IL NUOVO CIMENTO Vol. 35 C, N. 5 Suppl. 1 Settembre-Ottobre 2012
DOI 10.1393/ncc/i2012-11336-0

Colloquia: PR PS BB 2011

Investigating bias in semantic similarity measures for analysis of protein interactions

M. Mina([1]) and P. H. Guzzi([2])

([1]) *Department of Information Enginering, University of Padova Via Gradenigo 6/A, 35131, Padova, Italy*
([2]) *Department of Experimental Medicine and Clinic, University Magna Græcia Catanzaro, Italy*

ricevuto il 30 Settembre 2011; approvato l' 1 Dicembre 2011

Summary. — Protein interactions are fundamental blocks of almost all cellular processes, so the study of the set of protein interactions in a single organism (also referred to as Protein Interaction Networks - PIN) is an important step in the comprehension of mechanism at molecular level. Recently, the possibility to annotate such data using Gene Ontology and the consequent use of ontology-based analysis has been exploited, *e.g.* the use of semantic similarity (SS) measures. Whereas, SS measures present many challenges and different issues that have to be faced. In particular SS measures are affected from three main biases: i) annotation length, ii) evidence codes, and iii) shallow annotation. The common cause of such biases are the structure of GO and the corpora of annotations (GOA). Consequently, the impact of this variability has to be considered when developing novel algorithms for protein interactions analysis. Although the criticality of these aspects, there is a lack in the systematic analysis of the bias. Few works dealt with the three sources of bias most affecting SS measures. This paper demonstrates the existence of the bias that affect main SS on a set of well-known yeast complexes. It also provides some evidences about the variability of the bias effects over the proteome.

PACS 02.70.-c – Computational techniques; simulations.

1. – Introduction

The interactions among proteins, also referred to as protein-protein interactions (PPI), have a main role in almost all the processes carried out by cells [1]. Nowadays, thanks to the introduction of different technologies, many interactions are known, so the possibility to manage and analyse these data with computer-based tools arises. Recently, there has been a trend towards the integration of diverse information sources, *i.e.* different experimental data as well as prior and functional knowledge. Such knowledge is

often encoded into ontologies, that offer a formal framework to organize in a formal way biological knowledge that is often spread into multiple sources [2].

Gene Ontology (GO) [3] for instance provides a set of descriptions (namely GO Terms) of biological aspects, structured into three main taxonomies: Molecular function (MF), biological process (BP), and cellular component (CC). Each GO Term can be associated to a gene or protein in a process known as annotation, and the gene is said to be annotated with a GO Term. Such process determines a many to many relation among genes (or proteins) and GO Terms enabling enables the use of ontology-based analysis [4]. The comparison of different terms belonging to the same ontology had been defined and a number of different algorithms, referred to as semantic similarity (SS) measures, is available. SS measures usually takes in input GO terms of the same taxonomy and produces as output a value representing their similarity in the basis of different parameters such as the common ancestor of terms, or the information content of terms themselves [5]. The comparison of terms can be extended to sets of terms simply by adopting some mixing functions applied to all the pairwise term similarities. From this scenario, the possibility to compare two gene products using SS arises.

Consequently, many works have focused on: i) the definition of *ad hoc* semantic measures tailored to the characteristics of Gene Ontology; ii) the definition of measures of comparison among genes and proteins; and iii) the introduction of methodologies for the systematic analysis of metabolic networks [5]. Although these considerations semantic similarity measures present many challenges and different issues that have to be faced. It has been reported that SS measures are particularly affected by three factors: i) annotation length, ii) evidence codes, and iii) shallows annotations. These problems should be carefully considered when developing novel algorithms that use semantic similarities. Recent papers show the use of SSs to guide the alignment of pairwise interaction networks, or to identify hubs in PIN [6], as well as to predict protein interactions [7]. Unfortunately, to the best of our knowledge, the impact of the bias has not been carefully addressed.

The main contribution of this paper is to determine which factors effectively represent a bias that occurs in the analysis of PPI data and PINS using SS. In particular, we provide strong evidences that shallow annotation is still a problem heavily affecting SS measures, while we show that the impact of using different EC codes varies across the proteome.

This paper is structured as follows: sect. **2** introduces main concepts about SS, sect. **3** discusses main issues related to SS, sect. **4** presents some case studies, finally sect. **5** presents the conclusions.

2. – Semantic Similarity measures

In the biological field, a semantic similarity (SS) measure is a formal instrument enabling the representation of the relatedness of two or more terms belonging to the same ontology, or between proteins and gene products that are annotated with terms belonging to an ontology. Many semantic similarity measures are nowadays available. For lack of space we do not include any detailed infomation, limiting to the description of used SS measures, the interested reader can find a detailed comparison of SS measures in [5].

Here we just recall that SS measures based on *Information Content* or *Term Depth* evaluate the similarity on the basis of the specificity of the terms, that is, they try to score pairs of specific terms higher than pairs of generic terms. The Information Content of a term t of an ontology O is defined as $-\log(p(t))$, where $p(t)$ is the number of proteins

annotated with t and its descendants in the ontology O, divided by the number of all gene products that are annotated with a term of the same ontology O.

In this work we considered only two SS measures: Resnik BMA and SimGIC. Our choice is led by the results of several assessment works proposing Resnik and SimGIC as two of the most suited measures in the biological field.

Resnik's similarity measure sim_{res} of two terms T_1 and T_2 of GO is based on the determination of the Information Content (IC) of the their most informative common ancestor (MICA) [8]:

$$\text{sim}_{res} = \text{IC}(\text{MICA}(T_1, T_2)). \tag{1}$$

Since Resnik measares similarity between terms, we need a mixing strategy that combines the similarity scores between all the terms annotating two proteins. We used the Best Match Average (BMA) mixing strategy, (see [9]).

Even SimGIC is a measure based on IC, but instead of focusing on only the most informative common ancestor of a pair of terms, it considers the contributions of all the shared ancestors of the two sets of terms annotating two proteins

$$\text{SimGIC} = \frac{\sum_{t \in \{GO(A) \bigcap GO(B)\}} IC(t)}{\sum_{t \in \{GO(A) \bigcup GO(B)\}} IC(t)}, \tag{2}$$

$GO(x)$ is the set of terms annotating protein x and all their ancestors in the GO hierarchy. Since SimGIC considers all the terms annotating two protiens at once, no mixing strategy is needed.

3. – Issues related to Semantic Similarity issues

Some properties of GOs and GOAs actually represent issues for the definition of a fair similarity measure.

Annotation length. The number of term annotations is highly variable among proteins. This characteristics is an aspect of the more general feature of biological ontologies that is the *non uniform distribution of annotations within the same GO and over different GOs and species.*

It has been show that SS scores correlates with the number of annotations two proteins are annotated with [10]. Therefore, two protein pairs functionally related might score low if they have few annotations

Evidence codes. Annotations taken from GO can be derived in different ways, also referred to as evidence codes (EC). Without entering into details, they range between experimentally verified and electronically inferred annotations (IEA). Experimentally verified annotations are likely to be correct, but only cover a small fraction of proteins/terms. Electronically inferred annotations drastically extend the coverage, but at the expense of introducing a lot of noise. SS measures usually do not weight annotations on the basis of their ECs, and one has to choose between including unrealiable annotations to improve the quality of the annotation corpus, or ingoring them but drastically reducing the number of annotations considered.

Shallow annotations. Many proteins are annotated with very generic terms inside the GO. These annotations do not identify the specific role or function of the protein,

but only suggest the area in which the proteins operate. This effect particularly affects IEA annotations, that usually tend to be more generic than experimentally verified annotations.

4. – Case study on Semantic Similarity of CYC2008 Complexes

This section will present a case study on CYC2008 Complexes [11]. The CYC2008 is a comprehensive catalogue of 408 protein complexes in *S. cerevisiae* that are manually curated. These complexes are usually determined in small-scale experiment and published in the literature.

Protein complexes represent small subsets of interacting proteins that share a common biological goal, so they represent small subset of functionally related proteins. Protein complexes have been largely investigated in literature yielding to the introduction of many protein complexes prediction algorithms from PPI data. These algorithms are based on the search of small dense subgraphs [1] and usually search is refined by looking at biological properties of complexes (*e.g.* structural properties of proteins [12]). Consequently, the possibility to use SS as search parameter arises. In this scenario protein complexes can be seen as small dense regions that presents a high value of similarity. Unfortunately our analysis will show that choosing SS presents many challenges.

In the rest of the section we will elucidate how the inclusion of IEA annotations and the shallow annotation affect the semantic similarity of CYC2008 complexes. For each of the three biasing factors we verified whether or not they effectively affects the semantic similarity measures on the considered dataset. We verified that this bias is not a rare and isolated effect occurring only in some pathologic cases, but it is recurrent and still not handled properly. Finally, based on these evidences we draw some conclusions and suggestions regarding how to use semantic similarity measures when looking for patterns within PIN.

4˙1. *EC codes.* – As explained before, the strategy of assignment of GO terms determines a high variability on the reliability of annotations. Despite this, there are not common accepted mathematical models to score this reliability. A main distinction can be made by considering separately experimentally determined annotations and unsupervised electronically inferred annotations (IEA) [13].

Here we focus on the impact of the use of IEA annotations for scoring biological complexes to determine whether including IEA annotations changes semantic similarity scores. As a preliminary step, we verified to which extent the number of annotations significantly changes when considering IEA annotations. We found that while some complexes do not have almost any IEA annotation, others pass from few to many when including IEA annotations. Then we randomly selected some protein complexes that have a high number of IEA annotations. For all these complexes, we evaluated both Resnik BMA and SimGIC semantic similarity scores between proteins within the same complex, first ignoring IEA annotations, and then considering them. Finally, we verified whether similarity scores significantly changed.

Table I lists some of the complexes we considered. It shows that both for Resnik BMA and SimGIC the difference between the average of semantic similarity scores within single complexes is negligible (Delta columns).

We performed a more detailed comparison, considering similarity scores of single protein pairs instead of evaluating the average for each complex. Figures 1 and 2 present a comparison between similarity scores over different complexes evaluated with and without

TABLE I. – *Variation of average SS scores for some complexes with highest amount of IEA annotations.*

Complex data		GO	Average annotations			SimGIC (mean)			Resnik BMA (mean)		
Complex	Size		IEA	Non IEA	Delta	IEA	noIEA	Delta	IEA	noIEA	diff
6-Phosphofructokinase Complex	2	MF	9.00	3.50	5.50	0.96	1.00	0.04	0.90	1.00	0.10
Camp-dependent Protein Kinase	4	MF	7.75	2.50	5.25	0.53	0.52	0.01	0.49	0.42	0.07
Dash Complex	10	CC	11.60	3.90	7.70	0.90	0.88	0.02	0.76	0.78	0.02
Fatty Acid Sinthase Complex	2	MF	15.00	6.00	9.00	0.16	0.14	0.02	0.50	0.61	0.11
Gamma-Tubulin Complex	3	CC	9.67	3.00	6.67	1.00	1.00	0.00	1.00	1.00	0.00
Isocitrate Dehydrogenase	2	MF	9.00	2.00	7.00	1.00	1.00	0.00	1.00	1.00	0.00
Karyopherin Docking Subcomplex of NPC	3	BP	18.00	12.00	6.00	0.84	0.91	0.07	0.90	0.92	0.03
M-AAA Complex	2	MF	11.00	4.00	7.00	1.00	1.00	0.00	1.00	1.00	0.00
MCM2-7 Complex	6	MF	11.83	6.33	5.50	0.50	0.40	0.10	0.72	0.71	0.01
NSP1P Complex	4	BP	17.25	13.00	4.25	0.81	0.80	0.01	0.95	0.96	0.01
Nucleotide - Excision Repair Factor 3 Complex	7	BP	7.86	3.57	4.29	0.59	0.62	0.02	0.69	0.69	0.00

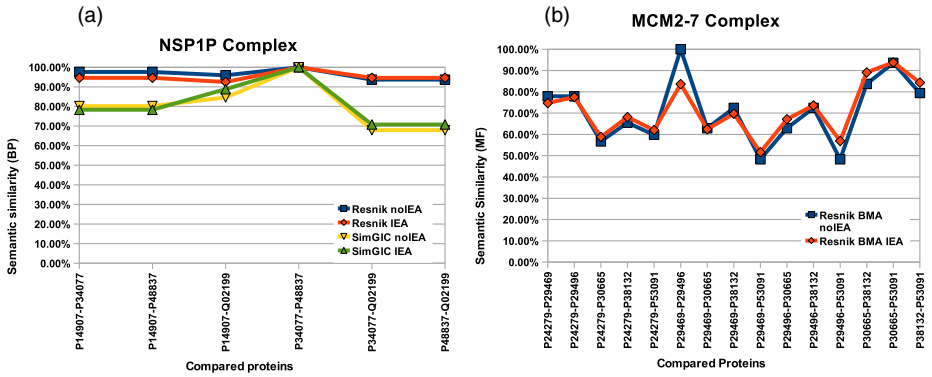

Fig. 1. – Comparison of Semantic Similarities on NSP1P and MCM2-7 complexes.

considering IEA annotations. It is clear that including IEA annotations almost does not influence the scores.

Since we considered complexes with the biggest number of IEA annotations, these evidences assume general validity. In fact complexes with few IEA annotations are likely to be less affected by this factor. Moreover, we considered cases from all the three ontologies (as reported in the GO column of table I).

Other assessment works reported a greater variability when using IEA annotations respect to ignoring them [13-15]. The fact is that these works focused on different groups of proteins. We verified that our analysis agrees with these results when dealing with similar datasets but for lack of space we do not report here these results. Consequently we can conclude that for manually annotated protein complexes of CYC2008 catalogue IEA annotations do not modify SS values. Therefore it is not possible to extend this consideration to arbitrary pairs of proteins, so we conclude that the influence of IEA annotations is not uniform within the proteome.

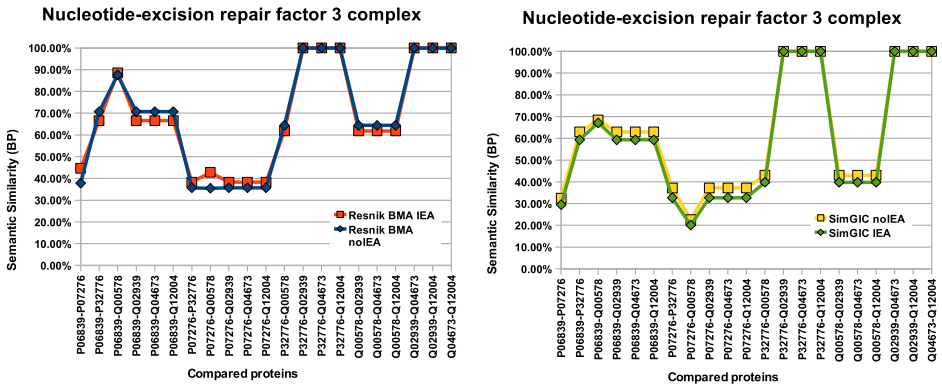

Fig. 2. – Comparison of Semanti Similarities on Excision-Repair complex.

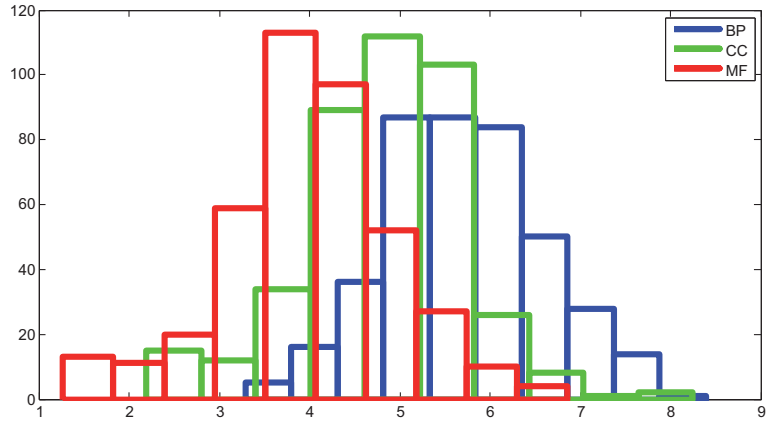

Fig. 3. – Distribution of IC among CYC complexes.

4˙2. *Shallow-annotation problem.* – Shallow annotation is one of the problems that has been explicitly taken into account during the design of several semantic similarity measures. Nevertheless, our analysis suggests that the variability of the specificity of terms used to annotate the proteins still affect semantic similarity measures, introducing a bias that allow complexes annotated with more general terms to get scores significantly higher than complexes finely annotated with specific terms. IC is a common score considered by semantic similarity measures to estimate the specificity of terms within an annotation corpus. For each complex we evaluated the complex average IC as the average of all the ICs of the terms annotated for the proteins of the complex.

As a preliminary step we analysed the distribution of complex average ICs in all the three ontologies, reported in fig. 3. These distibutions can be easily modeled with normal distributions. Many average ICs fall accumulate in the center of the distribution just because many complexes have proteins annotated with both very specific and very generic terms, producing averaged ICs in the middle of the distribution.

We focused on complexes with low variance, since those with high variance mix up terms with low and high ICs, leading to results difficult to interpret. In order to understand how IC influences semantic similarity scores we evaluated the relation between complex average ICs and average complex semantic similarity scores.

Table II reports the average semantic similarity (Resnik and SimGIC measures) scores for some complexes having different levels of average IC. Surprisingly, as clearly

TABLE II. – *Variation of average SS scores for some complexes according to their average IC in BP ontology.*

Complex	Size	mean IC	max IC	min IC	Resnik BMA	SimGIC
AP-2 Adaptor Complex	4	3.29	3.77	1.69	1.000	1.000
SMC5P-SMC6P Complex	8	3.97	4.48	3.49	0.893	0.934
EIF3	7	5.12	6.04	4.01	0.706	0.646
ARP2/3 Protein Complex	7	6.99	9.95	5.09	0.706	0.592
Signalosome Complex	6	7.49	8.16	5.64	0.749	0.631

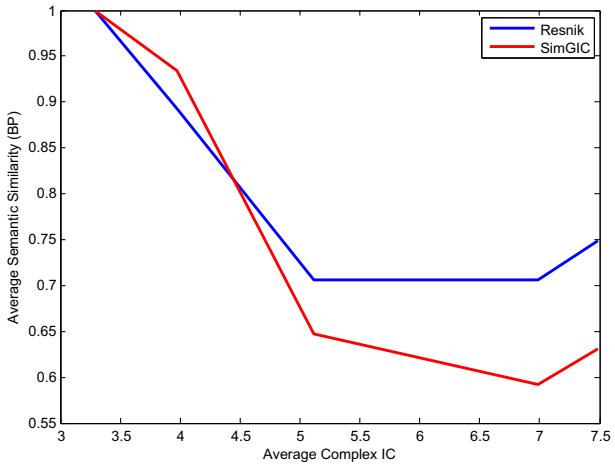

Fig. 4. – Correlation between average IC and SS scores.

represented in fig. 4 there is an inverse correlation between average IC and semantic similarity scores, regardless the measure considered.

These results are surprisingly for two main reasons. First of all, the similarity measures do not reflect the effective similarity between two proteins, since proteins annotated with the same generic terms are scored higher than proteins only partially annotated with the same specific terms. To demonstrate and clarify this point we evaluated the term enrichment of these complexes. Results show that complexes enriched for very specific terms obtain lower similarity scores only because some terms are not annotated for all the proteins within the complex, even though all the proteins within the complexes are annotated with at least one common specific term. Table III describes the terms enriched within the considered complexes. Complexes with highest average ICs are enriched with many and more specific terms than complexes with low average ICs. Therefore, these complexes should score higher than others. However, proteins within the complexes are also annotated with other specific terms. It should be noted that in general these terms are close within the GO.

5. – Conclusion

The application of semantic similarity measures for the analysis of PIN is becoming more and more popular, but SS are affected from evident biases that should be carefully investigated. Here we provided some evidences on these biases on some case studies focusing on the impact of the choose of IEA annotations and on the relation among IC and biological relevance of terms. We demonstrated that the use of IEA annotations is almost uninfluent when evaluating the functional similarity of proteins within the same biological complex. Furthermore, we show that this does not hold when considering other sets of proteins, consistently with results already present in literature. Moreover our results lead to the conclusion that actual semantic similarity measures tend to score higher complexes (and protein pairs) with some common but generic annotations rather than identifying common patterns between proteins annotated with specific but sometimes non overlapping terms. Finally, it should be noted that the two measures considered,

TABLE III. – *Complex Term Enrichment.*

Complex	IC	Enriched Term	Genome Frequency	P-value
AP-2	3.29	intracellular protein transport	331/7166, 4.6%	6.71E-005
		vesicle-mediated transport	384/7166, 5.4%	1.20E-004
		cellular protein localization	398/7166, 5.6%	1.40E-004
		cellular macromolecule localization	415/7166, 5.8%	1.60E-004
		protein transport	512/7166, 7.1%	3.80E-004
SMC5/6P	3.97	DNA recombination	212/7166, 3.0%	3.23E-010
		DNA repair	263/7166, 3.7%	1.49E-009
		response to DNA damage stimulus	311/7166, 4.3%	4.89E-009
ARP2/3	6.99	actin filament polymerization	16/7166, 0.2%	3.99E-018
		actin polymerization or depolymerization	21/7166, 0.3%	4.05E-017
		protein polymerization	26/7166, 0.4%	2.29E-016
		regulation/actin polymerization or depolymerization	15/7166, 0.2%	1.24E-014
		regulation/actin filament length	15/7166, 0.2%	1.24E-014
		regulation/actin filament polymerization	15/7166, 0.2%	1.24E-014
		regulation/protein polymerization	17/7166, 0.2%	3.08E-014
Signalosome	7.49	protein deneddylation	6/7166, 0.1%	8.69E-015
		cullin deneddylation	6/7166, 0.1%	8.69E-015
		adaptation/signaling pathway	21/7166, 0.3%	2.94E-011
		protein modification by small protein removal	31/7166, 0.4%	2.45E-010

Resnik and SimGIC, are based on IC and are between those most unaffected by the shallow-annotation problem. However, it seems that they are still unable to completely correct for this bias. Future semantic similarity measures should be designed keeping those problems into account. Future work will regard the extension of our analysis to other datasets (we plan to consider a cross-species comparison) and other measures.

* * *

Authors thanks C. GUERRA and M. CANNATARO for their suggestions and support during this work.

REFERENCES

[1] CANNATARO M., GUZZI P. H. and VELTRI P., *ACM Comput. Surv.*, **43** (2010) 1.
[2] BACLAWSKI K. and NIU T., *Ontologies for Bioinformatics (Computational Molecular Biology)* (The MIT Press) 2005.
[3] HARRIS M. A. *et al.*, *Nucl. Acids Res.*, **32** (2004) 258.
[4] DU PLESSIS L., ŠKUNCA N. and DESSIMOZ C., *Brief. Bioinf.* (2011)

[5] Pesquita C., Faria D., Falcão A. O., Lord P. and Couto F. M., *PLoS Comput. Biol.*, **5** (2009) e1000443.
[6] Cho Y. R. *et al.*, *BMC Bioinf.*, **8** (2007) 265.
[7] Wang H. *et al.*, *Pattern Rec. Lett.*, **31** (2010) 2073.
[8] Resnik P., *Using information content to evaluate semantic similarity in a taxonomy*, in *IJCAI-95* (1995) 448.
[9] Pesquita C., Faria D., Falcao A. O., Lord P. and Couto F. M., *PLoS Comput. Biol.*, **5** (2009) e1000443+.
[10] Wang J., Zhou X., Zhu J., Zhou C. and Guo Z., *BMC Bioinf.*, **11** (2010) 290.
[11] Pu S., Wong J., Turner B., Cho E. and Wodak S. J., *Nucl. Acids Res.*, **37** (2009) 825.
[12] Aloy P., Böttcher B., Ceulemans H., Leutwein C., Mellwig C., Fischer S., Gavin A. C. C., Bork P., Superti-Furga G., Serrano L. and Russell R. B., *Science*, **303** (2004) 2026.
[13] Benabderrahmane S., Smail-Tabbone M., Poch O., Napoli A. and Devignes M. D., *BMC Bioinf.*, **11** (2010) 588.
[14] Couto F., Silva M. and Coutinho P., *Data & Knowledge Engin.*, **61** (2007) 137.
[15] Pesquita C. *et al.*, *BMC Bioinf.*, **9** Suppl 5 (2008) S4.