

Assignment 5 SPARQL queries

Authors: David García Valcarce and Jimena Martín Reina

I would like you to create the SPARQL query that will answer each of these questions. **Please submit the answers as screenshots of the query plus the first 2-3 lines of response. Submit to Moodle.** NO programming is required! Use whatever SPARQL client interface you want (Jupyter, Yasgui, etc.) Thanks!

For many of these you will need to look-up how to use the SPARQL functions 'COUNT' and 'DISTINCT' (we used 'distinct' in class), and probably a few others...

UniProt SPARQL Endpoint: <http://sparql.uniprot.org/sparql> (note that you need to configure the endpoint to GET if you're using YASGUI)

Q1: 1 POINT How many protein records are in UniProt?

Your SPARQL query

Add common prefixes

```
1 > PREFIX x↔
64 SELECT (COUNT(?protein) AS ?count)
65 WHERE {
66   ?protein a up:Protein .
67 }
```

Submit Query

Results

[Sparql XML](#) [Sparql JSON](#) [CSV](#) [Show query](#) [Share](#)

count




"412838422"xsd:int

Q2: 1 POINT How many Arabidopsis thaliana protein records are in UniProt?

```

1 ▶ PREFIX x↔
64 SELECT (COUNT(?p) as ?count)
65 ▾ WHERE {
66     ?p a up:Protein;
67     up:organism <http://purl.uniprot.org/taxonomy/3702>
68 }

```


Table
Response
Pivot Table
Google Chart
Geo



Showing 1 to 1 of 1 entries (in 0.495 seconds)




	count
1	"272700"^^xsd:int

Q3: 1 POINT retrieve pictures of *Arabidopsis thaliana* from UniProt?

```

1 ▶ PREFIX x↔
64 SELECT ?picture
65 ▾ WHERE {
66     <http://purl.uniprot.org/taxonomy/3702> foaf:depiction ?picture
67 }
68

```


Table
Response
Pivot Table
Google Chart
Geo



Showing 1 to 2 of 2 entries (in 0.126 seconds)

	picture
1	https://upload.wikimedia.org/wikipedia/commons/3/39/Arabidopsis.jpg
2	https://upload.wikimedia.org/wikipedia/commons/thumb/6/60/Arabidopsis_thaliana_inflorescencias.jpg/800px-Arabidopsis_thaliana_inflorescencias.jpg

Q4: 1 POINT: What is the description of the enzyme activity of UniProt Protein Q9SZZ8

Your SPARQL query

Add common prefixes

```
1 1 > PREFIX x↔  
64  
65 SELECT ?description  
66 WHERE {  
67   VALUES ?protein { <http://purl.uniprot.org/uniprot/Q9SZZ8> }  
68   ?protein up:annotation ?annotation .  
69   ?annotation a up:Function_Annotation ;  
70               rdfs:comment ?description .  
71 }  
72
```

Submit Query

Results

[Sparql](#) [XML](#) [Sparql](#) [JSON](#) [CSV](#) [Show query](#) [Share](#)

description

"Nonheme diiron monooxygenase involved in the biosynthesis of xanthophylls. Specific for beta-ring hydroxylations of beta-carotene. Has also a low activity toward the beta- and epsilon-rings of alpha-carotene. No activity with acyclic carotenoids such as lycopene and neurosporene. Uses ferredoxin as an electron donor."[xsd:string](#)

Q5: 1 POINT: Retrieve the proteins ids, and date of submission, for 5 proteins that have been added to UniProt this year (HINT Google for “SPARQL FILTER by date”)

Your SPARQL query

Add common prefixes

```
1 1 > PREFIX x↔  
64  
65 SELECT DISTINCT ?proteinID ?submissionDate  
66 WHERE {  
67   ?proteinID up:date ?submissionDate .  
68   FILTER (STRSTARTS(STR(?submissionDate), "2023"))  
69 }  
70 LIMIT 5|
```

Submit Query

Results

[Sparql XML](#) [Sparql JSON](#) [CSV](#) [Show query](#) [Share](#)

proteinID	submissionDate
http://purl.uniprot.org/citations/36876065	"2023"xsd:gYear
http://purl.uniprot.org/citations/35942639	"2023"xsd:gYear
http://purl.uniprot.org/citations/SIPFCF38EF708ADD808	"2023-03"xsd:gYearMonth
http://purl.uniprot.org/citations/35640876	"2023"xsd:gYear
http://purl.uniprot.org/citations/36328097	"2023"xsd:gYear

Q6: 1 POINT How many species are in the UniProt taxonomy?

```
1 ► PREFIX x↔
64 SELECT (COUNT(DISTINCT ?species) as ?count)
65 ▼ WHERE {
66     ?species a up:Taxon
67 }
```

count

"2887005"xsd:int

Q7: 2 POINTS How many species have at least one protein record? (this might take a long time to execute, so do this one last!)

Your SPARQL query

Add common prefixes

```
1 ► PREFIX x↔
64
65 SELECT (COUNT(DISTINCT ?o) AS ?count)
66 ▼ WHERE {
67     ?p a up:Protein ;
68     up:organism ?o .
69 }
```

Submit Query

Results

[Sparql XML](#) [Sparql JSON](#) [CSV](#) [Show query](#) [Share](#)

count
"1305985"xsd:int

Q8: 3 POINTS find the AGI codes and gene names for all Arabidopsis thaliana proteins that have a protein function annotation description that mentions “pattern formation”

```

1  ► PREFIX x↔
64 SELECT DISTINCT ?geneName ?agi ?description
65 ▼ WHERE {
66     ?protein a up:Protein ;
67             up:organism <http://purl.uniprot.org/taxonomy/3702> ;
68             up:encodedBy ?gene ;
69             up:annotation ?annotation .
70     ?annotation a up:Function_Annotation ;
71             rdfs:comment ?description .
72     FILTER(CONTAINS(?description, "pattern formation"))
73     ?gene skos:prefLabel ?geneName ;
74             up:locusName ?agi .
75 }

```

geneName	agi	description
"RHD6" ^{xsd:string}	"At1g66470" ^{xsd:string}	"Transcription factor that is specifically required for the development of root hairs (PubMed:17556585). Acts with RSL1 to positively regulate root hair development (PubMed:17556585). Acts downstream of genes that regulate epidermal pattern formation, such as GL2 (PubMed:17556585). Targets directly RSL4, another transcription factor involved in the regulation of root hair elongation (PubMed:20139979). Acts with RSL1 as transcription factor that integrates a jasmonate (JA) signalling pathway that stimulates root hair growth (PubMed:31988260)." ^{xsd:string}
"GN" ^{xsd:string}	"At1g13980" ^{xsd:string}	"Activates the ARF proteins by exchanging bound GDP for free GTP. Plays a role in vesicular protein sorting. Acts as the major regulator of endosomal vesicle trafficking but is also involved in the endocytosis process. Could function redundantly with GNL1 in the retrograde Golgi to endoplasmic reticulum trafficking. Regulates vesicle trafficking required for the coordinated polar localization of auxin efflux carriers which in turn determines the direction of auxin flow. Mediates the sorting of PIN1 from endosomal compartments to the basal plasma membrane and the polarization of PIN3 to the bottom side of hypocotyl endodermal cells. Involved in the specification of apical-basal pattern formation in the early embryo and during root formation. Required for correct cell wall organization leading to normal cell adhesion during seedling development. Also plays an essential role in hydrotropism of seedling roots." ^{xsd:string}
"RPK1" ^{xsd:string}	"At1g69270" ^{xsd:string}	"Involved in the main abscisic acid-mediated (ABA) signalling pathway and in early ABA perception. Together with RPK2, required for pattern formation along the radial axis (e.g. the apical embryonic domain cell types that generate cotyledon primordia), and the apical-basal axis (e.g. differentiation of the basal pole during early embryogenesis)." ^{xsd:string}
"CUL3B" ^{xsd:string}	"At1g69670" ^{xsd:string}	"Component of the cullin-RING ubiquitin ligases (CRL), or CUL3-RBX1-BTB protein E3 ligase complexes which mediate the ubiquitination and subsequent proteasomal degradation of target proteins. The functional specificity of the CRL complex depends on the BTB domain-containing protein as the substrate recognition component. Involved in embryo pattern formation and endosperm development. Required for the normal division and organization of the root stem cells and columella root cap cells. Regulates primary root growth by an unknown pathway, but in an ethylene-dependent manner. Functions in distal root patterning, by an ethylene-independent mechanism. Functionally redundant with CUL3A." ^{xsd:string}
"RPK2" ^{xsd:string}	"At3g02130" ^{xsd:string}	"Key regulator of anther development (e.g. lignification pattern), including tapetum degradation during pollen maturation (e.g. germination capacity). Together with RPK1, required for pattern formation along the radial axis (e.g. the apical embryonic domain cell types that generate cotyledon primordia), and the apical-basal axis (e.g. differentiation of the basal pole during early embryogenesis)." ^{xsd:string}
"CUL3A" ^{xsd:string}	"At1g26830" ^{xsd:string}	"Component of the cullin-RING ubiquitin ligases (CRL), or CUL3-RBX1-BTB protein E3 ligase complexes which mediate the ubiquitination and subsequent proteasomal degradation of target proteins. The functional specificity of the CRL complex depends on the BTB domain-containing protein as the substrate recognition component. Involved in embryo pattern formation and endosperm development. Required for the normal division and organization of the root stem cells and columella root cap cells. Regulates primary root growth by an unknown pathway, but in an ethylene-dependent manner. Functions in distal root patterning, by an ethylene-independent mechanism. Functionally redundant with CUL3B." ^{xsd:string}
"TTL3" ^{xsd:string}	"At2g42580" ^{xsd:string}	"Involved in osmotic and salt stress tolerance. May play a role in the control of meristematic cell size during osmotic stress. May function as an adapter protein for BRL2 and may be required for signaling affecting leaf vascular tissue pattern formation." ^{xsd:string}
"IAMT1" ^{xsd:string}	"At5g55250" ^{xsd:string}	"Catalyzes the methylation of the free carboxyl end of the plant hormone indole-3-acetic acid (IAA). Converts IAA to IAA methyl ester (MeIAA). Regulates IAA activities by IAA methylation. Methylation of IAA plays an important role in regulating plant development and auxin homeostasis. Required for correct leaf pattern formation. MeIAA seems to be an inactive form of IAA." ^{xsd:string}
"ATML1" ^{xsd:string}	"At4g21750" ^{xsd:string}	"Probable transcription factor involved in cell specification and pattern formation during embryogenesis. Binds to the L1 box DNA sequence 5'-TAAATG[CT]A-3'. Plays a role in maintaining the identity of L1 cells, possibly by interacting with their L1 box or other target-gene promoters; binds to the LIP1 gene promoter and stimulates its expression upon imbibition (PubMed:24989044). Acts as a positive regulator of gibberellins (GAs)-regulated epidermal gene expression (e.g. LIP1, LIP2, LTP1, FDH and PDF1) (PubMed:24989044). Functionally redundant to PDF2 (PubMed:24989044). Seems to promote cell differentiation (PubMed:25564655)." ^{xsd:string}
"YDA" ^{xsd:string}	"At1g63700" ^{xsd:string}	"Functions in a MAP kinase cascade that acts as a molecular switch to regulate the first cell fate decisions in the zygote and the early embryo. Promotes elongation of the zygote and development of its basal daughter cell into the extra-embryonic suspensor. In stomatal development, acts downstream of the LRR receptor TMM, but upstream of the MKK4/MKK5-MPK3/MPK6 module to regulate stomatal cell fate before the guard mother cell (GMC) is specified. Plays a central role in both guard cell identity and pattern formation. This MAPK cascade also functions downstream of the ER receptor in regulating coordinated local cell proliferation, which shapes the morphology of plant organs. Upon brassinosteroid signaling, is inhibited by phosphorylation of its auto-inhibitory N-terminal domain by the GSK3-like kinase ASK7." ^{xsd:string}
"DEX1" ^{xsd:string}	"At3g09090" ^{xsd:string}	"Required for exine pattern formation during pollen development, especially for primexine deposition." ^{xsd:string}
"SCR" ^{xsd:string}	"At3g54220" ^{xsd:string}	"Transcription factor required for quiescent center cells specification and maintenance of surrounding stem cells, and for the asymmetric cell division involved in radial pattern formation in roots. Essential for cell division but not differentiation of the ground tissue. Also required for normal shoot gravitropism. Regulates the radial organization of the shoot axial organs. Binds to the promoter of MGP, NUC, RLK and SCL3. Restricts SHR movement and sequesters it into the nucleus of the endodermis." ^{xsd:string}
"ROPGAP3" ^{xsd:string}	"At2g46710" ^{xsd:string}	"Acts as a GTPase activator for the Rac-type GTPase by converting it to an inactive GDP-bound state. Involved in secondary wall pattern formation. In association with ROPGEF4, mediates local activation of ARAC10/ROP11 to initiate the distinct pattern of secondary cell walls in xylem cells." ^{xsd:string}
"RSL1" ^{xsd:string}	"At5g37800" ^{xsd:string}	"Transcription factor that is specifically required for the development of root hairs (PubMed:17556585). Acts with RHD6 to positively regulate root hair development (PubMed:17556585). Acts downstream of genes that regulate epidermal pattern formation, such as GL2 (PubMed:17556585). Acts with RHD6 as transcription factor that integrates a jasmonate (JA) signalling pathway that stimulates root hair growth (PubMed:31988260)." ^{xsd:string}
"ROPGEF7" ^{xsd:string}	"At5g02010" ^{xsd:string}	"Guanine-nucleotide exchange factor (GEF) that acts as an activator of Rop (Rho of plants) GTPases by promoting the exchange of GDP for GTP. In postembryonic roots, modulates root stem cell maintenance by regulating the expression of PLT1 and PLT2, which are key transcription factors that mediate the patterning of the root stem cell niche. May connect RopGEF-regulated Rac/Rop signaling and auxin-dependent PLT-regulated root pattern formation." ^{xsd:string}
"SWEET8" ^{xsd:string}	"At5g40260" ^{xsd:string}	"Mediates both low-affinity uptake and efflux of sugar across the plasma membrane. Required, in pollen, for microspore cell integrity and primexine pattern formation (PubMed:18434608, PubMed:25988582)." ^{xsd:string}
"SHR" ^{xsd:string}	"At4g37650" ^{xsd:string}	"Transcription factor required for quiescent center cells specification and maintenance of surrounding stem cells, and for the asymmetric cell division involved in radial pattern formation in roots. Essential for both cell division and cell specification. Regulates the radial organization of the shoot axial organs and is required for normal shoot gravitropism. Directly controls the transcription of SCR, and when associated with SCR, of MGP, RLK, TRI, NUC and SCL3."

From the MetaNetX metabolic networks for metagenomics database

SPARQL Endpoint: <https://rdf.metanetx.org/sparql>

(this slide deck will make it much easier for you!)

https://www.metanetx.org/cgi-bin/mnxget/mnxref/MetaNetX_RDF_schema.pdf

Q9: 4 POINTS: what is the MetaNetX Reaction identifier (starts with “mnxr”) for the UniProt Protein uniprotkb:Q18A79

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

Query Text

```
PREFIX meta: <https://rdf.metanetx.org/schema/>
PREFIX uniprot: <http://purl.uniprot.org/uniprot/>
SELECT DISTINCT ?reaction_identifier
WHERE{
?peptide meta:peptXref uniprot:Q18A79 .
?catalyzes meta:pept ?peptide .
?gpr meta:cata ?catalyzes ;
meta:reac ?reaction .
?reaction rdfs:label ?reaction_identifier .
FILTER regex( ?reaction_identifier, '^mnxr*', 'i') . }
```

Sponging:

Use only local data (including data retrieved before), but do not retrieve more

Results Format:

HTML

Execution timeout:

0

milliseconds (values less than 1000 are ignored)

Options:

- ☒ Strict checking of void variables
- ☐ Log debug info at the end of output (has no effect on some queries and output formats)
- ☐ Generate SPARQL compilation report (instead of executing the query)

(The result can only be sent back to browser, not saved on the server, see [details](#))

Run Query

Reset

reaction_identifier
"mnxr165934"
"mnxr145046c3"

FEDERATED QUERY - UniProt and MetaNetX

Q10: 5 POINTS: What is the official locus name, and the MetaNetX Reaction identifier (mnxr.....) for the protein that has “glycine reductase” catalytic activity in *Clostridium difficile* (taxon 272563). (this must be executed on the <https://rdf.metanetx.org/sparql> endpoint)

Query Text

```

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX mnx: <https://rdf.metanetx.org/schema/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX taxon: <http://purl.uniprot.org/taxonomy/>

SELECT DISTINCT ?locus ?reaction_id ?label
WHERE {
  SERVICE <https://sparql.uniprot.org/sparql>{
    SELECT ?uniprot_protein ?locus ?label
    WHERE{
      ?uniprot_protein a up:Protein ;
        up:organism taxon:272563 ;
        up:mnemonic ?locus ;
        up:enzyme ?enzyme .
      ?enzyme skos:prefLabel ?label .
      FILTER(CONTAINS(?label, "glycine reductase"))
    }
  }

  ?mnx_protein mnx:peptXref ?uniprot_protein .
  ?cata mnx:pept ?mnx_protein .
  ?gpr mnx:cata ?cata ;
    mnx:reac ?reaction .
  ?reaction rdfs:label ?reaction_id .
}

```

locus	reaction_id	label
"Q185M4_CL0D6"	"mnxr157884c3"	"glycine reductase"
"Q185M6_CL0D6"	"mnxr157884c3"	"glycine reductase"
"Q185M3_CL0D6"	"mnxr157884c3"	"glycine reductase"
"Q185M5_CL0D6"	"mnxr157884c3"	"glycine reductase"
"Q185M1_CL0D6"	"mnxr157884c3"	"glycine reductase"
"Q185M4_CL0D6"	"mnxr162774c3"	"glycine reductase"
"Q185M6_CL0D6"	"mnxr162774c3"	"glycine reductase"
"Q185M3_CL0D6"	"mnxr162774c3"	"glycine reductase"
"Q185M1_CL0D6"	"mnxr162774c3"	"glycine reductase"