answer

December 28, 2020

1 Question 1: How many protein records are in UniProt?

There are 322278756 records without distinct. So slowly when I run distinct query.

[4]:
```
%endpoint https://sparql.uniprot.org/sparql
%log debug
%show 20
%outfile query.log
```

```
Endpoint set to: https://sparql.uniprot.org/sparql
Logging set to DEBUG
Result maximum size: 20
Output file: /Users/pengxt/Documents/gitroom/homework/R-2020-12-16-sparql/query.log
```

[ ]:
```
# this is distinct version

PREFIX up: <http://purl.uniprot.org/core/>

SELECT (COUNT(DISTINCT ?protein) AS ?count)
WHERE
{
    ?protein a up:Protein .
}
```

[11]:
```
PREFIX up: <http://purl.uniprot.org/core/>

SELECT (COUNT(?protein) AS ?count)
WHERE
{
    ?protein a up:Protein .
}
limit 10
```

2 Question 2: How many Arabidopsis thaliana protein records are in UniProt?

The query run slowly.

[ ]:
```
# Taxon identifier of Arabidopsis thaliana protein is 3702

PREFIX up: <http://purl.uniprot.org/core/>
PREFIX taxon: <http://purl.uniprot.org/taxonomy/>


SELECT (COUNT(?protein) AS ?count)
WHERE
{
    ?protein a up:Protein .
    ?proten up:organism taxon:3702 .
}
limit 10
```

[ ]:
```
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX taxon: <http://purl.uniprot.org/taxonomy/>


SELECT (COUNT(distinct ?protein) AS ?count)
WHERE
{
    ?protein a up:Protein .
    ?proten up:organism taxon:3702 .
}
limit 10
```

3 Question 3: retrieve pictures of Arabidopsis thaliana from UniProt?

[29]:
```
# retrieve pictures from proteins and Arabidopsis thaliana
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX taxon: <http://purl.uniprot.org/taxonomy/>
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>


SELECT ?picture
WHERE
{
    ?x a up:Protein .
```

```
    ?x up:organism taxon:3702 .
    ?picture up:height ?height .
    ?picture up:width ?width
}
```

4 Question 4: What is the description of the enzyme activity of UniProt Protein Q9SZZ8?

the description of the enzyme activity of UniProt Protein Q9SZZ8 is Beta-carotene + 4 reduced ferredoxin [iron-sulfur] cluster + 2 H(+) + 2 O(2) = zeaxanthin + 4 oxidized ferredoxin [iron-sulfur] cluster + 2 H(2)O.

[8]:
```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX up: <http://purl.uniprot.org/core/>


SELECT ?protein ?enzyme ?activity ?description
WHERE
{
    ?protein a up:Protein .
    ?protein up:mnemonic "BCH1_ARATH" .
    ?protein up:enzyme ?enzyme .
    ?enzyme up:activity ?activity  .
    ?activity rdfs:label ?description .
}
```

5 Question 5: Retrieve the proteins ids, and date of submission, for proteins that have been added to UniProt this year

we only retrieve 10 records.

[2]:
```
PREFIX up:<http://purl.uniprot.org/core/>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>

SELECT ?protein ?created
WHERE
{
    ?protein a up:Protein .
    ?protein up:created ?created .
    FILTER (?created >= "2020-01-01"^^xsd:date) .
}
limit 10
```

6 Question 6: How many species are in the UniProt taxonomy?

There are 2615376 species in the UniProt taxonomy.

[18]:
```
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>

SELECT (count(distinct ?taxon) as ?count)
FROM <http://sparql.uniprot.org/taxonomy>
WHERE
{
    ?taxon a up:Taxon .
    #?taxon up:rank $species .

}
limit 10
```

7 Question 7: How many species have at least one protein record? (this might take a long time to execute, so do this one last!)

Query run so slowly.

[ ]:
```
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>

SELECT count(?taxon)
WHERE
{
    ?taxon a up:Taxon .
    ?protein up:organism ?taxon .
}
GROUP BY ?taxon
HAVING(count(?protein) > 0)
```

8 Question 8: find the AGI codes and gene names for all Arabidopsis thaliana proteins that have a protein function annotation description that mentions “pattern formation”

I only query 10 result.

[2]:
```
#AGI is Arabidopsis Gene Id
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX taxon: <http://purl.uniprot.org/taxonomy/>
```

```
SELECT ?geneName ?AGI
WHERE
{
    ?protein a up:Protein .
    ?protein up:organism taxon:3702 .

    ?protein up:alternativeName ?name .
    ?protein up:encodedBy ?gene .
    ?gene skos:prefLabel ?geneName .
    ?gene up:locusName ?AGI .
    ?protein up:annotation ?annotation .
    ?annotation a up:Function_Annotation .
    ?annotation rdfs:comment ?text .
    FILTER regex(?text, "pattern formation", "i")
}
limit 10
```

9 Question 9: what is the MetaNetX Reaction identifier (starts with "mnxr") for the UniProt Protein uniprotkb:Q18A79?

mnxr145046c3

[1]:
```
%endpoint https://rdf.metanetx.org/sparql

PREFIX uniprotkb: <http://purl.uniprot.org/uniprot/>
PREFIX mnx: <https://rdf.metanetx.org/schema/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>


select ?pept ?MNXR ?MNXR_ID
where
{
    ?pept mnx:peptXref uniprotkb:Q18A79 .
    ?cata mnx:pept ?pept ;
          rdfs:label ?cata_label .
    ?gpr mnx:cata ?cata ;
         mnx:reac ?reac .
    ?reac rdfs:label ?reac_label ;
          rdfs:comment ?reac_eq .
    ?mnet mnx:gpr ?gpr ;
          rdfs:label ?mnet_label .
    ?reac mnx:mnxr ?MNXR ;
          rdfs:label ?MNXR_ID .
}
```

```
Endpoint set to: https://rdf.metanetx.org/sparql
```

10 Question 10: What is the official Gene ID (UniProt calls this a "mnemonic") and the MetaNetX Reaction identifier (mnxr.....) for the protein that has "Starch synthase" catalytic activity in Clostridium difficile (taxon 272563).

sorry, we can't get any result. If we use "synthase" to match, we will find some result.

[8]:
```
# starch synthase activity: https://www.ebi.ac.uk/QuickGO/term/GO:0009011

PREFIX uniprotkb: <http://purl.uniprot.org/uniprot/>
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX mnx: <https://rdf.metanetx.org/schema/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX taxon: <http://purl.uniprot.org/taxonomy/>
PREFIX GO: <http://purl.obolibrary.org/obo/GO_>

select ?protein ?geneId ?enzyme ?MNXR_ID
where {
    service <https://sparql.uniprot.org/sparql> {
        ?protein a up:Protein ;
                 up:organism taxon:272563 ;
                 up:mnemonic ?geneId ;
                 up:classifiedWith|(up:classifiedWith/rdfs:subClassOf) GO:
 ↪0009011 ;
                 up:enzyme ?enzyme .
    }

    ?pept mnx:peptXref ?protein .
    ?cata mnx:pept ?pept ;
          rdfs:label ?cata_label .
    ?gpr mnx:cata ?cata ;
         mnx:reac ?reac .
    ?reac rdfs:label ?reac_label ;
          rdfs:comment ?reac_eq .
    ?mnet mnx:gpr ?gpr ;
          rdfs:label ?mnet_label .
    ?reac mnx:mnxr ?MNXR ;
          rdfs:label ?MNXR_ID .
}
limit 10
```

[ ]: