BRIEFINGS IN BIOINFORMATICS. VOL 13. NO 5. 569–585 doi:10.1093/bib/bbr066
Advance Access published on 2 December 2011

Semantic similarity analysis of protein data: assessment with biological features and issues

Pietro H. Guzzi, Marco Mina*, Concettina Guerra and Mario Cannataro*

Submitted: 5th August 2011; Received (in revised form): 30th September 2011

Abstract

The integration of proteomics data with biological knowledge is a recent trend in bioinformatics. A lot of biological information is available and is spread on different sources and encoded in different ontologies (e.g. Gene Ontology). Annotating existing protein data with biological information may enable the use (and the development) of algorithms that use biological ontologies as framework to mine annotated data. Recently many methodologies and algorithms that use ontologies to extract knowledge from data, as well as to analyse ontologies themselves have been proposed and applied to other fields. Conversely, the use of such annotations for the analysis of protein data is a relatively novel research area that is currently becoming more and more central in research. Existing approaches span from the definition of the similarity among genes and proteins on the basis of the annotating terms, to the definition of novel algorithms that use such similarities for mining protein data on a proteome-wide scale. This work, after the definition of main concept of such analysis, presents a systematic discussion and comparison of main approaches. Finally, remaining challenges, as well as possible future directions of research are presented.

Keywords: *Semantic similarity measures; protein data; biological features*

Corresponding author. Pietro H. Guzzi, Department of Medical and Surgical Sciences, University Magna Graecia of Catanzaro, Viale Europa (Loc. Germaneto), 88100 Catanzaro, Italy. E-mail: hguzzi@unicz.it, hguzzi@gmail.com
*These authors contributed equally to this work

Pietro H. Guzzi is an Assistant Professor of Computer Engineering at the University 'Magna Græcia' of Catanzaro, Italy, since 2008. He received his PhD in Biomedical Engineering in 2008, from Magna Græcia University of Catanzaro. He received his Laurea degree in Computer Engineering in 2004 from the University of Calabria, Rende, Italy. His research interests comprise bioinformatics, the analysis of proteomics data, and the analysis of protein interaction networks. Pietro is an ACM member and serves the scientific community as reviewer for many conferences. He is associate editor of Information Science journal, and of SIGBioinformatics Record.

Marco Mina is a Ph.D. student at the Department of Information Engineering, University of Padova, Italy, since 2010. He received the bachelor degree and the master degree in Computer Science and Engineering from the University of Padova, Italy, in 2009 and 2007, respectively. His research interests comprise bioinformatics, in particular the analysis of protein interaction networks and the integration of heterogeneous data.

Concettina Guerra is a professor at the Department of Information Engineering of the University of Padova, Italy, and at the College of Computing of the Georgia Institute of Technology, Atlanta, GA, USA. Her research activity is in the areas of Computational Biology, Bioinformatics and Computer Vision. Her recent interests fall in the domains of protein classification, recognition and docking and of comparative analysis of biological networks. She has been on the faculty of the University of Rome, Italy and of Purdue University, USA, for over a decade. She has visited extensively with US Institutions, including Rensseleaer Polytechnic and Carnegie Mellon University. Dr Guerra is a founding member of the steering committee of the International Symposium on 3D Data Processing Visualization and Transmission, that she co-chaired in 2002. She was Co-Director of the CIME School on Mathematical Methods for Protein Structure Analysis and Design (2000) and chairman of the fifth IEEE International Workshop on Computer Architectures for Machine Perception (2000), general chairman of the 10th International Conference on Research in Computational Molecular Biology, RECOMB06 and Co-Director of the series of Lipari Schools in Bioinformatics and Computational Biology.

Mario Cannataro is Associate Professor of Computer Engineering at the Magna Græcia University of Catanzaro, Department of Medical and Surgical Sciences, and an Associate Researcher at ICAR-CNR, Italy. He worked on parallel computing, massively parallel architectures, parallel implementation of logic programs and cellular automata. His current research explores bioinformatics, computational proteomics and genomics, medical informatics, grid and parallel computing and adaptive web systems. Dr Cannataro has published three books and more than 150 papers in international journals and conference proceedings. He is a Senior Member of ACM and a member of IEEE Computer Society and BITS (Italian Bioinformatics Society). Dr. Cannataro is a co-founder and a member of Exeura (www.exeura.com) and EasyAnalysis (www.easyanalysis.it).

INTRODUCTION

Bioinformatics approaches to the study of proteins lead to the introduction of different methodologies and related tools for the analysis of different types of data related to proteins, ranging from primary, secondary and tertiary structures to interaction data [1], not to mention functional knowledge.

One of the most advanced tools for encoding and representing functional knowledge in a formal way is the Gene Ontology (GO) [2, 3]. It is composed of three ontologies, named Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). Each ontology consists of a set of terms (GO terms) representing different functions, biological processes and cellular components within the cell. GO terms are connected to each other to form a hierarchical graph. Terms representing similar functions are close to each other within this graph.

Biological molecules are associated with GO terms that represent their functions, biological roles and localization. This process, usually referred to as annotation process, can be performed under the supervision of an expert or in a fully automated way. Obviously, computationally inferred annotations, commonly known as Electronically Inferred Annotation (IEA), are not as reliable as experimentally determined ones. For this reason, every annotation is labelled with an Evidence Code (EC) that keeps track of the type of process used to produce the annotation itself. Considering the release of annotations of April 2010, ~98% of all the annotations is an IEA annotation [4].

The term annotation corpus is commonly used to identify all the annotations involving a set of proteins or genes, usually referring the whole proteomes and genomes (i.e. the annotation corpus of yeast). For lack of space we do not further describe the Gene Ontology. A comprehensive review has been provided by du Plessis *et al.* [4].

The availability of well formalized functional data enabled the use of computational methods to analyse genes and proteins from the functional point of view. For example, a set of algorithms, known as functional enrichment algorithms, have been developed to determine the statistical significance of the presence (or the absence) of a GO term in a set of gene products. A detailed review of these algorithms can be found in Ref. [5].

An interesting problem is how to express quantitatively the relationships between GO terms. Several measures, referred to as (term) semantic similarity (SS) measures, have been introduced in the last decade. Given two or more GO terms, they try to quantify the similarity of the functional aspects represented by the terms within the cell. Exploiting annotation corpora, SS measures have been further extended to the evaluation of the similarity of genes and proteins on the basis of their annotations.

Many different works have focused on the following tasks: (i) the definition of ad-hoc SS measures tailored to the characteristics of GO; (ii) the definition of measures of comparison of genes and proteins; (iii) the introduction of methodologies for the systematic assessment of SS measures and (iv) the use of SS measures in many different contexts and applications.

Despite its relevance, the application of semantic similarity for the systematic analysis of protein data is still an open research area. There are, in fact, two main questions that have to be addressed: (i) the systematic assessment of SS with respect to other biological features, i.e. how much a high or a low value of SS is biologically meaningful; (ii) how reliable are the SS themselves, i.e. is there any systematic error or bias in the calculation of SS? Both these problems are relevant for the diffusion of SS measures; whereas in the first case several approaches have been proposed, comparing SS measures with a plethora of different biological features, only few works dealt with the second problem in a systematic way [6, 7, 8]. This article reviews SS measures and presents a comprehensive discussion of both problems that may stimulate further discussions.

This article is structured as follows: The next section briefly introduces and categorizes existing SS measures; the next two sections integrate the results of several assessment works and highlight the issues of current SS measures, whereas, the last two sections survey the main existing tools for SS calculation and finally present some possible future directions, respectively.

SS MEASURES

The landscape of current SS measures

A term SS measure is a formal instrument enabling the quantification of the relatedness of two or more terms within an ontology. Measures quantifying the similarity of two terms are often referred to as pairwise measures, whereas measures able to describe the relatedness of two sets of terms, yielding a global similarity of sets, are referred to as groupwise measures.

In the biological field, term similarity measures have been extended to objects (such as gene products and proteins) that are annotated with terms belonging to the ontology, allowing to draw conclusions on the relationship of two proteins relying on the similarity of GO terms.

In the following, we will give a generic but complete overview on the different strategies adopted to evaluate semantic similarity of terms and proteins, classifying similarity measures according to different properties. We will not describe all the measures in detail due to lack of space, but precise references are provided to papers describing them extensively.

Overview of term SS measures

SS measures can be categorized according to the properties of GO terms and annotation corpora on which they rely, and the strategies and models on which they are based. For instance in Ref. [6], a broad categorization is based on GO topological characteristics. Authors propose a first distinction between measures based on GO nodes properties and measures exploiting edge paths within the GO. However, this classification scheme does not cover some measures, such as those based on Vector Space Model, and it considers separately pairwise and groupwise approaches, even though they are substantially similar.

We propose a different categorization grouping together SS measures according to whether or not they consider some aspects or use some common strategies: (i) Term Information Content (IC), (ii) Term Depth, (iii) based on a common ancestor, (iv) based on all common ancestors, (v) Path Length and (vi) Vector Space Models (VSM).

Figure 1 and Table 1 present a detailed classification of Term SS measures according to their characteristics.

Measures based on Term Depth and IC evaluate terms similarity on the basis of the specificity of the terms. While the former assigns specificity to terms according to their depth in the GO directed acyclic graph (DAG), the latter considers the popularity of

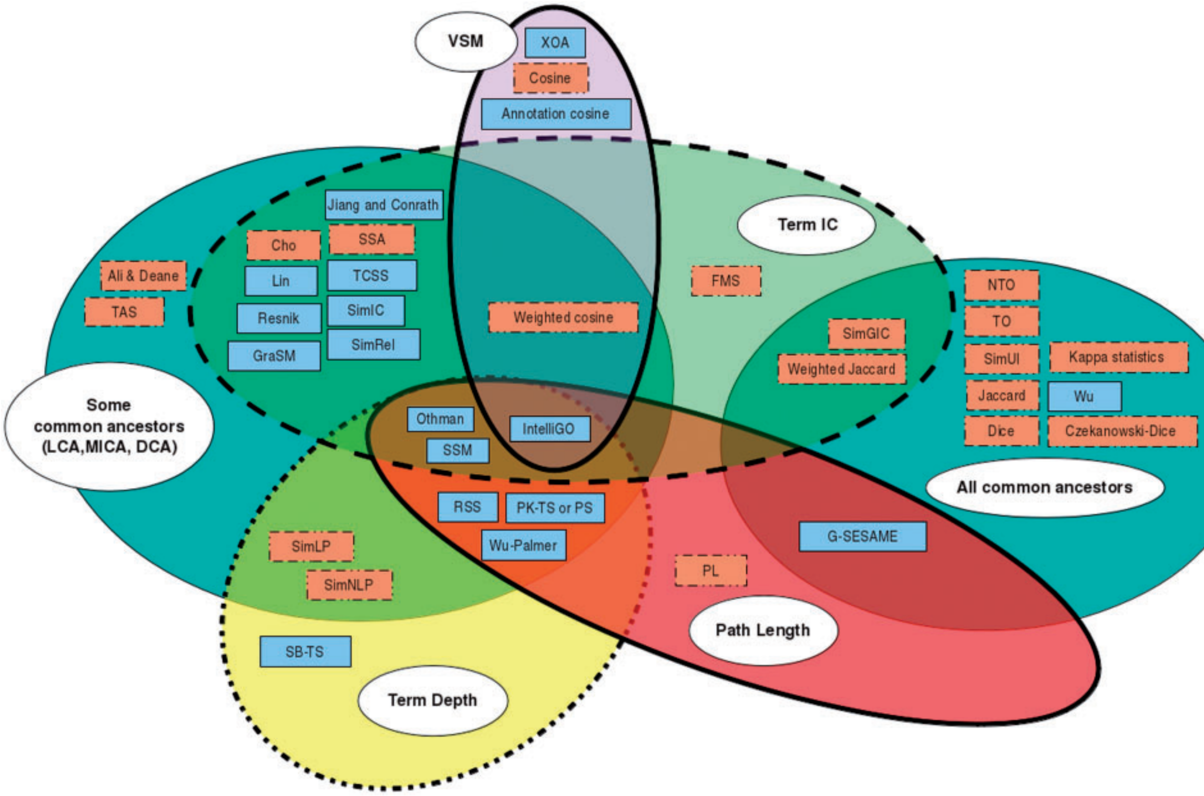

Figure 1: Classification of term SS measures. Each feature used for classifying measures in Table 1 is represented by a circle. Solid and dashed bordered rectangles represent pairwise and groupwise term measures, respectively. Each SS measure is assigned to sets according to its characteristics, summarized in Table 1.

Table 1: Summary of Term SS measures

Type	Name	References	Term IC	Some common ancestors (MICA)	All common ancestors	Path Length	Term Depth	VSM
Groupwise	Ali and Deane	[9]	No	Yes	No	No	No	No
Groupwise	Cho	[10]	Yes	Yes	No	No	No	No
Groupwise	Cosine	[11]	No	No	No	No	No	Yes
Groupwise	Czekanowski-Dice	[12]	No	No	Yes	No	No	No
Groupwise	Dice	[11]	No	No	Yes	No	No	No
Groupwise	FMS	[38]	Yes	No	No	No	No	No
Groupwise	IntelliGO	[13]	Yes	Yes	No	Yes	Yes	Yes
Groupwise	Jaccard	[11]	No	No	Yes	No	No	No
Groupwise	Kappa statistics	[14]	No	No	Yes	No	No	No
Groupwise	NTO	[15]	No	No	Yes	No	No	No
Groupwise	PL	[16]	No	No	No	Yes	No	No
Groupwise	simGIC	[17]	Yes	No	Yes	No	No	No
Groupwise	simLP	[18]	No	Yes	No	No	Yes	No
Groupwise	simNLP	[19]	No	Yes	No	No	Yes	No
Groupwise	simUI	[18]	No	No	Yes	No	No	No
Groupwise	SSA	[20]	Yes	Yes	Depends on measure used			
Groupwise	TO	[21]	No	No	Yes	No	No	No
Groupwise	TAS	[22]	No	Yes	No	No	No	No
Groupwise	Weighted cosine	[23]	Yes	No	No	No	No	Yes
Groupwise	WJ	[11]	Yes	No	Yes	No	No	No
Pairwise	Annotation cosine	[24]	No	No	No	No	No	Yes
Pairwise	G-SESAME	[25]	No	No	Yes	Yes	No	No
Pairwise	GraSM	[26]	Yes	Yes	No	No	No	No
Pairwise	Jiang and Conrath	[27]	Yes	Yes	No	No	No	No
Pairwise	Lin	[28]	Yes	Yes	No	No	No	No
Pairwise	Othman	[29]	Yes	Yes	No	Yes	Yes	No
Pairwise	PS or PK-TS	[30]	No	Yes	No	Yes	Yes	No
Pairwise	Resnik	[31]	Yes	Yes	No	No	No	No
Pairwise	RSS	[32]	No	Yes	No	Yes	Yes	No
Pairwise	SB-TS	[33]	No	No	No	No	Yes	No
Pairwise	simIC	[34]	Yes	Yes	No	No	No	No
Pairwise	simRel	[35]	Yes	Yes	No	No	No	No
Pairwise	SSM	[36]	Yes	Yes	No	Yes	Yes	No
Pairwise	TCSS	[37]	Yes	Yes	No	No	No	No
Pairwise	Wu	[38]	No	No	Yes	No	No	No
Pairwise	Wu-Palmer	[39]	No	Yes	No	Yes	Yes	No
Pairwise	XOA	[40]	Depends on measure used					Yes

Columns Term IC, Some common ancestor (MICA), All common ancestors, Path length, Term depth and VSM refer to the features of the measures described in the text. NTO, normalized term overlap; PL, path length; PS or PK-TS, pekar-staab term similarity; SSA, semantic similarity of annotations; TO, term overlap; TAS, total ancestry similarity; WJ, weighted Jaccard; XOA, cross ontological analysis.

the term (and its descendants) in an annotation corpus. More formally, given an annotation corpus, the IC of a term c is defined as

$$\mathrm{IC(c)} = -\log[\mathrm{p(c)}]$$

where p(c) is the fraction of gene products that are annotated with term c or its descendants in the annotation corpus.

Measures based on a common ancestor first select a common ancestor of two terms according to its properties, and then evaluates the semantic similarity on the basis of the distance among the terms and their common ancestor and the properties of the common ancestor. IC can be used to select the proper ancestor, yielding to the development of methods based on the information content of common ancestor: for instance, the Maximum Informative Common Ancestor (MICA)-based approaches select the common ancestor of two terms t_1 and t_2 with highest IC:

$$\mathrm{MICA}\ (t_1,t_2) = \underset{tj \in \text{ ancestors } (t_1,t_2)}{\arg\max,}\ \mathrm{IC}\ (t_j) \qquad (1)$$

Resnik's measure [31] (sim$_{\mathrm{Res)}}$, one of the most popular SS measures, is an exponent of this category.

The semantic similarity between two terms t_1 and t_2 is simply the IC of the MICA:

$$\mathrm{simRes}\ (t_1,t_2) = \mathrm{IC}[\mathrm{MICA}\ (t_1,t_2)] \tag{2}$$

Lin's measure [41], sim$_{\mathrm{Lin}}$, considers both the information content of the MICA and of the input terms:

$$\mathrm{sim}_{\mathrm{Lin}}\ (t_1,t_2) = \frac{\mathrm{IC}[\mathrm{MICA}(t_1,t_2)]}{\mathrm{IC}(t_1) + \mathrm{IC}(t_2)} \tag{3}$$

In a similar way, Jiang and Conrath's measure [27], simJC, takes into account the MICA and the input terms:

$$\mathrm{simJC}(t1,t2) = 1 - \mathrm{IC}(t1) + \mathrm{IC}(t2) - 2\times \mathrm{IC}[(\mathrm{MICA}(t1,t2)] \tag{4}$$

Also simGIC [42] is a measure based on IC, but instead of focusing on only the most informative common ancestor of a pair of terms, it considers all the common ancestors of two sets A and B of GO terms:

$$\mathrm{sim}_{\mathrm{GIC}}(A,B) = \frac{\sum_{t\in\{GO(A)\cap GO(B)\}} \mathrm{IC}(t)}{\sum_{t\in\{GO(A)\cup GO(B)\}} \mathrm{IC}(t)} \tag{5}$$

where GO(X) is the set of terms within X and all their ancestors in the GO hierarchy. In general, measures based on all common ancestors [18–20, 22, 26–32, 34–37, 39] collect all the ancestors of terms, and then evaluate the overlap between the two sets, sometimes using also other characteristics (e.g. IC) or edge distance to determine term similarity.

Completely different from previous approaches are those techniques based on Path Length [16, 25, 29, 30, 32, 36, 39]. In this case, similarity measures are correlated to the length of the path connecting the two terms. This approach usually considers the length of the path from terms to their common ancestor (lowest common ancestor –LCA—or maximum common ancestor—MCA).

Finally, VSM-based measures [11, 14, 23, 24] are based on a two-step strategy. Initially the annotation of proteins are represented as vectors, each GO Term is a component of such a vector. Then, the similarity is evaluated by considering the distance among vectors that are defined using topological considerations (e.g. the cosine of the induced angle), as well as semantic considerations, (e.g. IC-based).

Protein SS measures

The underlying idea to calculate protein semantic similarity is to evaluate the semantic similarity between all the terms annotating two proteins, and then combine them in some way.

Groupwise Term SS measures can be directly extended to measure protein similarity, simply considering as input the two sets of GO terms annotating the proteins. Instead, Pairwise Term SS measures evaluate similarity of pairs of terms and therefore, are not directly applicable to genes and proteins. Consequently it is necessary to define a strategy, called mixing strategy, that transforms all the pairwise term similarities into a single representative value. There are six mixing strategies reported in literature:

Average (avg): the average of all term pairwise similarities [43];

Maximum (max): the maximum of all term pairwise similarities [44];

Best Match Average (BMA): the average of similarity between best matching terms [45];

funSim: first protein semantic similarities in MF and BP ontologies are determined using max, avg or BMA mixing strategies and then they are combined together in a non-linear way [35];

Information Theory-based Semantic Similarity: best-matching pairs are filtered on the basis of their similarities, then the average is calculated [46];

FuSSiMeG: similar to max strategy; and the maximum of all term pairwise similarities weighted by the ICs of the terms is selected [26]

ASSESSMENT AND COMPARISON OF SS MEASURES

SS measures are substantially different from all the other classical measures such as sequence similarity because they use information regarding the functions and roles of proteins themselves. Conversely, classical similarity measures such as sequence similarity, rely on the assumption that sequence similarity implies functional similarity. Therefore, semantic similarity ideally should provide the true measure of functional similarity. However, the comparison of gene product using semantic similarity requires a well defined ontology and a complete and reliable annotation corpus. By far, the GO is still incomplete, and annotation corpora are extremely far from being complete and reliable [4, 47].

For these reasons, semantic similarity measures have to be compared with other measures and biological features in order to uncover the issues affecting current SS measures and design more and more reliable measures. In this section, we present an overview of all the assessments of SS measures performed in the literature relying on the following classical similarity measures and biological features: protein interaction, sequence similarity, Pfam-based and EC-based similarity, functional modules and complexes and expression profile similarity.

Use of semantic similarity to analyse protein interaction data

Several works investigated the use of SS measures to discern among interacting and non-interacting protein pairs under the assumption that interacting pairs may have higher similarity values. The papers we reviewed rely on different datasets and consider different SS measures, but all of them adopt the same assessment procedure: (i) build a positive set of interacting protein pairs and a negative set of non-interacting protein pairs; (ii) determine semantic similarities for each protein pair in both the positive and negative datasets and (iii) evaluate the discriminative power of each similarity measure, i.e. the ability of scoring protein pairs in the positive set higher than those in the negative set. All the protein pairs with semantic similarity above a certain threshold are assumed to be interacting. Specificity and sensitivity at different cut-off thresholds are collected and combined into a receiver operating characteristics (ROC) curve that is used to compare prediction performance with respect to the positive and negative sets.

SS measures have been verified to be good predictors of protein–protein interactions [48]. The robustness and generality of the analysis is certified by the heterogeneity of data used in different assessment methods. As expected, the best results are obtained when using BP ontology, while CC ontology has proven to be not particularly suited for this task [49]. Moreover, few works coherently report that ignoring IEA annotations leaves results almost unaffected [37, 49].

With few exceptions [50], most of the works identified Resnik as one of the best semantic similarity measures [34, 37, 49], especially when combined with the max mixing strategy [51]. This is not unexpected, since max strategy favours protein pairs sharing even only a part of their functions and, as we reported previously, two proteins are likely to interact even when they only have in common some of their aspects. SimIC [34] and Topological Clustering Semantic Similarity (TCSS) [37] achieve slightly better results than Resnik. In particular, in Ref. [37], a detailed comparison based on an extension of ROC analysis, highlighting the differences between Resnik and TCSS is provided.

Relation between semantic and sequence similarities

Sequence similarity has been one of the oldest approaches used to establish relations among genes. Nowadays it is clear that proteins with similar sequence are likely to accomplish similar functions. Therefore, there should be a good correlation between sequence and semantic similarity, at least when considering MF ontology. Three approaches have been used to assess semantic similarity using sequence similarity.

Many works determined Pearson's linear correlation between sequence and semantic similarity for a set of protein pairs [7, 15, 34, 37, 43, 52]. Usually a binning procedure is applied to raw data to reduce noise and allow a pattern to emerge. It consists in dividing sequence and semantic scores into some intervals, and using averages within each interval as data points. Pearson correlation has been evaluated first directly on raw scores, and later on binned data.

Analyses based on Pearson correlation rely on the assumption that the relationship between semantic and sequence similarity is linear. Unfortunately, this assumption proved to be false [17]. Pesquita *et al.* proposed an assessment based on non-linear regression, i.e. trying to fit data with a function that closely follows the behaviour of semantic similarity against sequence similarity [17].

Finally, two papers [34, 35] analysed the distribution of semantic similarity values in four different categories of protein pairs, corresponding to four different levels of evolutionary relationship ranging from no sequence similarity to orthology.

In general, all the works discovered a positive correlation between sequence and semantic similarity, low for Pearson correlation on raw data and much more defined on binned data. Non-linear regression analysis found that the normal cumulative distribution fits data for many different semantic similarity measures, confirming the positive yet, non-linear agreement between sequence and semantic similarity.

However, there are many cases in which the two measures disagree, especially at low levels of sequence similarity. In fact, even though functionally related proteins tend to have high sequence similarity, there is a fraction of protein pairs with high sequence similarity but no functional similarity [35], as well as protein pairs with different sequences but involved in the same biological process or even accomplishing the same function.

Sequence similarity has been used to compare different semantic similarity measures over several different datasets and conditions.

This evaluation is not as straightforward as it is for Protein-Protein Interactions (PPI) data, different strategies have been used, and not always coherent results are obtained. Interestingly, some of the most recent works did not report big performance variations between different measures. In general, it seems that BMA mixing strategy should be preferred to max and avg approaches. Resnik BMA, simGIC and sinIC BMA have often been identified as the best measures [17, 34]. TCSS reported the highest scores outperforming Resnik and simGIC, but due to its recent introduction, it has only been assessed in one work [37].

Pfam families

Proteins generally comprise one or more functional regions, commonly termed domains. Since proteins sharing the same domains are likely to have some common functional aspects, assessing SS measures using domain composition information is an appealing alternative to sequence similarity data. Couto *et al.* [26] showed that, especially when using MF ontology, semantic similarity significantly increases as the number of shared families between two proteins increases. Both Couto *et al.* [26] and CESSM [7], a tool for automatic comparison of SS measures performances, evaluate Pearson's linear correlation of semantic similarity and a Pfam-based similarity measure, and rank SS measures according to correlation levels.

Surprisingly, the two works disagree with each other: Couto concludes that Jang and Conrath measure with GraSM option is the best measure and outperforms Lin and Resnik measures, whereas according to CESSM, Jang and Conrath measures do not behave well at all. Benabderrahmane *et al.* [13] proposed a novel assessment strategy based on a set of proteins encompassing 10 different Pfam clans. They evaluated the ability of different semantic measures to discriminate between protein pairs within the same clan and protein pairs belonging to different clans. They report that their measure, IntelliGO, outperforms other SS measures. However, this conclusions are in disagreement with CESSM evaluation.

In conclusion, even though Pfam families are suited to assess and compare SS measures, current conclusions are not coherent, and further investigations should be performed before using Pfam families to compare SS measures.

Functional modules

In a protein interaction network, a functional module is a set of interacting proteins that share a common biological goal or play a biological role. For instance, a pathway or a protein complex is a functional module. A biological pathway is a number of biochemical steps, linked together, that perform a process inside cells. Since proteins within the same pathway are involved in the same biological process, they are likely to be annotated with the same or similar terms in the GO (at least in BP ontology) and therefore having high semantic similarity.

Guo *et al.* [49] analysed the distribution of Resnik scores when considering pairs of proteins belonging to the same pathway. They showed that all protein pairs within a Kyoto Encyclopedia for Genes and Genomes (KEGG) pathway have significantly higher similarity scores than randomly expected when considering BP ontology. On the other side, semantic similarity on MF and CC ontologies decays exponentially as proteins became farther within the same pathway.

Wang *et al.* [25] performed a manual validation and comparison of G- SESAME and Resnik measures (over the MF ontology) using Yeast pathways. They used the scores to hierarchically cluster genes within the same pathway, and by visually inspecting clustering results, they concluded that G-SESAME scores protein pairs consistently with human perception of protein relatedness.

Finally, Benabderrahmane *et al.* [13] evaluated the difference between similarity scores between protein pairs within the same pathway and protein pairs from different pathways.

Even protein complexes have been used to assess SS measures. Li *et al.* [34] and Wu *et al.* [32] first reconstructed Yeast PPI network relying on BP- and CC- based semantic similarity scores, and then mapped manually annotated Munich Information Center for Protein Sequences (MIPS) [53] complexes on their networks. They evaluated how many MIPS complexes were included in their reconstructed network. Wu's PPI network based on Relative Specificity Similarity (RSS) measure (and max mixing strategy) encompassed 120 out of 214 MIPS complexes, whereas Li's PPI network, based on simIC (and max mixing strategy), extended the coverage to 159 complexes. Such analysis revealed the applicability of SS measures for PPI network reconstruction problems and for biological clustering.

Expression profiles

Several studies compared SS measures to gene expression profile similarity. As for sequence similarity, for each gene pair within a set, both semantic similarity and expression profile correlation is evaluated.

Wang *et al.* [54] compared semantic similarity to expression profiles correlation for pairs of genes from Eisen dataset [55]. They verified that for all the tested similarity measures, high semantic similarity is significantly associated with strong expression correlation.

More recently, Sevilla *et al.* [44] reported that semantic similarity and expression profile correlation show low levels of correlation when considering raw data. As well as, in the case of sequence similarity, binning data dramatically improves correlation levels. High correlation levels have been reported also in Refs. [34, 37]. Remarkably, in [37] no binning procedure has been applied on data, leading to a more pure evaluation of agreement between the two measures.

In agreement with Wang *et al.*, a graphical inspection of Sevilla *et al.* results suggests that the correlation of gene expression and semantic similarity at low levels of semantic similarity values is negligible, whereas at higher levels of semantic similarity values, they are highly related. This behaviour has been confirmed by Xu *et al.* [51]; they show that semantic similarity linearly increases with respect to expression correlation when focusing only on gene pairs with high levels of expression correlation. On the other side, they show that for gene pairs with low levels of expression profile semantic similarity correlation is generally low and stationary.

Ranking SS measures according to their behaviour compared with expression profile similarity, Li *et al.* identify simIC as the best measure, followed by Resnik (coupled with max mixing strategy) [34], whereas in Ref. [37], TCSS is the best measure, followed by Resnik (always coupled with max mixing strategy).

SUMMARY

The several assessments reported provide a clear vision of the extent to which SS measures correlate with other biological features and similarity measures.

Semantic similarity proved to be a good predictor of PPIs. All the assessments followed a common and clear procedure. It seems that max mixing strategy, as well as groupwise term measures that favour protein pairs also when they share only a part of their functions, are the most suited for this task. According to the studies presented, Resnik, simIC and TCSS are the best measures in this case.

A different situation arises when considering assessments based on Pfam data. In this case, there are inconsistences between the conclusions reported in different assessments, with Resnik, J&C and IntelliGO being the best measures in some works and the worst in some others. Further studies should be performed in order to harmonize the different results and provide definitive conclusions.

Strong correlation has been reported in many assessments based on sequence or expression similarity data, indicating that in general, there is an agreement between these measures. However, even though protein pairs with high levels of sequence or expression similarity tend to show high semantic similarity, there are many protein pairs with low sequence or expression similarity but high semantic similarity. This highlights the fact that functional similarity extends beyond sequence or expression similarity, that are measures able to capture only some aspects of the biological similarity of two proteins. BMA mixing strategy and groupwise term measures that comprehensively compare all the aspects of two proteins, achieved the highest correlation levels in these assessments. In general, Resnik, simGIC, simIC and TCSS are identified as the best measures.

Assessing semantic similarity measures with biological pathways uncovers the ability of these measures in identifying strongly related protein pairs. In this case, Guo *et al.* [49] confirmed that SS measures for proteins within the same pathway behave as expected over the three different ontologies. Benabderrahmane [13] proposed a comparison based on biological pathways, but surprisingly many measures such as Resnik or simGIC did not behave as well as expected. As for Pfam data, further investigations should be provided.

In conclusion, different performances have been reported for SS measures in different contexts. Resnik, simGIC, simIC and TCSS are often identified as the best measures. Max mixing strategy should be used when looking for protein pairs sharing even only a portion of their functions, while BMA approaches should be considered when a more comprehensive comparison is required.

ISSUES OF SS MEASURES

In this section, the major problems affecting SS measures are discussed. We distinguish between: external issues related to the properties of annotation corpora (such as the shallow annotation problem), internal issues inherent to the design of the single measures (deriving for example from a misuse of the IC) and despite this distinction, external and internal issues present many links, e.g. measures that make wrong assumptions on the characteristics of annotation corpora.

External issues

With external issues we indicate all those problems related to annotation corpora and GO structure. The most relevant exponents of this class are the shallow annotation problem, the annotation lenght bias and the use of Evidence Codes.

Shallow annotation problem

Biologically speaking, terms within an ontology have different specificity, i.e. they describe more or less particular functions, processes and cellular components within the cell.

Many proteins are annotated with very generic terms inside the GO (shallow annotations). These annotations do not identify the specific role or function of the protein, but only suggest the area in which the proteins operate.

Shallow annotations heavily affect the fact of SS measures. Since proteins are often annotated with very generic terms in the GO, many proteins will share one or more very generic terms. However, the fact that two proteins share generic terms does not imply that they are closely related. SS measures have to keep into account this fact.

Annotation length bias

Annotations are not uniformly distributed among the proteins within an annotation corpus: some proteins are sensibly more annotated than others, and many proteins have just only one annotation term. Many works revealed a correlation between semantic scores and number of annotations [8, 26], clearly indicating that annotation length biases similarity scores. Moreover, the distribution of annotations sensibly changes among different organisms and GOs, and consequently some SS measures might behave differently on different annotation corpora.

Evidence codes

Protein annotations are assigned in many different ways [4]. A big portion of term annotations fall into the electronically inferred category. Experimentally verified annotations are likely to be correct, but only a small fraction of proteins are annotated through this process. Electronically inferred annotations drastically extend the coverage, but at the expense of introducing a lot of noise and the presence of more generic annotations. Table 2 summarizes the main characteristics of both classes of annotations. Thus, there is a trade-off between reliability of annotation and the number of annotated proteins.

The problem of whether or not to ignore or weight IEA annotations when using SS measures has been investigated in the literature and the impact on different datasets and various assessment data has also been analysed.

Table 3 summarizes the assessment proposed, the datasets used and the conclusions drawn in each work. The considered datasets significantly differ in terms of extension (ranging from few hundreds to several thousands of genes) and specificity (some of them are restricted to specific set of proteins, whereas some others use the entire Uniprot database).

Many works focus on generic datasets of randomly selected proteins from Uniprot database or other resources including thousands of proteins (i.e. DIP [56]). Couto *et al.* [26] considered the impact of ignoring IEA annotations on the correlation with Pfam family similarity between the most annotated proteins within Uniprot. Correlation is lower and much more variable when only manual annotations are considered. Even though they only consider J&C BMA, it is reasonable to assume that the results would be the same for other measures. CESSM [7] registers relevant

Table 2: Characteristics of IEA and non-IEA annotations

IEA	Non-IEA
Often generic annotations	Usually more specific annotations
More uniform distribution across the GO	Not uniform distribution across the GO
Covers a bigger set of proteins	Limited protein coverage.
	Few annotations per protein
Error-prone	High quality

Table 3: Conclusions of different assessment works about the usage of IEA annotations

References	Assessment	Dataset	Conclusions
Benabderrahmane et al. [13]	Pathways, Pfam families, CESSM	NCBI annotation files; human/MF: IEA ∼30.000 pp, non-IEA 16.243 pp human/BP: IEA ∼27.000 pp, non-IEA 21.462 pp yeast/MF: IEA ∼12.000 pp, non-IEA 9.564 pp yeast/BP: IEA ∼9.000 pp, non-IEA 18.496 pp	For Pfam families: significantly better results considering IEA. For KEGG Pathways: almost the same results for yeast, better using IEA for human
Jain and Bader [37]	PPI, Expression profiles, CESSM	GOA release 2010.	No significant variations registered. Using IEA annotations is recommended
Pesquita et al. [17]	Sequence similarity	GOA-UniProt [17] release of February 2007	Loss of resolution considering IEA for maximum and average mixing strategies. No big differences otherwise. Possible data circularity threat when using IEA/ISS.
Couto et al. [26]	Pfam families	500 proteins with the largest number of GO annotations from the December 2004 releases of UniProt and GO	Significantly better and more stable results using IEA annotations
Guo et al. [49]	PPI	1649 human proteins within KEGG pathways as positive interaction dataset; negative random set from Entrez Gene	No significant variations registered.
Lord et al. [43]	Sequence similarity	all human proteins in SwissProt (2002); at least 1 annotation in GO (TAS only); about 7000 proteins	Better results when using only TAS annotations and avg mixing strategy.
Pesquita et al. [7]	Pfam families, Sequence similarity, EC	random protein pairs from UniProt with at least 1 annotation in each GO with IC > 0.5, 1 annotation in EC classes and 1 pf	Performance decays on avg and max mixing strategies when considering IEA. No big variation in the other cases
Sevilla et al. [44]	Expression profiles	754 annotated genes from MARSHA Murine database (http://microarray.cnmcresearch.org/pgadatatable.asp); 753 genes with annotations from RAD database (http://cbil.upenn.edu)	Using IEA annotations improves results

variations in the correlation with Pfam family similarity when using max mixing strategy, with a significant decrease of performance when considering IEA annotations. This might be an effect of using the max strategy, since it is more susceptible to incorrect annotations. On the other side, no big variations appear when using different mixing strategies, such as best match approach, or groupwise SS measures, with slightly better results obtained when considering IEA annotations (in agreement with Couto et al. [26]).

The same conclusions follow from the assessments with sequence similarity data. One of the first assessments reports that using only Traceable Author Statement (TAS) annotations in combination with average mixing strategy leads to higher correlation levels [43]. Consistently, Pesquita et al. [17] showed that correlation with sequence similarity tends to be sensibly higher if IEA annotations are ignored when using max or average mixing strategy, providing a further evidence of the impact of false annotations on max mixing strategy. Performance are almost the same when considering groupwise measures or BMA strategy.

Again considering assessments on generic datasets, Jain and Bader's results [37] show that there is almost no difference between using or ignoring IEA annotations when comparing SS measures with PPI data both in human and yeast proteome, surprisingly even when using max mixing strategy.

Few works focused on specific sets of proteins. For example, Sevilla et al. [44] reported lower correlation levels with gene expression data when IEA annotations are not considered, supporting the recommendation of using IEA annotations.

An interesting analysis regards biological complexes and pathways. Guo *et al.* [49] reported that considering or ignoring non-TAS annotations does not make a big difference on their anaylsis on KEGG pathways. More recently, Benabderrahmane *et al.* [13] extended the analysis on KEGG pathways and specific Pfam families. They compared the ability of different types of annotation to capture the similarity between proteins within the same pathway or Pfam clan. Results vary according to the organism, ontology and assessment data considered, since the distribution of annotations and the ratio between IEA and non-IEA annotations sensibly varies within annotation corpora and between different species and ontologies. Considering the entire Gene Ontology annotation corpus (released in August 2011), we determined the ratios between IEA and non-IEA annotations for several different species (see Supplementary Data S2 and Ref. Rhee *et al*. [57]). Our results are consistent with those reported in [13]. For Yeast organism and BP ontology, where the ratio between non-IEA and IEA annotations is ~2:1, non-IEA annotations are enough to capture similarity of proteins within the same KEGG pathway, even though considering IEA annotations almost does not affect the results. A different behaviour is observed for human proteins on BP ontology (where the ratio is about 1:1), where neither IEA nor non-IEA are enough to describe proteins involved in KEGG pathways.

Consistently with CESSM [7] and Couto *et al.* [26], worse results are obtained when ignoring IEA annotation in the assessment with Pfam families on MF ontology. In this case, most of the annotations regarding MF ontology are electronically inferred, and therefore considering IEA annotations generally improves the quality of the similarity measures.

In order to extend this analysis, we repeated it on biological complexes of Yeast organism reported in CYC database [58] and Pfam families, testing Resnik BMA and simGIC measures instead [59]. Results are consistent with Benabderrahmane *et al.*

All the works presented, highlight an important fact: the impact of IEA annotations is not uniform. In particular cases (i.e. biological pathways) it is almost non-existent, obviously especially when the number of non-IEA annotations is high. In other cases, IEA annotations play an important role.

According to almost all the works reviewed, considering IEA annotations leads to better results, or at least does not influence results in a negative way. Therefore, in general we suggest to use IEA annotations, especially for large-scale studies. In fact, since including IEA annotations increases the number of proteins annotated, it allows to extend semantic similarity measures over some parts of the non-annotated proteome. This is particularly valid for MF ontology and species with a low rate of non-IEA to IEA annotations, such as human.

However, particular attention should be paid when using max-like mixing strategies, since they have proven to be badly affected by wrong annotations. Interestingly, this is not the case when semantic similarity measures are used to infer PPIs.

Internal issues

Almost every SS measure is built considering three aspects:

(i) define a measure of term specificity;
(ii) use GO structure to establish the similarity of terms and,
(iii) extend the similarity to protein pairs.

In the following, we discuss the main issues affecting SS measures that are related to these aspects, finally reporting some other minor issues.

Define a measure of term specificity

Conceptually, genes annotated with similar specific terms should score higher than genes annotated with similar but generic ones. Due to the shallow annotation problem, semantic similarity measures have to keep into account term specificity. However, dealing with it is not trivial. In fact by itself this information is not quantitative, and thus not computationally tractable. Two measures have been used to quantitatively represent term specificity: Term IC over the annotation corpus and Term Depth within the GO. Many SS measures rely only on one of the two properties, whereas few try to exploit both of them (Table 1). Both IC and Term Depth have advantages and drawbacks (Table 4).

Table 4: Comparison of advantages and drawbacks of using term IC and term depth

Term IC	Term depth
Annotation corpus dependent	Independent from annotations
High even for generic terms with few annotations (corpus bias)	Term specificity is not always related to term depth
Alleviates shallow annotation problem	Alleviates shallow annotation problem

IC is a measure based on the corpus of annotations and therefore, the IC score for the same term varies between different annotation corpora. As argued by Wang *et al.* [25] the similarity scores between two terms should be independent from the number of proteins annotated with them. Therefore, semantic similarity of GO terms should be based only on the structure of GO ontologies, independently on the distribution of annotations. Moreover, there are rarely used generic terms that unavoidably end up having high ICs. This means that in some cases IC does not reflect the biological specificity of the term. According to the nomenclature proposed by Mistry *et al.* [15], we will refer to this phenomenon as corpus bias.

On the other side, some similarity measures use Term Depth within the GO to estimate term specificity (Table 1). This strategy has the advantage of producing scores consistent over different annotation corpora. However, Term Depth fails whenever there are specific and generic terms at the same depth in the GO. This is not a rare case, since some regions of the GO are denser than others both in terms of nodes and edges.

In general, the analysis suggests that measures based on IC are more accurate than those based on Term Depth [17, 34]. The fact that Resnik, that only relies on IC, is often indicated as one of the best measures despite its simplicity is a further evidence of the good approximation of the biological specificity obtained using IC. Several other measures based on IC (i.e. simGIC, simIC, TCSS) reported even better correlations with several biological features. Other measures based on IC behave poorly because they do not integrate IC properly: Lin and J&C measures are relative measures that evaluate the distance between the IC of the terms and their MICA. When two proteins share an annotation, the distance is 0, leading to a similarity score equal to 1. Obviously, this is not the proper behaviour in the case of generic annotations. SimRel and simIC overcome this problem weighting Lin and J&C scores with a term directly proportional to the IC of the common ancestor of the two terms

As reported in Table 1, there are also some similarity measures that do not consider neither Term IC nor Term Depth. Those measures perform worse than the others in almost every assessment. It is likely that the reason for this behaviour is that they are unable to discern between generic and specific terms. Consequently, the trade-off among term specificity and shallow annotation problem should be deeper addressed. Further works should address this problem, trying to understand how to represent biological specificity.

Use GO structure to establish the similarity of terms

Despite the continuous process of GO updating, there are some common characteristics of the GO DAG that causes some issues that afflict almost all the SS measures. First of all, the distance (either if it is based on the difference between ICs or on the length of shortest path) is unable to capture the specifity of the terms [54]. Terms close to the root, i.e. topologically non specific, may be biologically specific. Thus, they could score high, as well as very specific terms.

Some regions of the GO are denser than others both in terms of nodes and edges. This might be a direct consequence of the incompleteness of the GO itself, or an intrinsic characteristics of the GOs. In any case, this characteristic influences measures that assume that the GO is uniform. Moreover, Term Depth and specificity is not a direct relationship. Terms in denser regions might be more specific than terms at the same depth but in sparser regions.

GO is not a balanced tree, having some regions denser than others. Thus, terms with the same depth in different regions of the GO are likely to have different specificity. This has impact on depth- or edge-based measures. Most sophisticated measures keep track of local density when evaluating term similarity (i.e. Othman, SSM) but unfortunately this problem has not been deeply addressed. TCSS somehow addresses this problem reshaping the GO exploiting terms IC in order to obtain a more balanced ontology.

It has been objected that all the common ancestors of two terms should be considered, since all of them contribute to the functional similarity of the term.

Nevertheless, measures based on only one common ancestor equally compete with measures based on all the common ancestors. However, using the set of shared ancestors without considering their degree of contribution to the terms considered is wrong as well, since far ancestors contribute less than closer ones [54]. G-SESAME addresses this issue by weighting the contribution of common ancestors. On the other side many measures do not consider the specificity of common ancestors. SimGIC is a remarkable example of measure that uses IC to weight common ancestors in order to deal with different specificity.

Extend the similarity to protein pairs

A common objection moved against extending pairwise term measures to proteins with mixing strategies is that they consider separately all the terms and then merge results, and this might lead to a loss of information. As an example, considering a pair of terms at time does not keep into account how the terms are spread over the GO. In general, many recent works suggest that using groupwise approaches leads to better results.

Pesquita *et al.* noticed that the average mixing strategy does not behave well in terms of agreement with sequence similarity, especially at high levels of sequence similarity [17]. At high levels of sequence similarity there is an increase of annotations, leading to worse performance in the case of average approach. In fact even proteins that share a large fraction of same terms may contain terms highly dissimilar among them thus decreasing the global average. In the non-IEA, average approach works better, probably because the number of annotations is more uniform. Average approaches scored the worst results even in PPI prediction.

The maximum approach is a simple way to assign protein–protein similarity score evaluating the most similar pairs of terms. It shows low resolution because it simply finds the best term pair, ignoring the others. Moreover, it seems to be tied to the number of annotations. In fact, the more annotations a protein has, the more likely is that it shares a term with another protein, resulting in a high score. In the non-IEA dataset, this issue is less notable since many proteins only have one annotation. Moreover, this approach is extremely sensible to false annotations, since an incorrectly annotated term might lead to a wrong high score. For PPI prediction, however, this approach seems to work well since proteins only need to share a common function to be likely to interact [37]. The Best Match Average approach shows the best behaviour. When comparing two proteins, it considers all terms but compares each term with the most similar term annotated for the other protein. It seems to be independent from the number of annotations of the protein (or at least, the most unaffected). In a more general framework, where a more general measure of functional similarity is required, the Best Match Average approach might be better than maximum, since it considers also the differences between two proteins.

Minor issues

It has been reported that the distribution of annotations and the types of annotations varies a lot among different organisms and GOs. It is not clear to which extent the behaviour of the SS measures differs in different scenarios, since all the assessments focus on Human and Yeast organisms. In Benabderrahmane *et al.*'s assessment [13] there is a strong variability among similarity measures, so it is not possible to make any conclusion about the use (or not) of IEA annotations on different species. This fact is a concrete evidence of the need of a systematic analysis on other organisms with different annotation distributions.

Resnik has been reported as one of the best measures, almost always outperforming Lin and J&C measures. Resnik only considers the IC of the MICA. Since many term pairs share the same MICA, this leads to step-like scores that tend to cluster at some discrete levels of similarity. This is not a desirable behaviour, especially for applications that need to rank protein pairs. Moreover, this effect would increase as the number of annotations increases, and therefore it can only get worse as the GO becomes more complete.

Finally, Resnik does not produce normalized scores by itself.

Summary

Considering the issues presented above, we identified some critical points:

It is absolutely necessary to use the concept of term specificity, whether it is represented by IC or Term Depth. It seems that using IC leads to better results, but the implications of using a measure independent from annotation corpus, such as Term Depth, are quite interesting.

Considering only one common ancestor might not be the optimal choice, since it discards the contribution of other terms. All the common ancestors should be considered instead, but appropriately weighted according to both their contribution to the considered terms and their specificity.

Using Path Length between pairs of terms might be a dangerous way, due to the unbalanced nature of GOs.

Finally, groupwise approaches should be preferred to pairwise approaches for evaluating protein pair similarity. However, most of groupwise approaches do not take into account term specificity and behave poorly. SimGIC is the only groupwise measure competing with pairwise approaches.

Actually, Resnik is one of the most considered semantic similarity measure, always included in assessment works and behaving properly most of the times. More recent approaches based on term specificity such as G-SESAME, simGIC, simIC and TCSS seem to outperform Resnik in several cases, but with the exception of simGIC they have not been included in many assessment or comparison works. Anyhow, we believe they represent the next generation of semantic similarity measures that should be used. All of them offer improvements over Resnik in different directions, resolving some of the issues presented above.

TOOLS AND APPLICATIONS FOR THE SEMANTIC ANALYSIS

This section presents some existing tools implementing SS measures. The current scenario is characterized from the absence of a tool that implements all the SS measures or that is easily extendible. Considering the distribution, tools are mainly available as web servers (Table 6) or as packages for the R platform (Table 5). However, FuSSiMeG, ProteInOn, FunSimMat, csbl.go and SemSim together cover almost all the similarity measures. In general, tools are based on GO and annotation corpora. Some tools, such as the web servers, include their own copy of annotation corpora and GO, offering user-friendly and ready-to-go solutions. However, they rely on maintainers for updated data, and generally do not offer many possibilities of customization or extension. On the contrary, other tools such as stand-alone R-packages, are generally more flexible and often easily extendable, but they require the intervention of expert users. Usually they require the user to provide annotations and ontologies as input data in more or less common formats. While this enables the full control over data used and guarantees the possibility to use most-updated data, the preparation of input datasets may result in an error-prone waste of time.

A possible future direction may regard the development of a comprehensive platform for the integrated semantic analysis of protein interaction networks.

Table 5: Packages for R

Functions	Measures	Input data
csbl.go [60]		
SS measures, Clustering based on SS	Resnik, Lin, JiangConrath, GRaSM, simRel, Kappa Statistics, Cosine, Weighted Jaccard, Czekanowski-Dice	Genes and Proteins annotations
GOSemSim [61]		
SS measures	Resnik, Lin, Jiang, simRel, G-SESAME	GO Terms
GOvis [62]		
SS measures	simLP, simUI	Entrez gene IDs, Gene ontology

Table 6: Web servers for calculation of semantic similarity measures

Web server	Functions	Measures
FuSSiMeG [47]	SS measures, statistical tests	Resnik, Lin, JiangCon- rath, GraSM
http://xldb.fc.ul.pt/biotools/rebil/ssm/		
ProteInOn [17]	SS measures, search for assigned GO Terms and annotated proteins, representative of GO Terms	Resnik, Lin, JiangCon- rath, simGIC, GraSM, simUI
xldb.di.fc.ul.pt/tools/proteinon/		
FunSimMat [63]	SS measures, disease-related genes prioritization	simRel, Lin, Resnik, JiangConrath
http://funsimmat.bioinf.mpi-inf.mpg.de/		
GOToolBox [12]	SS measures, clustering	Si, Sp, SCD
http://genome.crg.es/GOToolBox/		
G-SESAME [25]	SS measures, clustering	G-SESAME
http://bioinformatics.clemson.edu/G-SESAME		

None of these tools requires input annotations or GOs.

CONCLUSIONS

SS measures, i.e. the quantification of the similarity of two or more terms belonging to the same ontology, is a well established field. The application of SS to proteins as well as to protein interaction data is still a novel field, and there exist many open problems and challenges that should be addressed.

In this work, we presented a survey of main SS measures based on GO and the main issues discussed in the scientific community regarding: (i) the assessment of SSs in terms of biological features and (ii) the biases on the calculation of SSs that arise in the biological field.

The several assessments reported in this work provide a clear vision of the extent to which SS measures correlate with other biological features and similarity measures. Furthermore, we identified some critical points and issues regarding current measures that may stimulate discussion and research in the future. We concluded that Resnik, one of the most considered SS measures, behaves properly most of the times. More recent approaches based on term specificity such as G-SESAME, simGIC, simIC and TCSS seem to outperform Resnik in several cases. We believe they represent the next generation of SS measures that should be used, since all of them offer improvements over Resnik in different directions, resolving some of the issues presented above.

Finally, we point the attention to another problem that is emerging. Recently, semantic similarity measures have been used as input or validation data in several genome-wide and proteome-wide applications (i.e. PPI networks alignment problems), requiring the computation of semantic similarity between whole proteomes. Considering as an example the yeast organism, containing more than 5000 proteins, these applications require the calculation of more than 25 millions of protein similarities. So far, there is only one freely available tool, GS2 [64], that efficiently generates proteome-wide SS scores. Further work is necessary to design faster solutions for the calculation of semantic similarity measures.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- Comprehensive review of semantic similarity measures.
- Suggestions concerning the best uses of semantic similarity measures tailored to different contexts.
- Assessment with biological features.
- Critical discussion of common issues.
- Outline of future direction of research.

References

1. Cannataro M, Guzzi PH, Veltri P. Protein Interaction Data: technologies, databases and algorithms. *ACM Comput Sur* 2010;**43**:1–36.
2. Baclawski K, Niu T. *Ontologies for Bioinformatics (Computational Molecular Biology)*. Cambridge, MA: The MIT Press, 2005.
3. Harris MA, Clark J, Ireland A, *et al*. The gene ontology (go) database and informatics resource. *Nucleic Acids Res* 2004;**32**: 258–61.
4. du Plessis L, Škunca N, Dessimoz C. The what, where, how and why of gene ontology, a primer for bioinformaticians. *Brief Bioinform* 2011; doi: 10.1093/bib/bbr002.
5. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009;**37**: 1–13.
6. Pesquita C, Faria D, Falcao AO, *et al*. Semantic similarity in biomedical ontologies. *PLoS Comput Biol* 2009;**5**:e1000443.
7. Pesquita C, Pessoa D, Faria D, *et al*. CESSM: Collaborative Evaluation of Semantic Similarity Measures, JB2009: Challenges in Bioinformatics 2009.
8. Wang J, Zhou X, Zhu J, *et al*. Revealing and avoiding bias in semantic similarity scores for protein pairs. *BMC Bioinformatics* 2010;**11**:290.
9. Ali W, Deane CM. Functionally guided alignment of protein interaction networks for module detection. *Bioinformatics* 2009;**25**:3166–73.
10. Cho Y-R, Hwang W, Ramanathan M, *et al*. Semantic integration to identify overlapping functional modules in protein interaction networks. *BMC bioinformatics* 2007;**8**:265.
11. Popescu M, Keller JM, Mitchell JA. Fuzzy measures on the Gene Ontology for gene product similarity. *IEEE/ACM Trans Comput Biol Bioinform* 2006;**3**:263–7412.
12. Martin D, Brun C, Remy E, *et al*. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol* 2004;**5**:R101.
13. Benabderrahmane S, Smail-Tabbone M, Poch O, *et al*. IntelliGO: a new vector- based semantic similarity measure including annotation origin. *BMC Bioinformatics* 2010; **1**:588.
14. Huang DW, Sherman BT, Tan Q, *et al*. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol* 2007;**8**:R183.
15. Mistry M, Pavlidis P. Gene Ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics* 2008;**9**:327.
16. Al-Mubaid H, Nagar A. Comparison of four similarity measures based on GO annotations for Gene Clustering. Report no. 3, 2008 IEEE Symposium on Computers and Communications, 6–9 July 2008. Morocco: Marrakech.
17. Pesquita C, Faria D, Bastos H, *et al*. Metrics for GO based protein semantic sim- ilarity: a systematic evaluation. *BMC Bioinformatics* 2008;**9(Suppl 5)**:S4.
18. Gentleman A. *Visualizing GO Distances Using Bioconductor.* http://bioconductor.org/packages/2.3/bioc/html/GOstats.html (10 October 2011, date last accessed).
19. Ye P, Peyser BD, Pan X, *et al*. Gene function prediction from congruent synthetic lethal interactions in yeast. *Mol Syst Biol* 2005;**1**: 2005.0026.
20. Sheehan B, Quigley A, Gaudin B, *et al*. A relation based measure of semantic similarity for Gene Ontology annotations. *BMC Bioinformatics* 2008;**9**:468.
21. Lee HK, Hsu AK, Sajdak J, *et al*. Coexpression analysis of human genes across many microarray data sets. *Genome Res* 2004;**14**:1085.

22. Yu H, Jansen R, Gerstein M. Developing a similarity measure in biological function space. *Bioinformatics* 2007.

23. Chabalier J, Mosser J, Burgun A. A transversal approach to predict gene product networks from ontology-based similarity. *BMC Bioinformatics* 2007;**8**:235.

24. Bodenreider O, Aubry M, Burgun A. Non-Lexical approaches to identifying associative relations on the gene ontology. *Pac Symp Biocomput* 2005;91–102.

25. Wang JZ, Du Z, Payattakool R, *et al*. A new method to measure the semantic similarity of GO terms. *Bioinformatics* 2007;**23**:1274–81.

26. Couto FM, Silva MJ, Coutinho PM. Measuring semantic similarity between Gene Ontology terms. *Data Knowl Eng* 2007;**61**:137–52.

27. Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. Arxiv preprint cmp-lg/9709008 (1997), no. Rocling X. In *International Conference Research on Computational Linguistics (ROCLING X)*, September 1997, 9008.

28. Lin D. An information-theoretic definition of similarity. In: *Proceedings of the 15th International Conference on Machine Learning*, Vol. 1. Citeseer, 1998, 296–304.

29. Othman RM, Deris S, Illias RM. A genetic similarity algorithm for searching the Gene Ontology terms and annotating anonymous protein sequences. *J Biomed Inform* 2008;**41**: 65–81.

30. Pekar V, Staab S. Taxonomy learning factoring the structure of a taxonomy into a semantic classification decision, COLING '02. In: *Proceedings of the 19th International Conference on Computational Linguistics* 2002, Vol. 2, 786–92.

31. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. Arxiv preprint cmp-lg/9511007 1 (1995). In *Proceedings of the 14th International Joint Conference on Artificial Intelligence IJCAI 95*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 1995.

32. Wu X, Zhu L, Guo J, *et al*. Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. *Nucleic* Acids Res 2006; **34**:2137–50.

33. Yu H, Gao L, Tu K, Guo Z. Broadly predicting specific gene functions with expression similarity and taxonomy similarity. *Gene* 2005;**352**:75–81.

34. Li B, Wang JZ, Feltus FA, *et al*. Effectively integrating information content and structural relationship to improve the go-based similarity measure between proteins. Arxiv preprint arXiv:1001.0958 (2010), 1–54. http://arxiv.org/pdf/1001.0958.

35. Schlicker A, Domingues FS, Rahnenführer J, *et al*. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* 2006;**7**:302.

36. Couto FM. Implementation of a Functional Semantic Similarity Measure between Gene-Products. Phd Thesis, Lisbona University, 2003.

37. Jain S, Bader GD. An improved method for scoring protein- protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics* 2010;**11**:562.

38. Wu H, Su Z, Mao F, *et al*. Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Res* 2005;**33**:2822–37.

39. Wu Z, Palmer M. Verb semantics and lexical selection. In: *Proceedings of 32nd Annual Meeting of the Association for Computational Linguistics,* 1994;133–8.

40. Sanfilippo A, Posse C, Gopalan B, *et al*. Combining hierarchical and associative gene ontology relations with textual evidence in estimating gene and gene product similarity. *IEEE Trans Nanobiosci* 2007;**6**:51–9.

41. Dekang L. *An Information-theoretic Definition of Similarity*. San Francisco: Morgan Kaufmann, 1998.

42. Pesquita C, Faria D, Francisco M. Couto: measuring coherence between electronic and manual annotations in biological databases. *SAC* 2009;806–7.

43. Lord PW, Stevens R, Brass AL, *et al*. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 2003; **19**:1275–83.

44. Sevilla JL, Segura V, Podhorski A, *et al*. Correlation between gene expression and GO semantic similarity. *IEEE/ACM Trans Comput Biol Bioinform* 2005;**2**:330–8.

45. Azuaje F, Wang H, Bodenreider O. Ontology-driven similarity approaches to supporting gene functional assessment. In: *Proceedings of The Eighth Annual Bio-Ontologies Meeting, Citeseer,* 2005;9–10.

46. Tao Y, Sam L, Li J, *et al*. Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics* 2007;**23**:i529–38.

47. del Pozo A, Pazos F, Valencia A. Defining functional distances over gene ontology. *BMC Bioinformatics* 2008; **9**:50.

48. Ivanov A, Zgoda V, Archakov A. Technologies of protein interactomics: a review. *Russian Journal of Bioorganic Chemistry* 2011;**37**:4–16.

49. Guo X, Liu R, Shriver CD, *et al*. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics* 2006;**22**:967–73.

50. Chen G, Wang J, Li M. GO Semantic Similarity based analysis for human protein interactions. In: *International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*, *3–5 August 2009*. China: Shanghai, 2009;207–10.

51. Xu T, Du L, Zhou Y. Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data. *BMC Bioinformatics* 2008; **9**:472.

52. Lord PW, Stevens RD, Brass A, *et al*. Semantic similarity measures as tools for exploring the gene ontology. *Pac Symp Biocomput* 2003;**612**:601–12.

53. Mewes H, Frishman W, Güldener D, *et al*. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 2002; **30**:31–4.

54. Wang H, Azuaje F, Bodenreider O, Dopazo J. Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology CIBCB 04* 2004;25–31.

55. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998;**95**:14863–8.

56. Xenarios I, Ioannis, Rice, Danny W. Salwinski, *et al.,* DIP: the Database of Interacting Proteins. *Nucleic Acids Res* 2000; **28**:289–91.

57. Rhee SY, Wood V, Dolinski K, *et al*. Use and misuse of the gene ontology annotations. *Nat Rev Genet* 2008;**9**: 509–15.

58. Pu S, Wong J, Turner B, *et al*. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res* 2009;**37**:825–31.

59. Mina M, Guzzi PH. *PR PS BB Workshop on Pattern Recognition in Proteomics Structural Biology and Bioinformatics Ancona 13th Septembre 2011, ICIAP Annual conference*.

60. Ovaska K, Laakso M, Hautaniemi S. Fast gene ontology based clustering for microarray experiments. *BioData Min* 2008;**1**:11.

61. Yu G, Li F, Qin Y, *et al*. GOSemSim: an r package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 2010;**26**:976–8.

62. Ning Z, Jiang Z. GOVis, a gene ontology visualization tool based on Multi-Dimensional values. *Protein Pept Lett* 2010; **17**:675–80.

63. Schlicker A, Albrecht M. Funsimmat update: new features for exploring functional similarity. *Nucleic Acids Res* 2010; **38(Suppl 1)**:D244–8.

64. Ruths D, Ruths T, Nakhleh L. Gs2: an efficiently computable measure of go-based similarity of gene sets. *Bioinformatics* 2009;**25**:1178–84.