Highlight on semantic web technologies are effective to remove redundancies from protein-protein interaction databases and define reproducible interactomes

Marc Melkonian, Camille Juigné, Olivier Dameron, Gwenaël Rabut, Emmanuelle Becker

Highlight on Semantic Web technologies are effective to remove redundancies from protein-protein interaction databases and define reproducible interactomes

Marc Melkonian[1,2], Camille Juigné[1,3], Olivier Dameron[1], Gwenael Rabut[2] and Emmanuelle Becker[1]

[1] Univ Rennes, Inria, CNRS, IRISA - UMR 6074, F-35000 Rennes, France
[2] Univ Rennes, CNRS, IGDR-UMR 6290, Rennes, F-35000, France
[3] PEGASE, INRAE, Institut Agro, 35590, Saint Gilles, France

Corresponding author: marc.melkonian@irisa.fr

Abstract

Keywords Reproducibility, Interactome, Semantic Web

Information on protein-protein interactions (PPIs) is collected in numerous primary databases with their own curation process. To provide more exhaustive datasets, several meta-databases aggregate PPIs from multiple primary databases. However, aggregation of PPIs from different primary databases is not straightforward since distinct databases are often partly redundant and may have different PPI annotation policies. Mere aggregation can thus introduce a bias if these redundancies are not identified and eliminated, leading to systematically overestimating PPI reproducibility.

We propose a precise definition of two types of redundancies that can be observed between entries of PPI databases. We define explicit redundancy as the exact duplication of a PPI information in two distinct database entries. It occurs in particular when two databases have independently registered the same experimental evidence from a given publication and annotated it using an identical interaction detection method (IDM) term from the Proteomics Standards Initiative-Molecular Interactions (PSI-MI) ontology. Explicit redundancy is trivial and usually eliminated during the aggregation process of meta-databases. In contrast, we define implicit redundancy as the occurrence of two database entries referring to the same PPI reported in a given publication, but annotated with related IDM terms. Implicit redundancy occurs in particular when primary databases have different PPI annotation policies and preferentially use different types of IDM terms (for instance general vs specific). We show that both types of redundancies can be easily detected using Semantic Web technologies.

Using a dataset from the APID meta-database, we observed that while explicit redundancies are detected by the APID aggregation process, about 15% of APID entries are implicitly redundant. More than 90% of these redundancies result from the aggregation of distinct primary databases (inter-database redundancy), while the remaining occurs between entries of a single database (intra-database redundancy).

Further, we built for two species (yeast and human) a "reproducible interactome" with interactions that have been reproduced by multiple methods or publications. For both species, the size of the reproducible interactome is drastically impacted by removing redundancies (-59% and -56% for yeast and human, respectively), and we show that this is largely due to implicit redundancies. Thus, a significant number of PPIs currently considered as reproducible actually relies on database integration artefacts.

All software are are freely available at https://gitlab.com/nnet56/reproducible-interactome and data and results can be browsed and downloaded at https://reproducible-interactome.genouest.org/.