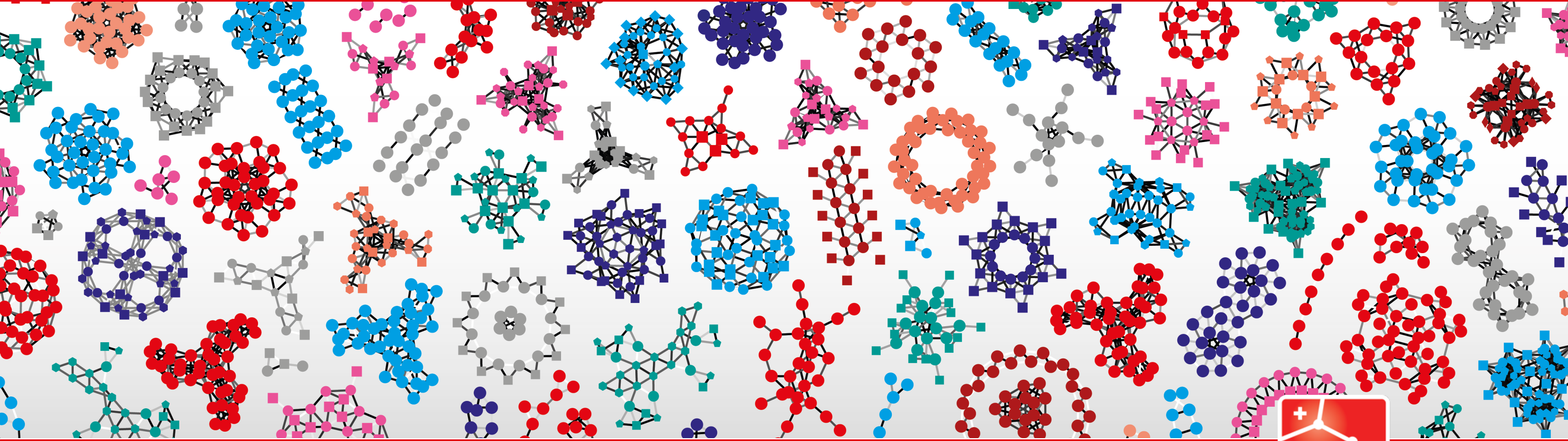


Swiss Institute of Bioinformatics

Querying the Orthologous MAtrix (OMA) Database

Tarcisio Mendes de Farias and Christophe Dessimoz

tarcisio.mendes@sib.swiss

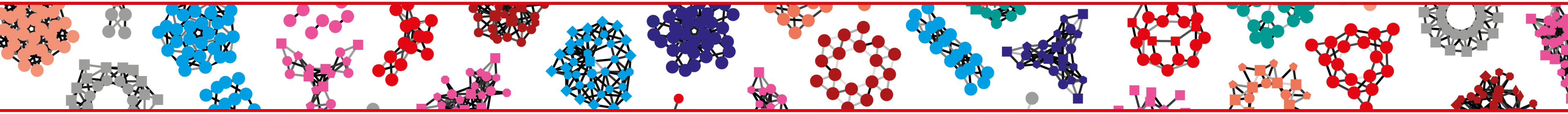
SIB Scientist

elixir SWITZERLAND | Unil UNIL | Université de Lausanne | oma BROWSER

www.omabrowser.org

www.sib.swiss

Overview



01 Introduction – What is the OMA database?

02 Data schema – ORTH Ontology

03 Querying OMA RDF data with SPARQL

04 Conclusion

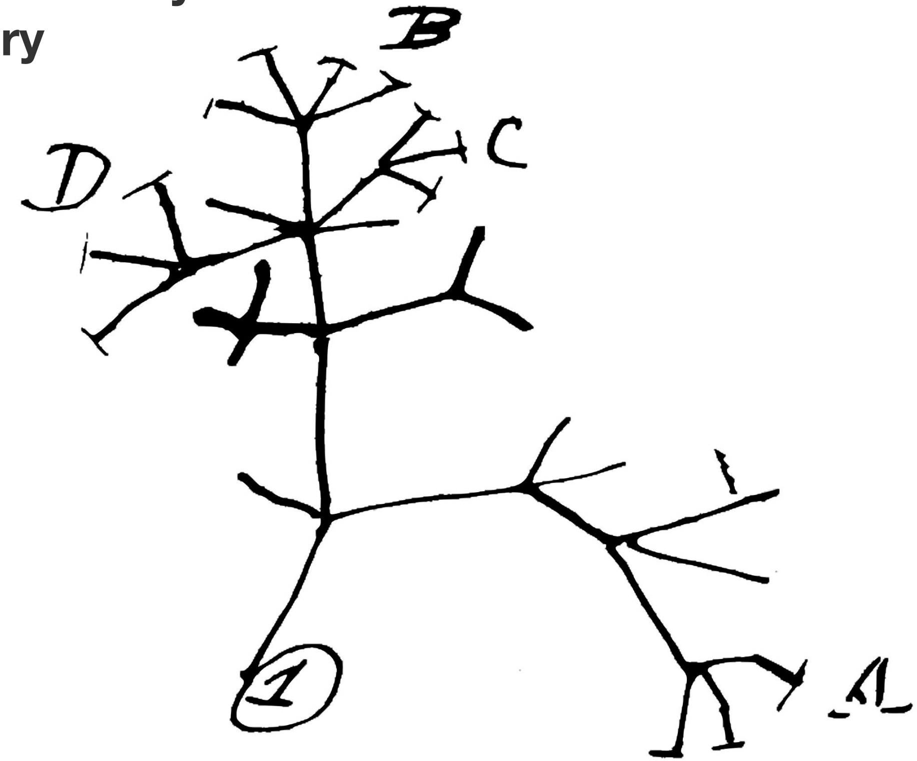
05 A federated query – OMA and Bgee databases

Introduction - OMA Database

- Gene classification based on evolutionary history is essential for many aspects of comparative and functional genomics.
- Evolutionary relations are often described as binary relations.
- Orthologous clusters
- Hierarchical Orthologous Groups (HOGs)

Homologous genes

Homologs are genes related by common ancestry



Homologs

Ortholog

Paralog

Xenolog

Co-ortholog

In-paralog

Out-paralog

…

Definitions

Orthology: A relation between pairs of genes that started diverging via evolutionary speciation

Paralogy: A relation between pairs of genes that started diverging via gene duplication

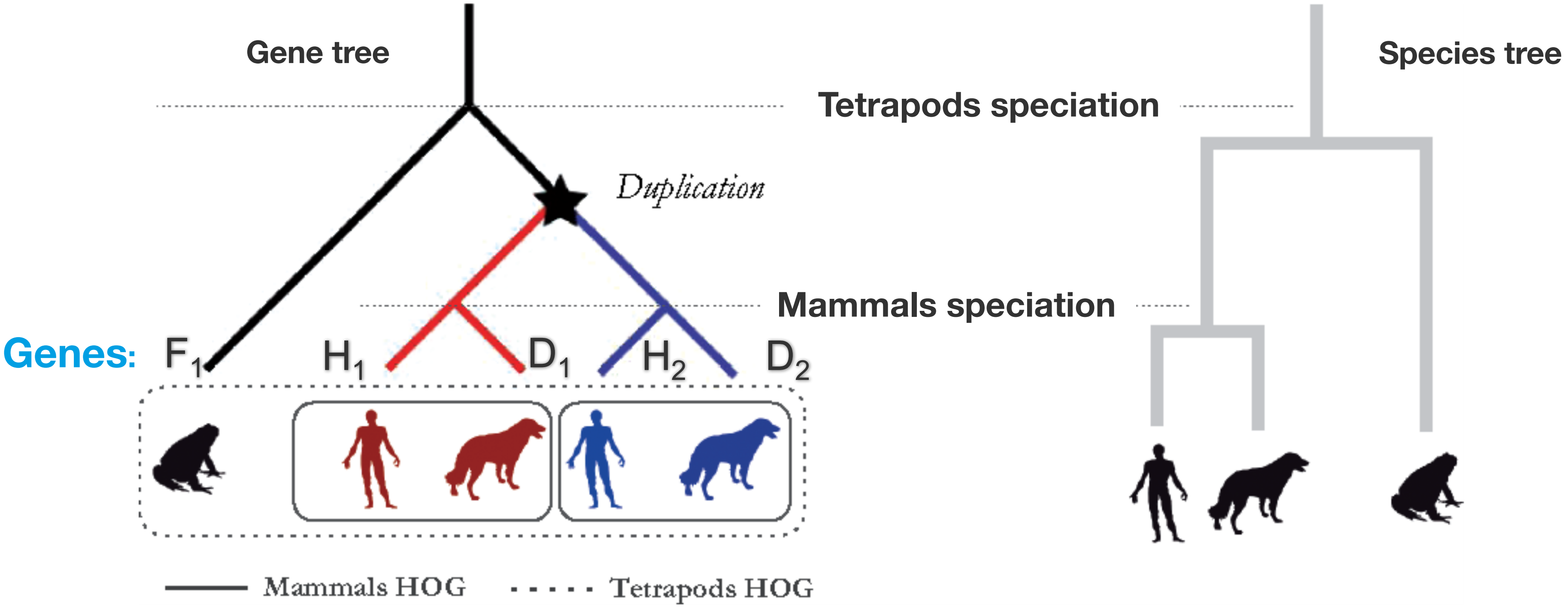
Orthologs: pairs of genes that started diverging via evolutionary speciation

Paralogs: pairs of genes that started diverging via gene duplication

⚠

Ortho = exact
Para = beside/next to

Hierarchical Orthologous Groups

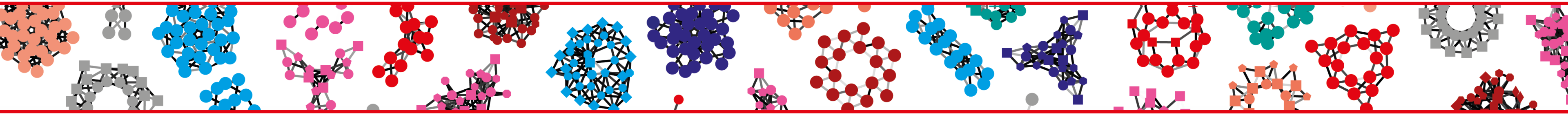


Each HOG is an ancestral gene at a given taxonomic level

Introduction - OMA Database

- **Orthologous gene inferences covering all three domains of life: Archaea, Bacteria, and Eukarya**

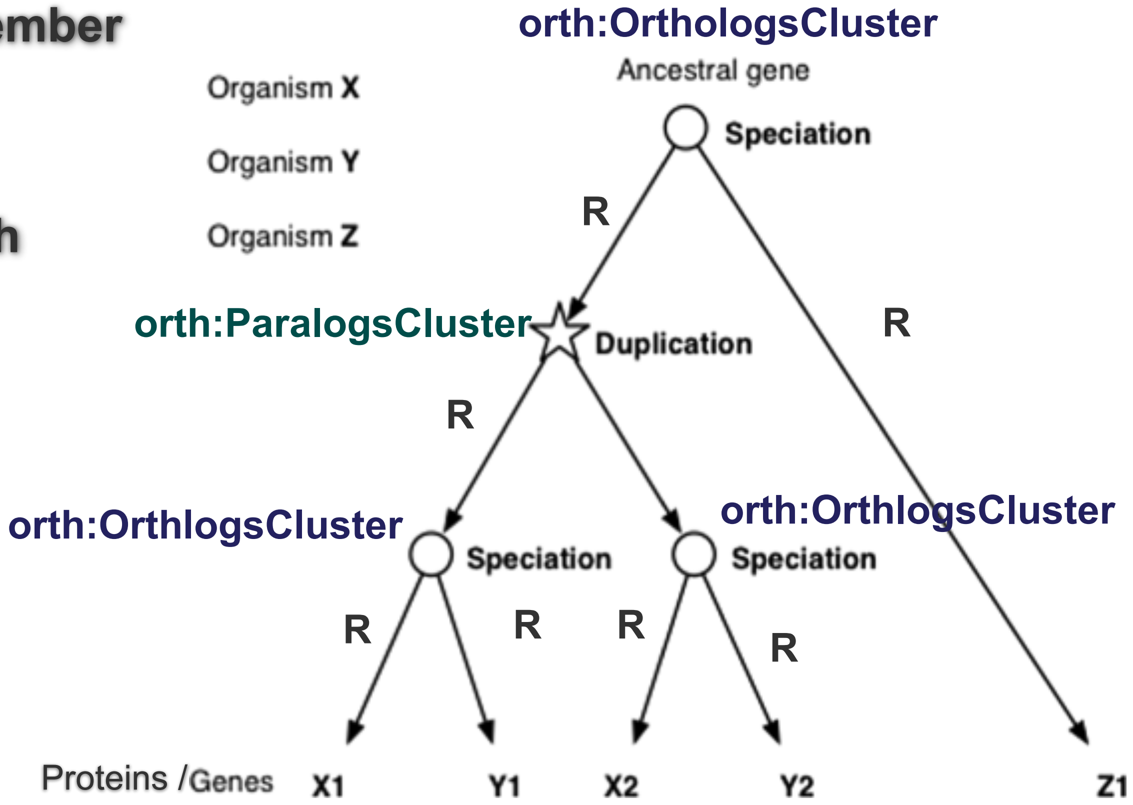
- **The 2018 OMA version has 2167 species and the HOGs can be queried through the SPARQL endpoint at**
  https://sparql.omabrowser.org/lode/sparql/
  https://sparql.omabrowser.org/sparql/
  (Virtuoso triple store)



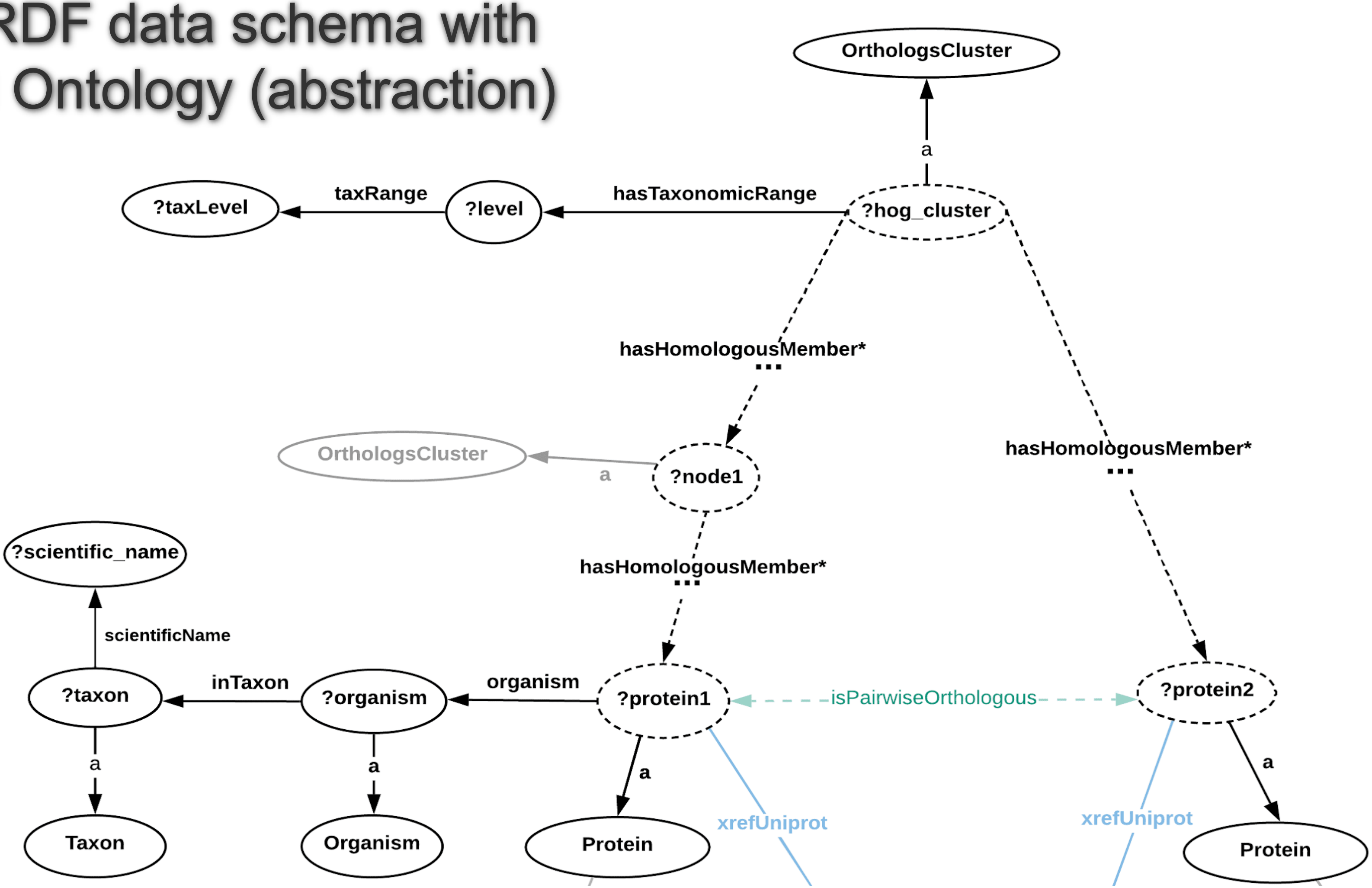
Data schema – ORTH Ontology

Representing HOGs with the ORTH Ontology

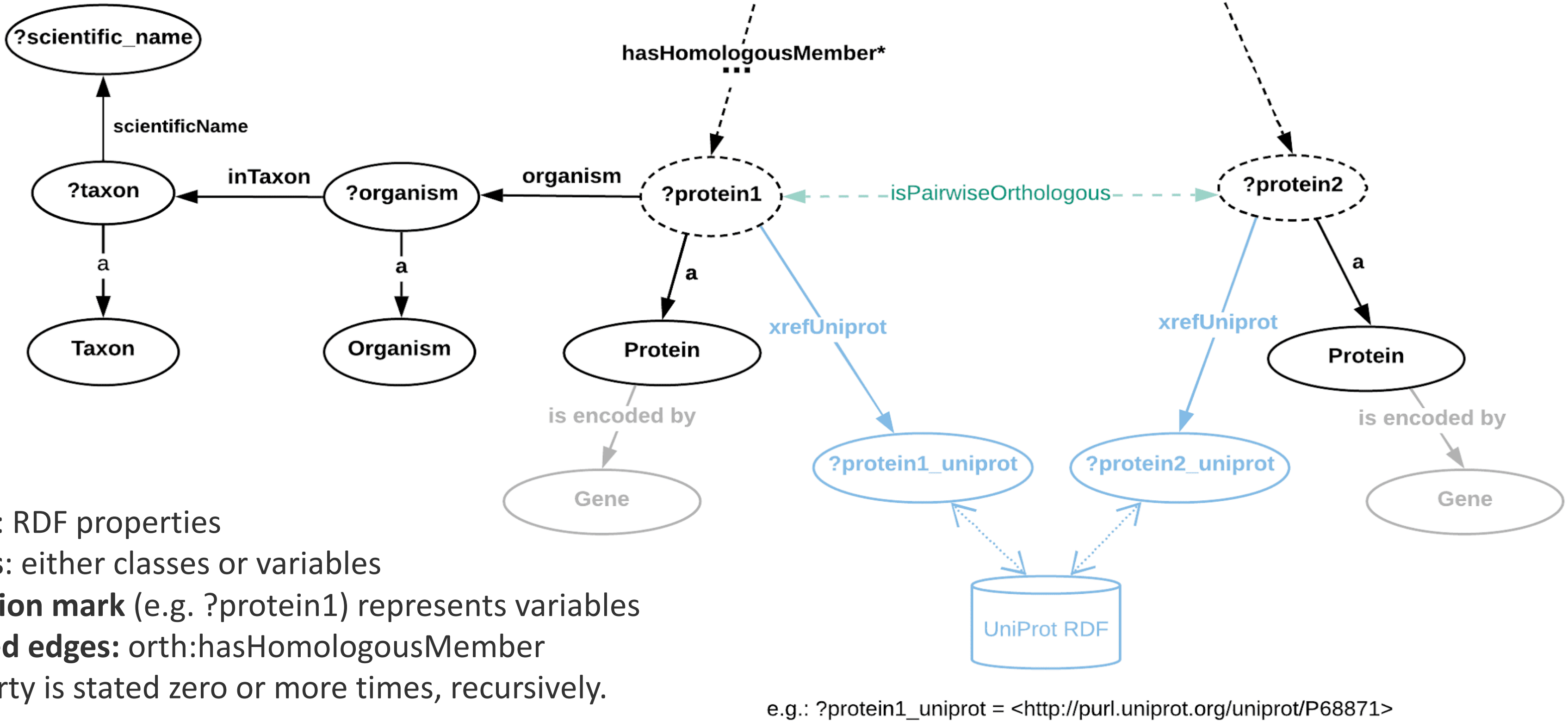
- R = orth:hasHomologousMember
- OMA HOGs represented with ORTH
- A three data structure
- OMA is protein-centric



OMA RDF data schema with ORTH Ontology (abstraction)



OMA RDF data schema with ORTH Ontology (abstraction)

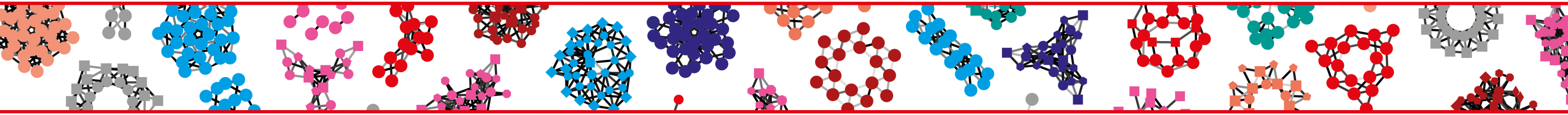


e.g.: ?protein1_uniprot = <http://purl.uniprot.org/uniprot/P68871>

LEGEND

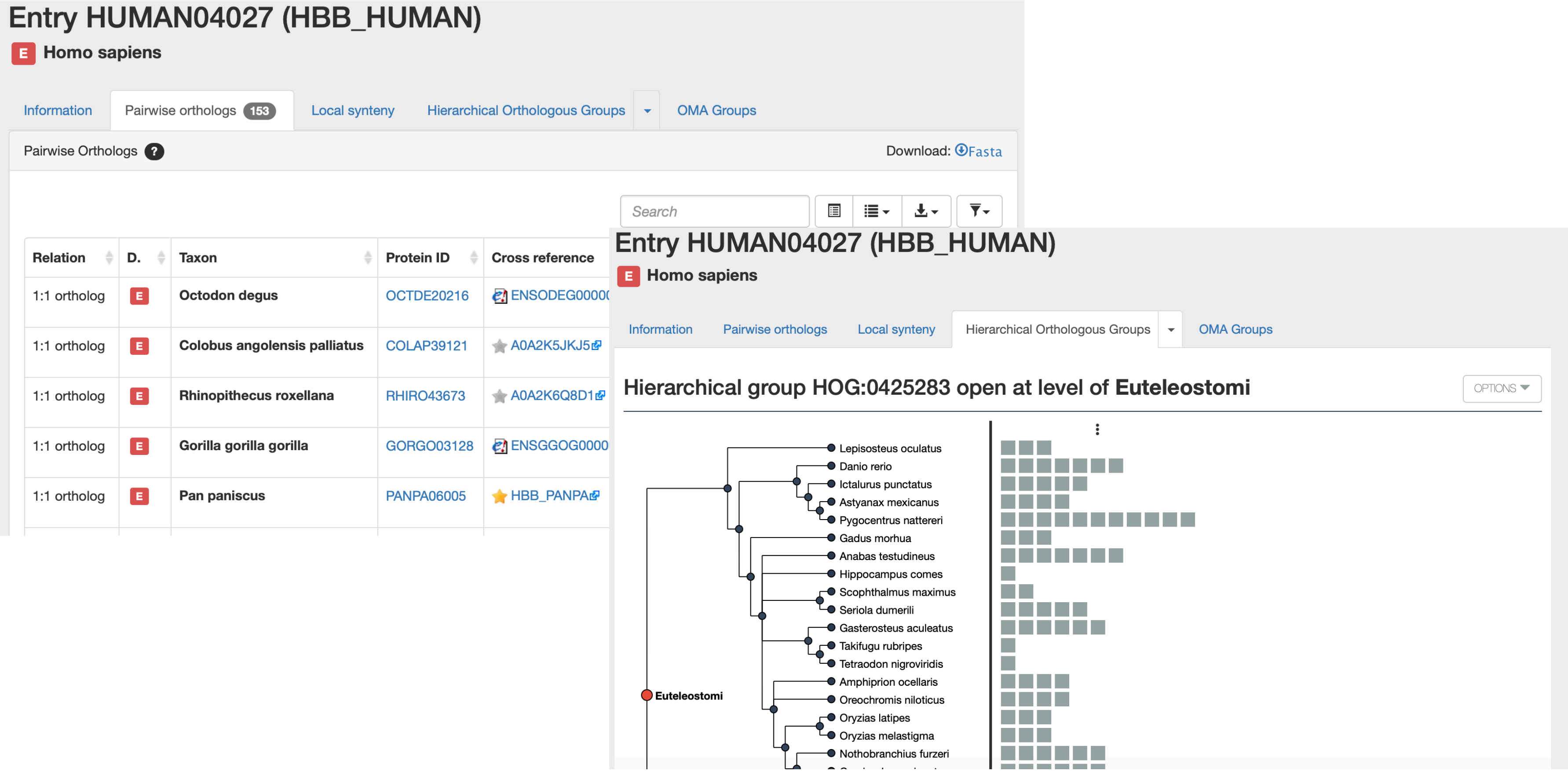
- **Edges**: RDF properties
- **Nodes**: either classes or variables
- **Question mark** (e.g. ?protein1) represents variables
- **Dashed edges:** orth:hasHomologousMember property is stated zero or more times, recursively.

*Note: URI prefixes were omitted



Querying OMA RDF data with SPARQL

OMA Browser (Webpages)



Entry HUMAN04027 (HBB_HUMAN)

E **Homo sapiens**

Relation	D.	Taxon	Protein ID	Cross reference
1:1 ortholog	E	**Octodon degus**	OCTDE20216	ENSODEG0000
1:1 ortholog	E	**Colobus angolensis palliatus**	COLAP39121	A0A2K5JKJ5
1:1 ortholog	E	**Rhinopithecus roxellana**	RHIRO43673	A0A2K6Q8D1
1:1 ortholog	E	**Gorilla gorilla gorilla**	GORGO03128	ENSGGOG0000
1:1 ortholog	E	**Pan paniscus**	PANPA06005	HBB_PANPA

Entry HUMAN04027 (HBB_HUMAN)

E **Homo sapiens**

Hierarchical group HOG:0425283 open at level of **Euteleostomi**

Retrieve pairwise orthologs from OMA HOGs with ORTH

- **isPairwiseOrthologousTo –> orth:hasOrtholog**

```
PREFIX orth: <http://purl.org/net/orth#>
SELECT ?seq_1 ?seq_2 {

    ?cluster a orth:OrthologsCluster.
    ?cluster orth:hasHomologousMember ?node_1.
    ?cluster orth:hasHomologousMember ?node_2.
    ?node_1 orth:hasHomologousMember* ?seq_1.
    ?node_2 orth:hasHomologousMember* ?seq_2.
    ?seq_1 a orth:Protein.
    ?seq_2 a orth:Protein.
    FILTER (?node_1 != ?node_2)
}
```

```
?seq_1 orth:hasOrtholog ?seq_2
```

Retrieve pairwise paralogs from OMA HOGs with ORTH

```
PREFIX orth: <http://purl.org/net/orth#>
SELECT ?seq_1 ?seq_2 {
   ?cluster a orth:ParalogsCluster.
   ?cluster orth:hasHomologousMember ?node_1.
   ?cluster orth:hasHomologousMember ?node_2.
   ?node_1 orth:hasHomologousMember* ?seq_1.
   ?node_2 orth:hasHomologousMember* ?seq_2.
   ?seq_1 a orth:Protein.
   ?seq_2 a orth:Protein.
   FILTER (?node_1 != ?node_2)
}
```

```
?seq_1 orth:hasParalog ?seq_2
```

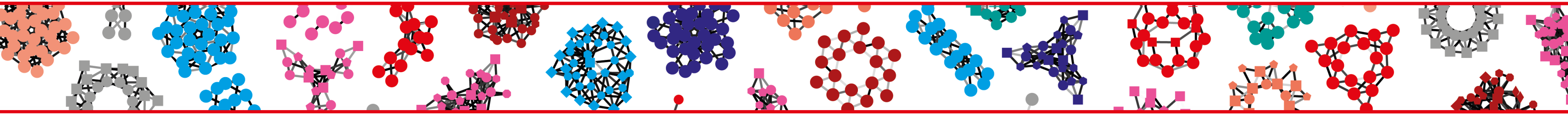
Retrieve Homologous Groups with ORTH

```
PREFIX lscr: <http://purl.org/lscr#>
PREFIX orth: <http://purl.org/net/orth#>

SELECT DISTINCT ?cluster ?protein2_OMA_URI ?protein2_uniprot_URI ?tax_name {
  VALUES(?protein1_uniprot_URI){(<http://purl.uniprot.org/uniprot/P68871>)}
  VALUES(?tax_name){("Primates")}
  ?cluster a orth:OrthologsCluster.
  ?cluster orth:hasHomologousMember* ?protein_OMA_1.
  ?cluster orth:hasHomologousMember* ?protein2_OMA_URI.
  ?protein_OMA_1 a orth:Protein.
  ?protein2_OMA_URI a orth:Protein.
  ?protein_OMA_1  lscr:xrefUniprot ?protein1_uniprot_URI.
  OPTIONAL{?protein2_OMA_URI  lscr:xrefUniprot ?protein2_uniprot_URI.}
  ?cluster orth:hasTaxonomicRange ?tax.
  ?tax  orth:taxRange ?tax_name. }
```

Retrieve OMA HOGs represented with ORTH

```
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX orth: <http://purl.org/net/orth#>
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX lscr: <http://purl.org/lscr#>
 SELECT DISTINCT ?root_hog ?species_name ?protein1_uniprot (?protein1 as
        ?protein1_OMA) ?taxLevel  {
          VALUES ?protein2_uniprot {<http://purl.uniprot.org/uniprot/P68871>}
          ?root_hog obo:CDAO_0000148 ?hog_cluster.          #has_Root
          ?hog_cluster orth:hasHomologousMember* ?node1.
          ?node1 a orth:OrthologsCluster.
          ?node1 orth:hasTaxonomicRange ?level.
          ?level orth:taxRange ?taxLevel .
          ?node1 orth:hasHomologousMember* ?protein1.
          ?hog_cluster orth:hasHomologousMember* ?protein2.
          ?protein1 a orth:Protein.
          ?protein1 orth:organism ?organism.
          ?organism obo:RO_0002162 ?taxon.
          ?taxon up:scientificName ?species_name.
          OPTIONAL{?protein1 lscr:xrefUniprot ?protein1_uniprot}.
          ?protein2 a orth:Protein.
          ?protein2  lscr:xrefUniprot ?protein2_uniprot. } ORDER BY ?taxLevel
```

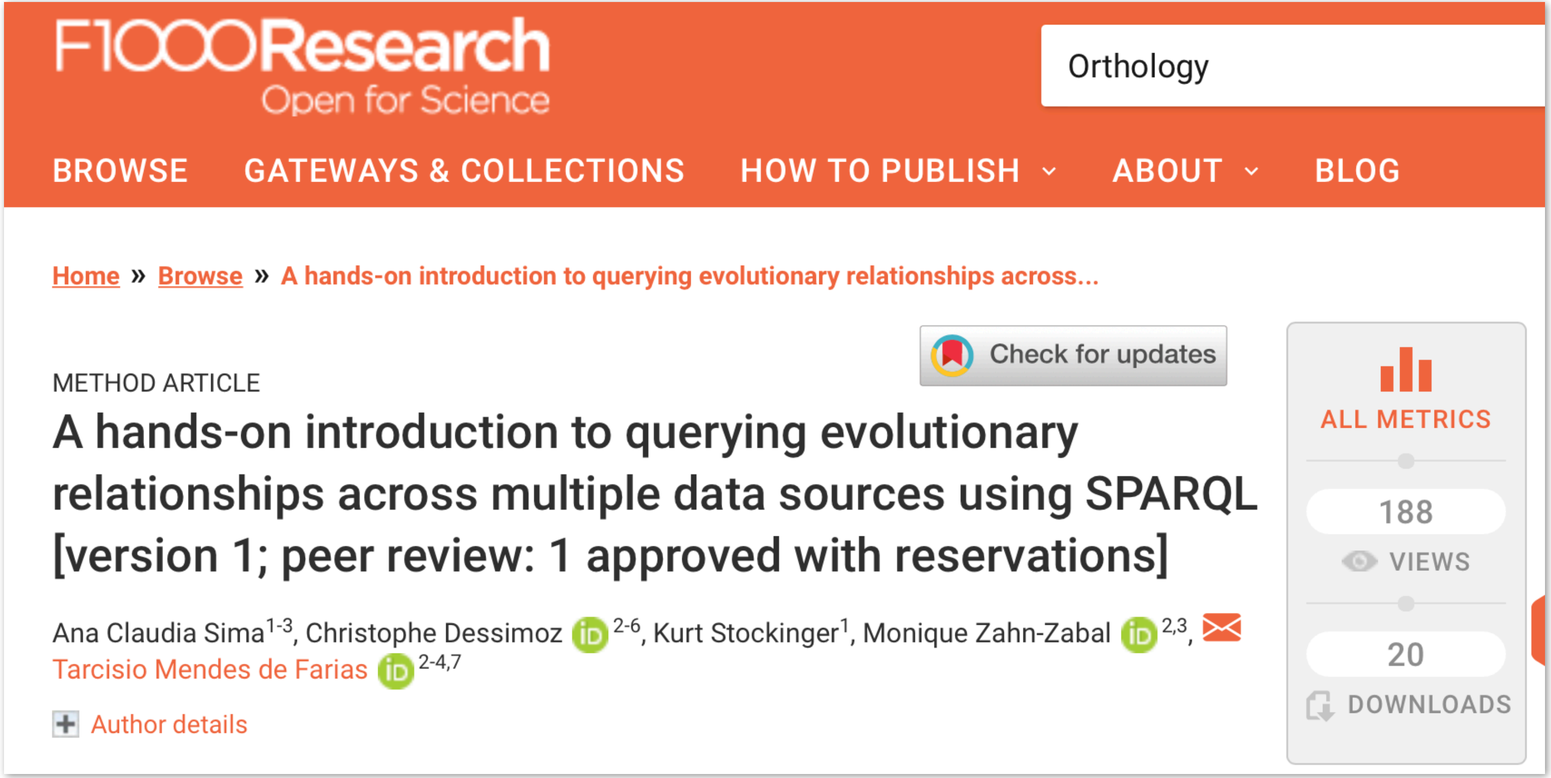


Conclusion

Conclusion

- In this tutorial, we learned how to query and retrieve orthology and paralogy information from the OMA HOGs
- We described the main part of the ORTH ontology used to represent the core data provided by OMA.
- We have shown how we take advantage of the OMA HOG structure to avoid the materialization of billion triples
- https://sparql.omabrowser.org/lode/sparql/

Tutorial for querying multiple orthology databases



Home » Browse » A hands-on introduction to querying evolutionary relationships across...

METHOD ARTICLE

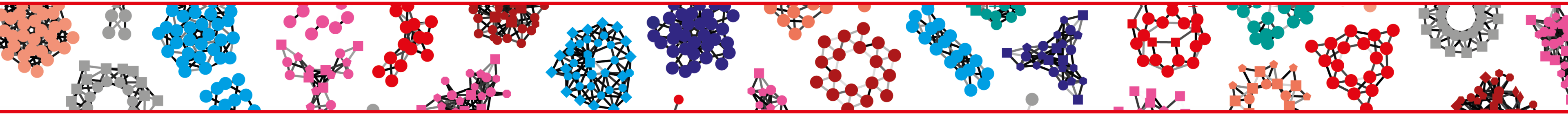
A hands-on introduction to querying evolutionary relationships across multiple data sources using SPARQL [version 1; peer review: 1 approved with reservations]

Ana Claudia Sima[1-3], Christophe Dessimoz [2-6], Kurt Stockinger[1], Monique Zahn-Zabal [2,3], Tarcisio Mendes de Farias [2-4,7]

188 VIEWS

20 DOWNLOADS

https://purl.org/orthology/paper



A Conjunctive Federated Query: OMA and Bgee databases

OMA-Bgee Federated Query Example – Part 1

```
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX genex: <http://purl.org/genex#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX orth: <http://purl.org/net/orth#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
```

#Which are the genes expressed in the Mouse's pancreas that are orthologous to the INS gene in the human?

```
SELECT DISTINCT ?protein1 (?protein2 as ?orthologous_to)  WHERE {
        SELECT * {
                ?cluster    a orth:OrthologsCluster .
                ?cluster    orth:hasHomologousMember ?node1 .
                 ?cluster    orth:hasHomologousMember ?node2 .
                ?node2    orth:hasHomologousMember* ?protein2 .
                ?node1     orth:hasHomologousMember* ?protein1 .
                ?protein1  sio:SIO_010079 ?gene1 ;
                           orth:organism ?organism1 .
                ?organism1 obo:RO_0002162 ?taxon1 .
                ?taxon1      up:scientificName 'Mus musculus'.     …
```

OMA-Bgee Federated Query Example – Part 2

```
        …
        ?protein2     rdfs:label 'INS';
                      orth:organism ?organism2.
        ?organism2  obo:RO_0002162 ?taxon2.  #in taxon property
        ?taxon2       up:scientificName 'Homo sapiens'.
         FILTER(?node1 != ?node2)
SERVICE <http://biosoda.expasy.org/rdf4j-server/repositories/bgeelight>{
           ?gene1 genex:isExpressedIn ?anat .
           ?anat    rdfs:label 'pancreas'^^xsd:string . }}}
```

Swiss Institute of Bioinformatics

OMA team members: Adrian Altenhoff, Victor Rossier, Christophe Dessimoz, David Dylus, David Moi, Alex , **Tarcisio Mendes de Farias**, Yannis Nevers and Natasha Glover

UNIL | Université de Lausanne

www.sib.swiss