Detection and correction of non-conformities and redundancies in complexes of molecules in BioPAX

Camille Juigné\textsuperscript{1,2} Olivier Dameron\textsuperscript{1} François Moreeew\textsuperscript{1,2} Florence Gondret\textsuperscript{2} Emmanuelle Becker\textsuperscript{1}

\textsuperscript{1}Univ Rennes, Inria, CNRS, IRISA - UMR 6074, F-35000 Rennes, France

\textsuperscript{2}PEGASE, INRAE, Institut Agro, 35590, Saint Gilles, France

July 5, 2022
Biological context: understand how biological systems adapt to their environment

Understand the organization of biological pathways at different scales
Biological context: metabolic pathways and complexes

Biological pathway

"a series of actions among molecules in a cell that leads to a certain product or a change in the cell" (NIH)
Biological pathway

"a series of actions among molecules in a cell that leads to a certain product or a change in the cell" (NIH)

Complexes and interactions in biology

- Chemical assembly of several molecules
- Can either participate in or control interactions
Biological context: complexes play a major role in pathways

What we consider

Reactions connect transcriptome and metabolome

Complexes are key participants that hide their components
Impact on the graph topology

What we consider

Reactions connect transcriptome and metabolome

Complexes are key participants that hide their components

Taking into account (in)valid complexes is required for analyzing:

- Interactions in which a molecule participates
- Molecules participating in an interaction
Computational context: Biological Pathway Exchange format

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

Demir et al. (2010)
Computational context: Biological Pathway Exchange format

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

Demir et al. (2010)

In the Reactome database complexes participate to 32% of the interactions
Issue: Complexes composed of other complexes

A complex cannot be composed of other complexes

The components of a complex cannot have a component
Issue: Complexes composed of other complexes

Complexes

Complex that has at least one component

Complex

Component 1
Component i
Component n

Complex that has no component (black-box complex)

Black-box Complex

Complex that has at least one component, none of which is a complex

Complex

Component 1
Not a complex
Component i
Not a complex
Component n
Not a complex
We observed some invalid complexes in Reactome (not detected by the BioPAX validator).
Objectives

1. Identify invalid complexes
2. Fix invalid complexes
3. Evaluation of the benefits of the procedure
Contrib 1: Identify and quantify invalid complexes

Invalid complexes are composed of $\geq 1$ complex with components

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX bp3: <http://www.biopax.org/release/biopax-level3.owl#>

SELECT DISTINCT ?invalidComplex
WHERE {
  ?invalidComplex rdf:type bp3:Complex .
}
```

Use case

- Homo sapiens: 39% complexes are invalid out of 14,840
- Mus musculus: 39% complexes are invalid out of 10,761
- Sus scrofa: 40% complexes are invalid out of 7,769

Complexes represent a large fraction of biological entities. Invalid complexes are present in large quantities in the data sets of different organisms.
Contrib 1: Identify and quantify invalid complexes

Invalid complexes are composed of $\geq 1$ complex with components

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX bp3: <http://www.biopax.org/release/biopax-level3.owl#>

SELECT DISTINCT ?invalidComplex
WHERE {
  ?invalidComplex rdf:type bp3:Complex .
}

---

Use case (v79)

**Homo sapiens**: 39% complexes are invalid out of 14,840
**Mus musculus**: 39% complexes are invalid out of 10,761
**Sus scrofa**: 40% complexes are invalid out of 7,769

Complexes represent a large fraction of biological entities
Invalid complexes are present in large quantities in the data sets of different organisms
Contrib 2: Fix the invalid complexes

How it should be

How it is

Collapse as direct components all the (in)direct components that do not have component
Contrib 2: Fix the invalid complexes

Stoichiometry has to accommodate the fact that components can occur at several places.

\[ S(Z) = \sum_{p \in \text{parent nodes}} S_p(Z) \cdot S(p) \]
Contrib 3: Homo sapiens Reactome use-case (repair)

All invalid complexes were fixed

Fixing invalid complexes increases the number of direct components
Number of molecules that are not complexes and participate in interactions at a distance of 1 or 2

DIST $\leq$ 2, COUNT: 4
MISSED: 2

2 molecules missed due to invalid complexes
Impact on the graph topology

Taking into account invalid complexes has a strong impact on the interaction graph topology
Side effect: detection of artificial redundancy (Homo Sapiens)

What we call redundant complexes:
Complexes that share the same components with the same stoichiometric coefficients and have the same cellular location

Before fixing, we identified 241 of these complexes
Side effect: detection of artificial redundancy (Homo Sapiens)

What we call redundant complexes:
Complexes that share the **same components** with the **same stoichiometric coefficients** and have the **same cellular location**

Before fixing, we identified **241** of these complexes

Fixing invalid complexes allowed to identify **92** additional redundant complexes (**+38%**)
Conclusion

• Semantically-rich queries for identifying and fixing invalid complexes that are reproducible on other databases
• Improves the conformity and the analysis of the graph by repairing the topology
• Will allow to apply reasoning methods on better quality data
• Will allow a better understanding of the regulation of complex phenotypes
• Side effect of allowing the detection of complex redundancies
• Essential methodology to analyse and advance in the knowledge of biological processes
Acknowledgments

Team Croissance

camille.juigne@irisa.fr
Issue: Complexes composed of other complexes

Complex component: defines the subunits of a complex (BioPAX v3 spec)

"This property should not contain other complexes, i.e. it should always be a flat representation of the complex. [...] Exceptions are black-box complexes (i.e. complexes in which the component property is empty), which may be used as component’s of other complexes because their parts are unknown."

We observed some invalid complexes in Reactome (not detected by the BioPAX validator)