P.38

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

One of the most important aspects in molecular biology is the ability to determine whether two molecules 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.
It is possible to compare these entities directly from their structure, by comparing aminoacid sequences or the graph representation of molecules, however, these methods do not always reflect biological similarity: for example, L-serine and D-serine have very different biological roles. Moreover, in some cases, such as when a comparison of biological roles is needed, there is not an easy way to extract a similarity measure: how to generate a mathematical representation of a function?
Ontologies can fill this gap. Ontologies represent knowledge by means of simple statements expressing a relation between concepts: for example, “<vasodilation> is part of <blood circulation>”, “<lithium sulfate> has role <antidepressant>”. Thus, ontologies allow computers to automatically explore the meaning behind concepts.
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. Since “<arm> is a <limb>” and “<leg> is a <limb>”, they are more similar than, e.g. an <arm> and <torso>. We can therefore create measures to compare the entities represented by these concepts. The usefulness of this technology, however, transcends the simple comparison of ontological concepts. With the help of Gene Ontology annotations (e.g. “<CFTR> has function <ion transporter>”), proteins themselves gain a machine-readable semantics that can be used to compare proteins not only by their sequence (using classical methods such as BLAST) but by their functions as well.