A federated query between neXtProt and OrthoDB retrieves 44 uncharacterized human proteins highly expressed in the brain and conserved in *Drosophila melanogaster.*

Lydie Lane[a,b], Kasun Samarasinghe[a] and Pierre-André Michel[b]

[a] *University of Geneva, Michel Servet 1, Geneva, 1204, Switzerland*
[b] *CALIPHO group, SIB Swiss Institute of Bioinformatics, Michel Servet 1, Geneva, 1204, Switzerland*

Abstract

To help researchers identify uncharacterized human genes that can be investigated using *Drosophila melanogaster* as a model organism, a new query federated between neXtProt and OrthoDB has been added on the neXtProt platform (NXQ_00300). The output of this query shows that there are 44 uncharacterized genes highly expressed in the human brain for which a homolog is found in *Drosophila melanogaster*.

Keywords

Human proteins, model organisms, SPARQL

1. Introduction

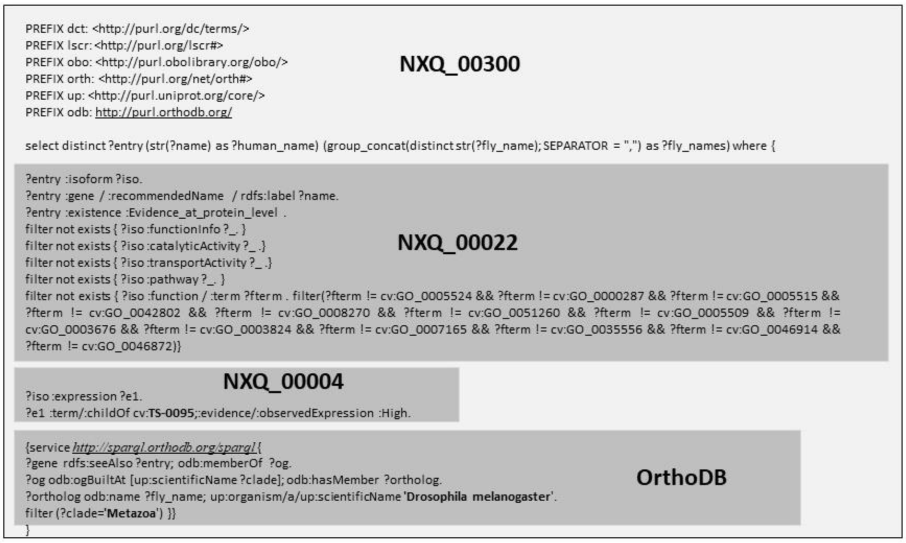
Currently, about 8% of the human protein-coding genes have no function annotated in neXtProt [1]. For half of these ~1500 uncharacterized genes, protein products have been confidently identified [2]. The HUPO Human Proteome Project recently launched an initiative to understand their functions [3]. Whereas some biological functions can be investigated in human cell lines, others require spatial and temporal integration of processes taking place in different cell types and can only be studied at the level of an organism. Investigation of such complex functions is usually performed using model organisms such as mouse or fly and the results are then extrapolated to the human protein. The choice of the model organism is governed by scientific, technical, economic, and ethical considerations. According to the current international regulations, organisms that do not experience pain, based on current scientific understanding, should be used whenever it is possible. In addition, laboratories tend to favor models that are low cost and easy to maintain. Of course, this is only possible if the protein of interest is conserved in such organisms. This information can be found in phylogenetic databases such as OrthoDB [4]. neXtProt maintains a list of ~200 tutorial SPARQL queries to support the research on human proteins. A federated query between neXtProt and OrthoDB has been added to help researchers finding a suitable model to characterize their proteins of interest.

2. Results and discussion

Since *Drosophila melanogaster* is a recognized model for neuroscience research [5], we built a query that retrieves the list of uncharacterized proteins detected by immunochemistry at high levels in the human brain that have homologs in *Drosophila melanogaster*. The query (NXQ_00300) has been added to the list of tutorial queries of the advanced neXtProt SNORQL query tool (https://snorql.nextprot.org/). It reuses parts of two pre-existing neXtProt tutorial queries : NXQ_00004, that retrieves proteins detected by immunochemistry at high levels in the brain (TS-0095 in neXtProt human anatomy vocabulary) and NXQ_00022, that retrieves entries lacking functional annotation [6].

SWAT4HCLS 2023: The 14th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences
EMAIL: lydie.lane@sib.swiss (A. 1); Kasun.Wijesiriwardana@unige.ch (A. 2); Pierre-Andre.Michel@sib.swiss (A. 3)
ORCID: 0000-0002-9818-3030 (A. 1); 0000-0002-0642-6841 (A. 2); ORCID: 0000-0002-7023-1045 (A. 3)

The *Drosophila* homologs are retrieved by the OrthoDB subquery. NXQ_00300 not only retrieves the human protein accession numbers but also the human and *Drosophila* gene names (**Figure 1**).



```
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX lscr: <http://purl.org/lscr#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX orth: <http://purl.org/net/orth#>
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX odb: http://purl.orthodb.org/

select distinct ?entry (str(?name) as ?human_name) (group_concat(distinct str(?fly_name); SEPARATOR = ",") as ?fly_names) where {

?entry :isoform ?iso.
?entry :gene / :recommendedName  / rdfs:label ?name.
?entry :existence :Evidence_at_protein_level  .
filter not exists { ?iso :functionInfo ?_. }
filter not exists { ?iso :catalyticActivity ?_ .}
filter not exists { ?iso :transportActivity ?_ .}
filter not exists { ?iso :pathway ?_. }
filter not exists { ?iso :function / :term ?fterm . filter(?fterm != cv:GO_0005524 && ?fterm != cv:GO_0000287 && ?fterm != cv:GO_0005515 &&
?fterm != cv:GO_0042802 && ?fterm != cv:GO_0008270 && ?fterm != cv:GO_0051260 && ?fterm != cv:GO_0005509 && ?fterm !=
cv:GO_0003676 && ?fterm != cv:GO_0003824 && ?fterm != cv:GO_0007165 && ?fterm != cv:GO_0035556 && ?fterm != cv:GO_0046914 &&
?fterm != cv:GO_0046872)}

?iso :expression ?e1.
?e1 :term/:childOf cv:TS-0095;:evidence/:observedExpression :High.

{service http://sparql.orthodb.org/sparql {
?gene rdfs:seeAlso ?entry; odb:memberOf ?og.
?og odb:ogBuiltAt [up:scientificName ?clade]; odb:hasMember ?ortholog.
?ortholog odb:name ?fly_name; up:organism/a/up:scientificName 'Drosophila melanogaster'.
filter (?clade='Metazoa') }}
}
```

Figure 1: SPARQL query NXQ_00300 at https://snorql.nextprot.org/

In a few seconds, query NXQ_00300 applied on neXtProt data release 2022-08-18 retrieves 44 human proteins. The results can be viewed as a html page or downloaded in json, xml or csv.

Users can explore Flybase [7] using the retrieved *Drosophila* gene names to find available mutants and their phenotypes. For some of the proteins, a synonym of the *Drosophila* gene symbol is retrieved from OrthoDB instead of the official gene name. For example, for KLHDC4 (NX_Q8TBB5), the retrieved name of the *Drosophila* homolog is anon-WO0172774.58 instead of CG4069. This example highlights the need of name standardization across the different resources. Fortunately, it is possible to explore Flybase using any name or synonym for a gene. NXQ_00300 can be adapted to any other tissue than the human brain by replacing TS-0095 by another term from the neXtProt human anatomy vocabulary, and/or to any other model organism by replacing "*Drosophila melanogaster*" by the scientific name of the organism of interest and 'Metazoa' by the appropriate clade.

3. Acknowledgements

We thank Amos Bairoch for his critical reading of the manuscript. The neXtProt and OrthoDB servers are hosted at SIB Swiss Institute of Bioinformatics in Switzerland

4. References

[1] M. Zahn-Zabal, P.-A. Michel, A. Gateau et al. (2019) The neXtProt knowledgebase in 2020: data, tools and usability improvements. Nucleic Acids Res., 48, D328–D334.

[2] S. Adhikari, E. C. Nice, E. W. Deutsch et al. (2020) A high-stringency blueprint of the human proteome. Nat. Commun., 11.

[3] G. S. Omenn, L. Lane, C. M. Overall et al. (2022) The 2022 Report on the Human Proteome from the HUPO Human Proteome Project. J. Proteome Res. doi: 10.1021/acs.jproteome.2c00498.

[4] D. Kuznetsov, F. Tegenfeldt, M. Manni, et al. (2022) OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. Nucleic Acids Res., 1, 13–14.

[5] V. Mariano, T. Achsel, C. Bagni, et al. (2020) Modelling Learning and Memory in Drosophila to Understand Intellectual Disabilities. Neuroscience, 445, 12–30.

[6] P. Duek, A. Gateau, A. Bairoch, et al. (2018) Exploring the Uncharacterized Human Proteome Using neXtProt. J. Proteome Res., 17, 4211–4226.

[7] A. Larkin, S. J. Marygold, G. Antonazzo, et al. (2021) FlyBase: Updates to the Drosophila melanogaster knowledge base. Nucleic Acids Res., 49, D899–D907.