We are IntechOpen,
the world's leading publisher of
Open Access books
Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Semantic Similarity in Cheminformatics

João D. Ferreira and Francisco M. Couto

Abstract

Similarity in chemistry has been applied to a variety of problems: to predict biochemical properties of molecules, to disambiguate chemical compound references in natural language, to understand the evolution of metabolic pathways, to predict drug-drug interactions, to predict therapeutic substitution of antibiotics, to estimate whether a compound is harmful, etc. While measures of similarity have been created that make use of the structural properties of the molecules, some ontologies (the Chemical Entities of Biological Interest (ChEBI) being one of the most relevant) capture chemistry knowledge in machine-readable formats and can be used to improve our notions of molecular similarity. Ontologies in the biomedical domain have been extensively used to compare entities of biological interest, a technique known as ontology-based semantic similarity. This has been applied to various biologically relevant entities, such as genes, proteins, diseases, and anatomical structures, as well as in the chemical domain. This chapter introduces the fundamental concepts of ontology-based semantic similarity, its application in cheminformatics, its relevance in previous studies, and future potential. It also discusses the existing challenges in this area, tracing a parallel with other domains, particularly genomics, where this technique has been used more often and for longer.

Keywords: semantic similarity, ontologies, ChEBI, prediction of molecule properties, databases

1. Introduction

With the unprecedented amount of data being generated today, we must start (and in some cases have already started) to rely on automatic systems to process, analyse, and understand all the scientific information that we produce. For some examples in chemistry, consider the number of drugs represented in DrugBank, which grew from 3909 in 2006 to 9688 [1], about 13% each year; the number of metabolites in the Human Metabolite Database grew from 2180 in 2007 to 114,100 in 2017 [2], approximately 39% per year (although at some point this database imported a large number of metabolites at once, artificially increasing this statistic); ChemSpider had 25 million compounds in 2010 [3] and now has 63 million (10% a year); and PubChem grew from 19 million compound structures in 2008 [4] to 96.5 million in August 2018 [5] (16% a year). These numbers usually grow exponentially [6], reflecting the fact that the amount of knowledge the scientific community produces is proportional to the amount of knowledge we discover.

With such high volumes of data, it is imperative that we categorise this information in ways that assist us in the tasks of consuming that information, specifically through categorisation schemas that abstract away the less useful details of reality and increase the manageability of this information. As we will see later in this chapter, ontologies can perform that goal: they are computational artefacts (files, tables in a database, etc.) whose goal is to encode real-world knowledge in machine-readable logical axioms that can be used by automatic systems to manipulate the knowledge inferred and potentially derivable from the data we have.

Furthermore, like most other scientific knowledge, chemistry ideas and notions are inferred from comparing entities and finding their similarities and differences. For instance, compound similarity has been used to (i) develop pharmacophores [7, 8], (ii) estimate whether a compound is harmful without in vivo experimentation [9], (iii) understand the evolution of metabolic pathways [10], (iv) predict adverse side effects of drugs [11], and (v) perform pharmacological profiling of compounds in drug design [12].

As we explore in this chapter, ontologies provide one way to measure similarity of chemistry entities (compounds, substances, mixtures, reactions, etc.), a technique known as ontology-based semantic similarity (shortened to semantic similarity in this chapter). This idea is already widely used in genomics and proteomics, but its full potential still needs to be brought over to other domains. While some research has successfully used this methodology in the cheminformatics domain (which we discuss below), there is still space for improvement and further methodological development.

In this chapter, we explore the ideas and concepts behind semantic similarity and chemistry ontologies, explore some past applications that use those concepts to further our knowledge of the chemical domain, and expose some limitations and challenges that this technique still needs to overcome for its whole potential to be released.

2. Measures of similarity in chemistry

Similarity, in its nature, is a notion that produces a number. In that sense, it is mathematical. However, chemical knowledge cannot be trivially reduced to mathematical form. For example, given two molecules, how should one compare them and assign a number to represent their similarity? And even if specific cases can be handled by humans, we still need an automatic way to perform comparison. However, to a certain extent, computers can only manipulate objects that can be represented mathematically (e.g., vectors) or as strings of characters (e.g., gene sequences, SMILES). But the algorithms that are used with these structures are context-free: they usually transform the structures without any knowledge of what they represent.

Many mechanisms exist to deal with this issue. For example, graph similarity can be used to find common substructures in two molecules as a basis for similarity calculations (see, e.g., [13, 14]), but these methods tend to be slow and computationally expensive. There is also the possibility to reduce a molecular structure into a *fingerprint*, which is a binary vector where each position represents the presence (with a 1) or absence (with a 0) of a certain feature in the structure. For example, the presence of a carboxyl group could be indicated with a 1 in some position of the vector. Similarity can then be computed by measuring the overlap in those vectors [15, 16].

These methods provide a high similarity value when the structures of the two molecules are high. Under the quantitative structure-activity relationship (QSAR)
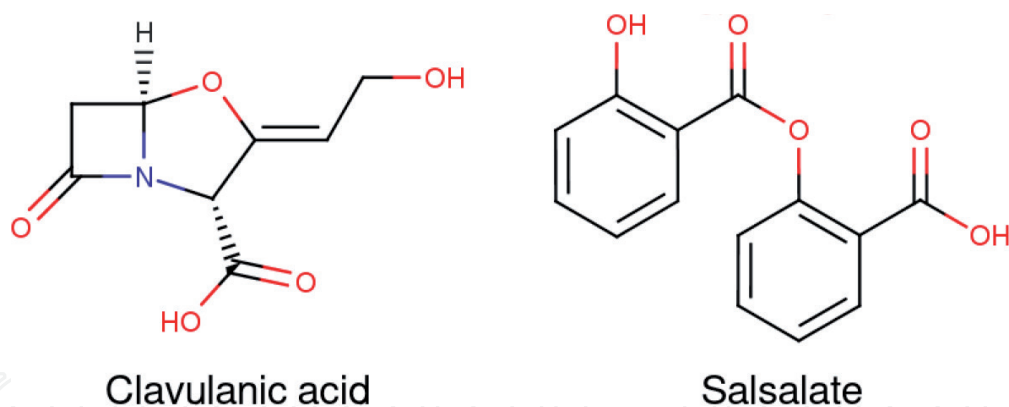
Figure 1.
Chemical structure of two semantically related compounds. The two molecular structures in the figure are quite different structures, and yet both present the same biological activity, namely, they inhibit β-lactamase enzymes.

premise, this means that, in general, two molecules with a high similarity score (as defined by these methods) tend to have similar biological role, similar chemical properties (such as melting point, optical parameters, and mass spectroscopy spectra), similar safety warnings, similar appearance, etc. But this is not always true. For instance, while L-amino acids are used to synthesise proteins, D-amino acids are much less frequent in nature, and their role is quite different [17]. From a biological point of view, they are distinct; however, to capture their structural differences, one needs to use three-dimensional methods, and even with that consideration, the structural similarity will be high, because both molecules have the same atoms and bonds. Another possibility includes simulation of docking with target proteins, but these methods are quite expensive computationally. Furthermore, not only can similar molecules perform different biological roles, different molecules can perform similar roles. For example, both clavulanic acid and salsalate are β-lactamase inhibitors, despite their different structures (see **Figure 1**).

Another way to measure similarity is by means of the semantics attached to the chemical compounds. Here, we use the term *semantics* to mean the knowledge that exists about a compound. This includes not only the structure of the molecule itself (e.g., the atomic connectivity, the number of oxygen atoms, the presence of triple bonds) but also other types of contextual knowledge, such as its chemical role (e.g., whether it is an electron donor, a solvent, or an explosive), biological role (e.g., whether it is a poison, a cofactor, or a vitamin), its applications (as a drug, fertiliser, fuel, etc.), its relationship to other molecules (such as being enantiomers, parent hydrides, etc.), and so on.

The difficulty with this is that knowledge is not directly machine-readable. Indeed, established facts have been traditionally published in plain text, which enables some humans to understand them; however, natural language processing techniques are not yet fully capable of converting scientific text into actionable formats (e.g., formats that allow automatic reasoning). Therefore, to enable the application of computerised processing power to knowledge manipulation, it is essential to find ways to represent knowledge in machine-readable formats.

3. Ontologies

Ontologies are the solution to this problem. An ontology is a representation of concepts from a domain of knowledge and the relationship between them and is usually visualised as a directed acyclic graph (DAG), where nodes are the concepts, edges are the relationships, and there are no cycles in the graph. See, for example,

Figure 2, a toy exampled based on a real-world ontology that encodes the fact that "acetate" is the conjugate base of "acetic acid" and that "acetic acid" is the conjugate acid of "acetate" and then organises these concepts in a hierarchy that contains concepts like "ion", "molecule", "organic acid", and "organic molecular entity", and ends up in the most generic "molecular entity" concept.

There are many ontologies whose purpose is to encode the chemical knowledge, but one of the most comprehensive and used is the ontology for Chemical Entities of Biological Interest (ChEBI) [18]. This ontology represents in a machine-readable format about 114 thousand concepts, including not only the chemical compounds but also their biological and chemical roles. Other ontologies that encode this or related domains include (*i*) Interlinking Ontology for Biological Concepts, (*ii*) Current Procedural Terminology, (*iii*) SNOMED CT, (*iv*) Chemical Information Ontology, and (*v*) Chemical Methods Ontology.

It is important to notice that, even though the notion of ontologies usually requires some logic concepts (such as axioms, predicates, etc.), some classification hierarchies are also sometimes named "ontologies". MeSH, the system used
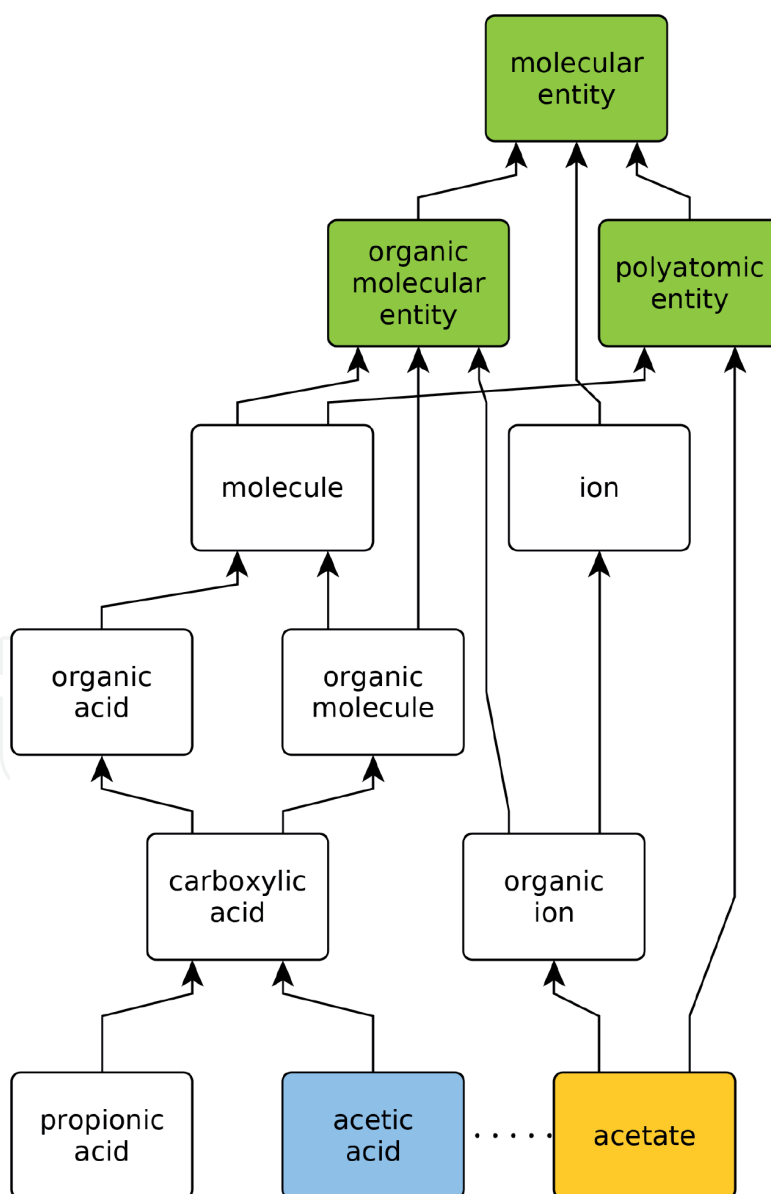
Figure 2.
A toy example of an ontology for chemical compounds, based on ChEBI. The ontology shows "is-a" relationships with solid lines, and a relationship between acid/base conjugates with a dotted line. The green shaded concepts are those that subsume both the yellow and the blue ones.

by PubMed to classify publications, is a hierarchy of concepts that possesses many of the same properties that ontologies do, namely, that it can be represented as a directed acyclic graph. However, one of the differences is that the relationship between two concepts does not always carry the same meaning. For example, "Head" is categorised under "Body Regions", and "Ear" is categorised under "Head", but while heads *are* body regions, ears *are not* heads; they are instead *parts* of the head. This illustrates the informality of MeSH: only one relationship type exists and it is used to express different notions. Another system in this category is the Anatomical Therapeutic Chemical (ATC) Classification System.

BioPortal [19], a repository of ontologies for the biomedical domain, contains a collection of 948 ontologies at the time of this writing. As an illustration of its magnitude, consider that 19 ontologies represent the concept "lidocaine". This reflects the effort being currently spent to represent human knowledge in machine-readable ontologies. In fact, while ontologies such as ChEBI are massive, BioPortal allows their users to submit new ontologies, even if small, focussed on a specific domain, and created with a specific application in mind other than pure knowledge representation (e.g., there is an ontology specific for cardiovascular drug adverse events, with 3 thousand concepts).

Other efforts have been set into place to aggregate ontologies in a single source of knowledge. For example, the Open Biological and Biomedical Ontology (OBO) Foundry [20] developed the OBO file format to represent ontologies and currently defines principles of quality for ontologies in biomedical domain that prescribe good practices for ontology development, such as being open, being reusable, being developed with collaboration in mind, containing both textual and logical definitions (for the benefit of both humans and machines), etc. They contain more than 200 ontologies as of this writing, 10 of which fully adhere to those principles (ChEBI being one of them). The OBO Foundry is tightly coupled with Ontobee [21], a web service that uses the principles of linked data to serve as a linked data server specifically targeted for ontologies and their concepts.

4. Semantic similarity

Using a formal representation of knowledge, computers are given the ability to manipulate concepts that are difficult to represent, in a way that preserves their "semantics". Ontologies provide the appropriate support for automatic manipulation of information. In this context, semantic similarity is a technique that assigns a numeric value to a pair of concepts based on the similarity of their meaning, extracted from the ontology.

For example, there is no directly obvious way to compare two roles. However, considering the illustration in **Figure 3**, it is possible to intuitively understand that, because both "hallucinogen" and "antifungal drug" are examples of "drugs", they are more similar than "hallucinogen" and "fossil fuel". This measure makes use of the meaning of the concepts, implicitly represented in the ontologies through the relations between the concepts. Ontologies function as a proxy for that meaning and enable its manipulation and ultimately comparison.

Several formulas and ideas have been proposed, implemented and tested in the past to compute semantic similarity. A full exposition on such measures and algorithms is beyond the scope of this chapter. The reader is encouraged to expand on this topic by reading works such as [22–25]. As such, the following is an abridged version of how ontology-based semantic similarity has been computed. In this discussion, consider the ontology in **Figure 3**.
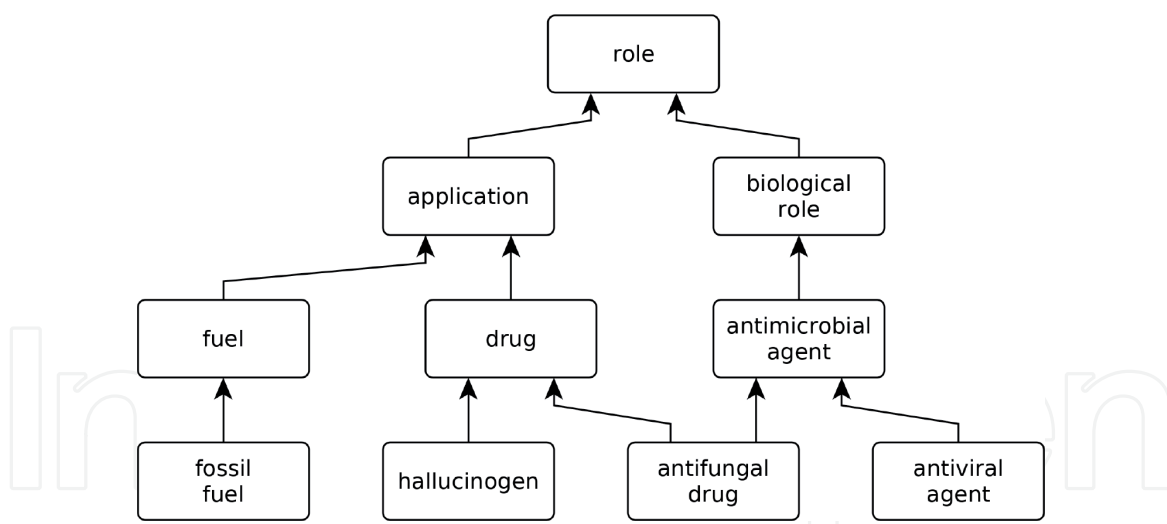
Figure 3.
A second toy example of an ontology representing chemical roles, also based on ChEBI.

Measures of similarity based on ontologies can roughly be divided into edge-based and node-based. An example of an edge-based measure is counting how many relations must be traversed to connect the two concepts being compared. Rada et al. [26] define distance as the number of edges in the smallest path between two nodes composed only of "is-a" relations. In this case, the distance between "hallucinogen" and "antimicrobial agent" would be three ("hallucinogen"→"drug"→"antifungal drug"→"antimicrobial agent"). While this type of approach is intuitive, it assumes that all nodes and all edges are equally important in terms of their semantics (e.g., that all edges weigh the same), which is generally not true in ontologies in life sciences. For instance, the "is-a" relation between "hallucinogen" and "drug" does not necessarily convey the same *amount of information* as the inverse "is-a" relation between "drug" and "antifungal drug".

One way to solve this is to introduce node-based methods, a technique that weighs nodes based on their *information content* (IC) [27]. The IC of a node is a numeric value based that reflects how informative its presence is and is calculated based on its frequency of use, since concepts that appear more frequently are generally less informative. The first formula proposed to measure IC was

$$IC(c) = -\log f(c) \qquad (1)$$

where $f(c)$ is the relative frequency with which the concept c and all its descendants appear in a corpus (in the example ontology, we can use the fraction of chemical concepts in ChEBI annotated as performing each of those roles). The intuition behind this idea is the following: consider a document (e.g., a scientific article) that uses the sentence "rodents have fur". The term "rodent" is used in such a way that every other concept that can be categorised under it also possesses the declared property. In fact, whenever a concept is used (in text, in logical axioms, etc.), it must be interpreted as including the set of all concepts recursively categorised under it.

The similarity between two concepts can be computed as the IC of the *most informative common ancestor* (usually abbreviated as MICA) between them

$$\text{sim}_{\text{Resnik}}(c_1, c_2) = IC(MICA(c_1, c_2)). \qquad (2)$$

This idea has been iterated upon with some additions and adaptations.

- The IC measure can be normalised so that it ranges from 0.0 to 1.0 (originally, the measure is unbounded above);

- The IC measure has been computed from multiple sources, such as (*i*) text corpora (as in the original), (*ii*) frequency of usage of the ontology concepts in external sources [28], or (*iii*) the ontology itself, where frequency can be computed based on the number of descendants (direct or indirect) of a concept [29], the number of leaf descendants of a concept [30], or other topological properties of the graph representation of the ontology [31].

- The semantic similarity measure itself can be normalised. Notice that the original measure gives the same similarity to the pair "application"/"biological role" (both generic concepts) and "fossil fuel"/"antiviral agent", which goes against the intuition that the first pair should be more similar. Lin [32] uses this idea to define

$$\text{sim}_{\text{Lin}}(c_1, c_2) = \frac{2 \cdot \text{IC}(\text{MICA}(c_1, c_2))}{\text{IC}(c_1) + \text{IC}(c_2)} \; ; \tag{3}$$

- The notion of shared information content (originally measured as the information content of the MICA of the two concepts) has also been tuned to take into account the fact that concepts can have multiple parents [33], which is necessary in many life science fields since it is in the nature of biomedical ontologies that some concepts are categorised under multiple parents, (see https://github.com/lasige-BioTM/DiShIn for an example of software that computes this type of measure) or the fact that ontologies have disjointness axioms that encode the fact that two concepts cannot share any descendants [34], also important because life science ontologies, and especially chemistry ones, make use of those types of axioms [35].

- The way to measure shared information content has also been completely re-implemented to use not the IC of the most informative common ancestor but a metric based on the set of all ancestors of the concepts [36].

These measures are able to compare one concept with another. It is also possible to compare sets of concepts. For this, one takes the matrix of pairwise similarities between concepts in the first set and concepts in the second set and mathematically manipulates it to produce a single number, taking, for example, the average, the maximum, or the "best match average", an approach that averages the highest values in each row and column [22]. There are other approaches that convert a set of concepts into the set of all their ancestors and take the intersection of those sets as a measure of similarity (two examples are simUI and simGIC [22]).

Finally, there is a difference in measuring the *similarity* or the *relatedness* between concepts. Similarity is a term that is generally applied to the notion that two concepts are "alike" and is usually computed based on "is-a" hierarchies; relatedness is more general: two related concepts can be related based on their categorisation on a hierarchy or on any number of other non-hierarchical relations. This distinction is important in chemistry, and ChEBI in particular, since many chemistry concepts are related via relations such as "has-role", "has-part", "is-enantiomer-of", etc.

Notice that when nothing is known about a chemical compound other than its structure, semantic methods can still be used, because one of the ways ontologies

(especially ChEBI) classify molecules is based on their structure. For example, ChEBI has a concept "carboxylic acid" which is an ancestor of all molecules that have one or more carboxylic acid groups (e.g., benzoic acid, all amino acids, all penicillins, etc.). This, however, is not conceptually different from measuring structural similarity, and such a setting would lack the enrichment provided by other types of knowledge (e.g., the knowledge of the chemical and biological roles of the molecule).

5. Applications

Since 2003, when Lord et al. [28] introduced the idea of ontology-based semantic similarity in the gene ontology (GO), several results have been achieved using this technique, proving beyond doubt that it is sound and useful and has real-life applications. In genomics and proteomics, semantic similarity based on GO has been used to (i) cluster proteins [37], (ii) find protein-protein interactions [38], (iii) interpret microarray data [39], (iv) predict protein functions [40], (v) prioritise candidate disease genes [41], etc. Other uses outside GO include predicting disease-related phenotypes [42] and predicting clinical diagnosis from a set of phenotype abnormalities [43].

The uses in chemistry-related areas have been scarce, but nonetheless existing and with real-world applications. We collected three research studies of semantic similarity in cheminformatics, which show its use in this area.

5.1 Predict biochemical properties of molecules

In 2010, ontology-based semantic similarity was applied to ChEBI [44] using a methodology named Chym. Chym shows for the first time that semantic similarity is useful in biomedical chemistry, by applying these ideas to predict whether a molecule (i) is capable of crossing the blood brain barrier, (ii) is a substrate of the P-glycoprotein, and (iii) binds to an oestrogen receptor. These properties are at least partially intrinsically related to the three-dimensional structure of the molecules and also of the proteins that perform the biochemical role in the organism. However, the work shows that structural similarity alone can be improved if it is coupled with semantic similarity.

Chym used daylight fingerprints for structural similarity and simUI and simGIC for semantic similarity, using ChEBI as the ontology. For all the three properties mentioned above, Chym was able to clearly outperform what were then the state-of-the-art prediction techniques for those properties.

Notice that this means that the two ideas presented here, structural similarity and semantic similarity, are not orthogonal and can be applied simultaneously with good results. This is not surprising, as ontologies can complement the knowledge that can be inferred form the structure alone, without needing to resort to wet-lab experiments.

5.2 Disambiguate chemical compound references in natural language

As the amount of textual chemistry information increases, particularly in the form of drug leaflets, articles, patents, and other types of communications, the need to develop mechanisms to automatically read these texts and extract tractable information from them increases as well. In this context, named entity recognition is a text mining task whose goal is to identify the entities mentioned in text.

There have been many attempts to create such systems in the chemical domain (see, e.g., the review [45]). In one of those attempts [46], semantic similarity has been used to improve the precision of existing methodologies by successfully identifying some false positives and removing them from the final result set. The idea of that work is that, within a scope of text (e.g., a sentence or a paragraph), chemical entities mentioned in that scope share some degree of semantic similarity that is higher than average. When entity recognition algorithms offer more than one possible ChEBI identifier for an excerpt of text, similarity with other ChEBI concepts can be used to disambiguate which is the correct entity.

5.3 Drug repurposing

Drug repurposing is the process by which drug that have therapeutic application are computationally tested to find other therapeutic applications. This reduces costs and improves the drug development pipeline and as such is important for the pharmaceutical industry.

The work presented in [47] couples similarity between the three-dimensional molecular structure with semantic similarity between the drug targets to find new indications for known drugs. The ontology used here is not a chemistry-specific one, but GO.

The main methodology of this work was:

1. Select a drug d and a potential target protein p.

2. Find drugs similar to this one (up to a threshold) with a structural similarity measure. Store these structural similarity values in a vector X str = $(d\,1, d\,2, ..., d\,m)$.

3. For each similar drug d_i, find its interacting proteins, compare them with p using GO-based semantic similarity, and sum the results. Call this value p_i. We have now a vector X sem = $(p\,1, p\,2, ..., p\,m)$.

4. The drug-protein association is assigned a score that depends on the correlation between the vectors X_{str} and X_{sem}. For a set of N proteins, each drug was then assigned a vector of N drug-protein association values, called the drug's "expression profile".

5. The drug-drug similarity measure was computed based on the correlation between the "expression profiles" of the two drugs.

The similarity between drugs was then used to construct a network of similarities, where clusters of highly connected drugs were indicative of potential drug repurposing.

A related work [48] also uses semantic similarity to predict drug-protein interaction. In this work, probabilistic similarity logic is used to construct models that are based on a notion of "similarity triads": triples of the form "drug-target-drug" with similar drugs or "target-drug-target" with similar targets. The whole work was based on the assumption that similar targets tend to interact with the same drug and similar drugs tend to interact with the same target. Here, several protein similarity methods (including semantic similarity based on GO) and drug similarity method (including semantic similarity based on ATC) were used to build a probabilistic model that predicts whether drugs and proteins interact.

6. Challenges and future work

Semantic similarity in cheminformatics has been slow to keep with the pace of equivalent research in other life science fields, such as genomics and proteomics. We posit that this is in some ways related to general and specific challenges associated with the application of this methodology in chemistry.

First, the state of ontology development and the more general knowledge representation area is very active, specifically in the biomedical fields. This means that many people have the motivation to develop their own ontology, with specific views of the reality embedded in it. However, as many people create their own knowledge representation artefacts, many different ontologies start to appear that overlap in domain, which means that it is not always obvious which ontology (or ontologies) to choose for a specific goal. Furthermore, these ontologies are not easy to reconcile, because they encode different and disjoint points of view. While efforts have been made to attenuate this problem, such as ontology matching (the process by which ontologies of the same domain are automatically merged into a single ontology) and the establishment of community standards (in chemistry, e.g., it is standard practice to reuse ChEBI concepts rather than create new concepts in new ontologies), the problem still persists.

Second, metrics of semantic similarity have been mostly developed and tested in the fields of natural language processing and genomics/proteomics. While these seem to have good enough results when used with ChEBI, we still do not know if they are the most adequate measures in this domain. Ferreira et al. [34] developed and validated a measure on the chemical domain, but more work needs to be done in this area. In particular, what role should the non-hierarchical relationship types ("is-enantiomer-of", "is-conjugate-acid-of", etc.) have in semantic similarity?

The third challenge is one of similarity profiles. It is not always obvious which details or properties of a molecule should be used for comparing. Should a pair of chemical compounds that differ only in the presence of an oxygen atom (e.g., methane vs. methanol) be more similar than a pair of molecules that differ only in charge (e.g., NO_2 vs. NO_2^-) or only in their three-dimensional conformation (e.g., L-serine vs. D-serine)? This problem must be solved based on context: determining what the similarity measure will be used for and then deciding which features are important. This includes deciding, for example, which relationship types should be taken into account, how to weight them, etc. Maggiora et al. [49] touch on the fact that chemoinformaticians and medicinal chemists typically perceive similarity differently and we need to find ways to capture those differences in actionable measures of similarity.

The fourth challenge is the necessity of taking into account multiple domains of knowledge: drugs interact with proteins, treat and cause diseases, are produced by different methods (industrial or otherwise), have side effects, participate in metabolic reactions, etc. These concepts from other domains can also be compared semantically (many are even already represented in appropriate ontologies, including diseases, proteins, types of molecular interaction, manufacturing procedures, side effects, and pathways). The question now is how to take advantage of these other ontologies in order to implement a useful and accurate measure of chemical similarity. This issue is even related to the previous one, since by tuning the weight of these other domains, we can create new profiles of similarity more pertinent to some goals than others.

Another challenge is the absence of a standardised way to *validate* the measures that are proposed. In practice, for each new measure being proposed by some research group, that same group validates the new measure by comparing them with previous ones or by using it to show that the new measure can find

hidden knowledge in some dataset. However, the *ad hoc* way these validations are performed means that frequently the measures are neither comparable nor interchangeable and that they can only be used for the goal used to validate them. Thus, a general but useful validation strategy should also be developed to bring cohesion to this field.

7. Conclusion

This chapter introduces the ideas behind ontology-based semantic similarity measures, how they are applied in life sciences, and some of their uses in chemistry-related research endeavours. The main idea that we exposed is that these methods, having been used in other biomedical fields, can help cheminformatics in several fronts. We described three applications of where this methodology has been applied directly in cheminformatics research efforts and expect that this number grows as more people are exposed to this idea and its use cases.

We also exposed some of the future challenges in this area, which can serve as a starting point to anyone wishing to improve on the work already published, and provided general guidelines that should be taken into account for the further improvement of cheminformatics as a scientific field. In particular, we emphasise the need to explore the multidomain potential in semantic similarity, as well as the need to standardise the ways to validate measures of semantic similarity.

Acknowledgements

Abbreviations

ATC	anatomical therapeutic chemical classification system
ChEBI	chemical entities of biological interest
DAG	directed acyclic graph
GO	gene ontology
IC	information content
MeSH	medical subject headings
MICA	most informative common ancestor
OBO	Open Biological and Biomedical Ontology
QSAR	quantitative structure-activity relationship
simGIC	similarity of graphs with information content
simUI	similarity with union and intersection
SMILES	simplified molecular-input line-entry system
SNOMED CT	systematised nomenclature of medicine—clinical terms

Author details

João D. Ferreira* and Francisco M. Couto
LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Portugal

*Address all correspondence to: jdferreira@fc.ul.pt

IntechOpen

References

[1] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Research. 2017;**46**(D1):D1074-D1082. Available from: 10.1093/nar/gkx1037

[2] Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, et al. HMDB 4.0: The human metabolome database for 2018. Nucleic Acids Research. 2017;**46**(D1):D608-D617. Available from: 10.1093/nar/gkx1089

[3] Pence HE, Williams A. ChemSpider: An online chemical information resource. Journal of Chemical Education. 2010;**87**(11):1123-1124. Available from: 10.1021/ed100697w

[4] Bolton EE, Wang Y, Thiessen PA, Bryant SH. Chapter 12—PubChem: Integrated platform of small molecules and biological activities. In: Wheeler RA, Spellmeyer DC, editors. Annual Reports in Computational Chemistry. Vol. 4. Amsterdam, The Netherlands: Elsevier; 2008. pp. 217-241. Available from: http://www.sciencedirect.com/science/article/pii/S1574140008000121

[5] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: Improved access to chemical data. Nucleic Acids Research. 2018;**47**(D1):D1102-D1109. Available from: 10.1093/nar/gky1033

[6] Larsen PO, von Ins M. The rate of growth in scientific publication and the decline in coverage provided by science citation index. Scientometrics. 2010;**84**(3):575-603. Available from: 10.1007/s11192-010-0202-z

[7] Penzotti JE, Lamb ML, Evensen E, Grootenhuis PDJ. A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein. Journal of Medicinal Chemistry. 2002;**45**(9):1737-1740

[8] Fukunishi Y, Mikami Y, Takedomi K, Yamanouchi M, Shima H, Nakamura H, et al. Classification of chemical compounds by protein-compound docking for use in designing a focused library. Journal of Medicinal Chemistry. 2006;**49**(2):523-533

[9] Richard AM, Gold LS, Nicklaus MC. Chemical structure indexing of toxicity data on the internet: Moving toward a flat world. Current Opinion in Drug Discovery & Development. 2006;**9**(3):314-325

[10] Tohsato Y, Nishimura Y. Metabolic pathway alignment based on similarity between chemical structures. Information and Media Technologies. 2008;**3**(1):191-200

[11] Huang LC, Wu X, Chen JY. Predicting adverse side effects of drugs. BMC Genomics. 2011;**12**(5):S11. Available from: 10.1186/1471-2164-12-S5-S11

[12] Nikolic K, Mavridis L, Djikic T, Vucicevic J, Agbaba D, Yelekci K, et al. Drug design for cns diseases: polypharmacological profiling of compounds using cheminformatic, 3D-QSAR and virtual screening methodologies. Frontiers in Neuroscience. 2016;**10**:265. Available from: https://www.frontiersin.org/article/10.3389/fnins.2016.00265

[13] Raymond JW, Gardiner EJ, Willett P. Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. Journal of Chemical Information and Computer Sciences. 2002;**42**(2):305-316. PMID: 11911700. Available from: 10.1021/ci010381f

[14] Gillet VJ, Willett P, Bradshaw J. Similarity searching using reduced graphs. Journal of Chemical Information and Computer Sciences. 2003;**43**(2):338-345. PMID: 12653495. Available from: 10.1021/ci025592e

[15] Rogers D, Hahn M. Extended-connectivity fingerprints. Journal of Chemical Information and Modeling. 2010;**50**(5):742-754. PMID: 20426451. Available from: 10.1021/ci100050t

[16] Daylight Chemical Information Systems, Inc. Daylight Theory Manual. Daylight Headquarters; 2011 [Online]. Available from: https://www.daylight.com/dayhtml/doc/theory/ [Accessed: 19 June 2019]

[17] Wolosker H, Dumin E, Balan L, Foltyn VN. D-amino acids in the brain: D-serine in neurotransmission and neurodegeneration. The FEBS Journal. 2008;**275**(14):3514-3526

[18] Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Research. 2015;**44**(D1):D1214-D1219. Available from: 10.1093/nar/gkv1031

[19] Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: Enhanced functionality via new web services from the National Center for biomedical ontology to access and use ontologies in software applications. Nucleic Acids Research. 2011;**39**(suppl 2):W541-W545. Available from: 10.1093/nar/gkr469

[20] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO foundry: Coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology. 2007;**25**(11):1251

[21] Xiang Z, Mungall C, Ruttenberg A, He Y. Ontobee: A linked data server and browser for ontology terms. In: Proceedings of the 2nd International Conference on Biomedical Ontology; 2011

[22] Pesquita C, Faria D, Falcao AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. PLoS Computational Biology. 2009;**5**(7):e1000443

[23] Harispe S, Sánchez D, Ranwez S, Janaqi S, Montmain J. A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. Journal of Biomedical Informatics. 2014;**48**:38-53

[24] Sánchez D, Batet M. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. Journal of Biomedical Informatics. 2011;**44**(5):749-759. Available from: http://www.sciencedirect.com/science/article/pii/S1532046411000645

[25] Ferreira JD. Semantic similarity across biomedical ontologies [PhD thesis]. Universidade de Lisboa; 2016. Available from: http://hdl.handle.net/10451/25070

[26] Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics. 1989;**19**(1):17-30

[27] Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research. 1999;**11**:95-130

[28] Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation. Bioinformatics. 2003;**19**(10):1275-1283.

Available from: 10.1093/bioinformatics/
btg153

[29] Seco N, Veale T, Hayes J. An
intrinsic information content metric
for semantic similarity in WordNet.
In: Proceedings of the 16th European
Conference on Artificial Intelligence;
ECAI'04; Amsterdam, The
Netherlands, The Netherlands: IOS
Press; 2004. pp. 1089-1090. Available
from: http://dl.acm.org/citation.cfm?
id=3000001.3000272

[30] Sánchez D, Batet M, Isern D.
Ontology-based information content
computation. Knowledge-Based
Systems. 2011;**24**(2):297-303.
Available from: http://www.
sciencedirect.com/science/article/pii/
S0950705110001619

[31] Zhou Z, Wang Y, Gu J. A new model
of information content for semantic
similarity in WordNet. In: 2008 Second
International Conference on Future
Generation Communication and
Networking Symposia; vol. 3; 2008.
pp. 85-89

[32] Lin D. An information-theoretic
definition of similarity. In: Proceedings
of the Fifteenth International
Conference on Machine Learning;
ICML '98; San Francisco, CA, USA:
Morgan Kaufmann Publishers
Inc.; 1998. pp. 296-304. Available
from: http://dl.acm.org/citation.
cfm?id=645527.657297

[33] Couto FM, Silva MJ. Disjunctive
shared information between
ontology concepts: Application to
gene ontology. Journal of Biomedical
Semantics. 2011;**2**(1):5. Available from:
10.1186/2041-1480-2-5

[34] Ferreira JD, Hastings J, Couto FM.
Exploiting disjointness axioms to
improve semantic similarity measures.
Bioinformatics. 2013;**29**(21):2781-2787.
Available from: 10.1093/bioinformatics/
btt491

[35] Hastings J, de Matos P,
Dekker A, Ennis M, Muthukrishnan V,
Turner S, et al. Modular extensions
to the ChEBI ontology. In: Cornet R,
Stevens R, editors. Proceedings of
the 3rd International Conference on
Biomedical Ontology (ICBO 2012);
KR-MED Series, Graz, Austria; 21-25
July 2012; vol. 897 of CEUR Workshop
Proceedings; CEUR-WS.org; 2012.
Available from: http://ceur-ws.org/Vol-
897/poster 7.pdf

[36] Batet M, Sánchez D, Valls A. An
ontology-based measure to compute
semantic similarity in biomedicine.
Journal of Biomedical Informatics.
2011;**44**(1):118-125. Ontologies for
Clinical and Translational Research.
Available from: http://www.
sciencedirect.com/science/article/pii/
S1532046410001346

[37] Brameier M, Wiuf C. Co-clustering
and visualization of gene expression
data and gene ontology terms for
Saccharomyces cerevisiae using self-
organizing maps. Journal of Biomedical
Informatics. 2007;**40**(2):160-173.
Available from: http://www.
sciencedirect.com/science/article/pii/
S153204640600061X

[38] Zhang SB, Tang QR. Protein-
protein interaction inference based on
semantic similarity of gene ontology
terms. Journal of Theoretical Biology.
2016;**401**:30-37

[39] Yang D, Li Y, Xiao H, Liu Q,
Zhang M, Zhu J, et al. Gaining
confidence in biological interpretation
of the microarray data: The functional
consistence of the significant
GO categories. Bioinformatics.
2007;**24**(2):265-271. Available from:
10.1093/bioinformatics/btm558

[40] Jiang Y, Oron TR, Clark WT,
Bankapur AR, D'Andrea D, Lepore R,
et al. An expanded evaluation of protein
function prediction methods shows
an improvement in accuracy. Genome

Biology. 2016;**17**(1):184. Available from: 10.1186/s13059-016-1037-6

[41] Liu B, Jin M, Zeng P. Prioritization of candidate disease genes by combining topological similarity and semantic similarity. Journal of Biomedical Informatics. 2015;**57**:1-5. Available from: http://www.sciencedirect.com/science/article/pii/S1532046415001458

[42] Xue H, Peng J, Shang X. Predicting disease-related phenotypes using an integrated phenotype similarity measurement based on HPO. BMC Systems Biology. 2019;**13**(2):34. Available from: 10.1186/s12918-019-0697-8

[43] Köhler S, Schulz MH, Krawitz P, Bauer S, Dlken S, Ott CE, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. The American Journal of Human Genetics. 2009;**85**(4):457-464. Available from: http://www.sciencedirect.com/science/article/pii/S0002929709003991

[44] Ferreira JD, Couto FM. Semantic similarity for automatic classification of chemical compounds. PLoS Computational Biology. 2010;**6**(9):1-11. Available from: 10.1371/journal.pcbi.1000937

[45] Eltyeb S, Salim N. Chemical named entities recognition: A review on approaches and applications. Journal of Cheminformatics. 2014;**6**(1):17. Available from: 10.1186/1758-2946-6-17

[46] Lamurias A, Ferreira JD, Couto FM. Improving chemical entity recognition through h-index based semantic similarity. Journal of Cheminformatics. 2015;**7**(1):S13. Available from: 10.1186/1758-2946-7-S1-S13

[47] Tan F, Yang R, Xu X, Chen X, Wang Y, Ma H, et al. Drug repositioning by applying "expression profiles" generated by integrating chemical

structure similarity and gene semantic similarity. Molecular BioSystems. 2014;**10**:1126-1138. Available from: 10.1039/C3MB70554D

[48] Fakhraei S, Raschid L, Getoor L. Drug-target interaction prediction for drug repurposing with probabilistic similarity logic. In: Proceedings of the 12th International Workshop on Data Mining in Bioinformatics; BioKDD '13. New York, NY, USA: ACM; 2013. pp. 10-17. DOI: 10.1145/2500863.2500870

[49] Maggiora G, Vogt M, Stumpfe D, Bajorath J. Molecular similarity in medicinal chemistry. Journal of Medicinal Chemistry. 2014;**57**(8):3186-3204. PMID: 24151987. Available from: 10.1021/jm401411z