evoKGsim
Evolving Semantic Representations for Protein-Protein Interaction Prediction

Rita T. Sousa, Sara Silva, Catia Pesquita

LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

Full Paper

LASIGE health and biomedical informatics

Problem

Knowledge graphs (KGs) represent an unparalleled opportunity for machine learning, given their ability to provide meaningful context. Since typical machine learning techniques are vector-based, semantic representations bridge the gap between KGs and data mining methods. Although KGs provide multiple perspectives over an entity, semantic representations are static and ignore that some semantic aspects (SAs) may be irrelevant to the downstream learning task.

Solution

evoKGsim [1,2] is a methodology that employs Genetic Programming (GP) to evolve similarity-based semantic representations for KGs, optimized for specific learning tasks. We implemented two variants of the methodology, one using taxonomic semantic similarity and the other using graph embeddings similarity. **evoKGsim** can be applied to the classification of pairs of KG entities.

Taxonomic semantic similarity (SS) as semantic representation

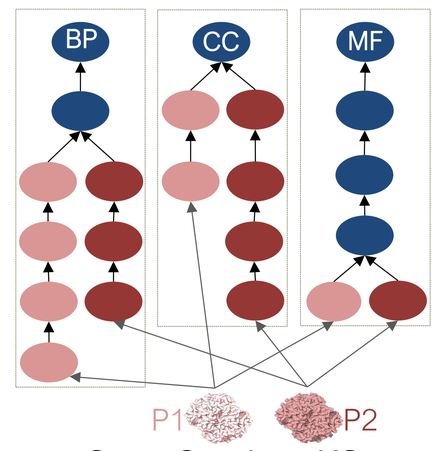
Taxonomic semantic similarity compares proteins based on the taxonomic relations within the Gene Ontology (GO) graph. The taxonomic semantic similarity, for each SA, was calculated using the pairwise approach **ResnikMaxSeco**.

$$SS_{SA}(P1,P2) = max\{sim(c_1,c_2): c_1 \in C(P1), c_2 \in C(P2)\}$$

$$C(P)\ is\ the\ set\ of\ classes\ annotated\ protein\ P$$

$$sim(c_1,c_2) = max\{IC(c): c \in \{A(n_1) \cap A(n_2)\}\}$$

$$A(c)\ is\ the\ set\ of\ ancestors\ of\ class\ c$$



Semantic representations

Computation of KG-based similarities for each SA (BP, CC, MF)

Evolution of semantic representations

GP for evolving possible combination of semantic aspects to support PPI prediction

Best combination of semantic aspects

Evaluation

Performance evaluation using the best combination of SAs

Graph embeddings similarity (ES) as a semantic representation

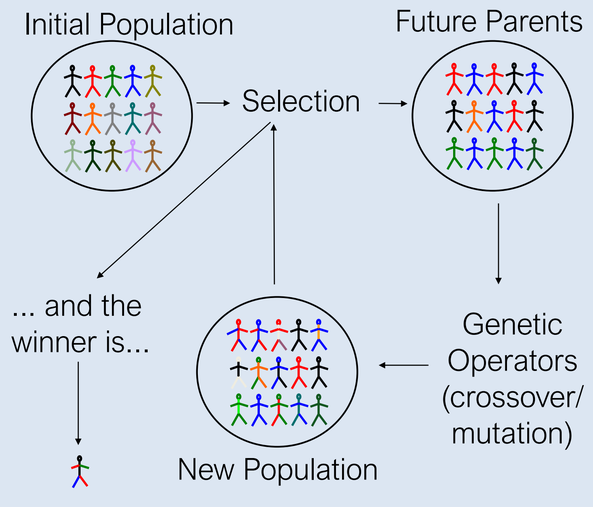
To calculate graph embeddings for each protein we employed the **RDF2Vec** approach.

1. GO KG is converted into a set of sequences of graph entities using random walks
2. The obtained sequences are used to train word2Vec
3. Each protein is represented as a vector of latent numerical features (protein embedding)

To compute the graph embeddings similarities, for each SA, we employ cosine similarity.

Evolution of semantic representations

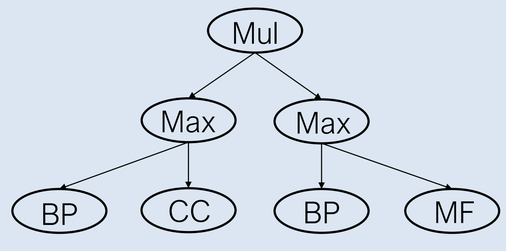
The evolution of semantic representations is guided by a fitness function that measures performance on a specific learning task.



Evaluation

evoKGsim was evaluated on several benchmark datasets for protein-protein interaction (PPI) prediction using the GO, with its three SAs: biological process (BP), cellular component (CC) and molecular function (MF).

Parse tree of one of the simplest combinations evolved in our experiments.



Max(BP,CC) × Max(BP,MF)

Results

Dataset	Size	Static-SS	evoKGsim-SS	Static-ES	evoKGsim-ES
STRING-EC	2245	0.834	**0.861**	0.815	**0.817**
STRING-DM	550	0.937	**0.945**	**0.900**	0.890
BIND-SC	1366	0.909	**0.912**	0.808	**0.812**
DIP/MIPS-SC	13807	0.843	**0.849**	0.804	**0.815**
STRING-SC	30384	0.835	**0.845**	0.795	**0.800**
DIP-HS	2739	0.882	**0.901**	0.698	**0.707**
STRING-HS	6912	0.853	**0.872**	0.766	**0.780**
GRID/HPRD-unbal-HS	31320	0.729	**0.735**	0.606	**0.609**
GRID/HPRD-bal-HS	31349	0.656	**0.665**	0.664	**0.665**

Median of weighted average F-measure for static baselines and for evoKGsim

Future Work

Despite the specific domain of evaluation employed, the **evoKGsim** methodology can also be extended to other semantic representations and generalized to other applications and domains (e.g., disease gene discovery, KG link prediction or recommendations).

References

[1] Sousa, R., Silva, S., Pesquita, C.: Evolving knowledge graph similarity for supervised learning in complex biomedical domains. BMC Bioinformatics (2020)

[2] Sousa, R., Silva, S., Pesquita, C. (2019,May). Evolving meaning: Using Genetic Programming to learn similarity perspectives for mining biomedical data. Paper presented at the meeting of European Society for Clinical Investigation, Coimbra, Portugal.

Acknowledgements: This work was supported by the Fundação para a Ciência e a Tecnologia (FCT) through the LASIGE Research Unit, ref. UIDB/00408/2020 and ref. UIDP/00408/2020. CP and RTS are funded by project ref. PTDC/EEI-ESS/4633/2014, and SS by projects ref. PTDC/CCI-INF/29168/2017 and ref. PTDC/CCI-CIF/29877/2017. SS and CP are also funded by project ref. DSAIPA/DS/0022/2018, and RTS by PhD grant SFRH/BD/145377/2019.