Integration and analysis of heterogeneous biological data through multilayer graph exploitation to gain deeper insights into feed efficiency variations in growing pigs

Camille Juigné

Composition du Jury :

Président :	**Mathieu EMILY**, Professeur, Institut Agro Rennes-Angers
Examinateur :	**Michel DUMONTIER**, Distinguished Professor, Maastricht University
Rapportrice et rapporteur :	**Andrea RAU**, Directrice de recherche, INRAE
	Fabien JOURDAN, Directeur de recherche, INRAE
Dir. de thèse :	**Florence GONDRET**, Directrice de recherche, INRAE
Co-dir. de thèse :	**Emmanuelle BECKER**, Professeure, Université de Rennes

L'INSTITUT agro Rennes Angers INRAE UMR PEGASE UMR IRISA Inria Rannvro BREIZH

Thesis general problem: Understand a complex phenotype through heterogeneous biological data

- To enhance our understanding of the complex biological phenomenon
  use case: feed efficiency
- Integration of heterogeneous data = linking data about various types of entities
- Through a computational method = Semantic Web and multilayer graphs

Use case: Feed efficiency in growing pigs

Feed efficiency

- The ability of pigs to turn feed nutriments into lean growth rate
  $\rightarrow$ while maintaining physiological functions and health
  $\rightarrow$ by reducing effluent discharge

Why is this biological question important?

- Feed represents between 60 and 70 % of the total cost of pork production
- Pig production is facing several issues related to competition with feed resources, and competitiveness due to global trade
- The increase in size of pig farm led to environmental issues related to storage, treatment and use of effluents

The need to get deeper insights into feed efficiency variations in growing pigs

↗ Feed efficiency

- A research priority to support sustainable meat production
- But a complex trait that integrates multiple biological pathways orchestrated in and by various tissues

Primary avenues for exploration

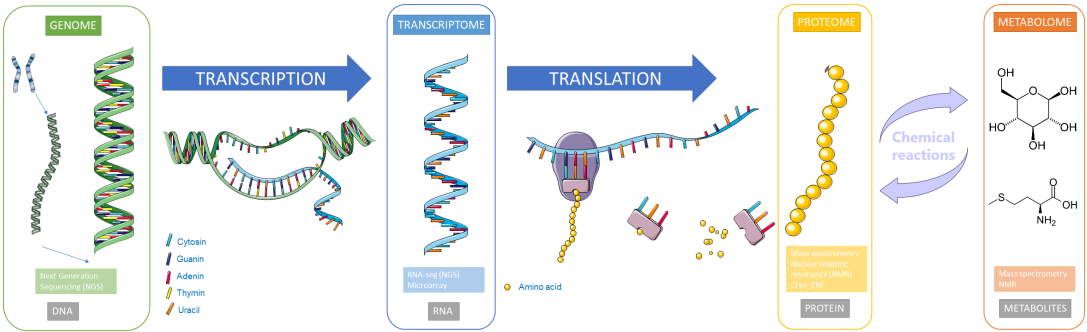
- -omics technologies: produce large amount of data without *a priori*
- blood samples: minimally invasive way to summarize the activities of various tissues within the body

Experimental biology for a better understanding of life

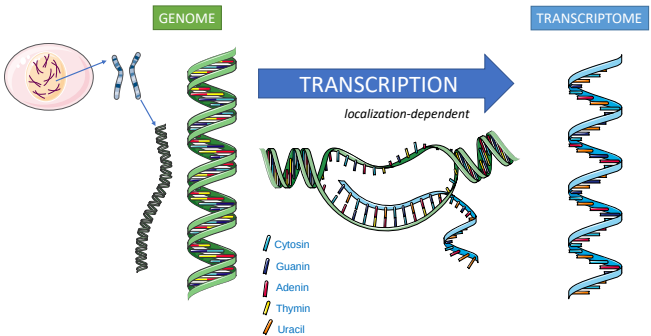
- Describing and understanding the biological mechanisms
- Investigate the different biological entities

Experimental biology for a better understanding of life

- Describing and understanding the biological mechanisms
- Investigate the different biological entities



Transcriptomics



Methods:

- Micro-arrays
- RNA-seq

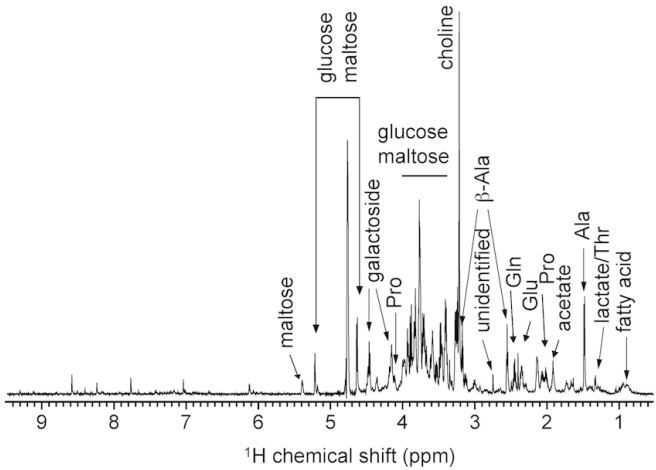
Data type:

- Gene expression level, transcript abundance (quantitative)

Analyses:

- Differential gene expression
- Functional enrichment
- Gene co-expression network

Metabolomics



Methods:

- Nuclear Magnetic Resonance
- Liquid Chromatography - Mass Spectrometry

Data types:

- Types and concentrations of metabolites (quantitative)
- Presence/absence (binary)

Analyses:

- Groups differentiation
- Biomarkers identification
- Assessing changes in the metabolic profile

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

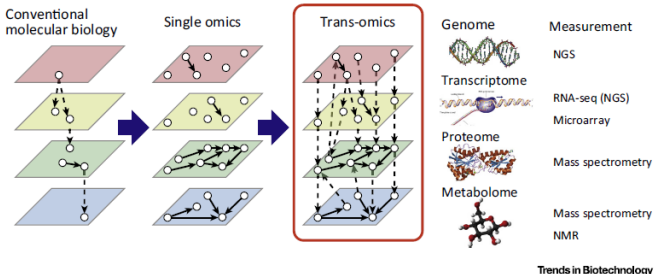


Fig. Linking the different levels of biological organization allows for a holistic view of biological entities (source: K. Yuri et al.)

Considering different levels of -omics **as a whole** will help to understand biological systems, especially by considering the cascade of events and the interactions between entities

Inherent heterogeneities in biological data

Heterogeneity of entity type

distinct biological entities: genes, transcripts, proteins, etc.

Heterogeneity of data type

in terms of the nature of the data itself: textual, binary, quantitative or qualitative

Technical heterogeneity

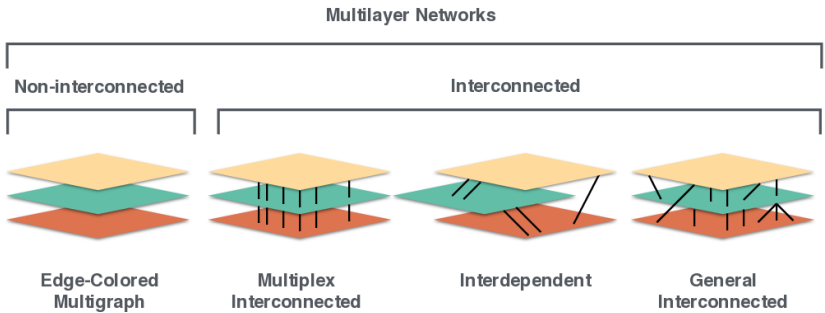
variations in measurement techniques, experimental protocols, and data formats

While it makes logical sense to consider biological data as a whole with interconnected elements, **the process of integrating these data is far from trivial**

Strategy: A comprehensive and systemic integration approach

Strategy adopted: systemic network-based integration

- Relationships between entities are preserved, allowing a holistic view
- Graphical representation facilitates understanding of relationships between data
- Adaptability to changes and addition of new data sources

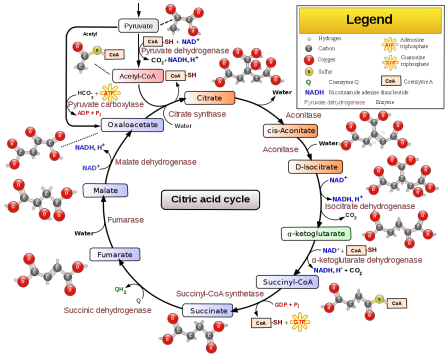


Manlio De Domenico, "Multilayer Networks Illustrated" (2020)

-Omic levels can be linked to each other by interactions

Biological pathway

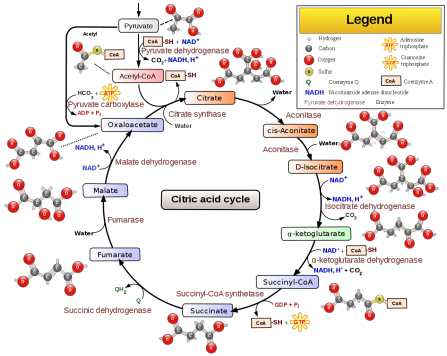
"*a* **series of actions** *among molecules in a cell that leads to a certain* **product or a change** *in the cell*" (NIH)

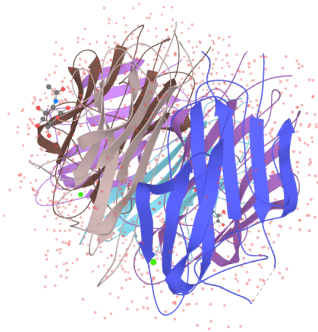


-Omic levels can be linked to each other by interactions

Biological pathway

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





R-HSA-173584

Complexes and interactions in biology

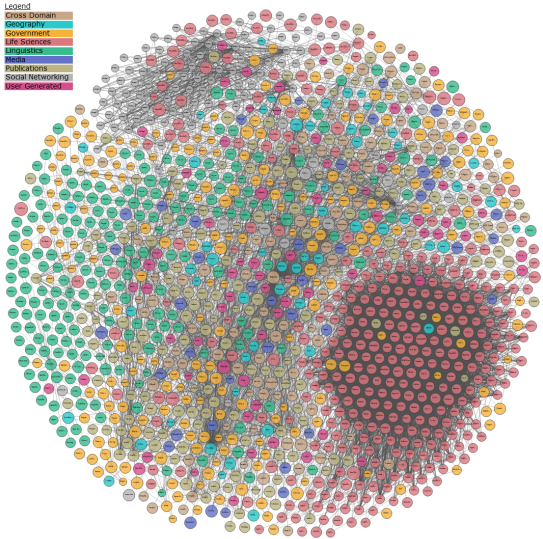
- Chemical assembly of **several molecules**
- Can either **participate in** or **control** interactions

Strategy: A Semantic Web based approach

Semantic Web : key principles

Representation of knowledge that can be understood by both humans and machines (semantic = meaning)

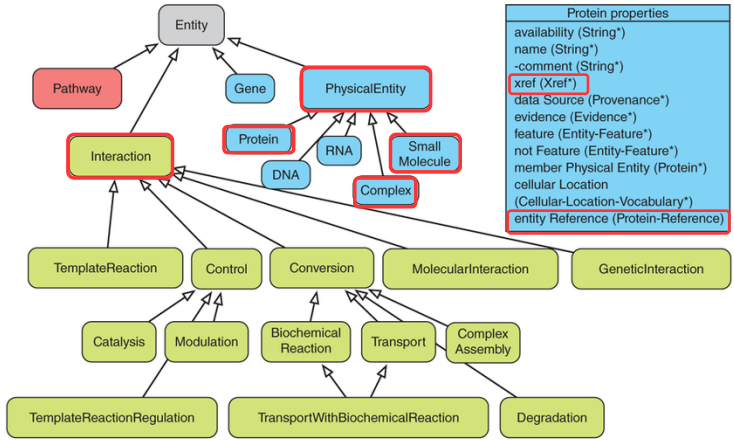
1. **RDF format**: simple way to represent knowledge (subject, predicate, object)
2. **OWL ontologies**: standardized vocabulary specific to a field
3. **SPARQL**: language for reasoning on data



Biological Pathway Exchange format (BioPAX)

Database of biological pathways in BioPAX

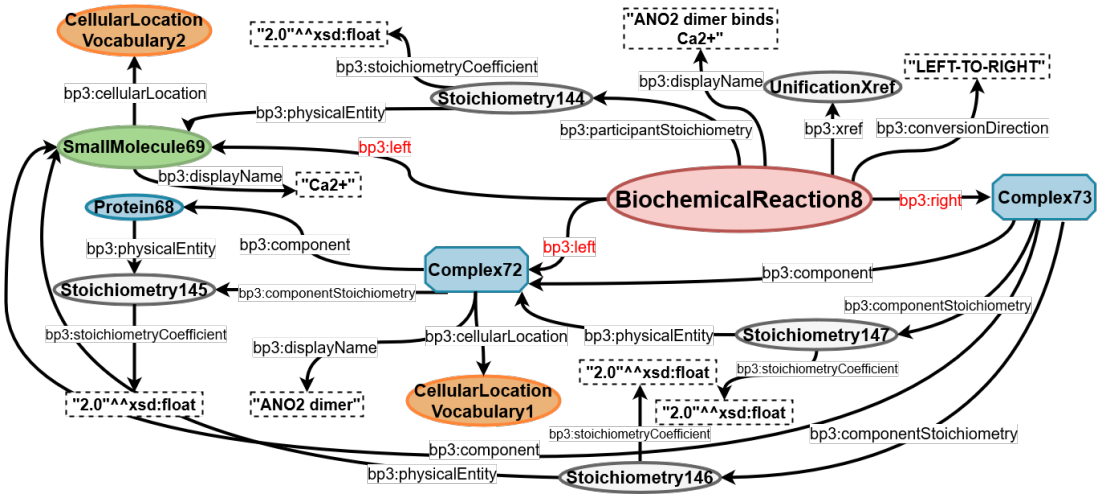
- Well established **ontology** to represent pathways at molecular and cellular levels
- **Reactome, KEGG, PathwayCommons...**
- Can be **mapped with other resources** such as ChEBI, UniProt, GO...



Demir et al. (2010)

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

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



The complexity of BioPAX reflects the complexity of biological reality

Thesis objective

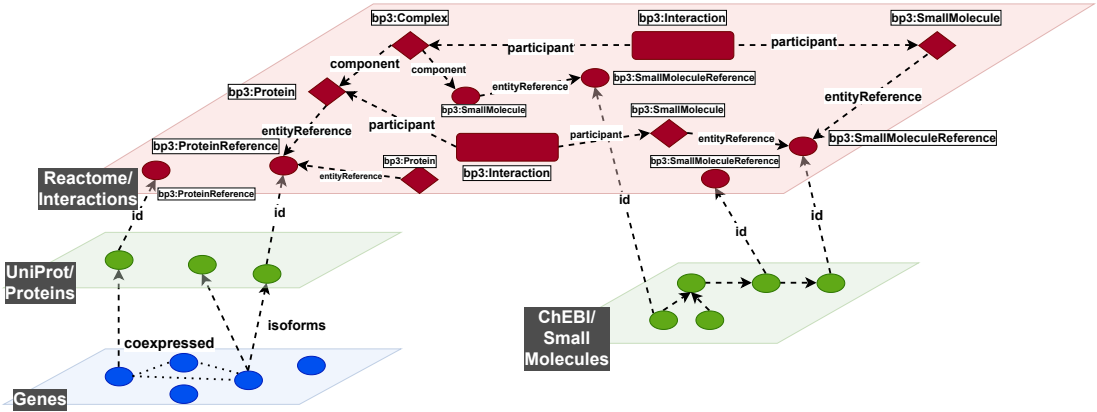
Better define key drivers of the phenotypic divergence in feed efficiency by

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

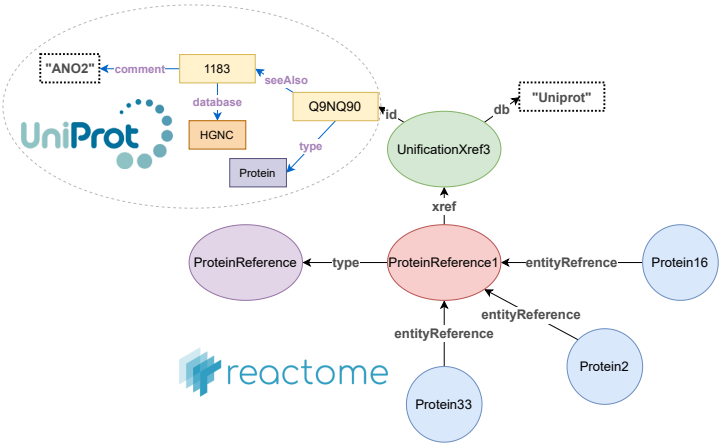
Overview

1. Introduction
2. **Contrib 1: Semantically rich queries for exhaustively connecting different -omics**
3. Contrib 2: Detect and fix non compliance with BioPAX specifications related to complexes
4. Contrib 3: A graph-based approach to identify complex connections in heterogeneous biological networks
5. Use-case: Application to feed efficiency data
6. Conclusion

Contrib 1: Semantically rich queries for exhaustively connecting different -omics



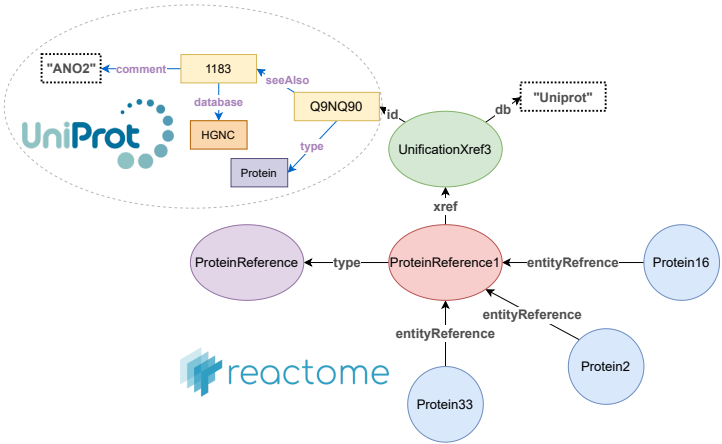
Retrieving `Proteins` in the Reactome database



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`

Retrieving `Proteins` in the Reactome database



Results in Reactome h. sapiens v81

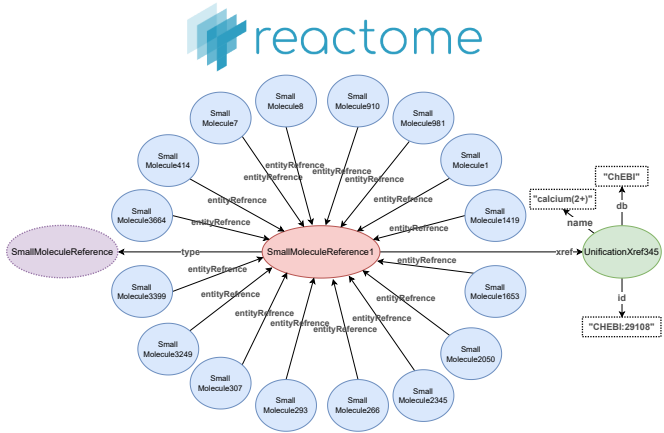
- 97% of the 11,685 `ProteinReferences` have a UniProt ID
- 89% of the 31,755 `Proteins` have a UniProt ID

Most Reactome proteins involved in reactions have a UniProt ID

Retrieving `SmallMolecules` in the Reactome database

Federated SPARQL query

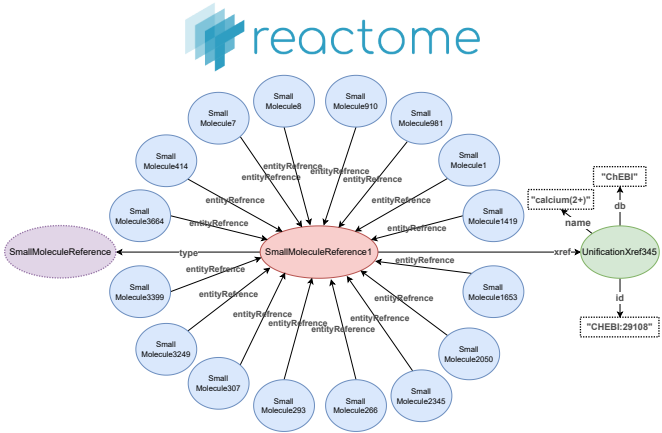
1. identify the target molecules in the ChEBI ontology (ChEBI SPARQL endpoint)
2. from a list of ChEBI IDs, locate the corresponding `SmallMoleculeReferences`
3. from these `SmallMoleculeReferences`, identify all the associated `SmallMolecules`



Retrieving `SmallMolecules` in the Reactome database

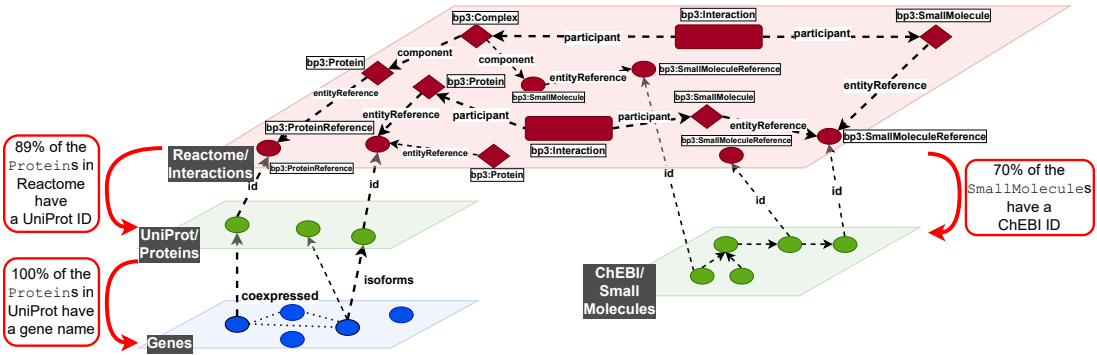
Results in Reactome h. sapiens v81

- 67% of the 2,878 `SmallMoleculeReferences` have a ChEBI ID
- 70% of the 5,049 `SmallMolecules` have a ChEBI ID



A significant number of Reactome metabolites are not identifiable in ChEBI

Contrib 1: Semantically rich queries for exhaustively connecting different -omics



Contrib 1: Outcomes and conclusions

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)

Overview

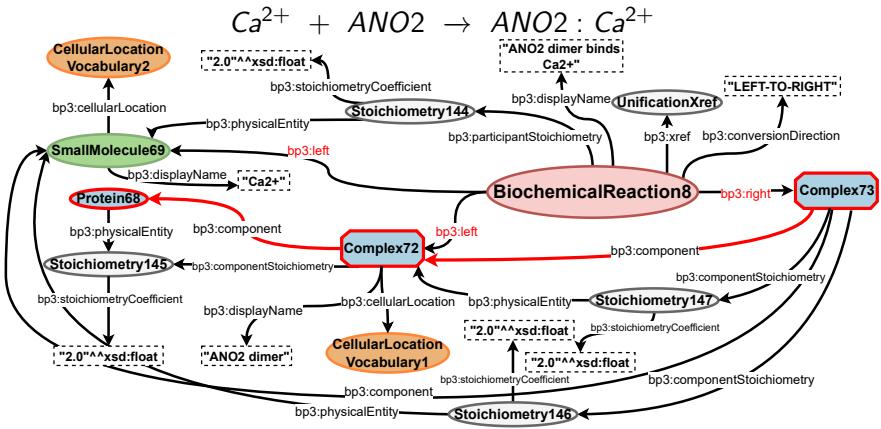
1. Introduction
2. Contrib 1: Semantically rich queries for exhaustively connecting different -omics
3. **Contrib 2: Detect and fix non compliance with BioPAX specifications related to complexes**
4. Contrib 3: A graph-based approach to identify complex connections in heterogeneous biological networks
5. Use-case: Application to feed efficiency data
6. Conclusion

Contrib 2: Detect and fix non compliance with BioPAX specifications related to complexes

(!) A complex cannot be composed of other complexes (!)

Contrib 2: Detect and fix non compliance with BioPAX specifications related to complexes

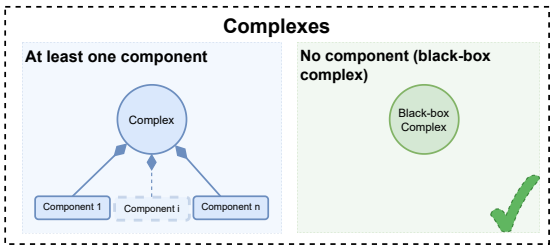
(!) A complex cannot be composed of other complexes (!)



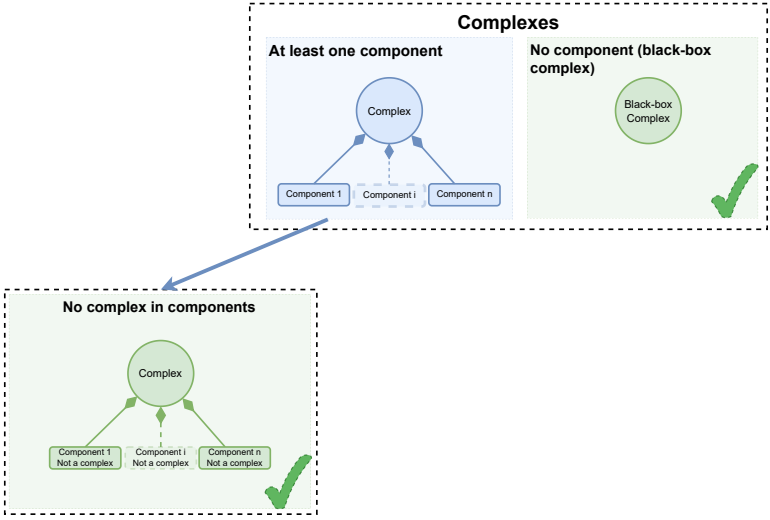
Contrib 2: Identify complexes composed of other complexes

A complex cannot be composed of other complexes

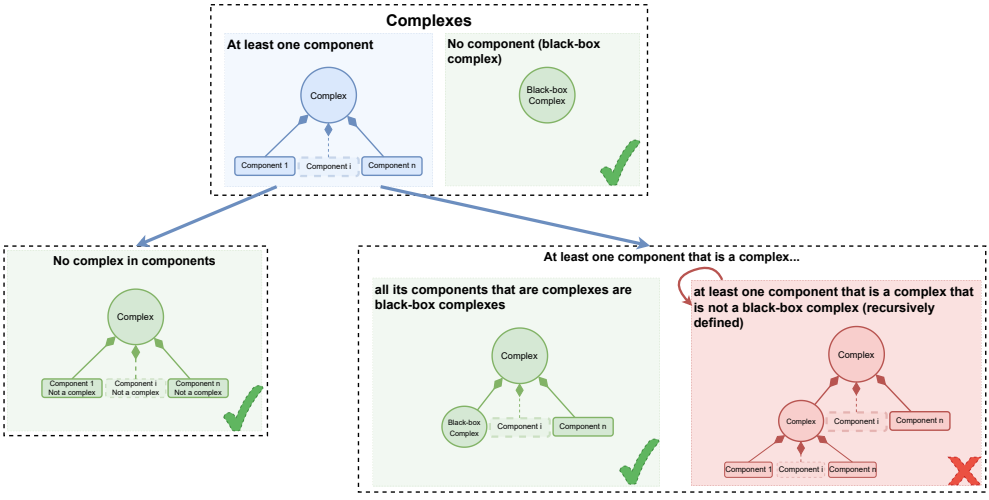
The components of a complex cannot have a component



Contrib 2: Identify complexes composed of other complexes



Contrib 2: Identify complexes composed of other complexes



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

Contrib 2: Identify and quantify invalid complexes

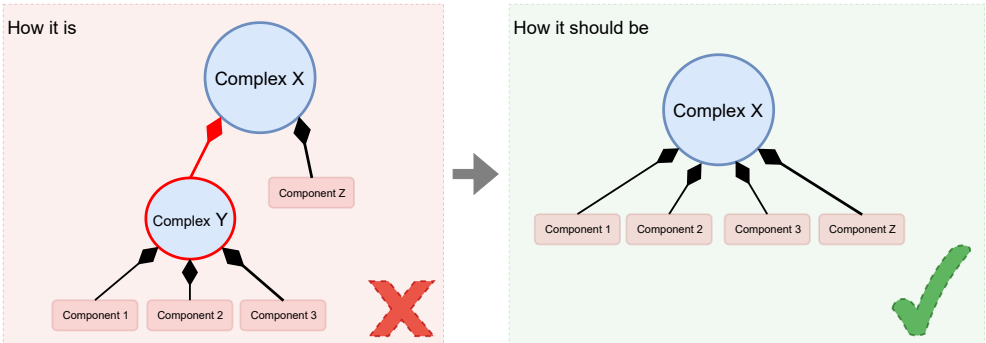
Complexes represent a large fraction of biological entities
Invalid complexes are present in large quantities in the data sets of different organisms

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**

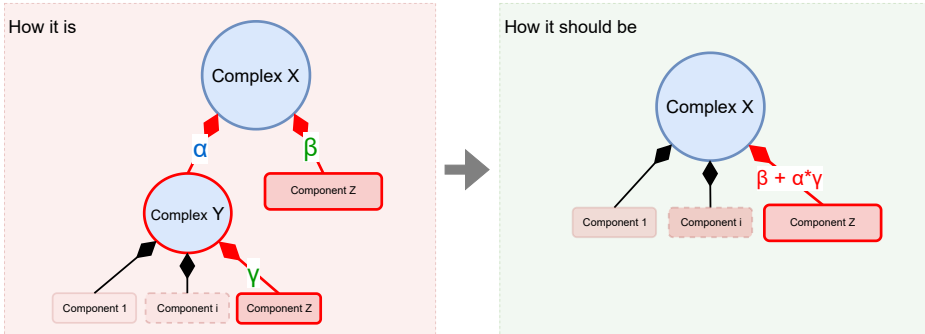
Invalid complexes composition reaches up to 10 levels in the tree of components

Contrib 2: Fix the invalid complexes



Collapse as direct components all the (in)direct components that do not have component

Contrib 2: Fix the invalid complexes

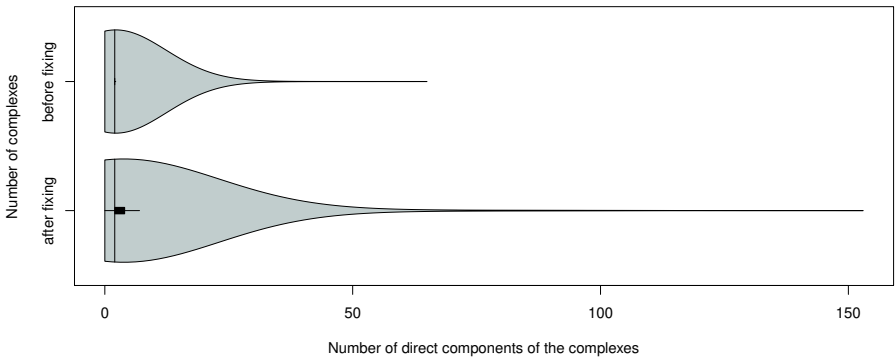


$$S(Z) = \sum_{p \in \textit{parent nodes}}^{P} S_p(Z) * S(p)$$

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

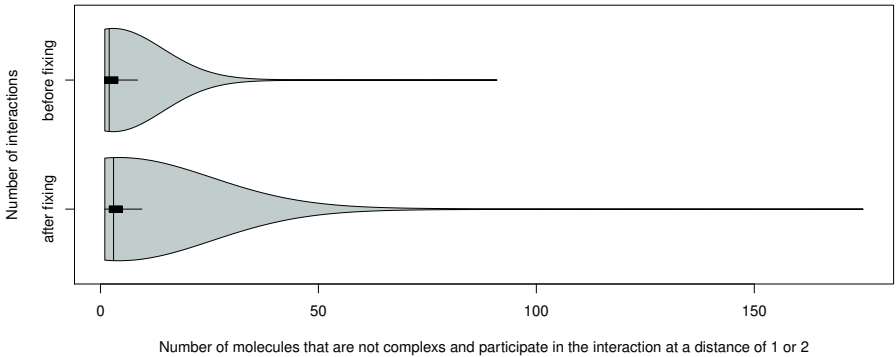
Contrib 2: Homo sapiens Reactome use-case (repair)

All invalid complexes were fixed



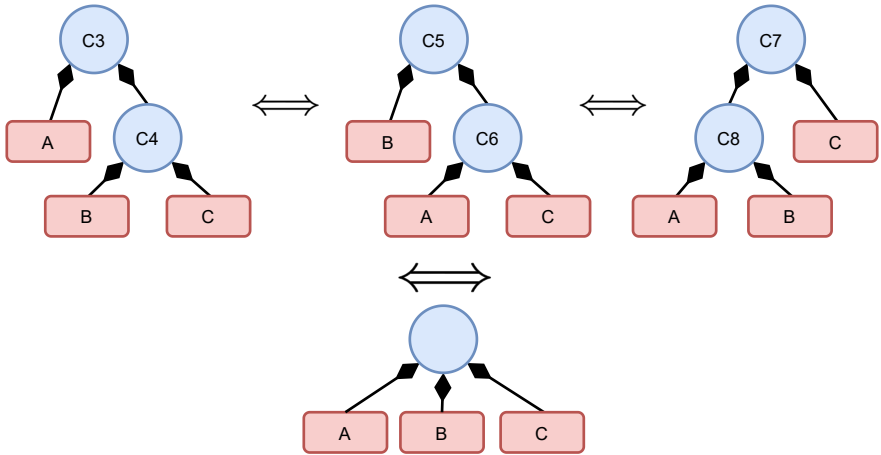
Fixing invalid complexes increases the number of direct components

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)



Fixing invalid complexes allowed to identify **333** redundant complexes (+38%)

Contrib 2: Outcomes and conclusions

Semantically-rich queries for

- identifying and fixing invalid complexes that are **reproducible** on other databases

Conclusions

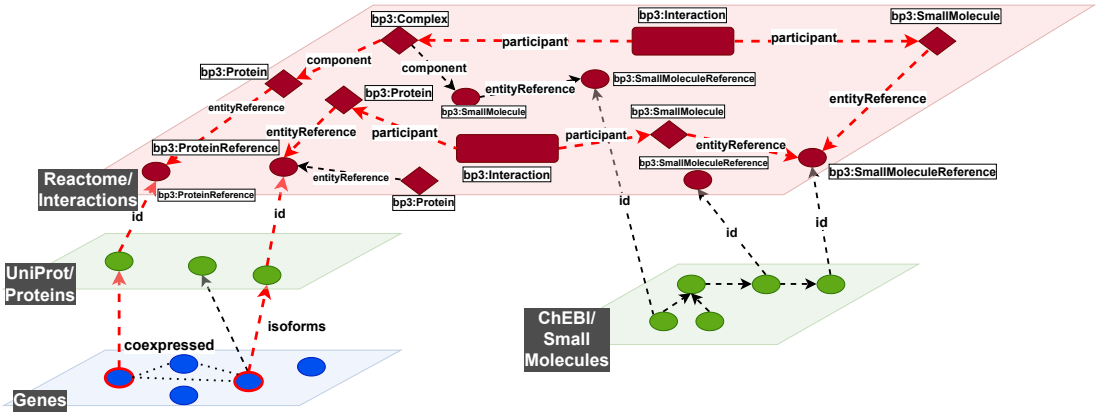
- Improves the conformity and the analysis of the graph by repairing the **topology**
- Will allow to **apply reasoning methods on better quality data**
- Side effect of **allowing the detection of complex redundancies**

Fixing molecular complexes in BioPAX standards to enrich interactions and detect redundancies using semantic web technologies. Camille Juigné, Olivier Dameron, François Moreews, Florence Gondret, Emmanuelle Becker. Bioinformatics, 2023.

Overview

1. Introduction
2. Contrib 1: Semantically rich queries for exhaustively connecting different -omics
3. Contrib 2: Detect and fix non compliance with BioPAX specifications related to complexes
4. **Contrib 3: A graph-based approach to identify complex connections in heterogeneous biological networks**
5. Use-case: Application to feed efficiency data
6. Conclusion

Contrib 3: A graph-based approach to identify complex connections in heterogeneous biological networks



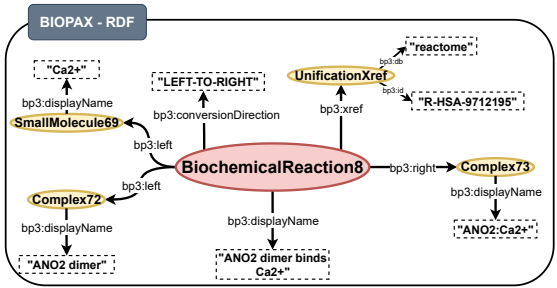
From BioPAX (RDF graph) to Neo4J (Labelled Property Graph) using NeoSemantics

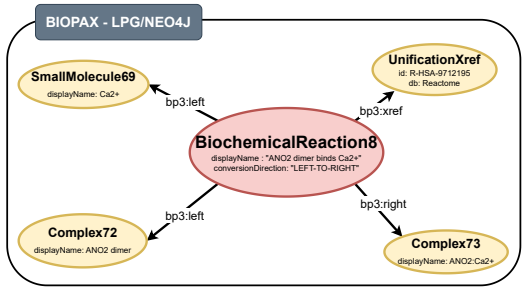
RDF - SPARLQ

- data integration
- symbolic reasoning

LPG/Neo4J - Cypher

- more complex analysis based on graph topology

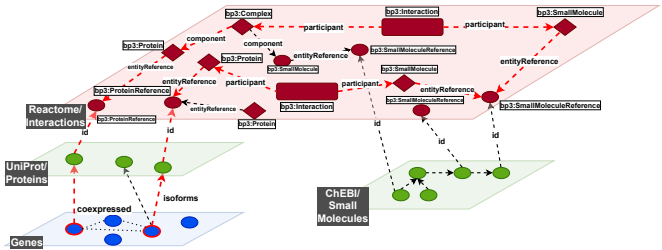




Contrib 3: Graph traversal

Comparing co-expression and random modules

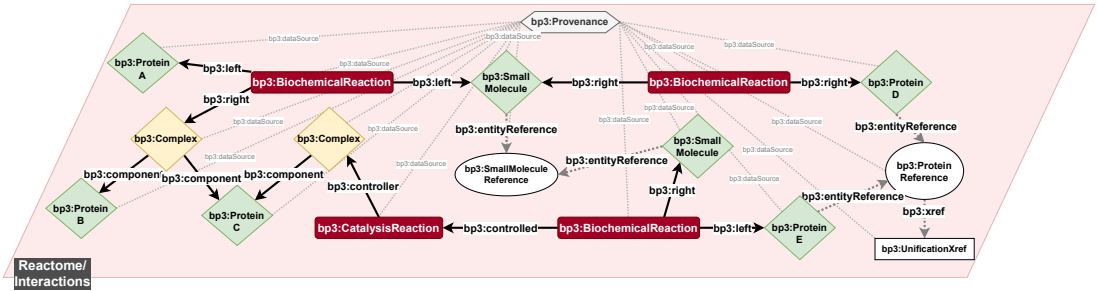
- length shortest paths connecting participants
- number of shortest paths connecting participants
- types of nodes that are traversed by the shortest path



length = 13 ; nb of interactions = 2 ; nb of small molecules = 2

We designed a graph traversal to perform these analyses using a Cypher query based on the BioPAX data schema

Contrib 3: Path filter for graph traversal



We selected a subset of edges (properties) to pass through that made biological sense

Contrib 3: Outcomes and onclusions

Cypher queries for

- conducting graph traversal based on the BioPAX data schema

Conclusions

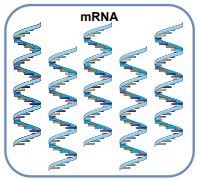
- Combining Semantic Web technologies with Neo4j using the Neosemantics plugin provides a robust framework for analyzing complex data
- Graph traversal will provide insight into the organization of biological entities of interest

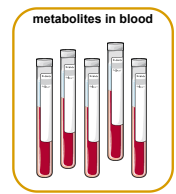
A graph-based approach to identify complex connections in heterogeneous biological networks. Camille Juigné, Océane Carpentier, Florence Gondret, Emmanuelle Becker, Olivier Dameron. To be submitted.

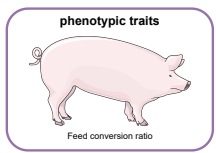
Overview

1. Introduction
2. Contrib 1: Semantically rich queries for exhaustively connecting different -omics
3. Contrib 2: Detect and fix non compliance with BioPAX specifications related to complexes
4. Contrib 3: A graph-based approach to identify complex connections in heterogeneous biological networks
5. **Use-case: Application to feed efficiency data**
6. Conclusion

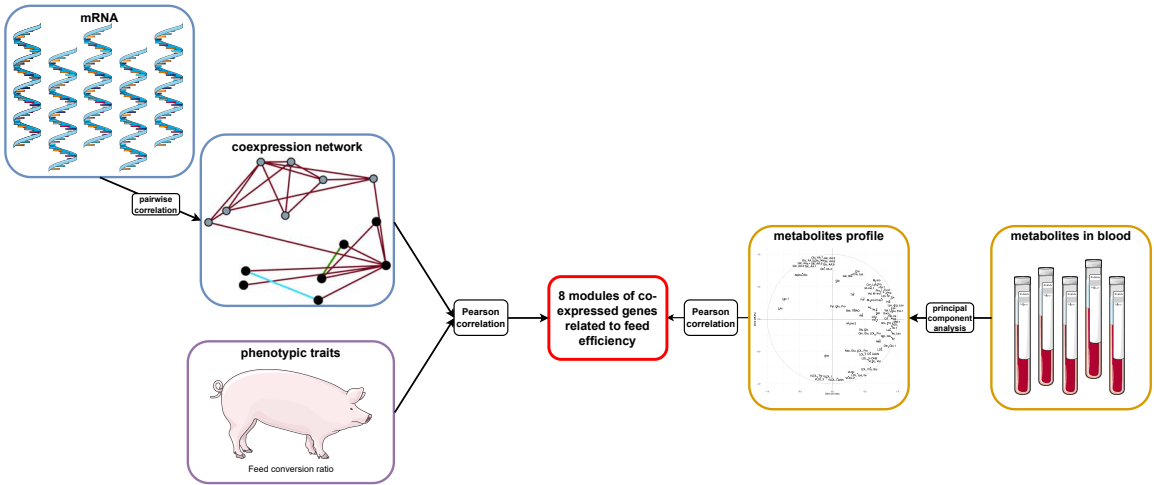
Use-case: Application to feed efficiency data



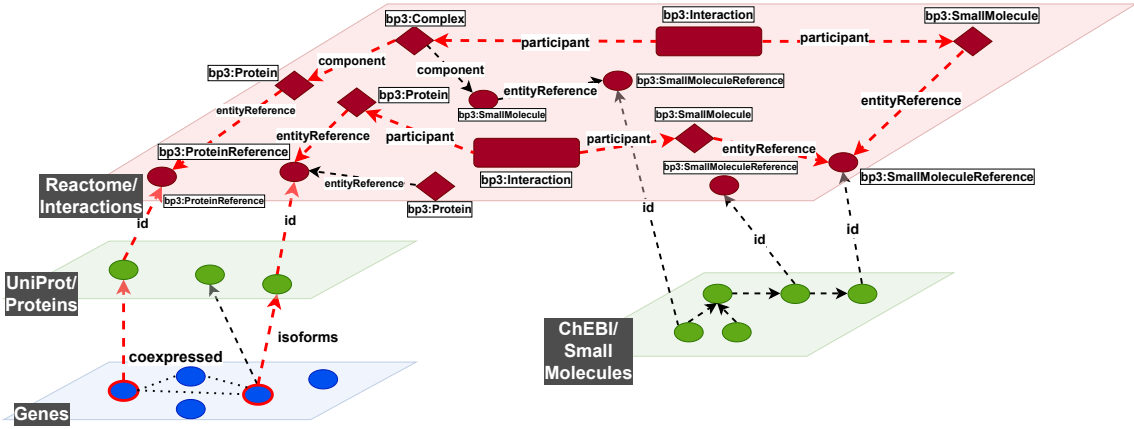




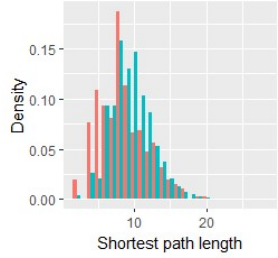
Use-case: Application to feed efficiency data

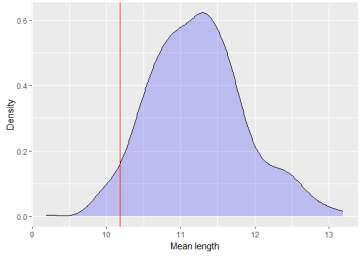


Exploring the connections within gene modules related to feed efficiency



Graph traversal results on the Royalblue module





	Royalblue module	Random modules (*)
Average shortest path length	8,5	9,7
Proportion of paths with biochemical reaction	41,5%	38,3%
Proportion of paths with small molecule	41,2%	25,0%

(*) same number of genes - average for 500 randomizations

The behavior of the Royalblue module significantly deviates from random

Use-case: Insights on feed efficiency in pigs

- Architecture of the trait: co-expressed and co-regulated gene modules identified related to feed efficiency
- Patterns with different structures than random in Reactome

Data not shown:

- These modules also regulate lean growth rate
- Among the biological processes over-represented within the modules, several are linked to immunity (+ cell development and protein localization)

Use-case: Insights on feed efficiency in pigs

- Interconnecting these modules with metabolic profiles suggests links between immunity and fatty acid % concentrations
- One of the regulatory pathways appears to be important: regulatory mechanisms - proteins G
- Relevant for future nutritional recommendations to obtain good synergy between production and health

🖃 Small networks of expressed genes in the whole blood and relationships to profiles in circulating metabolites provide insights in inter-individual variability of feed efficiency in growing pigs. Camille Juigné, Emmanuelle Becker, Florence Gondret. BMC Genomics, 2023.

General conclusion

A comprehensive and systemic method for complex phenotypes that are out of reach of traditional approaches that...

- bridges the gap between transcriptomics and metabolomics
- provides insights on complex phenotypes
- demonstrates that Semantic Web technologies can address the challenges of multi-omics integration
- offers generic, data-independent and reproducible methods and analyzes

Perspectives and potential future improvement and research directions

- Refining our graph traversal methods:
  - avoid traversing through small molecules acting as hubs in the graph (water, H+, ATP, NAD, etc.)
  - traverse the graph using alternative algorithms (e.g. random walk)
- Enhancing entity identification
- Enrich the existing graph with additional layers
- Applying our approach to another experimental dataset or a different biological question

Acknowledgment


Florence Gondret


Emmanuelle Becker


Olivier Dameron


François Moreews


Océane Carpentier

Comité de suivi de thèse


Yuna Blum


Christine Brun


Mathieu Emily


Symbiose


Pegase

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE – METAPROGRAMME DIGIT-BIO

Rannvro BREIZH