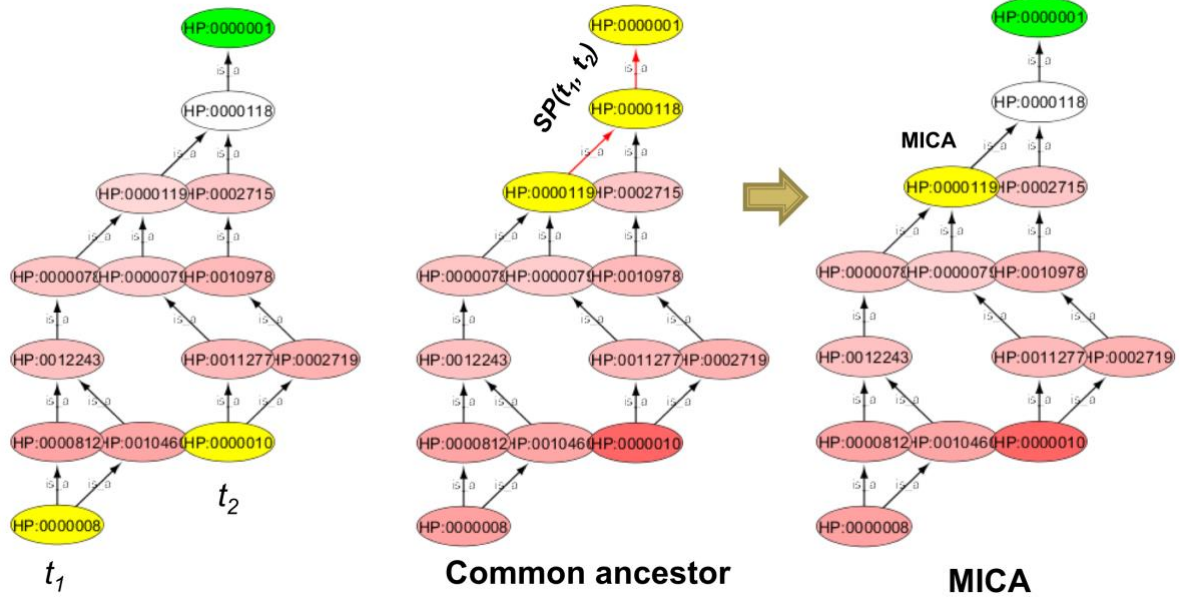


Figure 3: Shared/Common ancestors: Shared ancestors (HP:0000001, HP:0000118 and HP:0000119) and MICA term (HP:0000119) of two HPO terms (HP:0000008 and HP:0000010)

First, Resnik (Resnik, 1995) defined the similarity between two terms as following:

$$IC_t = -\log(p_t)$$

$$SimT_{Resnik}(t_1, t_2) = IC_{MICA}$$

Second, other between-term similarity measures proposed by (Lin, 1998), (Jiang and Conrath, 1997) and (Schlicker, et al., 2006) were defined respectively as following:

$$SimT_{Lin}(t_1, t_2) = \frac{2 \times IC_{MICA}}{IC_{t_1} + IC_{t_2}}$$

$$SimT_{JC}(t_1, t_2) = \frac{1}{IC_{t_1} + IC_{t_2} - 2 \times IC_{MICA} + 1}$$

$$SimT_{Rel}(t_1, t_2) = \frac{2 \times IC_{MICA} \times (1 - P_{MICA})}{IC_{t_1} + IC_{t_2}}$$

$$\text{where } P_{MICA} = 10^{-IC_{MICA}}$$

Similarly to the four between-term similarity measures, (Couto, et al., 2005) defined common disjunctive ancestors (CDA) of terms t_1 and t_2 to replace MICA in the four measures as following:

$$IC_{CDA} = \frac{\sum_{t \in CDA} IC_t}{|CDA|}$$

where CDA contain common disjunctive ancestors of terms t_1 and t_2

Figure 4 shows CDA terms (HP:0000118 and HP:0000119) of two HPO terms (HP:0000008 and HP:0000010)

Therefore, four more between-term similarity measures were defined: $SimT_{ResnikGraSM}$, $SimT_{LinGraSM}$, $SimT_{JCGraSM}$ and $SimT_{RelGraSM}$. In summary, a total of eight node-based between-term similarity measures were used in our study.

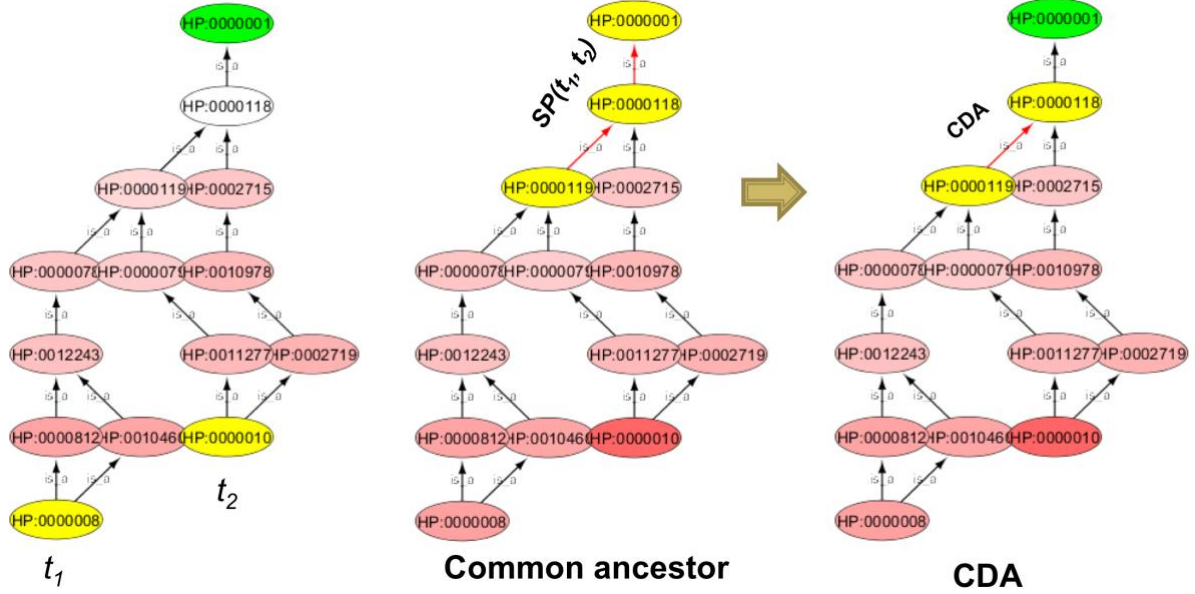


Figure 4: Shared/Common ancestors: Shared ancestors (HP:0000001, HP:0000118 and HP:0000119) and CDA terms (HP:0000118 and HP:0000119) of two HPO terms (HP:0000008 and HP:0000010)

B. Edge-based measures

In this section, we introduced two widely used edge-based between-term semantic similarity measures. The first one is simply defined as length of longest path from lowest common ancestor (LCA) to root, $SimT_{Wu}(t_1, t_2) = (LCA, root)$ (Wu, et al., 2005). Meanwhile, the second one (Yu, et al., 2005) additionally takes consider on length of longest path between each term to LCA and formally defined as following:

$$SimT_{Yu}(t_1, t_2) = \frac{L(LCA, root)}{L(LCA, root) + L(t_1, LCA) + L(t_2, LCA)}$$

where $L(t_1, t_2)$ length of the longest path between two terms t_1 and t_2

Figure 5 shows the longest path of two HPO terms (HP:0000008 and HP:0000010).

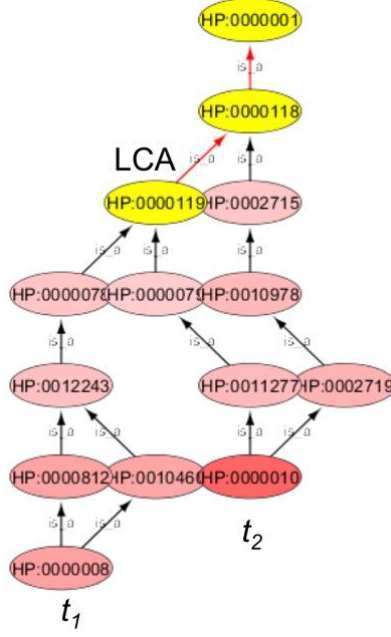


Figure 5: The longest path (red line) of two HPO terms (HP:0000008 and HP:0000010).

C. Hybrid-based measures

We implemented one hybrid measure, which was introduced in (Wang, et al., 2007). In the study, an ontology term t is represented as $DAG_t = (t, T_t, E_t)$ where T_t is the set of ontology terms in DAG_t , including term t and all of its ancestor terms in the ontology graph, and E_t is the set of edges connecting the ontology terms in DAG_t . To quantitatively compare two ontology terms, a semantic value of term t is defined as the aggregate contribution of all terms in DAG_t to the semantics of term t as following:

$$SV(t) = \sum_{t_i \in T_t} S_t(t_i)$$

where

$$S_t(t_i) = \begin{cases} S_t(t) = 1 \\ S_t(t_i) = \max\{w_e * S_t(c) | c \in children_{t_i}\} \text{ if } t_i \neq t \end{cases}$$

and $0 < w_e < 1$

Finally, a semantic similarity between two terms, t_1 and t_2 , is defined as follow:

$$SimT_{Hybrid}(t_1, t_2) = \frac{\sum_{t_i \in T_{t_1} \cap T_{t_2}} (S_{t_1}(t_i) + S_{t_2}(t_i))}{SV(t_1) + SV(t_2)}$$

Figure 6 shows DAGs of two HPO terms (HP:0000008 and HP:0000010).

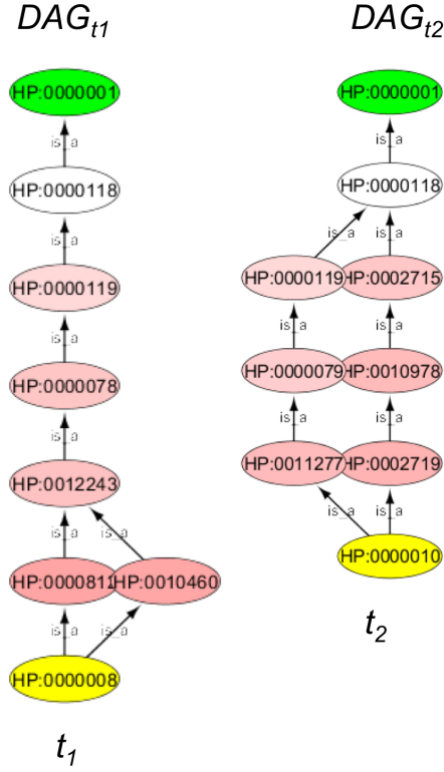


Figure 6: DAGs of two HPO terms (HP:0000008 and HP:0000010).

III. BETWEEN-ENTITY MEASURES

For assessment of similarity between two annotated entities, two main approaches have been proposed, i.e., pairwise and groupwise.

Assuming that two entities e_1 and e_2 are annotated by a set of term $T_1 = \cup_k^m t_{1k}$ and $T_2 = \cup_l^n t_{2l}$. In this section, we are going to introduce between-entity similarity measures

A. Pairwise

This approach calculates the similarity between two entities based on the similarity of every pair terms which annotate to the entities.

First, the similarity of every pair of terms is calculated to generate a similarity matrix as following (see Fig. 7):

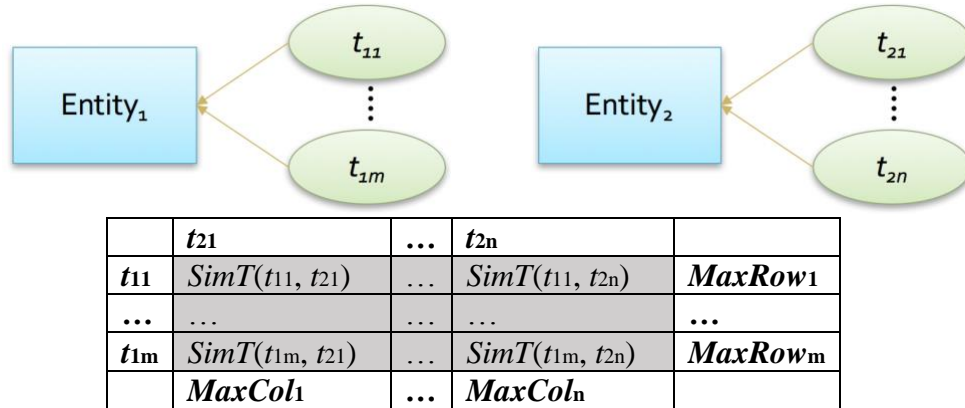


Figure 7. Example of a semantic similarity matrix, where shading cells are elements of the matrix, t_{ij} are column and cell headers, $MaxRow_i$ and $MaxCol_j$ are maximum values of row i and column j , respectively.

Then, four pairwise between-entity were defined as following:

The first two measures were simply either based on average (Lord, et al., 2003) or maximum (Sevilla, et al., 2005) similarity of all pairs.

$$SimE_{Avg} = \text{Avg}_{i=1\dots m; j=1\dots n} (SimT(t_{1i}, t_{2j}))$$

$$Max_{Max} = \text{Max}_{i=1\dots m; j=1\dots n} (SimT(t_{1i}, t_{2j}))$$

Meanwhile, Couto et al (Couto, et al., 2005) and Azuaje et al (Azuaje, et al., 2005) opted for a composite average in which only the best-matching term pairs are considered (best-match average):

$$SimE_{BMA} = \frac{\sum_{i=1\dots m} MaxRow_i + \sum_{j=1\dots n} MaxCol_j}{n_{Row} + n_{Col}}$$

Finally, Schlicker et al (Schlicker, et al., 2006) proposed a variation of the best-match average:

$$SimE_{RCMax} = \max \left(\text{Avg}_{i=0\dots m} (MaxRow_i), \text{Avg}_{j=0\dots n} (MaxCol_j) \right)$$

B. Groupwise

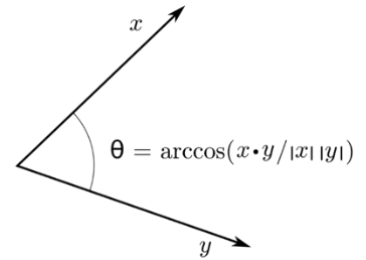
This approach is categorized into two main methods: i) vector-based and ii) graph-based. In which, two popular vector-based between-entity similarity measures Cosine (Huang, et al., 2007) and Kappa (Chaballier, et al., 2007) were defined based on cosine and kappa coefficients. More specifically, T_1 and T_2 were first represented as binary vectors as following:

$$v_1 = (v_{1_1}, \dots, v_{1_N}), v_2 = (v_{2_1}, \dots, v_{2_N})$$

$$v_{1_k}, v_{2_l} = \begin{cases} 1 & \text{if } t_k, t_l \in T \\ 0 & \text{otherwise} \end{cases}$$

where T is whole set of N terms in the ontology database.

Then, the cosine of the two vectors v_1 and v_2 is calculated:



$$SimE_{Cosine}(v_1, v_2) = \frac{\sum_{i=1}^N (v_{1_i} \times v_{2_i})}{\sqrt{\sum_{i=1}^N (v_{1_i} \times v_{1_i})} \times \sqrt{\sum_{i=1}^N (v_{2_i} \times v_{2_i})}}$$

Another vector-based between-entity similarity measure is based on Kappa coefficient. Formally, it is defined as following:

$$Kappa(v_1, v_2) = \frac{P_o - P_e}{1 - P_e}$$

where:

- P_o is observed proportionate agreement: $(P_{11} + P_{00})/P_t$
- P_e is overall probability of random agreement: $P_{e1} + P_{e0}$
- P_{e1} is probability that both of them = 1: $P_{v1=1} * P_{v2=1}$
- $P_{v1=1}$ is probability $v1=1$: $(P_{11} + P_{10})/P_t$
- $P_{v2=1}$ is probability $v2=1$: $(P_{11} + P_{01})/P_t$
- P_{e0} is probability that both of them = 0: $P_{v1=0} * P_{v2=0}$
- $P_{v1=0}$ is probability $v1=0$: $(P_{01} + P_{00})/P_t$
- $P_{v2=0}$ is probability $v2=0$: $(P_{10} + P_{00})/P_t$
- $P_t, P_{11}, P_{00}, P_{10}$ and P_{01} is total number of observations, number of observations where $v_{1_i} = v_{2_i} = 1$, number of observations where $v_{1_i} = v_{2_i} = 0$, number of observations where $v_{1_i} = 1$ and $v_{2_i} = 0$, and number of observations where $v_{1_i} = 0$ and $v_{2_i} = 1$, respectively.

		v_2		
		1	0	
v_1	1	P_{11}	P_{10}	
	0	P_{01}	P_{00}	
				P_t

For graph-based measures, T_1 and T_2 is extended with ancestors of terms in each set. Therefore, they are defined as follows:

$$T_1 = \bigcup_{k \in \mathcal{K}}^m (t_k \cup \text{anc}(t_k))$$

$$T_2 = \bigcup_{l \in \mathcal{L}}^n (t_l \cup \text{anc}(t_l))$$

Five more graph-based between-entity similarity measures were defined as following:

A measure is based on term overlap (Lee, et al., 2004)

$$SimE_{TO} = |T_1 \cap T_2|$$

A normalized version of $SimE_{TO}$ (Mistry and Pavlidis, 2008)

$$SimE_{NTO} = \frac{|T_1 \cap T_2|}{\min(|T_1|, |T_2|)}$$

A measure is based on Jaccard index (Gentleman, 2005; Martin, et al., 2004)

$$SimE_{UI} = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$$

An IC-based weighted version of $SimE_{UI}$ (Pesquita, et al., 2007)

$$SimE_{GIC} = \frac{\sum_{t_k \in |T_1 \cap T_2|} IC_{t_k}}{\sum_{t_l \in |T_1 \cup T_2|} IC_{t_l}}$$

Finally, a longest path-based between-entity similarity measure (Gentleman, 2005):

$$SimE_{LP} = \max_{t_k \in |T_1 \cap T_2|} L(t_k, \text{root})$$

In summary, a total of eleven between-entity similarity measures were used in our study.

IV. REFERENCES

- Azuaje, F., Wang, H. and Bodenreider, O. Ontology-driven similarity approaches to supporting gene functional assessment. *Proc. Of The Eighth Annual Bio-Ontologies Meeting* 2005;Michigan, 25 June.
- Chabalier, J., Mosser, J. and Burgun, A. A transversal approach to predict gene product networks from ontology-based similarity. *BMC Bioinformatics* 2007;8(1):235.
- Couto, F., Silva, M.r. and Coutinho, P. Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. In, *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*. Bremen, Germany: ACM; 2005. p. 343-344.
- Gentleman, R. Visualizing and distances using GO. URL <http://www.bioconductor.org/docs/vignettes.html> 2005;38.
- Huang, D., et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology* 2007;8(9):R183.
- Jiang, J.J. and Conrath, D.W. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In, *International Conference Research on Computational Linguistics (ROCLING X)*. 1997. p. 9008.
- Köhler, S., et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research* 2014;42(D1):D966-D974.
- Lee, H., et al. Coexpression Analysis of Human Genes Across Many Microarray Data Sets. *Genome Research* 2004;14(6):1085-1094.
- Lin, D. An Information-Theoretic Definition of Similarity. In, *Proceedings of the Fifteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc.; 1998.
- Lord, P.W., et al. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 2003;19(10):1275-1283.
- Martin, D., et al. GOTOolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biology* 2004;5(12):R101.
- Mistry, M. and Pavlidis, P. Gene Ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics* 2008;9(1):327.
- Pesquita, C., et al. Evaluating GO-based semantic similarity measures. In, *Proc. 10th Annual Bio-Ontologies Meeting*. 2007. p. 38.
- Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. In, *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1*. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.; 1995.
- Schlicker, A., et al. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* 2006;7(1):302.
- Sevilla, J.L., et al. Correlation between gene expression and GO semantic similarity. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* 2005;2(4):330-338.
- Wang, J.Z., et al. A new method to measure the semantic similarity of GO terms. *Bioinformatics* 2007;23(10):1274-1281.
- Wu, H., et al. Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic acids research* 2005;33(9):2822-2837.
- Yu, H., et al. Broadly predicting specific gene functions with expression similarity and taxonomy similarity. *Gene* 2005;352(0):75-81.