Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases

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

Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases

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

Databases and Ontologies

Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases

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

[1] Univ Rennes, Inria, CNRS, IRISA, F-35000, Rennes, France
[2] Univ Rennes, CNRS, IGDR - UMR 6290, F-35000, Rennes, France
[3] Pegase, Inrae, Institut Agro, 35590 Saint-Gilles, France.

* To whom correspondence should be addressed, equal contribution.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Information on protein-protein interactions is collected in numerous primary databases with their own curation process. Several meta-databases aggregate primary databases to provide more exhaustive datasets. In addition to exhaustivity, aggregation contributes to reliability by providing an overview of the various studies and detection methods supporting an interaction. However, interactions listed in different primary databases are partly redundant because some publications reporting protein-protein interactions have been curated by multiple primary databases. Mere aggregation can thus introduce a bias if these redundancies are not identified and eliminated. To overcome this bias, meta-databases rely on the Molecular Interaction ontology that describes interaction detection methods, but they do not fully take advantage of the ontology's rich semantics, which leads to systematically overestimating interaction reproducibility.

Results: We propose a precise definition of explicit and implicit redundancy, and show that both can be easily detected using Semantic Web technologies. We apply this process to a dataset from the APID meta-database and show that while explicit redundancies were detected by the APID aggregation process, about 15% of APID entries are implicitly redundant and should not be taken into account when presenting confidence-related metrics. More than 90% of implicit redundancies result from the aggregation of distinct primary databases, while the remaining occurs between entries of a single database. Finally, we build a "reproducible interactome" with interactions that have been reproduced by multiple methods or publications. The size of the reproducible interactome is drastically impacted by removing redundancies for both yeast (-59%) and human (-56%), and we show that this is largely due to implicit redundancies.

Availability: Software, data and results are available at `https://gitlab.com/nnet56/reproducible-interactome`, `https://reproducible-interactome.genouest.org/`, Zenodo (doi:10.5281/zenodo.5595037) and NDEx (doi:10.18119/N94302, doi:10.18119/N97S4D

Contact: emmanuelle.becker@irisa.fr, gwenael.rabut@inserm.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein-protein interactions (PPIs) play an ubiquitous and fundamental role in all biological processes. Description of PPIs is essential to understand how proteins operate at the molecular level and the construction of accurate

and comprehensive protein interaction networks (or interactomes) is an important aim of biological research (Bonetta, 2010; Cafarelli *et al.*, 2017; Luck *et al.*, 2020; Huttlin *et al.*, 2021).

PPIs can be probed using numerous interaction detection methods (IDMs), following biophysical (e.g. x-ray crystallography), biochemical (e.g. affinity purification) or genetic approaches (e.g. yeast two-hybrid). Importantly, since different IDMs probe PPIs in a different manner, they produce complementary results that often do not fully overlap. For instance, some IDMs are designed to detect binary interactions of proteins probed in pairs (e.g. yeast two-hybrid), while others probe interactions of protein groups assembled in complexes (e.g. affinity purification). Consequently, the biological interpretation of PPI networks depends on the underlying IDMs that have been used to produce them. Moreover, since IDMs can generate false positive and false negative interactions, multiple observations of a given PPI with different experimental techniques reinforce the confidence in this PPI. Accurate IDM annotation and interpretation is thus an important issue in interactome studies.

Information on published PPIs is collected in primary databases such as IntAct (Kerrien *et al.*, 2012), MINT (Calderone *et al.*, 2020), BioGRID (Oughtred *et al.*, 2019), DIP (Salwinski *et al.*, 2004) or HPRD (Keshava Prasad *et al.*, 2009). The major databases report IDMs using a controlled vocabulary defined by the Proteomics Standards Initiative-Molecular Interactions (PSI-MI) consortium (Sivade Dumousseau *et al.*, 2018). This vocabulary is structured in an ontology to represent the hierarchical relationships between IDM families by a directed acyclic graph.

Each primary database follows its own curation process with different literature mining, filtering, and reporting techniques. To address the resulting need for integration, several meta-databases aggregate information from multiple primary databases to provide more exhaustive PPI datasets. Some of these meta-databases, such as the Agile Protein Interactomes DataServer (APID) (Alonso-López *et al.*, 2016; Alonso-López *et al.*, 2019), HINT (Das and Yu, 2012) or mentha (Calderone *et al.*, 2013), focus exclusively on experimentally determined PPIs, while others, such as IID (Kotlyar *et al.*, 2019) or STRING (Szklarczyk *et al.*, 2019) also integrate predicted interactions, text mining results or other information.

The accurate aggregation of PPIs from multiple and partly redundant sources is not a trivial task (Turinsky *et al.*, 2010; Klapa *et al.*, 2013). Although the primary databases refer to the PSI-MI ontology, they do not necessarily select identical terms to annotate PPIs (Alonso-López *et al.*, 2019). Hence, a PPI observed in a single experiment reported in a given publication can be annotated with distinct IDM terms in different primary databases. Such annotation differences are usually not taken into account or corrected during the aggregation process.

APID, which unifies data from five of the largest PPI databases (Alonso-López *et al.*, 2016; Alonso-López *et al.*, 2019), implements an integration method that takes redundancy into account and enables to distinguish *'experimental evidences'* (i.e experimental observations reported in publications) from *'curation events'* (i.e. entries in PPI databases). For a given protein pair, multiple entries annotated with identical IDM and identical PubMed publication identifier (PMID) are considered as duplicates and counted as a single experimental evidence. In addition, IDMs are classified into *'binary'* and *'indirect'* methods and IDMs corresponding to related binary methods (e.g. 'two hybrid array' and 'two hybrid pooling approach') are assigned a common method type (e.g. 'two hybrid'). This common method type is then used instead of the original IDM to identify duplicate entries across multiple databases. This custom integration process is not fully satisfying since it is restricted to binary interactions and it does not take advantage of the PSI-MI ontology.

We propose a novel approach to integrate PPI information from primary databases. We define the conventional **explicit redundancy** and extend it with **implicit redundancy** based on parent-related terms in the PSI-MI ontology. We present a method relying on Semantic Web technologies that successfully detects and reconciles implicit redundancies in curation events compiled from multiple primary databases, opening the way to an improved automated curation process. Once curated for both explicit and implicit redundancies, the integrated set of experimental evidences can be used to determine the reproducible interactome supported by multiple experiments.

2 Approach

2.1 Explicit and implicit redundancy

Let us consider a pair of proteins (A, B) and count the number of non-redundant experiments reporting their interaction.

Primary databases such as BioGRID or IntAct can provide several entries corresponding to this protein pair. Usually, these entries differ in the IDM, the PMID, or both. An entry in these databases can thus be defined by a quadruplet

$$(A, B, M_i, P_x)$$

where A and B are the proteins, M_i is the IDM (such as 'affinity chromatography technology', 'anti-tag coimmunoprecipitation' or 'two hybrid', for the most frequent ones), and P_x is the PMID of the original article describing their interaction. When two entries only differ in the IDM, this should signify that the original article has observed the interaction using several experimental techniques. When two entries only differ in the PMID, this should signify that the interaction has been reproduced in two distinct studies using the same detection method.

For meta-databases such as APID, populated by aggregating curation events from other databases, an entry can be defined by a quintuplet

$$(A, B, M_i, P_x, D_a)$$

where D_a indicates the primary database indexing the interaction. Meta-databases can contain different types of redundancies:

- **Explicit redundancy** occurs when distinct entries referring to the same protein pair (A, B) and the same PMID P_x have an identical IDM M_i. This happens when two primary databases registered the same experimental evidence using the same IDM term. Explicit redundancies are detected and unified by APID and other meta-databases.
- **Implicit redundancy** occurs when distinct entries referring to the same protein pair and the same PMID have been annotated with different IDMs although they correspond to the same experimental evidence. In practice, this occurs when curators select IDM terms at different levels of the ontology, one being more general and the other more specific. For example, the interaction of the human proteins MDM2 and TP53 is listed in APID as (MDM2, TP53, 'anti tag Co-immunoprecipitation', PMID:17159902, INTACT:7156209) and also as (MDM2, TP53, 'affinity chromatography technology', PMID:17159902, BIOGRID:680279). Although biologists would naturally recognize one observation annotated twice at different granularities, the redundancy is not explicit. Implicit redundancy should not be confused with the common case where several experimental techniques are used in a single publication to validate a given PPI. Therefore, detecting implicit redundancies requires knowledge on IDMs.

Hereafter, we take advantage of the PSI-MI ontology to identify these two cases, as illustrated in Figure 1.

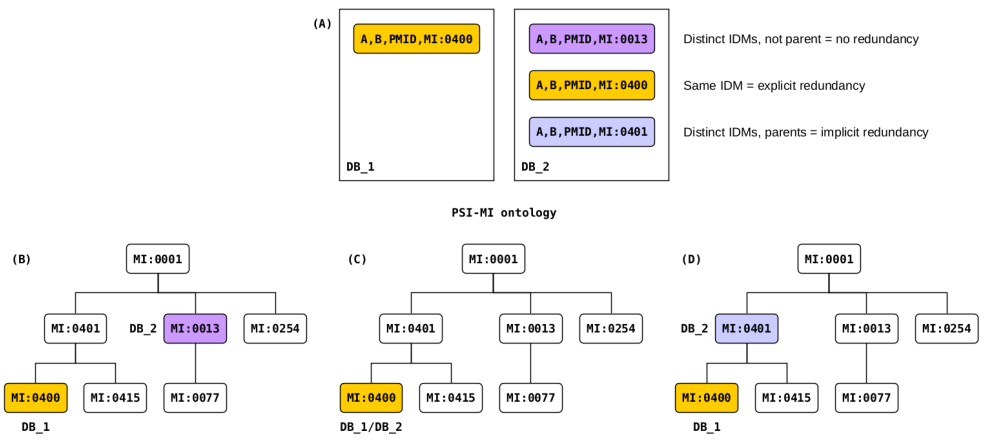

Fig. 1. Illustration of the different types of redundancy across primary databases. (A) Curation events from two databases (`DB_1` and `DB_2`). Depending on the IDM reported by `DB_2`, one can identify no redundancy (purple), explicit redundancy (yellow), or implicit redundancy (blue). Ontology representations of the different cases are presented in panels (B), (C) and (D).

2.2 Definitions

Following the notation introduced in 2.1, we consider two entries, E_i and E_j, of a meta-database, defined by their respective quintuplets of the form $(A, B, M_i, P_x, D_\alpha)$. Note that here we do not consider the experimental role of A and B, therefore all PPIs are symetric and the order of A and B is irrelevant.

E_i and E_j present explicit redundancy if and only if:

$$\begin{cases} E_i = & (A, B, M_i, P_x, D_a) \\ E_j = & (A, B, M_i, P_x, D_b) \end{cases}$$

E_i and E_j present implicit redundancy if and only if:

$$\begin{cases} E_i = & (A, B, M_i, P_x, D_a) \\ E_j = & (A, B, M_j, P_x, D_b) \\ M_i & \textit{is an ancestor of } M_j, \end{cases}$$

where an ancestor can be a direct or an indirect parent.

Ontologies such as PSI-MI (Sivade Dumousseau *et al.*, 2018) can be used to formalize the subsumption relations between IDMs. Note that with the notions provided in section 2.1, explicit and implicit redundancies might be observed among entries originating from different databases (inter-database redundancy, $D_a \neq D_b$) but also from the same database (intra-database redundancy, $D_a = D_b$). We will discuss later (Section 5.3) the meaning of intra-database redundancies, which can correspond either to multiple curation events, but also to variations of an IDM (for example, switching the experimental role (*'bait'* or *'prey'*) of the A and B proteins).

3 Methods

3.1 Source PPI datasets

PPI curation events integrated by APID were downloaded from the APID website on March 23, 2020, last update of APID in January, 2019) for two species (*Homo sapiens* and *Saccharomyces cerevisiae*) in the MITAB25 format (Kerrien *et al.*, 2007). These files aggregate the curated events from five primary databases in a standard format.

In MITAB25 formated data, each line represents a curation event. Interacting proteins are identified by their Uniprot accession numbers. The organism is identified with its NCBI taxonomy identifier. Various information on the experimental evidence is also provided, notably the PMID of the source publication and the PSI-MI code of the IDM used to detect the interaction. Some information such as the direction of the interaction (which protein was used as a *'bait'* and which as a *'prey'*) is not available in this format, but it is usually recorded in primary databases or in more recent MITAB formats (MITAB27). If necessary, missing information might be retrieved using the primary database interaction identifier which is provided and offers full tractability.

3.2 RDF schema and triplestore

The global RDF schema used to integrate all information is presented in Figure 2. It relies on the following ontologies:

- Biological Pathway Exchange (BioPAX) is an ontology developed as a standard for representing molecular interactions, including protein-protein interactions (Demir *et al.*, 2010). We followed the level 3 of the BioPAX specification.
- Proteomics Standards Initiative-Molecular Interactions (PSI-MI) is an ontology edited by the HUPO-PSI. It is dedicated to describe experimental IDMs (Sivade Dumousseau *et al.*, 2018). We used version 1.2.

Raw PPI curation events from the MITAB file were first imported into a MySQL database. A Perl script was used to connect to this database, to exclude curation events that are not considered by APID (see below), and to convert it into a RDF dataset following the BioPAX v3 standard. The resulting interaction data were merged with the PSI-MI ontology, available as an OWL file, into a triplestore powered by the Apache Foundation's JENA suite (v3.14.0). The complete workflow is described in Supplementary Figure S1.

In its integration process, the APID meta-database does not consider curation events annotated with IDMs that do not correspond to a specific experimental method (Alonso-López *et al.*, 2019). To be able to compare our results with APID, we also excluded from our analysis the very same curation events. These are the ones annotated with the IDMs 'molecular interaction', 'interaction detection method', 'biophysical', 'experimental interaction detection', 'inference', 'inferred by author', 'inferred by curator', 'in vitro', 'in vivo', 'unspecified method', or 'phenotype-based detection assay'.

3.3 SPARQL queries

Queries were run using SPARQL Protocol and RDF Query Language (SPARQL). The JENA suite was used to run the SPARQL queries. All queries used to detect redundancies are available in supplementary data

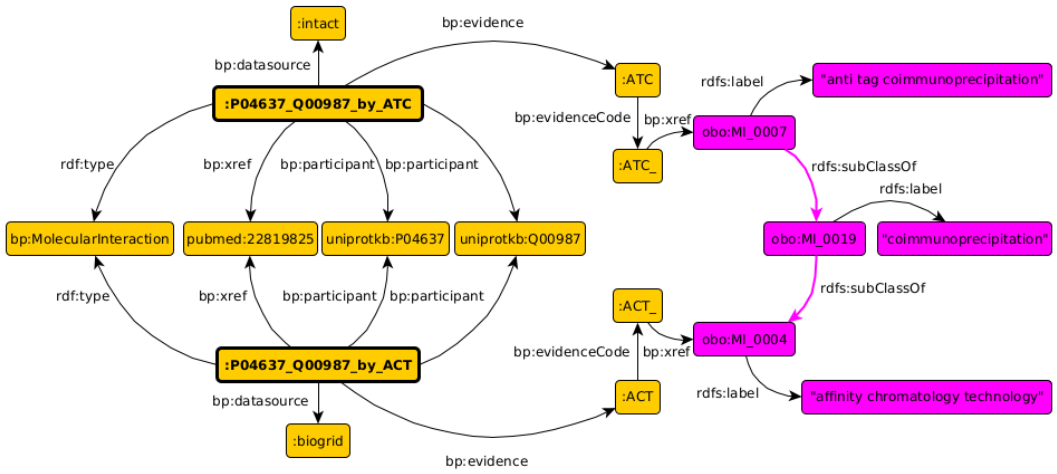

Fig. 2. Scheme representing two curation events reporting the interaction between the ubiquitin ligase MDM2 (UniprotKB: P04637) and the tumor protein 53 (UniprotKB: Q00987) in the BioPAX level 3 ontology (yellow nodes). These two curation events (highlighted in bold) were annotated by different databases (BioGRID and Intact). They refer to the same publication (PMID: 22819825), but different IDMs were used to annotate the interaction ('anti tag coimmunoprecipitation' and 'affinity chromatography technology'). The PSI-MI ontology (purple nodes) reveals that 'affinity chromatography technology' is an ancestor of 'anti tag coimmunoprecipitation', indicating an implicit redundancy between the two curation events.

(Figures S2, S3, S4, S5, S6, S7). As an example, Figure 3 presents the SPARQL query used to detect implicit redundancies in curation events, if one term is an ancestor of the other in the PSI-MI ontology. For each implicit redundancy detected, we conserved only the curation event with the most precise IDM.

```
SELECT DISTINCT ?p1 ?p2 ?pmid ?dm_name1
WHERE {
    ?ppi1 rdf:type bp:MolecularInteraction ;
          bp:participant ?p1, ?p2 ;
          bp:xref ?pmid ;
          bp:evidence ?dm_name1 .
    ?dm_name1 bp:evidenceCode ?m_vocab1 .
    ?m_vocab1 bp:xref ?dm_code1 .
    FILTER ( STR(?p1) < STR(?p2) )
    FILTER NOT EXISTS {
        ?ppi2 rdf:type bp:MolecularInteraction ;
              bp:participant ?p1, ?p2 ;
              bp:xref ?pmid ;
              bp:evidence ?dm_name2 .
        ?dm_name2 bp:evidenceCode ?m_vocab2 .
        ?m_vocab2 bp:xref ?dm_code2 .
        ?dm_code2 rdfs:subClassOf+ ?dm_code1 .
    }
}
```

Fig. 3. SPARQL query to select curation events without explicit nor implicit redundancies. (Note: prefixes are not shown)

3.4 Availability and implementation

The code is available at `https://gitlab.com/nnet56/reproducible-interactome`. The results are available at `https://reproducible-interactome.genouest.org/` and on the Zenodo open data repository (doi:10.5281/zenodo.5595037). The non-redundant interactomes are also accessible on the NDEx platform to facilitate their analysis and manipulation with classical algorithms (doi:10.18119/N94302 (human), doi:10.18119/N97S4D (yeast)).

4 Results

4.1 Overview of analyzed curation events

We analysed the same curation events as the APID database to assess the efficiency of redundancy detection methods. A summary of these curation events is presented in Table 1. The downloaded MITAB files contain $700,484$ curation events for *Homo sapiens* and $305,102$ for *Saccharomyces cerevisiae* (hereinafter referred to as human and yeast, respectively). Together, BioGRID and IntAct represent approximately 85% of all curation events in both species. The contribution of HPRD and BioPlex, restricted to human data, accounts for 13.9% of human curation events. For both species, most PPIs appear in only one or two curation events. PPIs reported by a single curation event represent 49.3% and 60.7% of interacting pairs in human and yeast, respectively.

4.2 Interaction detection methods (IDMs)

The most frequent IDMs in all curation events are listed in Table 1. Among them, 'affinity chromatography technology', 'tandem affinity purification', 'anti tag coimmunoprecipitation' and 'two hybrid' cover more than 58% of human and 76% of yeast curation events. Interestingly, these IDMs include terms with parent–child relationships in the PSI-MI ontology. For example, 'affinity chromatography technology' is a direct ancestor of 'anti tag coimmunoprecipitation'. The presence of such chains is suggestive of possible implicit redundancies between curation events, as defined in sections 2.1 and 2.2.

4.3 Quantification of implicit redundancies

Thanks to the expressiveness of the SPARQL language, we identified both explicit and implicit redundancies among curation events (example query in Figure 3). For constituting a non-redundant dataset, we selected the most precise curation events and discard the redundant and less precise ones since they do not add information.

The occurrence of redundancy among curation events is significant (Table 2). We detected and discarded $73,991$ (11.1%) and $40,266$ (13.7%) implicitly redundant curation events for human and yeast, respectively. Taking into account both explicit and implicit redundancies resulted in removing 30.9% of curation events for human and 35.4% for yeast.

Table 1. Human and yeast curation events (CEs) analysed in this study. Excluded Interaction Detection Methods (IDMs) concern 5.00% ($n = 35,000$) of all curation events in human and 3.87% ($n = 11,809$) in yeast. Only IDMs annotated with a frequency higher than 2% are shown.

Contributing Databases			**Most frequent Interaction Detection Methods**			**Currations events for** (P_a, P_b)		
Databases	**CEs**	**(%)**	**Interaction Detection Methods**	**Counts**	**(%)**	**Occurrences**	**Counts**	**(%)**
Human								
BioGRID	378,910	(54.1%)	Affinity chromatography technology	291,621	(41.63%)	One	161,031	(49.30%)
IntAct	215,577	(30.8%)	Two hybrid	71,969	(10.27%)	Two	91,742	(28.08%)
BIOPLEX!	55,151	(7.9%)	Anti tag coimmunoprecipitation	49,428	(7.06%)	[3-10[	69,015	(21.13%)
HPRD	42,327	(6.0%)	Pull down	42,423	(6.06%)	[10-50[	4,763	(1.46%)
DIP	8,519	(1.2%)	Biochemical	40,544	(5.79%)	≥ 50	113	(0.03%)
			Anti bait coimmunoprecipitation	27,745	(3.96%)			
			In vivo	21,118	(3.01%)			
			Two hybrid array	20,813	(2.97%)			
			Validated two hybrid	14,525	(2.07%)			
Yeast								
BioGRID	133,998	(43.9%)	Affinity chromatography technology	88,681	(29.07%)	One	83,799	(60.73%)
IntAct	130,025	(42.6%)	Tandem affinity purification	84,842	(27.81%)	Two	28,496	(20.65%)
DIP	41,079	(13.5%)	Anti tag coimmunoprecipitation	35,363	(11.59%)	[3-10[	21,792	(15.79%)
			Two hybrid	24,752	(8.11%)	[10-50[	3,799	(2.75%)
			Pull down	13,960	(4.58%)	≥ 50	99	(0.07%)
			Inferred by author	10,894	(3.57%)			
			Protein complementation assay	6,825	(2.24%)			
			Enzymatic study	6,817	(2.23%)			

Table 2. Impact of the removal of both explicit and implicit redundancies on the number of curation events and on the apparent size of the reproducible interactome, for human and yeast. (EEs: Experimental Evidences)

	Human	(%)	**Yeast**	(%)
Curation events				
Initial curation events	665,484	(100%)	293,293	(100%)
Curation events without explicit redundancies	534,140	(80.3%)	229,630	(78.3%)
Curation events without explicit and implicit redundancies	460,149	(69.1%)	189,364	(64.6%)
Apparent size of the reproducible interactome (PPIs supported by ≥ 2 EEs)				
Initial	159,192	(100%)	52,313	(100%)
Without explicit redundancies	111,009	(69,7%)	40,235	(76.9%)
Without explicit and implicit redundancies	70,554	(44,3%)	21,311	(40.7%)

Importantly, detection of redundancy between curation events has a strong impact on the apparent size of the reproducible interactome (i.e PPIs supported by at least two experimental evidences) (Table 2, Supplementary Figures S8 and S9). For human, the reproducible interactome drops from $159,192$ to $70,554$ PPIs (-55.7%: -30.3% due to explicit redundancies and -25.4% due to implicit ones). For yeast, the impact of redundancies is even worse, with a drop of the reproducible interactome from 52,313 PPIs to 21,311 after removal of both explicit and implicit redundancies (-59.3%: -23.1% due to explicit redundancies and -36.2% due to implicit ones). In other words, for human, discarding 11.1% of implicitly redundant curation events accounts for reducing by 25.4% the reproducible interactome. Similarly, for yeast, discarding 13.7% of implicitly redundant curation events accounts for reducing by 36.2% the reproducible interactome.

4.4 Implicit redundancies mostly result from the integration of the different primary databases

We then investigated whether implicit redundancy was already present in source databases (intra-database redundancy), or if it was a consequence of the integration of different source databases (inter-database redundancy). The vast majority originates from inter-database redundancies for both human (91.1%) and yeast (95.0%) (see Supplementary Tables S1 and S2). The couple of databases that generates the largest part of the implicit redundancies is BioGRID and IntAct. This is consistent with the fact that BioGRID and IntAct are the two most contributing source databases. Intra-database redundancies will be further discussed in section 5.3.

4.5 Frequently redundant identification methods

We computed the frequency of the pairs of detection methods involved in implicit redundancies. For human, the most frequent implicitly redundant couples of IDMs and their parent-child relationships in the PSI-MI ontology are displayed in Figure 4.

The most frequent couple is 'affinity chromatography technology' and 'anti tag coimmunoprecipition', which is responsible for $25,333$ redundancies. The term 'affinity chromatography technology' is also frequently observed with other descendants such as "pull down" ($n = 9,896$), 'anti bait coimmunoprecipitation' ($n = 6,617$), or "tandem affinity purification" ($n = 5,968$). Two-hydrid techniques are also introducing redundancies, for example with 'two hybrid', and its descendants 'two hybrid array' ($n = 16,113$), 'two hybrid prey polling approach' ($n = 11,238$), 'validated two hybrid' ($n = 11,123$), or 'two hybrid pooling approach' ($n = 10,713$). A similar situation is observed in yeast (the complete list of implicit redundancies for both human and yeast is available as Supplementary Tables S3 and S4). Implicit redundancies are thus widespread all along the PSI-MI ontology, and not limited to binary IDMs. This highlights the need for a general approach to reconcile

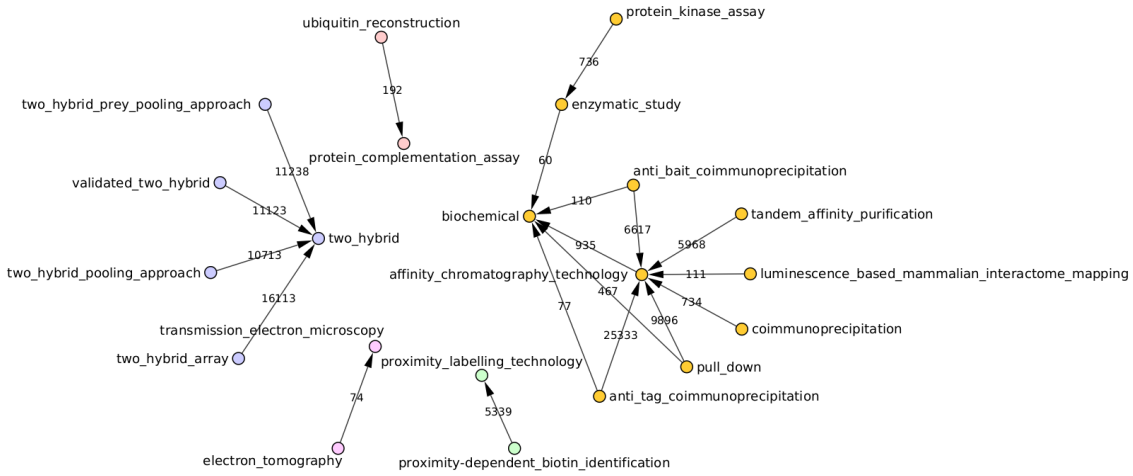

Fig. 4. Couples of related interaction detection methods (IDMs) from the PSI-MI ontology frequently identified in implicit redundancies of human PPIs. The arrows connect the most specific to the most general term according to the PSI-MI ontology. Only implicit redundancies with at least 50 occurrences are shown. Nodes connected to a common IDM are represented with the same color.

curation events during the integration of multiple primary databases.

The fact that implicit redundancies are observed between very different terms of the PSI-MI ontology suggests that different primary databases have different policies for annotating IDMs, as previously noted for IntAct and BioGRID (Alonso-López *et al.*, 2019). We therefore further analysed the IDMs used by each primary database.

We observed that IntAct and DIP use a wide range of IDMs for both human and yeast PPIs (165 for IntAct and 89 for DIP) while BioGRID, HPRD and BioPlex use much fewer (12, 3 and 1 IDMs, respectively) and more general IDMs. Hence, the strong discrepancies in database annotation policies are the source of inter-database implicit redundancies.

Overall, we observed that implicit redundancy (i) occurs between a wide range of the PSI-MI ontology terms, regardless of the species, (ii) mostly results from the integration of different primary databases with different annotation policies, and (iii) happens for all database combinations.

5 Discussion

The construction of a reliable interactome demands to combine interaction data produced by several independent experimental evidences and IDMs in order to reduce false positives. Since experimental evidences are curated and stored in several primary databases, a unification of these databases is required. The Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI) developed the PSICQUIC specification and web services that facilitate data retrieval from multiple databases and assist their integration but do not elaborate on redundancy detection (del Toro *et al.*, 2013). In several (meta-) databases, PPIs are annotated with a confidence score, which is calculated using the number of independent experimental evidences and the nature of IDMs (Villaveces *et al.*, 2015). To be relevant, these algorithmic require reliable, non-redundant, datasets of experimental evidences. Therefore, several primary databases have decided to coordinate their curation efforts in the frame of the IMEx consortium in order to provide a single non-redundant set of homogeneously annotated protein interaction data (Orchard *et al.*, 2012; Porras *et al.*, 2020).

Here, we propose a formalisation of both explicit and implicit redundancy between experimental evidence entries in order to integrate PPIs from any database that uses the PSI-MI ontology. Knowledge about IDMs is extracted from the PSI-MI ontology, while the method to identify redundancies is based on Semantic Web technologies.

5.1 The Semantic Web is adapted for identifying implicit redundancies

Alonso-López *et al.* (2019) pointed two problems related to redundancy identification: (i) there may be a parent-child relationship between IDM terms, and (ii) the path from a child term to its ancestors may not be unique due to multiple inheritance. We propose the notion of **implicit redundancy** to address the logical implications of two database entries describing the interaction of the same protein pair with IDMs that have a descendant-ancestor relationship. The Semantic Web is designed to perform integrated reasoning on data annotations and ontologies. In particular, it makes handling simple and multiple hierarchies straightforward. In the raw data of APID that aggregates BioGRID, IntAct, HPRD, BioPlex and DIP, we were able to identify both explicit and implicit redundancies. Our work reveals that implicit redundancies are a widespread phenomenon resulting from the different curation choices of the various databases and that it is of similar importance than explicit redundancies. Therefore, we demonstrated the relevance of both the notion of implicit redundancy and of the choice of the Semantic Web as a technical framework for addressing the redundancy identification problem. Moreover, new explicit and implicit redundancies will continue to occur over the natural updates of the various databases.

The PSI-MI ontology that describes the IDMs is evolving. For example, during the time of our project, we noticed that the term 'three hybrid', which was initially a child of the term 'two hybrid', is now a child of 'transcriptional complementation assay'. This modification is highly relevant since 'two hybrid' is a binary identification method, whereas 'three hybrid' is not, and having a non-binary identification method as a direct child of a binary one was not consistent. Therefore, just like the databases are regularly updated, the ontologies are also corrected and enriched, which also has an incidence on redundancies. By allowing to automate redundancy detection as the integration of databases scales up, the Semantic Web facilitates the reliable interpretation of the results in the perspective of the construction of a reproducible interactome.

5.2 Widespread inter-databases implicit redundancies

Implicit redundancies primarily arise from the integration of different databases (91.1% and 95.0% of inter-database redundancies for human

and yeast, respectively). In our study, we clearly highlight that this is due to the granularity of IDMs used in the primary databases. Indeed, while some databases like IntAct refer to numerous detailed terms from the PSI-MI ontology (165 and 89 terms used to annotate human and yeast PPIs, respectively), other databases like BioGRID merely use general and high level terms (only 12 terms used for both human and yeast).

Therefore, if the integration of different PPI databases is necessary to better cover the interactome, a particular attention has to be paid to detect the widespread inter-database implicit redundancies. A simple method could be to define priorities between databases depending on whether they use precise or general terms to annotate PPIs. In case of multiple curations events referring to the same proteins and the same PMID, the ones from the database with the highest priority would be selected. However, this would be an approximate approach whereas we propose an exact solution, robust to possible changes of annotation policy by primary databases.

Primary databases of the IMEx consortium coordinate and share their curation efforts to produce a non-redundant dataset of PPI experimental evidences (Orchard *et al.*, 2012). IMEx members use common curation rules to harmonize their annotation process. The unicity of the curation events is ensured by allowing PPIs from a given PMID to be annotated only once, and all data are centralized in IntAct. Both this work from the IMEx consortium and ours emphasize the need for a general approach to assemble non-redundant PPI datasets.

5.3 Intra-database redundancies

Our analysis also identified a significant number of apparently redundant curation events within primary databases (Supplementary Figures S8 and S9). Such intra-database redundancy may originate from multiple independent annotations of identical experimental evidences within primary databases, as noted by Alonso-López *et al.* (2019). Yet, further inspection of such curation events indicates that intra-database redundancy primarily occurs when independent experiments from the same publication have been annotated in a given database with identical or related IDMs, leading to apparent explicit or implicit intra-database redundancies. For instance, we observed that the vast majority of the explicit intra-database redundancies originating from BioGRID are due to PPIs probed with both partners as baits and preys (6229 out of 8696 explicit redundancies involving exactly two curation events for yeast and 12283 out of 15385 for human). Intra-database redundancy can also occur when a PPI has been identified with a high-throughput experiment and then validated using the same or a related method performed at low-throughput. Hence, this currently leads to the unification of curation events that actually report distinct experimental evidences. To correct this, our method could be extended by taking into account additional information, such as the experimental role of each protein.

5.4 Towards a reproducible interactome

The size of the reproducible interactome is drastically impacted by removing redundancies for both human (-55.7%) and yeast (-59.3%), and we show that this is largely due to implicit redundancies. Indeed, we observe that filtering the curation events involved in implicit redundancy (11 to 14 %) leads to a drastic (25 to 36 %) reduction of the apparently reproducible interactome. This implies that a large number of PPIs currently considered as reproducible actually relies on integration artefacts. Thus, more experimental data are still needed to further improve the size and confidence level of the reproducible interactome. Information on PPIs that have not yet been reproduced can help to prioritize such experiments. Knowledge-based methods as presented in this article will be necessary to support the integration of the continuously increasing experimental evidences and publications.

Acknowledgements

The GenOuest platform provided computational support and Web hosting.

Funding

This work has been supported by Univ Rennes with a Defi Emergent 2019 grant to EB and GR.

References

Alonso-López, D. *et al.* (2016). APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. *Nucleic Acids Res*, **44**(W1), W529–535.

Alonso-López, D. *et al.* (2019). APID database: redefining protein–protein interaction experimental evidences and binary interactomes. *Database*, **2019**. baz005.

Bonetta, L. (2010). Interactome under construction. *Nature*, **468**(7325), 851–852.

Cafarelli, T. *et al.* (2017). Mapping, modeling, and characterization of protein–protein interactions on a proteomic scale. *Current Opinion in Structural Biology*, **44**, 201–210.

Calderone, A. *et al.* (2013). mentha: a resource for browsing integrated protein-interaction networks. *Nat Methods*, **10**(8), 690–691.

Calderone, A. *et al.* (2020). Using the MINT Database to Search Protein Interactions. *Curr Protoc Bioinformatics*, **69**(1), e93.

Das, J. and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst Biol*, **6**, 92.

del Toro, N. *et al.* (2013). A new reference implementation of the PSICQUIC web service. *Nucleic Acids Res*, **41**(Web Server issue), W601–606.

Demir, E. *et al.* (2010). The BioPAX community standard for pathway data sharing. *Nature biotechnology*, **28**(9), 935–942.

Huttlin, E. L. *et al.* (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, **184**(11), 3022–3040.

Kerrien, S. *et al.* (2007). Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol*, **5**, 44.

Kerrien, S. *et al.* (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Res*, **40**(Database issue), D841–846.

Keshava Prasad, T. S. *et al.* (2009). Human Protein Reference Database–2009 update. *Nucleic Acids Res*, **37**(Database issue), D767–772.

Klapa, M. I. *et al.* (2013). Reconstruction of the experimentally supported human protein interactome: what can we learn? *BMC systems biology*, **7**, 96.

Kotlyar, M. *et al.* (2019). IID 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res*, **47**(D1), D581–D589.

Luck, K. *et al.* (2020). A reference map of the human binary protein interactome. *Nature*, **580**(7803), 402–408.

Orchard, S. *et al.* (2012). Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods*, **9**(4), 345–350.

Oughtred, R. *et al.* (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Res*, **47**(D1), D529–D541.

Porras, P. *et al.* (2020). Towards a unified open access dataset of molecular interactions. *Nature communications*, **11**(1), 6144.

Salwinski, L. *et al.* (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, **32**(Database issue), D449–451.

Sivade Dumousseau, M. *et al.* (2018). Encompassing new use cases - level 3.0 of the HUPO-PSI format for molecular interactions. *BMC Bioinformatics*, **19**(1), 134.

Szklarczyk, D. *et al.* (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*, **47**(D1), D607–D613.

Turinsky, A. L. *et al.* (2010). Literature curation of protein interactions: measuring agreement across major public databases. *Database*, **2010**. baq026.

Villaveces, J. M. *et al.* (2015). Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. *Database*, **2015**. bau131.