A Collection of Benchmark data sets for Knowledge Graph-based Similarity in the Biomedical Domain

Carlota Cardoso[1], Rita T. Sousa[1], Sebastian Köhler[2], Catia Pesquita[1]

cmacardoso@ciencias.ulisboa.pt

[1] LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2] Monarch Initiative, monarchinitiative.org

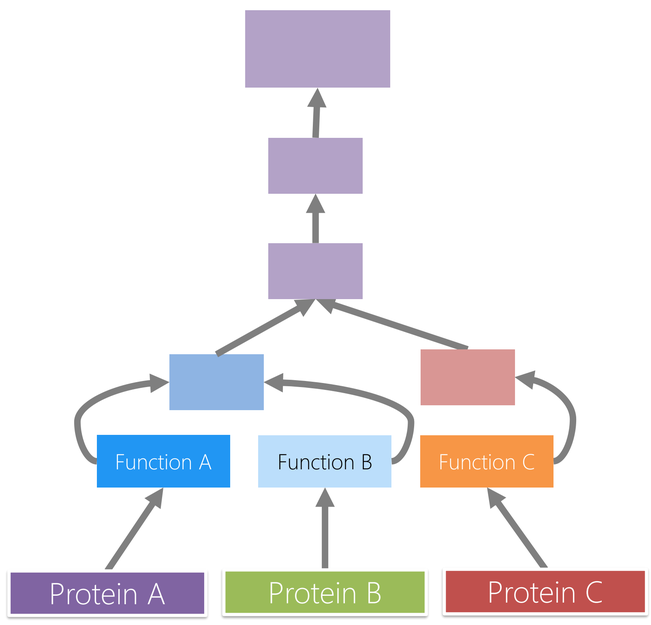
LASIGE health and biomedical informatics

INTRODUCTION

Biomedical research produces large amounts of data about the function, regulation and interaction of genes and proteins. The size of the data and its underlying complexity were strong motivators for the adoption of ontologies. The knowledge provided by linking entities and ontologies can be represented in graph form, Knowledge Graph (KG).

KGs provide the structure that supports the comparison of entities through semantic similarity (SS).

A semantic similarity measure is a function that returns a numerical value reflecting the closeness in meaning between entities. There are several measures available, each with their distinguishing characteristics.



$$sim(■, ■) > sim(■, ■)$$

MOTIVATION

Evaluating the reliability of biomedical SS measures is still an open question since:

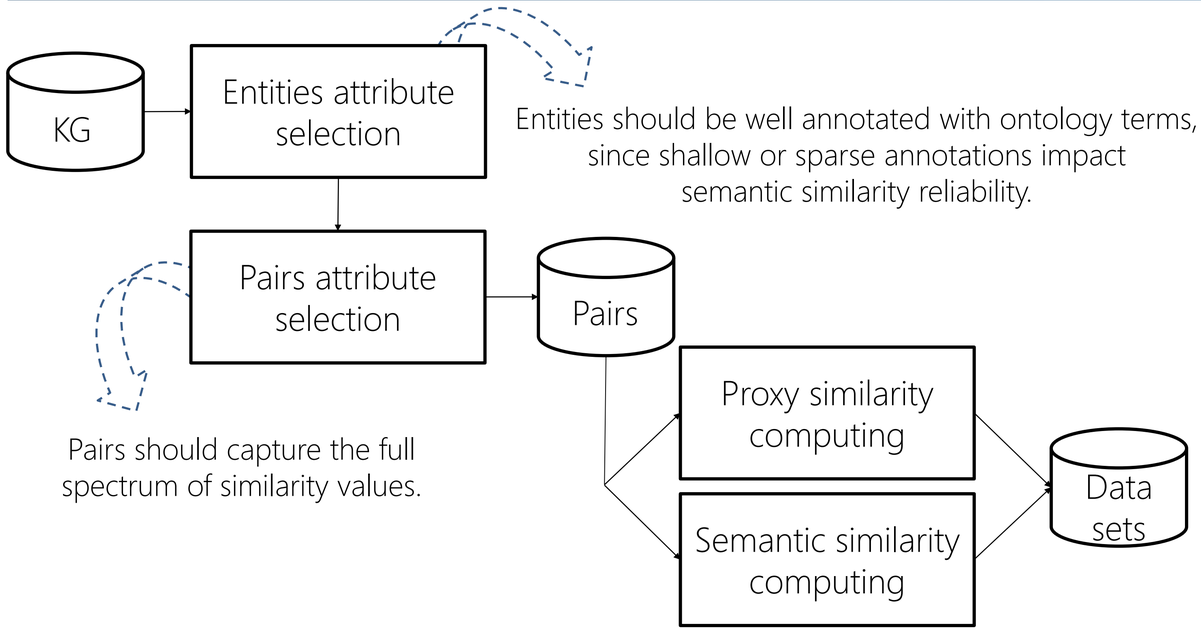
- there is no gold standard for similarity, due to data size and diversity;
- crowdsourcing is not suitable in expert domains.

One solution can be to compare them to similarity proxies. In this domain proxies can be, for instance, the sequence similarity of two genes or the number of metabolic pathways common to two diseases.

GOAL

Develop benchmark data sets to support similarity proxy-based evaluation of semantic similarity measures for biomedical entities at a large scale.

METHODOLOGY



The proposed methodology was applied to the Gene Ontology and the Human Phenotype Ontology to produce benchmark data sets based on different similarity proxies:

- Protein-protein interaction (PPI)
- Protein molecular function (PMF)
- Gene role in phenotypic abnormalities (GP)

RESULTS

We developed 21 benchmark data sets comprising 2 annotation levels and 31k different genes and proteins from 4 species. The data set sizes range from 270 to 158k pairs.

	E. coli	*D. melanogaster*	*H. sapiens*	*S. cerevisiae*	*All species*
PPI	428/738	270/397	42k/45k	22k/35k	64k/80k
PMF	2k/4k	52k/53k	60k/60k	31k/42k	142k/158k
GP			12k		

For each pair, 4 representative semantic similarity measures and relevant similarity proxies were calculated.

CONCLUSION

The developed data sets can support:

- Evaluation of semantic similarity measures
- Protein-protein interaction prediction
- Correlation visualization

Even though these data sets are domain-specific, they can be used for the evaluation of general-purpose semantic similarity measures or adapted to any domain where a similarity proxy can be created.

Data sets and KG data available for the scientific community at https://github.com/liseda-lab/kgsim-benchmark/tree/master

Acknowledgements: This work was supported by the Fundação para a Ciência e a Tecnologia (FCT) through the LASIGE Research Unit (UIDB/00408/2020 and UIDP/00408/2020) and the SMILAX project (PTDC/EEI-ESS/4633/2014). RTS is also funded by the FCT grant (SFRH/BD/145377/2019).