Toward an Integrative Resource of Information Content-based Gene Ontology Measures

Horace Pene Akata Ekete (horace@aims.ac.za)
African Institute for Mathematical Sciences (AIMS)

Supervised by: Dr. Gaston Kuzamunu Mazandu
University of Cape Town, South Africa

22 May 2014

Submitted in partial fulfillment of a structured masters degree at AIMS South Africa

Abstract

Several approaches have been proposed to measure functional similarity between proteins, enabling comparisons between genes or gene products on the basis of their functional annotations. In terms of their conceptions, these approaches are divided into two main groups, namely annotation- and topology-based approaches. Annotation-based approaches have been extensively used in several biomedical and biological applications. In the case of topology-based approaches, each approach has been set with its specific functional similarity measure depending on biological applications for which it has been designed. However, it is not known which measure is the best for a given approach. In addition, existing online tools do not implement functional similarity measures other than the one suggested for the topology-based approaches. We use different types of biological data, including sequence and protein family domain similarity, gene expression and protein-protein interaction data, to assess these measures. We have shown that the performance of each measure depends on the biological applications and a given approach does not always yield the best performance when used with its specific measure. Finally, we have included all these measures in an existing web tool, IT-GOM, which integrates several functional similarity measures, to ensure that protein functional similarity data are conveniently accessible and can effectively be used in protein analyses at the functional level.

Resumé

Plusieures approches ont été proposées pour mésurer la similarité fonctionnelle entre protéines, permettant les comparaisons entre gène ou produit de gènes sur base de leurs descriptions fonctionnelles. Selon leur conceptions, ces approches sont divisées en deux groupes, nommés approches basées sur la description et la topologie. Les approches basées sur la description ont été largement utilisées dans plusieurs applications biomédicales et biologiques. Dans le cas des approches basées sur la topologie, chaque approche a été mise au point avec ses mésures fonctionnelles spécifiques dépendant des applications biologiques pour lesquelles elle a été conçue. Cependant, il n'est pas connu quelle mésure est la meilleure pour une approche donnée. En outre, les outils en ligne existant n'implémentent pas les autres mésures fonctionnelles que celles suggérées pour les approches basées sur la topologie. Nous utilisons les différents types de données, comportant les données de similarité de séquence et famille de protéine, l'expression du gène et l'interaction de protéine-protéine, pour évaluer ces mésures. Nous avions montré que la performance de chaque mésure dépend des applications biologiques et une approche donnée ne donne toujours pas la meilleure performance quand elle est utlisée avec ses mésures spécifiques. Enfin, nous avions inclut toutes ces différentes mésures dans un outil web existant, IT-GOM, qui intègre plusieurs mésures de similarité fonctionnelle, de sorte que les données de similarité fonctionnelle de proténes soient accessible et utilisées éfficacement dans l'analyse des protéines au niveau fonctionnel.

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Horace Pene Akata Ekete, 22 May 2014

Contents

1. Introduction

In biology, there are different representations of functional terms due to the use of various notations from different sources. If each source can use its own notations or vocabularies to describe processes in which genes or proteins are involved, it would be difficult to integrate different vocabularies and to process the dataset produced. Thus, there is a need for standard notations or vocabularies to uniformly represent gene or protein functional vocabularies. This led to the creation of the Gene Ontology (GO) (Ashburner et al., 2000), which produces a standardized gene and protein functional scheme.

1.1 Overview of Gene Ontology

In biology, ontologies are expected to produce an efficient and standardized functional scheme for describing genes and gene products (Mazandu and Mulder, 2012). The Gene Ontology is a database of standardized and controlled set of terms, words and phrases used for representing information about gene and proteins (Ashburner et al., 2000). The main objective of GO is to:

- produce a controlled vocabulary for terms used to characterize protein functions.
- standardize functional representations of gene products in order to facilitate the knowledge sharing and processing.

GO is organized as a directed acyclic graph (DAG), in which two terms are topologically linked by the relation "is_a" or "part_of" meaning that a term child is a subclass (instance) or a component of a parent term. Other relationships between terms exist but do not have influence on the topology of the GO DAG as they are essentially biological.

The Gene Ontology is divided into three ontologies or domains, Biological Process (BP), Molecular Function (MF), and Cellular Component (CC).

- **Biological Process** describes general biological roles in which protein contributes. e.g.metabolism, transport, etc;
- **Molecular Function** refers to a specific task performed by a gene or protein at the molecular level. For example binding, catalysis, etc.
- **Cellular Component** characterizes locations where protein activities are occurring, or different parts of a cell like membrane or organelles.

Each ontology or domain is engineered as a directed acyclic graph (DAG) in which the nodes are GO terms and the edges are relations between terms. A GO term can have several parents and children. A term with no child is called a **leaf** and only one term does not have a parent, and this term is the root of the ontology. All terms more close to the root in the same path of a GO term constitute the **ancestors** of this GO term.

The following section explains the relationships between terms:

1. **"is_a"** relation: indicates that a child is a subclass or an instance of a parent. In this relation, a child has all the characteristics of its parent.

2. **"part_of"** relation: means that a child is a component of a parent. Here a child will have some characteristics of its parent but not all.

Each GO term is characterized by its ID, name, ontology (namespace), definition and relationships with other terms. It can be possible to find the secondary IDs, synonyms, comments, etc.

- **The ID of a GO term**. Each GO term is characterized by a unique identifier, that starts with GO: followed by a zeros padded integer of seven digits. Example of Identifier: GO:0004325.
- **The name of a GO term** is a text describing the term.
- **The namespace of GO term** provides the ontology or domain in which a term belongs. This ontology can be Molecular Function, Biological Process or Cellular Component.
- **The definition of GO term** provides the biological meaning of the term.

Beside these relationships, there also exist some other characteristics that appear in the structure of a term, including:

1. **Secondary Ids** occurring when one or more terms with identical meaning are merged into a single term.

2. **Synonym(s)** of a GO term are other biological definitions of the GO term.

3. **Database cross-references** or dbxrefs, refers to other databases or sources where the term was used.

4. **Comment** provides additional information about the GO term and its usage in terms of number of times the GO term was used to annotate proteins.

Example

This example of GO term, is retrieved from AmiGO at [1]

Id: GO:0048311,

name: mitochondrion distribution Ontology,

namespace: biological process,

def: "Any process that establishes the spatial arrangement of mitochondria between and within cells. Source: GOC:jid ",

synonym: "distribution of mitochondria, mitochondrial distribution, positioning of mitochondria".

Finally, note that the Gene Ontology is a dynamic structure that allows new terms to be easily added as new proteins are discovered and characterized.

1.2 Applying GO in protein analyses

The use of GO data in protein analyses has largely contributed to the improved outcomes of these analyses (Mazandu and Mulder, 2013a). Several term and protein functional similarity measures have been suggested and enable the integration of biological knowledge contained in the GO directed acyclic graph (DAG) structure into different analyses. These measures are designed using relations between terms in the GO DAG.

[1] http://amigo1.geneontology.org/cgi-bin/amigo/term_details?term=GO:0048311

In fact, relations between terms are used to assess the specificity or information content (IC) scores of terms using their positions in the GO DAG. From these IC scores, term semantic similarity and protein or gene functional similarity measures are built.These measures are divided into two main approaches, namely, annotation- and topology-based. In topology-based approaches, GO term IC score depends only on the intrinsic topology of GO structure while in annotation-based approaches IC value depends also on the frequencies at which terms occur in the corpus under consideration.

In this work, we assess the existing information content based functional similarity measures, in order to learn which are the best performing protein analysis at the functional level.

The two main aims of this essay are:

- Assessing the performance of different functional similarity measures in the context of the existing topology- and annotation-based approaches.
- Integrating these measures in the existing tool GO-based functional analysis using term information content measures. This allows the inference of the best functional similarity measures for each topology-based approach and provides more flexibility by allowing users to choose an appropriate model for their application when using a topology-based approach.

The rest of this essay is organized as follows: Chapter 2 provides an overview of different semantic similarity measures suggested in the context of Gene Ontology. In chapter 3, we assess the different functional similarity measures using different types of biological data. Chapter 4 describes how the online web tool, IT-GOM, which integrates large number of protein functional similarity measures, has been extended. We conclude this essay in chapter 5.

2. GO Semantic Similarity Approaches

Several semantic similarity approaches have been developed to quantify semantic similarity scores between terms in the GO structure, enabling comparisons of proteins at the functional level using their GO annotations.

In fact, these semantic similarity approaches have enabled the integration of biological knowledge contained in the GO structure for protein or gene analyses. These different GO semantic similarity approaches are classified into two main categories, namely path or edge- and node-based categories. The edge-based group uses the distance of a term to the root of the ontology in terms of the minimum number of edges separating the term to the root, i.e., the length of the shortest path between the term and the root of the ontology to compute the specificity score of the term in the GO structure. This category has been criticized for producing uniform IC values, which do not reflect the specificity scores of terms. Thus, node-based category has been introduced and uses the position of the term in the ontology structure to compute information content (IC) or semantic value (SV) to score the specificity of the term. This group has been shown to outperform edge-based group and produce scores reflecting the specificity of terms in the ontology.

In this essay, we focus on the node-based category and this chapter explores existing GO-semantic similarity measures suggested in the context of the node- or IC-based group and their associated protein functional similarity measures. Note that different IC-based approaches are classified into two families, namely annotation- and topology-based approaches. The topology-based approaches use only the intrinsic topology of the GO structure whereas annotation-based approaches also use the annotation statistics related to terms in the corpus under consideration.

2.1 Annotation-Based Approaches

The first approach was introduced by Lord (Lord et al., 2003), and deployed in many biological applications. As pointed out previously, annotation-based approaches depend on the frequency at which terms occur in the corpus or dataset under consideration.

The information content or semantic value of a given term t is computed using the following formula:

$$\mathsf{IC}(t) = -\ln(p(t)), \tag{2.1.1}$$

With $p(t)$ being the frequency at which the term t occurs in the protein dataset considered. It is the proportional relation between the occurrence frequency of GO term t, represented by $f(t)$, and that of the root r, represented by $f(r)$, of the ontology considered. Hence,

$$p(t) = \frac{f(t)}{f(r)}. \tag{2.1.2}$$

In the context of GO DAG, a given child should have all the characteristics of its parents, thus the frequencies of a child contribute to that of its parents. Thus, denoting $A(t)$ as the number of proteins annotated with term t in the dataset considered, the frequency of a term t is given by

$$f(t) = \begin{cases} A(t) & \text{if } t \text{ is a leaf} \\ A(t) + \sum_{z \in C_h(t)} A(z) & \text{otherwise,} \end{cases} \tag{2.1.3}$$

where $C_h(t)$ is the set of GO terms having t as a parent.

2.1.1 Resnik and Lin's approach. According to Resnik's approach, the similarity between two terms is the information content of their most informative common ancestor (MICA). Thus, it is given by the following formula:

$$\mathrm{Sim}_{\mathrm{Res}}(s,t) = \mathrm{IC}(a) = \max\{\mathrm{IC}(x), x \in A_s \cap A_t\}, \tag{2.1.4}$$

with a, the most informative common ancestor of s and t, $\mathrm{IC}(x)$ information content value of the term x and $A_x = A \cup \{x\}$ containing the set A of ancestors of the term x and x itself.

The Lin semantic similarity approach takes the information content (IC) of the most informative common ancestor (MICA) of terms being compared and normalized by the average of IC values of these terms. Thus, the similarity between two terms is given by:

$$\mathrm{Sim}_{\mathrm{Lin}}(s,t) = \frac{2 \times \mathrm{IC}(a)}{\mathrm{IC}(s) + \mathrm{IC}(t)}. \tag{2.1.5}$$

Let us note that Lin's approach produces normalized scores (range between 0 and 1) and satisfies the semantic similarity axioms, saying that the similarity between a term and itself should be equal to 1. But Resnik's approach does not and its scores are not normalized.

Some studies (Jain and Bader, 2010; Couto et al., 2003; Pesquita et al., 2008) have normalized Resnik's approach by using either the possible upper bound of IC values (Pesquita et al., 2008), referred to as the Nunif strategy, or the highest IC score in the ontology under consideration (Jain and Bader, 2010; Couto et al., 2003), referred to as the Nmax strategy. Thus, the normalized Resnik semantic similarity scores between two terms are given by

$$\mathrm{S}_{\mathrm{Nunif}}(s,t) = \frac{\mathrm{IC}(a)}{\log_2^N}, \tag{2.1.6}$$

and

$$\mathrm{S}_{\mathrm{Nmax}}(s,t) = \frac{\mathrm{IC}(a)}{\mathrm{IC}_{\max}}, \tag{2.1.7}$$

where N is the number of annotated proteins in the dataset under consideration, $\mathrm{IC}_{\max} = \max\{\mathrm{IC}(x) : x \in N_t\}$ with N_t as the set of all terms used in the annotation set for ontology under consideration.

Note that the Resnik approach is also inconsistent with the hierarchy considered (Mazandu and Mulder, 2013b) as illustrated in figure 2.1. According to the Resnik approach, the semantic similarity score between node 2 and 3 is equal to that between 2 and all descendants of node 3, which is the IC of node 1. This is not consistent for a hierarchical structure in which a child term is expected to be more specific

or to have the higher IC value than its parents. One expects the semantic similarity scores between nodes 2 and all descendants of 3 to be less than that between nodes 2 and 3, that is, node 3 should be semantically similar to node 2 than to any of its descendants.

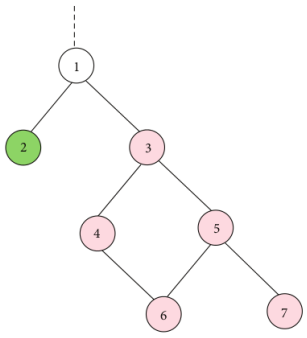

Figure 2.1: Illustration of the inconsistency of the Resnik approach (Mazandu and Mulder, 2013b)

From the above inconsistency, the GO-universal normalization concept referred to as Nunivers (Mazandu and Mulder, 2013b) has been proposed, where the semantic similarity score between terms s and t normalized by the maximum IC values of these terms, given by

$$\mathrm{Sim}_{\mathrm{Nunivers}}(s,t) = \frac{\mathrm{IC}(a)}{\max\{\mathrm{IC}(s), \mathrm{IC}\,(t)\}}. \tag{2.1.8}$$

2.1.2 Improving GO Annotation Semantic Similarity Measures. In order to correct the limitations of annotation-based approaches as described above, some propositions were proposed. Among them, we quote the Relevance similarity measure introduced by Schlicker et al (Schlicker et al., 2006), the Information coefficient developed by Li et al. (Li et al., 2010), and the Graph-based similarity measure suggested by Couto et al. (Couto et al., 2005).

A. Relevance semantic similarity approach. This approach was developed to correct the fact that the Lin approach leads to the overestimation of the similarity value of the terms close to the root. According to Schlicker et al. (Schlicker et al., 2006), this issue is solved by taking into account the IC of the most informative common ancestor using its relative frequency $p(a)$. Therefore, the Sim_{Lin} is weighted with $1 - p(a)$, and the Relevance semantic similarity score between two terms is given by

$$\mathrm{Sim}_{\mathrm{Rel}}(t_1, t_2) = \frac{2 \times \mathrm{IC}(a)}{\mathrm{IC}(s) + \mathrm{IC}(t)} \times (1 - p(a)), \tag{2.1.9}$$

with a being the most informative common ancestor of t_1 and t_2, $p(a)$ the relative frequency obtained by the expression $\exp(-IC(a))$.

The interesting observation with this measure is that the similarity value ranges between 0 and 1. But the semantic similarity score between two identical terms can not be 1.

B. Information Coefficient Semantic Similarity Measure. This measure emphasizes the impact of the location of terms and the relationships between terms in the GO DAG when computing GO term semantic similarity. Thus, the semantic similarity information content is obtained as follows:

$$\mathrm{Sim}_{\mathrm{IC}}(s,t) = \frac{2 \times \mathrm{IC}(a)}{\mathrm{IC}(s) + \mathrm{IC}(t)} \times \left(1 - \frac{1}{1 + \mathrm{IC}(a)}\right). \tag{2.1.10}$$

The authors claim that this score reflects the IC of a GO term and the topology of the GO DAG. However, despite these different corrections, the semantic similarity score between two identical terms can not be 1.

C. Graph-based Semantic Similarity (GraSM). In all the previous approaches, the similarity between two terms is obtained using the MICA between terms. It may overestimate the semantic similarity value. Thus, the GraSM approach (Couto et al., 2005) has been proposed, in which the IC of the most specific ancestor is obtained by computing the average of the IC of all disjunctive common ancestors (DCA), in which case the semantic similarity score between terms s and t is computed as follows:

$$\mathrm{Sim}_{\mathrm{GraSM}}(s,t) = \frac{\sum_{a \in \mathrm{DCA}(s,t)} \mathrm{IC}(a)}{|\mathrm{DCA}(s,t)|}, \tag{2.1.11}$$

where $|\mathrm{DCA}(s,t)|$ is the number of DCA between s and t.

Two ancestors x and y of a term t are said to be disjunctive, if there exists a path between t and x which passes by y. Hence, GraSM computes the similarity of two terms as the average of the information content of their disjunctive common ancestors (DCAs).

2.1.3 Illustration of GO Annotation-based Approches. In order to show how the different annotation-based approaches work, we illustrate this example by computing the Resnik and Lin approaches, improvements of Lin, including $\mathrm{Sim}_{\mathrm{Rel}}$ and $\mathrm{Sim}_{\mathrm{IC}}$ and different normalization proposed in order to keep the Resnik semantic similarity approach between 0 and 1. We use the sub-GO DAG shown in figure 2.2, downloaded from the AmiGO tool [1] with GO term mitochondrion distribution, having the GO Id GO:0048311. The table 2.1 shows the IC of each sub-GO term and their normalized scores.

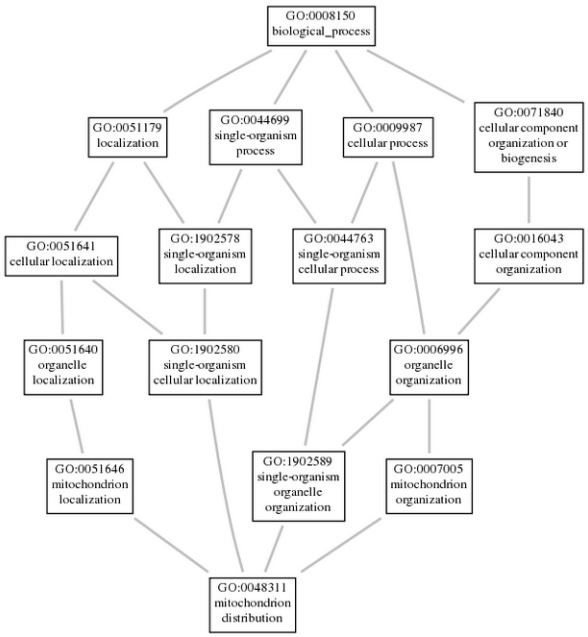

Figure 2.2: Sub-graph of mitochondrion distribution

From figure 2.2, let us assume that GO:0008150 = r, GO:0009987 = m, GO:0044763 = a, GO:0006996 = b, GO:0007005 = c, GO:1902589 = d, GO:0048311 = l.

[1] http://amigo.geneontology.org/amigo

To find the similarity between the terms single organism organelle organization (c), and mitochondrion organization (d), the Lin and Resnik measures use the most informative common ancestor, which is organelle organization (b).

$$\mathrm{Sim_{Res}(c,d)} = \mathrm{IC(b)} = 8.55707.$$

$$\mathrm{Sim_{Lin}(c,d)} = \frac{2 \times \mathrm{(b)}}{\mathrm{IC(c)} + \mathrm{IC(d)}} = \frac{2 \times 8.55707}{11.93007 + 9.85620} = 0.78554.$$

$$\mathrm{Sim_{Nunivers}(c,d)} = \frac{\mathrm{IC(a)}}{\max\{\mathrm{IC(c)}, \mathrm{IC(d)}\}} = \frac{8.55707}{\max\{11.93007, 9.85620\}} = 0.71726.$$

$$\mathrm{Sim_{Rel}(c,d)} = \frac{2 \times \mathrm{IC(b)} \times (1 - p(\mathrm{b}))}{\mathrm{IC(c)} + \mathrm{IC(d)}} = \frac{2 \times 8.55707 \times (1 - 0.00019)}{11.93007 + 9.85620} = 0.78539.$$

$$\begin{aligned}\mathrm{Sim_{IC}(c,d)} &= \frac{2 \times \mathrm{IC(b)}}{\mathrm{IC(c)} + \mathrm{IC(d)}} \times \left(1 - \frac{1}{1 + \mathrm{IC(b)}}\right) = \frac{2 \times 8.55707}{11.93007 + 9.85620} \times \left(1 - \frac{1}{1 + 8.55707}\right) \\ &= 0.70334.\end{aligned}$$

The GraSM approach will now use the disjunctive common ancestors between mitochondrion organization and single organism organelle organization, instead of considering the most informative common ancestor. Thus, it uses the path (GO:0007005, GO:0006996, GO:0009987, GO:0044763). Therefore for GraSM, we use the average of the IC of their common disjunctive ancestors: $\mathrm{GO:0006996}$ (organelle organization) and $\mathrm{GO:0009987}$ (cellular process).

Term	$p(t)$	IC	$\mathrm{IC_n}(t)$	$\mathrm{IC_{max}}(t)$
GO:0008150	1.00000	0.00000	0.00000	0.00000
GO:0009987	0.26070	1.34438	0.09122	0.06094
GO:0044763	0.03354	3.39490	0.23036	0.15390
GO:0006996	0.00019	8.55707	0.58064	0.38792
GO:1902589	0.00005	9.85620	0.66879	0.44682
GO:0007005	6.58925e-06	11.93007	0.80952	0.54083
GO:0048311	1.56291e-07	15.67154	1.06340	0.71045

Table 2.1: Results of GO annotation-based approaches.

All the annotation-based approaches rely on the annotation statistics related to terms and they are unable to quantify the specificity or IC score of terms that have not been used to annotate a protein or terms that do not occur in the corpus under consideration. Thus, topology-based approaches have been developed in order to overcome this issue.

2.2 GO Topology-Based Approaches

In the GO DAG, a given term is expected to have a unique information content value that should not depend on the corpus under consideration. Thus, the mapping between proteins and the GO annotations

provided by the GO annotation (GOA) project (Barrell et al., 2009) allows us to solve this issue of the uniqueness of the IC of a given term. However, the fact that the IC for annotation-based approaches depends on the annotation statistics related to terms may induced biased IC values since a term can be rarely used, but not necessarily very specific considering its position in the GO DAG (Mazandu and Mulder, 2013b). Then when a term has not been used, it will be difficult to know its IC value. Thus, we need an approach that depends only on the intrinsic topology of the GO DAG, and this refers to a topology-based approach.

Topology-based approaches overcome the effect of annotation dependence (of annotation-based measures) to provide an effective way to measure similarity between proteins based only on the GO DAG, helping to obtain a fixed and well-defined information content for a given GO term independent of the corpus under consideration (Mazandu and Mulder, 2013b). These topology-based approaches include the Zhang et al. (Zhang et al., 2006) , Wang et al. (Wang et al., 2007) , and GO-universal approach introduced by Mazandu and Mulder (Mazandu and Mulder, 2012).

2.2.1 Zhang Semantic Similarity Measure. As said in the annotation-based approaches, the IC of a term is given by

$$\mathrm{IC}(t) = -\ln(p(t)), \tag{2.2.1}$$

where $p(t)$ is the frequency at which the term t occurs in the dataset considered, this frequency is called D-value. In the case of the topology-based approach introduced by Zhang et al., this frequency is then calculated independently for each ontology and given by

$$p(t) = \frac{f(t)}{f(r)}, \tag{2.2.2}$$

where $f(r)$ is the frequency (count) of the root term in the ontology under consideration and $f(t)$ the count of the term t, depending only on the children of a given GO term and equal to the sum of counts of all its children. $f(t)$ is computed using a recursive formula starting from leaves in the hierarchical structure and given by

$$f(t) = \begin{cases} 1 & \text{if } t \text{ is a leaf} \\ \sum\limits_{z \in C_h(t)} f(z) & \text{otherwise.} \end{cases} \tag{2.2.3}$$

Thus, these semantic similarity between two terms is computed as follows:

$$\mathrm{Sim}_{\mathrm{Zhang}}(s,t) = \frac{2 \times \mathrm{IC}(a)}{\mathrm{IC}(s) + \mathrm{IC}(t)}. \tag{2.2.4}$$

2.2.2 Wang Semantic Similarity Measure. Here, the specificity score of a GO term is computed by using all its parents and relationships between them. In other words, we will consider the structure from the root to the term considered. Thus, the semantic value of a given term x is computed using the semantic contribution factors we of edges linking the term x to a parent t, denoted $\mathrm{S_x}(t)$, and given by

$$\mathrm{S_x}(t) = \begin{cases} 1 & \text{if } t = \mathrm{x} \\ \max\{we * \mathrm{S_x}(t') : t' \in C_h(t)\} & \text{otherwise,} \end{cases} \tag{2.2.5}$$

with $C_h(t)$, the set of children of the term t, we as the semantic contribution factor for "is_a" and "part_of" relations set to 0.8 and 0.6, respectively.

Thus the information content or semantic value of a term x is calculated as follows:

$$\mathrm{IC_w(x)} = \sum_{t \in A_\mathrm{x}} \mathrm{S_x}(t), \tag{2.2.6}$$

where $A_\mathrm{x} = A \cup \{\mathrm{x}\}$ and A denotes the set of ancestors of the term x and $\mathrm{S_x(t)}$ the relation between the term x and its parent t.

The semantic similarity $(\mathrm{Sim_w})$ between two terms t_1 and t_2 is given by

$$\mathrm{Sim_w}(s,t) = \frac{\sum_{\mathrm{x} \in A_s \cap A_t} (\mathrm{S}_s(\mathrm{x}) + \mathrm{S}_t(\mathrm{x}))}{\mathrm{IC_w}(s) + \mathrm{IC_w}(t)}, \tag{2.2.7}$$

where A_i is the set of ancestors of the term.

Let us note that when a GO contains a great number of nodes, it might become expensive to compute the similarity between terms. Furthermore, the semantic similarity score is a function of semantic contribution factors of the relationship.

2.2.3 GO-universal Approach. According to this approach, the IC of a given term is similar to the biological content of the term converted into a numeric value called "topological information", and computed using its immediate parents' topological positions (Mazandu and Mulder, 2013b). The topological position characteristic of a term t is computed by taking into account all its parents included in the set $P_t = \{z : (z,t) \in L_{GO}\}$, with L_{GO} the set of all edges (links) in the GO DAG (Mazandu and Mulder, 2013b). This topological position is represented by $p(t)$ and given by

$$p(t) = \begin{cases} 1 & \text{if t is a root} \\ \prod_{z \in P_t} \frac{p(z)}{|C_h(z)|} & \text{otherwise,} \end{cases} \tag{2.2.8}$$

with $|C_h(z)|$ the number of children having the term z as parent.

Hence, this implies that the $\mathrm{IC}(t) = -\ln(p(t))$.

and the GO term semantic similarity $(\mathrm{Sim_{univ}})$ between t_1 and t_2 is computed as follows:

$$\mathrm{Sim_{univ}}(s,t) = \frac{\mathrm{IC}(a)}{\max\{\mathrm{IC}(s), \mathrm{IC}(t)\}}, \tag{2.2.9}$$

with a being the most informative common ancestor shared by terms s and t.

2.2.4 Illustrating GO Topology-Based Measures. Let us consider the directed acyclic graph as shown in the below figure 2.3. our graph contains 6 terms with s and y respectively the root and the leaf.

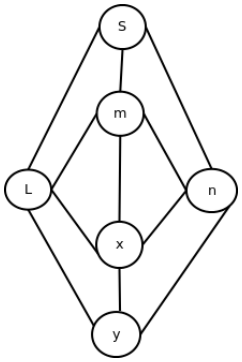

Figure 2.3: Hierarchical structure illustrating topology-based approaches.

- According to the Zhang et al. approach, the D-value of the term y, denoted by $\mathrm{f(y)}$ and that of the term x, denoted by $\mathrm{f(x)}$ is 1, because y is a leaf and x has only one child, that is y.

  The term (L) has two children (x) and (y), hence we obtain its count value as follows,

  $$\mathrm{f(L) = f(x) + f(y) = 1 + 1 = 2.} \tag{2.2.10}$$

  Likely,

  $$\mathrm{f(s) = f(L) + f(m) + f(n) = 2 + 5 + 2 = 9.} \tag{2.2.11}$$

  Therefore, the D-value of (y) is

  $$\mathrm{D(y) = \frac{f(y)}{f(s)} = \frac{1}{9} = 0.11111.} \tag{2.2.12}$$

  By using the same principles the value of the other terms is shown in the table 2.2

- According to the Wang et al. approach, the S-values of the term y, assuming that all relationships between terms are "is_a", i.e, the link semantic value is 0.8, are computed as follows:

  $\mathrm{S_y(y) = 1.}$

  $\mathrm{S_y(x) = \max\{we * S_y(y)\} = \max\{we \times S_y(y)\} = 0.8.}$

  $\mathrm{S_y(n) = 0.8.}$

  $\mathrm{S_y(l) = \max\{we * S_y(x)\} = \max\{0.64, 0.8\} = 0.8.}$

  $\mathrm{S_y(m) = \max\{0.8 * S_y(x), 0.8 * S_y(l), 0.8 * S_y(n)\} = \max\{0.64, 0.64, 0.64\} = 0.64.}$

  $\mathrm{S_y(s) = \max\{0.8 * S_y(l)\}, 0.\,8 * S_y(m), 0.\,8 * S_y(n) = \max\{0.64, 0.512, 0.64\} = 0.64.}$

Hence, the IC of the term (y) is given by

$$\mathrm{IC_w(y)} = \sum_{t \in H} S_{\mathrm{y}}(t), \tag{2.2.13}$$

with H as the set of ancestors of term y, including y itself. H describes the set of all the terms in the ontology considered.

The table 2.2 indicates the IC value of other terms, obtained by using the same principle.

- According to the GO-universal approach, the topological characteristic position of the root s is $p(t) = 1$, and its IC will be $\mathrm{IC}(\mathsf{s}) = 0$. For other terms, the topological characteristic positions are found using the formula (2.2.8), and are obtained as follows:

  $p(\mathsf{m}) = \frac{p(\mathsf{s})}{3} = \frac{1}{3}$.

  $p(\mathsf{n}) = \frac{p(\mathsf{s})}{3} \times \frac{p(\mathsf{m})}{3} = \frac{1}{3} \times \frac{1}{9} = \frac{1}{27} = p(l)$.

  $p(\mathsf{x}) = \frac{p(\mathsf{n})}{2} \times \frac{p(\mathsf{m})}{3} \times \frac{p(l)}{2} = \frac{1}{9} \times \frac{1}{54} \times \frac{1}{729} = 2.822541e{-}06$.

  $p(\mathsf{y}) = \frac{p(\mathsf{n})}{2} \times \frac{p(\mathsf{l})}{2} \times \frac{p(\mathsf{x})}{1} = \frac{1}{54} \times \frac{1}{54} \times \frac{1}{354294} = 9.67941e{-}10$.

Term	level	count	Zhang et al. D-value	Zhang et al. $\mathrm{IC_z}$	Wang et al. $\mathrm{IC_w}$	GO Universal $p(z)$	GO Universal IC
s	0	9	1	0.00000	1	1.00000	0.00000
m	1	5	$5/9$	0.58778	1.8	0.33333	1.09861
n	2	2	$2/9$	1.50407	2.6	0.03703	3.29583
l	2	2	$2/9$	1.50407	2.6	0.03703	3.29583
x	3	1	$1/9$	2.19722	4.01	2.8e-06	12.77788
y	4	1	$1/9$	2.19722	4.68	9.6e-10	20.75585

Table 2.2: Table illustrating different IC values for the hierarchy structure in figure 2.3.

The Zhang et al. approach has obtained the same value of the IC between the terms x and y, this can have the impact in the computation of semantic similarity score. The Wang et al. approach uses the contribution factor we when computing the similarity between a term and its child. However, it does not take into account the position of terms in the ontology considered. These issues are solved by the GO-universal approach by considering the topological characteristic position of each term and this enable to find the different values of IC for the terms x and y in the context of our example.

2.3 Protein Functional Similarity Measures

One of the main objectives of this project is to compare the similarity between proteins at functional level. This chapter explores protein functional similarity measures proposed in the context of GO.

Protein can be annotated by a set of terms since it can perform more than one biological functions from them. Thus, a protein functional similarity can be measured by combining the GO term semantic similarity annotated to these proteins by using basic statistical measures, such as Best-Match Average (BMA), Average Best-Matches (ABM), Average (Avg), Maximum (Max) (Mazandu and Mulder, 2013a; Seuneu and Mulder, 2010).

2.3.1 Average Method (Avg). The similarity between two proteins p_1 and $p2$ is computed by calculating the average of all pairs of GO terms (s, t) with $s \in \mathrm{GO}_{p_1}$ and $t \in \mathrm{GO}_{p_2}$.

The formula below describes how to compute that similarity.

$$\mathrm{Avg}(p_1, p_2) = \frac{1}{|\,\mathrm{GO}_{p_1}\,| \times |\,\mathrm{GO}_{p_2}\,|} \sum_{(s,t)\in \mathrm{GO}_{p_1} \times \mathrm{GO}_{p_2}} \mathrm{S}(s,t), \tag{2.3.1}$$

with $\mathrm{S}(s,t)$ being the similarity between GO terms s and t, GO_{p_1} a set of terms annotating proteins p_1, GO_{p_2} a set of terms which annotate proteins p_2, $|\,\mathrm{GO}_{p1}\,|$ and $|\,\mathrm{GO}_{p2}\,|$ are the number of terms annotating the first and second proteins p_1 and p_2, respectively.

2.3.2 Maximum Method (max). The maximum method computes the similarity between proteins by finding the maximum of the similarity of all pairs. This maximum is given by the following formula.

$$\mathrm{Max}(p_1, p_2) = \max\{\mathrm{S}(s,t) : s \in \mathrm{GO}_{p_1} \ \ and\ t \in \mathrm{GO}_{p_2}\}. \tag{2.3.2}$$

2.3.3 Best-Match Average Method (BMA). The Best-Match Average Method computes the similarity between proteins p_1 and p_2 by finding the average of the mean of the maximum semantic similarity value between terms s and t annotating p_1 and p_2, respectively, and that between terms s and t annotating p_1 and p_2, respectively. This is given by the following formula:

$$\mathrm{BMA}(p_1, p_2) = \frac{1}{2}\left(\frac{1}{|GO_{p_1}|}\sum_{s\in GO_{p_1}} \max_{t\in GO_{p2}} S(s,t) + \frac{1}{|GO_{p_2}|}\sum_{s\in GO_{p_2}} \max_{t\in GO_{p1}} S(s,t)\right). \tag{2.3.3}$$

2.3.4 Average Best-Matches. The Average Best-Matches measure computes the functional similarity between two proteins p_1 and p_2 as follows:

$$\mathrm{ABM}(p_1, p_2) = \frac{1}{|\,\mathrm{GO}_{p_1}\,| + |\,\mathrm{GO}_{p_2}\,|}\left(\sum_{s\in \mathrm{GO}_{p_1}} \max_{t\in \mathrm{GO}_{p_2}} \mathrm{S}(s,t) + \sum_{s\in \mathrm{GO}_{p_2}} \max_{t\in \mathrm{GO}_{p_1}} \mathrm{S}(s,t)\right). \tag{2.3.4}$$

Instead of performing all the previous computations in order to find protein functional similarity scores, one can directly use term IC values to compute protein functional similarity scores. These methods are referred to as direct IC-based methods and are derived from the Jaccard Index based on the Tversky ratio model of similarity (Tversky, 1977). These measures include SimGIC (Pesquita et al., 2008, 2007), SimDIC, SimUIC, and SimUI (Mazandu and Mulder, 2013a).

2.3.5 SimGIC. The Sim GIC uses the following formula to calculate the functional similarity between proteins,

$$\mathrm{SimGIC}(p_1, p_2) = \frac{\sum_{t\in \mathrm{A}\cap \mathrm{B}} \mathrm{IC}(t)}{\sum_{t\in \mathrm{A}\cup \mathrm{B}} \mathrm{IC}(t)}, \tag{2.3.5}$$

with A as the set of proteins annotating $p1$, and B that of proteins annotating $p2$. The same definitions will be used for SimGIC, SimDIC, SimUIC, SimUI.

2.3.6 SimDIC. The SimDIC uses the following formula to calculate the functional similarity between proteins:

$$\mathrm{SimDIC}(p_1, p_2) = \frac{2 \times \sum_{t\in \mathrm{A}\cap \mathrm{B}} \mathrm{IC}(t)}{\sum_{t\in \mathrm{A}} \mathrm{IC}(t) + \sum_{t\in \mathrm{B}} \mathrm{IC}(t)}. \tag{2.3.6}$$

2.3.7 SimUIC. The SimGIC uses the following formula to calculate the functional similarity between proteins,

$$\mathrm{SimUIC}(p_1, p_2) = \frac{\sum_{t \in \mathrm{A} \cap \mathrm{B}} \mathrm{IC}(t)}{\max\left\{ \sum_{t \in \mathrm{A}} \mathrm{IC}(t), \sum_{t \in \mathrm{B}} \mathrm{IC}(t) \right\}}. \tag{2.3.7}$$

2.3.8 SimUI. . The SimUI is a particular case of SimGIC and the functional similarity between terms is given as follows:

$$\mathrm{SimUI}(p1, p2) = \frac{|\,\mathrm{A} \cap \mathrm{B}\,|}{|\,\mathrm{A} \cup \mathrm{B}\,|}, \tag{2.3.8}$$

where $|A \cap B|$, denotes the number of terms annotating $p1$ and $p2$, and $|A \cup B|$, the number of terms annotating $p1$ or $p2$.

2.4 Summary

In this chapter, we have explored the annotation and topology-based approaches and different semantic similarity measures. We have observed that GO annotation-based approaches are limited since they have the total dependence on the annotation statistics, thus, Lin's axiom is not satisfied. Whereas, GO topology-based approaches have overcame these limitations by using only the structure of GO.

3. Assessing GO Semantic Similarity Measures Performance

In chapter 2, we presented all Information content-based approaches (annotation- and topology-based approaches) and semantic similarity measures between terms. They have been compared on the basis of their conception and formula.

In this chapter, we evaluate these different approaches and functional similarity measures using external datasets downloaded from the Collaborative Evaluation of Semantic Similarity Measures (CESSM) tool online (Pesquita et al., 2009) and different protein-protein interaction databases.

Proteins of GO annotations were obtained from the GOA-uniProtKB project, release 2014-04 of April 9, for three ontologies namely, Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) from the GO database.

3.1 Using Sequence, Pfam domain, and EC similarity data

A protein sequence or very often a protein domain, which is a particular sub-sequence that determines protein functional properties, is used to assign a function to the protein under consideration (Mazandu and Mulder, 2011). An Enzyme Commission (EC) number is used to describe an enzyme in biochemical reactions. Thus, sequence, Pfam domain and EC dataset are protein functional data that can be used to assess different protein functional similarity measures.

CESSM is an online tool for automated evaluation of GO-based semantic similarity measures, that enables the comparison of new measures against previously published ones in terms of performance against sequence similarity, Pfam domain and EC (Enzyme Commission) (Pesquita et al., 2008).

Figure 3.1 below shows the comparison of performance of different annotation and topology-based approaches for GO BP, MF, and CC ontologies using Pearson correlation with Enzyme Commission (EC), Pfam and Sequence similarity. These correlations between FSMs were obatined by CESSM online tool. The measures producing higher correlation scores are the best.

The figure contains three sub-figures. For each sub-figure, the Y-axis represents the value of correlation scores (Sequence, Pfam, and EC) and the X-axis contains all the approaches (Annotation and Topology-based) with their corresponding models for all the ontologies (BP, MF, and CC) as described in the legend.

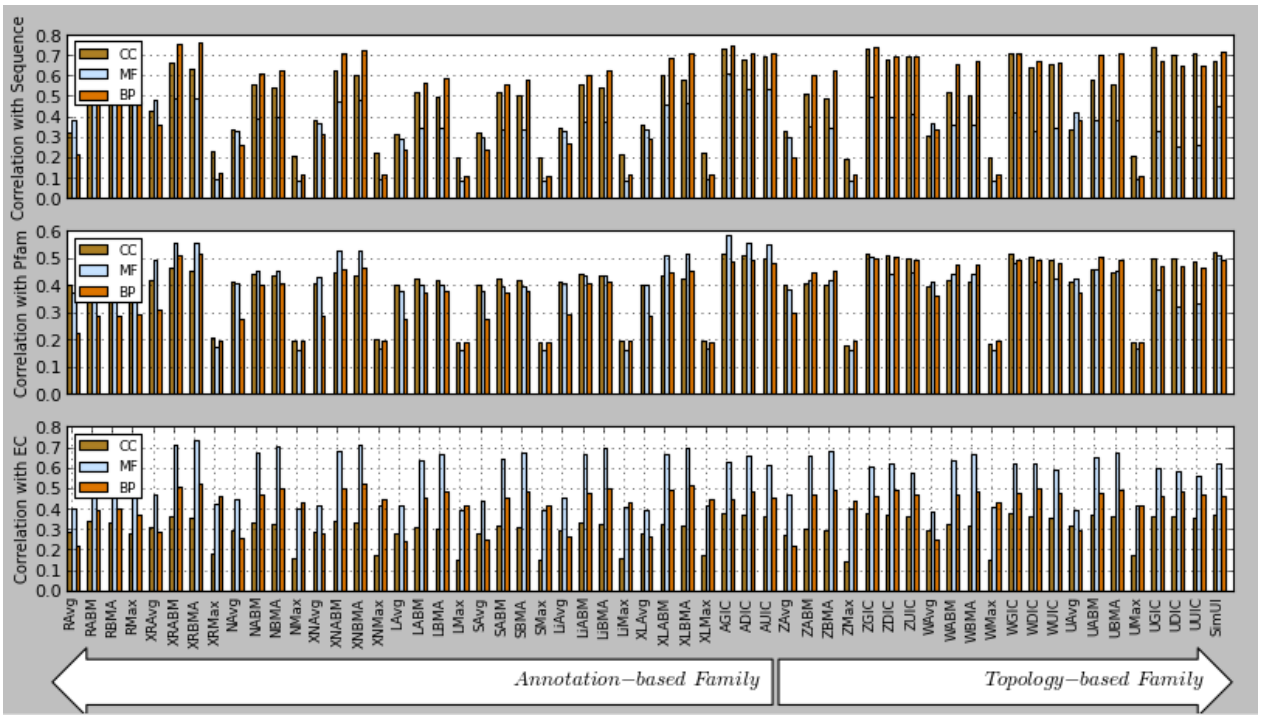

Figure 3.1: Performance evaluation in terms of Pearson's correlation scores.

For X-axis labels R, N, L, Li, S, X, A, Z, W and U represent the term semantic similarity approaches and stand for Resnik, Nunivers, Lin, Li, Relevance, XGraSM, Annotation-based, Zhang, Wang and GO-universal, respectively. The suffix GIC, UIC and DIC represent SimGIC, SimUIC, and SimDIC measures respectively. In cases where the prefix X is used, it is immediately followed by the approach prefix.

3.1.1 Performance of Annotation-based approaches. For GOBP Ontology

For the Resnik approach, the BMA model performs better than the others for Sequence, Pfam, and EC similarities. Thus, BMA is a good method when using the Resnik approach.

For the Lin approach, the BMA perfoms better for the Sequence , Pfam, and EC similarity measures. Thus, it is a good method compared to others.

For the Relevance approach, the BMA model is still better with Sequence, Pfam, and EC similarity measures. It is a good method for this approach.

For Li et al, XGraSM-Resnik, XGraSM-Lin, XGraSM-Nunivers approaches, the BMA method performs better with Sequence, Pfam, and EC measures since it has the highest value of correlation.

The observations above reveal that the most appropriate method for annotation-based approaches is BMA, which performs well for Sequence, Pfam and EC similarity measures. In the context of direct IC-based methods, SimDIC is more appropriate than SimGIC or SimUIC since it has been shown to provide overall best performance.

We suggest using the BMA method for all the annotation-based approaches when we are computing the similarity scores with Sequence, Pfam and EC similarity under the BP ontology. If this is in agreement with what we know, it is not the case for direct IC-based methods where SimDIC provide an overall best performance instead of SimGIC.

For MF Ontology

For the Resnik approach, the Max method performs well in Sequence and Pfam similarity except for BMA that performs well for EC.

For the Lin approach, BMA is shown to have the highest correlation value for Sequence, Pfam and EC data except for ABM which produces the same correlation score as BMA for Sequence and Pfam data. Hence, BMA or ABM can be suggested for Sequence similarity and BMA for all the similarities.

For the Nunivers approach, BMA is still better than Avg, ABM, and Max methods for Sequence, Pfam and EC data.

For Lin, Relevance, XGraSM-Resnik, XGraSM-Lin and XGraSM-Nunivers approaches, ABM and BMA are better with Sequence and Pfam similarity. In addition, BMA shows the best performance for EC data, it is the best method.

SimGIC has the highest correlations with Sequence and Pfam similarity, except for EC that performs well by with.

For CC Ontology

For the Resnik approach, ABM has the highest correlation with Sequence similarity, Pfam and EC.Thus it is the best method.

For the Lin, Nunivers, Relevance, XGraSM-Resnik, XGraSM-Lin and XGraSM-Nunivers approaches, ABM is still the best with the highest correlation for Sequence similarity, Pfam and EC. Therefore ABM is suggested when performing semantic similarity for the CC ontology.

3.1.2 Performance of topology-based approaches. For GO BP Ontology.

For the GO-universal approach, ABM performs better by showing the highest correlations with Sequence similarity and Pfam, and BMA has the highest correlations with EC similarity. The SimDIC measure performs better in Pfam and EC similarity.

For the Wang et al. approach, the BMA method performs well since it has the highest correlation with all similarity measures (Sequence similarity, Pfam, and EC). The SimGIC method is good for Sequence and Pfam similarity measures, except for EC where SimDIC performs well. The authors of this approach suggest the use of ABM, which does not perform better than BMA in this context.

For the Zhang et al. approach, BMA has the highest correlations with all the similarity measures (Sequence similarity, Pfam, and EC). It is the best method compared to others. SimDIC performs well with Pfam and EC similarity. However, the authors suggest the use of the Avg measure as for the Resnik approach.

Therefore, BMA is shown to be the best method on topology-based approaches for the BP ontology.

For GO MF Ontology.

For the GO-universal approach, ABM performs well with Pfam, Avg performs well with Sequence similarity, and BMA performs well with EC. SimGIC performs well with the highest correlations in all similarity measures. However, BMA was initially suggested.

For the Wang et al. approach, the BMA method performs well with Pfam and EC, Avg has the highest correlation sequence similarity and the ABM method performs well with Pfam. Thus the best method is BMA when compared to the ABM, Max and Avg. SimGIC also perfoms well.

For the Zhang et al. approach, ABM has the highest correlation with Sequence similarity and Pfam, BMA has the highest correlation with Sequence similarity and EC. SimGIC has the highest correlation with Sequence similarity and Pfam.

For GO CC Ontology.

For the GO-universal, Wang et al., Zhang et al., approaches, the best method is ABM with the highest correlation in Sequence similarity, Pfam, and EC. The same applies for SimGIC.

SimUI, has the highest correlation with Sequence similarity in the BP ontology, it performs well with Pfam under the MF ontology, and has the highest correlation with EC in the CC ontology.

Different observations above are summarized in table 3.1, 3.2 and 3.3, for BP, CC and MF ontologies, respectively.

Biological process (BP)

Approach	Sequence	Pfam	EC	Best
Annotation-based	BMA, SimGIC	BMA, SimDIC	BMA, SimDIC	BMA, SimDIC
GO-universal	ABM, SimGIC	ABM, SimDIC	BMA, SimDIC	ABM, SimDIC
Wang et al.	BMA, SimGIC	BMA, SimGIC	BMA, SimDIC	BMA, SimGIC
Zhang et al	BMA	BMA, SimDIC	BMA, SimDIC	BMA, SimDIC
Observation	BMA and SimDIC			

Table 3.1: Table summarizing performing measures for the BP ontology.

Cellular Component (CC)

Approach	Sequence	Pfam	EC	Best
Annotation-based	ABM, SimGIC	ABM, SimGIC	ABM, SimGIC	ABM, SimGIC
GO-universal	ABM, SimGIC	ABM, SimGIC	ABM, SimDIC	ABM, SimGIC
Wang et al.	ABM, SimGIC	ABM, SimGIC	ABM, SimGIC	ABM, SimGIC
Zhang et al	ABM, SimGIC	ABM, SimGIC	ABM, SimGIC	ABM, SimGIC
Observation	ABM and SimGIC			

Table 3.2: Table summarizing performing measures for the CC ontology.

Molecular Function (MF)

Approach	Sequence	Pfam	EC	Best
Annotation-based	BMA, ABM, SimGIC	BMA, ABM, SimGIC	BMA	BMA, SimGIC
GO-universal	Avg, SimGIC	ABM, SimGIC	BMA, SimGIC	SimGIC
Wang et al.	Avg, SimGIC	BMA, ABM, SimGIC	BMA, SimGIC	BMA, SimGIC
Zhang et al	ABM, BMA, SimGIC	ABM, SimGIC	BMA , SimDIC	BMA, SimGIC
Observation	BMA and SimGIC			

Table 3.3: Table summarizing performing measures for MF ontology.

3.2 Using protein-protein interaction and expression data

We have also assessed different measures in terms of their ability to capture functional coherence in a human protein-protein interaction (PPI) network based on how interacting proteins are functionally related to each other. Human PPI datasets were downloaded from several different PPI databases, including the IntAct, DIP, BIND, MIPS, MINT, BioGRID databases, and integrated into a single network in which only interactions predicted by at least two different approaches and found in the STRING dataset are considered, to reduce the impact of false positives. This produced a human PPI network with 6031 interactions from which a total of 5365 and 5648 interactions with both interacting partners were among 31098 and 34125 proteins annotated with respect to the GO BP and CC ontologies, respectively.

We have considered the set of these 5365 and 5648 interactions are considered as a positive set, while the negative set consists of the same number of interactions randomly selected among annotated human proteins pairs. This is consistent as the chance of randomly selecting a detected PPI is very small (less than 0.0011%). We only considered proteins annotated with BP and CC terms in the network produced since two proteins that interact physically are more likely to be involved in similar biological processes or localized in the same cellular component, but there is no guarantee that they share molecular functions (Mazandu and Mulder, 2012). The classification power of different functional similarity measures was tested using Receiver Operating Characteristic (ROC) curve analysis, which assesses the Area Under the Curve (AUC), plotting the true positive rate or sensitivity vs the false positive rate or 1-specificity. This AUC value is used as a measure of discriminative power and a realistic classifier must have an AUC larger than 0.5.

For expression data, we use the human co-expression network retrieved from the Bossi et al. (Bossi and Lehner, 2009) and the STRING human network. We retrieved 7228 co-expressed protein pairs of which a total of 6998 pairs have both proteins found among 31098 human proteins annotated with BP terms. We are only considering the BP ontology as co-expressed genes are more likely to share common processes and may at least belong to the same pathway or contribute to a similar biological process (Mazandu et al., 2011). We partitioned these co-expressed proteins into different clusters using the Blondel et al. method (Blondel et al., 2008) and the corresponding partition is considered to be a ground truth, i.e., the true partition of the actual co-expressed network. Thereafter, the interactions from the co-expressed network are weighted using functional similarity scores and proteins clustered using the same clustering method. We assessed the clustering power of a given functional similarity measure by comparing this clustering result to the ground truth using Normalized Mutual Information (NMI) and Rank Index (RI) of pairwise cluster memberships (Steinhaeuser and Chawla, 2010).

The following table shows us the combination between the Ground Truth and clustering.

Clustering	Ground Truth	
	$G(P_i) = G(P_j)$ or (P)	$G(P_i) \neq G(P_j)$ or (N)
$C(P_i) = G(P_j)$ or (P)	a	c
$C(P_i) \neq C(P_j)$ or (N)	b	d

Table 3.4: The combination of proteins interaction in Ground Truth and Clustering.

The Rand Index is obtained by

$$\text{RandIndex} = \frac{a+d}{a+b+c+d} = \frac{a+d}{\binom{n}{2}}, \tag{3.2.1}$$

where a is the pairs of proteins i and j belonging to the same Ground Truth and Clustering, b the pairs of proteins i and j which are located in the same Ground Truth but different clusterings, c the pairs of proteins i and j within the same Clustering but different Ground truth, and finally, d the pairs of proteins i and j situated in different Ground Truth and Clustering network. The Rand Index captures the measure to which the two partitions approach one another. It is also called Accuracy (Steinhaeuser and Chawla, 2010).

Similar to the Rand Indices, NMI assumes that the network was partitioned into Ground truth (true partition) and in each Ground truth i, every node or element v has been assigned a label. For a given clustering p having kp partitions, each with n_l^p proteins, $l = 1, ..., kp$, the entropy $\mathcal{H}(p)$ of p is given by

$$\mathcal{H}(p) = -\sum_{l=1}^{kp} \frac{n_l^p}{n} \log\left(\frac{n_l^p}{n}\right), \tag{3.2.2}$$

and the mutual information between Ground truth (true partition) G and clustering network structure C can thus be computed as

$$\mathrm{I}(G,C) = \sum_{i=1}^{kG}\sum_{j=1}^{kC} \log\left(\frac{\frac{n_{ij}^{GC}}{n}}{\frac{n_i^G}{n} \times \frac{n_j^C}{n}}\right), \tag{3.2.3}$$

where n_i^G is the number of nodes (elements) in the Ground truth i, n the total number of elements in all the Ground truth (true partitions) located in the network and n_j^G that of nodes (elements) in the clustering j. Normalizing by the maximum value, NMI is given by the following expression:

$$\text{NMI} = -\frac{2\sum_{i=1}^{kG}\sum_{j=1}^{kC} n_{ij}^{GC} \log\left(\frac{n_{ij}^{GC} \times n}{n_i^G \times n_j^C}\right)}{\sum_{i=1}^{kG} n_i^G \log\left(\frac{n_i^G}{n}\right) + \sum_{j=1}^{kC} n_j^C \log\left(\frac{n_j^C}{n}\right)}, \tag{3.2.4}$$

with $n_{i,j}^{GC}$, the number of nodes (elements) which are in the intersection between Ground Truth i and Clustering j, kG and kC the number of true partitions in the network and that of clustering respectively.

After the computation of the AUCs using R software, and that of RI and NMI using python code, the results obtained are shown in figure 3.2.

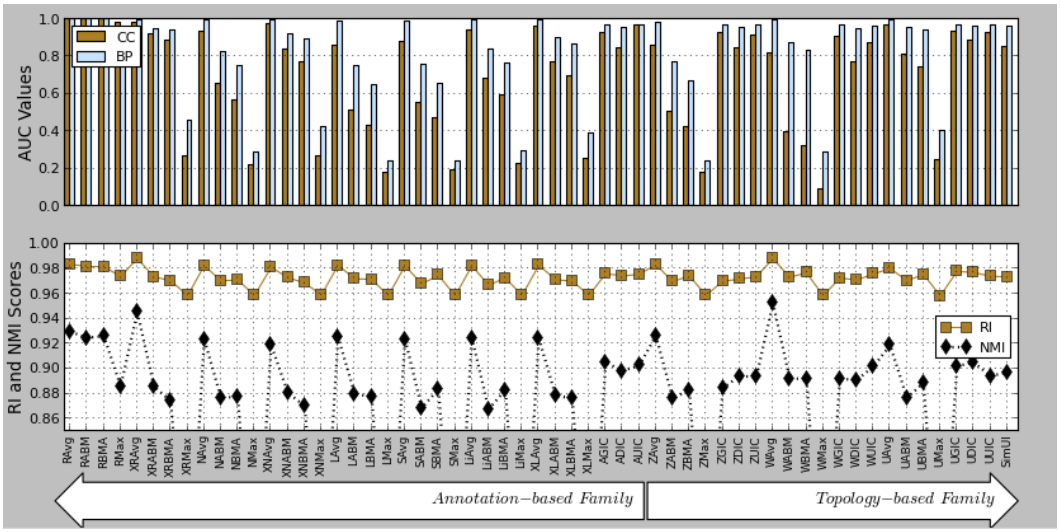

Figure 3.2: Result for Area under ROC curves (AUCs) and Network Clustering for the human PPI dataset on the BP and CC Ontologies.

For Biological (BP) and Cellular Component (CC) Figure 3.2, shows that Avg is the best method that can be used to compute the functional similarity scores since it produces the highest value for the Area under ROC Curves (AUCs), Rand Index (RI) and Normalized Mutual Information (NMI) and for any approach or measure.

3.3 Summary of CESSM, PPI and gene expression data results

From the CESSM online tool results, we observed that BMA and SimGIC are the best for all the ontologies (BP, MF, and CC).

From the Area Under Consideration (AUC) and Network Clustering (Rand Index and Normalized Mutual Information) showing Protein-Protein Interaction (PPI), It has been demonstrated that Average (Avg) is the best method to measure protein functional similarity since it has the highest score for AUC, RI and NMI. The classification of performance of measures was tested using the Area Under ROC curve analysis (AUCs) using the R programming language, which measures the Area between positive and negative rates. Also, RI and NMI were determined using python programming for the clustering of a protein network in order to obtain Protein-Protein interaction.

3.4 Summary

In this chapter, we have assessed the performance of different measures proposed for GO Annotation and Topology-based approaches. This made possible using the CESSM online tool in terms of Sequence, Pfam, and EC similarities. AUCs, RI and NI were used to determine the Human protein-protein interaction network. Performance results obtained from the CESSM for the different methods, BMA and SimGIC are the best methods to compute functional similarity between proteins. From AUCs, RI, and NI, Avg performs better than the other measures for protein-protein Interaction.

4. Extending Existing GO IC-based Online Tools

This chapter covers the extended IT-GOM-tool, a tool that incorporates all known GO information content-based semantic similarity measures, including topology- and annotation-based approaches for effective exploration of different semantic similarity measures that have been suggested in the context of GO. The tool was implemented by (Mazandu and Mulder, 2013a) and its initial version did not include all known functional similarity measures. In the previous chapter, we have shown that the performance of each measure depends on the biological application or data and a given approach does not always yield the best performance when used with the measure that has been designed for it. Thus, in this project we extend the tool by integrating all known information content (IC) based semantic similarity measures, including topology- and annotation-based approaches, enabling effective exploration of different protein functional similarity measures and providing researchers with the freedom to choose the most relevant measure for their specific applications.

4.1 Description of IT-GOM Tool

IT-GOM consists of a web front-end accessible at [1] , a web service interface and a repository for calculating GO term and protein semantic similarity measures at the back-end. The web front is displayed in Figure 4.1 (A) available in the supplementary file and different semantic similarity measures implemented by IT-GOM are shown in (B), and described in chapter 2.

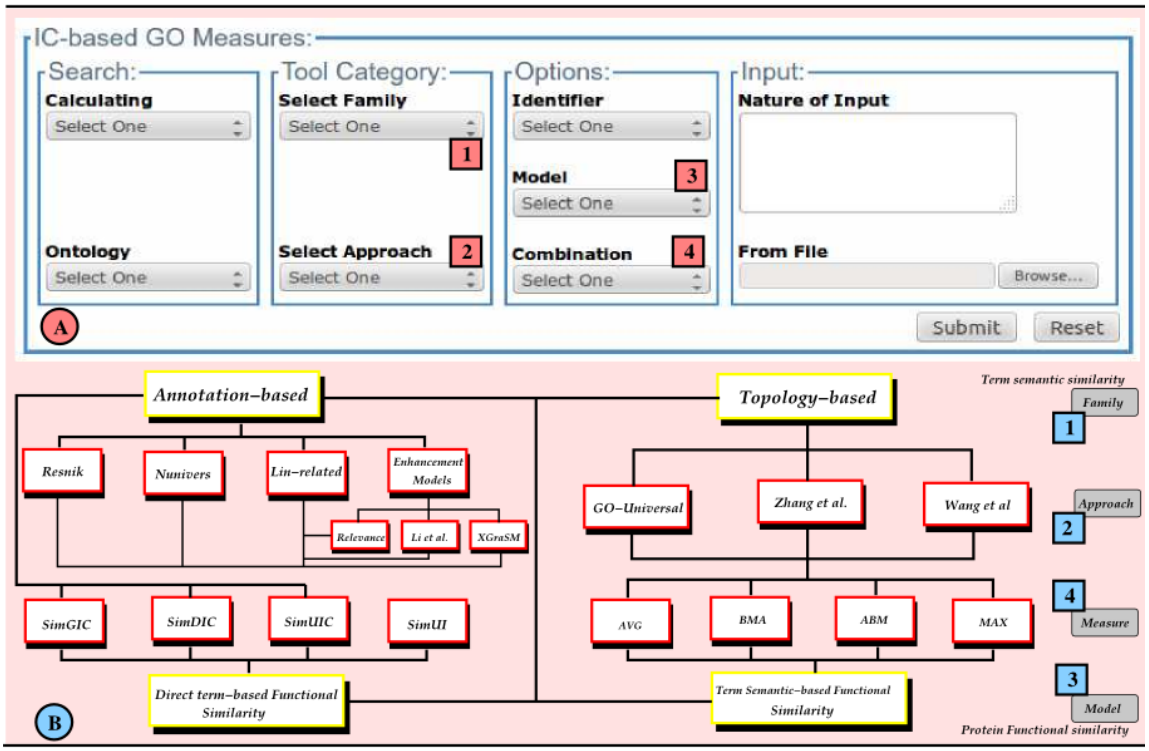

Figure 4.1: New IT-GOM web front. The interface (A) offers different input possibilities in terms of semantic similarity measures shown in section (B).

[1] http://web.cbio.uct.ac.za/ITGOM/tools/itgom.php

The back-end is composed of a set of query processing programs using Python to implement 57 different functional similarity measures. Each of the eight annotation-based and three topology-based approaches, namely Resnik, XGraSM-Resnik, Nunivers, XGraSM-Nunivers, Lin, XGraSM-Lin, Relevance (SimRel), Li et al. (SimIC), Wang et al., Zhang et al. and GO-universal, is implemented with four known IC-based non-direct functional similarity measures (Avg, Max, BMA and ABM) (Mazandu and Mulder, 2013a). Note that the Jiang and Conrath approach is not implemented explicitly since it has been shown to be a particular case of the Lin approach (Mazandu and Mulder, 2013b). IT-GOM also includes the three IC-based direct term functional similarity measures: SimGIC, SimDIC and SimUIC for the annotation-based family and each of the three topology-based approaches, and SimUI which is a particular case of SimGIC with all term IC values set to 1 (Mazandu and Mulder, 2012).

The IT-GOM interface allows easy and comprehensive navigation, query and exploration of GO term and protein semantic similarity measures. This 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. Comprehensive summary reports are generated and made available in table format and can be downloaded as a tab-delimited text file or printed. More details on the use of the tool are provided in the help page on the website.

4.2 Implementation of New Measures

The table 4.1 shows us all the methods which are now available in the IT-GOM tool. The letter "x" indicates the relevant approaches implemented in the existing one with the corresponding functional similarity. Whereas, the letter "y" indicates those implemented in our project for the extended IT-GOM.

		Functional similarity measures							
		Directed 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 l.					x	x	x	x
	Relevance					x	x	x	x
Topology-based					x				
	Zhang et al.	y	y	y		y	x	y	y
	Wang et al.	y	y	y		y	x	y	y
	GO-Universal	y	y	y		x	y	y	y

Table 4.1: Different GO term semantic similarity approaches and functional similarity measures implemented in the existing IT-GOM and those implemented in our project.

4.3 Summary

This chapter, we have explained the existing IT-GOM tool. The extended contains all the topology-based approaches with all its specific methods that we have implemented in this project. This ensures that GO semantic similarity data are conveniently accessible to users and can effectively be used to investigate functional similarity between proteins based on their GO annotations.

5. Conclusion

The aim of this essay has been to assess the performance of functional similarity measures and extend the IT-GOM online tool in order to enable users to choose the appropriate measure for their biological applications. We have used sequence, protein domain (pfam) and Enzyme commission (EC) similarity data downloaded from CESSM to assess these different measures using Pearson correlation score. In addition, we have used human protein-protein interaction and gene expression data downloaded from several different protein-protein interaction databases, to assess the performance of these measures in terms of their power to capture functional coherence in a human protein-protein interaction (PPI) network. These measures have been tested using Receiver Operator Characteristic (ROC) curve analysis, which assesses the Area Under the Curve (AUC), plotting the true positive rate or sensitivity vs the false positive rate. For the gene expression, we have assessed the clustering power of a given functional similarity measure by comparing the clustering result obtained by weighting protein relationships in the network by functional similarity scores to the ground truth using Normalized Mutual Information (NMI) and Rank Index (RI). Results show that BMA is the best measure for sequence, Pfam and EC similarity data under BP and MF ontologies, while ABM and SimGIC are better under CC ontology. For human protein-protein interaction and gene expression data, Avg has been shown to outperform all other measures. Finally, we have extended the existing web tool IT-GOM, by including all these measures that were not implemented in order to ensure that protein functional similarity data are conveniently accessible and can effectively be used in protein analyses at the functional level.

Acknowledgements

I would like to thank God, the source of intelligence and wisdom, for his protection and love in my life,

I would like to express my deep gratitude to Doctor Gaston Mazandu, my research supervisor, for his patient guidance, availability, help, enthusiastic encouragement and useful critiques of this research work,

my family particularly my mother Pauline Mbumba, brothers and sister Esperant, Dieudonne, Delord, Lazare and Aisance for all their support and love,

all staff of AIMS particularly Prof. Neil Turok, Prof Barry Green, Prof Jeff Sanders, A'eeda Mpofu and Jan Groenewald for this AIMS initiative, for making possible my presence at AIMS, their kindness and encouragement,

all my AIMS lecturers especially Frances for their kindness and for teaching me,

all the tutors of AIMS especially, Martha for their assistance and encouragement,

all my lecturers in DRC especially Prof Kafunda P., Prof Manya L. and Prof Mbuyi E. for their advice and for teaching me,

all my classmates at AIMS and in DRC for their assistance and encouragement.

Finally, I wish to express my appreciation to all my friends, brothers and sisters not included above who have assisted, encouraged me from applying to AIMS till completing the AIMS program. This goes most specially and importantly to Maurice Felo, Loris Matanda, Cathy Lapika, Audrey Moswa, HK Love, Rachel Bulungu, Carine Longonda, Sylvie Mpwekela, Rachel Dunia, Junior david Panzi, Freddy Kalambayi, Daddy Kanika, Jack Liuta, Blaise Kanyamanda, and all the members of Campus Crusade for Christ and DEV.

References

R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, et al. UniProt: the universal protein knowledgebase. *Nucleic acids research*, 32:D115–D119, 2004.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25:25–29, 2000.

D. Barrell, E. Dimmer, R. P. Huntley, D. Binns, C. O'Donovan, and R. Apweiler. The GOA database in 2009—an integrated Gene Ontology Annotation resource. *Nucleic acids research*, 37:D396–D403, 2009.

V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvreet. Fast unfolding of communities in large networks. *J. Stat. Mech*, 10008:1–12, 2008.

A. Bossi and B. Lehner. Tissue specificity and the human protein interaction network. *Molecular Systems Biology*, 5:260, 2009.

F. M. Couto, M. J. Silva, and P. M. Coutinho. Implementation of a functional semantic similarity measure between gene-products. 2003. URL https://docs.di.fc.ul.pt/jspui/bitstream/10455/2935/1/03-29.pdf.

F. M. Couto, M. J. Silva, and P. M. Coutinho. Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. pages 343–344, 2005. URL http://pdf.aminer.org/000/095/325/semantic_similarity_over_the_gene_ontology_family_correlation_and_selecting.pdf.

S. Jain and G. D. Bader. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC bioinformatics*, 11:562, 2010.

B. Li, F. Luo, J. Wang, F. A. Feltus, and J. Zhou. Effectively integrating Information Content and structural relationship to improve the GO-based Similarity Measure between Proteins. *International conference on Bioinfomatics computational biology , BIOCOMP*, pages 166–172, 2010.

P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19: 1275–1283, 2003.

G. K. Mazandu and N. J. Mulder. Scoring protein relationships in functional interaction networks predicted from sequence data. *PloS one*, 6:e18607, 2011.

G. K. Mazandu and N. J. Mulder. A topology-based metric for measuring term similarity in the gene ontology. *Advances in bioinformatics*, 2012:17, 2012.

G. K. Mazandu and N. J. Mulder. DaGO-Fun: tool for Gene Ontology-based functional analysis using term information content measures. *BMC bioinformatics*, 14:284, 2013a.

G. K. Mazandu and N. J. Mulder. Information content-based Gene Ontology semantic similarity approaches: toward a unified framework theory. *BioMed research international*, 2013:11, 2013b.

G. K. Mazandu, K. Opap, and N. J. Mulder. Contribution of microarray data to the advancement of knowledge on the *Mycobacterium tuberculosis* interactome: Use of the random partial least squares approach. *Infection, Genetics and Evolution*, 11(4):725–733, 2011.

C. Pesquita, D. Faria, H. Bastos, A. Falcão, and F. Couto. Evaluating GO-based semantic similarity measures. pages 37–40, 2007.

C. Pesquita, D. Faria, H. Bastos, A. E. Ferreira, A. O. Falcão, and F. M. Couto. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC bioinformatics*, 9:S4, 2008.

C. Pesquita, D. Pessoa, D. Faria, and F. Couto. Cessm: collaborative evaluation of semantic similarity measures. *JB2009: Challenges in Bioinformatics*, 157:1–5, 2009.

A. Schlicker, F. S. Domingues, J. Rahnenführer, and T. Lengauer. A new measure for functional similarity of gene products based on Gene Ontology. *BMC bioinformatics*, 7:302, 2006.

M. Seuneu and N. Mulder. Assessing Functional Similarity Measures for Protein Analysis at Functional Level, 2010. URL http://archive.aims.ac.za/postgraduate-diploma-essays/2009-10/milaine.pdf. Essay towards Postgraduate Diploma, African Institute for Mathematical Sciences (South Africa).

K. Steinhaeuser and N. V. Chawla. Identifying and evaluating community structure in complex networks. *Pattern Recognition Letters*, 31:413–421, 2010.

A. Tversky. Features of similarity. *Psychological review*, 84:327, 1977.

J. Z. Wang, Z. Du, R. Payattakool, S. Y. Philip, and C.-F. Chen. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23:1274–1281, 2007.

P. Zhang, J. Zhang, H. Sheng, J. J. Russo, B. Osborne, and K. Buetow. Gene functional similarity search tool (GFSST). *BMC bioinformatics*, 7:135, 2006.