Semantic similarity definition

Francisco M. Couto*

LaSIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, 1749- 016 Lisboa, Portugal

Andre Lamurias

Department of Computer Science, Aalborg University, Denmark

Pedro Ruas

LaSIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, 1749- 016 Lisboa, Portugal

Abstract

In bioinformatics, semantic similarity has been used to compare different types of biomedical entities, such as proteins, compounds and phenotypes, based on their biological role instead on what they look like. This manuscript presents a definition of semantic similarity between biomedical entities described by a common semantic base (e.g. knowledge graph, ontology) following an information-theoretic perspective of semantic similarity. It defines the amount of information content two entries share in a semantic base, and, by extension, how to compare biomedical entities represented outside the semantic base but linked through a set of annotations. Software to check how semantic similarity works in practice is available at: `https://github.com/lasigeBioTM/DiShIn/`.

Keywords: Biomedical Ontologies, Knowledge Graphs, Biomedical Vocabularies, Functional Analysis, Information Content, Information Theory, Knowledge Organization Systems, Semantic Similarity, Semantic Web, Similarity Measures

☆Accepted manuscript: *in press*

*Corresponding author

Email addresses: `fcouto@di.fc.ul.pt` (Francisco M. Couto), `andrel@cs.aau.dk` (Andre Lamurias), `psruas@fc.ul.pt` (Pedro Ruas)

1. Introduction

The biological role of an entity is considered to be its semantics, which has been increasingly represented through common vocabularies. The entries in these vocabularies represent biological features, that are often connected with each other by semantic relations, such as subsumption. The availability of these common vocabularies, and their usage to semantically annotate entities enabled the development of computational semantic similarity measures (Batet et al., 2014). Before defining semantic similarity, we should start by defining why bioinformatics needs semantic similarity in the first place, then what it is, to finally describe how it can be calculated.

1.1. Why?

Biomedical entities, such as proteins or chemical compounds, are frequently compared to each other to find similarities that may enable us to transfer knowledge from one another. In the case of proteins, one of the most popular techniques is to calculate sequence similarity by locating short matches between sequences and then generate local alignments (Smith and Waterman, 1981). In the case of compounds, one of the most popular techniques is to calculate the number of 2D substructural fragments (molecular fingerprints) that they have in common (Willett, 2011). The above techniques are popular mainly because they can be implemented by high-performance tools, such as BLAST (Altschul et al., 1997), and are based on simple, unambiguous and widely available digital representations. However, these digital representations result from observations of how these biomedical entities look like, and not about their semantics. This means that we cannot have a direct insight into their biological role. Sequence similarity and common fingerprints measure how close two entities are in terms of what they look like, which may differ from their biological role.

There is a relationship between what an entity looks like and its biological role, i.e. proteins with similar sequence tend to have similar molecular functions,

as well as with compounds with similar molecular shapes. However, there are many exceptions. For example, crystallins have a high sequence similarity to several different enzymes due to evolution, but in the eye lens their role is to act as structural proteins, not enzymes (Petsko and Ringe, 2004). Another example is caffeine and adenosine. These two molecules have a similar shape, so similar that caffeine is able to bind to adenosine receptors (Gupta and Gupta, 1999). However, adenosine induces sleep and suppresses arousal while caffeine makes you more awake and less tired. Semantic similarity addresses the above exceptions, by comparing biomedical entities based on what they do and not on what they look like. This means that when looking for similar compounds to caffeine, other central nervous system stimulants, such as doxapram, will appear before adenosine that has the opposite effect.

1.2. What?

Digital representations of biomedical entities based on structure can normally be expressed using a simple syntax. For example, ASCII strings are used to represent the nucleotide sequences of genes, the amino acid sequences of proteins, and also the structure of compounds using SMILES. Semantics is however more complex since it may have different interpretations according to a given context. For example, the meaning of a biological role of a given gene may differ from a biological or medical perspective. For humans the easiest way to represent semantics is to use free text due to its flexibility to express any concept. For example, short text comments are usually valuable semantic descriptions to understand the meaning of a piece of information. However, for computers free text is not the most effective form of encoding semantics, making semantic similarity measurement between different text descriptions almost unfeasible.

In recent years, the biomedical community made a substantial effort in representing the semantics of biomedical entities by using common vocabularies, which vary from simple terminologies to highly complex semantic models. These vocabularies are instantiated by Knowledge Organization Systems (KOS) in the form of knowledge graphs, classification systems, thesauri, lexical databases,

gazetteers, and taxonomies, and ontologies(Barros et al., 2016). Perhaps the most well-known KOS is the Gene Ontology, which has been extensively used to annotate gene-products with terms describing their molecular functions, biological processes and cellular components, and the source of most semantic similarity studies in bioinformatics. This manuscript will denote a KOS used in a semantic similarity measure as its Semantic-Base (SB). Semantic similarity measures become feasible when a biomedical community accepts a SB as a standard to represent the semantics of the entities in their domain. Semantic similarity is therefore a measure of how close are the semantic representations of different biomedical entities in a given SB. This means that the semantic similarity between two entities depends on their SB representation and also on a similarity measure that calculates how close these representations are in the SB.

1.3. How?

We may think that given a SB, we should be able to find the optimal quantitative function to implement semantic similarity. However, the notion of semantic similarity is dependent on what are the objectives of the study. For example, a biologist and a physician may have two different expectations about the semantic similarity between the biological roles of two genes.

In bioinformatics, ontologies have been the standard SB for calculating semantic similarity. An ontology is a formal representation of a set of objects or entities (the "universe of discourse" or the domain) and the relationships between them (Gruber, 1993). Its structure is simultaneously human and computer readable, which allows the use of automatic approaches for inference.

Recently the term *knowledge graph* has been gaining prominence and is often used interchangeably with the term *ontology*. A knowledge graph is a collection of assertions or statements from which can be derived a graph-like structure. The exact definition of a knowledge graph is not consensual in the literature, but a common distinction is based on its comparatively larger size, the more heterogeneous source of data used for its construction, which in many cases is

not manually curated, and the inclusion of a more flexible formal schema. A knowledge graph can have an underlying ontology to provide a logical backbone for the assertions layer, but can also only include a unique layer integrating both the assertions and the logical schema (Fensel et al., 2020). Despite the differences, an ontology and a knowledge graph can both be represented and analysed as a graph structure.

The SB provides an unambiguous context on where semantic representations can be interpreted. A semantic representation is sometimes referred as a set of annotations, i.e. a link between the entity and an entry in the SB. Each entity can have multiple annotations. This means that the similarity measure may be applied to multiple entries in the SB. There are also different types of annotations. For example, an annotation can represent a finding with experimental evidence, or just a prediction from a computational method. Semantic similarity can explore the different types of annotations, for example to filter out annotations in which we have lower confidence.

It is possible to bridge the gap between entities expressed through natural language in text and the entries of the SB using automatic approaches. This type of approach, designated by Named Entity Linking, links entities to the entry (or entries) that best describe its semantic meaning. There are several challenges for Named Entity Linking, including:

Name variations: like abbreviations, acronyms, alternate spellings or synonyms. These types of variation are especially frequent in the biomedical domain. For example, the gene "CF transmembrane conductance regulator", whose mutations are responsible for cystic fibrosis, has the following associated symbols: CFTR, CF, MRP7, ABC35, ABCC7, CFTR/MRP, TNR-CFTR and dJ760C5.1; the terms "Kawasaki disease", "Kawasaki syndrome", "Mucocutaneous Lymph Node Syndrome" and "MLNS" refer to the same disease.

Ambiguity or polysemy: this means that a given word can have different meanings depending on the context where it appears. For example, the entity

"iris" can correspond to an animal eye's anatomical structure or to a plant genus.

Unlinkable: absence of entries in the target SB to describe a given entity. Biomedical knowledge is ever-evolving, which means that manually-curated structures, such as ontologies, are prone to be out-of-date. This hinders the semantic representation of entities described in free text.

Semantic relatedness between two entries in a given SB can be based on several types of relations described in the SB. Semantic similarity is a specific case of semantic relatedness since it is only based on hierarchical relations between SB entries, such as hyponymy/hyperonymy (subsumption) or synonym relations.

A similarity measure is a quantitative function between entries in the SB, which explores the relations between its entries to measure their closeness in meaning. An entry is normally connected to the other entries by different types of relations represented in the SB. The similarity measure calculates the degree of shared meaning between two entries, resulting in a numerical value. For example, this can be performed by identifying a path between both entries in the SB, and calculating the semantic gap encoded in that path. This means that a semantic similarity measure can be defined by the SB and the quantitative measure used, which will be formulated in the following sections.

2. Semantic Base

Definition 1 (Semantic-Base). A Semantic-Base is a tuple $SB =< E, R >$, such that E is the set of entries, and R is the set of relations between the entries. Each relation is pair of entities (e_1, e_2) with $e_1, e_2 \in E$.

When using biomedical ontologies, the entries represent the classes, terms or concepts. This definition ignores the type of relations that may be present in the ontology, since semantic similarity measures are normally restricted to subsumption relations (*is-a*). Nevertheless, a measure may use other type of

relation, or even use different types of relations. In the present manuscript, we focus on semantic similarity measures that are based on hierarchical relations. The interpretation of its results should take this into consideration. One of the reasons why subsumption relations are used is because they are transitive, i.e. if $(e_1, e_2) \in R$ and $(e_2, e_3) \in R$ then we can implicitly assume that (e_1, e_3) is also a valid relation. This enables us to define the ancestors and descendants of a given entry.

Definition 2 (Ancestors). Given a SB represented by the tuple $< E, R >$, and T the transitive closure of R on the set E (i.e. the smallest relation on E that contains R and is transitive), the Ancestors of a given entry $e \in E$ are defined as $Anc(e) = \{a : (e, a) \in T\}$

Definition 3 (Descendants). Given a SB represented by the tuple $< E, R >$, and T the transitive closure of R on the set E, the Descendants of a given entry $e \in E$ are defined as $Des(e) = \{d : (d, e) \in T\}$

There are multiple successful semantic similarity measures being used in bioinformatics. Many of them are inspired on the contrast model proposed by Tversky (1977), in the sense that they balance the importance of common features versus the exclusives. Thus, a semantic similarity measure can be categorized by how it defines the common features, and how it calculates the importance of each feature. The first step in most measures is to find the common ancestors in the SB to define the common features.

Definition 4 (Common Ancestors). Given a SB represented by the tuple $< E, R >$, the Common Ancestors of two entries $e_1, e_2 \in E$ is defined as $CA(e_1, e_2) = Anc(e_1) \cap Anc(e_2)$.

3. Information Content

This manuscript follows an information-theoretic perspective of semantic similarity (Sánchez and Batet, 2011). To calculate the importance of each entry

the measures identify the information content of each entry. Resnik (1995) defined the information content of an entry based on the notion of the entropy of the random variable X known in information theory (Ross, 2009). The intuition is to measure the surprise evoked by having an entry $e \in E$ in the semantic representation.

Definition 5 (Information Content). Given a SB represented by the tuple $< E, R >$, and a probability function $p : E \rightarrow]0, 1]$, the information content of an entry $e \in E$ is defined as $IC(e) = -log(p(e))$.

The probability function should be defined in a way that bottom-level entries in the SB become more informative than top-level entries, making the $IC(e)$ correlated with the specificity of e in the SB.

The definition of the probability function p can follow two different approaches:

Intrinsic: p is based only on the internal structure of the SB.

Extrinsic: p is based on the frequency of each entry in an external dataset.

Considering the graph represented in Figure1 as our SB, and assuming an intrinsic approach $p(e) = \frac{Desc(e)+1}{|E|}$, then we have all the bottom entries with p equal to $\frac{1}{8}$, $p(coinage) = \frac{4}{8}$, $p(precious) = \frac{5}{8}$, and $p(metal) = \frac{8}{8}$. Thus, we have $IC(metal) < IC(precious) < IC(coinage) < IC(platinum) \ldots < IC(copper)$. Note also that the addition of 1 to avoid having a zero probability for the entries without descendants.

Definition 6 (Frequency). Given a SB represented by the tuple $< E, R >$, and an external dataset D , and a predicate $refer(d, e)$ that is true when a data element $d \in D$ refers the entry $e \in E$, then the frequency of a given entry in that dataset is defined as

$$F_D(e) = |\{d : refer(e_1, d) \wedge d \in D \wedge e_1 \in Desc(e) \cup \{e\}\}|$$

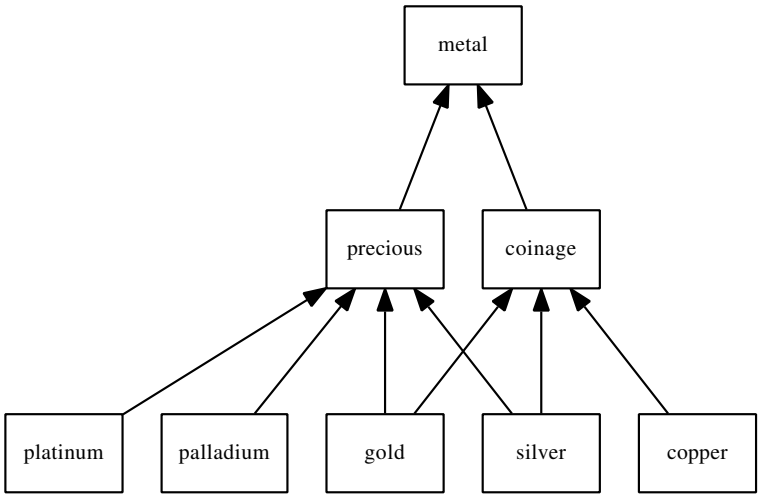

Figure 1: This graph represents an example of a classification of metals with multiple inheritance, since *gold* and *silver* are considered both precious and coinage metals.

Note that when using subsumption relations, i.e. an occurrence of an entry, it is also an implicit occurrence of all its ancestors.

Definition 7 (Extrinsic Probability). Given a SB represented by the tuple $< E, R >$, and a frequency measure F_D the extrinsic probability function of an entry $e \in E$ is defined as

$$p(e) = \frac{F_D(e) + 1}{max\{F_D(e_1) : e_1 \in E\} + 1}$$

.

Note that top-level entries have high frequency values due the occurrences of their descendants, so their IC is close to zero. Note again the addition of 1 this time in both parts of the fraction to avoid having a zero probability.

Considering again the graph represented in Figure1 as our SB, and assume an external dataset D containing exactly one occurrence of each entry, then we have all the bottom entries with F_D equal to $\frac{2}{9}$, $F_D(coinage) = \frac{5}{9}$, $F_D(precious) = \frac{6}{9}$, and $F_D(metal) = \frac{9}{9}$. Thus, we again have $IC(metal) <$

$IC(precious) < IC(coinage) < IC(platinum) \ldots < IC(copper)$. We will assume this IC instantiation for the remainder examples in this manuscript.

4. Shared Ancestors

Not all ancestors are relevant when calculating semantic similarity since some of them are already subsumed by others and do not represent any new information. So normally the measures select only the most informative ones.

Definition 8 (Most Informative Common Ancestors). Given a SB represented by the tuple $< E, R >$, and an IC measure, the Most Informative Common Ancestors of two entries $e_1, e_2 \in E$ is defined as

$$MICA(e_1, e_2) = \{a : a \in CA(e_1, e_2) \wedge IC(a) = max\{IC(a_1) : a \in CA(e_1, e_2)\}\}$$

Considering again the graph represented in Figure1 as our SB, and the extrinsic IC defined above, then we have $MICA(platinum, copper) = \{metal\}$, $MICA(silver, gold) = \{coinage\}$, and $MICA(platinum, gold) = \{precious\}$.

Sometimes the most informative common ancestors are not sufficient, since they may neglect multiple inheritance relations. Thus, instead of $MICA$, the measures can use the disjunctive common ancestors (Couto and Silva, 2011).

Definition 9 (Disjunctive Common Ancestors). Given a SB represented by the tuple $< E, R >$, and an IC measure, and a function $PD : E \times E \times E \to \mathbb{N}$, that calculates the difference between the number of paths from the two entries to one of their comon ancestors, the Disjunctive Common Ancestors of two entries $e_1, e_2 \in E$ is defined as

$$\begin{aligned} DCA(e_1, e_2) = \{a \ : \\ & a \in CA(e_1, e_2) \wedge \\ & \forall_{a_x \in CA(e_1, e_2)} PD(e_1, e_2, a) = PD(e_1, e_2, a_x) \\ & \Rightarrow IC(a) > IC(a_x)\} \end{aligned}$$

Considering again the graph represented in Figure1 as our SB, the extrinsic IC defined above, then we have $DCA(silver, gold) = \{coinage, precious\}$, and $DCA(platinum, gold) = \{precious, metal\}$.

5. Shared Information

The importance of common features is defined by the shared IC present in the common ancestors, normally its average.

Definition 10 (Shared Information Content). Given a SB represented by the tuple $< E, R >$, and an IC measure, the Shared Information Content of two entries $e_1, e_2 \in E$ is defined as $IC_{shared}(e_1, e_2) = \overline{\{IC(a) : a \in DCA(e_1, e_2)\}}$.

Note that DCA can be replaced by $MICA$, however since all ancestors in $MICA$ have the same IC value by definition only that IC value is used in practice.

Considering again the graph represented in Figure1 as our SB, the extrinsic IC defined above, then when using $MICA$ we have $IC_{shared}(platinum, gold) = -log(\frac{6}{9})$. If we use DCA then we have $IC_{shared}(platinum, gold) = (-log(\frac{6}{9}) - log(\frac{9}{9}))/2$.

More recently, Ferreira et al. (2013) proposed the usage of the disjointness axioms in semantic similarity by defining the disjoint shared information content. The idea is that if we know that two entries are disjoint, then we should decrease their amount of shared information.

Definition 11 (Disjoint Shared Information Content). Given a SB represented by the tuple $< E, R >$, a set of axioms A, and an IC_{shared} measure, the Disjoint Shared Information Content of two entries, $e_1, e_2 \in E$ is defined as $IC_{dshared}(e_1, e_2) = IC_{shared}(e_1, e_2) - k(e_1, e_2)$ with $k : E \times E \rightarrow \mathbb{N}$ satisfying the following conditions: i) $k(e_1, e_2) > 0$ if e_1 and e_2 are disjoint according to A; ii) $k(e_1, e_2) = 0$ if otherwise.

6. Similarity Measure

Definition 12 (Semantic Similarity Measure). Given a SB represented by the tuple $< E, R >$, a Semantic Similarity Measure is a quantitative function $SSM : E \times E \rightarrow \mathbb{R}$.

Note that a semantic similarity measure is not expected to be instantiated by the inverse of a metric or distance function, but the following conditions are normally satisfied:

non-negativity: $SSM(e_1, e_2) \geq 0$ with $e_1, e_2 \in E$;

symmetry: $SSM(e_1, e_2) = SSM(e_2, e_1)$ with $e_1, e_2 \in E$.

Many measures are also normalized, i.e. $SSM(e_1, e_2) \in [0..1]$ with $e_1, e_2 \in E$; and $SSM(e, e) = 1$ with $e \in E$.

The seminal work based on Resnik's measure (Resnik, 1995) was one of the first measures to be successfully applied to a biomedical ontology, namely the Gene Ontology. (Lord et al., 2003). The measure was defined as:

$$SSM_{resnik}(e_1, e_2) = IC_{shared}(e_1, e_2)$$

Another well-known measure, was defined by Lin et al. (1998) as:

$$SSM_{lin}(e_1, e_2) = \frac{2 \times IC_{shared}(e_1, e_2)}{IC(e_1) + IC(e_2)}$$

where the denominator represents the exclusive features.

Note that both measures are independent of using $MICA$ or DCA as the common features.

Considering again the graph represented in Figure1 as our SB, the extrinsic IC defined above, and $MICA$, then we have $SSM_{resnik}(platinum, gold) = -log(\frac{6}{9})$ and $SSM_{lin}(platinum, gold) = (2 \times -log(\frac{6}{9}))/(-log(\frac{2}{9}) - log(\frac{2}{9}))$.

Table 1 shows the SSM described in this manuscript.

Resnik	$SSM_{resnik}(e_1, e_2) = IC_{shared}(e_1, e_2)$
Lin	$SSM_{lin}(e_1, e_2) = \frac{2 \times IC_{shared}(e_1, e_2)}{IC(e_1) + IC(e_2)}$
Jaccard	$SSM_{jaccard}(AS(b_1), AS(b_2)) = \frac{\sum\{IC(e):e \in \{Anc(e_1):e_1 \in AS(b_1)\} \cap \{Anc(e_2):e_2 \in AS(b_2)\}\}}{\sum\{IC(e):e \in \{Anc(e_1):e_1 \in AS(b_1)\} \cup \{Anc(e_2):e_2 \in AS(b_2)\}\}}$

Table 1: Caption

7. Entity Similarity

Until now we only defined SSM in terms of entries, but a biomedical entity may not be directly represented in the SB, but instead linked to the SB through annotations. For example in the case of proteins, they are not represented as entries of the Gene Ontology but through annotations. In opposition, chemical compounds are represented as entries of the ontology Chemical Entities of Biological Interest (ChEBI).

Definition 13 (Annotation). Given a SB represented by the tuple $< E, R >$ and a set of biomedical entities B, a predicate $annotates(b, e)$ that is true when the entity $b \in B$ is annotated with the entry $e \in E$, then the annotation set of a biomedical entity (or concept) $b \in B$ is defined as

$$AS(b) = \{e : e \in E \wedge annotates(b, e)\}$$

This definition ignores the type of annotation, e.g. with experimental or computational evidence, since the similarity measure calculation is usually independent of this information. It is up to the user to decide which type of annotations to include. In the case of biomedical entities identified in a piece of text, a Named Entity Linking approach can annotate a given biomedical entity a with the entry e (or entries) in the SB that best represents its meaning.

To compare biomedical entities we need to extend the SSM definition so it applies to the two sets of entries of each entity, instead of a single entry for each entity. For readability we will use the same function name SSM, to represent different functions according to the input domain, i.e. two entries or two sets of entries.

There are multiple successful instantiations of entity semantic similarity measures and most of them use two aggregate functions (e.g. average, maximum) on the results from comparing each pair of entries annotated to each entry.

Definition 14 (Aggregate Measure). Given a SB represented by the tuple $< E, R >$, a set of biomedical entities B, two aggregate functions f and g, and two biomedical entities $b_1, b_2 \in B$ the Aggregate Similarity Measure is defined as

$$SSM_{aggregate}(AS(b_1), AS(b_2)) = f(\{g(\{SSM(e_1, e_2) : e_1 \in AS(b_1)\}) : e_2 \in AS(b_2)\})$$

Considering again the graph represented in Figure1 as our SB, f as the average function, g as the maximum function, two entities containing metals $B = \{\alpha, \beta\}$, and their respective annotation set $AS(\alpha) = \{platinum, palladium\}$ $AS(\beta) = \{copper, gold\}$, then we have

$$\begin{aligned} SSM_{aggregate}(\{platinum, palladium\}, \{copper, gold\}) = avg\{ \\ max\{SSM(platinum, copper), SSM(platinum, gold)\}, \\ max\{SSM(palladium, copper), SSM(palladium, gold)\}\} \end{aligned}$$

Another popular approach is to apply the Jaccard coefficient to all common entries vs. the exclusive ones.

Definition 15 (Jaccard Measure). Given a SB represented by the tuple $< E, R >$, a set of biomedical entities B, an annotation set AT, and two biomedical entities $b_1, b_2 \in B$ the similarity measure is defined as

$$SSM_{jaccard}(AS(b_1), AS(b_2)) = \frac{\sum\{IC(e) : e \in \{Anc(e_1) : e_1 \in AS(b_1)\} \cap \{Anc(e_2) : e_2 \in AS(b_2)\}\}}{\sum\{IC(e) : e \in \{Anc(e_1) : e_1 \in AS(b_1)\} \cup \{Anc(e_2) : e_2 \in AS(b_2)\}\}}$$

Considering the example above of α and β when using Jaccard we will have

$$SSM_{jaccard}(\{platinum, palladium\}, \{copper, gold\}) = \frac{IC(precious) + IC(metal)}{IC(coinage) + IC(precious) + IC(metal)}$$

7.1. Vector representations

Although not the scope of the present manuscript, which focuses on measures based on Information Content, a currently active research direction investigates the use of vector-based representations or embeddings for entity comparison, including the determination of similarity.

For instance, Zhong et al. (2019) proposed GO2Vec, a method that first converts the terms and annotations present in the Gene Ontology into vectors and then leverages these representations to calculate the similarity between Gene Ontology terms and to determine the functional similarity between proteins.

Littmann et al. (2021) proposed a method based on language models to annotate proteins with Gene Ontology terms. The authors used SeqVec and ProtBERT to encode proteins in fixed-length vectors. Then, the similarity between proteins was calculated using the Euclidian distance between the respective vectors. The Gene Ontology terms can thus be transferred from an annotated protein to a similar non-annotated protein.

Other vector-based approaches are developed for the task of predicting protein-protein interactions (PPIs), such as Zhong and Rajapakse (2020); Yang et al. (2020); Nasiri et al. (2021); Smaili et al. (2018). The approach proposed by Zhong and Rajapakse (2020) first builds Gene Ontology annotation graphs that include relations between terms and between terms and proteins, then learns vector representations for the nodes, and then the similarity between proteins is measured based on the distance between the respective vector representations, calculated through the modified Hausdorff distance. Yang et al. (2020) also proposed a method to learn graph representations that are used for predicting PPIs. Proteins are encoded based on their raw sequence and graph information and then the resulting embeddings are used to train a feed-forward neural network that predicts an interaction between a given pair of proteins. Nasiri et al. (2021) proposed a method based on the DeepWalk algorithm, which uses random walks to derive vector representations for the nodes of a graph, which are used to predict PPIs. Smaili et al. (2018) proposed Onto2Vec, which generates protein vector-based representations. First, representations for proteins and

Gene Ontology terms are built according to the ontology structure and UniProt protein annotations. Proteins are represented as the sum of the vectors for each of the terms present in annotations and the distance between a pair of proteins is given by the cosine similarity between the respective vectors.

8. Future Directions

This manuscript is focused on defining semantic similarity using a single KOS, however a large amount of biomedical resources use multiple KOS describing a single domain from different perspectives or even distinct domains. Calculating semantic similarity using multiple KOS as SB is a complex problem, and only a few works have addressed it (Solé-Ribalta et al., 2014). Thus, a future formulation of multiple-domain semantic similarity is much required.

Another issue is the incompleteness of KOS. They normally represent work in progress, being updated as our knowledge of the domain becomes more sound and comprehensive. Keeping a KOS up-to-date is also a daunting task in terms of human effort, especially in large KOS, so we should always expect to have a delay until new knowledge is incorporated. This means that the common features identified in a KOS may be incomplete, and the exclusives features may not even be exclusive in the future. If a biomedical entity is not annotated with a specific feature, that does not mean that the entity does not have that feature, it only means that we do not know if it has or not. Thus, a future formulation of semantic similarity that takes into account the incompleteness of KOS is also much required. A type of approach to deal with this problem of KOS incompleteness attempts to associate biomedical entities with features or entries in the KOS that only partially describe their meaning (Ruas and Couto, 2022). The assumption is that it is preferable to have an entity partially annotated instead of simply discarding the entity and losing all of its respective semantic information.

9. Closing Remarks

This manuscript presented a definition of semantic similarity following an information-theoretic perspective that covers a large number of the measures currently being used in bioinformatics. It defined the amount of information content two entries share in a SB, and how it can be extended to compare biomedical entities represented outside the SB but linked through a set of annotations.

The manuscript aims at providing a generic and inclusive formulation that can be helpful to understand the fundamentals of semantic similarity and at the same time be used as a guideline to distinguish between different approaches. The formulation did not aim at providing a one size fits all definition, i.e. trying to represent all measures being proposed.

The manuscript presented well-known measures in bioinformatics, Resnik, Lin and Jaccard coefficient, according to the proposed definitions. It also presented their results when applied to simple example of a classification of metals, which is used along the text to clarify the definitions being presented. Finally, a software repository [1] is available to test and learn more on how semantic similarity works in practice.

10. Acknowledgement

This work has been supported by FCT through Deep Semantic Tagger (DeST) Project under Grant PTDC/CCIBIO/28685/2017 (http://dest.rd.ciencias.ulisboa.pt/), in part by LASIGE Research Unit under Grants UID/CEC/00408/2013, UIDB/00408/2020, and UIDP/00408/2020, and in part by FCT and FSE through Ph.D. Scholarships under Grants PD/BD/106083/2015 and 2020.05393.BD.

[1] `https://github.com/lasigeBioTM/DiShIn/`

Nomenclature

CA	Common Ancestors
CE	Common Entries
DCA	Disjunctive Common Ancestors
F_D	Frequency in a external dataset D
IC	Information Content
$IC_{dshared}$	Disjoint Shared Information Content
IC_{shared}	Shared Information Content
$MICA$	Most Informative Common Ancestors
SSM	Semantic Similarity Measure
KOS	Knowledge Organization Systems
SB	Semantic-Base

References

Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 25, 3389–3402.

Barros, M., Couto, F.M., et al., 2016. Knowledge representation and management: a linked data perspective. IMIA Yearbook , 178–183.

Batet, M., Harispe, S., Ranwez, S., Sánchez, D., Ranwez, V., 2014. An information theoretic approach to improve semantic similarity assessments across multiple ontologies. Information Sciences 283, 197–210.

Couto, F., Silva, M., 2011. Disjunctive shared information between ontology concepts: application to Gene Ontology. Journal of Biomedical Semantics 2, 5.

Fensel, D., Şimşek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., Toma, I., Umbrich, J., Wahler, A., 2020. Introduction: What is a knowledge graph?, in: Knowledge Graphs. Springer International Publishing, pp. 1–10.

Ferreira, J.D., Hastings, J., Couto, F.M., 2013. Exploiting disjointness axioms to improve semantic similarity measures. Bioinformatics 29, 2781–2787.

Gruber, T.R., 1993. A translation approach to portable ontology specifications. Knowledge Acquisition 5, 199–220. URL: http://tomgruber.org/writing/ontolingua-kaj-1993.pdf.

Gupta, B.S., Gupta, U., 1999. Caffeine and Behavior: Current Views & Research Trends: Current Views and Research Trends. CRC Press.

Lin, D., et al., 1998. An information-theoretic definition of similarity., in: ICML, Citeseer. pp. 296–304.

Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T., Rost, B., 2021. Embeddings from deep learning transfer go annotations beyond homology. Scientific Reports 11. doi:10.1038/s41598-020-80786-0.

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

Nasiri, E., Berahmand, K., Rostami, M., Dabiri, M., 2021. A novel link prediction algorithm for protein-protein interaction networks by attributed graph embedding. Computers in Biology and Medicine 137. doi:10.1016/j.compbiomed.2021.104772.

Petsko, G.A., Ringe, D., 2004. Protein structure and function, New Science Press.

Resnik, P., 1995. Using information content to evaluate semantic similarity in a taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence .

Ross, S., 2009. A First Course in Probability 8th Edition. Pearson.

Ruas, P., Couto, F.M., 2022. Nilinker: Attention-based approach to nil entity linking. Journal of Biomedical Informatics 132, 104137. URL: https://www.sciencedirect.com/science/article/pii/S1532046422001526, doi:https://doi.org/10.1016/j.jbi.2022.104137.

Sánchez, D., Batet, M., 2011. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. Journal of biomedical informatics 44, 749–759.

Smaili, F.Z., Gao, X., Hoehndorf, R., 2018. Onto2vec: Joint vector-based representation of biological entities and their ontology-based annotations. Bioinformatics 34, i52–i60. doi:10.1093/bioinformatics/bty259.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. Journal of molecular biology 147, 195–197.

Solé-Ribalta, A., Sánchez, D., Batet, M., Serratosa, F., 2014. Towards the estimation of feature-based semantic similarity using multiple ontologies. Knowledge-Based Systems 55, 101–113.

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

Willett, P., 2011. Similarity searching using 2d structural fingerprints. Chemoinformatics and computational chemical biology , 133–158.

Yang, F., Fan, K., Song, D., Lin, H., 2020. Graph-based prediction of protein-protein interactions with attributed signed graph embedding. BMC Bioinformatics 21, 1–16. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03646-8, doi:10.1186/S12859-020-03646-8/TABLES/4.

Zhong, X., Kaalia, R., Rajapakse, J.C., 2019. Go2vec: Transforming go terms and proteins to vector representations via graph embeddings. BMC Genomics 20. doi:10.1186/s12864-019-6272-2.

Zhong, X., Rajapakse, J.C., 2020. Graph embeddings on gene ontology annotations for protein–protein interaction prediction. BMC Bioinformatics 21. doi:10.1186/s12859-020-03816-8.

11. Further Reading

- Batet, M., Sánchez, D., 2015. A review on semantic similarity, Encyclopedia of Information Science and Technology, Third Edition. IGI Global, pp. 7575-7583.
- Couto, F.M., Pinto, H.S., 2013. The next generation of similarity measures that fully explore the semantics in biomedical ontologies. Journal of bioinformatics and computational biology 11, 1371001.
- Harispe, S., Ranwez, S., Janaqi, S., Montmain, J., 2015. Semantic similarity from natural language and ontology analysis. Synthesis Lectures on Human Language Technologies 8, 1-254.
- Pedersen, T., Pakhomov, S.V., Patwardhan, S., Chute, C.G., 2007. Measures of semantic similarity and relatedness in the biomedical domain. Journal of biomedical informatics 40, 288-299.
- Pesquita, C., Faria, D., Falcao, A., Lord, P., Couto, F., 2009. Semantic similarity in biomedical ontologies. PLoS computational biology 5, e1000443

12. Author Biography and Photograph

Francisco M. Couto is currently an associate professor with habilitation at Universidade de Lisboa (Faculty of Sciences) and a researcher at LASIGE. He

graduated (2000) and has a master (2001) in Informatics and Computer Engineering from the IST. He concluded his doctorate (2006) in Informatics, specialization Bioinformatics, from the Universidade de Lisboa. He was an invited researcher at EBI, AFMB-CNRS, BioAlma during his doctoral studies. His main research contributions cover several key aspects of bioinformatics and knowledge management, namely in proposing and developing: various text mining solutions that explore the semantics encoded in ontologies; semantic similarity measures and tools using biomedical ontologies; and ontology and linked data matching systems. Until August 2022, he published 2 books; was co-author of 10 chapters, 62 journal papers (47 Q1 Scimago), and 32 conference papers (10 core A and A*); and was the supervisor of 10 PhD theses and of 51 master theses. He received the Young Engineer Innovation Prize 2004 from the Portuguese Engineers Guild, and an honorable mention in 2017 and the prize in 2018 of the ULisboa/Caixa Geral de Depósitos (CGD) Scientific Prizes.

Andre Lamurias is a post-doctoral researcher in the Department of Computer Science at Aalborg University, since 2021. In 2020, he worked as a research scientist at Priberam Labs, and he has previously been a researcher at LASIGE, Faculty of Sciences, University of Lisboa. He completed his PhD in 2019 at University of Lisbon, where he developed text-mining approaches for disease network discovery and systems biology. Prior to this, he obtained his Master's degree in Bioinformatics and Computational Biology from the same institution. His research focus is on information extraction applied to biomedical data, such as scientific papers, electronic health records, and genomics.

Pedro Ruas is a researcher at LASIGE and is currently enrolled in the Informatics PhD programme at Universidade de Lisboa. His PhD thesis focuses on improving text mining approaches for biomedical data, in particular approaches for recognizing biomedical entities in scientific literature and clinical text and linking them to biomedical data repositories. His research focuses on information extraction from biomedical text.