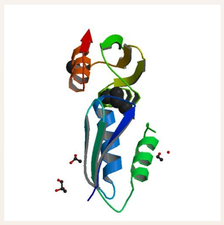
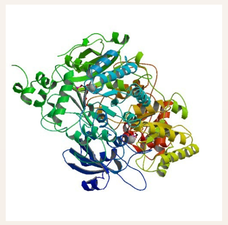
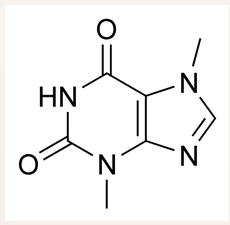
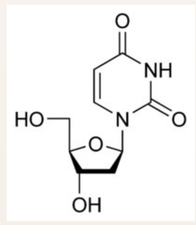
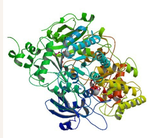
Semantic Similarity in the Biomedical Domain

João D. Ferreira & **Francisco M. Couto**
joao.ferreira@lasige.di.fc.ul.pt

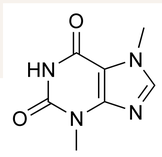
Are these things similar?

How to compare them?

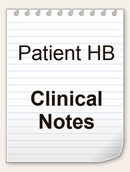


- Compared proteins by sequence: **BLAST**



- Compare chemicals by structure: **Graph comparison algorithms**

These methods may not reflect biological similarity:
e.g. L-serine and D-serine are almost identical, but they have very different biological roles

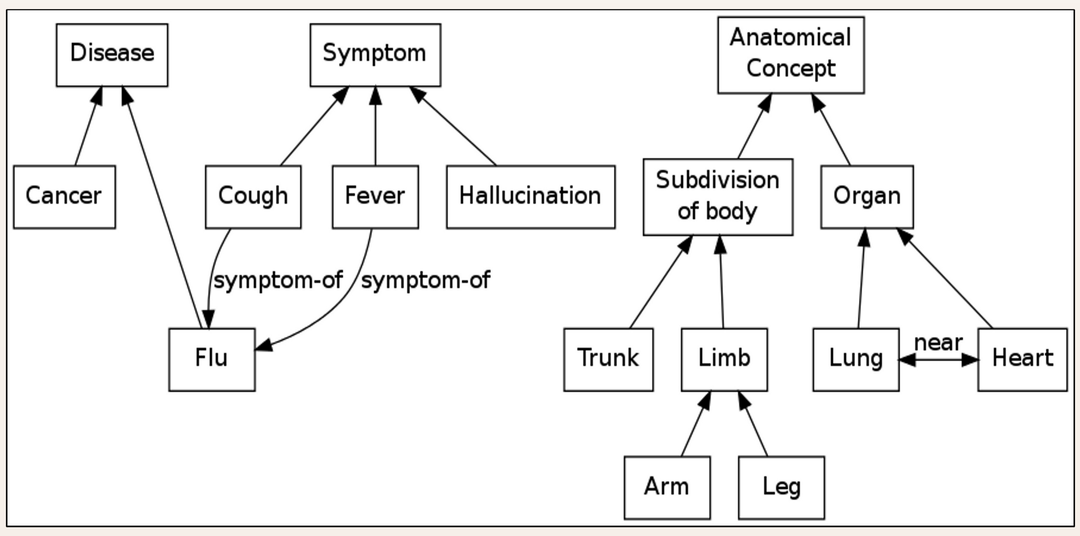


What about things without **mathematical** or other **formal** representation?

Ontologies

Ontologies represent **knowledge** in a formal way
1) “an Arm is a Limb”
2) “Fever is a symptom of Flu”

Computers can manage these facts and explore them



Semantic Similarity

By using the ontologies, we can compare pairs of concepts:

Arm** is more similar to **Leg** than to **Trunk

More than similarity, we can get degrees of **relatedness**:

Flu** is more related to **Fever** than to **Hallucination
Heart** is more related to **Lung** than to **Pancreas

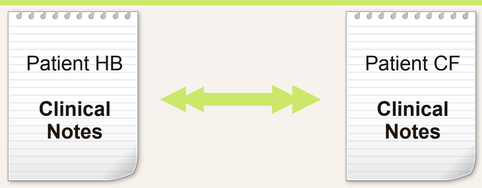
Comparing complex concepts

Proteins are annotated with GO terms

- GO:0006984
- GO:0000133
- GO:0000633

- GO:0018655
- GO:0005441
- GO:0005090

Clinical terms are annotated with anatomical terms, symptoms, metabolites, …



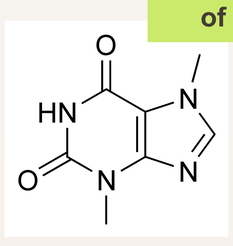
- Lung
- Fever
- Low carnitine

- Heart
- Sneeze
- Rash

It's possible to compare concepts such as **proteins** and **clinical cases**

My contributions

ChEBI – Chemical Entities of Biological Interest



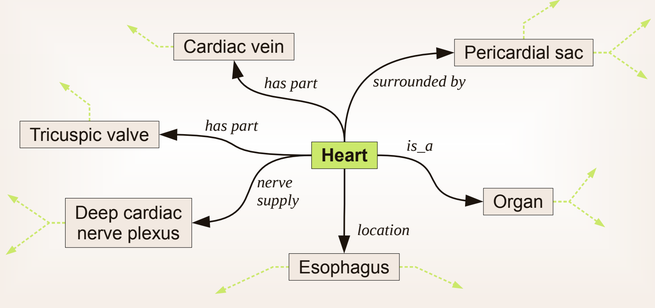
Can this pass the blood brain barrier?

Semantic similarity can be used to improve classification of molecules [1].

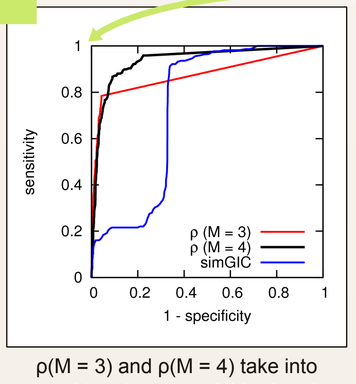
Improvements in accuracy:
81.5% → 90.0%

FMA – Foundational Model of Anatomy

Can we detect related anatomical structures by comparing their **neighborhood**? [2]



ROC analysis



ρ(M = 3) and ρ(M = 4) take into consideration the neighborhood. simGIC is a baseline that only considers class-subclass relations

The closer to point (0,1), the higher the performance of the measure

[1] Ferreira JD, Couto FM (2010). Semantic **Similarity for Automatic Classification of Chemical Compounds**. PLoS Computational Biol 6(9): e1000937.

[2] Ferreira JD, Couto FM (2010). **Generic semantic relatedness measure for biomedical ontologies**. Proceedings of the ICBO 2011.