Towards predicting essential proteins via federated SPARQL queries

Petros Liakopoulos[2], Borbala Banfalvi[1], Xinyi Wang[1], Sina Majidian[2], Tarcisio Mendes de Farias[2,3], Christophe Dessimoz[2,3] and Ana Claudia Sima[3]

[1]*Department of Genetics, Evolution, Environment, University College London, UK*
[2]*Department of Computational Biology, University of Lausanne, Switzerland*
[3]*SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland*

Abstract

We showcase the role of federated queries in reproducible science, by attempting to replicate an existing study to predict essential proteins through an integrative approach. More precisely, we compute orthologous scores of proteins using OMA, as well as the expression breadth of their orthologs using Bgee, which we federate via SPARQL, in order to rank candidate essential proteins across model organisms. We highlight challenges of this attempt, in particular the granularity of the data available in RDF, performance limitations, but also the current absence of a protein-protein interaction dataset in RDF.

The SIB provides a growing catalog of interoperable datasets available through public SPARQL endpoints[1]. For the purpose of this paper we make use of the OMA database of evolutionary relationships [1] and the Bgee gene expression database [2], both available in RDF through public SPARQL endpoints. Furthermore, in order to add richer information (e.g. functional annotations of genes) we also incorporate RDF data from UniProt [3].

The goal of this paper is to analyse to what extent it is possible, using today's bioinformatics databases available in RDF, to predict essential proteins in a target organism. These are proteins that are key to ensuring the survival of the organism. The authors of [4] demonstrate that their prediction can be achieved through an integrative analysis incorporating orthology, gene expression and protein-protein interaction (PPI) data. However, given that PPI databases are not yet available as RDF, we focus on querying orthology and gene expression data for reproducing the integrative analysis in this study.

In similar lines to the methodology described in [4], our federated SPARQL query aims to compute the *orthologous score* of a protein (number of reference species that have an ortholog for this protein), along with the *expression breadth* of its orthologs. This is the number of tissues where the gene coding for a protein is highly expressed (*i.e.*, expression level above 99 in Bgee, with a high confidence score associated). The simplifying assumption we make is: proteins that have a high orthologous score (*i.e.*, conserved across many species), and whose orthologs are broadly expressed in the reference species, are good candidates for essential proteins.

For an overview of computational methods for predicting essential proteins, we refer the

SWAT4HCLS 2023: The 14th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences
ana-claudia.sima@sib.swiss (A. C. Sima)

[1]An overview of these resources can be seen at https://www.expasy.org/search/sparql

reader to [5]. Furthermore, a more extensive catalog of questions and corresponding SPARQL queries targeting OMA, Bgee and UniProt is available online[2] as part of our previous work.

We provide the federated SPARQL query covering this use case in our github repository at https://github.com/dssib/swat23-data-in-use. Furthermore, in the accompanying Jupyter notebook in the repository, we also provide a step-by-step construction of the query.

The federated query we considered here is merely a starting point for reproducing the analysis in the initial study and is currently limited by the performance of servers involved (*i.e.*, we can only currently run the query using *limits*). In the extended discussion in our Jupyter notebook, we also highlight the difficulty in writing federated queries with aggregations. However, compared to the methodology of the study, ours would have a few important advantages: the SPARQL queries are in theory fully reproducible, allowing any reader to directly obtain the datasets of interest from the most recent versions of the databases. Moreover, the intersection of these databases is performed implicitly, via the federation, as opposed to the manual work involved in merging datasets downloaded from disparate resources (*e.g.* InParanoid, Gene Expression Omnibus in the case of our considered study), an effort that would need to be repeated by every reader, given that the finalised dataset is not references in the paper.

Ranking candidate genes by expression level and expression breadth could be an interesting direction for future work. The availability of a Protein-Protein Interaction database in RDF, through a public SPARQL endpoint, would increase the potential to reproduce the results from the study by incorporating network topological features, as well as the co-expression within interaction networks. All in all, federated SPARQL queries remain an interesting, currently under-explored avenue for the future of reproducible research, in particular in the case of integrative analyses. We are working towards compiling a catalog of questions from scientific publications that involve multiple databases jointly, with their corresponding federated queries.

References

[1] A. M. Altenhoff, C.-M. Train, K. J. Gilbert, I. Mediratta, T. Mendes de Farias, D. Moi, Y. Nevers, H.-S. Radoykova, V. Rossier, A. Warwick Vesztrocy, et al., Oma orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more, Nucleic acids research 49 (2021) D373–D379.

[2] F. B. Bastian, J. Roux, A. Niknejad, A. Comte, S. S. Fonseca Costa, T. M. De Farias, S. Moretti, G. Parmentier, V. R. De Laval, M. Rosikiewicz, et al., The bgee suite: integrated curated expression atlas and comparative transcriptomics in animals, Nucleic Acids Research 49 (2021) D831–D847.

[3] Uniprot: the universal protein knowledgebase in 2021, Nucleic acids research 49 (2021) D480–D489.

[4] X. Zhang, W. Xiao, X. Hu, Predicting essential proteins by integrating orthology, gene expressions, and ppi networks, PloS one 13 (2018) e0195410.

[5] C. Dong, Y.-T. Jin, H.-L. Hua, Q.-F. Wen, S. Luo, W.-X. Zheng, F.-B. Guo, Comprehensive review of the identification of essential genes using computational methods: focusing on feature implementation and assessment, Briefings in bioinformatics 21 (2020) 171–181.

[2] https://biosoda.expasy.org/build_biosodafrontend/