An adaptive semantic similarity measure over the Gene Ontology

Pietro Galliani

Advisor: Novi Quadrianto

August 30, 2015

Abstract

I introduce a simple, adaptive semantic similarity measure between proteins annotated with Gene Ontology terms. I show that, with respect to the problem of distinguishing between interacting and noninteracting protein pairs and over four datasets of four different species (baker's yeast, humans, E. coli and house mice), this measure outperforms common non-adaptive measures of semantic similarity. Then I investigate the connection between the efficacy of "cross-training" this measure and the philogenetic relationships between the corresponding organisms.

1 Introduction

1.1 The Gene Ontology

The Gene Ontology (GO) consists of a vast (43594 terms as of August 2015) *controlled vocabulary* of species-neutral attributes to be used in the description of the properties of genes and gene products [1]. These attributes are connected by various binary *relationships*, thus making the ontology into a *directed acyclic graph*. Roughly speaking, the Gene Ontology can be subdivided into three components, or *domains*:

1. **Biological Process** — multi-step biological events occurring in an organism, with a definite beginning and end. Examples: GO:0048568 (embryonic organ development), GO:0043500 (muscle adaptation), GO:0046034 (ATP metabolic process);

2. **Molecular Function** — elemental, molecular-level events, such as GO:0003824 (catalytic activity), GO:0043178 (alcohol binding) or GO:0019239 (deaminase activity);

3. **Cellular Component** — parts of a cell, such as GO:0005840 (ribosome), GO:0005886 (plasma membrane) or GO:0030133 (transport vesicle).

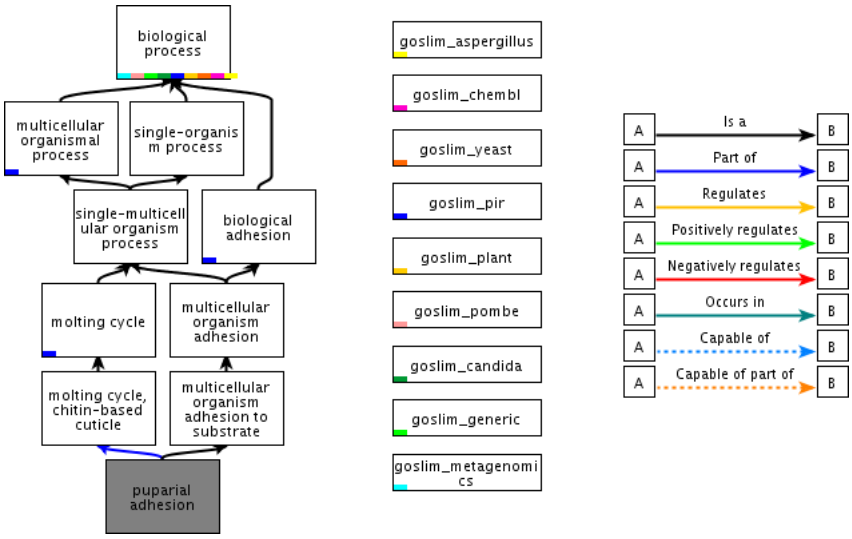


Figure 1: **Left:** A part of the Gene Ontology - all the ancestors of the GO term GO:0007594 (puparial adhesion). **Center:** slim GO sets (collections of high-level GO terms). **Right:** the relations occurring in the Gene Ontology. **Our problem:** two proteins or genes, each one annotated with a set of GO terms. Do they interact? **Source:** QuickGO browser (`http://www.ebi.ac.uk/QuickGO/`).

Most major species-specific databases and research initiatives contribute to the Gene Ontology and rely on it for the *annotation* of genes and gene products, thus making the Gene Ontology itself into a very valuable tool for the sharing of information across different biological scientific communities and for the automated analysis and comparison of genes or gene products.

In this work, I will investigate the topic of *semantic similarity measures* over the Gene Ontology. For instance, let $\mathcal{A}_1$ and $\mathcal{A}_2$ be the the sets of Gene Ontology terms associated with the two proteins P_1 and P_2: what can we then say about the *degree of similarity* between P_1 and P_2?

The obvious reply to this, of course, is that the question is much too vague to have a meaningful answer: there are many different senses in which two proteins may be said to be similar or dissimilar — do they have similar ammino-acid sequences? Do they have comparable 3D shapes, or do they perform conceptually analogous functions inside of the organism?

Any choice (and many more besides) can be appropriate for some purposes, and entirely unreasonable for others: for example, if we wish to estimate the likelihood that the reduced or defective activity of certain enzyme (let us say, as a consequence of a given genetic disease) may affect the quantity of some other protein present in the organism then we are not as interested in the shapes of the two proteins as in whether – for instance – they are involved in different steps of the same overall metabolic process. On the other hand, if we want to understand whether two genes may be derived from a common ancestor then we will not particularly care about that, and we will be interested instead in estimating the similarity between their nucleotide sequences.

This leads directly to a first (perhaps trivial, but worth spelling out anyway) observation:

- *There is no such thing as an "universal similarity measure" between sets of GO annotations: different purposes ask for different measures.*

At this point, it would be only natural to feel a strong temptation to go looking for a very general sort of measure, one which — through the tuning of some, hopefully few, parameters — could be adapted to any and all purpose. But given the sheer variety of different specific purposes that may lurk behind the generic notion of "estimating the degree of similarity between two genes or gene products", such an enterprise would risk achieving generality only at the expense of simplicity and conceptual coherence. There is certainly value in the theoretical study of similarity measures between groups of terms of ontologies, and there is also value in the comparative study of different similarity measures between sets of GO annotations[1]; but designing a specific measure with "generality" as one of the objectives would run the very real risk of increasing complexity for no real practical gain. Therefore, in this work I will hold to the following principle:

- *When designing a similarity measure between genes/gene products,* ***one*** *intended purpose is more than enough.*

[1] For an interesting work along these lines, see for example [9].

But which purpose should we select for our soon-to-be similarity measure? In this work, for reasons of practicality and simplicity, we will focus on the problem of guessing whether two proteins are going to *interact* which each other. The *Database of Interacting Protein* [14, 13] is a readily available source of curated data that may be used to train and evaluate the similarity; and the question of whether the activity of a protein is likely to influence the activity of another is of clear applied interest. Furthermore, the fact that this is a simple binary question allows us to give an exceedingly simple "high-level" definition of our intended similarity measure *logSim*:

- Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be two sets of GO annotations. Then

  $$\mathit{logSim}(\mathcal{A}_1, \mathcal{A}_2) = \text{Prob}(P_1 \text{ and } P_2 \text{ interact } | \mathcal{A}_1 \text{ annotates } P_1, \mathcal{A}_2 \text{ annotates } P_2).$$

In this way, we turned our initial, and very generic, aim of designing a "good" similarity measure between sets of GO annotations into the concrete problem of estimating the probability that two proteins may interact with each other, or, in other words, into an instance of *classification-based similarity learning*, a subject with close ties with the general (and highly studied) problem of (supervised) *distance metric learning* [12, 16, 15].

But wait, we forgot to answer a fundamental question — interact *where*? After all, proteins do not exist in static and blessed isolation, but take part to all sorts of activities and transformations inside of an organism; and therefore, the identity of the organism itself has no small effect on the probability that two proteins may interact, given their GO annotations. Let us therefore revise slightly our definition:

- Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be two sets of GO annotations, and let $\mathbf{s}$ be a species. Then

  $$\mathit{logSim}_{\mathbf{s}}(\mathcal{A}_1, \mathcal{A}_2) = \text{Prob}(P_1 \text{ and } P_2 \text{ interact in } \mathbf{s} | \mathcal{A}_1 \text{ annotates } P_1, \mathcal{A}_2 \text{ annotates } P_2).$$

Clearly, we are bravely running away from generality. And we are not quite done running yet! Indeed, there is yet another observation worth making: not all Gene Ontology terms are created equal — at least, not insofar as similarity *with respect to them* is indicative of the existence of interactions between two proteins. If we had an arbitrarily large amount of data to work with, we might attempt to estimate the *degree of relevance* of every single GO term to the problem of inferring the likelihood that two proteins interact inside of a given organism; but this is clearly unfeasible, given the sheer size of the Gene Ontology and the dimensions of the datasets we will work with (the biggest one of them, corresponding to *S. cerevisiae* — that is, the common baker's yeast — will consist of 5145 pairs of reliably interacting proteins). What to do, then? Well, one possibility that comes to mind is to choose a smaller set of "high-level", generic GO terms and estimate the relative importances of the sets consisting of them *and all their subterms* to the problem of learning if two proteins are interacting.

As for the choice of the terms, the GO Consortium itself offers *slim ontologies*, subsets of the GO created for the purpose of permitting a broader classification of genes and gene products than the one that the whole GO would offer. Some of these slim ontologies are custom-made for particular organisms or classes of organisms, or for specific purposes such as metagenomics; but here at last we will make a stand in favor of generality, and choose to use the *generic slim* — a set $\mathrm{GO}_{\mathrm{slim}}$ of 149 high-level, non species-specific GO terms chosen by the GO Consortium and intended to be "suitable for most purposes".[2] Using the same slim ontology for all species which we will consider will permit us to evaluate the efficacy of *cross-training* by trying to estimate whether two proteins interact in a species while using a model (and a weighing the terms of the ontology) developed for a different species. The degree up to which this will degrade the performance of our measure will then offer an estimate of the degree up to which species identity affects the relationship between GO annotations and likelihood of protein interaction.

1.2 Semantic Similarity Measures for GO

The literature on semantic similarity over the Gene Ontology is vast, and it is far beyond the scope of this work to present an exhaustive summary of all the proposed measures.[3] In this section, therefore, I will limit myself to describe a few of the most used and historically significant ones, which I will then employ as a basis for comparison with my proposed approach.

One element which is common to all these measures is the (information theory-derived) notion of the *Information Content* of a GO term. In brief, given a term t, one defines

$$IC(t) = -\log\left(\frac{\mathrm{freq}(t)}{\max(\{\mathrm{freq}(t') : t' \text{ GO term}\})}\right)$$

where where $\mathrm{freq}(t)$ is the *frequency* of the term t in our annotation data, that is, the number of times in which it *or any of its subterms* occurs in our set of annotations. Intuitively speaking, a term's information content is higher the less often it and its successors occur in the data: in particular, if the term or it successors never occur then its information content is undefined, while if every occurring term is a subterm of the given term (e.g., the given term is at the root of our ontology) then its information content is zero.

It is common practice to relativize the above expression to the three domains of the Gene Ontology (that is, Biological Process, Molecular Function and Cellular Component), thus obtaining

$$IC(t) = -\log\left(\frac{\mathrm{freq}(t)}{\max(\{\mathrm{freq}(t') : t' \text{ GO term over the same domain as } t\})}\right)$$

and to compute the similarities described below with respect to only one of these domains: indeed, these domains differ both in size (e.g., the Biological Process

[2] `http://geneontology.org/page/go-slim-and-subset-guide`, retrieved 24 August 2015.
[3] For a survey on this topic, see [10].

domain contains more than 27000 terms, while the Molecular Function contains less than 4000) and relative importance (e.g., under many informal notions of "similarity" the fact that two proteins belong to the same cellular subcompartment does not really say that much about their relative degree of similarity). Furthermore, it is also commonplace to normalize the information content so that it takes values between 0 and 1. Thus, the definition of information content which we will use in this work is

$$IC(t) = -\frac{\log\left(\frac{\mathrm{freq}(t)}{\max(\{\mathrm{freq}(t'):t' \text{ GO term over the same domain as } t\})}\right)}{\log(|\{t' : t' \text{ GO term over the same domain as } t\}|)}. \quad (1)$$

We can now briefly recall and comment the definitions of the semantic similarity measures which we will discuss in this work:

- *Resnik's measure* [11] is a simple notion of semantic similarity which has proven itself surprisingly reliable and has often been used in practice [10]. In its simplest form, Resnik's measure computes the degree of similarity of two *terms* t_1 and t_2, rather that two set of terms $\mathcal{A}_1$ and $\mathcal{A}_2$, and defines it as

  $$sim_R(t_1, t_2) = IC(LCA(t_1, t_2)) :$$

  the degree of similarity between t_1 and t_2 is given by the information content of the *lowest common ancestor* of t_1 and t_2 in our ontology. This definition, however, is not directly applicable to GO terms: indeed, the Gene Ontology is not a tree but merely a directed acyclic graph, and therefore there does not necessarily exist a *unique* lowest common ancestor between two terms t_1 and t_2. A simple fix consists in choosing instead the *most informative common ancestor* $MICA(t_1, t_2)$, that is, the common ancestor t of t_1 and t_2 which maximizes $IC(t)$. This is the approach which we will consider in this work: therefore, we will have that

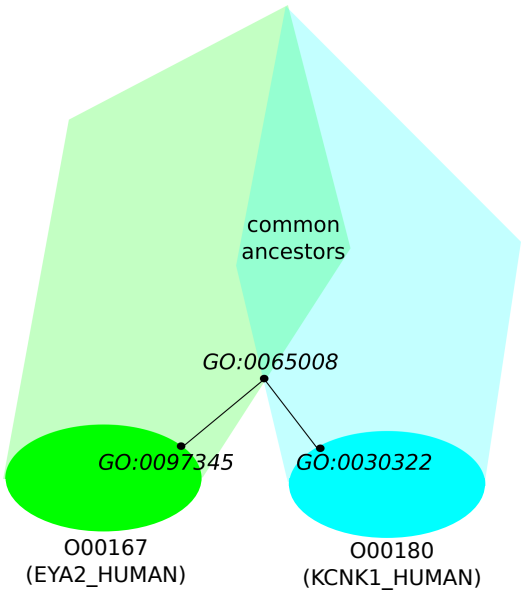
  $$sim_R(t_1, t_2) = IC(MICA(t_1, t_2)). \quad (2)$$

  Other, more sophisticated approaches [2, 3] average out the information content of various common ancestors of t_1 and t_2, chosen so that they are independent (for certain given notions of independence); but I will not pursue that line of investigation further in this work.

  Another difficulty — common also to Lin's and Jiang/Conrath's measures, which we will later discuss — is that we need to find a way to "lift" our similarity measure from pairs of terms t_1, t_2 to pairs of *sets* of terms $\mathcal{A}_1$, $\mathcal{A}_2$. The three most common approaches to this are the following:

  Average: $sim_R^{\mathrm{avg}}(\mathcal{A}_1, \mathcal{A}_2) = \frac{1}{|\mathcal{A}_1|\cdot|\mathcal{A}_2|} \sum_{t_1 \in \mathcal{A}_1} \sum_{t_2 \in \mathcal{A}_2} sim_R(t_1, t_2)$;

  Maximum: $sim_R^{\max}(\mathcal{A}_1, \mathcal{A}_2) = \max(\{sim_R(t_1, t_2) : t_1 \in \mathcal{A}_1, t_2 \in \mathcal{A}_2\})$;



ID	Name	IC
GO:0097345	mitochondrial outer membrane permeabilization	0.6673
GO:0030322	stabilization of membrane potential	0.7642
GO:0065008	regulation of biological quality	0.2517

$sim_R(\text{O00167}, \text{O00180}) = IC(\text{GO:0065008}) = 0.2517$
$sim_L(\text{O00167}, \text{O00180}) = 2 \cdot IC(\text{GO:0065008})/(IC(\text{GO:0097345}) + IC(\text{GO:0097345})) = 0.3517$
$sim_{JC}(\text{O00167}, \text{O00180}) = 1 + IC(\text{GO:0065008}) - (IC(\text{GO:0097345}) + IC(\text{GO:0097345}))/2 = 0.5360$

Figure 2: Resnik, Lin, and Jiang-Conrath similarities between proteins O00167 (Eyes absent homolog 2) and O00180 (Potassium channel subfamily K member 1). GO:0065008 is the most informative GO term whose descendants annotate both proteins: in particular, GO:0097345 annotates protein O00167 and GO:0030322 annotates protein O00180.

Best Matching Average:

$$sim_R^{\mathrm{bma}}(\mathcal{A}_1, \mathcal{A}_2) = \frac{1}{2|\mathcal{A}_1|} \sum_{t_1 \in \mathcal{A}_1} \max(\{sim_R(t_1, t_2) : t_2 \in \mathcal{A}_2\}) + \frac{1}{2|\mathcal{A}_2|} \sum_{t_2 \in \mathcal{A}_2} \max(\{sim_R(t_1, t_2) : t_1 \in \mathcal{A}_1\}).$$

Any of these approaches has advantages or disadvantages depending on the intended purpose of the similarity measure. When it comes to detecting interactions between proteins, however, the Maximum approach tends to outperform the others: this is fairly reasonable, since it can be enough for two proteins to have much in common under *one single aspect* of their behaviour for them to interact with each other.

- One fairly counterintuitive aspect of Resnik's measure is that the similarity between two terms is only a function of their LCA/MICA, and not of their distance from it. In particular, we always have that

$$sim_R(t_1, t_2) = sim_R(MICA(t_1, t_2), MICA(t_1, t_2))$$

from which it follows almost at once that $sim_R(t, t) = IC(t)$ can be smaller than one. Lin's measure [7] remedies to this by normalizing Resnik's measure as follows:

$$sim_L(t_1, t_2) = \frac{2 \cdot sim_R(t_1, t_2)}{IC(t_1) + IC(t_2)}.$$

From this definition, it follows at once that $sim_L(t, t) = 1$ for all GO terms t. This, however, can cause Lin's measure to overestimate the degree of similarity between shallowly annotated terms: indeed, if two proteins share even a very shallow annotation — for instance, GO:0000003 (reproduction) or GO:0032502 (developmental process) — then their maximum Lin similarity will be one.

In order to lift Lin's measure to sets of annotations, we choose t_1 and t_2 as the two terms which maximize $sim_R(t_1, t_2)$, and not $sim_L(t_1, t_2)$. The reason for this is simple: since some very general GO terms — such as the ones just mentioned — are shared by a great number of proteins, $\max\{sim_L(t_1, t_2) : t_1 \in \mathcal{A}_1, t_2 \in \mathcal{A}_2\}$ would be nearly always one. This is not the case if we instead define

$$sim_L(\mathcal{A}_1, \mathcal{A}_2) = \frac{2 \cdot sim_R(t_1, t_2)}{IC(t_1) + IC(t_2)} \tag{3}$$

for $(t_1, t_2) \in \operatorname{argmax}\{sim_R(t_1, t_2) : t_1 \in \mathcal{A}_1, t_2 \in \mathcal{A}_2\}$. Thus, in order to compute the Lin similarity between two sets of annotations we first find the two annotations $t_1 \in \mathcal{A}_1$ and $t_2 \in \mathcal{A}_2$ which share the most informative MICA — in other words, we focus on the aspect of the gene ontology

with respect to which the two annotations share the most information — and then we divide the information content of this MICA for the average information content of t_1 and t_2. Informally speaking, if under the point of view with respect to which $\mathcal{A}_1$ and $\mathcal{A}_2$ share the most information we have that $\mathcal{A}_1$ and $\mathcal{A}_2$ are in entire agreement then the Lin distance between $\mathcal{A}_1$ and $\mathcal{A}_2$ will be precisely one; but if under that aspect $\mathcal{A}_1$ and/or $\mathcal{A}_2$ would carry finer-grained information than the one carried by the most informative term attributable to both $\mathcal{A}_1$ and $\mathcal{A}_2$ then the Lin similarity will be smaller than one.

- Jiang-Conrath's measure [5] is, under some points of view, conceptually similar to Lin's measure in that it attempts to estimate the *information content loss* between the most common ancestor of t_1 and t_2 and t_1 and t_2 themselves. In its original formulation, Jiang-Conrath's measure is a *distance*, not a *similarity*, and it is defined as

$$\begin{aligned} \mathrm{d}_{JC}(t_1, t_2) &= IC(t_1) + IC(t_2) - 2 \cdot IC(MICA(t_1, t_2)) \\ &= IC(t_1) + IC(t_2) - 2sim_R(t_1, t_2) : \end{aligned}$$

  if the two terms t_1 and t_2 are identical, their distance is zero, whereas if they are not identical it is given by $IC(t_1) - IC(MICA(t_1, t_2))$ (that is, the information loss between t_1 and the most informative common ancestor of t_1 and t_2) plus $IC(t_2) - IC(MICA(t_1, t_2))$ (that is, the information loss between t_2 and the most informative common ancestor of t_1 and t_2).

  In order to turn this distance into a similarity measure, we follow Jiang and Conrath's suggestion in [5] and define

$$sim_{JC}(t_1, t_2) = 1 + sim_R(t_1, t_2) - \frac{IC(t_1) + IC(t_2)}{2};$$

  and then we lift this expression much in the same way in which we lifted Lin's similarity measure, thus obtaining

$$sim_{JC}(\mathcal{A}_1, \mathcal{A}_2) = 1 + sim_R(\mathcal{A}_1, \mathcal{A}_2) - \frac{IC(t_1) + IC(t_2)}{2} \tag{4}$$

  where $t_1 \in \mathcal{A}_1, t_2 \in \mathcal{A}_2$ are picked in order to maximize $IC(MICA(t_1, t_2))$.

- The simGIC measure [8] is quite different from the above-mentioned measures in that it does not rely on the notion of "most informative common ancestor". Instead. it is defined as

$$\mathrm{simGIC}(\mathcal{A}_1, \mathcal{A}_2) = \frac{\sum\{IC(t) : t \in \mathcal{A}_1^{\uparrow} \cap \mathcal{A}_2^{\uparrow}\}}{\sum\{IC(t) : t \in \mathcal{A}_1^{\uparrow} \cup \mathcal{A}_2^{\uparrow}\}}$$

  where $A^{\uparrow} = A \cup \{t : \exists t' \in \mathcal{A} \text{ s.t. } t \text{ is an ancestor of } t'\}$. Informally speaking, this measure computes the total information content of the most detailed *set of attributes* compatible with both $\mathcal{A}_1$ and $\mathcal{A}_2$, and divides it by

the total information content of the least detailed set of attributes which entails both $\mathcal{A}_1$ and $\mathcal{A}_2$. Despite its apparent simplicity, this measure proved itself to be in close relation with measures of *sequence similarity* between genes or gene products [9].

2 LogSim: an Adaptive, Logistic Regression-Based Similarity Measure

2.1 The Similarity

In this section, I will describe in some detail a novel, adaptive similarity measure based on logistic regression. The main ideas behind it, as briefly mentioned in the introduction, will be the following:

1. Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be two sets of GO terms, and let $\mathbf{s}$ be a species. Then $logSim(\mathcal{A}_1, \mathcal{A}_2 \mid \mathbf{s})$ represents an estimation of the probability that two proteins $\mathbf{P}_1$ and $\mathbf{P}_2$, annotated with GO terms $\mathcal{A}_1$ and $\mathcal{A}_2$ respectively, interact in a specimen of $\mathbf{s}$.[4]

2. Let $\mathrm{GO}_{\mathrm{slim}}$ be the set of all 149 terms of the *generic slim gene ontology* discussed in the introduction. Then our similarity measure will operate as follows:

   (a) First, it will compute a *non-adaptive* measure of similarity with respect to *all* sub-ontologies generated by the sub-terms of all terms in the generic slim gene ontology;

   (b) Then it will combine these similarities into a single similarity measure, which will constitute an estimate of the probability mentioned in the previous point.

3. In order to learn how to weigh the respective contributions of the various sub-ontologies, we will need to *train* our similarity on a series of interacting and non-interacting pairs of annotated proteins.

Three aspects of the above description which are yet unspecified are the following:

1. How are we going to compute the similarity of two annotated proteins with respect to a sub-ontology?

2. How are we going to combine the "local" similarity measures computed with respect to sub-ontologies into a global similarity measure?

3. How are we going to learn the weights for the similarity combination operation mentioned in the previous point?

[4] Whenever the identity of $\mathbf{s}$ is clear we will tacitly omit it from expressions.

As for the first point, a straightforward choice is to adapt the definition of (Max) Resnik Similarity: indeed, as already mentioned, this measure appears to perform fairly well in a variety of circumstances, and its intuitive interpretation — a measurement of the maximum amount of information *shared* between attributes assigned to the two proteins — fits very well our intended purpose of detecting interactions between proteins. The fact that, differently from the other measures considered, $sim_R(t,t)$ and $sim_R(\mathcal{A},\mathcal{A})$ are not necessarily one also fits our intended application: indeed, it is not always necessarily the case that a protein can interact with *itself*. The main drawback of Resnik's method for our purposes would be the fact that the *identity* of the shared information between the two sets of annotations also affects the probability of interaction — that is, a term may be very informative (that is, few terms are annotated by it) without indicating a high likelihood of interaction between proteins which share it. But this is precisely the kind of issue that our approach is designed to resolve: indeed, the terms of the slim GO ontology correspond to different conceptual categories of GO terms, and thus by attributing different weights to them we may be able to account to the different degrees up to which they affect the likelihood that two proteins may be interacting.

Thus, for every slim term $gs \in \mathrm{GO}_{\mathrm{slim}}$ and for any two sets of annotations $\mathcal{A}_1$ and $\mathcal{A}_2$, we will define the *gs-local* Resnik similarity between $\mathcal{A}_1$ and $\mathcal{A}_2$ as

$$sim_R(\mathcal{A}_1,\mathcal{A}_2|gs) = \max\{IC(t) : \exists t_1 \in \mathcal{A}_1, t_2 \in \mathcal{A}_2 \text{ s.t. } t \prec t_1, t \prec t_2, gs \prec t\} \tag{5}$$

where $t_1 \prec t_2$ stands for "t_1 is an ancestor of t_2".

As for the other two points, *logistic regression* is a natural choice. A simple, well-studied and widespread classification technique, logistic regression models the probability of an event $\mathbf{E}$ (in our case, "the two proteins interact") as follows:

$$\mathrm{Prob}(\mathbf{E}) = \frac{1}{1+e^{-(\vec{w}\cdot\vec{x}_{\mathbf{E}}+w_0)}}$$

where $\vec{x}_{\mathbf{E}}$ is a tuple of numerical features relevant to the event $\mathbf{E}$, $\vec{w}$ is a tuple of weights associated to each feature, and w_0 is an additional "bias" term. Every weight w_i (for $i > 0$) specifies the degree up to which the value of the i-th variable (that is, in our case, the value of $sim_R(\mathcal{A}_1,\mathcal{A}_2|gs_i)$) affects the likelihood that the two proteins interact. In particular, if $\vec{w}\cdot\vec{x}_{\mathbf{E}} = 0$ then $\mathrm{Prob}(\mathbf{E}) = 1/(1+e^{-w_0})$ is constant. If $w_0 = 0$, then whenever $\vec{w}\cdot\vec{x}_{\mathbf{E}} = 0$ then $\mathrm{Prob}(\mathbf{E}) = 0.5$: otherwise, the value of w_0 specifies the probability assigned to the event $\mathbf{E}$ = the two proteins interact in the absence of relevant information.

The values of the weights $\vec{w}$ and w_0 will be learned through *maximum likelihood estimation*, that is, we will attempt to maximize the overall probability of our training dataset (consisting of the partial similarity measures corresponding to interacting and non-interacting tuples). A minor point should be clarified here: it would seem at first sight reasonable to add the extra condition that the weights $\vec{w}$ must always be greater than 0. This, however, would be a mistake: indeed, the terms of the GO slim ontology are not independent — some are subterms of other, and a negative value of a subterm may be used to indicate

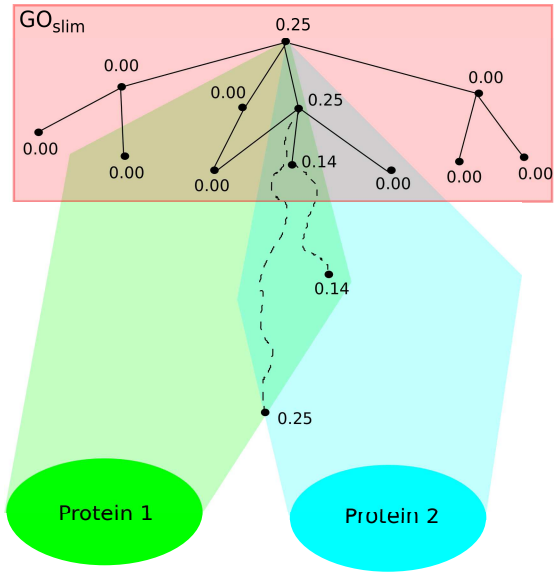


Figure 3: Local Resnik similarities computed with respect to the subcomponents of the GO ontology corresponding to $\mathrm{GO}_{\mathrm{slim}}$ terms.

that similarity with respect to it is not as indicative of interaction as similarity with respect to the rest of the superterm.

Finally, we should ask ourselves which metrics we should use to evaluate our similarity measure. In this work, we will focus on the following four metrics:

- *Precision*: the fraction of accepted protein pairs (that is, proteins to which our model assigns a probability of interaction greater than 0.5) which are indeed interacting;
- *Recall*: the fraction of interacting protein pairs which are accepted by our model;
- F_1 *score*: the harmonic mean of precision and recall;
- *Area under the Receiver Operating Characteristic (ROC) curve*: the probability that a random interacting pair of proteins will score a higher similarity value than a random non-interacting pair of proteins.

In order to evaluate our method, we will also need to compute the above metrics for the other similarity measures considered in this work. Computing the area under the ROC curve for these measures is straightforward; however, computing precision and recall for them will be more delicate — since these measures are not meant to express probabilities, it will not do to simply take as interacting all protein pairs whose similarity (according to these measures)

is greater than 0.5 and to reject all those which have similarity less than 0.5. Instead, I will first train logistic regression classifiers to try to infer whether two proteins do or do not interact according to the values of these individual similarity measures *alone*; then I will evaluate the precision and the recall of this classifier, which will provide a fair assessment of the ability of the underlying similarities to discriminate between interacting and non-interacting pairs of proteins.

In the next subsections, we will go in some more detail through the operations involved in gathering and preparing the data, training our classifier/similarity measure, and testing it.

2.2 Data

- The gene ontology itself (as well as the list of the generic slim GO terms) was downloaded from the website of the Gene Ontology Consortium.[5]

- The annotation data for the four species considered — Homo sapiens, Mus musculus (mice), Saccharomyces cerevisiae (baker's yeast) and Escherichia Coli — was also downloaded from the Gene Ontology site.

- The interaction data was downloaded from the site of the DIP project.[6]. For any species, the DIP Project offers both a *core* set, which contains only the most reliable of the suspected protein-protein interactions, and a *full* set, which also contains less reliable interactions. Both sets were downloaded: the core sets were used as sources of protein-protein interactions for the training and validation of the model, whereas — as we will discuss in the next section — the full sets were employed to generate examples of *non*-interacting proteins. It is perhaps also worth pointing out here that in the case of the house mice (M. musculus), the core and the full sets were identical. Finally, we removed from these sets all the pairs containing proteins for which we could not find any *GO* annotation.

- For M. musculus and S. cerevisiae, the annotations obtained from the Gene Ontology website identified the proteins through their species-specific IDs (that is, through their MGI and SGD codes respectively). However, in the interaction files obtained from the DIP the proteins are always reported by means of their UniProt identifiers. Therefore, cross-references files between uniprot and MGI/SGD were downloaded from the uniprot website[7] and used to translate the interaction data.

2.3 Preprocessing, Training, and Testing

In order to train and test our similarity measure, we needed to also generate a set of *non*-interacting proteins for each species. This was made in three steps:

[5] `http://geneontology.org/`
[6] `http://dip.doe-mbi.ucla.edu`
[7] `http://www.uniprot.org/help/?query=biocuration&fil=section%3Ahelp`

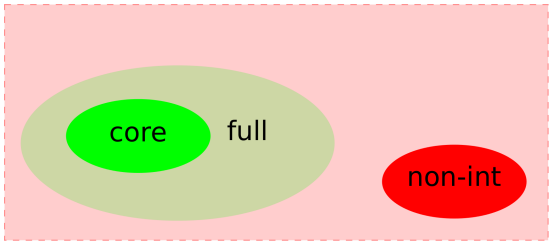


Figure 4: Relationship between core interaction datasets, full interaction datasets, and (generated) non-interaction datasets.

1. Create a list of all proteins which participate in *some* interaction (according to the more reliable, "core" set of interactions);

2. Generate random pairs of proteins from the above created list, and add it to the set of *non-interacting* proteins;

3. Remove from the above set any pair of proteins $(P1, P2)$ such that either it or its reverse $(P2, P1)$ is in the more general, "full" set of interactions.

Finally, for any pair of proteins $P1$ and $P2$, in the list of core interactions or in the list or non-interacting proteins, we collected the corresponding sets of GO annotations $\mathcal{A}_1$, $\mathcal{A}_2$ and we computed the local similarities $sim_R(\mathcal{A}_1, \mathcal{A}_2|gs)$ for any generic GO slim term gs. Thus, for every species considered and every interacting (core) and non-interacting pair of proteins for that species we stored 149 values.

Then, in order to extract training and testing datasets for a given species, we will simply split the set of (core) interacting pairs (or better, the corresponding vectors of local Resnik similarities) into two random, equally-sized subsets.[8] Then we will add to each subset an equal number of non-interacting pairs, and finally we will shuffle randomly both lists (while keeping track of which similarity vector originates from an interacting pair and which one originates from a non-interacting pair in a separate list of *labels*.

Then we will train our Logistic Regression classifier over the training set and labels, and will test its performance with respect to the testing set of vector and the testing labels.

In order to evaluate the behaviour of our similarity measure (and, for comparison, that of the other, non-adaptive ones), we repeated the procedure of splitting the datasets, training the classifier, testing it 1000 times per experiment and recording at each time the relevant statistics (that is, precision, recall, ROC area and F_1 score).

[8] If the number of interacting pairs is odd, we will ignore a random one of them.

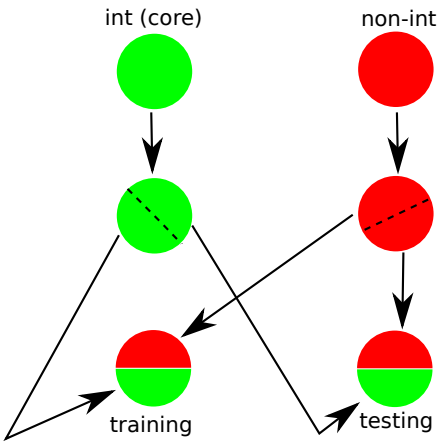


Figure 5: Generation of training and testing datasets: randomly split the (core) interaction dataset into two parts, then add to either an equal amount of number of pairs from the non-interaction dataset (while keeping the training and the testing sets disjoint).

3 Results

3.1 Saccharomyces cerevisiae

With 5145 verified ("core") interactions out of 22056, our interaction dataset for baker's yeast is the richest among those which we will consider in this work. Table 1 summarizes the results of our experiments: as we can see, our measure is the best one in terms of precision, F_1 score and area under the ROC curve, but it is outperformed by Resnik's, Lin's and Jiang-Conrath's (BIO) measures in terms of recall. However, these two measures are far surpassed by our measure in terms of precision and ROC area. All of these observations are strongly statistically significant: indeed, the t-test for dependent paired samples, applied to the scores of logSim and the best non-adaptive measure (that is, simGIC (CEL) in the case of precision, Lin (BIO) in the case, of recall and F_1 score, and Jiang (BIO) in the case of ROC AUCH), returns t-statiscs 64.49, -107.38, 209.53 and 624.35 respectively, and the corresponding p-values are all 0.0 (that is, below machine precision).

In Figure 7.A, on the left, we can the ROC curves for the logSim measure and for the non-adaptive measures computed over the biological process sub-ontology (that is, the one with respect to which they have the best scores). The width of the lines corresponds to one standard deviation, over and below the value. On the right, instead, we have a representation of the distribution of similarity scores for interacting and non-interacting pairs of proteins in the testing set after training the classifier over the corresponding training set; and as we can see, our classifier is capable of discriminating between interacting and

	precision	recall	F_1 score	ROC AUC
logSim	**0.8943 ± 0.0058**	0.7812 ± 0.0074	**0.8339 ± 0.0045**	**0.9021 ± 0.0032**
Resnik (BIO)	0.8196 ± 0.0077	0.7814 ± 0.0079	0.8000 ± 0.0042	0.7956 ± 0.0053
Lin (BIO)	0.8110 ± 0.0061	**0.8012 ± 0.0069**	0.8061 ± 0.0041	0.7938 ± 0.0053
Jiang (BIO)	0.7611 ± 0.0055	0.7960 ± 0.0069	0.7781 ± 0.0043	0.8198 ± 0.0046
simGIC (BIO)	0.8757 ± 0.0062	0.6724 ± 0.0086	0.7606 ± 0.0056	0.7819 ± 0.0053
Resnik (MOL)	0.7189 ± 0.0110	0.6056 ± 0.0110	0.6572 ± 0.0054	0.2868 ± 0.0055
Lin (MOL)	0.6959 ± 0.0060	0.6227 ± 0.0075	0.6572 ± 0.0052	0.2799 ± 0.0055
Jiang (MOL)	0.5354 ± 0.0038	0.7085 ± 0.0074	0.6099 ± 0.0045	0.5550 ± 0.0055
simGIC (MOL)	0.8261 ± 0.0084	0.4472 ± 0.0084	0.5803 ± 0.0072	0.2816 ± 0.0055
Resnik (CEL)	0.8444 ± 0.0055	0.6785 ± 0.0071	0.7524 ± 0.0051	0.7725 ± 0.0052
Lin (CEL)	0.7789 ± 0.0053	0.7618 ± 0.0069	0.7702 ± 0.0044	0.7497 ± 0.0053
Jiang (CEL)	0.6862 ± 0.0049	0.7764 ± 0.0067	0.7285 ± 0.0043	0.7623 ± 0.0049
simGIC (CEL)	0.8796 ± 0.0065	0.6448 ± 0.0080	0.7441 ± 0.0054	0.7776 ± 0.0052

Table 1: Comparison between similarity measures. S. cerevisiae, mean ± standard deviation

non-interacting pairs of proteins in common yeast with fair reliability.

3.2 Homo sapiens

Our interaction dataset for H. sapiens is the second largest among the ones we will work with, and it contains 4463 interactions (of which 4316 belong to the "core" set). Table 2 summarizes the results of 1000 trials, much in the same way of Figure 1. Again, our measure is the one which scores the highest with respect to precision, F_1 score and ROC area, and it is second only to the Lin and Jiang-Conrath measures over the biological process subontology with respect to recall (and, again, these measures fare considerably worse than logSim with respect to the other metrics). Again, by applying the t-test for dependent paired samples to the logSim measure and to the best non-adaptive measure (for every metrics) we can see that the two distributions are statistically different, with t-statistics 70.61, -62.26, 213.83 and 531.06 respectively and all p-values evaluating to 0.0.

As discussed previously, Lin and Jiang-Conrath's measures tend to assign high similarity values to proteins which share annotations, even relatively shallow ones; and thus, it comes to no great surprise that their use leads to an improvement of the recall (more interacting proteins being recognized as such) at expense of the precision (more non-interacting proteins being mistakenly recognized). As in the case of the common yeast, non-adaptive similarities computed on the basis of the biological process sub-ontology perform better than the ones computed on the basis of the cellular component sub-ontology, which perform better yet than the ones computed on the basis of the molecular function sub-ontology. Figure 7.B shows the ROC curves for this dataset, for the logSim measure and for the non-adaptive measured based on the biological process sub-ontology, as well as the similarity distribution over interacting and non-interacting pairs of a testing set.

	precision	recall	F_1 score	ROC AUC
logSim	**0.8379 ± 0.0071**	0.7701 ± 0.0085	**0.8025 ± 0.0051**	**0.8802 ± 0.0038**
Resnik (BIO)	0.7831 ± 0.0067	0.7423 ± 0.0076	0.7621 ± 0.0049	0.8146 ± 0.0049
Lin (BIO)	0.7450 ± 0.0062	**0.7869 ± 0.0076**	0.7654 ± 0.0047	0.8004 ± 0.0051
Jiang (BIO)	0.7210 ± 0.0063	0.7857 ± 0.0083	0.7519 ± 0.0048	0.8070 ± 0.0050
simGIC (BIO)	0.8193 ± 0.0077	0.6546 ± 0.0097	0.7277 ± 0.0062	0.8033 ± 0.0049
Resnik (MOL)	0.7004 ± 0.0071	0.6295 ± 0.0076	0.6630 ± 0.0058	0.6744 ± 0.0066
Lin (MOL)	0.6581 ± 0.0062	0.6820 ± 0.0089	0.6698 ± 0.0058	0.6170 ± 0.0068
Jiang (MOL)	0.6019 ± 0.0053	0.6946 ± 0.0088	0.6449 ± 0.0054	0.6450 ± 0.0063
simGIC (MOL)	0.8064 ± 0.0095	0.4499 ± 0.0103	0.5774 ± 0.0086	0.6625 ± 0.0066
Resnik (CEL)	0.7862 ± 0.0074	0.5962 ± 0.0074	0.6781 ± 0.0061	0.7499 ± 0.0056
Lin (CEL)	0.6643 ± 0.0056	0.7618 ± 0.0087	0.7097 ± 0.0051	0.6882 ± 0.0062
Jiang (CEL)	0.6248 ± 0.0047	0.7676 ± 0.0076	0.6889 ± 0.0047	0.6926 ± 0.0059
simGIC (CEL)	0.7756 ± 0.0084	0.6016 ± 0.0102	0.6775 ± 0.0069	0.7466 ± 0.0059

Table 2: Comparison between similarity measures. H. sapiens, mean ± standard deviation

3.3 Escherichia coli

With only 1325 core interactions (out of 6971 total), the dataset for E. coli offers significantly fewer interactions from which to learn the parameters of our similarity measure than those of S. cerevisiae or H. sapiens. Thus, we might expect that our approach will find more difficulties in this setting. As the following results show, this is up to some degree true: not only is the performance of our measure significantly degraded in comparison to the previous cases, but now non-adaptive measures exist outperform it with respect to precision and to recall. However, it is worth remarking that no *single* measure outperforms logSim with respect to *both* precision and recall, and that with respect to F_1 score and ROC area our measure still significantly outperforms all the other ones (t-statistics 94.49 and 243.65 respectively, p-values still 0.0).

	precision	recall	F_1 score	ROC AUC
logSim	0.7847 ± 0.0162	0.6516 ± 0.0183	**0.7117 ± 0.0115**	**0.8009 ± 0.0093**
Resnik (BIO)	0.7185 ± 0.0145	0.5903 ± 0.0155	0.6480 ± 0.0116	0.6105 ± 0.0122
Lin (BIO)	0.7050 ± 0.0142	0.6176 ± 0.0178	0.6582 ± 0.0118	0.6085 ± 0.0124
Jiang (BIO)	0.6542 ± 0.0115	0.6540 ± 0.0152	0.6540 ± 0.0106	0.7039 ± 0.0111
simGIC (BIO)	**0.8686 ± 0.0156**	0.4636 ± 0.0162	0.6043 ± 0.0142	0.6205 ± 0.0125
Resnik (MOL)	0.7301 ± 0.0343	0.5819 ± 0.0623	0.6442 ± 0.0222	0.4695 ± 0.0133
Lin (MOL)	0.6580 ± 0.0111	0.6976 ± 0.0156	0.6771 ± 0.0105	0.4362 ± 0.0130
Jiang (MOL)	0.5616 ± 0.0084	**0.7450 ± 0.0134**	0.6404 ± 0.0089	0.6043 ± 0.0114
simGIC (MOL)	0.8197 ± 0.0162	0.4371 ± 0.0176	0.5699 ± 0.0161	0.4653 ± 0.0134
Resnik (CEL)	0.6780 ± 0.0188	0.5984 ± 0.0557	0.6338 ± 0.0250	0.4071 ± 0.0130
Lin (CEL)	0.6458 ± 0.0134	0.6354 ± 0.0368	0.6402 ± 0.0234	0.3774 ± 0.0126
Jiang (CEL)	0.5323 ± 0.0080	0.7108 ± 0.0128	0.6087 ± 0.0085	0.5574 ± 0.0113
simGIC (CEL)	0.6840 ± 0.0160	0.4276 ± 0.0152	0.5261 ± 0.0139	0.3929 ± 0.0130

Table 3: Comparison between similarity measures. E. coli, mean ± standard deviation

It might be possible — and experiments along these lines would seem to

confirm this — to improve somewhat the performance of our system in this setting by fiddling with the strength of regularization. However, we will not pursue this line of thought here, as it would make it harder to later compare the weights estimated by our system for different species.

3.4 Mus musculus

The house mouse (Mus musculus) presents perhaps the most challenging interaction dataset among those considered, with merely 1178 interacting pairs (quite a bit less than ten pairs per feature!) and no distinction between "core" and "full" pairs. Nonetheless, our approach continues to perform adequately. Indeed, essentially the same observations made in the case of E. coli hold for this case too: other similarity measures can outperform (in a statistically significant way, but by comparatively small amount) logSim in precision, but only at the cost of a steep loss in recall score, or vice versa, but no measure surpasses logSim in both precision and recall; and furthermore, logSim outperforms all the other measures in terms of F_1 score or ROC area (t-statistics 46.06 and 196.76, p-values 2.68E-249 and 0.0). The fact that logSim performs acceptably even in such a difficult scenario is a testament to the solidity of this approach, although — as image 7.D confirms — the reliability with which our system fulfills the task of distinguishing between interacting and non-interacting mouse proteins is less than entirely overwhelming.

	precision	recall	F_1 score	ROC AUC
logSim	0.7790 ± 0.0144	0.7057 ± 0.0200	**0.7403 ± 0.0124**	**0.8159 ± 0.0096**
Resnik (BIO)	0.7209 ± 0.0135	0.6952 ± 0.0169	0.7076 ± 0.0107	0.7217 ± 0.0123
Lin (BIO)	0.6941 ± 0.0119	0.7527 ± 0.0183	0.7221 ± 0.0103	0.7169 ± 0.0125
Jiang (BIO)	0.6669 ± 0.0119	0.7422 ± 0.0166	0.7024 ± 0.0098	0.7395 ± 0.0114
simGIC (BIO)	0.7840 ± 0.0159	0.5938 ± 0.0209	0.6755 ± 0.0141	0.7279 ± 0.0122
Resnik (MOL)	0.6677 ± 0.0131	0.6219 ± 0.0161	0.6439 ± 0.0118	0.6087 ± 0.0133
Lin (MOL)	0.6533 ± 0.0119	0.6732 ± 0.0164	0.6630 ± 0.0111	0.5807 ± 0.0136
Jiang (MOL)	0.5891 ± 0.0109	0.6984 ± 0.0170	0.6390 ± 0.0097	0.6364 ± 0.0121
simGIC (MOL)	**0.7904 ± 0.0183**	0.4295 ± 0.0206	0.5562 ± 0.0179	0.6104 ± 0.0136
Resnik (CEL)	0.7182 ± 0.0149	0.5963 ± 0.0168	0.6515 ± 0.0124	0.6984 ± 0.0125
Lin (CEL)	0.6467 ± 0.0109	0.7129 ± 0.0149	0.6781 ± 0.0101	0.6624 ± 0.0127
Jiang (CEL)	0.6188 ± 0.0099	**0.7532 ± 0.0154**	0.6793 ± 0.0093	0.6836 ± 0.0117
simGIC (CEL)	0.7705 ± 0.0168	0.5766 ± 0.0174	0.6593 ± 0.0129	0.7325 ± 0.0123

Table 4: Comparison between similarity measures. M. musculus, mean ± standard deviation

4 Cross-Species Training

Figure 8 represents the average weights assigned to each one of the 149 terms of the GO slim sub-ontology according to the four datasets. At a cursory first observation, it would seem that there are some similarities between the weights assigned to the GO slim terms in these four cases: for example, one may notice the peak occurring in all weight distribution at the GO slim term no. 77, which corresponds to the high-level term GO:0008150 (biological process), which is the root of the whole "biological process" sub-ontology. Thus, it would seem that a high similarity value over that sub-ontology can be often indicative of the existence of interactions, which is indeed supported also by the relative success of non-adaptive similarity measures based on the biological process sub-ontology with respect to all four species considered (see tables 1,2,3 and 4). On the other hand, all distributions assign very small values to term no. 131, corresponding to GO:0043226 (organelle). Thus, it would seem that not much can be inferred about the common activity (or lack thereof) of two proteins by the fact that they are both active in organelles of the cell. On the other hand, term no. 41, that is, GO:0005829 (cytosol) takes a high positive weight in the E. coli dataset, but not in other datasets. If the author may offer a speculation, this might perhaps be related to the fact that in prokaryotes, such as E. coli, most of the metabolic processes of the cell occur inside the cytosol, whereas in eukaryotes, like the other species considered in this work, it is more common for reactions to take place inside dedicated organelles.

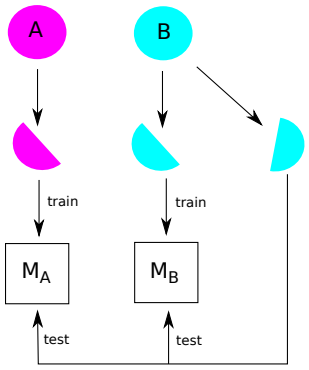


Figure 6: Cross-species training and testing: train two models on 1000 pairs from training sets for species A and B, test both of them on the testing set for species B. Then record $F_1(M_A)/F_1(M_B)$: the smaller it is, the greater it is the performance loss from training on a different species.

A natural question that may be asked at this point, therefore, is the degree up to which the information learned about interacting and non-interacting proteins from the evaluation of a dataset associated to a species applies also to different species. This may be considered an instance of the general problem of *domain*

adaptation [4]: how is the behaviour of a learning algorithm affected when the training and the testing data do not come from the same domain?[9]

In order to answer this, we proceed as follows: given two distinct species A, B in {S. cerevisiae, H. sapiens, E. coli, M. musculus}, let us split their datasets into a training set and a testing set, as usual, let us extract a fixed number of elements (in this work, one thousand) from the two training sets and let us train two models over them. Then let us extract another random testing set from the dataset for B, with possibility of overlap with the training set. This possibility inflates the scores of the model for B, of course; but nonetheless, it is the correct choice, as it guarantees the independence of the training set for B and the testing set for A *and* B: if we required that this latter set were disjoint from the training set for B, then even in the case $A = B$ we would obtain statistically different scores for the model trained over A and the one trained over B, which would be unacceptable. Furthermore, different species present ortholog proteins, which share a common evolutionary origin and tend to have similar annotations and to be involved into similar interactions [6] — and since we do not search and exclude orthologs of pairs in the testing sets from the training set for A, it would not be correct methodology to exclude possible repeats of the testing set from the training set for B. In any case, we then test the performance of both the model for A and the model for B against the testing set, we compute the F_1 scores, and we record the gains (or, typically, losses) resulting from using a model trained over A rather than one trained over B:

$$\text{gain}_{F_1} = \frac{F_1 \text{ score for } M_A}{F_1 \text{ score for } M_B}$$

We repeat this procedure one thousand times, every time for newly generated training and testing sets; then we record the means of the previously-mentioned gain scores and we test against the null hypothesis "gain= 1" (that is, no significant statistical difference was found between the scores computed over M_A and those computed over M_B).

Table 5 summarizes our results concerning this cross-trained F_1 measure. First of all, as we can see, the null hypothesis is rejected (at a very low significance level) whenever the training species is different from the testing species; and in this case, the score is always lower than the one obtained by training with the testing species (that is, the gain values are all smaller than one). This confirms that our model learns and makes use of species-specific information about which aspects of the Gene Ontology are more or less relevant to two proteins interacting or not interacting. Furthermore, it is also instructive to order, for any choice of "training" species, the "testing" species in decreasing gain order:

- S. cerevisiae → M. musculus → H. sapiens → E. coli
- H. sapiens → M. musculus → S. cerevisiae → E. coli
- E. coli → S. cerevisiae → M. musculus → H. sapiens

[9] However, in our case — differently from many others — we are not so much interested in reducing the resulting performance loss as in *measuring* it.

	S. cere	H. sapi	E. coli	M. musc
S. cere	1.0004 (0.16)	0.9867 (1.17E-215)	0.8371 (0)	0.9876 (3.63E-066)
H. sap	0.9520 (0)	1.0004 (0.28)	0.7194 (0)	0.9788 (2.63E-137)
E. coli	0.9874 (3.01E-228)	0.9456 (0)	1.0008 (0.29)	0.9500 (0)
M. musc	0.9442 (0)	0.9888 (4.31E-131)	0.7317 (0)	1.0005 (0.44)

Table 5: Cross-species F_1 gains. Between parentheses: p-values against hypothesis "gain = 1.0". Row = training, column = testing.

- M. musculus → H. sapiens → S. cerevisiae → E. coli

It is easy to see that whenever the training species is eukaryotic (that is, it is not E. coli), the gain is at its smallest when the testing species is E. coli. This suggests either that our system, when training over the E. coli dataset, captures unique features of prokaryotic organisms which it can then use to predict more accurately whether proteins interact or do not interact *within the context of a prokaryotic cell*, or that when training over eukaryotic species we capture unique features of eukaryotic organisms which prove themselves useless or even *detrimental* when used to try to infer whether proteins interact within a prokaryotic cell. A more careful analysis would be required in order to verify the degree up to which either of these hypotheses applied, although the fact that our approach appears to perform better on eukaryotic organisms than on E. coli suggests that the latter one may be better supported. In any case, the suggestion that the prokaryotic nature of E. coli is at the root of the relative failure of models trained on eukaryotic organisms to predict interactions between proteins in E. coli seems fairly reasonable and well supported by the data.

Similarly, it is also worth noticing that the two multicellular animals (and mammals) which we considered — that is, Homo sapiens and Mus musculus — are always put next to each other in the above lists, and in particular are each other's "best" other target species.

The very limited number of species considered in this work (as well as the fact that the same observation could not be replicated when computing the gain/loss with respect to ROC AUC scores: see Table 6) makes it impossible to reason more in depth about the ways in which the biological differences between different groups of organisms affect the learned weights to Gene Ontology slim terms; however, this data is at the very least highly suggestive of the hypothesis that such influences exist and are of some relevance for the detection of interactions between proteins. This lends support to our original point about the limits of generic or species-independent similarity measures and the advantages of tailoring any semantic similarity measure to a specific objective and category of organisms.

	S. cere	H. sapi	E. coli	M. musc
S. cere	1.0000 (0.9)	0.9878 (1.04E-284)	0.9858 (1.90E-123)	0.9731 (1.51E-272)
H. sap	0.9898 (1.10E-304)	0.9999 (0.57)	0.9730 (7.59E-249)	0.9852 (2.77E-127)
E. coli	0.9913 (1.14E-232)	0.9765 (0)	1.0011 (0.04)	0.9623 (0)
M. musc	0.9873 (0)	0.9919 (1.97E-169)	0.9677 (2.29E-313)	1.0004 (0.44)

Table 6: Cross-species ROC-AUC gains. Between parentheses: p-values against hypothesis "gain = 1.0". Row = training, column = testing.

5 Conclusions and Further Work

In this work, we developed and tested a simple adaptive semantic similarity measure for the detection of interactions between proteins on the basis of their gene ontology annotations. Much could be done in order to improve further the performance of the measure: for example, by employing *regularization* one can somewhat improve the performance of the system over the smaller datasets. A simple but interesting idea along these lines consists in a *graded* regularization, which would penalise the weights of deep, "specific" terms of our GO slim sub-ontology more than those of shallow, "generic" ones. Preliminary investigations along these lines have yielded mildly promising results; however, the main purpose of this work was not to provide a new way to automatically detect interactions between proteins (that task would be better faced by considering *more* information about proteins than merely their GO annotations — for example, their amminoacid sequence, or their geometric configurations) but rather to investigate the way in which a measure can be made to adapt itself to the peculiarities of a given species. It would be an interesting — and fairly ambitious — enterprise to develop similar adaptive measures for other possible purposes among the ones discussed in the introduction, then compare the resulting weight distributions, both between different measures over the same species and over different species with respect to the same measure, and attempt to learn insights about the relationships between different GO terms or groups of terms and the properties of the proteins which they annotate.

But this goes perhaps a little to far for now. More modestly, what can be concluded from the investigations described in this work is that contextual information — for example, the identity of the species we are investigating — can be used to good effect in order to fine-tune the parameters of a semantic similarity measure; and that, moreover, the way in which such a measure adapts itself to different scenarios can offer insights about properties and peculiarities of the organisms themselves.

Acknowledgments

I thank my advisor Dr. Novi Quadrianto, who supervised this work, for his invaluable advice and encouragement. Furthermore, I thank Dr. Di Lena and Dr. Casadio for mentioning to me the topic of similarity measures over the Gene Ontology in Spring 2014, thus sparking my interest in the problem.

References

[1] Gene Ontology Consortium et al. The gene ontology project in 2008. *Nucleic acids research*, 36(suppl 1):D440–D444, 2008.

[2] Francisco M Couto, Mário J Silva, and Pedro M Coutinho. Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 343–344. ACM, 2005.

[3] Francisco M Couto, Mário J Silva, et al. Disjunctive shared information between ontology concepts: application to gene ontology. *J. Biomedical Semantics*, 2:5, 2011.

[4] Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, pages 101–126, 2006.

[5] Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.

[6] Ben Lehner and Andrew G Fraser. A first-draft human protein-interaction map. *Genome biology*, 5(9):R63, 2004.

[7] Dekang Lin. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304, 1998.

[8] Catia Pesquita, Daniel Faria, Hugo Bastos, André Falcão, and Francisco Couto. Evaluating go-based semantic similarity measures. In *Proc. 10th Annual Bio-Ontologies Meeting*, volume 37, page 38, 2007.

[9] Catia Pesquita, Daniel Faria, Hugo Bastos, António EN Ferreira, André O Falcão, and Francisco M Couto. Metrics for go based protein semantic similarity: a systematic evaluation. *BMC bioinformatics*, 9(Suppl 5):S4, 2008.

[10] Catia Pesquita, Daniel Faria, Andre O Falcao, Phillip Lord, and Francisco M Couto. Semantic similarity in biomedical ontologies. *PLoS Comput Biol*, 5(7):e1000443, 2009.

[11] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 1*, pages 448–453. Morgan Kaufmann Publishers Inc., 1995.

[12] Michael M Richter. *Classification and learning of similarity measures*. Springer, 1993.

[13] Lukasz Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. The database of interacting proteins: 2004 update. *Nucleic acids research*, 32(suppl 1):D449–D451, 2004.

[14] Ioannis Xenarios, Lukasz Salwinski, Xiaoqun Joyce Duan, Patrick Higney, Sul-Min Kim, and David Eisenberg. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic acids research*, 30(1):303–305, 2002.

[15] Liu Yang. An overview of distance metric learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2007.

[16] Liu Yang and Rong Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2, 2006.

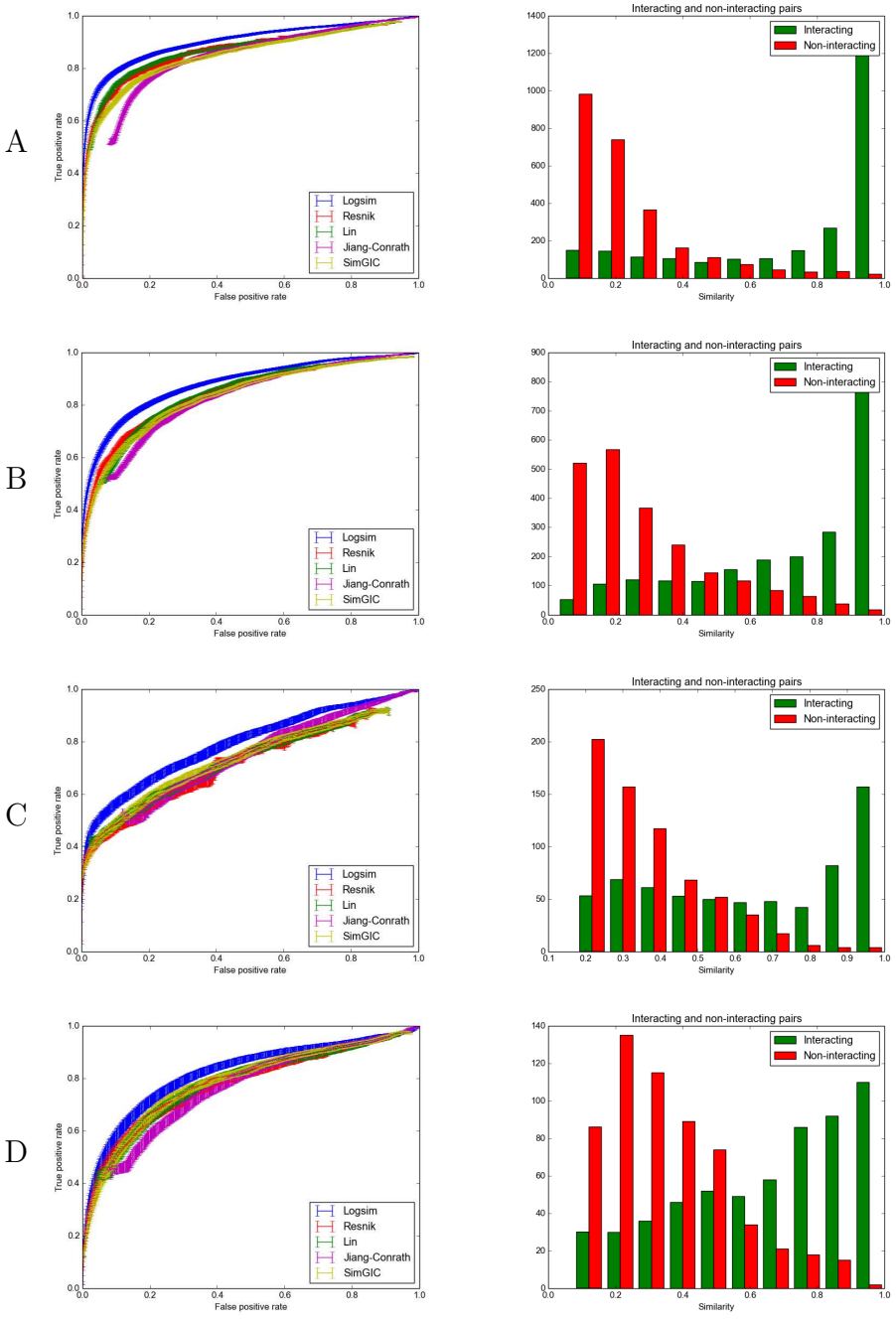


Figure 7: LogSim: ROC curves (left: width = 2*standard dev) and similarity scores for interacting/noninteracting proteins (right). A = S. cerevisiae, B = H. sapiens, C = E. coli, D = M. musculus.

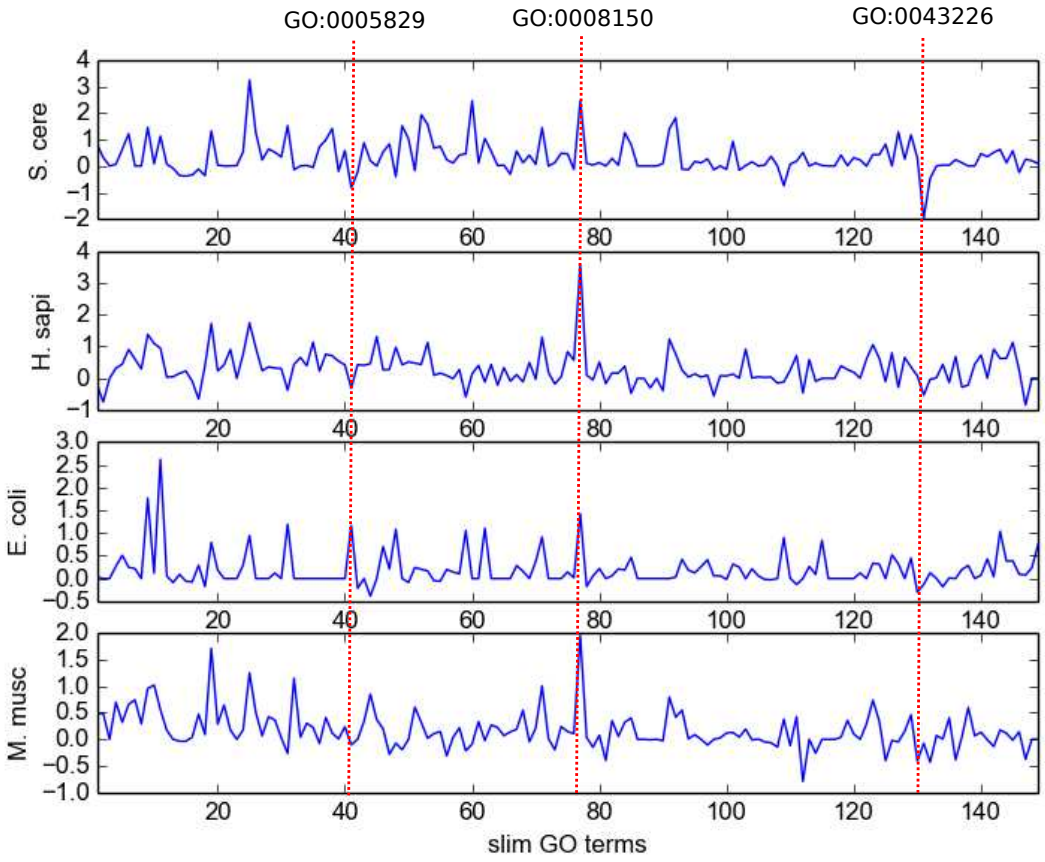


Figure 8: Average weights assigned to the 149 slim terms in the four datasets. Note that all weight distribution assign high importance to GO:0008150 (biological process) and low importance to GO:0043226 (organelle), but the one for E. coli disagrees with the others about the importance of term GO:0005829 (cytosol). This may be related to the fact that in prokaryotes, most metabolic processes occur directly inside the cytosol.