Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 98 (2016) 275 – 282

Procedia Computer Science

The 6th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2016)

Exploiting Biomedical Web Resources: a Case Study

Nicoletta Dessì * , Barbara Pes

Dipartimento di Matematica e Informatica, Università di Cagliari, Via Ospedale 72, 09125 Cagliari, Italy

Abstract

An increasing number of web resources continue to be extensively used by healthcare operators to obtain more accurate diagnostic results. In particular, health care is reaping the benefits of technological advances in genomic for facing the demand of genetic tests that allow a better comprehension of diagnostic results. Within this context, Gene Ontology (GO) is a popular and effective mean for extracting knowledge from a list of genes and evaluating their semantic similarity. This paper investigates about the potential and any limits of GO ontology as support for capturing information about a set of genes which are supposed to play a significant role in a pathological condition. In particular, we present a case study that exploits some biomedical web resources for devising several groups of functionally coherent genes and experiments about the evaluation of their semantic similarity over GO. Due to the GO structure and content, results reveal limitations that not affect the evaluation of the semantic similarity when genes exhibit simple correlations but influence the estimation of the relatedness of genes belonging to complex organizations.

Peer-review under responsibility of the Program Chairs

Keywords: Information and Knowledge Processing, Bioinformatics; Gene Ontology; Semantic similarity of gene sets.

1. Introduction

In recent years, the advent of high-throughput technologies (such as next generation sequencing) and the consequent production of lists of genes associated with specific conditions is stressing the need of recognizing groups of functionally coherent genes in order to construct networks of genes with high pair-wise similarity[1] and characterize these networks with a particular transcriptional behavior[2]. Given difficulties in establishing these

* Corresponding author. Tel.: +0-39-070-6758758; fax: +0-39-070-6758505.
E-mail address: dessi@unica.it

Peer-review under responsibility of the Program Chairs
doi:10.1016/j.procs.2016.09.043

relationships from comparative experiments on the sequence or structure between genes[2], biomedical researchers started to explore new promising ways to compare genes on functional level, including the development of methods for the exploitation of knowledge from ontologies that provide effective descriptions of biomedical events, avoid the short-comings of natural language descriptions (namely ambiguity, subjectivity and lack of structure) and consequently enable automated annotation and automated reasoning over annotations[3]. One of the main contributions in molecular biology has been Gene Ontology (GO) [4], which is dedicated to the functional annotation of gene products in a cellular context[5].

In this paper we explore the potential and any limits of GO ontology in supporting geneticists for capturing additional knowledge about a set of genes which are supposed to play a significant role in a pathological condition. In particular, we try to give suggestions about the extent to which GO ontology can be trusted for detecting the similarity within a group of functionally coherent genes. Our exploration relies on two web resources made freely and easily available by the large multidisciplinary community of biomedical researchers: the HUGO Gene Nomenclature Committee[6] and Reactome[7]. These resources, namely organizations from now on, provide access to a rich catalogues of biomedical and genomic data including information about functional groups of genes i.e. genes acting through the production of specific products and gene-co-function networks. So, we are trusted that GO should also offer a good support in detecting the functional coherence between genes attributed to the same group by the above organizations. For testing our idea, we carried on experiments on the relatedness of gene groups using two classical and popular similarity measures. Results reveal that GO is effective in detecting functional groups of genes, but the hierarchical structure of its catalogue limits the discovery of complex relationships.

The paper is organized as follows. Section 2 describes the web resources and analyses the semantic similarity measures we considered. Next, in the section 3, we present the datasets used in estimating the semantic similarity, the organization our experiments and we also analyze and discuss the results. The section 4 introduces the related work. Finally, section 5 presents our conclusions and the lines of our future research work.

2. Evaluation of the semantic similarity on GO ontology

This section briefly summarizes the basics about the GO structure and the evaluation of the semantic similarity. GO ontology is the result of a collaborative project to provide a controlled vocabularies of terms that describe specific aspects of a gene product's biology. The structure of GO can be described in terms of a graph where each GO term is a node and the relationships between the terms are edges between the nodes. The GO structure is loosely hierarchical as the relationships between different GO terms can be either is-a (parent-child) or part-of (part-whole) relationships, the leaves being the most specific terms. Fig.1 (left) shows an example of the GO structure. Each node is a term and has a unique identifier. Continue arrows indicate "is-a" relationships and dashed arrows denote "part-of" relationships between the terms. In GO there are three types of gene ontologies each describing a particular biological aspect of genes. Specifically, Cellular Component (CC) ontology represents the structural organization of genes, Molecular Function (MF) ontology depicts the specific activities that the gene entails and Biological Processes (BP) ontology describes the series of events that are influenced by the gene. These ontologies are disjoint meaning that no "is-a" relationship operate between terms from different ontologies.

The graph of GO serves as a platform for annotating a term with genes involved in the event that that term describes. Fig.1 (right) shows a toy example of such annotation where the letters in the rectangles indicate the GO terms and the letters in the oval shapes indicate genes directly annotated to GO terms. In Fig. 2 the gene "g7" annotated to the term "GO: 005" and the gene "g8" labeled with the term "GO: 001" are considered as correlated because both are annotated to terms which are semantically alike.

In particular, GO has drawn more and more attention from the bioinformatics researchers as a support for assessing the similarity between two genes by measuring the distance between their respective GO terms. A lot of research work has focused on defining semantic similarity measures tailored to the characteristics of GO[8,9,10] that can be broadly classified into the following categories[11]:

- Information Content (IC) measures - These are the earlier developed methods that evaluate the semantic similarity between two genes by considering the frequencies of their annotations within GO terms and their lower

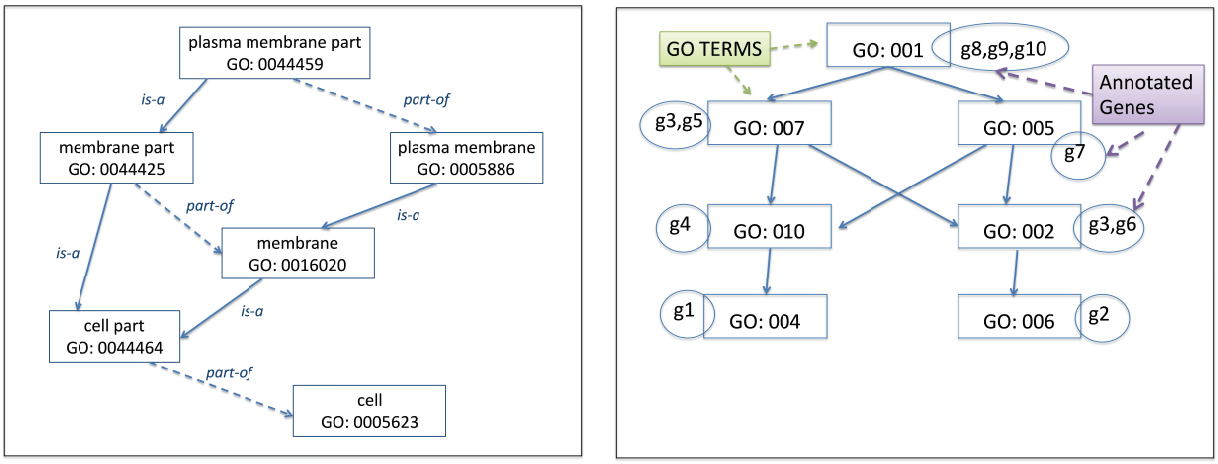

Fig. 1 On the left, an example of GO graph showing a small set of terms from the ontology; on the right, a toy example of GO annotations.

common ancestor (LCA) in the GO graph. For example, in Fig.1 (right), the term GO: 005 is the LCA of the terms GO: 004 and GO: 00 . The IC content of a term t depends on the number of the genes annotated by t and is defined as follows:

$$IC(t) = -\log (G(t)/G(r))$$

where G(t) and G(r) are the number of genes annotated to the term t and to the root term r (including all of its descendants) respectively.

- Graph-based measures - These methods compute the semantic similarity using the topology of the GO graph structure and consider the length or the types of edges by which terms are linked. They make use of the graph structure that is associated with each pair of genes. With reference to Fig.1 (right), for assessing the semantic similarity between the gene "g2" and the gene"g4", graph-based measures consider the ancestors of the term GO: 006 (i.e. GO: 006, GO: 002, GO: 005, GO: 007 , GO: 001) and the ancestors of the term GO: 00 (i.e., GO: 010, GO: 007, GO: 001, GO: 005).

In evaluating the semantic similarity of two genes, a *first and prominent issue* is the choice of the semantic similarity measure to compare the terms to which gene products are annotated. The results from available methods differ in their scale and distribution because each method solely relies on only one or few types of relationships while neglecting the others. There are many cases where a method seems to fail to uncover similarity and others where it suggests a similarity that does not exist. Because no method is perfect, devising the most accurate approach for asserting the semantic similarity between two genes it remains challenging.

A *second critical* aspect is that "is-a" and "part-of" are the only relationships extensively used in GO to express that two concepts are alike. This limited form of representing the knowledge doesn't account for the existence of complex relationships such as "has-part-in" and "is-a-way-of-doing" , which typically correlate a group of genes , for example genes belonging to the same pathway. An *additional critical aspect* is related to the organization of GO into three ontologies . Indeed, this organization reflects the notion that the three ontologies are independent, when, in reality, they represent biological aspects that are strongly correlated. A *final and important issue* is the incompleteness of the GO data. Indeed, not all genes are well enough understood to accurately annotate them to the existing GO nodes and, within the same ontology, a single gene might be annotated at several GO terms. Continuous additional work is necessary because the biomedical knowledge is continuously updated and individual curators must agree on stating how establishing relationships between genes.

Given the importance of evaluating the semantic similarity of a group of genes and the popularity of GO, the geneticists should be aware of the extent to which these issues could affect the GO assertions about the similarity of

two genes. With the aim of giving a contribute in this directions, we carried on experiments presented in the next section.

3. Experiments

3.1. Material and methods

HUGO Gene Nomenclature Committee[6] is a worldwide consortium that is responsible for approving unique symbols and names for human loci, including protein coding genes, RNA genes and pseudo-genes. This organization makes available several online repositories of gene families. Specifically, a family is an efficient and informative way to name related genes, and already this classification works well for a number of established gene families. A gene family index is available that is ordered alphabetically according to family names. Where applicable the family pages include a curated display of hierarchical relationships between families and allows users to browse easily through each hierarchy.

Reactome[7] is an open-source, open access, manually curated and peer-reviewed pathway database. Pathways are large-scale organizations of genes that perform a variety of functions and have complex interactions between them. Information about pathways is limited as it strongly depends on the advances in current knowledge of molecular and cellular biology. In Reactome, pathway annotations are authored by expert biologists, in collaboration with Reactome editorial staff and cross-referenced to many bioinformatics databases.

Being asserted by expert curators, genes belonging to the same pathway or the same family share coherently related functions or participate in the same biological process. As a consequence, they are supposed to have significant higher similarity than expected by chance in terms of GO annotations. Our experiments aim to verify this expectation on genes belonging to 4 families from Hugo and to 4 pathways from Reactome.

Table 1. shows the families and the pathways we considered and how they vary in number of subfamilies (if any) and in the number of considered genes. The last three columns show the coverage of each group i.e. the percentage of the genes annotated in each GO sub-ontology within each single group. From now on we refer as to "group" both a subfamily or a pathway.

Table 1. Selected genes within organizations and their coverage in GO sub-ontologies

FAMILY	# GENES	#SUB_FAMILIES	BP	MF	CC
ABC	33	7	93%	93%	59%
Class_B GPR	15	5	89%	89%	89%
Ligand	107	11	96%	93%	96%
Myosin_superfamily	34	12	83%	83%	83%
PATWAY					
Apoptotic Cleavage	35	0	99%	95%	99%
GABR-Gaba Receptors	67	0	99%	99%	99%
Plateled-derived grow factor A	25	0	93%	94%	98%
Pyruvate	28	0	99%	99%	99%

Because genes belonging to the same group are certainly related, our basic idea is that a similarity method should assign high scores in evaluating their similarity on GO. The mould of our experiments is the evaluation of some "side effects" that could influence this evaluation such as a low coverage of a group, the limits of GO due its representation of knowledge etc. So, differently for most of the literature works that aim to assert the best semantic similarity measure, we considered only two measures. Specifically, among IC-based measures, we considered the Resnik measure[12] that expresses the similarity between two terms as the IC of their lower ancestor. Within this

measure, the terms sharing the same LCA have the same similarity, even if they are at different levels in the GO graph. Among the graph-based measures, we considered the Wang measure[13] that takes into account all the parents of the candidate terms. This measure stresses the difference between two genes that share a common ancestor LCA and two genes that don't share an LCA.

3.2. Data Processing

First, we evaluated the pair-wise semantic similarity of all pairs of genes belonging to the same group on the three GO ontologies (i.e. CC, MF, BP) separately. As we considered two semantic similarity measures, this process resulted in six sequences of values for each single group. Each sequence expresses the distribution of the semantic similarity evaluated by a single method on a specific sub-ontology and within genes belonging to the same group. The mean of the values in a sequence was assumed as representative of the semantic similarity of the corresponding group . Finally, only for families, the semantic similarity of each family was asserted by averaging the semantic similarity of its sub-families.

3.3. Results

For each family and pathway, Fig. 2 to 4 shows the values of the semantic similarity estimated by each measure on the three GO ontologies separately. In the follows, we explore these results from the point of view of the issues presented in the section 2. Specifically, it is clear that the behavior of the semantic similarity is not linear, regardless of the similarity measure used. Indeed, the Wang measure improves Resnik's measure in the majority of families and pathways. This desirable behavior depends on the circumstance that Wang measure considers the global structure of GO and benefits from the properly structured and annotated knowledge of GO.

Results also indicate that the Wang measure seems to be more suitable for devising genes that cooperate within a pattern. These results agree with[14] that compares over GO both the Resnik and Wang measures on three model organisms.

As regard the coverage , it seems that the low coverage (under 55%) of the family ABC does not affect evaluation of the semantic similarity. This indicates that the annotated genes of this family are well correlated in GO. So, the evaluation of their semantic similarity is very high and ,consequently, it originates an high similarity value for all the family ABC, although half of its genes were not considered. Conversely, despite the high coverage of genes belonging to the considered pathways, all the three GO ontologies fail in expressing high values of semantic similarity between genes. These results are due to the hierarchical organization of GO annotations and the use of only two relations (i.e. part-of and is-a) for representing relationships between genes. Because pathways are more complex organizations than families, this aspect severely limit the representation of interactions between its genes.

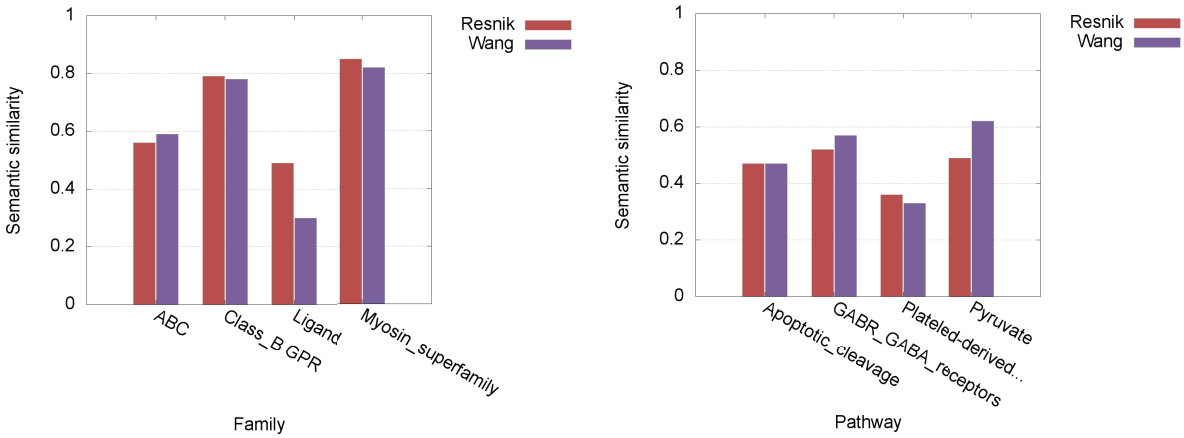

Fig. 2 . The semantic similarity of families (on the left) and pathways (on the right) expressed by the Biological Processes sub-Ontology.

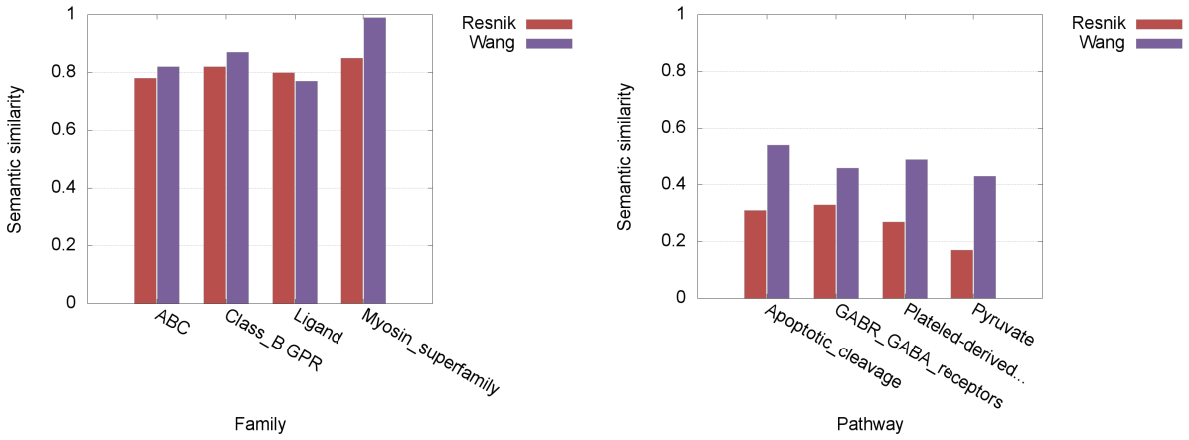

Fig. 3. The semantic similarity of families (on the left) and pathways (on the right) expressed by the Molecular Function sub-Ontology

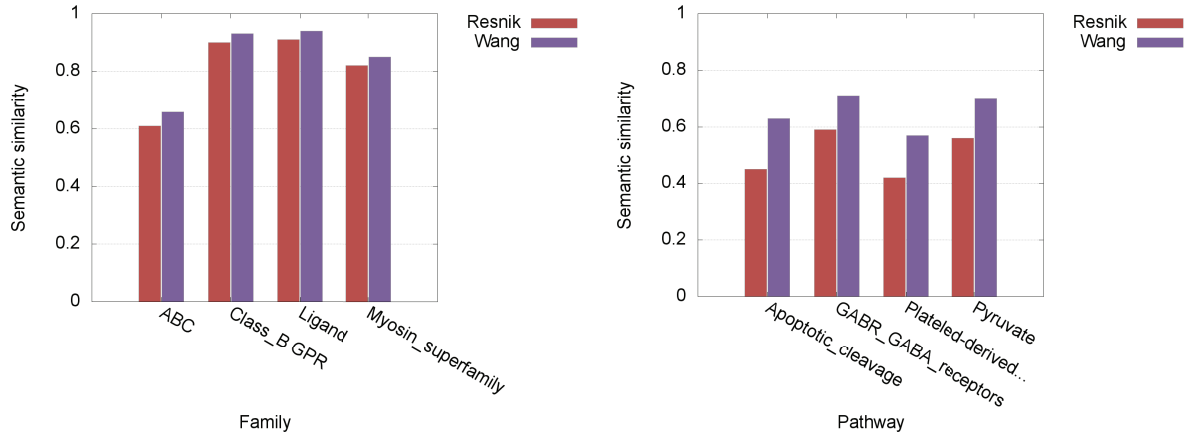

Fig. 4. The semantic similarity of families (on the left) and pathways (on the right) expressed by the Cellular Components sub-Ontology.

Finally, results reveal strong differences within the evaluations performed on the 3 ontologies and a considerable level of asymmetry within families. For example, the similarity of genes belonging to the Ligand family seems to be better evaluated by MF and CC ontologies. This effect is due to the fact that each family can be better featured by a specific category of biological events (i.e. biological processes, molecular functions and cellular component) according to the biological role of its genes.

Unfortunately, as stressed by recent research[14], it does not make sense to combine two or three similarity evaluations performed on different GO aspects into a single measure and difficulties arise from the adoption of approaches that try to integrate multiple measures[14]. So, the choice of the most appropriate ontology remains a specific responsibility of the biologist.

4. Related work

There are an increasing number of application scenarios where biomedical ontologies are exploited to complement and enrich knowledge acquired from experiments and domain experts. Among possible applications,

the semantic similarity is a popular approach to evaluate the relatedness of two genes by considering how close two concepts are from each other in some ontology.

Tailored to the characteristics of GO, new semantic similarity measures are constantly appearing in the literature[15,16]. Hybrid approaches[17] try to integrate state-of-art gene-to-gene similarity measures on GO and demonstrate that this integration improves the performance of single methods, almost on some datasets.

A novel approach is proposed in[18] that incorporates information from gene co-function networks in addition to using GO structure and annotation. The study demonstrates the benefits of such integration in terms of performance of semantic similarity measures.

Few work[19] has been done for evaluating how a specific method performs in recognizing whether genes share or not similar functions and/or participate or not in the same biological event. Such a kind of similarity analysis would be of great importance especially in high dimensional gene lists, where the application of semantic similarity may help the selection of functionally related genes.

5. Conclusions

Over recent years a number of semantic similarity measures have been proposed but the questions of which measure performs better and what are the advantages and limitations in using GO were still open. The work presented in this paper is a preliminary approach to give a contribution in this direction.

We have investigated about the performance of two semantic similarity measures by assessing how well they capture the expected relationship between genes belonging to well known functional organizations. The influence of electronic annotations was assessed on several groups of genes, while the effect of the GO structure was investigated over all the three GO sub-ontologies.

The loss of annotations seems not to influence the semantic similarity of a group with a consistent number of annotated and well-correlated genes. Moreover, it as been observed that, as electronic annotations grow in quantity and quality, the cost of ignoring them will eventually outweigh the gain[15].

Our experiments demonstrate that the evaluation of the semantic similarity on GO presents some limitations related both to completeness and coherency of its taxonomical knowledge that does not provide extensive coverage about gene functions and processes. These limitations do not affect the evaluation of the semantic similarity in organizations where genes exhibit simple correlations like the Hugo families. Conversely, these limitations influence the evaluation of the semantic similarity of complex organizations such as pathways.

As a consequence, we suspect that, due to the knowledge organization implemented in GO, there is a level beyond which the accuracy of the semantic similarity may not be extended. Of course, this suspect needs to be confirmed by further investigation since our experiments are limited and preliminaries to further developments.

Beside the benefits deriving from the exploitations of GO knowledge, our results seem to suggest that the use of electronic annotations should be carefully considered.

Future work will include the evaluation of additional similarity measures as well as investigations about the semantic similarity of further functionally coherent groups defined by biomedical experts such as protein families, enzymes etc. Leveraging on our previous work[20,21], we are also planning to investigate about strategies for complementing the GO knowledge such as the exploitation of knowledge from the most important and largest collections of biomedical documents freely available on the Internet.

References

1. Jang H, Lim J, Lim J-H, Park S-J, Lee K-C, Park S-H. Finding the evidence for protein-protein interactions from PubMed abstracts. Bioinformatics 2006; 22:e220-e226.
2. Kang N, Van Mulligen EM, Kors JA. Comparing and combining chunkers of biomedical text. J Biomed Inform 2011;44:354-360.
3. Azuaje F, Al-Shahrour F, Dopazo J. Ontology-driven approaches to analyzing data in functional genomics. Methods Mol Biol 2006;316:67-86.
4. http://geneontology.org/
5. Harris MA et al. The Gene Ontology (GO) database and informatics resource.Nucleic Acids Res 2004;32(Database issue):D258-61.
6. Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2013. Nucleic Acids Res 2013;41(Database issue):D545-52.
7. http://www.reactome.org/pages/about/reactome/

8. Guzzi PH, Mina M, Guerra C, Cannataro M. Semantic similarity analysis of protein data: assessment with biological features and issues. Briefings in Bioinformatics 2012; 13(5): 569-585.
9. Pesquita, C., Faria, D., Bastos, H., Ferreira, A., Falcao, A., Couto, F.: Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinform. 2008; 9 (Suppl. 5), S4
10. Richards AJ,Muller B., Shotwell M., Cowart LA, Rohrer B., Lu X. Assessing the functional coherence of gene sets with metrics based on the Gene Ontology graph, Bioinformatics 2010; 26(12): i79-i87.
11. Pesquita, C., Faria, D., Falcao, A., Lord, P., Couto, F. Semantic similarity in biomedical ontologies. PLoS computational biology 2009; 5(7):e1000443.
12. Resnik, P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J Art Intell Res 1999;11:95–130.
13. Wang, J., Du, Z., Payattakool, R., Yu, P., Chen, C. A new method to measure the semantic similarity of GO terms. Bioinformatics 2007; 23(10):1274-1281.
14. Peng, J., Li, H., Jiang, Q., Wang, Y., Chen, J. An integrative approach for measuring semantic similarities using gene ontology. BMC Systems Biology 2014; 8 (Suppl 5):S8.
15. Camon EB, Barrell DG, Dimmer EC, Lee V, Magrane M, Maslen J, Binns D, Apweiler R: An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. BMC Bioinformatics 2005; 6 Suppl 1:S17.
16. Yang, H., et al.: Improving GO semantic similarity measures using downward random walks. Bioinformatics 2012; 28, 1383–1389.
17. Peng, J., Wang, Y., Chen, J.: Towards integrative gene functional similarity measurement, BMC Bioinformatics 2014;15(Suppl 2):S5.
18. Peng, J., Uygun, S., Kim, T., Wang, Y., Rhee, S.Y., Chen, J.: Measuring semantic similarities by combining gene ontology annotations and gene co-function networks. BMC Bioinformatics 201;16:44.
19. Pedersen, T., Pakhomov, S.V.S, Patwardhan, S., Chute, C.G.: Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics 2007; 40(3), 288-99.
20. Dessì, N., Pascariello, E., Pes, B.Integrating Ontological Information About Genes, Proceedings of the IEEE 23rd International WETICE Conference 2014; 417- 422.
21. Dessì, N., Dessì S., Pascariello, E., Pes B. Exploring the Relatedness of Gene Sets. Proceedings of the 11th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics 2014; 44-56.