

# Formal Representation of BioPAX data as OWL ontologies



**Michel Dumontier<sup>1</sup> and Robert Hoehndorf<sup>2</sup>**

<sup>1</sup>Department of Biology, School of Computer Science, Institute of Biochemistry, Carleton University; Ottawa Institute of Systems Biology; Ottawa-Carleton Institute of Biomedical Engineering

<sup>2</sup> Department of Genetics, Cambridge University

# Why?

- Our research aims to support personalized medicine by leveraging knowledge for effective therapies.
- Our approach involves, in part, the use of formal representations to semantically integrate everything from molecular entities (dna/genes, rna, proteins, small molecules) and the roles they play in metabolism, behaviour, health and disease

# Formalization

- Formalization is the process by which we map a conceptualization into a logical representation
- We describe terms using a formal language with a well defined semantics such that we can *logically* combine terms to form expressions that have an unambiguous interpretation, and hence can be automatically reasoned about.

# Have you heard of OWL?



# OWL - The Web Ontology Language

**Enhanced vocabulary (strong axioms)** to express knowledge relating to classes, properties, individuals and data values

- quantifiers (existential, universal, cardinality restriction)
- negation
- disjunction
- property characteristics
- complex classes in domain and range restrictions
- property chains

# OWL can help you create rich, machine-understandable descriptions!

- transform our expert knowledge into axioms and expressions that can be automatically reasoned about
  - a transcription factor is
    - a protein
    - that *binds to* DNA
    - and *regulates* the expression of a gene.

Axiom: ‘transcription factor’ equivalentTo:  
‘protein’

and ‘has disposition’ some ‘to bind to DNA’

and ‘is participant in’ some ‘regulation of gene expression’

- can we mine 'omic datasets to discover which proteins are transcription factors?

# Powerful Reasoning over OWL ontologies

- **Consistency:** determines whether the ontology contains contradictions.
- **Satisfiability:** determines whether classes can have instances.
- **Subsumption:** is class C1 implicitly a subclass of C2?
- **Classification:** repetitive application of subsumption to discover implicit subclass links between named classes
- **Realization:** find the most specific class that an individual belongs to.

# Doesn't BioPAX already provide pathway data?

Absolutely.

Pathways are formalized as individuals composed of pathways and/or interactions and having molecular participants. The participants are the experimental forms of “reference” entities, and are distinguished by the side of the chemical reaction (LEFT/RIGHT) they appear.



# So what's the problem?

- BioPAX was developed specifically for the *interchange of data* from different databases – and is validated using custom software - in this way it behaves very much like an XML Schema
- Lacking the motivation for the semantic integration of pathways, the BioPAX formalization lacks the ontological commitment required to uncover new or conflicting knowledge.

# Use Cases

- Structural
  - Discover all the participants in a pathway
  - Discover physical entities that are involved in more than one pathway
  - Discover equivalent pathways obtained from different sources
  - Discover overlapping pathways from different database providers
  - Discover a path from a starting substrate to a desired product
- Biological
  - Find pathways that are involved in sugar metabolism (by looking at the parts of reactants)
  - Find pathways that *generate* or *consume* energetic intermediates (ATP, GTP, NADH, NADPH)
  - Identify pathway participants that may be affected by drugs targeting cardiac disease
  - etc

# Conceptualization

- A pathway is composed of a set of biochemical interactions and reactions
- Sometimes the events are indirectly specified indirectly when pathways have other pathways as parts.
- Normally specifies molecular events, but is often interpreted in the context of pools of molecules (wrt to kinetics)

# Formalization

## **Pathway** subClassOf

(‘has part’ some (Pathway or Interaction or Reaction))

and ‘occurs in’ some ‘physical entity’ // organism, compartment

## **Interaction** subClassOf

process

and ‘has participant’ min 2 ‘physical entity’

and ‘occurs in’ some ‘physical entity’

## **Biochemical Reaction** subClassOf

process

and ‘realizes’ some (‘reactant role’ and ‘is role of’ some ‘physical entity’)

and ‘realizes’ some (‘product role’ and ‘is role of’ some ‘physical entity’)

and ‘realizes’ some (‘enzyme role’ and ‘is role of’ some ‘physical entity’)

# Reasoning Support

## **Role chain:**

'has part' o 'has participant' -> 'has participant'

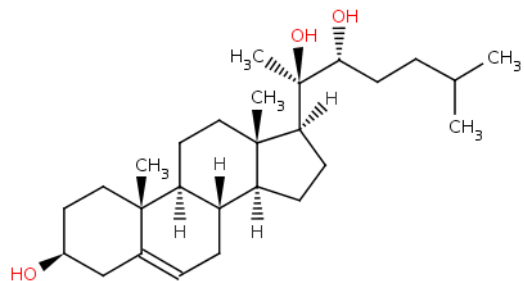
'realizes' o 'is role of' -> 'has participant'

# Materials

- Pathway commons:
  - 1623 pathways, 585k interactions, 105k entities, 564 organisms
  - Biogrid, intact, **Reactome**, mint, hprd, humancyc
  - Available as BioPAX (errors), tab files (not working); web services (incomplete)
- Other sources: Kegg (no longer free to download), Reactome, Biocyc (need cycL parser), WikiPathways (requires further investigation)

# Methods

- Data
  - Used the Reactome BioPAX file from Pathway Commons
- Ontology
  - Custom PHP Script
    - ARC2 library to parse BioPAX RDF/XML into a MySQL database
    - Executed SPARQL queries against the triple store
    - Custom PHP OWL API to generate OWL Axioms
- Reasoning
  - Protégé 4.1
  - TrOWL/FaCT++ reasoners



(20R,22R)-20,22-dihydroxycholesterol

Class hierarchy: 'Steroid hormones'

- Thing
  - Pathway
    - 'Steroid hormones'
    - CPATH-789250
    - CPATH-789251
    - CPATH-789252
    - CPATH-789253
    - CPATH-789254
    - CPATH-789321
    - CPATH-791030
    - CPATH-791031
    - CPATH-791032

Query: has-participant some chebi:1294

Execute Add to ontology

Query results

Sub classes (2)

- CPATH-794421
- CPATH-794422

With role chain:

Class hierarchy: 'Steroid hormones'

- Thing
  - Pathway
    - 'Steroid hormones'
    - CPATH-789250
    - CPATH-789251
    - CPATH-789252
    - CPATH-789253
    - CPATH-789254
    - CPATH-789321
    - CPATH-791030
    - CPATH-791031
    - CPATH-791032
    - CPATH-791033
    - CPATH-791034
    - CPATH-791035
    - CPATH-791036
    - CPATH-791037

Query: has-participant some chebi:1294

Execute Add to ontology

Query results

Sub classes (4)

- 'Steroid hormones'
- CPATH-789254
- CPATH-794421
- CPATH-794422





# DL Reasoners provide an Explanation

The screenshot shows a DL Reasoner interface. On the left, a class hierarchy for 'Steroid hormones' is displayed, with 'Steroid hormones' selected. The hierarchy includes 'Thing', ':Pathway', and several CPATH identifiers. On the right, a query window shows the query: 'has-participant some chebi:1294'. Below the query, the 'Query results' section lists sub-classes: 'Steroid hormones', 'CPATH-789254', 'CPATH-794421', and 'CPATH-794422'. A red arrow points to the question mark icon next to 'CPATH-789254'.

The screenshot shows an 'Explanation for CPATH-789254 SubClassOf has-participant some chebi:1294' window. The window contains the following axioms:

- CPATH-789254 SubClassOf has-component-part some CPATH-794421
- CPATH-794421 SubClassOf realizes some (role:ReactantRole and (is-role-of some chebi:1294))
- has-component-part o has-participant SubPropertyOf has-participant
- realizes o is-role-of SubPropertyOf has-participant

An 'OK' button is located at the bottom of the window.

# Data quality Issues?


(generated from PathwayCommons XREF entries)

- CPATH-791039
- CPATH-791040
- CPATH-791095
- CPATH-791097
- CPATH-792052
- CPATH-792412
- CPATH-792415
- CPATH-792416
- CPATH-792417
- CPATH-794100
- CPATH-794110
- CPATH-794111
- CPATH-794112
- CPATH-794113
- CPATH-794114
- CPATH-794115
- CPATH-794116
- CPATH-794117
- CPATH-794118
- CPATH-794119
- CPATH-794120
- CPATH-794121
- CPATH-794122
- CPATH-794123
- CPATH-794124
- CPATH-794125

Equivalent classes +	
Superclasses +	
● has-product some chebi:15355	Ⓜ ✕ ○
● has-product some chebi:15366	Ⓜ ✕ ○
● has-product some chebi:15377	Ⓜ ✕ ○
● has-product some chebi:15378	Ⓜ ✕ ○
● has-product some chebi:15422	Ⓜ ✕ ○
● has-product some chebi:15551	Ⓜ ✕ ○
● has-product some chebi:15553	Ⓜ ✕ ○
● has-product some chebi:15627	Ⓜ ✕ ○
● has-product some chebi:15647	Ⓜ ✕ ○
● has-product some chebi:15650	Ⓜ ✕ ○
● has-product some chebi:15729	Ⓜ ✕ ○
● has-product some chebi:15756	Ⓜ ✕ ○
● has-product some chebi:16196	Ⓜ ✕ ○
● has-product some chebi:16467	Ⓜ ✕ ○
● has-product some chebi:16469	Ⓜ ✕ ○
● has-product some chebi:16761	Ⓜ ✕ ○
● has-product some chebi:16978	Ⓜ ✕ ○

# Next Steps

- Distinguish between individual molecular transformations and mass action kinetics
- Process all BioPAX events to get a richer knowledge base
  - Understand the source of the data quality problems
- Formalize the XREFS to other ontological sources, and incorporate into the final ontology



Thank You

Michel Dumontier

michel\_dumontier@carleton.ca

*Publications:* <http://dumontierlab.com>

*Presentations:* <http://slideshare.com/micheldumontier>



MITACS



Canada Foundation for Innovation  
Fondation canadienne pour l'innovation



canarie



Carleton  
UNIVERSITY



Health  
Canada



Ontario