BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 21 2013, pages 2781–2787
doi:10.1093/bioinformatics/btt491

Databases and ontologies

Advance Access publication September 3, 2013

Exploiting disjointness axioms to improve semantic similarity measures

João D. Ferreira[1,*], Janna Hastings[2,3,4] and Francisco M. Couto[1]

[1]Department of Informatics, Faculdade de Ciências da Universidade de Lisboa, 1749-016 Lisboa, Portugal, [2]Cheminformatics and Metabolism, EMBL–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK, [3]Swiss Center for Affective Sciences, University of Geneva, 7, rue des Battoirs, 1205 Geneva, Switzerland and [4]Evolutionary Bioinformatics Group, Swiss Institute of Bioinformatics, Biophore - CH-1015 Lausanne, Switzerland

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Representing domain knowledge in biology has traditionally been accomplished by creating simple hierarchies of classes with textual annotations. Recently, expressive ontology languages, such as Web Ontology Language, have become more widely adopted, supporting axioms that express logical relationships other than class–subclass, e.g. disjointness. This is improving the coverage and validity of the knowledge contained in biological ontologies. However, current semantic tools still need to adapt to this more expressive information. In this article, we propose a method to integrate disjointness axioms, which are being incorporated in real-world ontologies, such as the Gene Ontology and the chemical entities of biological interest ontology, into semantic similarity, the measure that estimates the closeness in meaning between classes.

Results: We present a modification of the measure of shared information content, which extends the base measure to allow the incorporation of disjointness information. To evaluate our approach, we applied it to several randomly selected datasets extracted from the chemical entities of biological interest ontology. In 93.8% of these datasets, our measure performed better than the base measure of shared information content. This supports the idea that semantic similarity is more accurate if it extends beyond the hierarchy of classes of the ontology.

Contact: joao.ferreira@lasige.di.fc.ul.pt

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 20, 2012; revised on August 15, 2013; accepted on August 16, 2013

1 INTRODUCTION

Semantic similarity has direct application to the class–subclass hierarchy of many biomedical ontologies, such as the Gene Ontology (GO; Lord *et al*., 2003), the chemical entities of biological interest ontology (ChEBI; Ferreira and Couto, 2010) and the human phenotype ontology (Köhler *et al*., 2009). Semantic similarity assigns a quantitative measure of similarity between two entities in an ontology, which has seen multiple applications in semantic web and bioinformatics contexts (Grego and Couto, 2013).

The state-of-the-art in knowledge representation in the biomedical domain is evolving to make use of ontology languages such as the Web Ontology Language (OWL). OWL allows for more logically expressive axioms than the simple class–subclass hierarchy and the relational statements favored in early bio-ontology releases (McGuinness and van Harmelen, 2004). Following this trend, there is a need to adjust the current similarity measures to conform to current practices in ontology development (Couto and Pinto, 2013). Ontologies such as ChEBI and GO now contain disjointness axioms, which express for a pair of classes the constraint that an instance of one of them cannot also be an instance of the other [Although the terms 'class' and 'concept' are usually interchangeable in literature, the former is favored by the Semantic Web and OWL language communities and the latter by the description logic (DL) community. In this article, we use the term 'class'.]. The constraint also restricts subclasses from being a subclass of both of the disjoint classes. If such shared instances or subclasses are detected by an ontology reasoner, the reasoner will flag the ontology as *inconsistent*, which can be used by ontology developers as a validation step to prevent errors in ontology development.

In this article, we propose that disjointness axioms can also enhance the information that is exploited by similarity measures. Figure 1 illustrates this situation. In this snippet, it is stated that no instance of *Rectangle* can simultaneously be an instance of *Trapezoid*. However, given the open-world assumption that underlies ontologies, there can be instances of *Rectangle* that are also instances of *Parallelogram* (in fact, it is a consequence of the relevant geometric definitions that all squares are both rectangles and parallelograms). (Informally, the open-world assumption states that what is not known to hold does not give any information about what is known *not* to hold. One consequence is that if an ontology does not contain subclasses for a given class, it cannot be assumed that no such subclasses exist.) For this reason, the similarity between *Rectangle* and *Parallelogram* should intuitively be higher than the similarity between *Rectangle* and *Trapezoid*. Using σ to represent the two-argument function that returns the similarity between two classes:

$$\sigma(Rectangle, Parallelogram) > \sigma(Rectangle, Trapezoid) \quad (1)$$

*To whom correspondence should be addressed.

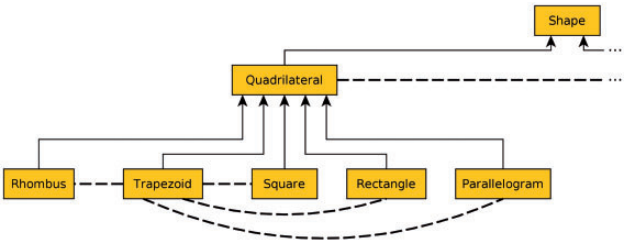

Fig. 1. A snippet of a hypothetical shape ontology. Arrows represent class–subclass relationships and dashed lines represent disjointness axioms. We use the term *Trapezoid* to mean a quadrilateral with two parallel sides and two obtuse angles. Note that proper shape ontology would classify *Square* as a subclass of *Rectangle*, *Rhombus* and *Parallelogram*. However, for the sake of the argument being exposed, we assume that such information is as yet unknown by the ontology creators

Several current semantic similarity measures make use of the idea of *information content* (IC) applied to the classes of the ontology (Resnik, 1995; Sánchez and Batet, 2011). The IC is a number that reflects how specific the class is. For example, in the illustration in Figure 1, *Shape* is the least specific class, receiving a lower IC than the other classes. There have been many proposals for how best to measure the IC of a class [See, e.g. Seddiqui and Aono (2010) and Van Buggenhout and Ceusters (2005)].

Another notion commonly used in semantic similarity is the *most informative common ancestor* (MICA; Resnik, 1995). This applies to a pair of classes x and y, and is defined as the class with the highest IC from the set of all ancestors of both x and y:

$$\mathrm{MICA}(x,y) = \arg\max_c\{\mathrm{IC}(c) \mid c \in \mathrm{A}(x) \cap \mathrm{A}(y)\} \quad (2)$$

where $\mathrm{A}(x)$ is the set of ancestors (super-classes) of x, including x.

The first semantic similarity measure to make use of IC, by Resnik (1995), estimates similarity as the IC of the MICA between x and y. The motivation behind this choice of the formula is simple: x and y share a certain amount of information, and the MICA is one way to estimate this shared information. Many semantic similarity measures are based on this notion of shared IC (Jiang and Conrath, 1997; Lin, 1998; Pesquita *et al*., 2008). For example:

$$\sigma_{\mathrm{Resnik}}(x,y) = \mathrm{IC}(\mathrm{MICA}(x,y)) \quad (3)$$

$$\sigma_{\mathrm{Lin}}(x,y) = \frac{2 \times \mathrm{IC}(\mathrm{MICA}(x,y))}{\mathrm{IC}(x) + \mathrm{IC}(y)} \quad (4)$$

On the other hand, work has been published recently showing a new approach to the problem of finding the best way to measure shared IC between two classes. Although shared IC has been assumed to be best estimated as $\mathrm{IC}(\mathrm{MICA}(x,y))$ (Resnik, 1995), Couto and Silva (2011) suggest DiShIn, which behaves as an *add-on* to the measure of IC, and contributes to a better measure of shared IC by exploring *multiple parentage* to ensure that all the shared information across multiple ancestors is taken into account.

Just as was done for DiShIn, instead of proposing a semantic similarity measure, we propose an *add-on* that can be used by existing measures, such as the ones in Equations (3) and (4). Our *add-on* refines the estimation of shared information between the two classes by incorporating the disjointness axioms in the ontology. We call the new shared IC measure $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}(x,y)$, which will be based on a prior measure of shared IC, denoted by $\mathrm{IC}^{\mathrm{s}}(x,y)$. We stress that any measure of shared IC can be used as a base to $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}$, not just the one proposed by Resnik, as is the case with DiShIn.

Given the example presented in Figure 1 and the inequality of Equation (1), it would be desirable for the measure of shared IC to decrease for classes that are known to be disjoint, to formalize the intuition that disjoint classes are less similar, as they cannot share instances. Furthermore, to respect the open-world assumption that often accompanies ontologies, the measure should stay unchanged when two classes are not known to be disjoint.

With this novel measure of shared IC, we intend to show that semantic similarity can take advantage of the disjointness axioms of an ontology, thus providing evidence that future measures should consider them in evaluating the closeness in meaning between two classes.

2 CHEBI

For the evaluation of our proposal, we have computed shared IC for ChEBI, the ontology of Chemical Entities of Biological Interest (Degtyarenko *et al*., 2008). It is worth, as such, to introduce the reader to the state of disjointness information that this ontology includes. In the Open Biological and Biomedical Ontologies community (in which ChEBI is embedded), there is a tacit agreement that it is good practice to ensure that sibling terms are mutually disjoint. This is, however, not the case for ChEBI: mid-level chemical classes, which constitute most of ChEBI, are generally not pairwise disjoint, as chemical classification is compositional, i.e. classes often reflect parts or properties of molecules that may co-occur in many different combinations in fully specified molecules (Hastings *et al*., 2012b).

In an ontology of chemical compounds, a leaf class can, in theory, be regarded as disjoint with the other leaf classes. For example, *α-D-glucose* is disjoint with *histidine*. However, ChEBI is not a complete ontology for chemistry, and some of the leaves it contains do not follow this rule. For example, *aminophospholipid*, defined as 'a phospholipid that contains one or more amino groups', is a leaf in ChEBI at present. However, this class represents the molecules that contain specific substructures and, as such, it is not necessarily disjoint with the other leaves. Given that ChEBI is a work in progress, where new knowledge is added after careful manual duration, this has resulted in *aminophospholipid* being presently a leaf. Other such cases can be found, rendering even the theoretical rule that all leaves are disjoint not applicable.

Thus, in what follows, we have not attempted to automatically enhance the number of disjointness axioms available in ChEBI. Rather, we have used only those axioms that have explicitly been added to the ontology.

3 METHODS

3.1 Shared information using disjointness

We propose the new measure of shared IC:

$$\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}(x,y) = \mathrm{IC}^{\mathrm{s}}(x,y) - k(x,y) \quad (5)$$

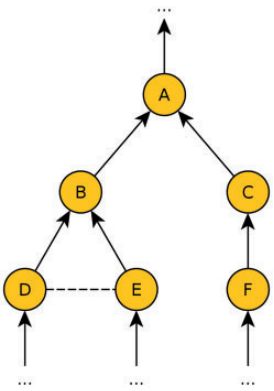

Fig. 2. An illustration of an ontology containing disjointness axioms. Arrows between classes represent class–subclass relationships and dashed lines represent disjointness

where $IC^s(x, y)$ is any measure of shared IC between x and y, $k(x, y) > 0$ if x and y are disjoint and $k(x, y) = 0$ otherwise.

Two points were crucial in the development of our measure. First, we note that, as is, this equation presents a *discontinuity*. In the hypothetical ontology of Figure 2, this measure leads to $IC^s_{disj}(D, E) < IC(B)$. Depending on the value $k(D, E)$, this could lead to $IC^s_{disj}(D, E) < IC(A) = IC^s_{disj}(D, F)$, which, however, should not be possible, as D and E share more information than D and F. Therefore, k was bounded according to the IC of the most informative ancestor of the MICA, which, in this case, results in $k(D, E) \leq IC(B) - IC(A)$.

The second major point associated with our measure is an operational notion: the likelihood of two classes sharing ancestors that are not asserted as such. We call this the potential for *implicit common ancestors* (ICAs). Take as an example the ontology snippets of Figure 3. In situation B, given the open-world assumption, there is a small chance that Y turns out to be a subclass of X', while in situation A that cannot happen, as Y is inferred to be disjoint with X'. This suggests that there is a lower potential for ICA between the classes X and Y in situation A, as the disjointness is declared between the direct subclasses of M.

Rather than modeling the potential for ICA, we model the *unlikelihood* of ICA as the function $f(x, y)$, which returns higher values for situations with lower potential for ICA:

$$f(x, y) = \max\left\{\frac{1}{p(a, b)} \mid a \in A(x) \wedge b \in A(y) \wedge J(a, b)\right\} \cup \{0\} \quad (6)$$

where $A(x)$ is the set of ancestors of x (including x), $J(a, b)$ is true when a and b are disjoint (either by assertion of inference), and false otherwise and $p(a, b)$ is the length of the shortest path from a to b. The path length takes into account only the class–subclass relations, not the disjointness arcs (the dashed edges of the figures).

Using the example ontologies in Figure 3, we can illustrate this definition by calculating $f(X, Y)$. In B, $J(a, b)$ is true only for $(a, b) = (X, Y)$; the shortest path from X to Y, using only class–subclass relations, is $X \to X' \to M \to Y' \to Y$, which has length 4. Therefore, $f(X, Y) = \max\{\frac{1}{4}, 0\} = \frac{1}{4}$. In A, $J(a, b)$ is true for $(a, b) \in \{(X, Y), (X, Y'), (X', Y), (X', Y')\}$. These correspond to paths of length $\{4, 3, 3, 2\}$, respectively, leading to $f(X, Y) = \max\{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 0\} = \frac{1}{2}$. Finally, for non-disjoint terms, such as X' and Y' in situation B, $J(a, b)$ is always false: therefore, the first set of the union becomes empty, resulting in $f(x, y) = 0$.

The general procedure to calculate $IC^s_{disj}(x, y)$ is, therefore:

(1) Determine $M = MICA(x, y)$;

(2) Determine $Z = \arg\max_c\{IC(c) \mid c \in A(M)\}$, *i.e.* the most informative ancestor of M;

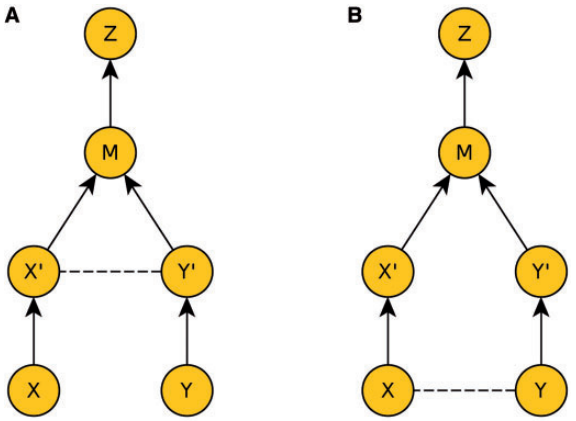

Fig. 3. This image illustrates the notion of the potential for ICAs between two classes. In both cases, $MICA(X, Y) = M$, and the most informative ancestor of M is Z. The difference is in the location of the disjointness axiom. In situation A, there is a lower likelihood of ICA between X and Y, because the axiom of disjointness is closer to their common ancestry

(3) Estimate the unlikelihood of ICA, $f(x, y)$, as described in (6);

(4) Calculate $k(x, y) = f(x, y) \cdot (IC(M) - IC(Z))$;

(5) Calculate $IC^s_{disj}(x, y) = IC^s(x, y) - k(x, y)$.

With this procedure, the new shared IC is estimated as a weighted average between $IC(M)$ and $IC(Z)$, where a higher f (lower potential for ICA) leads to a shared IC closer to $IC(Z)$ and lower f (higher potential for ICA) leads to a shared IC closer to $IC(M)$. This means that the shared IC decreases by a larger amount when there is a smaller potential for ICAs. Note that if the two classes are not disjoint, $k(x, y) = 0$ and $IC^s_{disj} = IC^s$, which satisfy the open-world assumption.

3.2 Assessment

The assessment was done in the following three steps: (i) increase in correlation coefficient, (ii) effect of the number of disjointness axioms and (iii) the effect of our measure in several random datasets.

3.2.1 Increase in correlation coefficient First, we applied our new measure of shared IC to a subset of ChEBI. Disjoint axioms were supplied by the ChEBI development team (Hastings *et al.*, 2012a, 2013), and the main ontology was directly extracted from the official web page (http://www.ebi.ac.uk/chebi/downloadsForward.do) on October 18, 2012 (corresponding to version 96 of the ontology).

To avoid any bias to an external corpus, IC for a class c was calculated with an intrinsic measure based on the total number of direct and inferred subclasses of c (Van Buggenhout and Ceusters, 2005):

$$IC(c) = -\frac{1}{\log N} \cdot \log\frac{|D(c)|}{N} \quad (7)$$

where $D(c)$ is the set of subclasses of c (including c) and N is the total number of classes in the ontology. For example, leaves of the ontology (those classes without any descendants) have the maximum possible IC, 1.0. It is worth noting again that this is but one of the many possible ways of measuring IC, and that our measure can be adapted to any of them. We also used, for this assessment, the classical notion of shared IC proposed by Resnik (1995):

$$IC^s(x, y) = IC(MICA(x, y)) \quad (8)$$

The subset of chemical classes from ChEBI used in this assessment (see Supplementary Material A) was randomly selected by first choosing

a pair of asserted disjoint classes in the ontology, A' and B', and then choosing two classes A and B, respectively, descendants of A' and B', both fulfilling two conditions:

- **Classes, not leaves**: as many of ChEBI's leaves represent fully specified chemical compounds but we lack a trivial way to detect whether they do so (see Section 2), we decided not to use the leaves in the testing dataset.
- **Classes with sufficient structural information**: a class was included in the dataset if either (i) it contains a Simplified Molecular-Input Line-Entry System (SMILES) representation of its chemical structure or (ii) at least 80% of its leaf descendants contain such a representation. This allowed us to compare semantic similarity with a purely structural measure, as explained later in the text. Only classes in the *chemical entity* branch of ChEBI can fulfill this condition.

These selection criteria were applied until 40 distinct classes were found. Therefore, the resulting set contained some pairs of classes that are disjoint and some that are not. We chose to create a dataset bounded by a number of classes rather than use all disjointness axioms at once because it would be much larger and therefore analysis would take more time.

To assess the usefulness of the disjointness axioms, we calculated the Pearson's correlation coefficient between the outcome of IC^s_{disj} and a purely structural measure of similarity between every pair of compounds in the dataset created previously. Semantic similarity, in general, is not intended to replace structural measures of similarity but to complement them with a knowledge-oriented perspective. Thus, it may seem strange, at first, that we use the correlation between structural similarity and IC^s_{disj} as a way to validate our measure. However, ChEBI's *chemical entity* branch models chemistry knowledge largely based on the structural properties of the molecules. As such, it is to be expected that measures of semantic similarity between classes from this branch of the ontology reflect, to some extent, the structural similarity between them. Therefore, in the particular case of this branch of ChEBI, semantic similarity should also reflect structural similarity, and, as such, it is valid to assume that an ontology-based measure with a higher correlation to structural similarity is better at estimating similarity than a measure with lower correlation to structural similarity.

To measure structural similarity, we used PubChem's fingerprint method (Bolton *et al*., 2008). To compare classes x and y, we extracted the SMILES representations associated with their leaf descendants, using a best-match average approach to average over all the similarities, as follows:

(1) For class x, choose the leaf descendant classes that contain SMILES information, $\{x_1, \ldots, x_n\}$. If x has SMILES information, assume $n = 1$ and $x_1 = x$. Do the same for y to achieve the set $\{y_1, \ldots, y_m\}$.
(2) Generate a PubChem fingerprint for each x_i and y_j.
(3) Compare all the x_i fingerprints with all the y_j fingerprints, with the Tanimoto coefficient (Flower, 1998), generating the matrix of structural similarities $s(x_i, y_j)$.
(4) For each i, find $f_x(i) = \max_j\{s(x_i, y_j)\}$; and for each j, find $f_y(j) = \max_i\{s(x_i, y_j)\}$.
(5) Assign $\frac{\sum_i f_x(i) + \sum_j f_y(j)}{n+m}$ to the structural similarity between x and y.

In summary, for the dataset created above, we compared all compounds with all the other compounds (780 distinct pairs) using three measures: structural similarity, classical IC^s and IC^s_{disj}. We proceeded by using Wolfe's t-test (Rosner, 2010; Wolfe, 1976) to determine the statistical significance of the increase from the correlation coefficient between "structural similarity" and IC^s to the correlation coefficient between "structural similarity" and IC^s_{disj}.

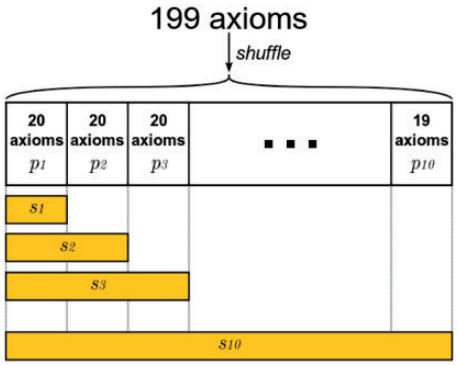

Fig. 4. The process used to assess the effect of the number of axioms on the correlation coefficient between structural and semantic similarity. The axioms are randomly partitioned into clusters p_1–p_{10}. Consecutively, each of these clusters is joined with the previous ones to create a set $s_i = \bigcup_{j=1}^{i} p_j$, which is then used to compute the increase in correlation coefficient

It is important to notice here that we do our analysis over the raw value of IC^s_{disj}, rather than any one measure of similarity based on this value [such as in Equation (4)]. This was done to show that we can increase the actual utility factor of the measure of shared IC rather than the utility of a specific measure of similarity.

3.2.2 Effect of the number of axioms The second assessment step was aimed at measuring the effect of the number of disjointness axioms in the increase of correlation. To this end, we created 10 sets of axioms based on the full set of 199 axioms of disjointness to which we had access. The first set contained 20 random axioms, the second contained these same 20 axioms plus 20 other random ones, etc, until the final set, which contained all the 199 axioms (see Fig. 4 for a graphical representation of this procedure). For each of these sets, we ran the IC^s_{disj} algorithm and plotted a graph showing the increase in correlation versus the number of axioms. To remove any bias that could have resulted from the random method used to create these groups, we ran the same experiment 20 times.

3.2.3 Effect on other datasets As the third assessment step, we studied the increase in correlation coefficient on other datasets, as the dataset created for the first step resulted from a random selection process. Following the same selection process presented above, we created 550 more datasets (all with either 40 or 41 compounds) and compared the correlation coefficient as previously explained.

3.2.4 Implementation The semantic-related algorithms we used were implemented by us in Java, based on the OWL–API (Horridge and Bechhofer, 2011). Python was used to calculate Pearson correlation coefficients and to compute the P-values of the Wolfe's t-test. We used `scipy` (http://www.scipy.org/) package for this effect.

4 RESULTS AND DISCUSSION

We present three main results stemming from the comparison of structural and semantic similarity measures. Our main assumption is, as stated above, that in the *chemical entity* branch of ChEBI, a measure that correlates more strongly with structural similarity is performing better at estimating the real similarity than a measure with a lower correlation.

4.1 Increase in correlation coefficient

Our first conclusion is that exploring the axioms of disjointness leads to an increase in the correlation between structural and semantic similarity.

The Pearson's correlation coefficient between the structural measure and IC^s is 0.69883, and after taking the disjointness axioms into account, the correlation for structural similarity versus IC^s_{disj} becomes 0.71571. This represents an increase of 0.01688. Despite the small absolute increase, this value is statistically significant, with a *P*-value of 4.5×10^{-8}.

The small increase of the correlation can be attributed to at least three factors:

- As the annotation of disjointness is still incomplete in ChEBI, we have access to only a small subset of all the *real* disjointness axioms that could be expressed in ChEBI, which means that the shared IC changes only for a fraction of all the class pairs (39% from the sample selected). As more axioms of this kind are included in ChEBI, we expect both this fraction and the difference between correlation coefficients to increase.
- Although highly correlated, structural similarity and semantic similarity measures are inherently different, and as such, there is a maximum bound on the actual correlation that can be expected between the two. Also, different classes within ChEBI can be expected to show a lower increase, whereas others show a higher increase.
- Disjointness is only one of the logical axiom types that are used to express class definitions in an OWL ontology. In fact, ChEBI contains a number of other properties that are also used to capture the meaning of its classes, e.g. the property **has-tautomer**, which connects together closely structurally related chemicals, and **has-role**, which connects a chemical class to its biological activity.

4.2 Effect of the number of axioms

To clarify the first item above, we performed the second assessment step, which aims to simulate the development of the ChEBI ontology with respect to the number of disjointness axioms. For each of the 20 runs, we studied the difference between the correlation coefficients as the number of disjointness axioms increases, and plotted a graph with this information.

The graphs in Figure 5 show the result of some of these experiments. These graphs illustrate that not all disjointness axioms are important for a given dataset. In fact, only for some of the sets of axioms is the correlation coefficient significantly affected, which suggests that those sets contained the axioms that change the logical meaning behind the classes in the dataset. The graphs present a very obvious trend (see Fig. 6 for an average of the graphs of all the 20 experiments) that indicates an increase of the correlation, which, again, indicates that the disjointness axioms improve the correctness of the measure of semantic similarity.

4.3 Effect on other datasets

As the dataset created for the purpose of the results presented before resulted from a random selection process, we also studied the effect of considering the axioms of disjointness in other 550 datasets. The graph of Figure 7 is a histogram that represents the difference in the Pearson's correlation coefficient for all these datasets. As is visible in that graph and in Table 1, the vast

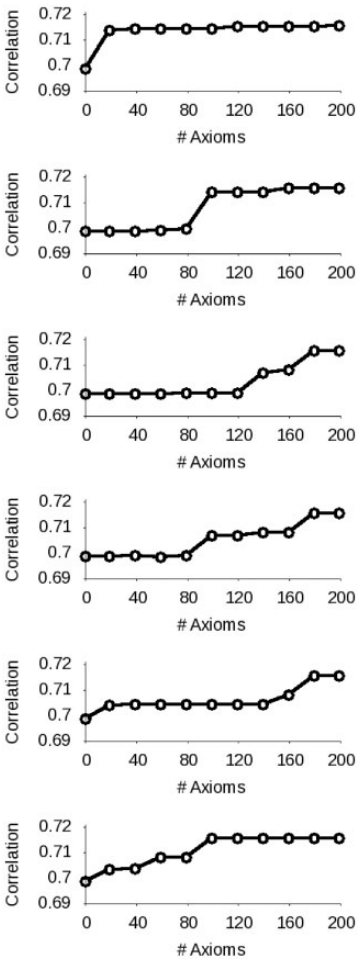

Fig. 5. These graphs illustrate the effect of the number of disjointness axioms on the correlation coefficient between structural and semantic similarity. In each graph, the abscissa is the number of axioms used by the semantic similarity measure, and the ordinate is the correlation coefficient. The correlation coefficient for 0 axioms is always equal to the correlation measured with the classical IC^s, which is 0.69883; the correlation coefficient for the maximum number of axioms corresponds to the value 0.71571 presented in Section 4.1. These graphs are representative of the behavior obtained in all of the 20 runs

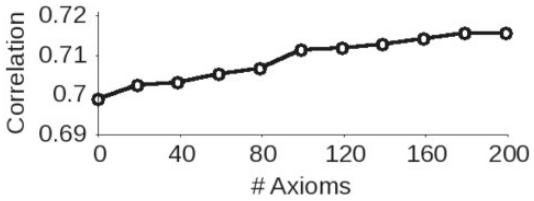

Fig. 6. This graph shows the average of all the graphs produced in the 20 runs of Section 4.2. Although these values do not have any statistical significance in themselves, they clearly show the trend that the more disjointness axioms are considered, the better is the correlation between structural and semantic similarity

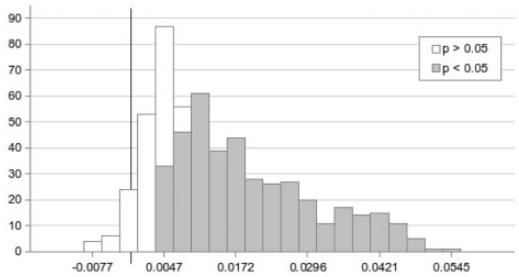

Fig. 7. Distribution of the difference in correlation coefficient for 550 random datasets. The majority of the cases show a positive difference. We used Wolfe's *t*-test to calculate the *P*-value associated with the hypothesis that the increase was due to random chance, and marked with a darker shade the amount of datasets for which $P < 0.05$. The vertical line shows the 0 of the axis, i.e. where the two correlation coefficients are the same

Table 1. Statistics related to the histogram of Figure 7

Characteristic	Number of datasets	Percentage datasets
Increase in correlation	516	93.8%
$P < 0.05$	399	72.5%

Note: The last column shows the frequency relative to all the 550 datasets created.

majority of the datasets are associated with an increase in the correlation coefficient. In fact, the effect of considering the disjointness axioms for the semantic similarity only impacts negatively 6.2% of the datasets. We observed a mean correlation increase of 0.0149, with a standard deviation for that value of 0.0130. Furthermore, in 72.5% of the datasets, the increase in correlation is significant at a confidence value of 0.05.

5 LIMITATIONS

Although the work presented here shows with statistical strength the utility of IC^{s}_{disj} when measuring shared IC between two classes, it can still be improved. We presented the *discontinuity* problem, and how to avoid it by restricting k so that shared IC never reduces below IC(Z) (where Z is the most informative ancestor of the MICA). This can lead to some other problems. For example, future changes to the ontology can lead to unexpected changes in IC^{s}_{disj}. Consider the ontology change of Figure 8. Assuming 1000 classes in the ontology, $IC(B) \approx 0.77$ and $IC(A) = 0$. After the step illustrated in the figure, $IC(X) \approx 0.72$. This means that $IC^{s}_{disj}(E, F)$ increases unexpectedly from 0.38 to 0.74 because of a very small change in the ontology. These kinds of top-level additions, however, are not very common, and as such, the magnitude of this particular *jump* in similarity is not expected to happen very often.

A second point of future development in our measure concerns Equation (6), used to model the potential for ICAs. Our approach depends on the edge distance between two classes;

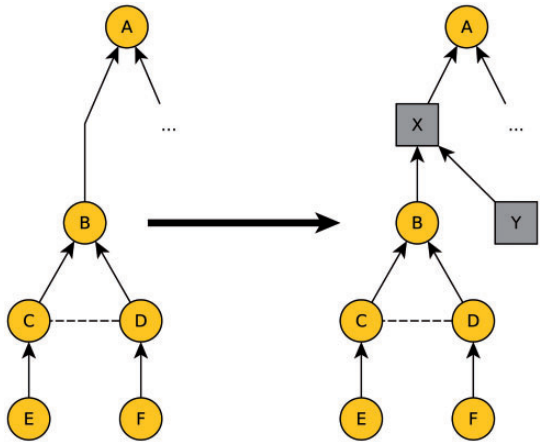

Fig. 8. A hypothetical developing step in one ontology. From one iteration to the next, the ontology gained a new term between *A* and *B*. Before this change, the similarity between *E* and *F* depends on the difference IC(B) − IC(A); after the change it depends on IC(B) − IC(X)

however, it may be possible to explore the semantics of the edges themselves to refine this measure.

Another important point to notice in this work is that the measure of IC influences the results obtained with IC^{s}_{disj}. In this case, IC was calculated with the information contained in the ontology alone. It would be informative to see the effect of changing the IC measure used with IC^{s}_{disj} to a more realistic one.

Finally, this assessment is valid in ChEBI because we may assume that semantic similarity correlates with structural similarity in the particular branch we used. For other ontologies, where such assumption is not valid, the most promising way to validate would be to rely on external and curated gold standards.

6 CONCLUSION

Recently, Couto and Pinto (2013) presented the benefits that result from using DL axioms in the calculation of similarity. To the best of our knowledge, this article is the first attempt to include DL axioms into ontology-based similarity measures in the biomedical domain.

Accordingly, the main purpose of this work was to test whether exploiting the disjointness axioms of an ontology increases the performance of shared IC measures. We developed an *add-on* that can be used with any measure of shared IC, called IC^{s}_{disj}, which satisfies the designated requirements set forth in the beginning of the work, particularly that its value should decrease for disjoint pairs of classes.

The assessment of our measure, which is based on the Pearson's correlation coefficient between structural similarity and semantic similarity, has shown that there is, in fact, an improvement of the measure of shared IC because its correlation with structural similarity in an ontology that encodes structural knowledge increases as the number of disjointness axioms increase.

This new approach is able to successfully explore more than just the subsumption hierarchy of an ontology, relying

additionally on a partial subset of the description logic axioms that are included in the ontology to further refine the comparison of two classes.

To the best of our knowledge, this represents the first attempt to use description logic expressivity in semantic similarity in the biomedical domain. We demonstrated our hypothesis that disjointness axioms contain informative data that can be correctly explored by semantic similarity measures, even with a naïve approach. More sophisticated approaches may include the exploration of the semantics of edges, other types of IC based on external corpus, etc.

In conclusion, this work strongly suggests that future measures of semantic similarity should consider the full logical formalism of the ontologies that they use to establish a measure of similarity that more accurately reflects the reality of the domain of knowledge therein modeled.

Funding: Fundação para a Ciência e Tecnologia PhD (SFRH/BD/69345/2010 to J.D.F.); Lasige Multiannual Funding Programme, SOMER project (PTDC/EIA-EIA/119119/2010) and by the European Commission (EU-OPENSCREEN project to J.H.).

Conflict of Interest: none declared.

REFERENCES

Bolton,E.E. *et al.* (2008) PubChem: integrated platform of small molecules and biological activities. In: Wheeler,R.A. and Spellmeyer,D.C. (eds) *Annual Reports in Computational Chemistry*. Vol. 1, chapter 12. American Chemical Society, Washington, DC, pp. 217–241.

Couto,F.M. and Pinto,H.S. (2013) The next generation of similarity measures that fully explore the Semantics in Biomedical Ontologies. *J. Bioinform. Comput. Biol.*, **11**, 1371001.

Couto,F.M. and Silva,M.J. (2011) Disjunctive shared information between ontology concepts: application to Gene Ontology. *J. Biomed. Semantics*, **2**, 5.

Degtyarenko,K. *et al.* (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344.

Ferreira,J.D. and Couto,F.M. (2010) Semantic similarity for automatic classification of chemical compounds. *PLoS Comput. Biol.*, **6**, e1000937.

Flower,D. (1998) On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.*, **38**, 379–386.

Grego,T. and Couto,F.M. (2013) Enhancement of chemical entity identification in text using semantic similarity validation. *PLoS One*, **8**, e62984.

Hastings,J. *et al.* (2012a) Modular extensions to the ChEBI ontology. In: *International Conference on Biomedical Ontologies*. Graz, Austria.

Hastings,J. *et al.* (2012b) Structure-based classification and ontology in chemistry. *J. Cheminform.*, **4**, 8.

Hastings,J. *et al.* (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, **41**, D456–D463.

Horridge,M. and Bechhofer,S. (2011) The OWL API: a java API for OWL ontologies. *Semant. Web*, **0**, 1–11.

Jiang,J.J. and Conrath,D.W. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: *International Conference Research on Computational Linguistics, Rocling X*. Taiwan.

Köhler,S. *et al.* (2009) Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.*, **85**, 457–64.

Lin,D. (1998) An information-theoretic definition of similarity. In: *15th International Conference on Machine Learning*. Vol. 1, Madison, Wisconsin, USA, pp. 296–304.

Lord,P.W. *et al.* (2003) Semantic similarity measures as tools for exploring the gene ontology. In: *Proceedings of the Pacific Symposium on Biocomputing*. Vol. 8, Lihue, Hawai'i, pp. 601–612.

McGuinness,D.L. and van Harmelen,F. (2004) OWL web ontology language overview. *W3C Recomm.*, **10**, 10.

Pesquita,C. *et al.* (2008) Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, **9(Suppl. 5)**, S4.

Resnik,P. (1995) Using information content to evaluate semantic similarity in a taxonomy. In: *Proceedings of the 14th International Joint Conference on Artificial intelligence*. Vol. 1, Montréal, Canada.

Rosner,B. (2010) Statistical Inference for Correlation Coefficients. In: *Fundamentals of Biostatistics*. 7th edn. chapter 11, Cengage Learning, pp. 466.

Sánchez,D. and Batet,M. (2011) Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective. *J. Biomed. Inform.*, **44**, 749–59.

Seddiqui,H. and Aono,M. (2010) Metric of intrinsic information content for measuring semantic similarity in an ontology. In: *Proceedings of the Seventh Asia-Pacific Conference Modelling (Apccm)*. Darlinghurst, Australia, pp. 89–96.

Van Buggenhout,C. and Ceusters,W. (2005) A novel view on information content of concepts in a large ontology and a view on the structure and the quality of the ontology. *Int. J. Med. Inform.*, **74**, 125–32.

Wolfe,D.A. (1976) On testing equality of related correlation coefficients. *Biometrika*, **63**, 214–215.