



HAL
open science

A method to identify target molecules and extract the corresponding graph of interactions in BioPAX

Camille Juigné, Olivier Dameron, Florence Gondret, Emmanuelle Becker

► To cite this version:

Camille Juigné, Olivier Dameron, Florence Gondret, Emmanuelle Becker. A method to identify target molecules and extract the corresponding graph of interactions in BioPAX. BBCC2022 - Bioinformatics and Computational Biology Conference, Dec 2022, Virtual, Italy. hal-03876091

HAL Id: hal-03876091

<https://inria.hal.science/hal-03876091>

Submitted on 15 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A method to identify target molecules and extract the corresponding graph of interactions in BioPAX

Camille JUIGNÉ^{1,2} Olivier DAMERON¹ Florence GONDRET²
Emmanuelle BECKER¹

(1) Univ Rennes, Inria, CNRS, IRISA - UMR 6074, F-35000 Rennes, France

(2) PEGASE, INRAE, Institut Agro, F-35590 Saint Gilles, France

General context: From single omics to multi omics analysis

- High-throughput techniques generate a large quantity of data
- Each modality is analyzed statistically, independently from the others
- The modalities are not independent

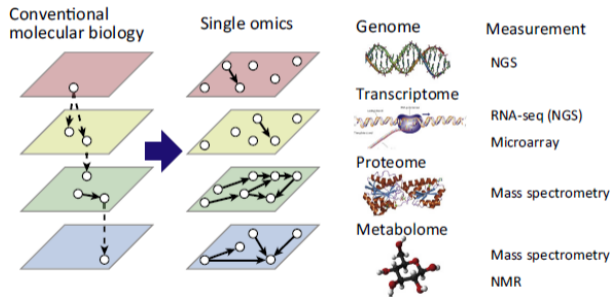


Fig. Linking the different levels of biological organization allows for a holistic view of biological entities (source: K. Yuri et al.)

Considering the different levels of omics as a whole will help to understand biological systems

General context: From single omics to multi omics analysis

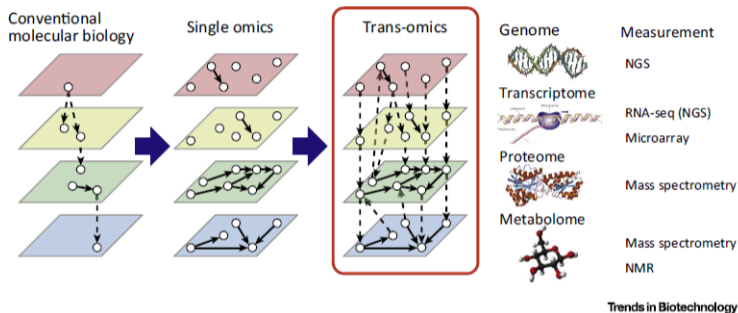


Fig. Linking the different levels of biological organization allows for a holistic view of biological entities (source: K. Yuri et al.)

This systemic representation may provide a better knowledge of:

- the cascade of events
- the upstream regulators
- the complexity of biological processes

General context: Better understanding of a phenotype

Starting hypothesis

Key drivers of the phenotypic divergence can be better defined by

- considering the different levels of organization between biological entities
- integrating experimental data and knowledge bases

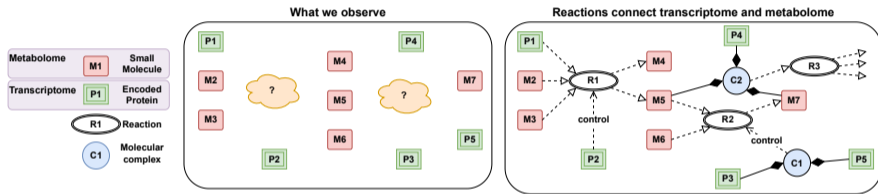
In this situation, we have

- transcriptomic experimental data
- metabolomic experimental data

To provide an holistic view, it is necessary to obtain an extensive description of where and how these molecules participate and interact with each others in biological pathways

General context: Better understanding of a phenotype

These -omic levels can be linked to each other by interactions through the proteomic level



Knowledge about interactions provide the structure that underlies the dependencies between modalities

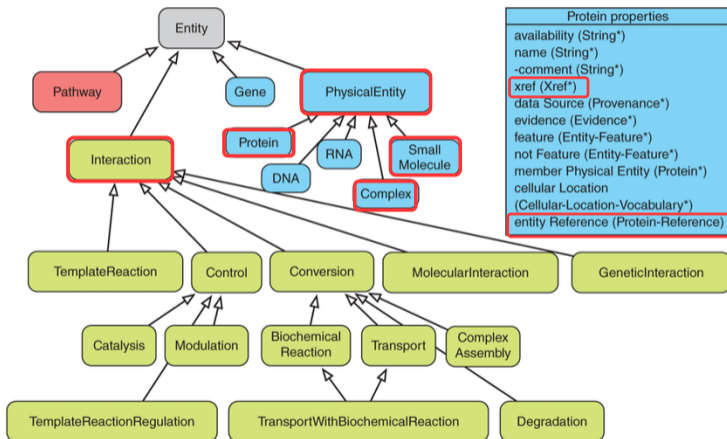
Our contribution

A methodology to map the results of high-throughput data obtained at different omics levels on a graph representing metabolism

Biological Pathway Exchange format (BioPAX)

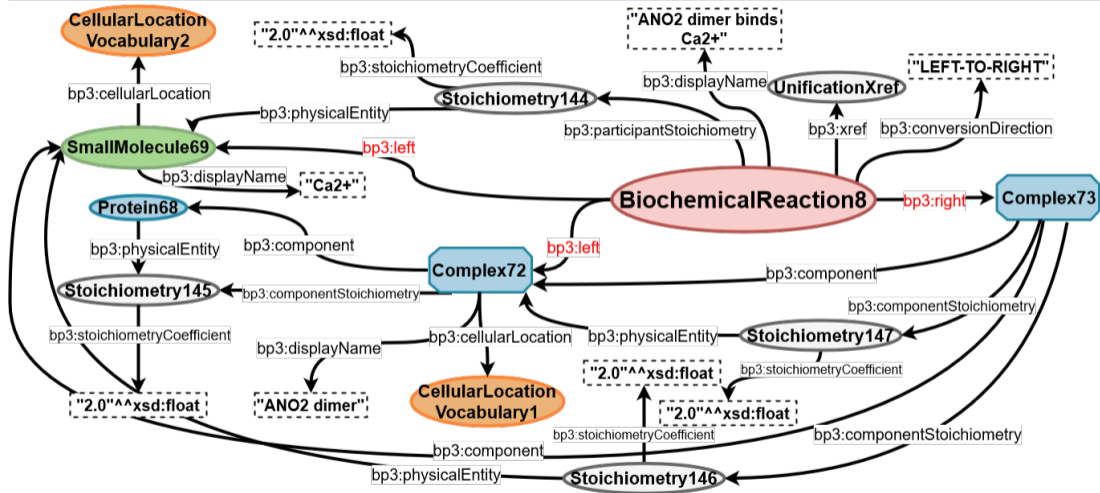
Database of biological pathways in BioPAX

- **Reactome, KEGG, PathwayCommons...**
- Well established **ontology** to represent pathways at molecular and cellular levels
- Represented in **graphs** (RDF and OWL)
- Can be queried with **SPARQL**
- Can be **mapped with other resources** such as ChEBI, UniProt, GO...



Demir et al. (2010)

BioPAX: Example $Ca^{2+} + ANO2 \rightarrow ANO2 : Ca^{2+}$



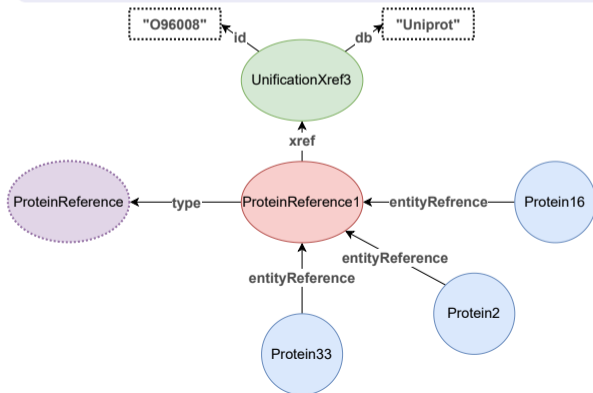
The complexity of BioPAX reflects the complexity of biological reality

BioPAX: EntityReference Class (utility class)

Definition

An entity reference is a **grouping of several physical entities** across different contexts and molecular states, that share common physical properties

Physical entities are linked to entity reference by the `entityReference` property

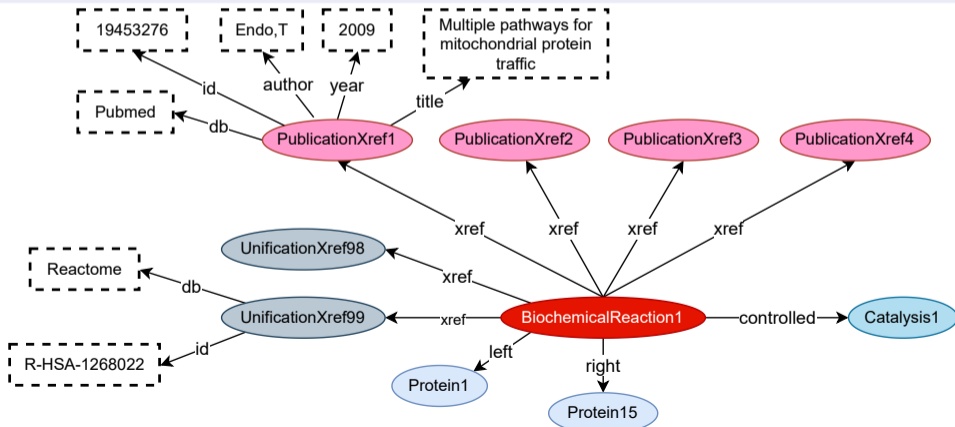


For proteins and small molecules, there is a corresponding node where all the non-changing aspects of the entity are stored

BioPAX: Xref Class (utility class)

Definition

A reference from an instance of a class to an object in an external resource (Physical) entities are linked to xref by the xref property



BioPAX: Interaction class

Definition

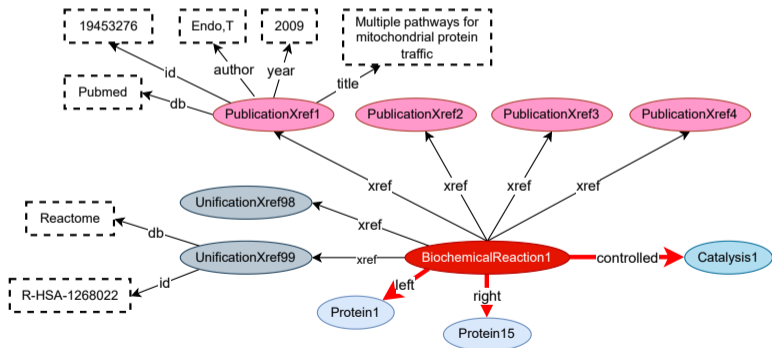
A biological relationship between two or more entities.

Subclasses

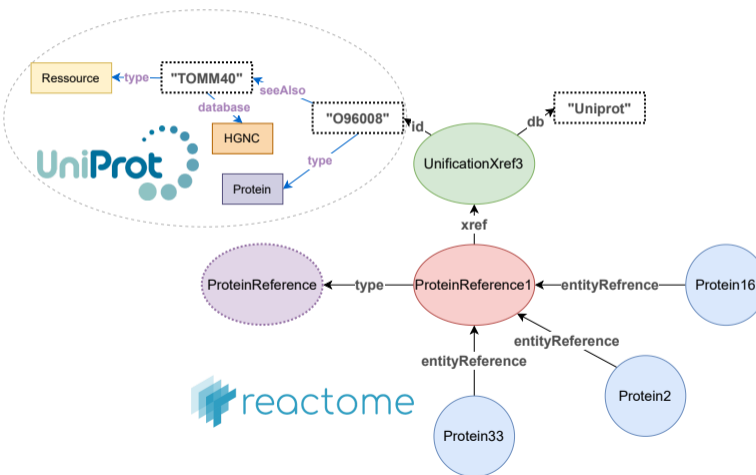
Control, Conversion, GeneticInteraction, MolecularInteraction, TemplateReaction

The participant property

Multiple sub-properties:
left, *right*, *controller*, *controlled*



Methods - Retrieving Proteins from the Entity and Xref classes in the Reactome database



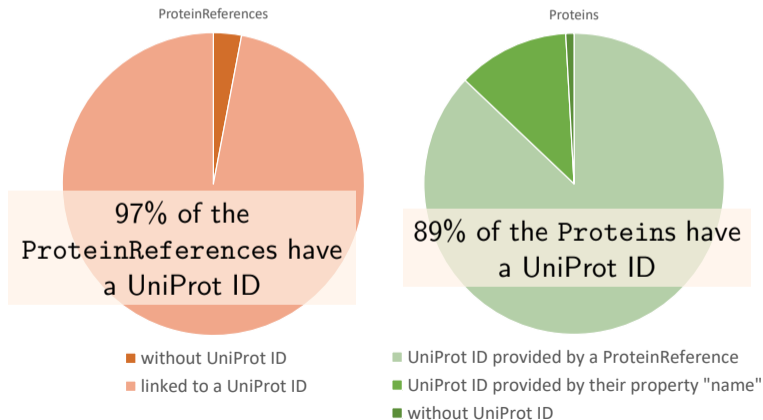
Federated SPARQL query

1. from a list of HGNC IDs, identify the corresponding UniProt IDs (UniProt SPARQL endpoint)
2. from a list of UniProt IDs, locate the corresponding ProteinReferences
3. from these ProteinReferences, identify all the associated Proteins

Results - Proteins linked to a UniProt ID

Proteins and ProteinReferences in Reactome (h. sapiens v81)

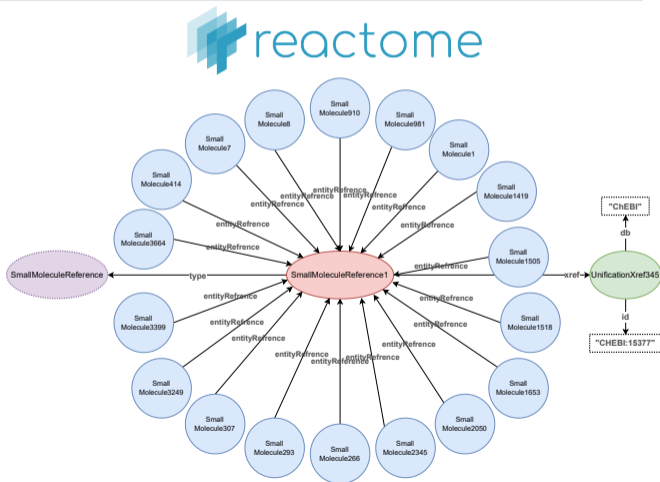
11,685 ProteinReferences and 31,755 Proteins



Methods - Retrieving SmallMolecules from Entity and Xref classes in the Reactome database

Federated SPARQL query

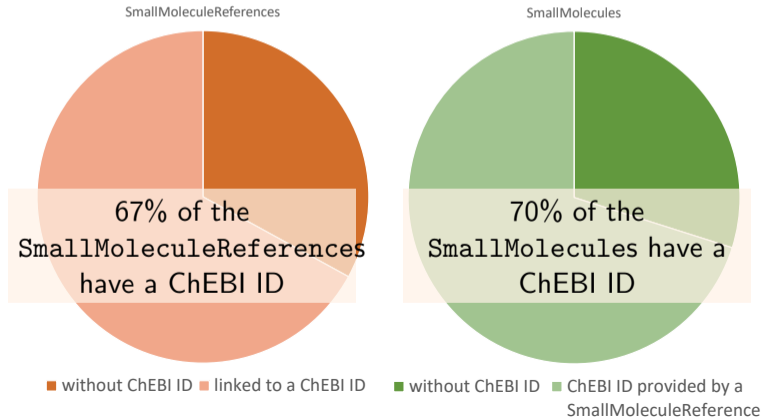
1. (*in progress*) identify the target molecules in the ChEBI ontology (ChEBI SPARQL endpoint)
2. from a list of ChEBI IDs, locate the corresponding SmallMoleculeReferences
3. from these SmallMoleculeReferences, identify all the associated SmallMolecules



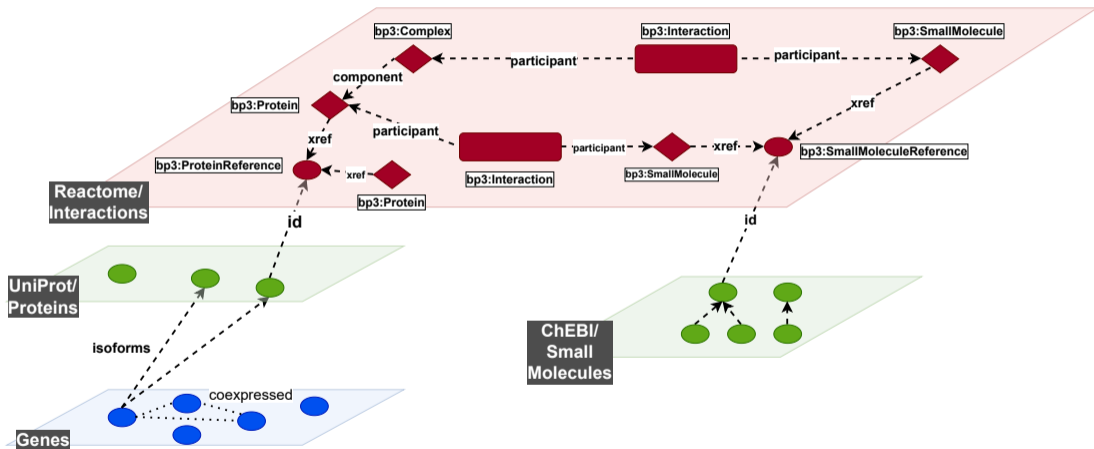
Results - SmallMolecules linked to a ChEBI ID

SmallMolecules and SmallMoleculesReferences in Reactome (h. sapiens v81)

2,878 SmallMoleculeReferences and 5,049 SmallMolecules



Methods - Integrated data schema



Methods - Identifying the Interactions in which target molecules are involved

22,237 Interactions having from 1 to 62 participants (median of 2)

Type of interaction participants	Total number of participants
Complex	21,896
SmallMolecule	17,000
Protein	14,197
BiochemicalReaction	8,279
PhysicalEntity	1,900
Dna	1,361
Rna	310
Degradation	7
TemplateReaction	7

We expect that for a majority of the Interactions, it is possible to identify the participants by their UniProt or ChEBI IDs (or by those of their components)

Methods - Identifying and extraction of subgraphs of Interactions involving target molecules

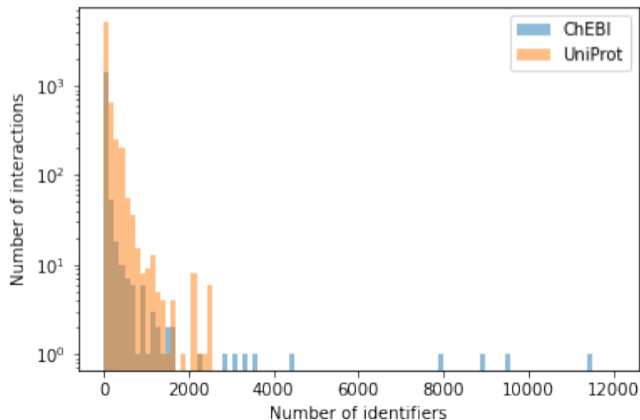
→ All targeted physical entities have been located

SPARQL query

- From a list of identifiers, identifies all the corresponding physical entities in the graph
- Identifies all the interactions in which these entities participate
 - as a direct participant
 - as a component of a complex
- To export the interactions subgraph, retrieves all properties related to these interactions

Results - Extraction of subgraphs of Interactions involving target molecules

Distribution of the number of interactions
a ChEBI or UniProt ID is involved



We confirm that 85% of the 22,237 Interactions have at least one participant having a UniProt or a ChEBI ID

Contributions

A method and its implementation

- to integrate simultaneously metabolomic, proteomic and transcriptomic data
- to extract subgraphs of interest from BioPAX databases...
- ... enriched with knowledge bases (UniProt, ChEBI)

It underlines the importance

- of developing and using tools with such semantic richness
- to step up the efforts to **link the different ontologies and databases** (systematically using universal identifiers)

Perspectives

- This opens new perspectives to **find networks** between molecules at different levels of cell organisation
- This integrative approach should allow to better define **regulators**

Acknowledgments



UMR **PEGASE**



contact : camille.juigne@irisa.fr
github : [cjuigne/data_integration_biopax](https://github.com/cjuigne/data_integration_biopax)