Briefings in Bioinformatics, 18(5), 2017, 886–901

doi: 10.1093/bib/bbw067
Advance Access Publication Date: 26 July 2016
Paper

Gene Ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery

Gaston K. Mazandu, Emile R. Chimusa and Nicola J. Mulder

Corresponding authors. Gaston K. Mazandu. Tel.: +27 21 650 5789; Fax: +27 86 275 6014; E-mail: gmazandu@cbio.uct.ac.za; Nicola J. Mulder. E-mail: nicola.mulder@uct.ac.za

Abstract

Gene Ontology (GO) semantic similarity tools enable retrieval of semantic similarity scores, which incorporate biological knowledge embedded in the GO structure for comparing or classifying different proteins or list of proteins based on their GO annotations. This facilitates a better understanding of biological phenomena underlying the corresponding experiment and enables the identification of processes pertinent to different biological conditions. Currently, about 14 tools are available, which may play an important role in improving protein analyses at the functional level using different GO semantic similarity measures. Here we survey these tools to provide a comprehensive view of the challenges and advances made in this area to avoid redundant effort in developing features that already exist, or implementing ideas already proven to be obsolete in the context of GO. This helps researchers, tool developers, as well as end users, understand the underlying semantic similarity measures implemented through knowledge of pertinent features of, and issues related to, a particular tool. This should empower users to make appropriate choices for their biological applications and ensure effective knowledge discovery based on GO annotations.

Key words: gene ontology; semantic similarity tools; protein functional similarity; protein functional analysis; Gene Ontology annotations

Introduction

Advances in high-throughput biology technologies have resulted in an exponential increase in biological data, including genomic sequence and functional data for several targeted organisms. Outcomes of these high-throughput experiments are usually long lists of genes, which are often used to identify trends or patterns relevant to the biological experiment being performed. However, providing the biological interpretation or significance of these lists of candidate genes/proteins using biological information stored in bioinformatics resources, such as the Gene Ontology (GO) [1] and Gene Ontology Annotation (GOA) [2–4], remains a major challenge. GO provides a way of consistently describing genes and their products in any organism and produces a well-adapted platform to computationally process data at the functional level [5]. GO is a widely adopted resource that has been successfully deployed in several biological and biomedical applications, ranging from theoretical to experimental and computational biology [6]. The GO structure is composed of three distinct components (http://geneontology.org/page/ontology-documentation), biological processes (BP), molecular function, cellular component, each organized (http://geneontology.org/page/ontology-structure) as a directed acyclic graph (DAG) in which each edge represents an 'is_a', a 'part_of', 'regulates' ('positively_regulates' and 'negatively_regulates'), 'has_part' or 'occurs_in' relationship [7] between GO terms

Gaston K. Mazandu received his PhD in Bioinformatics from University of Cape Town. He is an honorary senior member of the Computational Biology Division at the Institute of Infectious Disease and Molecular Medicine at University of Cape Town (UCT) and a Researcher at the African Institute for Mathematical Sciences (AIMS).
Emile R. Chimusa received his PhD in Bioinformatics from University of Cape Town. He is currently a Senior Lecturer at the Division of Human Genetics, Department of Pathology, University of Cape Town (UCT).
Nicola J. Mulder is a Professor at the Department of Integrative Biomedical Sciences and the Head of the Computational Biology Division, University of Cape Town (UCT).
Submitted: 27 January 2016; **Received (in revised form):** 12 June 2016

(ftp://ftp.geneontology.org/pub/go/www/GO.draft-page.shtml). Incorporating the GO structure in GOA-based protein analyses has significantly improved the outcomes of these analyses [5, 6]. Thus, several GO semantic similarity measures [8–23] have been proposed in recent years and have enabled the integration of biological knowledge embedded in the GO structure into different biological analyses.

Preliminary data processing using computational or bioinformatics methods often involves enrichment of gene lists, e.g. to identify differentially expressed genes, and yield lists of genes or proteins relevant to a particular biological question. Using biological knowledge embedded in the GO structure has enabled further comparison or classification of protein lists to understand the biological phenomena underlying the experiment under consideration and identify pertinent processes. Semantic similarity measures are used to tackle challenges related to knowledge discovery based on the GOAs. They have been deployed in several biological applications, including (i) identifying and ranking genes or proteins associated with biological experiments [24], (ii) retrieving BP most pertinent to the biological experiments [5], (iii) describing and characterizing the structure of protein data sets and elucidating protein or gene set patterns to improve biological interpretation [25–30], (iv) predicting and validating protein–protein interactions [18, 31–34], protein functions [35, 36] and cellular localization [37], as well as imputing missing values [38], (v) constructing mappings between GO terms and user-defined categories, such as GO Slim categories [39], and (vi) integrating heterogeneous GOA pipelines [40]. This suggests that these measures are currently playing an important role in improving protein analyses at the functional level for various high-throughput biological data sets.

A number of publicly available GO semantic similarity tools (ftp://ftp.geneontology.org/pub/go/www/GO.tools_by_type.semantic_similarity.shtml) were independently designed to facilitate exploration of different semantic similarity measures [5], including web-based online tools and standalone software. In 2009 and 2011, approximately 10 such tools were identified by Pesquita *et al.* [41] and Guzzi *et al.* [42], respectively. Most of these tools, including, but not limited to, FuSSiMeG [12, 43], ProteInOn [44] and G-SESAME [45], are context-dependent tools, i.e. developed for a new measure that was being introduced and proven to be more efficient than previous ones. Most of these web-based tools implement only annotation-based measures, which require protein GOA data for computing term information content (IC) values, except G-SESAME [45], which implements the Wang *et al.* approach [14], and DaGO-Fun [5] implementing the Zhang *et al.* [10], Wang *et al.* [14] and GO-universal [19] approaches. Currently, the Zhang *et al.*, Wang *et al.* and GO-universal approaches are the only topology-based measures (where a term IC value is computed based on the intrinsic structure of GO), which have been extensively applied to GO. One of the characteristics common to several tools is that they are organism-specific, and therefore, a given term in the GO structure may have different IC values depending on the organism under consideration.

In the past few years, GO semantic similarity has been an active research area, focusing mainly on developing new approaches, adapting existing approaches and combining several existing approaches to design new ones. This confusing proliferation of semantic similarity approaches has resulted in two main issues: (i) old and existing approaches are considered as novel, and obsolete ideas or approaches proven to be ineffective are brought back because of the lack of a comprehensive analytical (mathematical) comparison and experimental (biological) evaluation of these different approaches, and (ii) difficulties for the end users to choose the most appropriate approach or similarity measure for their biological data sets or research question. Recently, these two issues have been tackled by suggesting a unified description of these approaches in a well-defined mathematical framework to provide a theoretical basis for validating these approaches [6, 46], and conducting a performance evaluation of different IC-based functional similarity measures using different types of biological data. The evaluation results indicate the best functional similarity measure for each term IC and semantic similarity approach, thus helping researchers choose the most appropriate measure for their needs [6, 47].

In this article, we survey the different tools that have been developed in the context of GO semantic similarity measures. We first identify 14 GO semantic similarity tools, different IC models, term semantic similarity approaches and functional similarity measures implemented in each tool, and provide GO semantic similarity-based biological applications supported by each of these tools. We classify existing GO semantic similarity measures and present mathematical expressions and properties related to these measures for optimal use to enable effective biological knowledge discovery. We also provide a summary of performance evaluation results from previous studies [6, 15, 47] and suggest possible existing tools that may be used for retrieving semantic similarity scores for a given biological application or data type. This will assist researchers, software developers and end users identify characteristics specific to each tool for users to easily access their features and for software developers to identify challenges and advances made in this area to avoid redundant effort in developing objects that already exist or implementing ideas already proven to be obsolete. We focus on several important questions and issues, enabling the end users to choose the most effective tool for their biological questions and the software developers to focus on how to implement useful tools that are practical, easy to use and able to meet input requirements of current genome- and proteome-wide applications.

GO semantic similarity measures

Broadly, a term IC model and term semantic similarity approach provide scoring mechanisms to numerically quantify the specificity of terms and the closeness or relatedness and difference between terms within an ontology [6, 47]. In biological research, especially in the context of GO, these IC models and term similarity approaches have been extended to quantify semantic similarity between proteins, referred to as functional similarity between gene products (proteins) annotated with GO terms, to analyze or infer relationships between proteins at the functional level based on their GOAs. In the following sections, we briefly describe and then characterize different functional similarity measures based on the strategies adopted to compute their scores. We review some advantages and drawbacks of the semantic similarity measures according to different mathematical properties and biological performances of these measures, and we classify these similarity measures based on different properties. Finally, we focus on several critical, but largely unanswered, questions and issues associated with different measures. We hope that discussing these questions and issues will be considered when using a given measure and enable end users to choose or apply a measure that is more likely to produce unbiased results for their biological questions.

Models for computing GO semantic similarity scores: review and classification of different measures

Several IC models and term semantic similarity approaches have been introduced or updated for use in the context of GO (see Supplementary File for mathematical descriptions of different IC models and semantic similarity measures). Note that when computing semantic similarity scores, only 'is_a' and 'part_of' relationships between GO terms are often used. From their conception, term IC models can be divided into two families [6]: annotation- and topology-based IC families. While the topology-based family exploits only the intrinsic topology of the GO DAG, the annotation-based family also requires the addition of (protein) annotation data. With the exception of the topology-based model proposed by Wang *et al.* [14], all other approaches compute the IC of terms in a similar way (i.e. using log function) despite their conceptual differences. The IC value of a term x is generally calculated as follows:

$$IC(x) = -\ln(p(x)) \quad (1)$$

where p(x) is the relative frequency of the term x for the annotation-based family or the topological feature of the term x modeling term specificity and depends on the model under consideration. Models include GO-universal [19], Zhang *et al.* [10], Seco *et al.* [48] or Sánchez *et al.* [49] (see Supplementary File for different mathematical expressions of these models). Note that the Seco *et al.* and Zhang *et al.* models have a similar concept of term specificity and they often fail to distinguish different term-specificity levels, which may bias IC values produced. The Zhou *et al.* [50] and Seddiqui *et al.* [51] models attempted to correct these potential biases by weighing IC values and term depths or number of relations a term has. Meng *et al.* [52] introduced a model capturing the term specificity without depending on a tuning factor. See section 1 of the Supplementary File for mathematical expressions of these IC models, which are shown in Figure 1.

Broadly speaking, there exist three main categories of GO term similarity approaches: edge- (or path-), IC-based (or node-based) and 'hybrid' edge-node based. In edge-based approaches, semantic similarity score is a function of the number of edges (or nodes) on paths connecting terms. Those depending on term IC values are referred to as node- or IC-based, whereas models combining edge- and node-based approaches to compute scores between terms are known as hybrid-based models. Considering the large number of term semantic similarity models that have been suggested, we only provide a unified formula from which most of edge- and IC-based similarity models can be retrieved. We refer the interested reader to the Supplementary File for explicit mathematical expressions of these term semantic similarity models. A unified framework, which captures most of the node- and edge-based term semantic similarity measures, is given by the following formula [6]:

$$S(x,y) = \frac{\varepsilon\mu(A_x \cap A_y)}{\alpha\mu(A_x \cap A_y) + \beta\mu(A_x) + \zeta\mu(A_y)} \quad (2)$$

where ε is the correction factor reducing the impact of bias in scoring term commonality $\mu(A_x \cap A_y)$ with $0 < \varepsilon \leq 1$, which may depend on common ancestors between x and y. A_z is the subsumer of the term z, i.e. $A_z = AU\{z\}$ with A the set of ancestors of the term z. α, β, $\zeta \geq 0$ are three free parameters with $\alpha + \beta + \zeta \geq 1$ and the symmetry property of the measure can be tuned according to the parameters β and ζ. The measure of the description of terms x and y given by $\alpha\mu(A_x \cap A_y) + \beta\mu(A_x) + \zeta\mu(A_y)$ defines a normalization model, which depends on parameters α, β and ζ.

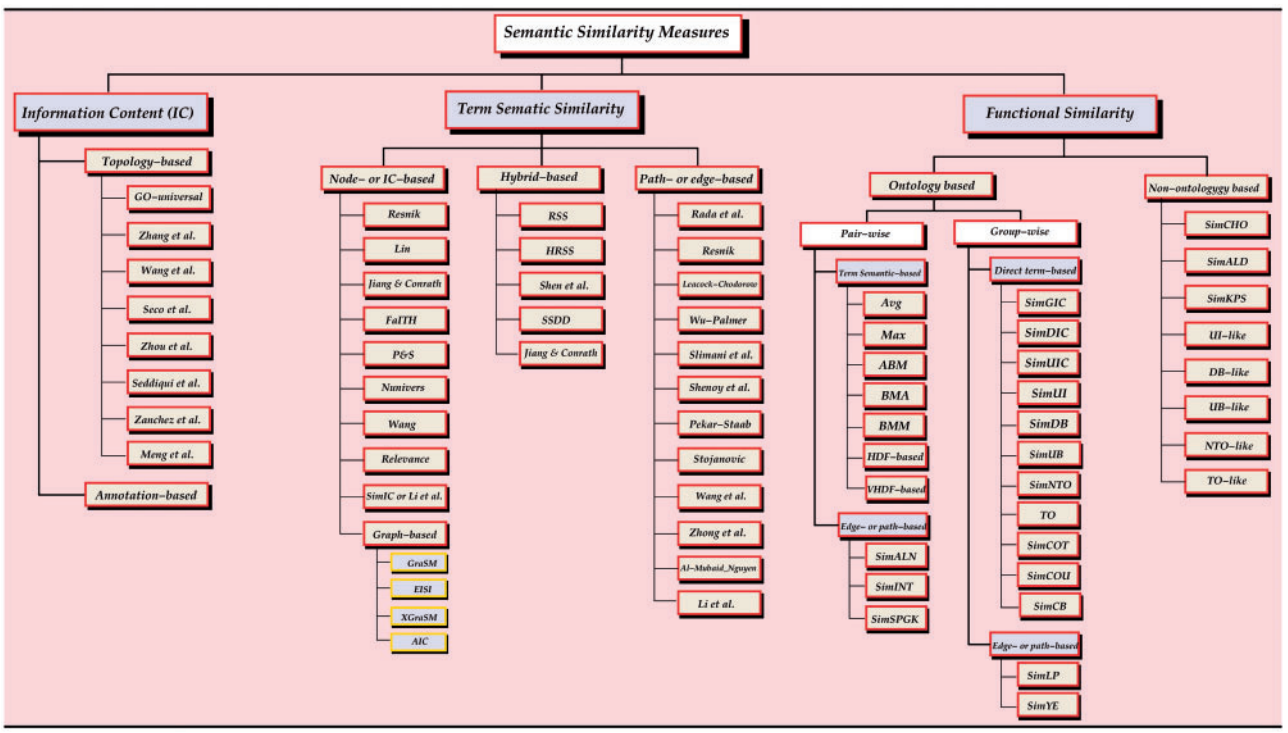

Figure 1. Classification of semantic similarity measures. Flowchart of different measures classified according to their conception. IC values are produced using topology- or annotation-based models. Using term IC values or path length/distance (shortest path length) between two terms (nodes) and depth of terms in the structure lead to different term semantic similarity and functional similarity models based on the structure of the ontology. There also exist functional similarity measures, which do not consider the structure of the ontology (non-ontology structure-based measures). A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

As an illustration, if we define $\mu(A_x) = \max\{IC(t)$: for t in $A_x\} = IC(x)$ and set $\alpha = 0$, $\beta = \zeta = 0.5$ and $\varepsilon = 1$ we get the Lin-based term similarity model, which is given by:

$$S(x,y)=\frac{2*\mu(A_x\cap A_y)}{\mu(A_x)+\mu(A_y)}=\frac{2*IC(c)}{IC(x)+IC(y)}$$

where $IC(c) = \mu(A_x \cap A_y) = \max\{IC(t)$: for t in $A_x \cap A_y\}$ and c is clearly the most informative common ancestor for terms x and y.

On the other hand, functional similarity measures are models that are used to quantify protein pairwise functional similarity scores based on their GOAs. They can be partitioned into two main classes based on the strategy adopted to quantify the similarity scores between proteins: ontology structure and non-ontology structure-based measures. Ontology structure-based measures are those that consider the structure of the ontology, while the non-ontology structure measures are those quantifying the similarity score based on the protein annotation data set only, without referring to or knowledge of the ontology from which these annotations are structured. Figure 1 provides a general overview of semantic similarity measure classification and Figure 2 describes these semantic similarity measures in terms of inputs that each category requires for calculating scores.

The ontology structure-based functional similarity measures use the concept of term IC and ontology edge or node (term) counts. In this category, there are two types of measures, pairwise and group-wise (or set-based) measures. Pairwise measures combine the semantic similarity scores between terms annotating these proteins or sets of GO terms using basic statistical measures of closeness (mean, max, min, etc.). These include Best Match Average [15, 19], Best Match Maximum [13], Average Best Matches [14, 18], Average [8] and Maximum [9], denoted BMA, BMM, ABM, Avg and Max, respectively. These measures are also known as term semantic-based (or non-direct, pairwise or indirect) measures. In this category of measures, special measures derived from term distance scores from the Hausdorff (HDF) distance [53–55] have been suggested [17, 56], and are implemented in several semantic similarity tools [57–59]. Mathematical expressions of these different measures can be found in section 3.1 of the Supplementary File and for any two proteins p and q, we get:

$$\mathrm{Avg}(p,q) \le \mathrm{BMA}(p,q) \le \mathrm{BMM}(p,q) \le \mathrm{Max}(p,q) \qquad (3)$$

This suggests that these measures often produce different scores with the Max measure producing the largest and the Avg measure giving the lowest functional similarity scores. There are also pairwise edge-like measures, including SimALN suggested by Ali-Mubaid and Nagar [60, 61], which uses the concept of the average shortest path length between pairwise GO terms annotating proteins, IntelliGO, introduced by Benabderrahmane *et al.* [62], referred to as SimINT, and the shortest path graph kernel (spgk), or SimSPGK, used by Alvarez *et al.* [63]. The SimALN measure depends on a scaling parameter adjusting the contribution of the average path length to the score produced (see Supplementary File, section 3.2, for more details).

In general, these statistical measures of closeness, including max, min and mean, are known to be sensitive to scores that lie at abnormal distances from the majority of scores, or outliers. This means that pairwise measures, such as Avg, Max, etc. may produce biases, which affect functional similarity scores [19], as illustrated in [15]. Thus, other functional similarity measures, such as SimGIC [15], SimDIC, SimUIC [5, 19, 47] and Cosine [41, 64], which use term-specificity scores directly (e.g. IC) to compute functional similarity scores from their GOAs, were introduced. SimGIC, SimDIC and SimUIC use the Jaccard index [65]. The Cosine measure uses a normalized dot product (usual and Tanimoto normalization models: See Supplementary File, section 3.3, for more information) to estimate functional similarity scores by defining feature vectors of length equal to the total number of terms annotating the two proteins considered. Each vector component describes the absence or presence of a term in the set of terms annotating the protein and is weighted by its IC value. These protein functional similarity measures, except the Cosine measure, which uses a usual normalization model, are unified based on the Tversky ratio framework given by [65, 66]:

$$\mathrm{Sim}_{Tversky}(p,q)=\frac{m(A_p\cap A_q)}{m(A_p\cap A_q)+\alpha m(A_p-A_q)+\beta m(A_q-A_p)} \qquad (4)$$

where $m(A)$ is a measure scoring the feature of the set A, which is the set of properties describing A, with $m(\phi) = 0$ and $m(A) \le m(B)$ if $A \subseteq B$, A_ω the set of terms annotating the protein ω including ancestors of these terms, $\alpha, \beta \ge 0$ are two free parameters in the score of the description of proteins p and q given by $m(A_p \cap A_q)+\alpha m(A_p - A_q)+\beta m(A_q - A_p)$. This defines a normalization model that is a function of these parameters for a given

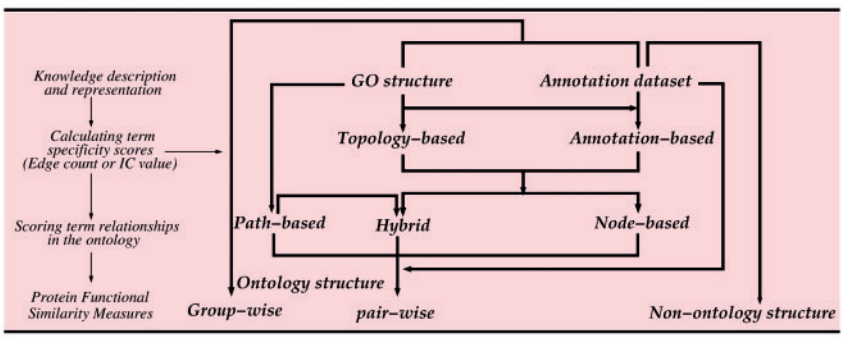

Figure 2. Summary of GO semantic similarity models and inputs. General description of term IC, term semantic similarity and functional similarity models and inputs required to compute their scores (see Figure 1 for their classification). The arrows point to models requiring elements from the arrow source to calculate concept scores, e.g. for a group-wise functional similarity measure under an ontology structure-based model, one needs protein GO annotation, GO structure and IC score datasets.

measure. Note that the symmetry property of a measure can be tuned according to these parameters.

In general, the following two measures are used to define m, given a set *A*:

$$m_2(A) = < s_A, s_A > \text{ and } m_1(A) = < s_A, 1_A > \quad (5)$$

where $< >$ is the inner (dot) product, s_A the vector of length $|A|$, the number of elements in the set A, having IC values of terms in the set A as components and 1_A is a vector of equal length to $|A|$ with all components set to 1. Note that all functional similarity measures use m_1, except the Cosine-based (SimCOU and SimCOT) measures, which use m_2 and we can see that for two sets *A* and *B*, we have:

$$\begin{aligned} &2*m(A \cup B) + m(A - B) + m(B - A) = m(A) + m(B) \text{ and} \\ &m(A \cup B) = m(A \cap B) + m(A - B) + m(B - A) = m(A) + m(B) - m(A \cap B) \end{aligned} \quad (6)$$

For example, using the relation (4), with $\alpha = \beta = 1$ and for a given protein p, defining $m(A_p) = m_1(A_p)$ *by*:

$$m(A_p) = < s_{A_p}, 1_{A_p} > = \sum_{t \in A_p} \mathrm{IC}(t)$$

with s_{Ap} the vector whose components are IC values of terms annotating the protein p, we get the SimGIC measure, which is given by:

$$\begin{aligned} \mathrm{SimGIC}(p,q) &= \frac{m_1(A_p \cap A_q)}{m_1(A_p \cap A_q) + m_1(A_p - A_q) + m_1(A_p - A_q)} \\ &= \frac{m_1(A_p \cap A_q)}{m_1(A_p \cup A_q)} = \frac{\sum_{t \in A_p \cap A_q} \mathrm{IC}(t)}{\sum_{t \in A_p \cup A_q} \mathrm{IC}(t)} \end{aligned}$$

Similarly, we obtain SimCOT by setting $\alpha = \beta = 1$ (using $m_2(A)$), and SimDIC by taking $\alpha = \beta = 0.5$. In addition, by setting $\alpha = 1$ and $\beta = 0$ we get the degree of inclusion (SimNTO) or extension (SimUIC) for A_p, representing the proportion of A_p overlapping with or overshadowing A_q. The usual normalization model used in the context of SimCOU is given by $\sqrt{m(A) * m(B)}$and because $m(A) \geq m(A \cap B)$ and $m(B) \geq m(A \cap B)$, it follows that:

$$\sqrt{m(A) * m(B)} \leq \frac{m(A) + m(B)}{2} \leq \max\{m(A), m(B)\} \leq m(A \cup B) \quad (7)$$

This means that for two proteins p and q annotated with GO terms, we have:

$$\mathrm{SimCOU}(p,q) \geq \mathrm{SimDIC}(p,q) \geq \mathrm{SimUIC}(p,q) \geq \mathrm{SimGIC}(p,q) \quad (8)$$

This indicates that depending on the normalization model used, different scores are produced with SimCOU producing largest scores and SimGIC yielding the smallest. Other types of functional similarity measures, which use the structure of the ontology and depend on the number of edges or term depths, referred to as group-wise edge-like measures, include the SimLP measure [67]. The SimLP measure computes the functional similarity scores between two proteins as the longest path length in the intersection graph produced by the two sub-graphs derived from GO terms annotating the two proteins under consideration. The SimLP measure is not normalized, but Ye *et al.* [35] suggested a normalized version, denoted SimYE, by considering the maximum and minimum depths in the ontology under consideration.

Other ontology structure-based measures are particular cases of the SimGIC, SimDIC and SimUIC measures by assigning an equal IC score to all terms in the GO-DAG. For example, the SimUI measure [67], which refers to the union-intersection protein similarity measure, is a particular case of SimGIC with equal IC values assigned to all GO terms [19, 47]. Similarly, one can define particular cases based on SimDIC (Dice) and SimUIC (Universal) [5], denoted by SimDB and SimUB, respectively. A variant of SimUB, known as normalized term overlap (TO) and referred to as SimNTO, has been proposed [16, 68, 69]. The SimUI, SimDB, SimUB and SimNTO measures are normalization models of the TO measure and are thus equivalent. The only difference between them is the normalization scheme used by each of these measures (see Supplementary File, section 3.3, for more details). For these specific cases, $m_1 = m_2 = m$ and a given set A by:

$$m(A) = < 1_A, 1_A > = |A| \quad (9)$$

This suggests that these measures are particular cases where IC values are 1 for all terms. In this context, SimUI corresponds to the Tanimoto normalization model-based Cosine similarity measure (SimCOT). Moreover, using the usual normalization schemes can lead to another measure, denoted SimCB, which is equivalent to SimUI, SimDB, SimUB and SimNTO. As for relation (8) and knowing $\min\{m(A), m(B)\} \leq \sqrt{m(A) * m(B)}$, it is clear that for given two proteins p and q, we have:

$$\mathrm{SimNTO}(p,q) \geq \mathrm{SimCB}(p,q) \geq \mathrm{SimDB}(p,q) \geq \mathrm{SimUB}(p,q) \geq \mathrm{SimUI}(p,q) \quad (10)$$

showing that SimNTO produces the largest scores and SimUI the smallest. It is worth mentioning that even though assigning an equal IC value to all terms in the ontology is not a realistic assumption in the context of the GO DAG [5, 19, 47], this category of measures can still be used as alternative measures in practice as the SimUI measure, for example, showed relatively good performance when applied to different biological data [47].

The non-ontology structure-based measures include Cho *et al.* [70], Ali and Diane [71] and Kappa-Statistics [72] functional similarity measures, referred to as SimCHO, SimALD and SimKPS, respectively (see Supplementary File, section 3.5, for more details). Other types in this category of measures can be derived from SimUI, SimUB, SimDB and SimNTO when used with only the terms annotating the two proteins, without considering other terms in the two sub-graphs of the GO terms. One such measure is the TO-like functional similarity measure, which is denoted TO-like and was introduced by Lee *et al.* [25]. Here, the functional similarity of a protein pair is scored simply by the number of terms shared by the proteins and a score of zero is assigned when there is no term shared. The normalized versions of this TO-like measure are denoted NTO-, UI-, UB- and DB-like measures. The UB-like measure computes the average number of matched GO terms between protein pairs. In addition, a variant of the TO-like measure exists, which assigns a score of 1 if the two proteins share at least one term and 0 otherwise. It is important to know that these non-ontology structure-based measures can only be recommended when it is established that the ontology under consideration is incomplete and terms occurring in the data set, but missing in the ontology, cannot be removed.

Questions and issues related to GO semantic similarity measures

The use of protein annotation data sets and GO structure

Protein annotation data sets are used for computing term IC values in the context of annotation-based family and functional similarity scores. However, this makes the calculation of term IC values dependent on the corpus (annotation data set), whereas a term in the GO structure is expected to have a unique IC value, which should be independent of the corpus under consideration [6]. In this case, the performance of the associated-term semantic similarity approach and functional similarity measure will depend on the corpus under consideration because of its dependence on the frequencies of GO term occurrences in the corpus. Annotations may be unbalanced in their distribution across the GO structure, thus compromising these annotation-based approaches, specifically for organisms with sparse GOAs. This may negatively affect their performance [19]. It has been suggested that the use of the GOAs provided by the GOA project [2, 73–75] can solve this issue [5, 6], but the shallowness of annotation artifacts will still persist when comparing pairs of proteins annotated with few terms [16]. Thus, it is recommended to use topology-based IC family methods to produce a fixed and well-defined IC value for a given GO term, independent of the corpus under consideration, using only the GO structure. Even in the case of functional similarity measures, it is important to make use of the GO structure. For example, it was proven that the group-wise-based measures, such as SimGIC, perform better than the pairwise-based measures, and even within the pairwise-based measures, those using the GO structure perform better than those using only the most informative common ancestor. These include measures that use Graph-based Similarity models, such as Disjunct Common Ancestors [43], known as GraSM, Exclusively Inherited Shared Information, referred to as EISI [76], eXtended GraSM, denoted XGraSM [5, 6], and Aggregate Information Content, referred to as AIC [77]. This indicates that the use of the GO structure in the calculation of semantic similarity scores can improve the outcome of further analyses. This justifies or adds value to the argument that the non-ontology structure measures, such as SimCHO, SimALD and SimKPS, should be used only when necessary as pointed out previously.

Inferring or adapting a measure to the GO context—mapping a distance to a semantic similarity measure

Several semantic similarity measures are derived from a distance (metric) or a variant distance. The concept 'variant' distance is often used, as these measures do not satisfy the triangle inequality. One of the metrics (or distances) used for this purpose with about 24 different variant distances for object matching is the HDF distance. Based on their behavior in the presence of noise, the best variant distance, called the modified HDF distance for object matching, has been shown to be more robust to outliers [53]. The BMM measure (see Supplementary File, section 3.1) corresponds to the functional similarity measure derived from it. Other ontology structure-based pairwise measures, such as the BMA and ABM measures, are derived from the variant HDF measures. We have identified the four most used transformations for mapping distances or variant distances to semantic similarity measures:

i. A linear transformation $L(x) = 1 - x$ if x is a normalized score, otherwise $L(x) = d - x$ where d is the possible maximum distance value between the concept under consideration. In this case, the normalized semantic similarity score is given by:

$$L_n(x) = 1 - \frac{x}{d}$$

ii. A log transformation $L_{\log}(x) = \log(d) - \log(x)$, and normalized as follows:

$$L_{n\log}(x) = 1 - \frac{\log(x)}{\log(d)}$$

in which case, $\log(x+1)$ is used instead of $\log(x)$ to avoid $\log(0)$ and $\log(d+1)$ instead of $\log(d)$ to prevent negative semantic similarity scores.

iii. An exponential function of additive inverse of distance score x, $L_{\exp}(x) = \exp(-x)$.

iv. A simple rational function $L_r(x) = x^{-1}$, in which case its variant $L_r(x) = (1+x)^{-1}$ is used to avoid a division by zero.

Note that these different transformations lead to different scores as shown in Figure 3, indicating that a given mapping function can be overestimating or underestimating semantic similarity scores, which likely influences the performance of a given measure. Linear transformation is the most used and recommended compared with log, exponential and simple rational transformation, which are nonlinear. For example, using the linear transformation we see that the Lin approach can be derived from the Jiang and Conrath semantic distance (see Supplementary File, subsection 2.3.5). On the other hand, even though these different nonlinear transformations force the semantic similarity measures derived to satisfy mathematical properties of a semantic similarity measure [36], namely non-negativity, symmetry, normalized, reflexivity (i.e. $S(a,a) = 1$), identity of indiscernible (i.e. $S(a,b)$ reaches the maximum value of 1 if and only if $a = b$) and integrity (i.e. $S(a,b) \leq S(a,a)$), these nonlinear transformations do not effectively reflect the distance measures from which they are derived, i.e. how close or similar two concepts are does not directly provide an indication of how distant or dissimilar these concepts are. This is mostly owing to the fact that the triangle inequality (sub-additivity) property of a distance is a linear property, suggesting that the use of a

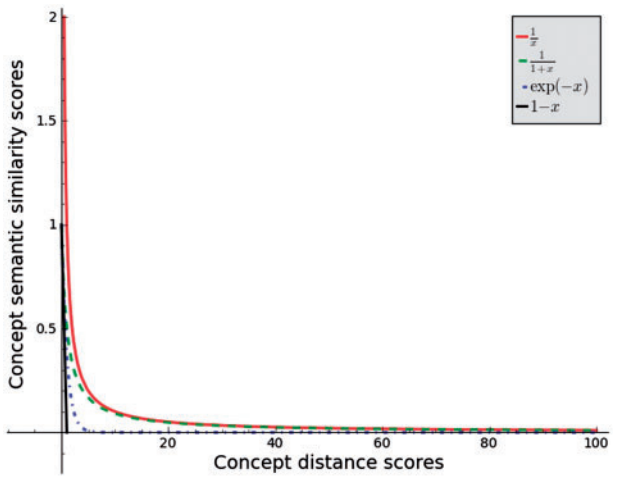

Figure 3. Graphical representation of semantic distance and similarity scores. Comparing different semantic similarity scores obtained from semantic distance scores using linear transformation for a normalized distance, exponential and rational functions for transforming semantic distance scores. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

nonlinear transformation may not necessarily satisfy this triangle inequality property, and thus may not define a distance. This holds especially for the currently used nonlinear transformations. It is worth mentioning the strategy used by Jeang and Chen [58] to derive a semantic similarity measure from a variant HDF-like distance suggested by Lerman and Shakhnovich [56] and implemented under the KU-GOAL tool. This Jeang and Chen model (see Supplementary File, section 3.1) possibly originated from the fact that BMM, BMA and ABM measures have similar patterns to variants of the HDF-based distance from which they can be derived, but this is not the case for this specific variant distance measure. Even though this measure may satisfy mathematical properties of a semantic similarity measure, it cannot in any way map or characterize the variant distance measure from which it has been derived. They are only comparable in terms of patterns and there is no mathematical relation linking them, eliciting a need for a biological assessment to determine how well it performs compared with other measures.

Impact of correction factors and normalization models

As indicated previously, normalization models from relations (2) and (4) depend on parameters α, β and ζ, and different correction or tuning factors from relation (2) yield different models producing different scores. It has been shown that the correction factor and normalization model impact the performance of a given semantic similarity measure [6]. The correction or tuning factors for GraSM, EISI and XGraSM, for example, are explicitly expressed (see Supplementary File, section 2.1.2) and simply adjust the contribution of control parameters involved in the semantic similarity score assessment. These factors can depend on the concepts under consideration in the ontology, their positions in the structure of the ontology (ancestors, edges connected to them, etc.) or can be set by the user. For example, the Shortest Semantic Differentiation Distance (SSDD) [78], Information Coefficient (SimIC) [79] and Relevance [13] and graph-based similarity approaches, including GraSM [43], EISI [76] and XGraSM [5, 6], use a contribution factor, which depends on the term positions in the topology, whereas in the Wang *et al*. [14], Nguyen and Al-Mubaid [80] and Li *et al*. edge-based [81] approaches, this contribution factor must be set by the user. In each model, optimal values of different factors were suggested, which provide good performance for an empirical finding for a specific setting. However, these values lack a theoretical or mathematical basis, thus cannot be generalized, and more importantly, it is not clear for which values of these semantic factors the semantic similarity measure yields the best performance independently of the biological application. Note that SimIC and Relevance correction factors (see Supplementary File, subsection 2.1.2) should be considered with caution, as they may violate reflexivity, identity of indiscernible and integrity properties of semantic similarity measures.

Several models have been suggested for normalizing a semantic similarity measure. All these models have a common strategy consisting of multiplying the unnormalized score by an inverse value of a greater similarity score that can be derived from the two concepts for which the semantic similarity score is being calculated. The most used values are the maximum, average and minimum scores between the two concepts. For example, the Resnik, Nunivers, GO universal, Wang *et al*., SimGIC, SimUIC and SimYE models use the maximum score, the Lin, SimDIC and SimCB measures use the average approach and the minimum score is used by the SimNTO measure.

When normalizing a semantic similarity measure, it is worth mentioning that the use of the highest score as for the Resnik approach [82] should be considered with caution [6]. The same for a normalization factor, which does not depend on the two concepts under consideration, as it may violate the reflexivity and integrity properties of semantic similarity measures. In addition, adding an excess value, δ, to the fraction components (e.g. to avoid log of or division by zero) may lead to overestimation ($\delta > 0$) or underestimation ($\delta < 0$) of the semantic similarity scores produced. This shows that the contribution of the root of the ontology should be reduced to zero, otherwise it obviously leads to overestimating similarity scores as these roots are assumed to be meaningless with the lowest term IC value, which is zero for the IC models using log function in their conception. As an illustration, when computing functional similarity scores or ontology structure-based group-wise measures, such as SimUI, SimDB, SimUB and SimNTO, or term semantic similarity scores using a graph-based enhancement strategy (XGraSM), only informative common ancestors should be used. This means roots of ontologies should be removed from the intersection and the union of ancestors [68]. Otherwise, this mathematically tends to overestimate similarity scores because, given two real numbers a and b, such that $0 < a \leq b$, we have:

$$\frac{a}{b} \leq \frac{a+\delta}{b+\delta} \text{ for any real } \delta \geq 0$$

Theoretically, this means that removing roots in these computations may improve the performance of these measures, as adding roots obviously leads to over-estimating similarity scores, as these roots are assumed to be meaningless. Including the root of the ontology in the calculation may therefore bias the similarity scores produced.

Influence of GO evidence codes

Several annotation pipelines, including manual and electronic pipelines, were designed to predict or assign GOAs to proteins [40] with associated confidence scores embodied through evidence codes indicating how these annotations were assigned. The GOA project integrates these protein annotations into a single set of high-quality electronic and manual associations (annotations) of GO terms to UniProt Knowledgebase (UniProtKB) entries. Electronic annotations represent more than 99% of the GOA data set (http://www.ebi.ac.uk/GOA/uniprot_release) and they receive an Inferred from Electronic Annotation (IEA) evidence code. This is assumed to be the lowest confidence, as annotations produced are not manually curated or experimentally validated. In the context of GO semantic similarity measures, some researchers have removed IEAs from the calculations of the semantic similarity scores. The SimINT measure [62] takes into account the confidence level of different GO evidence codes, thus enabling a user to weigh evidence codes associated with GO terms annotating the proteins. However, as the IEA evidence code is highly dominant in protein annotations, the setting of evidence code weight is likely to have a negligible impact. IEAs completely dominate manual annotation in terms of number of annotations available for a particular genome and this is the most likely future trend [40]. The use of all evidence codes equally has shown to yield an improved measure performance [41, 42], confirming that it is better to make use of all the GO evidence codes. In fact, the IEA annotations are becoming more and more accurate mainly because most of the annotation data sets are generated from conversion maps, namely

SPKW2GO, SPSL2GO, EC2GO, HAMAP2GO, UniPathway2GO and InterPro2GO, which themselves are manually curated to ensure high-accuracy annotations from the electronically inferred GOA set [40].

Exploring existing GO semantic similarity tools

Considering the large number of GO semantic similarity measures that have been introduced, it is tedious for an end-user or tool developer to identify particular features of a given tool without an appropriate summary of all available tools. This is particularly important, as for some tools, such as GossTo [83] and FuSSMeg [12, 43], it is unclear how term semantic similarity scores are combined to produce functional similarity scores between genes or proteins. In general, it is not clear how existing tools deal with some issues related to semantic similarity measures as discussed in the previous section. In this section, we describe the general features of existing tools in terms of IC models, term semantic similarity approaches and functional similarity measures, and discuss limitations associated with these tools. Thus, readers can more easily grasp the key spirit of the 14 existing tools without trying to search through each tool's Web site, with often limited documentation, to determine its features.

Features related to different GO semantic similarity tools

Several tools have been developed for producing GO term and protein semantic similarity scores [41, 42] to facilitate the use of different semantic similarity approaches for different biological data types. These include online web tools and software packages implemented in R, Python and Java programming languages. Tables 1 and 2 list each tool with the IC models, term semantic similarity approach and functional similarity measures it supports. Table 3 provides a mapping between different notations as used by different tools. Approximately 14 GO semantic similarity tools that are currently available in the community are collected in this survey. Tools are uniquely characterized according to their underlying GO semantic similarity measures and associated models implemented. The comprehensive collection, unique tool feature identification and associated questions or issues provide a comprehensive and up-to-date view on the advantages, issues and recent trends. This survey will help tool designers or developers and experienced end users understand the underlying features and pertinent details of each tool, enabling them to make the best choices for their particular research interests.

Most existing tools do not implement the IC topology-based model, except GOSemSim [86], SML [87], A-DaGO-Fun [59], DaGO-Fun [5], FastSemSim (https://sourceforge.net/p/fastsemsim/home/) and G-SESAME [45] implementing the Wang *et al.* approach [14]. The SML Java software, which is not specific to GO semantic similarity measures, additionally implements the Sánchez *et al.* [49], Zhou *et al.* [50] and Seco *et al.* [48] IC models, but it is not clear if the tool allows the use of these IC models for GO. Note that the Zhang *et al.* IC model [10] can be considered to be the improved non-normalized version of the Seco *et al.* IC model, as adding a positive value (+1) to the D-value may bias IC scores. This Zhang *et al.* approach is implemented in the DaGO-Fun and A-DaGO-Fun tools using the average normalization model (Lin-like normalization), which showed best performance in WordNet [48]. In addition, these two tools also implement the GO-universal approach.

GO semantic similarity tools: consideration and recommendation

Issues related to organism-based IC value

One of the main issues with the existing GO semantic similarity tools is that several tools, including FastSemSim, GOSemSim, GOssTO [83] and csbl.go [64], are organism-based tools. Regardless of species coverage, these tools may not support less

Table 1. Current GO semantic similarity web tools

Tool	GO-semantic similarity features implemented					
	Model	Approach	FSM	Input size	URL	Reference
DaGO-Fun	Wang *et al.*, Zhang *et al.*, GO-universal; Annotation-based	Wang *et al.*, Zhang *et al.*, GO-universal; Resnik, Lin, Nunivers, XGraSM, Relevance, SimIC	Avg, Max, BMA, ABM; SimGIC, SimDIC, SimUIC, SimUI	3000	http://web.cbio.uct.ac.za/ITGOM/	[5]
FunSimMat	Annotation-based	Resnik, Lin, Jiang, Relevance	Avg, Max, BMM; SimGIC, SimUI, SimNTO	Unlimited	http://www.funsimmat.de/	[68, 84]
FuSSiMeg	Annotation-based	Resnik, Lin, Jiang, GraSM	Max	1	http://xldb.di.fc.ul.pt/rebil/ssm/	[12, 43]
G-SESAME	Wang *et al.*, Annotation-based	Wang *et al.*, Resnik, Lin, Jiang, AIC	ABM	1	http://bioinformatics.clemson.edu/G-SESAME/	[45]
GOToolBox	—	—	SimDB	Unlimited	http://genome.crg.es/GOToolBox/	[85]
GOssTo	Annotation-based	Resnik, Lin, Jiang, Relevance, GraSM	Max, SimGIC, SimUI	Unlimited	http://www.paccanarolab.org/gosstoweb/	[83]
KU-GOAL	Annotation-based	Resnik, Lin, Jiang, Relevance	ABM, HDF, VHDF; TO–like, UB–like	Unlimited	http://www.ittc.ku.edu/chenlab/goal/index.php	[58]
ProteInOn	Annotation-based	Resnik, Lin, Jiang, GraSM	BMA, SimGIC, SimUI	1000*	http://lasige.di.fc.ul.pt/webtools/proteinon/	[44]

Existing GO semantic similarity tools for computing semantic similarity score and some of their features are provided. FSM stands for functional similarity measure(s) and input size provides acceptable number of protein pairs. An asterisk on a given number indicates acceptable number of proteins and in this case, protein functional similarity scores of all protein pairs built from the provided set or list of proteins are computed. '—' indicates that the semantic similarity measure is not supported by the tool.

Table 2. Current GO semantic similarity software packages

Tool	Format	GO-semantic similarity features implemented				
		Model	Approach	FSM	URL	Reference
A-DaGO-Fun	Python	Wang *et al.*, Zhang *et al.*, GO-universal; Annotation-based	Wang *et al.*, Zhang *et al.*, GO-universal; Resnik, Nunivers, Lin, Relevance, SimIC, XGraSM	Avg, Max, BMA, BMM, ABM, HDF, VHDF; SimGIC, SimDIC, SimUIC, SimUI, SimCOU, SimCOT, SimDB, SimUB, SimNTO	http://web.cbio.uct.ac.za/ITGOM/adagofun	[59]
csbl.go	R	Annotation-based	Resnik, Lin, Jiang, GraSM, Relevance	BMM, SimGIC, SimDB, SimCOU, SimKPS	http://csbi.ltdk.helsinki.fi/csbl.go/	[64]
FastSemSim	Python	Wang *et al.*; Annotation-based	Wang *et al.*; Resnik, Lin, Jiang, Relevance, SimIC	Avg, Max, BMA, SimGIC, SimUI, SimCB, SimDB, SimUB, SimNTO, TO	https://sourceforge.net/p/fastsemsim/home/	—
GOSim	R	Annotation-based	Resnik, Lin, Jiang, Relevance, GraSM	Avg, Max, BMA, BMM, HDF, SimCOU, SimCOT	http://www.bioconductor.org/packages/release/bioc/html/GOSim.html	[57]
GOssTo	Java	Annotation-based	Resnik, Lin, Jiang, Relevance, GraSM	Max, SimGIC, SimUI	http://www.paccanarolab.org/gosstoweb/	[83]
GOvis	R	—	—	SimLP, SimUI	http://bioconductor.org/packages/2.3/bioc/html/GOstats.html	[67]
SemSim	R	Wang *et al.*; Annotation-based	Wang *et al.*; Resnik, Lin, Jiang, Relevance	Avg, Max, ABM, BMM	http://bioconductor.org/packages/2.6/bioc/html/GOSemSim.html	[86]
SML	Java	Wang *et al.*, Sanchez *et al.*, Zhou *et al.*, Seco *et al*; Annotation-based	Wang *et al.*; Resnik, Lin, Jiang, Relevance, SimIC, GraSM; Wu *et al.*, Pekar *et al.*, Leacok *et al.*, Li *et al.*, Rada	Avg, Max, BMA, BMM; SimGIC, SimALD, SimLP, TO-like	http://www.semantic-measures-library.org	[87]

Current GO semantic similarity software packages for computing semantic similarity score and some of their features are provided. FSM stands for functional similarity measure(s) and '—' indicates that the semantic similarity measure is not supported by the tool or there is no paper published for tool.

popular species and are limited for newly annotated organisms. This argument holds for all existing tools, except A-DaGO-Fun, as they only work for proteins in the GOA data set [59]. Furthermore, as pointed out previously, the use of organism-based IC values raises several issues, including (i) non-uniqueness and inconsistency of term IC values, remarkably for unbalanced annotation across the GO structure, (ii) persistence of annotation shallowness, especially when comparing pairs of proteins annotated with few terms, and (iii) no generalizable and generally inconclusive performance evaluation results. Even though the use of the whole set of annotations provided by the GOA-UniProtKB project [2, 73–75] may solve this problem, this is only at the cost of an increase in the running time and the complexity of these annotation-based approaches. This is expected to worsen as the number of protein annotations increases, which would potentially hamper the performance of these approaches in their running time, as processing the annotation file would take a lot of time before being able to compute the IC values [47]. Thus, it is recommended to precompute GO term IC values to enable rapid response to user queries as applied in many tools, including GOSim, FunSimMat, DaGO-Fun and A-DaGO-Fun. Furthermore, it is worth considering implementing topology-based approaches to ensure IC values produced are fixed and well-defined for a given GO term and independent of the corpus under consideration.

When retrieving GO semantic similarity scores, one needs to keep in mind that GO is incomplete and its structure is unbalanced. GO is updated daily to incorporate new knowledge as it accumulates, which could result in changes in term IC values along the updated branches. However, none of the GO semantic similarity tools update their data sets daily to accommodate daily modifications of the GO structure. This raises the need for analyzing the impact of the change of the GO structure on the term IC and semantic similarity values. Moreover, the GO structure is unbalanced and as such, terms at the same level (depth) do not necessarily have the same specificity, and edges at the same level do not necessarily represent the same distance. This makes edge- or path-based models ineffective [41] in the context of GO, as they assume that nodes or terms at the same levels have the same semantic distance to the root of the hierarchy, producing a biased semantic similarity between terms. In general, models incorporating ancestors and descendant term features in their conception were shown to perform well in practice [6, 15, 47]. Finally, while recent tools, including DaGO-Fun and FunSimMat, have collectively made substantial progress, compared with earlier tools, such as FuSSiMeg, especially in terms of input size, most of these tools are not robust enough for exploring GO semantic similarity measures in the context of high-throughput functional analysis.

Cross comparison and implementing redundant and possible missing semantic similarity measures

Despite the wide range of GO semantic similarity measures and the existence of several approaches to meet requirements of these applications, there is no tool available that integrates all these possible semantic similarity measures. Even though the

Table 3. Mapping between different GO semantic similarity measures and their corresponding labels in the tools identified

FSM	A-DaGO-Fun	DaGO-Fun	KU-GOAL	GOssTO	FunSimMat	G-SESAME	ProteInOn	FuSSiMeg	GOToolBox	SML	FastSemSim	GOSemSim	csbl.go	GOSim	GOvis
BMA	bma	x	–	–	–	–	x	–	–	x	BMA	–	–	funSimAvg	–
BMM	bmm	–	–	–	O	–	–	–	–	x	–	rcmax	O	funSimMax	–
ABM	abm	x	AveMax	–	–	O	–	–	–	–	–	BMA	–	–	–
Avg	avg	x	–	–	avg	–	–	–	–	AVERAGE	Average	avg	–	mean	–
Max	max	x	–	O	max	–	–	O	–	MAX	x	max	–	max	–
HDF	hdf	–	HdfDist[1]	–	–	–	–	–	–	–	–	–	–	hausdorff	–
VHDF	vhdf	–	AveNMS	–	–	–	–	–	–	–	–	–	–	–	–
SimGIC	gic	x	–	simGIC	GIC	–	x	–	–	GIC	x	–	Weighted Jaccard	–	–
SimDIC	dic	x	–	–	–	–	–	–	O	–	–	–	–	–	–
SimUIC	uic	x	–	–	–	–	–	–	–	–	–	–	–	–	–
SimCOU	cou	–	–	–	–	–	–	–	–	–	–	–	Cosine	(dot, sqrt)	–
SimCOT	cot	–	–	–	–	–	–	–	–	–	–	–	–	(dot, Tanimoto)	–
SimLP	–	–	–	–	–	–	–	–	–	LP	–	–	–	–	LP
SimUI	ui	x	–	simUI	UI	–	x	–	–	–	Jaccard	–	–	–	UI
SimDB	db	–	–	–	–	–	–	–	–	–	Dice	–	Czekanowski-Dice	–	–
SimUB	ub	–	–	–	–	–	–	–	–	–	–	–	–	–	–
SimCB	–	–	–	–	–	–	–	–	–	–	Cosine	–	–	–	–
SimNTO	nto	–	–	–	NTO	–	–	–	–	–	SimNTO	–	–	–	–
TO	–	–	–	–	TO	–	–	–	–	–	SimTO	–	–	–	–
TO-like	–	–	Match[1]	–	–	–	–	–	–	LEE	–	–	–	–	–
UB-like	–	–	AveMatch	–	–	–	–	–	–	–	–	–	–	–	–
SimALD	–	–	–	–	–	–	–	–	–	ALI_DEANE	–	–	–	–	–
SimKPS	–	–	–	–	–	–	–	–	–	–	–	–	Kappa	–	–
GO-universal	u	x	–	–	–	–	–	–	–	–	–	–	–	–	–
Wang *et al.*	w	x	–	–	–	O	–	–	–	HYBRID_WANG	GSESAME	Wang	–	–	–
Zhang *et al.*	z	x	–	–	–	–	–	–	–	–	–	–	–	–	–
Resnik	r	x	x	x	Res	–	x	x	–	NODE_RESNIK	x	x	x	x	–
Lin	l	x	x	x	x	–	x	x	–	NODE_LIN	x	x	x	x	–
Nunivers	n	x	–	–	–	–	–	–	–	–	–	–	–	–	–
Jiang and Conrath	–	–	Jiang	Jiang	x	–	x	x	–	NODE_JIANG_CONRATH	Jiang-Conrath	Jiang	x	x	–
Relevance	s	SimRel	Schlicher	–	simRel	–	–	–	–	NODE_SCHLICHER	SimRel	Rel	x	relevance	–
SimIC	li	x	–	–	–	–	–	–	–	–	x	–	–	–	–
GraSM	–	–	–	x	–	–	DCA	x	–	–	–	–	x	Couto	–
XGraSM	(x)	x	–	–	–	–	–	–	–	–	–	–	–	–	–
Resnik edge-based	–	–	–	–	–	–	–	–	–	EDGE_RESNIK	–	–	–	–	–
Rada *et al.*	–	–	–	–	–	–	–	–	–	EDGE_RADA	–	–	–	–	–
Wu&Palmer	–	–	–	–	–	–	–	–	–	EDGE_WU_PALMER	–	–	–	–	–
Slimani *et al.*	–	–	–	–	–	–	–	–	–	EDGE_SLIMANI	–	–	–	–	–
Stojanovic *et al.*	–	–	–	–	–	–	–	–	–	EDGE_STOJANOVIC	–	–	–	–	–
Li *et al.* edge-based	–	–	–	–	–	–	–	–	–	EDGE_LI	–	–	–	–	–
Pekar&Staab	–	–	–	–	–	–	–	–	–	EDGE_PEKAR_STAAB	–	–	–	–	–
Leacok&Chodorow	–	–	–	–	–	–	–	–	–	EDGE_LEACOCK_CHODOROW	–	–	–	–	–

Known and currently used GO semantic similarity measures with their corresponding labels in the existing tools. '–' indicates that the semantic similarity measure is not supported by the tool while 'O' indicates that the measure may possibly be supported, but the notation is not provided, 'x' means that the tool uses the same notation as indicated and (x) indicates that the tool uses x as a symbol of the measure. A couple in the case of GOSim indicates the measure (method) and the normalization model used and the superscript 1 indicates a special variant of the measure is implemented.

Table 4. Mapping biological applications (data types), semantic similarity measures and tools

Applications			Measures Suggested	Semantic Similarity Tools suggested							
				Organism-dependent				Organism-independent			
				UniProt ID	Gene name	Entrez Gene	Other	UniProt ID	Gene name	Entrez Gene	Other
Classification	Similarity-based	Sequence	Resnik-Max BMA ABM SimGIC	SML GossTo	csbl.go	SemSim	n/a	FuSSiMeg[1] FunSimMat ProteInOn DaGO-Fun A-DaGO-Fun	G-SESAME[1] FunSimMat DaGO-Fun A-DaGO-Fun	GOSim A-DaGO-Fun	A-DaGO-Fun
		Domain (Pfam)	Resnik-Max BMA ABM SimGIC	SML GOssTo	csbl.go	SemSim	n/a	FuSSiMeg[1] FunSimMat ProteInOn DaGO-Fun A-DaGO-Fun	G-SESAME[1] FunSimMat DaGO-Fun A-DaGO-Fun	GOSimA-DaGO-Fun	A-DaGO-Fun
		Enzyme Commission (EC)	SimGIC SimDIC BMA	SML GOssTo	csbl.go	SemSim	n/a	FunSimMat ProteInOn DaGO-Fun A-DaGO-Fun	FunSimMat DaGO-Fun A-DaGO-Fun	GOSimA-DaGO-Fun	A-DaGO-Fun
	Ortholog		BMA	SML	csbl.go	SemSim	n/a	FunSimMat ProteInOn DaGO-Fun A-DaGO-Fun	FunSimMat DaGO-Fun A-DaGO-Fun	GOSimA-DaGO-Fun	A-DaGO-Fun
	Protein–protein interactions		SimGICAvg	SML	csbl.go	SemSim	n/a	FunSimMat ProteInOn DaGO-Fun A-DaGO-Fun	G-SESAME[1] FunSimMat DaGO-Fun A-DaGO-Fun	GOSimA-DaGO-Fun	A-DaGO-Fun
	Gene co-expression features		SimGIC SmUIC Resnik-BMA Avg	SML GOssTo	csbl.go	SemSim	n/a	FunSimMat ProteInOn DaGO-Fun A-DaGO-Fun	G-SESAME[1] FunSimMat DaGO-Fun A-DaGO-Fun	GOSimA-DaGO-Fun	A-DaGO-Fun
	Categorization		XGraSM-Resnik Resnik GO-universal	SML GOssTo	csbl.go	SemSim	n/a	FunSimMat ProteInOn DaGO-Fun A-DaGO-Fun	FunSimMat DaGO-Fun A-DaGO-Fun	GOSimA-DaGO-Fun	A-DaGO-Fun
Prediction	Annotation		GO-universal Lin Resnik	SML GOssTo	csbl.go	SemSim	n/a	DaGO-Fun A-DaGO-Fun	DaGO-Fun A-DaGO-Fun	A-DaGO-Fun	A-DaGO-Fun
	Cellular localization		Resnik GO-universal	SML GossTo	csbl.go	SemSim	n/a	FunSimMat ProteInOn DaGO-Fun A-DaGO-Fun	G-SESAME[1] FunSimMat DaGO-Fun A-DaGO-Fun	GOSimA-DaGO-Fun	A-DaGO-Fun
			SimGIC	SML	csbl.go	SemSim	n/a	FunSimMat	G-SESAME[1]	GOSimA-DaGO-Fun	A-DaGO-Fun

(continued)

Table 4. Continued

Applications	Measures Suggested	Semantic Similarity Tools suggested							
		Organism-dependent				Organism-independent			
		UniProt ID	Gene name	Entrez Gene	Other	UniProt ID	Gene name	Entrez Gene	Other
Protein–protein interaction	Avg	GossTo				ProteInOn	FunSimMat		
						DaGO-Fun	DaGO-Fun		
						A-DaGO-Fun	A-DaGO-Fun		
Disease and drug targets	GO-universal	SML	csbl.go	SemSim	n/a	FunSimMat	G-SESAME1	GOSimA-DaGO-Fun	A-DaGO-Fun
	Wang *et al.*					ProteInOn	FunSimMat		
	Resnik					DaGO-Fun	DaGO-Fun		
						A-DaGO-Fun	A-DaGO-Fun		

Each 'biological application' is mapped to the possible semantic similarity models providing better performances and potential tools that may be used to retrieve semantic similarity values. 'n/a' indicates that there is no specific tool when the user ID system is different to the three ID systems provided, in which case he/she may need to map their protein ID system to UniProt, gene name or entrez gene, and then make use of a tool according to ID system adopted. Superscript 1 indicates that the tool can be limited as it can only accept one protein or term pair.

existing tools cover a wide range of semantic similarity measures, researchers often need to implement some measures themselves, use different tools for different approaches or download the individual software packages, making extraction and comparison of these scores difficult and time-consuming. Software packages are more flexible and generally extendible, but are not plausible for some end users, as they often require programming skills. On the other hand, using different tools for different measures raises several issues, including (i) scaling issue: different tools may implement the same measure or approach, but using different features, e.g. normalization models, correction factors, considering the root in the set of term's ancestors or not, etc. (ii) data set issue: different tools often use different GO and GOA data set releases and thus lead to scores that are not comparable.

Considering advances made in this area, several models do not need to be implemented, as they have been proven to be ineffective, especially in the context of GO, even though some of these outdated ideas are being brought back. For example, in the context of GO, edge-based models were proven ineffective [41], but some tools, such as SML and GOvis, still implement these measures. In addition, in some tools, including GOssTO [83], semantic similarity scores are not always normalized (range between 0 and 1) and also inferring semantic similarity scores from semantic distance scores is performed using different mapping functions, making comparison between approaches difficult. For example, linear transformation provides a realistic function that maps semantic distance to semantic similarity scores, but GOSim [57] uses the exponential function of the additive inverse of distance scores and KU-GOAL [58] implements a functional similarity measure that is supposed to be derived from a variant of HDF distance. However, this measure does not map its corresponding distance. Finally, there are some redundancies in semantic similarity in several tools owing to the lack of understanding of different measures. For example, most of the existing tools implement the Lin [88] and Jiang and Conrath [89] approaches, but the Jiang and Conrath semantic distance is just the non-normalized distance derived from the Lin similarity measure (see Supplementary File, subsection 2.3.5, or refer to [6] for details) and all normalization schemes that have been proposed have failed to improve the performance of this approach [15]. This shows that the Jiang and Conrath-based semantic similarity approach should be under the Lin approach label and indicated as such in the tool documentation.

There is clearly a need for a tool that integrates all possible relevant GO semantic similarity measures to make semantic similarity score retrieval easy, efficient and effective. Even though A-DaGO-Fun [59] integrates all known GO IC-based semantic measures and a research study was conducted to assess the most used IC-based measures [47], they do not include non-ontology structure-based measures and term similarity approaches derived from distance scores, such as SSDD [78] and Shen *et al.* [90] based term semantic similarity approaches. These are worth assessing and possibly worth including in a GO semantic similarity tool.

Mapping gene or protein identifiers

Existing tools, except for A-DaGO-Fun, use a specific gene identifier (ID) system in integrating and constructing their database. In general, online tools and Java packages use the UniProt ID system, although the DaGO-Fun tool also uses gene names in addition to UniProt ID and FunSimMat accepts the UniProt ID and returns scores with gene names. The R packages

implementing GO semantic similarity measures use NCBI Entrez Gene IDs, except for the csbl.go package, which uses Gene names. This means that each tool uses its own gene ID system in the back-end, and only existing GO annotated organisms are integrated for semantic similarity score calculations. Understanding gene ID to annotation content and gene ID to gene ID mapping is an important initial step for retrieving semantic similarity scores from these tools. However, this can be a serious issue when the user gene IDs cannot be efficiently mapped to the gene ID system used by the tool or when the user gene IDs are redundant or originate from different sources. The user should only rely on the existing cross reference mapping systems, such as UniProt (http://www.uniprot.org/uploadlists/), to align data sets with the tool ID requirements when exploring semantic similarity measures. Even though these platforms effectively address the gene ID cross-mapping issue, it is important to know that there are still some referencing problems across platforms. For example, UniProt does not reference all known ID systems; therefore, in the cross-reference mapping, some annotations may be left out of semantic similarity score computations without the user's awareness, resulting in an incomplete or even a failed semantic similarity score retrieval. In addition, the existing tools generally do not provide help on how to deal with the gene ID to annotation and gene ID to gene ID mapping issues.

Choice of the most appropriate measures and tools

The appropriate use of functional similarity measures depends on the applications [5, 6, 47], as the measures perform differently for different applications. The choice of the semantic similarity measure, therefore, depends on two key aspects: (i) biological question being studied or application being implemented, and (ii) biological quality and mathematical properties of the selected measure. The biological application is based on a type of biological data set, the features of which determine which measure to use. This issue has recently been addressed in [47], in which an extensive performance evaluation of 57 commonly used measures was done using different biological data. It is also important for the selected similarity measure to have good mathematical properties to ensure that results obtained are consistent and can be generalized. As an illustration, in most cases, the Resnik approach is known to perform well, but because this approach does not satisfy some mathematical properties, the performance may be contextual or depend on the normalization factor, and hence cannot be generalized.

GO semantic similarity tools are available in two main formats: software packages, mostly implemented in the R programming language, and web-based online tools (see Tables 1 and 2). These tables also suggest the choice of possible tools, which may be used to retrieve semantic similarity scores given the measure selected for the biological application under consideration. Most of these tools do not support the topology-based family of GO term IC, except SML [87], ADaGO-Fun [59], DaGO-Fun [5] and GOSemSim [86], which support the Wang *et al.* approach. SML also supports the Seco *et al.* approach, ADaGO-Fun and DaGO-Fun support the GO-universal and the Zhang *et al.* approaches. Web-based online tools are user-friendly, but rely on the provider for updates and as a consequence, many are outdated as the GO and GOA data sets are dynamic, updated daily and monthly, respectively. Furthermore, these online tools are limited to a fixed number of options and generally do not offer many possibilities for customization or extension [42]. On the other hand, software packages are generally more flexible and often easily extendible to a user's needs, but they require programming expertise, making them less appealing for some end users. These tools together cover a wide range of semantic similarity measures, but the choice should be made with caution, as the selected tool may not implement exactly the mathematical measure selected with parameters required as pointed out previously. Table 4 maps potential GO semantic similarity measures to biological data types based on performance evaluation results from previous studies [6, 15, 47], assisting GO users in the choice of appropriate measures for a given biological application or data type. Moreover, this table also suggests possible tools that can be used for retrieving similarity scores.

Most of the existing tools also include some applications related to semantic similarity measures, such as protein identification (FunSimMat, DaGO-Fun, A-DaGO-Fun), protein classification (G-SESAME, DaGO-Fun, A-DaGO-Fun), term enrichment (GOSim, DaGO-Fun, A-DaGO-Fun) and protein interactions (ProteInOn). Finally, note that there are also tools that do not produce semantic similarity scores explicitly, but use these scores for specific applications. These include DynGO [91], UTMGO [92] and GFSST [10] for term and protein identification, and Categoriser [39] for term and gene or protein assignment into user-defined biological groups.

Conclusions and perspectives

The recent research trend in semantic similarity measures has consisted of developing novel measures or applying those suggested in WordNet to GO. The current proliferation of these measures initially aims at producing well-adapted and improved measures that are biologically (providing good performance on biological data) and mathematically (satisfying theoretical properties) attractive. However, these measures were developed independently and thus, they fail to answer the initial question of elucidating the most appropriate measure for a given biological question or application. In this study, we have provided an overview of existing semantic similarity measures and focus on analyzing features and issues related to the existing tools to help users choose a relevant measure and/or tool for effective biological knowledge discovery based on GOAs. We have categorized different tools and described their general features in terms of semantic similarity measures (Tables 1 and 2), and we have provided notations adopted by each tool for measures it implements (Table 3). Furthermore, we have classified these measures (Figure 1) and described them in terms of inputs required to retrieve their scores (Figure 2). We have also provided a summary of performance evaluation results and suggested possible tools that may be used for retrieving semantic similarity scores for a given biological application or data type (Table 4). It is worth mentioning that even though several studies [6, 15, 47] have focused on assessing a large number of these measures in the context of GO using different types of biological data, they do not consider hybrid, non-ontology structure-based and other recent measures, such as the Jeang and Chen model (see Supplementary File, section 3.1). This suggests that there is a need to consider performance analyses of these measures to assess how efficient and effective they are in comparison with those that have already been evaluated. Our future work includes evaluating the performance of these measures, which is important as it may provide an indication of whether these measures should be included in the current or future tools.

While substantial progress has been made in the area of GO semantic similarity measures and tools, it is still far from reaching a peak in terms of integration and standardization of these

measures. These measures are currently being deployed into several biological applications and existing software and web-based online tools are helping with retrieving semantic similarity scores. However, the new trend is to make these measures practical and useful for biological questions taking into account the current status of high-throughput technologies with various types of data sets being produced. Owing to the complexity of these high-throughput biology data, in its current state, the next generation of semantic similarity tools should strive for a comprehensive score retrieval environment that will provide a more efficient means to generate semantic similarity scores using appropriate tools and measures, addressing different issues raised above. Tools with such capabilities could make semantic similarity score retrieval easier, and more focused in the context of high-throughput biology data. Furthermore, we can expect the development of new tools to continue, owing to the unmet needs. This study sets a standard and qualitative strategy for assessing new tools and potential measures to orient or guide the growth of the field. Finally, the growing availability of genome scale data sets has raised the issue of quantifying gene set semantic similarity scores rather than just protein pairwise semantic similarity scores. This has not been given enough focus and so far there is one tool, GS^2 [93], which is a non-ontology structure-based method generating gene or protein set semantic similarity scores. It is increasingly important to design an appropriate measure that is customizable, comprehensible and takes into account the structure of the ontology. This is useful, especially when assessing genome- or proteome-wide applications, such as validating protein–protein interaction networks and functionally classifying proteins [62].

Key Points

- Comprehensive summary of semantic similarity measures and tools.
- Consistent classification of existing and future semantic similarity measures.
- Discussion of issues related to semantic similarity measures and tools.
- Assessment of features associated with different semantic similarity tools.
- Prediction of future trends and research directions.

Supplementary data

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

Acknowledgements

The authors thank everyone involved with free software, from the core developers to those who contributed to the documentation. Many thanks to the authors of the freely available libraries, in particular, the GO Consortium et maintainers of the GOA-UniProtKB data set at the European Bioinformatics Institute (EBI), for their effort in keeping the GO and GOA data sets up to date.

Funding

The authors appreciate financial support received from the South Africa National Research Foundation (NRF). Some of the authors are funded in part by Government of Canada via the International Development Research Centre (IDRC) through the African Institute for Mathematical Sciences—Next Einstein Initiative (AIMS-NEI).

References

1. GO-Consortium. The Gene Ontology: enhancements for 2011. *Nucleic Acids Res* 2012;**40**:D559–64.
2. Huntley RP, Sawford T, Mutowo-Muellenet P, *et al*. The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Res* 2014;**43**:D1057–63.
3. Camon E, Magrane M, Barrell D, *et al*. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res* 2003;**13**(4):662–72.
4. Camon E, Barrell D, Lee V, *et al*. The Gene Ontology Annotation (GOA) Database - an integrated resource of GO annotations to the UniProt knowledgebase. *In Silico Biol* 2004;**4**(1):5–6.
5. Mazandu GK, Mulder NJ. DaGO-Fun: tool for Gene Ontology-based functional analysis using term information content measures. *BMC Bioinformatics* 2013;**14**:284.
6. Mazandu GK, Mulder NJ. Information content-based Gene Ontology semantic similarity approaches: toward a unified framework theory. *BioMed Res Int* 2013;**2013**:11.
7. GO-Consortium. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res* 2009;**38**(1):D331–5.
8. Lord PW, Stevens PW, Brass A, *et al*. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics* 2003;**19**(10):1275–83.
9. Sevilla JL, Segura V, Podhorski A, *et al*. Correlation between gene expression and go semantic similarity. *IEEE/ACM Trans Comput Biol Bioinform* 2005;**2**(4):330–8.
10. Zhang P, Jinghui Z, Huitao S, *et al*. Gene functional similarity search tool (GFSST). *BMC Bioinformatics* 2006;**7**:135.
11. Pesquita C, Faria D, Bastos H, *et al*. Evaluating GO-based Semantic Similarity Measures. In: *Proceedings of the 10th Annual Bio-Ontologies (BIOONTOLOGIES)*, 2007. Vienna: Australia, pp. 37–40.
12. Couto FM, Silva MJ, Coutinho PM. Measuring semantic similarity between Gene Ontology terms. *Data Knowledge Eng* 2007;**61**(1):137–52.
13. Schlicker A, Domingues FS, Rahnenfuhrer J, *et al*. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* 2006;**7**:302.
14. Wang JZ, Du Z, Payattakool R, *et al*. A new method to measure the semantic similarity of GO terms. *Bioinformatics* 2007;**23**(10):1274–81.
15. Pesquita C, Faria D, Bastos H, *et al*. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics* 2008;**9**(Suppl 5):S4.
16. Mistry M, Pavlidis P. Gene Ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics* 2008;**9**:327.
17. del Pozo A, Pazos F, Valencia A. Defining functional distances over Gene Ontology. *BMC Bioinformatics* 2008;**9**:50.
18. Jain S, Bader GD. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics* 2010;**11**:562.
19. Mazandu GK, Mulder NJ. A topology-based metric for measuring term similarity in the Gene Ontology. *Adv Bioinformatics* 2012;**2012**:17.
20. Gan M. Correlating information contents of gene ontology terms to infer semantic similarity of gene products. *Comput Math Methods Med* 2014;**2014**:891842.

21. Pesaranghader A, Matwin S, Sokolova M, *et al*. simDEF: definition-based semantic similarity measure of gene ontology terms for functional similarity analysis of genes. *Bioinformatics* 2016;**32**:1380–7.
22. Peng J, Uygun S, Kim T, *et al*. Measuring semantic similarities by combining gene ontology annotations and gene co-function networks. *BMC Bioinformatics* 2015;**16**:44.
23. Bandyopadhyay S, Mallick K. A new path based hybrid measure for gene ontology similarity. *IEEE/ACM Trans Comput Biol Bioinform (TCBB)* 2016;**11(1)**, 116–27.
24. Schlicker A, Lengauer T, Albrecht M. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. *Bioinformatics* 2010;**26**(18):i561–7.
25. Lee HK, Hsu AK, Sajdak J, *et al*. Coexpression analysis of human genes across many microarray data sets. *Genome Biology* 2004;**14**(6):1085–94.
26. Wang H, Azuaje F, Bodenreider O, *et al*. Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. *Proc IEEE Sympo on Comput Intell Bioinforma Comput Biol* 2004;**2004**:25–31.
27. Chabalier J, Mosser J, Burgun A. A transversal approach to predict gene product networks from ontology-based similarity. *BMC Bioinformatics* 2007;**2**:235.
28. Guo X, Liu R, Shriver CD, *et al*. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics* 2006;**22**(8):967–73.
29. Wang J, Xie D, Lin H, *et al*. Filtering Gene Ontology semantic similarity for identifying protein complexes in large protein interaction networks. *Proteome Science* 2012;**10**(Suppl 1):S18.
30. Bettembourg C, Diot C, Dameron O. Optimal threshold determination for interpreting semantic similarity and particularity: application to the comparison of gene sets and metabolic pathways using GO and ChEBI. *PLoS One* 2016;**10**(7):e0133579.
31. 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.
32. Kouhsar M, Zare-Mirakabad F, Jamali Y. WCOACH: protein complex prediction in weighted PPI networks. *Genes Genet Syst* 2016;**90**(5):317–24.
33. Meng J, Liu D, Luan Y. Inferring plant microrna functional similarity using a weighted protein-protein interaction network. *BMC Bioinformatics* 2015;**16**:360.
34. Cui G, Kim B, Alguwaizani S, *et al*. Assessing protein-protein interactions based on the semantic similarity of interacting proteins. *Int J Data Min Bioinform* 2015;**13**(1):75–83.
35. Ye P, Peyser BD, Pan X, *et al*. Gene function prediction from congruent synthetic lethal interactions in yeast. *Mol Syst Biol* 2005;**2005**:1–12.
36. Mazandu GK, Mulder NJ. Function prediction and analysis of *Mycobacterium tuberculosis* hypothetical proteins. *Int J Mol Sci* 2012;**13**(6):7283–302.
37. Lei Z, Dai Y. Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. *BMC Bioinformatics* 2006;**7**:491.
38. Yang Y, Xu Z, Song D. Missing value imputation for microRNA expression data by using a GO-based similarity measure. *BMC Bioinformatics* 2016;**17**(Suppl 1):10.
39. Na D, Son H, Gsponer J. Categorizer: a tool that categorizes genes into user-defined biological groups based on semantic similarity. *BMC Genomics* 2014;**15**:1091.
40. Mazandu GK, Mulder NJ. The use of semantic similarity measures for integrating heterogeneous Gene Ontology annotation pipelines. *Front Genet* 2014;**5**:264.
41. Pesquita C, Faria D, Falcão AO, *et al*. Semantic similarity in biomedical ontologies. *PLoS Comput Biol* 2009;**5**(7):e1000443.
42. Guzzi PH, Mina M, Guerra C, *et al*. Semantic similarity analysis of protein data: assessment with biological features and issues. *Brief Bioinform* 2012;**13**(5):569–85.
43. Couto F, Silva M, Coutinho P. Semantic similarity over the Gene Ontology: family correlation and selecting disjunctive ancestors. In *ACM CIKM - Conference in Information and Knowledge Management*, 2005, Bremen, Germany.
44. Faria D, Pesquita C, Couto FM, *et al*. ProteInOn: a web tool for protein semantic similarity. http://xldb.fc.ul.pt/xldb/publications/Faria.etal:ProteInOnAWeb:2007_document.pdf. 2007.
45. Du Z, Li L, Chen CF, *et al*. G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery. *Nucleic Acids Res* 2009;**37**(2):D345–9.
46. Harispe S, Ranwez S, Janaqi S, *et al*. Semantic measures for the comparison of units of language, concepts or instances from text and knowledge representation analysis: a comprehensive survey and a technical introduction to knowledge-based measures using semantic graph analysis. *arXiv*. 2013. 1310.1285 (2013) version 2.3–06.12.2013.
47. Mazandu GK, Mulder NJ. Information content-based gene ontology functional similarity measures: which one to use for a given biological data type? *PLoS One* 2014;**9**(12):e113859.
48. Seco N, Veale T, Hayes J. An intrinsic information content metric for semantic similarity in wordnet. In: *16th European Conference on Artificial Intelligence, ECAI 2004, IOS Press*, Valencia, Spain, 2004, 1089–90.
49. Sánchez D, Batet M, Isern D. Ontology-based information content computation. *Knowledge-Based Syst* 2011;**24**:297–303.
50. Zhou Z, Wang Y, Gu J. A new model of information content for semantic similarity in WordNet. In: *Second International Conference on Future Generation Communication and Networking Symposia, FGCNS 2008, IEEE Computer*, Washington, USA, 2008;**3**, pp. 85–9.
51. Seddiqui MH, Aono M. (2010) Metric of intrinsic information content for measuring semantic similarity in an ontology. In: *Proceedings of 7th Asia-Pacific Conference on Conceptual Modeling*, Brisbane, Australia, 2010;**110**,89–96.
52. Meng L, Gu J, Zhou Z. A new model of information content based on concept's topology for measuring semantic similarity in WordNet. *Int J Grid Distrib Comput* 2012;**5**(3):81–94.
53. Dubuisson MP, Jain AK. A modified Hausdorff distance for object matching. In *ICPR94*, Jerusalem, Israel, 1994,**1**, pp. A:566–8.
54. Memoli F, Sapiro G. Theoretical and computational framework for isometry invariant recognition of point cloud data. *J Foundations Comp Math* 2005;**55**:313–47.
55. Bronstein AM, Bronstein MM, Mahmoudi M, *et al*. A Gromov-Hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching. *Int J Computer Vision* 2010;**89**:266–86.
56. Lerman G, Shakhnovich BE. Defining functional distance using manifold embeddings of gene ontology annotations. *Proc Natl Acad Sci* 2007;**104**(27):11334–9.
57. Fröhlich H, Speer N, Poustka A, Beißbarth T. GOSim–an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics* 2007;**8**:166.
58. Jeong JC, Chen XW. A new semantic functional similarity over Gene Ontology. *IEEE/ACM Trans Comput Biol Bioinform* 2014;**12**(2):322–34.
59. Mazandu GK, Chimusa ER, Mbiyavanga M, *et al*. A-DaGO-Fun: an adaptable Gene Ontology semantic similarity based functional analysis tool. *Bioinformatics* 2016;**32**:477–9.

60. Al-Mubaid H, Nagar A. Comparison of four similarity measures based on GO annotations for gene clustering. In *IEEE Symposium on Computers and Communications, 6–9 July 2008*, Morocco: Marrakech, Report no. 3.
61. Al-Mubaid H, Nagar A. A new path length measure based on GO for gene similarity with evaluation using SGD pathways. In: *21st IEEE International Symposium on Computer-Based Medical Systems, CBMS '08*, 2008, Jyvaskyla, Finland, pp. 590–5.
62. Benabderrahmane S, Smail-Tabbone M, Poch O, *et al*. IntelliGO: a new vector-based semantic similarity measure including annotation origin. *BMC Bioinformatics* 2010;**11**:588.
63. Alvarez MA, Qi X, Yan C. A shortest-path graph kernel for estimating gene product semantic similarity. *J Biomed Semant* 2011;**2**:3.
64. Ovaska K, Laakso M, Hautaniemi S. Fast gene ontology based clustering for microarray experiments. *BioData Min* 2008;**1**:11.
65. Tversky A. Features of similarity. *Psychol Rev* 1977;**84**(4):327–52.
66. Cross V, Hu X. Fuzzy set and semantic similarity in ontology alignment. In: *WCCI 2012 IEEE World Congress on Computational Intelligence*, 2012, Brisbane, Australia, pp. 1454–61.
67. Gentleman R. Visualizing and distances using GO. 2005. http://bioconductor.org/packages/2.6/bioc/vignettes/GOstats/inst/doc/GOvis.pdf.
68. Schlicker A, Albrecht M. FunSimMat update: new features for exploring functional similarity. *Nucleic Acids Research* 2010;**38**(suppl 1):D244–8.
69. Wang J, Zhou X, Zhu J, *et al*. Revealing and avoiding bias in semantic similarity scores for protein pairs. *BMC Bioinformatics* 2010;**11**:290.
70. Cho YR, Hwang W, Ramanathan M, *et al*. Semantic integration to identify overlapping functional modules in protein interaction networks. *BMC Bioinformatics* 2007;**8**:265.
71. Ali W, Deane CM. Functionally guided alignment of protein interaction networks for module detection. *Bioinformatics* 2009;**25**:3166–73.
72. 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.
73. Camon E, Magrane M, Barrell D, *et al*. The Gene Ontology Annotation (GOA) Database: sharing knowledge in uniprot with Gene Ontology. *Nucleic Acids Res* 2004;**32**:D262–6.
74. Barrell D, Dimmer E, Huntley RP, *et al*. The GOA database in 2009-an integrated Gene Ontology Annotation resource. *Nucleic Acids Res* 2009;**37**:D396–403.
75. Dimmer EC, Huntley RP, Alam-Faruque Y, *et al*. The UniProt-GO Annotation database in 2011. *Nucleic Acids Res* 2012;**40**: D565–70.
76. Zhang SB, Lai JH. Semantic similarity measurement between gene ontology terms based on exclusively inherited shared information. *Gene* 2015;**558**(1):108–17.
77. Song X, Li L, Srimani PK, *et al*. Measure the semantic similarity of GO terms using aggregate information content. *IEEE/ACM Trans Comput Biol Bioinform* 2014;**11**(3), pp. 468–76.
78. Xu Y, Guo M, Shi W, *et al*. A novel insight into gene ontology semantic similarity. *Genomics* 2013;**101**:368–75.
79. Li B, Luo F, Wang JZ, *et al*. Effectively integrating information contentand structural relationship to improve the GO-based similarity measure between proteins. In: *BIOCOMP'10*, 2010, Las Vegas, Nevada (USA), pp. 166–72.
80. Al-Mubaid H, Nguyen HA. A cluster-based approach for semantic similarity in the biomedical domain. In *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. New York, NY: IEEE Computer Society, 2006, 2713–7.
81. Li Y, Bandar ZA, McLean D. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans Knowl Data Eng* 2003;**15(4)**:871–82.
82. Resnik P. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J Artif Intell Res* 1999;**11**:95–130.
83. Caniza H, Romero AE, Heron S, *et al*. GOssTo: a user-friendly stand-alone and web tool for calculating semantic similarities on the Gene Ontology. *Bioinformatics* 2014;**30**:2235–6.
84. Schlicker A, Albrecht M. FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Res* 2008;**36**:D434–9.
85. Martin D, Brun C, Remy E, *et al*. GOToolBox: functional analysis of gene data sets based on Gene Ontology. *Genome Biol* 2004;**5**(12):1901–8.
86. 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**(7):976–8.
87. Harispe S, Ranwez S, Janaqi S, *et al*. The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics* 2014;**30**:740–2.
88. Lin D. An information-theoretic definition of similarity. In: *Proceedings of the Fifteenth International Conference on Machine Learning*, Madison, Wisconson (USA), 1998, pp. 296–304.
89. Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. In: *Proceedings of the 10th International Conference on Research in Computational Linguistics*, Taiwan, 1997, pp. 19–33.
90. Shen Y, Zhang S, Wong HS. A new method for measuring the semantic similarity on gene ontology. In: *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Hong Kong, 2010, pp. 533–8.
91. Liu H, Hu ZZ, Wu CH. DynGO: a tool for visualizing and mining of gene ontology and its associations. *BMC Bioinformatics* 2005;**6**:201.
92. Othman RM, Deris S, Illias RM. UTMGO: a tool for searching a group of semantically related Gene Ontology terms and application to annotation of anonymous protein sequence. *Int J Med Health Biomed Bioeng Pharm Eng* 2007;**1**(3):177–85.
93. Ruths D, Ruths T, Nakhleh L. GS^2: an efficiently computable measure of GO-based similarity of gene sets. *Bioinformatics* 2009;**25**(9):1178–84.