Semantic Similarity in the Biomedical Domain

João D. Ferreira & Francisco M. Couto

joao.ferreira@lasige.di.fc.ul.pt (corresponding author),
fcouto@di.fc.ul.pt

Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa

Abstract

One of the most important aspects in biomedical informatics is the ability to determine whether two entities are related to each other. For instance, in genetics, similarity between genetic products is often associated with one or more functions being shared among them; in chemistry, similarity in molecular structure correlates to a similar biological role; in medicine, similarity in two clinical cases is a strong argument towards a similar diagnosis.

But finding a way to compare these entities is not trivial and strongly depends on the entities being compared. For gene products, similarity can be calculated by comparing theirs sequences; for chemical compounds, by comparing the graphs that represent their structure. While these methods have been shown to be useful, they may fail in some cases (e.g., L-serine and D-serine, although extremely similar in structure, have very different biological roles). Moreover, in some other cases, like the medical example given above, there is not an easy way to extract a similarity measure: how to generate a numeric value to the degree of similarity between two clinical cases?

This gap can be filled in by the notion of ontologies. These are machine-readable representations of knowledge, a way to make computers aware of given facts, which are, themselves, simple statements expressing a relation between concepts: for example, “an Arm is a Limb”, “a Hand is adjacent to an Arm”, or “Fever is a symptom of the Flu”. Thus, ontologies allow computers to understand the meaning behind the concept. By explicitly providing the relations between these concepts to a computer, we open up the possibility to employ computational power to automatically explore them.

One of the technologies enabled by the use of ontologies is indeed the calculation of similarity between the concepts they represent, a technology also known as semantic similarity [1]. Since an Arm is a Limb and a Leg is also a Limb, they are more similar than, e.g., an Arm and a Torso. Or, because Fever is a symptom of the Flu, the two concepts are more related to each other than Fever and Leg. By exploiting ontologies and the relations they contain, we can therefore create measures that can compare these concepts and the entities they represent. In the biomedical field, several fields have already benefited from this notion of semantic similarity, including (i) protein functions; (ii) chemical compounds; (iii) symptoms; and (iv) anatomical concepts.

Despite these previous achievements, the usefulness of this technology lies well beyond the mere capability of comparing ontological concepts. Consider the Gene Ontology, which contains a few thousand gene functions. This ontology can be used to “annotate” proteins, i.e., it can be used to explicitly state that “protein A has function X”. By doing so, proteins are brought up to the world of machine-readable semantics. Computers can leverage on these annotations to compare proteins not only by their sequence (using classical methods such as BLAST) but by their functions as well. Imagine that a clinical case is annotated with some anatomical concepts, several symptoms, and a set of altered chemical compounds in blood screenings. By using semantic similarity over all these concepts, we can achieve the goal of objectively measuring the degree of relatedness between two clinical cases, eventually leading to diagnostic and treatment of the patient.

In the course of my PhD program, I have created and applied two measures of semantic similarity in the biomedical domain: one for chemical compounds, using the ontology for Chemical Entities of Biological Interest (ChEBI) and another for anatomical concepts, using the Foundational Model of Anatomy (FMA) ontology. To evaluate the similarity measure in ChEBI, I applied it a classification problem. The results show that semantic similarity can be used to enhance the performance of predicting whether small molecules are able to cross the blood-brain barrier, thus establishing that the measure does indeed reflect biological similarity.

The work done on semantic similarity for the FMA [2] consists of a generic measure of semantic relatedness that can potentially be applied to any ontology. More interestingly, it can be applied to multiple ontologies, using for that effect links between two ontologies. For example, the concepts of the Human Phenotype Ontology (HPO), an ontology containing human symptoms, make some cross-references to anatomical concepts on FMA. By treating these as bridges between FMA and HPO, it is possible to extract relations between anatomical concepts and symptoms, thereby bringing the notion of semantic similarity to the multi-domain level.

References

1. Pesquita C, et al. (2009). Semantic Similarity in Biomedical Ontologies. PLoS Comput Biol, 5(7): e1000443.
2. Ferreira JD & Couto FM (2010). Generic semantic relatedness measure. Proceedings of the ICBO 2012.