DaGO-Fun: Tool for Gene Ontology-based functional analysis using term information content measures

Gaston K. Mazandu & Nicola J. Mulder

Abstract— **The use of Gene Ontology (GO) data in protein analyses have largely contributed to the improved outcomes of these analyses. Several GO semantic similarity measures have been proposed in recent years and provide tools that allow the integration of biological knowledge embedded in the GO structure into different biological analyses. There is a need for a unified tool that provides the scientific community with the opportunity to explore these different GO similarity measure approaches and their biological applications. We have developed DaGO-Fun, an online tool available at `http://web.cbio.uct.ac.za/ITGOM`, which incorporates many different GO similarity measures for exploring, analyzing and comparing GO terms and proteins within the context of GO. It uses GO data and UniProt proteins with their GO annotations as provided by the Gene Ontology Annotation (GOA) project to precompute GO term information content (IC), enabling rapid response to user queries. The DaGO-Fun online tool presents the advantage of integrating all the relevant IC-based GO similarity measures, including topology- and annotation-based approaches to facilitate effective exploration of these measures, thus enabling users to choose the most relevant approach for their application. Furthermore, this tool includes several biological applications related to GO semantic similarity scores, including the retrieval of genes based on their GO annotations, the clustering of functionally related genes within a set, and term enrichment analysis.**

BACKGROUND

During the last decade several Gene Ontology (GO) semantic similarity approaches [1]–[10] have been introduced for assessing the specificity of and relationship between GO terms based on their position in the GO Directed Acyclic Graph (DAG) [11]–[13]. Terms in the GO DAG are semantically and topologically linked by the relations 'is_a' and 'part_of', expressing relations between a given child term and its parents. Semantic similarity approaches are based on these relations between terms and enable efficient exploitation of the enormous corpus of biological knowledge embedded in the GO DAG by comparing GO terms and proteins at the functional level. GO semantic similarity measures have been widely used in different contexts of protein analysis, including gene clustering, gene expression data analysis, prediction and validation of molecular interactions, and disease gene prioritization [9], [14].

Initially, path- or edge-based approaches, which use a distance or the number of edges between terms in the ontology structure, were introduced [15], [16]. For these approaches, the similarity score between two terms is proportional to the number of edges on the shortest path between these two terms. Path-based approaches were criticized for being limited to edge counting, ignoring positions of terms in the structure and producing uniform similarity scores [9]. Thus, information content based approaches, which rely on a numerical value to convey the description and specificity of a GO term using its position in the structure, were introduced [1]. This numerical value is called information content (IC) or semantic value, and depending on the conception of the term IC, these approaches are divided into two main families, annotation-based and topology-based families. Those depending only on the intrinsic topology of the GO structure are referred to as topology-based approaches while those using the frequencies at which terms occur in the corpus under consideration are referred to as annotation-based approaches.

Annotation-based approaches have been widely analyzed, deployed in many biological applications and were shown to outperform path-based models [17]. Most of them are adapted from Resnik [18], Lin [19] or Jiang & Conrath's [20] methods, and are referred to as classical IC-based similarity approaches. These classical approaches use the most informative common ancestor (MICA) between terms to assess their semantic similarity. Beyond these classical approaches, several other IC-based GO semantic similarity approaches and enhancements have been suggested in order to improve annotation-based measures. These include the graph-based similarity measure (GraSM), developed by Couto et al. [7], which uses all the disjunctive common ancestors (DCA) instead of MICA, the relevance similarity approach proposed by Schlicker et al. [4], and the information coefficient idea of Li et al. [10] to correct the overestimation of similarity scores in Lin's metric. However, the reliance of these approaches on the annotation statistics of the terms biases the scores produced [21]. Topology-based approaches, including the GO-universal metric [22], and the Zhang et al. [3] and Wang et al. [5] methods, were proposed to remove the effect of annotation dependence.

The main use of GO semantic similarity measures is the computation of protein semantic similarity or functional similarity between proteins based on their GO annotations. The completion of several genome sequencing projects has generated immense quantities of sequence data. Subsequently, with

Nicola J. Mulder & Gaston K. Mazandu, Computational Biology Group/Department of Clinical Laboratory Sciences, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, South Africa. Email: {Nicola.Mulder@, gmazandu@cbio.}uct.ac.za.

This work is supported by the Claude Leon Foundation Postdoctoral Fellowship and the National Research Foundation (NRF) in South Africa through the Computational Biology (CBIO) research group.

the continuous development of new high-throughput methods the amount of functional data has increased dramatically, justifying the development of dedicated methods and tools that help extract information from these data. GO [11] has successfully provided a way of consistently describing genes and proteins and a well adapted platform to computationally process data at the functional level. Protein functional similarity methods are counted among tools that allow integration of the biological knowledge contained in the GO DAG, and have contributed to the improvement of biological analyses [17]. These protein functional similarity measures have been used in several applications, including microarray data analysis [23], protein-protein interaction assessments [17], clustering and identification of functional modules in protein-protein interaction networks [24], and putative disease gene identification [25].

As well as different GO semantic similarities, several functional similarity approaches have been proposed. Some of them depend directly on the GO term IC, referred to as Direct Term- or graph-based approaches, and others are constructed via computation of GO term semantic similarity measures, referred to as Term Semantic-based approaches. The former includes approaches derived from the Jaccard, Dice and universal indices based on the Tversky ratio model of similarity [26], referred to as SimGIC [8], [27], SimDIC and SimUIC [22], respectively. The latter approach includes the average (Avg) [1], best-match average (BMA) [8], [22], average best matches (ABM) [5], [24], and the maximum (Max) [2] combinations of GO term similarities for calculating protein functional similarities where proteins are annotated to multiple GO terms. The recent proliferation of these measures in the biomedical and bioinformatics areas was accompanied by the development of tools (`http://www.geneontology.org/GO.tools_by_type.semantic_similarity.shtml`) that facilitate effective exploration of these measures.

These tools include software packages and web-based online tools. Most of the software packages are implemented in the R programming language [28], [29], among which we have SemSim [30], GOSim [31], and csbl.go [23]. There are also online tools, such as ProteInOn [32] and G-SESAME [33]. In addition, an integrated online tool exists, the Collaborative Evaluation of Semantic Similarity Measures (CESSM) [34], for automated evaluation of GO-based semantic similarity approaches, enabling the comparison of new measures against previously published annotation-based GO similarity measures. Evaluation is done in terms of performance with respect to sequence, Pfam and EC similarity. Note that most of the online tools do not support topology-based approaches. The G-SESAME online tool, designed by Du et al. [33] in the context of the Wang et al. approach, supports only classical Resnik [18], Jiang & Conrath [20], and Lin [19] similarity measures for protein or gene clustering applications.

The appropriate use of functional similarity measures depends on the applications [9], [24] since the measures perform differently for different applications. A given measure can yield good performance for one application, but performs poorly for another. Numerous online tools have been developed, but to the best of our knowledge there is no single tool that exhaustively integrates the IC-based functional similarity metrics in order to provide researchers with the freedom to choose the most relevant approach for their specific applications. Here, this is solved through the DaGO-Fun online tool, which integrates up to 27 functional similarity measures, including topology- and annotation-based approaches. This tool also includes some important biological applications directly linked to the use of GO semantic similarity measures, namely the identification of genes based on their GO annotations, the clustering of functionally related genes within a set, and GO term enrichment analysis.

I. Methods

The DaGO-Fun tool integrates GO IC-based semantic similarity measures, allowing researchers to explore and choose an appropriate measure for their analysis. The resulting GO similarity scores are retrieved from the DaGO-Fun database implemented using MySQL and accessible via a web interface. The whole system is implemented using a LAMP (Linux-Apache-MySQL and PHP/Python) platform. This means that the DaGO-Fun tool is implemented under free software (GNU General Public Licence) using a Linux Apache server with a database structured in a relational model using MySQL, with the web interface implemented in PHP-HTML.

The back-end is composed of a set of query processing programs implemented in Python. The user input data are GO terms or UniProt proteins [35]–[37] and their GO annotations from the GOA project [38]–[41]. The database contains about 2×10^7 proteins with GO annotations and $38\,877$ GO terms ($25\,178$ biological process, $10\,426$ molecular function and $3\,273$ cellular component terms) from the GO database. The current version of DaGO-Fun uses UniProt and GOA-UniProtKB release 2013-01 of Jan 9, 2013 and GO version 1.3499 downloaded on 19-January-2013. The database will be updated using an automated scheme every three months.

A. IC-based GO Semantic Similarity Measures

We have implemented two main families of IC-based GO semantic similarity measures: annotation and topology-based families. The annotation-based methods are constrained by the annotation statistics related to terms, while topology-based measures use the intrinsic topology of the GO DAG. In terms of GO term IC, the DaGO-Fun tool includes both families and for the topology-based family, the tool implements three approaches; Zhang et al. [3], Wang et al. [5] and the GO-universal approach [22]. These topology-based family measures each has a specific scheme for computing GO term semantic similarity and functional similarity scores. The annotation-based family has been widely studied and several GO term semantic similarity and protein functional similarity approaches have been introduced.

The GO term semantic similarity approaches include traditional Resnik and Lin measures and two approaches that have been suggested to improve the performance of the Lin measure, namely Relevance (SimRel) [4] and Information

Coefficient (SimIC) [10] similarity measures. Note that in the DaGO-Fun tool, the Jiang & Conrath similarity approach is under the Lin approach label as it is just the non normalized distance derived from the Lin similarity measure. Furthermore, all other normalization schemes that have been proposed have failed to improve the performance of this approach [8]. For similarity measures which are not normalized or whose values do not range between 0 and 1, we have normalized them using the uniformized information content [8], [21], [24], to enable users to compare these data. A value close to one indicates high similarity and close to zero indicates low similarity between proteins at the functional level.

These annotation-based GO term similarity approaches are combined using statistical measures of closeness, such as average (Avg), maximum (Max), best-match average (BMA) and averaging all the best matches (ABM) for calculating protein functional similarity scores. The difference between ABM and BMA approaches is subtle in their conception and scores produced by these two approaches differ. The ABM [5], [24] for two annotated proteins is the mean of best matches of GO terms of each protein against the other, given by the following formula:

$$\text{ABM}(p,q)= \frac{1}{n+m}\left(\sum_{t\in T_p^X}\max_{s\in T_q^X}\mathcal{S}(s,t)+\sum_{t\in T_q^X}\max_{s\in T_p^X}\mathcal{S}(s,t)\right) \tag{1}$$

The Best Match Average (BMA) [8], [22] for two annotated proteins p and q is the mean of the following two values: average of best matches of GO terms annotated to protein p against those annotated to protein q, and average of best matches of GO terms annotated to protein q against those annotated to protein p, given by the following formula:

$$\text{BMA}(p,q)= \frac{1}{2}\left(\frac{1}{n}\sum_{t\in T_p^X}\max_{s\in T_q^X}\mathcal{S}(s,t)+\frac{1}{m}\sum_{t\in T_q^X}\max_{s\in T_p^X}\mathcal{S}(s,t)\right) \tag{2}$$

In equations (1) and (2), $\mathcal{S}(s,t)$ is the semantic similarity score between terms s and t, T_r^X is a set of GO terms in X representing the molecular function (MF), biological process (BP) or cellular component (CC) ontology annotating a given protein r and $n = \left|T_p^X\right|$ and $m = \left|T_q^X\right|$ are the number of GO terms in these sets. These two approaches produce different scores and they are equal only when $n = m$, which is not often the case in a set of annotated genes or proteins.

A well known issue with all these statistical measures of closeness is that they are sensitive to scores that lie at abnormal distances from the majority of scores, or outliers. This means that these measures may produce biases which affect protein functional similarity scores [22]. The functional similarity approach, SimGIC [8], [27], which uses the IC of terms directly to compute protein functional similarity from their GO annotations, was introduced, and uses the Jaccard index. The DaGO-Fun tool also supports two other protein similarity measures relying on GO term IC [22]: SimDIC (Czekanowski or Lin like measure), which uses the Dice index, and SimUIC, which uses a universal index, given by the following formula:

$$\text{SimDIC}(p,q)=\frac{2\times\sum\limits_{x\in\mathcal{A}_p^X\cap\mathcal{A}_q^X} IC(x)}{\sum\limits_{x\in\mathcal{A}_p^X} IC(x)+\sum\limits_{x\in\mathcal{A}_q^X} IC(x)} \tag{3}$$

$$\text{SimUIC}(p,q)=\frac{\sum\limits_{x\in\mathcal{A}_p^X\cap\mathcal{A}_q^X} IC(x)}{\max\left\{\sum\limits_{x\in\mathcal{A}_p^X} IC(x),\sum\limits_{x\in\mathcal{A}_q^X} IC(x)\right\}} \tag{4}$$

where $\mathcal{A}_r^X$ is a set of GO terms together with their ancestors in X representing the ontology (MF, BP or CC) annotating a given protein r. Note that these two measures are still to be evaluated and compared to the existing functional similarity measures.

The DaGO-Fun tool implements 27 functional similarity measures (see Table 1). Each of the four annotation-based GO term similarity approaches, namely Resnik, Lin, relevance and Li et al., is implemented with four known IC-based non-direct functional similarity measures (Avg, Max, BMA and ABM). DaGO-Fun also includes the three IC-based direct term functional similarity measures; SimGIC, SimDIC and SimUIC). It implements XGraSM (eXtended GraSM) in which, instead of considering only the disjunctive common ancestors (DCA), as is the case for the original GraSM, all informative common ancestors (ICA) are considered when computing semantic similarity between two different GO terms and the score between a term and itself is set to 1. This XGraSM approach has been shown to outperform the GraSM approach [21]. Note that finding the disjunctive common ancestors (DCA) between two GO terms makes the original GraSM approach computationally unattractive. Unfortunately, this computational complexity is not proportional to the improvement in performance, and thus, this approach is not included in the DaGO-Fun tool.

On the topology-based approaches, the DaGO-Fun tool implements each approach with its associated functional similarity measure as suggested by the authors of the approach (shown in Figure 1). Thus, the GO-universal approach is implemented with the best match average (BMA) and the Wang et al. approach uses the average best matches (ABM). For the Zhang et al. approach, the DaGO-Fun tool uses averaging best matches (ABM) as it has been shown to improve the performance of this approach [24]. The SimUI approach refers to the union-intersection protein similarity measure, which is also implemented in the GOstats package of Bioconductor [31]. It is a particular case of SimGIC (using the Jaccard index) which assumes that all GO terms occur at equal frequency, in which case, only the topology of the GO DAG is needed [22].

B. Implementation of the DaGO-Fun tool

Protein annotations were retrieved from GOA-UniProtKB at `http://www.ebi.ac.uk/GOA` using UniProt protein

TABLE I: Different GO term semantic similarity approaches and functional similarity measures implemented in DaGO-Fun. The letter ‘x’ indicates that the relevant approach is implemented in DaGO-Fun with the corresponding functional similarity measure.

		Functional Similarity Measures							
		Direct Term-based				Term Semantic-based			
	Approaches	SimGIC	SimDIC	SimUIC	SimUI	BMA	ABM	Avg	Max
Annotation-based		x	x	x					
	XGraSM					x	x	x	x
	Resnik					x	x	x	x
	Lin					x	x	x	x
	Li et al.					x	x	x	x
	Relevance					x	x	x	x
Topology-based					x				
	Zhang et al					x			
	Wang et al.						x		
	GO-universal					x			

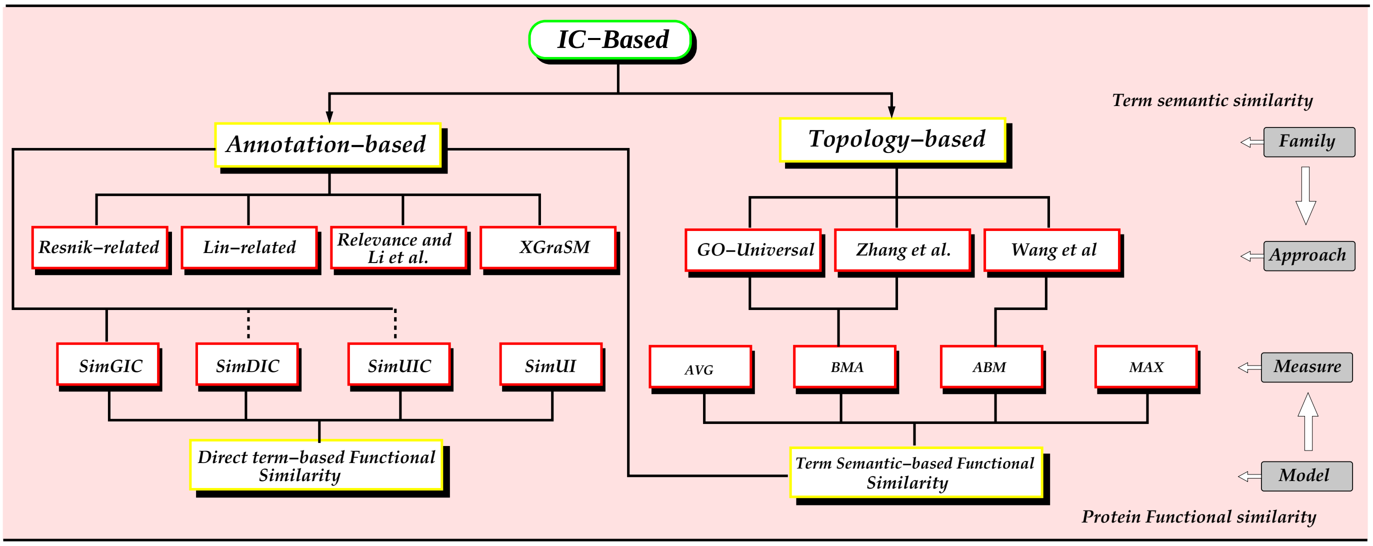


Fig. 1: **Flowchart of all GO measures implemented in DaGO-Fun.** The solid line indicates that the performance of a given measure has already been assessed and the dashed line stands for measures or approaches that have to be evaluated.

accession (ID), gene name and description. GO term topological features (term parents and level) were extracted from the GO database. These data are integrated into a MySQL database of biological concepts present in DaGO-Fun, and used to produce GO term IC, GO term semantic similarity and protein functional similarity scores. The GO term IC scores are integrated into the precompiled dictionaries in the DaGO-Fun tool. The tool is based on a client-server model and is accessible at `http://web.cbio.uct.ac.za/ITGOM` by any user with a standard web browser. The user interface in DaGO-Fun allows easy and comprehensive navigation, query and exploration of GO term, protein semantic similarity scores, and includes biological applications, as shown in Figure 2. This web interface allows the user to input queries in two main dynamic and customizable steps from the search to the user input options before submitting an application for processing.

1) Setting parameters step: The DaGO-Fun tool provides a comprehensive searching scheme. The user selects the task to be processed, which includes the ontology (Biological Process, Molecular Function or Cellular Component) under consideration, and chooses the GO semantic similarity measure family (annotation or topology-based). After this, he/she can select one from a list of available models, which is restricted according to the selected family. Finally, some additional options are available only when dealing with proteins, depending on the user’s choices. If the user selects the annotation-based family then more information is requested about the class (direct IC or non direct IC) of the approach selected and how the IC or GO term similarity scores should be combined. The engine changes further steps to guide the user’s choices by only

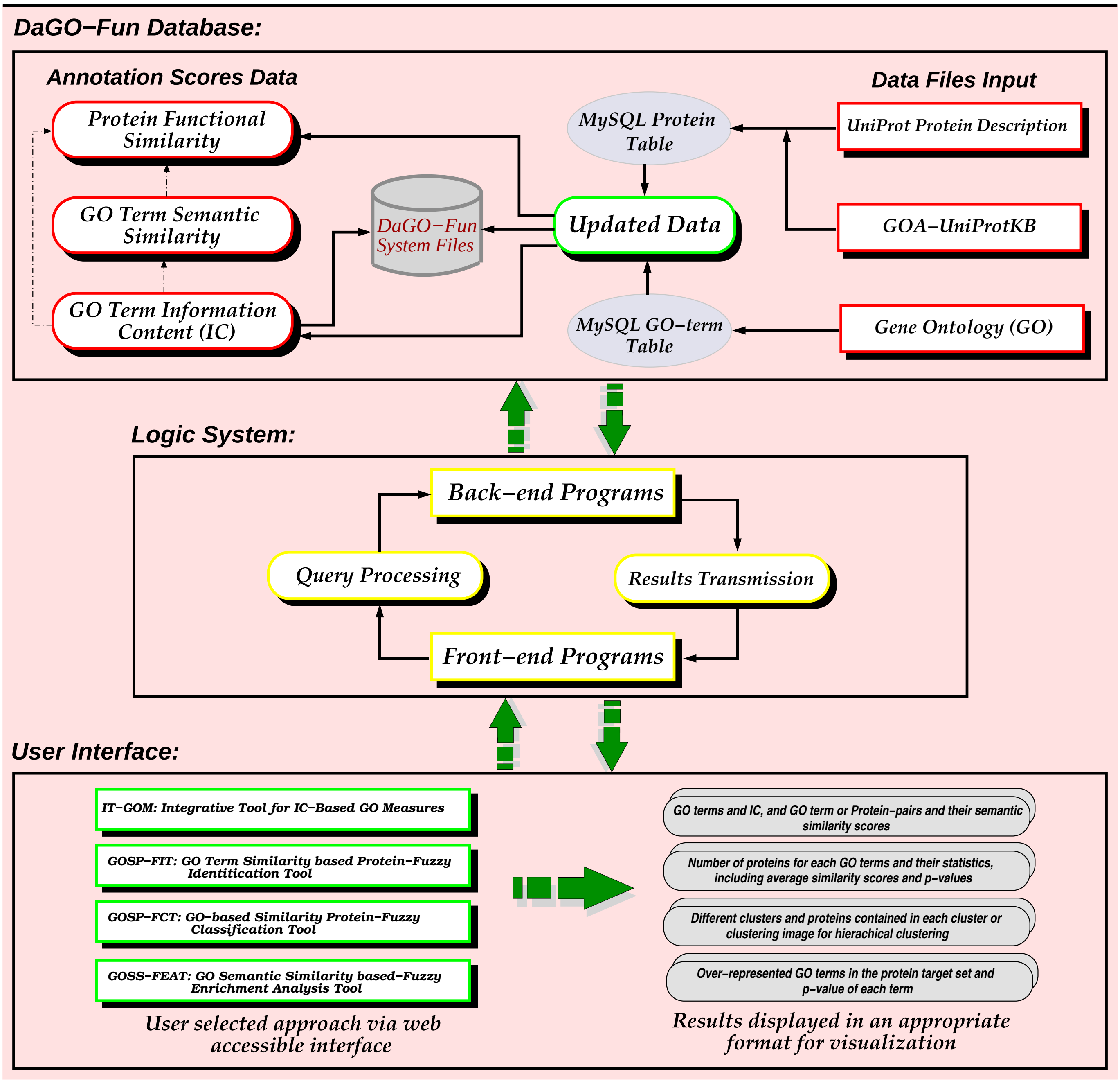

DaGO−Fun Database:
Annotation Scores Data
Protein Functional Similarity
GO Term Semantic Similarity
GO Term Information Content (IC)
DaGO−Fun System Files
MySQL Protein Table
Updated Data
MySQL GO−term Table
Data Files Input
UniProt Protein Description
GOA−UniProtKB
Gene Ontology (GO)
Logic System:
Back−end Programs
Query Processing
Results Transmission
Front−end Programs
User Interface:
IT−GOM: Integrative Tool for IC−Based GO Measures
GOSP−FIT: GO Term Similarity based Protein−Fuzzy Identitication Tool
GOSP−FCT: GO−based Similarity Protein−Fuzzy Classification Tool
GOSS−FEAT: GO Semantic Similarity based−Fuzzy Enrichment Analysis Tool
User selected approach via web accessible interface
GO terms and IC, and GO term or Protein−pairs and their semantic similarity scores
Number of proteins for each GO terms and their statistics, including average similarity scores and p−values
Different clusters and proteins contained in each cluster or clustering image for hierachical clustering
Over−represented GO terms in the protein target set and p−value of each term
Results displayed in an appropriate format for visualization

Fig. 2: **The DaGO-Fun system architecture.** The user selects the application and enters the input (GO Ids, Protein Accessions or Gene names, GO Id pairs and protein or gene name pairs). The application is processed from the DaGO-Fun system and results are displayed in a comprehensive format for visualization.

making available the options relevant to the current choice.

2) User Input step: After selecting appropriate parameters, the user enters their queries in a text area or from a file, and the size of the input allowed depends on the applications. Note that the DaGO-Fun tool currently includes four applications, namely: Term and protein semantic similarity measures (IT-GOM), Protein Fuzzy-Identification (GOSP-FIT), Term Fuzzy-Enrichment Analysis (GOSS-FEAT) and Protein Fuzzy-Classification (GOSP-FCT). Here, the fuzzy concept is related to the fact that the results or outputs of a given query are a function of a certain agreement score or level.

- For IT-GOM at `http://web.cbio.uct.ac.za/ITGOM/tools/itgom.php`: up to 3000 pairs of GO Ids, UniProt protein accessions or gene names can be submitted for GO term similarity and functional similarity querying. For GO term IC, the user can enter up to 5000 GO Ids.

- A list of at most 20 GO Ids belonging to the same GO ontology is recommended when using GOSP-FIT at `http://web.cbio.uct.ac.za/ITGOM/tools/gotspfit.php`.
- For GOSS-FEAT at `http://web.cbio.uct.ac.za/ITGOM/tools/gossfeat.php`: a target list of at most 2000 protein UniProt accessions or gene names is recommended.
- Finally, a list of no more than 200 protein UniProt accessions or gene names is recommended for GOSP-FCT at `http://web.cbio.uct.ac.za/ITGOM/tools/gospfuct.php`.

These cut-offs are mainly due to the limitations of the computational resources available but also to the visualization constraints and algorithm complexity, for example when running hierarchical clustering in GOSP-FCT.

3) Outputs: Comprehensive summary reports generated from the DaGO-Fun tool are made available in table format. An example of a result report is shown in Figure 3 and this report can be downloaded as a tab-delimited text file or printed. Users can query specific links directly, leading to the reported GO terms or proteins. Note that proteins are linked to their annotations via QuickGO at EBI (`http://www.ebi.ac.uk/QuickGO`), and for GO term semantic similarity and information content queries, GO Ids are linked to their characteristics and their sub-GO graphs displayed using AmiGO at `http://amigo.geneontology.org`. A given concept (protein accession or GO Id) can also be linked to more detailed results related to the concept. More details on the use of the tool are provided in the help page on the website.

C. GO term statistics

The DaGO-Fun tool uses binomial test for the retrieval of genes based of their GO annotations (GOSP-FIT) and hyper-geometric test for term enrichment analysis (GOSS-FEAT), adjusted using the Bonferroni multiple testing correction. Note that using the hyper-geometric distribution, the p-value, which is the probability of observing at least ℓ genes from a target gene set of size n by chance, knowing that the reference dataset, considered as a background distribution, contains m such annotated genes out of N genes is given by the following formula:

$$P\left[X \geq \ell\right] = 1 - \sum_{k=0}^{\ell-1} \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \qquad (5)$$

The random variable X represents the number of genes within a given target gene subset, annotated with a given GO term. Note that we are dealing with very large population size (organism's genome, proteome or set of annotated proteins in the GOA file), in which case the size of the target gene or protein subset is very small compared to the population size. Thus, the p-value can also be approximated by or modeled using the binomial distribution [42] by taking the relative frequency of occurrence of each GO term in the reference dataset as an estimator of the probability p of observing the GO term under consideration. In this case, a gene taken at random from the reference dataset is an event with two possible outcomes, namely success (1), if the gene is annotated with the GO term, and failure (0) otherwise. Thus, the probability of obtaining at least ℓ successes in n trials or observing at least ℓ genes annotated with the GO term under consideration among n genes in the target set is given by the following formula:

$$P\left[X \geq \ell\right] = 1 - \sum_{k=0}^{\ell-1} \binom{n}{k} p^{k}\left(1-p\right)^{n-k} \qquad (6)$$

In these cases, the lower the p-value, the less likely it is that the observed frequency of the term is due to chance, and thus the more meaningful the term is in the target gene set. Thus, GO terms in the dataset under consideration can be ranked based on their p-values using the fact that the lower the p-value, the more significant the observed GO term is.

Note that as the biological applications implemented depend on the agreement level, the frequency of occurrence of a term through a gene or protein g is in fact fuzzy-frequency of this term modeled using GO similarity score $\mathcal{A}_g$, of the term to the set of GO terms annotating the gene, given by the following formula:

$$\mathcal{A}_g\left(t\right) = \mathcal{S}\left(t, T_g^X\right) \qquad (7)$$

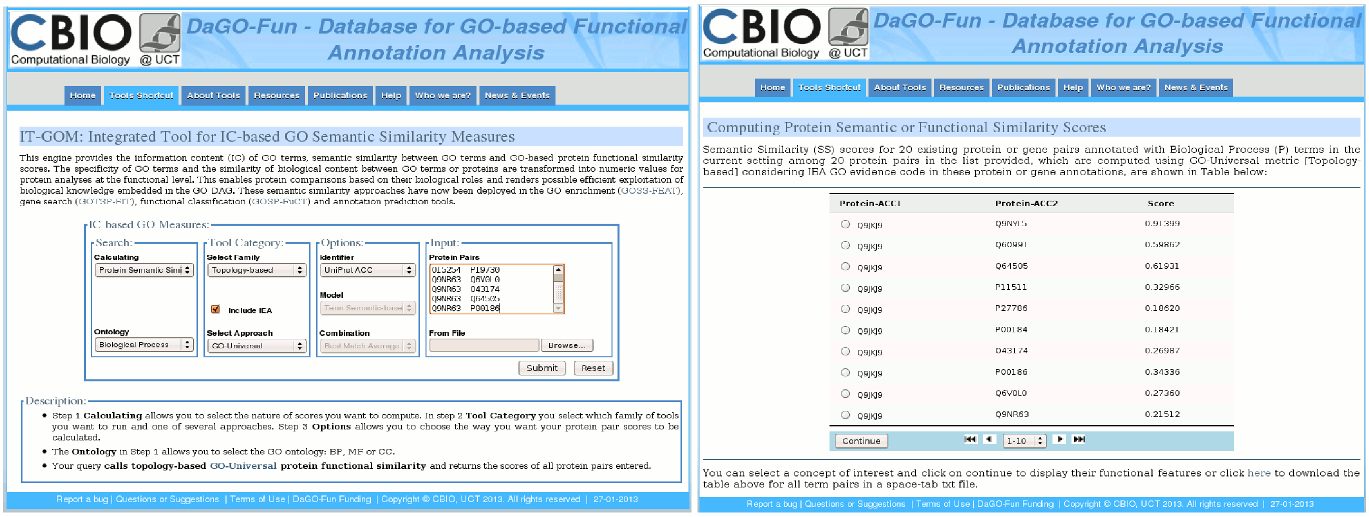
T_g^X is a set of GO terms in the ontology X annotating the gene g and $\mathcal{S}\left(t, T_g^X\right) = \max\left\{\mathcal{S}\left(t,s\right) : s \in T_g^X\right\}$ [22], with $\mathcal{S}\left(t,s\right)$ representing the semantic similarity score between GO terms t and s. We say the gene g is not annotated with t or t does not occur through the gene g if $\mathcal{A}_g\left(t\right) = 0$, g is fully annotated with t or t fully occurs if $\mathcal{A}_g\left(t\right) = 1$ and g is fuzzy annotated with t or t fuzzy occurs if $0 < \mathcal{A}_g\left(t\right) < 1$. Thus, the fuzzy occurrence of a given term induces the possibility of a term occurrence through a given protein in the annotation data under consideration. Specifically, the fuzzy frequency of occurrence of the GO term t in a set of genes $\mathcal{C}$ from a given experiment, denoted $ff\left(t\right)$, is calculated using the following formula:

$$ff\left(t\right) = \sum_{g \in \mathcal{C}} \delta_g\left(t\right) \qquad (8)$$

where δ_g is the $g-$function indicator given by

$$\delta_g\left(t\right) = \begin{cases} 1 & \text{if } \mathcal{A}_g\left(t\right) \geq c \\ 0 & \text{otherwise} \end{cases}$$

$c > 0$ is the agreement level or customized agreement at which the GO term t is considered to be a possible annotation of the gene g. The value of $c = 0.3$ is considered to be a default value of the agreement level, and its associated fuzzy frequency is referred to as realistic or moderate frequency. This is strong or high frequency if $c = 0.7$ and perfect frequency if $c = 1$, which corresponds to the traditional approaches.



CBIO
Computational Biology @ UCT
DaGO-Fun - Database for GO-based Functional Annotation Analysis
Home
Tools Shortcut
About Tools
Resources
Publications
Help
Who we are?
News & Events
Submit
Reset
Browse...
Continue
1-10
Report a bug | Questions or Suggestions | Terms of Use | DaGO-Fun Funding | Copyright © CBIO, UCT 2013. All rights reserved | 27-01-2013

IT-GOM: Integrated Tool for IC-based GO Semantic Similarity Measures

This engine provides the information content (IC) of GO terms, semantic similarity between GO terms and GO-based protein functional similarity scores. The specificity of GO terms and the similarity of biological content between GO terms or proteins are transformed into numeric values for protein analyses at the functional level. This enables protein comparisons based on their biological roles and renders possible efficient exploitation of biological knowledge embedded in the GO DAG. These semantic similarity approaches have now been deployed in the GO enrichment (GOSS-FEAT), gene search (GOTSP-FIT), functional classification (GOSP-FuCT) and annotation prediction tools.

IC-based GO Measures:

Search:
Calculating: Protein Semantic Simi
Ontology: Biological Process

Tool Category:
Select Family: Topology-based
☑ Include IEA
Select Approach: GO-Universal

Options:
Identifier: UniProt ACC
Model: Term Semantic-base
Combination: Best Match Average

Input:
Protein Pairs:
```
O15254   P19730
Q9NR63   Q6V0L0
Q9NR63   O43174
Q9NR63   Q64505
Q9NR63   P00186
```
From File

Description:

- Step 1 **Calculating** allows you to select the nature of scores you want to compute. In step 2 **Tool Category** you select which family of tools you want to run and one of several approaches. Step 3 **Options** allows you to choose the way you want your protein pair scores to be calculated.
- The **Ontology** in Step 1 allows you to select the GO ontology: BP, MF or CC.
- Your query **calls topology-based GO-Universal protein functional similarity** and returns the scores of all protein pairs entered.

Computing Protein Semantic or Functional Similarity Scores

Semantic Similarity (SS) scores for 20 existing protein or gene pairs annotated with Biological Process (P) terms in the current setting among 20 protein pairs in the list provided, which are computed using GO-Universal metric [Topology-based] considering IEA GO evidence code in these protein or gene annotations, are shown in Table below:

Protein-ACC1	Protein-ACC2	Score
○ Q9JKJ9	Q9NYL5	0.91399
○ Q9JKJ9	Q60991	0.59862
○ Q9JKJ9	Q64505	0.61931
○ Q9JKJ9	P11511	0.32966
○ Q9JKJ9	P27786	0.18620
○ Q9JKJ9	P00184	0.18421
○ Q9JKJ9	O43174	0.26987
○ Q9JKJ9	P00186	0.34336
○ Q9JKJ9	Q6V0L0	0.27360
○ Q9JKJ9	Q9NR63	0.21512

You can select a concept of interest and click on continue to display their functional features or click here to download the table above for all term pairs in a space-tab txt file.

Fig. 3: **Application example of querying IT-GOM and an output summary.** The left figure shows the DaGO-Fun interface providing the query form with user input data and the figure on the right displays the results table of protein similarity scores produced by the selected algorithm.

TABLE II: Results obtained after running the GOSP-FIT for specific GO Ids and using different GO term semantic similarity approaches, namely GO-universal (GA), Wang et al (WA), Zhang et al. (ZA), Resnik (RA), Lin (LA) and Li et al. (LLA).

			Number of proteins detected							
GO ID	Level	GO Name	GA	WA	ZA	RA	LA	LLA	p-value	Corrected p-value
GO:0044255	4	cellular lipid metabolic process	154	1590	1907	225	1652	1286	0.00e+00	0.00e+00
GO:0071236	6	cellular response to antibiotic	123	307	545	277	739	588	2.42e-14	9.68e-14
GO:0051409	3	response to nitrosative stress	91	226	426	243	435	418	1.07e-14	4.26e-14
GO:0052099	6	acquisition by symbiont of nutrients from host via siderophores	47	128	463	2	418	269	2.75e-14	1.10e-13

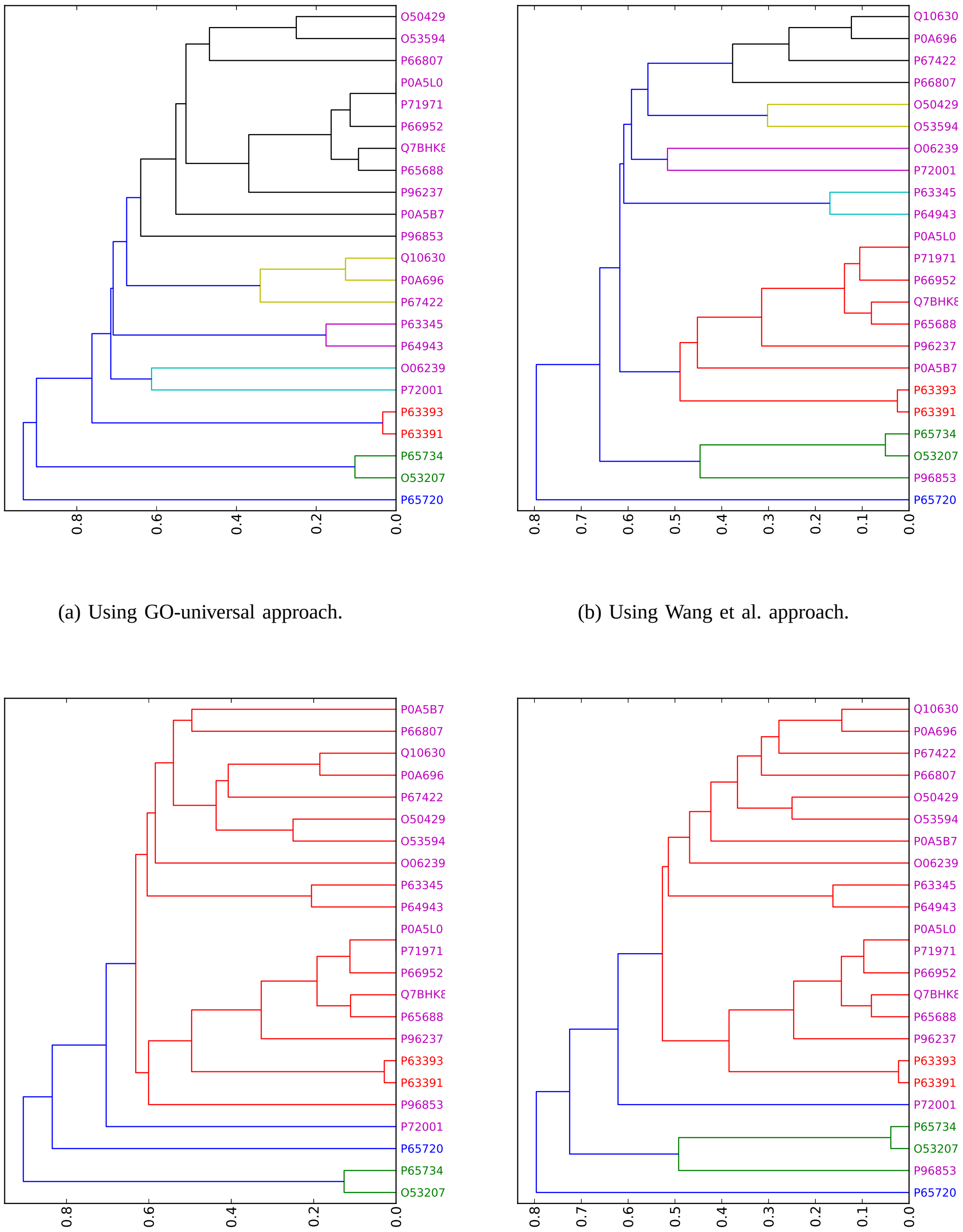
II. Results and Discussion

In this section we provide and discuss briefly some illustrations of biological applications included in the DaGO-Fun tool, namely GO Term Similarity based Protein-Fuzzy Identification Tool (GOSP-FIT), GO based Similarity Protein-Fuzzy Classification Tool (GOSP-FCT) and GO Semantic Similarity based-Fuzzy Enrichment Analysis Tool (GOSS-FEAT). We ran these applications on the *Mycobacterium tuberculosis* (MTB) genome using different GO semantic similarity approaches and analyzed the results obtained. MTB is an intracellular pathogen that causes tuberculosis (TB), one of the most threatening infectious diseases considering the severity of its impact on human populations [43]. To be successful, MTB must, at each step of the infection, express a set of genes that enables it to survive and persist inside its host macrophages, defeating antibacterial mechanisms of host cells and evading the antibiotic actions of drugs. Thus, it is believed that besides some basic biological processes, these genes or proteins must be involved in critical biological processes, such as *response to nitrosative stress* (GO:0051409), *cellular response to antibiotic* (GO:0071236), *acquisition by symbiont of nutrients from host via siderophores* (GO:0052099), *cellular lipid metabolic process* (GO:0044255), etc. We used these GO biological process terms as initial data or input for running different biological applications in the DaGO-Fun tool at moderate agreement, unless otherwise stated.

A. Performing DaGO-Fun applications

Using the biological process terms listed above, we ran GOSP-FIT to identify proteins involved in a process similar to the input processes, using the GO-universal metric, Wang et al., Zhang et al, Resnik, Lin and Lin with Li et al. enhancement similarity measures. Results are shown in Table 2. We see that, except for GO-universal and Resnik approaches, other approaches tend to select more proteins for a given term. This is an indication that these approaches are overestimating GO term similarity scores. It is already known that the Lin approach overestimates similarity scores between terms, which is why the enhancement of this measure has been suggested through the information coefficient idea of Li et al. [10] and the relevance similarity approach proposed by Schlicker et al. [4] to correct these overestimated scores. From the number of proteins detected by Lin and its enhancement proposed by Li et al., we observe that this enhancement is trying to reduce the impact of Lin similarity score overestimation even though overall these measures are still overestimating similarity scores. Finally, note that one can display all proteins identified for a given term by selecting the row of the term and clicking on the 'Continue' button.

Before running other applications, we first identified in the MTB genome all genes or proteins involved in the GO annotations under consideration. A total of 23 proteins have been identified with 18 proteins (O53594, P66807, P0A696, P0A5L0, Q10630, P72001, P96853, O06239, P65688, P64943, O50429, P66952, P63345, P96237, P67422, Q7BHK8, P0A5B7, P71971) for GO:00051409, one protein (P65720) for GO:0071236, 2 (P65734, O53207) for GO:0044255, and 2 (P63391, P63393) for GO:0052099. We used these proteins as input data for running GOSP-FCT using hierarchical clustering under the customized agreement level. Results are depicted in Figure 4 and indicate that the clustering outcome depends



(a) Using GO-universal approach.

(b) Using Wang et al. approach.

(c) Using Resnik approach.

(d) Using Li et al. approach.

Fig. 4: **Clustering results obtained by running the hierarchical clustering program using different similarity metrics under the DaGO-Fun tool.** Protein label is colored according to the process in which the protein is involved. Magenta for proteins involved in GO:00051409, blue for GO:0071236, green for GO:0044255 and red for GO:0052099.

TABLE III: Running the GOSS-FEAT for specific GO Ids and using different GO term semantic similarity approaches.

Approach	GO-ID	GO Name	Level	Reference Fuzzy Frequency	Target Fuzzy Frequency	p-value	Corrected p-value
GO-Universal	GO:0051409	response to nitrosative stress	3	91	18	0.00e+00	0.00e+00
	GO:0006979	response to oxidative stress	3	90	18	8.08e-13	5.17e-11
	GO:0052572	response to host immune response	7	92	8	1.16e-07	7.45e-06
Wang et al.	GO:0006979	response to oxidative stress	3	226	18	0.00e+00	0.00e+00
	GO:0046677	response to antibiotic	4	164	18	0.00e+00	0.00e+00
	GO:0001666	response to hypoxia	5	163	18	0.00e+00	0.00e+00
	GO:0006974	response to DNA damage stimulus	5	419	18	0.00e+00	0.00e+00
	GO:0009432	SOS response	5	316	18	0.00e+00	0.00e+00
	GO:0034605	cellular response to heat	5	351	18	0.00e+00	0.00e+00
	GO:0071500	cellular response to nitrosative stress	5	365	18	0.00e+00	0.00e+00
	GO:0075136	response to host	5	293	12	9.10e-08	5.82e-06
Resnik	GO:0009432	SOS response	5	294	18	0.00e+00	0.00e+00
	GO:0034605	cellular response to heat	5	296	18	0.00e+00	0.00e+00
	GO:0071500	cellular response to nitrosative stress	5	294	18	0.00e+00	0.00e+00
	GO:0009267	cellular response to starvation	6	294	18	0.00e+00	0.00e+00
	GO:0071456	cellular response to hypoxia	7	370	18	0.00e+00	0.00e+00
	GO:0071732	cellular response to nitric oxide	7	369	18	0.00e+00	0.00e+00
	GO:0006284	base-excision repair	8	361	18	0.00e+00	0.00e+00
	GO:0006289	nucleotide-excision repair	8	361	18	0.00e+00	0.00e+00
	GO:0006307	DNA dealkylation involved in DNA repair	9	424	18	0.00e+00	0.00e+00
	GO:0052059	evasion or tolerance by symbiont of host-produced reactive oxygen species	11	319	18	1.63e-12	1.04e-10
	GO:0052060	evasion or tolerance by symbiont of host-produced nitric oxide	11	319	18	1.63e-12	1.04e-10
	GO:0051701	interaction with host	4	35	6	1.05e-07	6.74e-06

strongly on the similarity approach used. Here, again we see that the GO-universal approach performs better than other approaches, producing a clustering image which is consistent with mapping between GO terms and identified proteins, as indicated above. It is worth mentioning that two other clustering approaches are implemented under the DaGO-Fun tool, namely the graph spectral or kmeans clustering approach and the community detecting model [44], which is referred to as a model-based approach. For the kmeans clustering approach, the user is required to provide the expected number of clusters of his/her model. For these two approaches, results are displayed in a table format in which each cluster is mapped to its related proteins.

Finally, we ran GOSS-FEAT, taking as the target set a list of 18 proteins annotated to GO:0051409 in order to identify the most statistically relevant biological processes in which these proteins are involved. We used the GO-universal metric, Wang et al and Resnik approaches and results are shown in Table 3. Once again, these results depend on the semantic similarity measure used and looking at these results, only the GO-universal approach was able to output the GO term used to identify proteins used as the target set, namely *response to nitrosative stress* GO:0051409. This application suggests that the GO-universal approach may constitute an effective solution to the GO metric problem for the next generation of functional similarity metrics [22].

B. Other GO semantic similarity tools and DaGO-Fun

As mentioned previously, there have been numerous tools developed for producing GO term and protein semantic similarity scores. These include web interfaces and software tools very often implemented in the R programming language. These tools, together with functional similarity measures they support, are shown in Table 4. As pointed out previously, each approach performs differently for different applications. For example, the maximum approach achieves good performance for prediction of protein-protein interactions compared to other approaches [24]. The best-match average approaches perform better in protein function prediction and validation [9], and protein or gene clustering, while the average approach is good for detecting similar protein sequences from their GO annotations [1]. The existing tools allow researchers to browse the specific approaches separately for their proteins of interest, but an integrated tool for exploring all the IC-based similarity approaches to allow researchers to choose the most relevant approach for their applications did not exist previously. DaGO-Fun solves this by allowing researchers to browse the integrated set of all IC-based GO semantic similarity approaches. The similarity scores produced are scaled (normalized) to enable comparison between different approaches, and in the future we will work on enabling multiple options to be run, with a summary or merging of results where possible.

In terms of input size, the G-SESAME and FuSSiMeg web tools accept only one pair of GO terms or proteins. The ProteInOn tool may take up to 1000 GO terms or proteins according to its authors, for which the tool outputs all pairs of similarity scores, and the FunSimMat tool has unlimited input size. We aim to let the DaGO-Fun tool calculate results for as many user inputs as possible, however, because of limitations in computational resources, we have to balance the maximum

TABLE IV: IC-based GO semantic similarity tools and functional similarity measures (FSM) they support.

		GO-Semantic Similarity features implemented		
Tool	Format	Family	Approach	FSM
G-SESAME	Web	Topology-based Annotation-based	Wang et al. Classical Resnik, Lin and Jiang & Conrath	ABM Average
ProteInOn	Web	Annotation-based	Classical Resnik, Lin and Jiang & Conrath GraSM	BMA and SimGIC
FuSSiMeg	Web	Annotation-based	Classical Resnik, Lin and Jiang & Conrath GraSM	Max
FunSimMat	Web	Annotation-based	Classical Resnik, Lin and Jiang & Conrath SimRel (Enhancement)	ABM
SemSim	R	Topology-based Annotation-based	Wang et al. Method Classical Resnik, Lin and Jiang & Conrath SimRel (Enhancement)	ABM Average
csbl.go	R	Annotation-based	Classical Resnik, Lin and Jiang & Conrath GraSM and SimRel	SimGIC average

number of GO terms, and GO term and protein pairs for each user query. Thus, the DaGO-Fun tool accepts up to 5000 GO terms when retrieving GO term IC scores, in which case the tool will display only 10 of them per page, but all GO term features can be retrieved by downloading them in a text file. For GO term semantic similarity scores as well as for protein functional similarity scores, the user can enter at most 3000 pairs. Entries beyond the maximum limitations will be ignored. Unfortunately if you have cases where your data exceeds these limitations, it is necessary to divide the input data, run the DaGO-Fun tool separately, and merge the results at the end of the process. Alternatively you can contact the authors who are willing to collaborate and run large data sets for analysis.

III. Conclusions

We have developed the DaGO-Fun tool, a customized web-based GO semantic similarity resource. This user-friendly online interface produces GO term information content (IC), GO term semantic similarity and protein functional similarity scores, which may assist experimental and computational biologists in several applications involving protein analyses at the functional level. These include gene list enrichment, protein function prediction and comparison, clustering genes or proteins based on their GO annotation information, and ranking disease candidate proteins or identification of novel disease candidate proteins. This tool will be updated quarterly (every three months) using an automated scheme in order to remain up to date to meet requirements of ever increasing applications in the biomedical field. The DaGO-Fun tool is freely available, meaning that one is free to copy, distribute, display and make unrestricted non-commercial use of it under the GNU General Public Licence provided that it is done with appropriate citation of the tool and its components.

Despite the wide range of IC-based GO semantic similarity applications and the existence of several approaches to meet requirements of these applications, there was no tool available that integrates all these IC-based approaches. Thus, researchers had to implement these approaches themselves, use different tools for different approaches, or download the individual software packages, making extraction and comparison of these scores difficult and time-consuming. The DaGO-Fun tool overcomes these issues, providing easy retrieval of IC-based GO term semantic similarity and protein functional similarity scores within a large protein annotation dataset from GOA-UniProtKB. It ensures that GO semantic similarity data are conveniently accessible to researchers and can effectively be used to investigate functional similarity between proteins based on their GO annotations. In addition, we implemented some biological applications of these semantic similarity measures, including protein classification and identification based on their GO annotations, and term enrichment analysis.

Future work includes facilitating the search for functional similarity between sets of GO terms. In this case, the user will have to provide pairs of sets of GO terms using a specified key linking the sets. This will undoubtedly improve the flexibility of the DaGO-Fun tool, by allowing users to produce functional similarity scores for their own predicted set of genes given their GO annotations. We will assess the relevance of two IC-term based functional similarity approaches introduced here, namely SimDIC and SimUIC and evaluate the use of annotation-based functional similarity approaches in the context of the GO term IC topology-based family. Finally, we will be expanding the DaGO-Fun tool to include some other applications of GO semantic similarity in protein analyses, such as protein function prediction, annotation system comparisons, and disease protein prioritization.

Conflict of interest statement. None declared.

IV. Authors contributions

NJM generated and supervised the project, and finalized the manuscript. GKM designed and implemented the tool, and wrote the manuscript. All authors read and approved the final manuscript, and NJM approved the production of this paper.

V. Acknowledgements

Any work dependent on open-source software owes debt to those who developed these tools. 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. We would also like to thank Mr Ayton Meintjes for his availability in helping transfer DaGO-Fun resources to our main server and his practical advice concerning the tool management. This work has been supported by the Claude Leon Foundation Postdoctoral Fellowship and the National Research Foundation (NRF) in South Africa through the Computational Biology (CBIO) research group at the Institute of Infectious Disease and Molecular Medicine, University of Cape Town.

References

[1] Lord PW, Stevens PW, Brass A, Goble CA: **Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation**. *Bioinformatics* 2003, **19(10)**:1275–1283.

[2] Sevilla JL, Segura V, Podhorski A, Guruceaga E, Mato JM, Martinez-Cruz LA, Corrales FJ: **Correlation between Gene Expression and GO Semantic Similarity**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) archive* 2005, **2(4)**:330–338.

[3] Zhang P, Jinghui Z, Huitao S, Russo J, Osborne B, Buetow K: **Gene functional similarity search tool (GFSST)**. *BMC Bioinformatics* 2006, **7**:135.

[4] Schlicker A, Domingues FS, Rahnenfuhrer J, Lengauer T: **A new measure for functional similarity of gene products based on Gene Ontology**. *BMC Bioinformatics* 2006, **7**:302.

[5] Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF: **A new method to measure the semantic similarity of GO terms**. *Bioinformatics* 2007, **23(10)**:1274–1281.

[6] Couto F, Silva M, Coutinho P: **Measuring semantic similarity between gene ontology terms**. *Data Knowledge Eng* 2007, **61(1)**:137–152.

[7] Couto F, Silva M, Coutinho P: **Semantic similarity over the gene ontology: Family correlation and selecting disjunctive ancestors**. In *CIKM '05 Proceedings of the 14th ACM international conference on Information and knowledge management* 2005:343–344.

[8] Pesquita C, Faria D, Bastos H, Ferreira AEN, Falcão AO, Couto FM: **Metrics for GO based protein semantic similarity: a systematic evaluation**. *BMC Bioinformatics* 2008, **9(Suppl 5)**:S4.

[9] Pesquita C, Faria D, Falcão AO, Lord P, Couto FM: **Semantic Similarity in Biomedical Ontologies**. *PLoS Comput Biol* 2009, **5(7)**:e1000443.

[10] Li B, Wang JZ, Feltus FA, Zhou J, Luo F: **Effectively integrating information content and structural relationship to improve the GO-based similarity measure between proteins**. *ArXiv e-prints* 2010, :1001.0958.

[11] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene Ontology: tool for the unification of biology**. *Nat Genet* 2000, **25(1)**:25–29.

[12] GO-Consortium: **The Gene Ontology in 2010: extensions and refinements**. *Nucleic Acids Research* 2009, **38**:D331–D335.

[13] GO-Consortium: **The Gene Ontology (GO) project in 2006**. *Nucleic Acids Research* 2006, **34**:D322–D326.

[14] Schlicker A, Albrecht M: **FunSimMat: a comprehensive functional similarity database**. *Nucleic Acids Research* 2008, **36(Database issue)**:D434–D439.

[15] Pekar V, Staab S: **Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision**. In *Proceedings of the 19th international conference on Computational linguistics. Morristown, NJ, USA: Association for Computational Linguistics, Volume 1* 2002:1–7.

[16] Rada R, Mili H, Bicknell E, Blettner M: **Development and application of a metric on semantic nets**. In *IEEE Transaction on Systems, Man, and Cybernetics, Volume 19(1)* 1989:17–30.

[17] Guo X, Liu R, Shriver C, Hu H, Liebman M: **Assessing semantic similarity measures for the characterization of human regulatory pathways**. *Bioinformatics* 2006, **22(8)**:967–973.

[18] Resnik P: **Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language**. *Journal of Artificial Intelligence Research* 1999, **11**:95–130.

[19] Lin D: **An Information-Theoretic Definition of Similarity**. In *Proceedings of the Fifteenth International Conference on Machine Learning* 1998:296–304.

[20] 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* 1997:19–33.

[21] Mazandu GK, Mulder NJ: *Information content-based Gene Ontology semantic similarity approaches: Toward a unified framework theory*. BioMed Research International 2013. [In Press].

[22] Mazandu GK, Mulder NJ: **A topology-based metric for measuring term similarity in the Gene Ontology**. *Adv Bioinformatics* 2012, **2012**:Ariticle ID 975783, 17 pages.

[23] Ovaska K, Laakso M, Hautaniemi S: **Fast gene ontology based clustering for microarray experiments**. *BioData Mining* 2008, **1**:11.

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

[25] Schlicker A, Lengauer T, Albrecht M: **(2010) Improving disease gene prioritization using the semantic similarity of gene ontology terms**. *Bioinformatics* 2010, **26(18)**:i561–i567.

[26] Tversky A: **Features of similarity**. *Psychological Review* 1977, **84(4)**:327–352.

[27] Pesquita C, Faria D, Bastos H, Falcão AO, Couto FM: *Evaluating GO-based Semantic Similarity Measures* 2007, [`http://xldb.fc.ul.pt/xldb/publications/Pesquita.etal:EvaluatingGO-basedSemantic:2007_document.pdf`].

[28] R Development Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria 2010, [`http://www.R-project.org`]. [3-900051-07-0].

[29] R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2011, [`http://www.R-project.org`]. [ISBN 3-900051-07-0].

[30] Yu G, Li F, Qin Y, Bo X, Wu Y, Wand S: **GOSemSim: an R package for measuring semantic similarity among GO terms and gene products**. *Bioinformatics* 2010, **26(7)**:976–978.

[31] Gentleman R: *Visualizing and Distances Using GO* 2005, [`http://bioconductor.org/packages/2.6/bioc/vignettes/GOstats/inst/doc/GOvis.pdf`].

[32] Faria D, Pesquita C, Couto FM, Falcão AO: *ProteInOn: A Web Tool for Protein Semantic Similarity* 2007, [`http://xldb.fc.ul.pt/xldb/publications/Faria.etal:ProteInOnAWeb:2007_document.pdf`].

[33] Du Z, Li L, Chen CF, Yu PS, Wang JW: **G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery**. *Nucleic Acids Research* 2009, **37(2)**:D345–D349.

[34] Pesquita C, Pessoa D, Faria D, Couto F: **CESSM: Collaborative Evaluation of Semantic Similarity Measures**. *JB2009: Challenges in Bioinformatics* 2009, [`http://xldb.fc.ul.pt/xldb/publications/Pesquita:CESSMCollaborative:2009_document.pdf`].

[35] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS: **UniProt: the Universal Protein knowledgebase**. *Nucleic Acids Research* 2004, **32**:D115–D119.

[36] UniProt-Consortium: **The Universal Protein Resource (UniProt) in 2010**. *Nucleic Acids Research* 2010, **38**:D142–D148.

[37] Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E: **Infrastructure for the life sciences: design and implementation of the UniProt website**. *BMC Bioinformatics* 2009, **10**:136.

[38] Camon E, Magrane M, Barrell D, Binns D, Fleischmann W, Kersey P, Mulder N, Oinn T, Maslen J, Cox A, Apweiler R: **The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro**. *Genome Research* 2003, **13(4)**:662–672.

[39] Camon E, Barrell D, Lee V, Dimmer E, Apweiler R: **The Gene Ontology Annotation (GOA) Database - An integrated resource of GO annotations to the UniProt Knowledgebase**. *In Silico Biology* 2004, **4(1)**:5–6.

[40] Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R: **The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology**. *Nucleic Acids Research* 2004, **32**:D262–D266.

[41] Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R: **The GOA database in 2009-an integrated Gene Ontology Annotation resource**. *Nucleic Acids Research* 2009, **37**:D396–D403.

[42] Teerapabolarn K: **Binomial approximation to the generalized hypergeometric distribution**. *International Journal of Pure and Applied Mathematics* 2013, **83(4)**:559–563.

[43] Mazandu GK, Mulder NJ: **Generation and analysis of large-scale data-driven *Mycobacterium tuberculosis* functional networks for drug target identification**. *Adv Bioinformatics* 2011, **2011**:Article ID 801478.

[44] Blondel VD, Guillaume JL, Lambiotte R, Lefebvreet E: **Fast unfolding of communities in large networks**. *J. Stat. Mech* 2008, **10008**:1–12.