# Supplementary Material

for tutorial "A hands-on introduction to querying evolutionary relationships across multiple data sources using SPARQL"

Table 1. "Cheat sheet" for RDF data available in the four sources considered in this tutorial. (\*) GO annotations can be retrieved from the UniProt RDF store through UniProt cross-references.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Information provided /  Source in RDF | EBI | OMA | OrthoDB | MBGD |
| Protein-centric |  | √ |  |  |
| Gene-centric | √ |  | √ | √ |
| Sub-gene level |  |  |  | mbgd:domainBegin, mbgd:domainEnd |
| Orthology property | sio:SIO\_000558  (isOrthologousTo) | orth:hasHomologousMember | orthodb:memberOf | orth:hasHomologous |
| Cross-Ref to UniProt | ensemblterms:DEPENDENT | lscr:xrefUniprot | orthodb:xref /; orthodb:xrefResource ?xref. ?xref a orthodb:Uniprot. | mbgd:uniprot |
| Cross-Ref to Ensembl |  | sio:SIO\_010079/lscr:xrefEnsemblGene |  |  |
| Taxonomic Range | N/A | √ | √ (Clade) | Incomplete  (not always available) |
| GO Annotations | N/A\* | N/A\* | From UniProt per gene  (+ per group) | *goa:goa ?goa*  (per cluster) |
| Evidence Codes | N/A | N/A | N/A | ?goa rdfs:subPropertyOf+ *goa:Type* |
| Additional info | Exons, locations, rich set of external references |  | Translated Sequence Length, Evolutionary rate | Chromosome Location, Length, Cluster description |

### 

Supplementary Protocols

### **Retrieving MBGD Homologous Groups** (members can be either paralogs or orthologs between them)

HOGs in MBGD do not provide taxonomic levels at the root level of a HOG. However, the taxon NCBI identifiers of subHOGs (i.e. sublevels) can be extracted from the cluster URI and the root cluster can be identified by parsing its URI that contains the keyword “default”. Taxons are converted into valid UniProt taxonomy URIs in the query below with the BIND assignment expression. MBGD also provides orthology information of gene domains. To retrieve homologous groups related to subdomains, *orth:Gene* can be replaced with *orth:Subgene* (i.e. domain).

|  |
| --- |
| PREFIX mbgdr: <http://mbgd.genome.ad.jp/rdf/resource/>  PREFIX mbgd: <http://purl.jp/bio/11/mbgd#>  PREFIX orth: <http://purl.org/net/orth#>      SELECT DISTINCT ?hog ?gene1\_uniprot (?gene1 as ?gene1\_MBGD) ?taxLevel\_uniprot {  VALUES ?protein2\_uniprot {<http://purl.uniprot.org/uniprot/B8GA25>}  ?hog a orth:OrthologsCluster.  ?hog orth:hasHomologous+ ?node1.  ?node1 orth:hasHomologous\* ?gene1.  ?hog orth:hasHomologous\* ?gene2.  ?gene1 a orth:Gene.  OPTIONAL{?gene1 mbgd:uniprot ?gene1\_uniprot}.  ?gene2 a orth:Gene.  ?gene2 mbgd:uniprot ?protein2\_uniprot.    BIND(IF(CONTAINS(STR(?node1),"cluster"), IRI(CONCAT("http://purl.uniprot.org/taxonomy/",STRBEFORE(STRAFTER(STR(?node1),"\_tax"),"\_")) ), 0) as ?taxLevel\_uniprot)  FILTER not exists{?x orth:hasHomologous ?hog}  } ORDER BY ?taxLevel\_uniprot |

### **Aggregation queries**

### **Retrieve top 10 entries (for a given species, e.g. 'Drosophila melanogaster'), with most orthologs in OMA, MBGD, OrthoDB and EBI:**

Here we provide additional examples showing how to retrieve the top 10 entries with most orthologs in OMA and MBGD for a given species, e.g. 'Drosophila melanogaster'. These examples make use of a few more complex SPARQL features, such as aggregation (count) and ordering by a criterion in order to select the top N results.

**a)** **OMA**

|  |
| --- |
| PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>  PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>  PREFIX obo: <http://purl.obolibrary.org/obo/>  PREFIX oma: <http://omabrowser.org/ontology/oma#>  PREFIX orth: <http://purl.org/net/orth#>  PREFIX taxon: <http://purl.uniprot.org/taxonomy/>  PREFIX up: <http://purl.uniprot.org/core/>    select ?PROTEIN (**count**(distinct ?IS\_ORTHO\_TO\_PROTEIN) **as ?num\_orthologs** )  where {  ?cluster a orth:OrthologsCluster.  ?cluster orth:hasHomologousMember ?node1.  ?cluster orth:hasHomologousMember ?node2.  ?node1 orth:hasHomologousMember\* ?IS\_ORTHO\_TO\_PROTEIN.  ?node2 orth:hasHomologousMember\* ?PROTEIN.  ?PROTEIN a orth:Protein.  ?PROTEIN orth:organism/obo:RO\_0002162/up:scientificName ?species.  ?IS\_ORTHO\_TO\_PROTEIN a orth:Protein.  ?IS\_ORTHO\_TO\_PROTEIN orth:organism/obo:RO\_0002162/up:scientificName  ?other\_species .  values(?species ){( 'Drosophila melanogaster' )}  filter(?node1 != ?node2)  } group by ?PROTEIN order by DESC(?**num\_orthologs**) **limit 10** |

**b)** **MBGD**

Note: this query might take a long time to execute.

|  |
| --- |
| PREFIX orth: <http://purl.org/net/orth#>  PREFIX mbgd: <http://purl.jp/bio/11/mbgd#>  PREFIX taxon: <http://ddbj.nig.ac.jp/ontologies/taxonomy/>    select ?GENE (**count**(distinct ?IS\_ORTHO\_TO\_GENE) **as ?num\_orthologs** )  where {  ?cluster a orth:OrthologsCluster.  ?cluster orth:hasHomologous ?node1.  ?cluster orth:hasHomologous ?node2.  ?node1 orth:hasHomologous\* ?IS\_ORTHO\_TO\_GENE.  ?node2 orth:hasHomologous\* ?GENE.  ?GENE a orth:Gene.  ?GENE **mbgd**:organism/**mbgd**:inTaxon/taxon:scientificName ?species.  ?IS\_ORTHO\_TO\_GENE a orth:Gene.  ?IS\_ORTHO\_TO\_GENE mbgd:organism/mbgd:inTaxon/taxon:scientificName  ?other\_species .  values(?species ){( 'Drosophila melanogaster' )}  filter(?node1 != ?node2)  } group by ?GENE order by DESC(?**num\_orthologs**) **limit 10** |

**c) OrthoDB**

|  |
| --- |
| PREFIX : <http://purl.orthodb.org/>  select ?gene\_a (**count**( distinct ?gene\_h) as ?**num\_orthologs**)  where {  ?gene\_a a :Gene.  ?gene\_h a :Gene.  ?gene\_a up:organism/a [up:scientificName "Drosophila melanogaster"].  ?gene\_h up:organism/a ?tax.  ?gene\_a :memberOf ?og.  ?gene\_h :memberOf ?og.  ?og a :OrthoGroup.  } group by ?gene\_a order by desc(?**num\_orthologs**) limit 10 |

**d) EBI**

|  |
| --- |
| PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>  PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>  PREFIX obo: <http://purl.obolibrary.org/obo/>  PREFIX sio: <http://semanticscience.org/resource/>  PREFIX skos: <http://www.w3.org/2004/02/skos/core#>  SELECT ?gene (**count** (DISTINCT ?ortholog) as ?**num\_orthologs**) {  ?gene sio:SIO\_000558 ?ortholog .  ?gene obo:RO\_0002162 ?taxon .  ?taxon skos:prefLabel ?name.  ?ortholog obo:RO\_0002162 ?ortholog\_taxon .  VALUES(?name){( 'Drosophila melanogaster' )}  FILTER (?taxon != ?ortholog\_taxon)  } group by ?gene order by desc(?**num\_orthologs**) limit 10 |