bioRxiv preprint doi: https://doi.org/10.1101/709899; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Enzyme annotation in UniProtKB using Rhea

Anne Morgat[1]*, Thierry Lombardot[1], Elisabeth Coudert[1], Kristian Axelsen[1], Teresa Batista Neto[1], Sebastien Gehant[1], Parit Bansal[1], Jerven Bolleman[1], Elisabeth Gasteiger[1], Edouard de Castro[1], Delphine Baratin[1], Monica Pozzato[1], Ioannis Xenarios[2], Sylvain Poux[1], Nicole Redaschi[1], Alan Bridge[1], and The UniProt Consortium[1,3,4,5]

[1]Swiss-Prot group, SIB Swiss Institute of Bioinformatics, Rue Michel Servet 1, 1211-4 Geneva, Switzerland,

[2]University of Lausanne, 1015 Lausanne, Switzerland,

[3]European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK,

[4]Protein Information Resource, Georgetown University Medical Center, 3300 Whitehaven Street NW, Suite 1200, WA 20007, USA

[5]Protein Information Resource, University of Delaware, 15 Innovation Way, Suite 205, Newark DE 19711, USA

*To whom correspondence should be addressed.

Abstract

Motivation: To provide high quality computationally tractable enzyme annotation in UniProtKB using Rhea, a comprehensive expert-curated knowledgebase of biochemical reactions which describes reaction participants using the ontology ChEBI (Chemical Entities of Biological Interest).

Results: We replaced existing textual descriptions of biochemical reactions in UniProtKB with their equivalents from Rhea, which is now the standard for annotation of enzymatic reactions in UniProtKB. We developed improved search and query facilities for the UniProt website, REST API, and SPARQL endpoint that leverage the chemical structure data, nomenclature, and classification that Rhea and ChEBI provide.

Availability and Implementation: UniProtKB at https://www.uniprot.org/; UniProt REST API at https://www.uniprot.org/help/api; UniProt SPARQL endpoint at https://sparql.uniprot.org/sparql; Rhea at https://www.rhea-db.org/.

Contact: anne.morgat@sib.swiss

1 Introduction

The UniProt Knowledgebase (UniProtKB, at https://www.uniprot.org) is a reference resource of protein sequences and functional annotation that covers proteins from all branches of the tree of life (The UniProt Consortium, 2019). UniProtKB includes an expert curated core of over 560,000 reviewed UniProtKB/Swiss-Prot protein sequence entries that is supplemented by over 158 million unreviewed UniProtKB/TrEMBL entries annotated by automatic systems (release 2019_05 of June 5 2019). UniProtKB/Swiss-Prot curation focuses on experimentally characterized proteins from a broad range of taxa, including proteins of human origin (Breuza, et al., 2016) as well as proteins of bacteria, archaea, viruses, and plants. Approximately half of all protein sequence entries in UniProtKB/Swiss-Prot describe enzymes, whose function has traditionally been annotated using reference vocabularies such as the hierarchical enzyme classification of the Enzyme Nomenclature Committee of the IUBMB (often referred to as EC numbers) (Bairoch, 2000; McDonald, et al., 2009; McDonald and Tipton, 2014). In this article, we describe the introduction of a new reference vocabulary for the annotation of enzymes in UniProtKB – the Rhea knowledgebase of biochemical reactions (https://www.rhea-db.org) (Lombardot, et al., 2019; Morgat, et al., 2017). Rhea is a

bioRxiv preprint doi: https://doi.org/10.1101/709899; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

comprehensive expert-curated knowledgebase that uses the chemical ontology ChEBI (Chemical Entities of Biological Interest, https://www.ebi.ac.uk/chebi/) (Hastings, et al., 2016) to describe reaction participants, their chemical structures, and chemical transformations. Rhea provides stable unique identifiers and computationally tractable descriptions for around 12,000 unique biochemical reactions, including reactions of the IUBMB enzyme classification, as well as thousands of additional enzymatic, transport reactions, and spontaneous reactions. The introduction of Rhea as the reference vocabulary for enzyme annotation in UniProtKB will significantly improve the coverage and precision of enzyme annotation and will allow users of UniProtKB to leverage chemical structure data and classifications to answer biological questions.
In the following we describe the re-annotation of legacy enzyme data in UniProtKB using Rhea and a selection of the new tools and services we have developed to exploit this enhanced enzyme dataset.

2 Methods

2.1 Rhea as a reference vocabulary for enzyme annotation in UniProtKB

Prior to this work, UniProtKB used the hierarchical enzyme classification of the Enzyme Nomenclature Committee of the IUBMB (hereafter referred to simply as the enzyme classification) as the main reference vocabulary for enzyme annotation. The enzyme classification uses a hierarchy of exactly four levels to classify enzymes according to the chemistry of representative reactions, which are described in text form (Bairoch, 2000). In this work we introduce Rhea reaction identifiers as the reference vocabulary for enzyme annotation in UniProtKB, with the corresponding four digit enzyme class (EC number) selected from a distinct mapping of Rhea reactions to EC numbers (maintained by Rhea at https://www.rhea-db.org/download). The EC number annotations for Rhea reactions are optional in UniProtKB, as Rhea contains thousands of reactions not described by the IUBMB Enzyme Classification. Note that UniProtKB will continue to describe some enzymatic reactions in text form, particularly those reactions whose specific chemistry is not yet well characterized and which cannot be described using ChEBI.

2.2 Migration of legacy enzyme annotation in UniProtKB to Rhea and re-annotation

In order to lay the groundwork for the integration of Rhea in UniProtKB, we first mapped legacy textual descriptions of enzymatic reactions in UniProtKB to Rhea reaction identifiers. We accomplished this using the ENZYME database (Bairoch, 2000), which links textual reaction descriptions (as used in UniProtKB) to their corresponding EC numbers, and the Rhea database, which links EC numbers to their corresponding Rhea reactions. We checked and validated all such mappings of UniProtKB enzyme annotation – EC number – Rhea identifier derived in this way. A small number of legacy UniProtKB enzyme annotations were not based on EC numbers, and we mapped these textual descriptions manually to Rhea identifiers where possible, creating new Rhea reactions where needed. We then used the completed mapping to replace the legacy textual descriptions of enzymatic reactions in UniProtKB by Rhea annotations, and to update all automatic annotation rules that are used to add enzyme annotations to UniProtKB/TrEMBL records, including those from HAMAP (Pedruzzi, et al., 2015) and PROSITE (Sigrist, et al., 2013). Mapping of all EC number annotations is now complete, while the mapping of additional legacy enzyme data described in natural language in UniProtKB/Swiss-Prot (mainly in “FUNCTION” annotations) is still ongoing.

2.3 UniProt tools and services that use Rhea

We modified the UniProt data model and output formats – text, XML, and RDF – to include Rhea reaction data and references to ChEBI. We modified the UniProt website https://www.uniprot.org/ (Jain, et al., 2009), UniProt REST API

bioRxiv preprint doi: https://doi.org/10.1101/709899; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

https://www.uniprot.org/help/api and SPARQL endpoint https://sparql.uniprot.org/to support searches using Rhea and ChEBI identifiers as well as ChEBI names, synonyms, and chemical structures represented as InChIKeys. The InChIKey, a simple hash representation of chemical structures, provides a convenient means to search and map chemical structure databases. The first layer of the InChIKey encodes information on connectivity, the second layer encodes information on stereochemistry, and the third encodes information on charge. A more complete description of the InChIKey is available at https://www.inchi-trust.org/.

3 Results

3.1 Annotation of UniProtKB using Rhea

We performed a complete re-annotation of legacy UniProtKB enzyme data using Rhea (as described in Methods), and now use Rhea as the primary reference vocabulary for enzyme annotation in UniProtKB. UniProtKB/Swiss-Prot currently includes annotations for 5,795 unique Rhea reactions (around 49% of all Rhea reactions), which feature in 215,789 distinct UniProtKB/Swiss-Prot protein records (38.5% of all UniProtKB/Swiss-Prot records are annotated with Rhea) (release 2019_05 of June 5 2019). Of the 5,795 Rhea reactions used in UniProtKB/Swiss-Prot, 4,828 reactions (around 85% of all reactions) are linked to EC numbers. We are currently working to improve the coverage of the 6,140 Rhea reactions not currently represented in UniProtKB/Swiss-Prot through a variety of approaches. These approaches include continuing expert curation of new literature, re-annotation of free text legacy annotations in UniProtKB/Swiss-Prot entries, and integration of data from other resources that use Rhea. One such resource of note is the SwissLipids knowledgebase (Aimo, et al., 2015), which contains annotations for more than 2,000 unique Rhea reactions that are not yet represented in UniProtKB. We will describe these and other annotation efforts in more detail in forthcoming publications.

3.2 UniProt tools and services that use Rhea

Below we describe how users can navigate and exploit Rhea data using the UniProt website, REST API, and SPARQL endpoint.

3.2.1 Rhea and the UniProt website

The UniProt website https://www.uniprot.org constitutes the main point of entry for most UniProt users. The UniProtKB entry view provides a summary of annotated Rhea reactions for each enzyme (Figure 1). Users can choose to reveal the two-dimensional structures of reaction participants for each annotated reaction, and click on the reactions and their participants to launch searches in UniProtKB or Rhea or link out to ChEBI.

bioRxiv preprint doi: https://doi.org/10.1101/709899; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

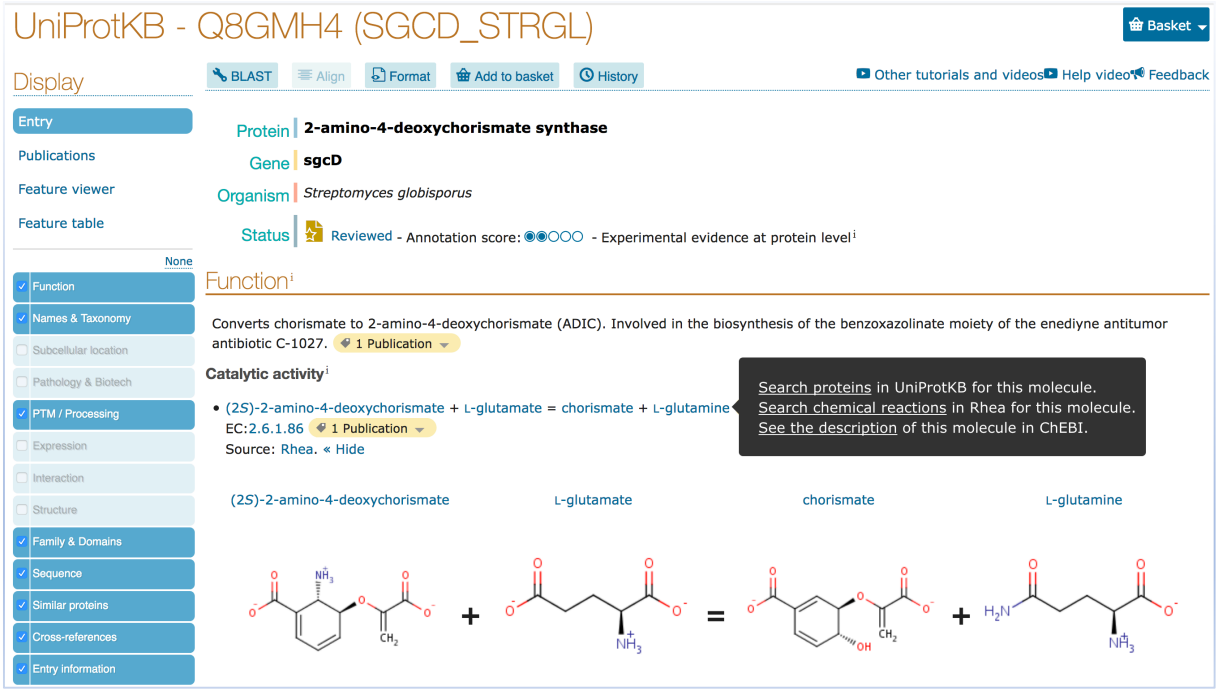


Figure 1. UniProtKB entry view showing Rhea annotation for the Streptomyces globisporus enzyme 2-amino-4-deoxychorismate synthase (UniProt:Q8GMH4 annotated with RHEA:25512). The search and link-out options available for each reaction participant are illustrated using “L-glutamine”; users can search in UniProtKB or Rhea, or link out to ChEBI to learn more about the metabolite in question. We omit most sections for clarity.

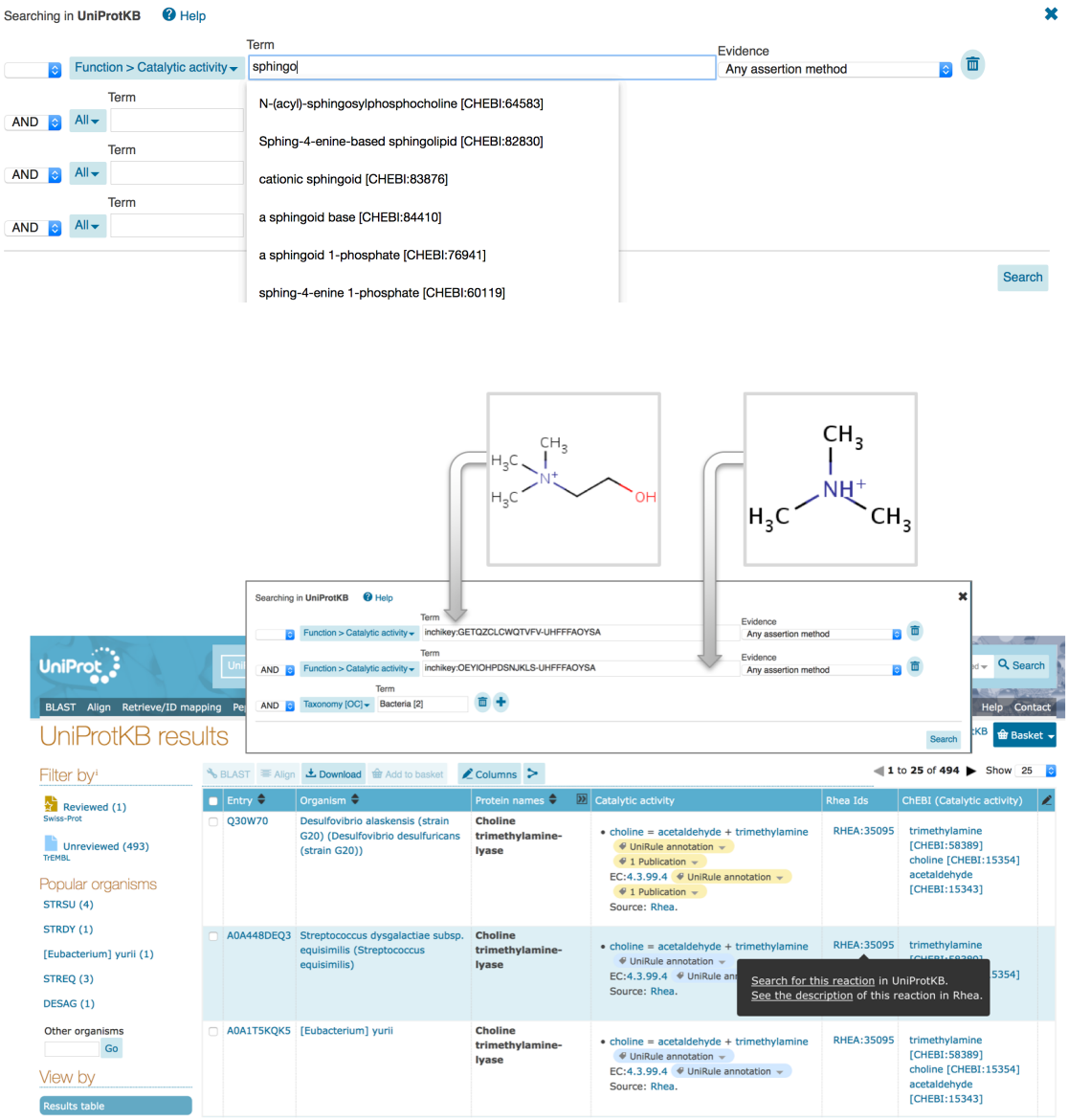
Figure 2 illustrates selected advanced search options for reactions, chemical names and structures in UniProtKB. Users can search for identifiers from Rhea as well as identifiers, names, synonyms (Figure 2a), and chemical structures (encoded as InChIKeys) from ChEBI (Figure 2b). The complete ChEBI ontology is indexed to support hierarchical searches, while ChEBI identifiers entered by users are mapped to those of the major species at pH 7.3, the form used in Rhea, using the mapping provided at https://www.rhea-db.org/download. InChIKeys provide a means to query UniProtKB using chemical structure data from other resources, including reference knowledgebases such as the Human Metabolome Database (HMDB) (Wishart, et al., 2018) or LIPID MAPS (Fahy, et al., 2009). Users of these and other resources can simply convert their structures to InChIKeys and use them to query UniProtKB. Searches may be performed using the complete InChIKey (to find exact structure matches), or using the first and second layers of the InChIKey (to find molecules with matching connectivity and stereochemical orientation, irrespective of charge, as in Figure 2b), or using only the first layer of the InChIKey (to find molecules with matching connectivity).

Figure 2b illustrates an InChIKey search for microbial enzymes that metabolize choline (HMDB00097, InChIKey=OEYIOHPDSNJKLS-UHFFFAOYSA-N) and trimethylamine (TMA) (HMDB0000906, InChIKey=GETQZCLCWQTVFV-UHFFFAOYSA-N). Host gut microbes produce TMA from choline, while the human liver enzyme FMO3 then converts TMA to the pro-atherogenic molecule trimethylamine N-oxide (TMAO) (HMDB0000925) (Canyelles, et al., 2018; Chhibber-Goel, et al., 2017). This InChIKey search allows users of HMDB to exploit protein annotation in UniProtKB to connect the bacterial choline trimethylamine-lyase *cutC* of *Desulfovibrio alaskensis* (UniProtKB:Q30W70) and homologs to human enzymes such as FMO3 by their common metabolites.
For those chemical structures that have no corresponding match annotated in UniProtKB – no matching InChIKey first layer – users might choose instead to search for relevant

bioRxiv preprint doi: https://doi.org/10.1101/709899; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

information about the chemical classes to which these structures belong. Chemical structures can be mapped to their corresponding ChEBI classes using tools such as ClassyFire (Djoumbou Feunang, et al., 2016), and the ChEBI classes then used to search UniProtKB. Interested readers can find further information about chemical data search in UniProtKB in the online documentation provided at https://www.uniprot.org/help/chemical_data_search.

a)



b)

Figure 2. Advanced search in UniProtKB for enzymatic reactions.
a) Advanced search using chemical nomenclature. The autocomplete feature is shown.
b) Advanced search using InChIKeys of choline (HMDB00097) and trimethylamine (TMA) (HMDB0000906) to identify bacterial enzymes metabolizing both compounds. The result table can be customized to display Rhea reaction data, which can be used to launch further searches and link out precisely as in the entry view.

3.2.2 Rhea and the UniProt REST API

The UniProt website serves a REST API (https://www.uniprot.org/help/api) that allows users to query and process data programmatically. The REST API has been modified to handle

bioRxiv preprint doi: https://doi.org/10.1101/709899; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Rhea and ChEBI identifiers, as well as ChEBI names, synonyms, and structural data. The sample query below will retrieve the aforementioned bacterial enzymes metabolizing choline and trimethylamine using their respective InChIKeys (see also https://tinyurl.com/y2mcjotd).

```
annotation:(type:"catalytic activity" inchikey:OEYIOHPDSNJKLS-UHFFFAOYSA)
annotation:(type:"catalytic activity" inchikey:GETQZCLCWQTVFV-
UHFFFAOYSA) taxonomy:"Bacteria [2]"
```

As before, this particular query uses only the first and second layers of the InChIKey to allow permissive matching between different charge state representations. Users can modify REST queries like that shown above in order to specify both the required data output (annotation fields) and format (such as .tab, .xls, .rdf, and others) (for more details see https://www.uniprot.org/help/api_queries).

3.2.3 Rhea and the UniProt SPARQL endpoint

The UniProt SPARQL endpoint allows users to perform complex federated queries that combine RDF data from UniProt with that from other SPARQL endpoints. Like the UniProt website and REST API, the UniProt SPARQL endpoint now supports queries using Rhea and ChEBI identifiers, as well as ChEBI names, synonyms, and structural data. Figure 3 provides a sample SPARQL query that combines the UniProt, Rhea (Lombardot, et al., 2019) and ChEMBL (Gaulton, et al., 2017) SPARQL endpoints. The aforementioned pro-atherogenic metabolite TMAO, produced from choline and TMA by a combination of microbial and human enzymes, may promote atherosclerosis by perturbing enterohepatic cholesterol metabolism and macrophage cholesterol transport (Canyelles, et al., 2018; Chhibber-Goel, et al., 2017). This query will find drugs in ChEMBL that target human enzymes acting on sterols (members of the ChEBI class CHEBI:15889).

```
PREFIX rh:<http://rdf.rhea-db.org/>
PREFIX ch:<http://purl.obolibrary.org/obo/>
PREFIX ch2:<http://purl.obolibrary.org/obo/chebi#>
PREFIX up:<http://purl.uniprot.org/core/>
PREFIX keywords:<http://purl.uniprot.org/keywords/>
PREFIX taxon:<http://purl.uniprot.org/taxonomy/>
PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
PREFIX cco: <http://rdf.ebi.ac.uk/terms/chembl#>

SELECT DISTINCT ?protein ?proteinFullName ?activityType
?standardActivityValue ?standardActivityUnit ?chemblMolecule
?chemlbMoleculePrefLabel
WHERE {
  SERVICE <https://sparql.rhea-db.org/sparql> {
    ?reaction rdfs:subClassOf rh:Reaction .
    ?reaction rh:status rh:Approved .
    ?reaction rh:side ?reactionSide .
    ?reactionSide rh:contains ?participant .
    ?participant rh:compound ?compound .
    { ?compound rh:chebi ?chebi .
      ?chebi rdfs:subClassOf+ ch:CHEBI_15889 . }
    UNION
    { ?compound rh:chebi ?chebi .
      ?chebi2 rdfs:subClassOf ?chebiRestriction .
      ?chebiRestriction a owl:Restriction .
      ?chebiRestriction owl:onProperty ch2:has_major_microspecies_at_pH_7_3 .
      ?chebiRestriction owl:someValuesFrom ?chebi .
```

bioRxiv preprint doi: https://doi.org/10.1101/709899; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

```
      ?chebi2 rdfs:subClassOf+ ch:CHEBI_15889 . }
  }
  ?ca up:catalyzedReaction ?reaction .
  ?a up:catalyticActivity ?ca .
  ?protein up:annotation ?a .
  ?protein up:organism taxon:9606 .
  ?protein up:classifiedWith keywords:1185 .
  ?protein up:recommendedName ?proteinRecName .
  ?proteinRecName up:fullName ?proteinFullName .
  SERVICE <https://www.ebi.ac.uk/rdf/services/sparql> {
    ?activity a cco:Activity .
    ?activity cco:hasMolecule ?chemblMolecule .
    ?chemblMolecule cco:highestDevelopmentPhase 4 .
    ?chemblMolecule rdfs:label ?chemblMoleculeLabel .
    ?chemblMolecule skos:prefLabel ?chemlbMoleculePrefLabel .
    ?activity cco:hasAssay ?assay .
    ?activity cco:standardType ?activityType .
    ?activity cco:standardValue ?standardActivityValue .
    ?activity cco:standardUnits ?standardActivityUnit .
    ?assay cco:hasTarget ?target .
    ?target cco:hasTargetComponent ?targetcmpt .
    ?targetcmpt cco:targetCmptXref ?protein .
    ?protein a cco:UniprotRef .
  }
}
```

Figure 3. A sample federated SPARQL query that leverages Rhea annotation in UniProtKB. The query retrieves information about drugs that target enzymes involved in human sterol (CHEBI:15889) metabolism from the UniProt, Rhea and ChEMBL SPARQL endpoints, federating the three SPARQL endpoints with two SERVICE calls.

Other federated queries might explore metabolism in the context of genomic organization and variation using resources such as Ensembl (Zerbino, et al., 2018) or how metabolism evolves using resources such as OMA (Altenhoff, et al., 2018) and OrthoDB (Kriventseva, et al., 2019). We describe more sample queries in the documentation available at the UniProt SPARQL endpoint.

4 Conclusions and futures directions

Here we describe the introduction of Rhea as the reference vocabulary for enzyme annotation in UniProtKB. This development will significantly increase the utility of UniProtKB for applications such as the integrative analysis of metabolomics and other 'omics data (Kale, et al., 2016; Sud, et al., 2016), as illustrated here, as well as the study of enzyme chemistry and evolution (Rahman, et al., 2016; Tyzack, et al., 2019), the construction and annotation of metabolic models (Cottret, et al., 2018; King, et al., 2016; Moretti, et al., 2016), and the engineering of pathways for biosynthesis and bioremediation (Duigou, et al., 2019). Rhea will also provide a basis to further improve links with other knowledge resources, such as the Gene Ontology (The Gene Ontology Consortium, 2019) and Reactome (Fabregat, et al., 2018), which both plan to use Rhea as the reference for reaction chemistry (Chris Mungall and Peter D'Eustachio, personal communication). We also plan to extend the use of ChEBI within UniProtKB to describe all small molecule chemical structure data, including functionally important ligands and post-translational modifications (see https://www.uniprot.org/docs/ptmlist). This will further improve links and interoperability between UniProtKB and other resources that use ChEBI such as the metabolomics data repository MetaboLights (Kale, et al., 2016), the IMEx molecular interaction databases

bioRxiv preprint doi: https://doi.org/10.1101/709899; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(Orchard, et al., 2012) and the Complex Portal (Meldal, et al., 2015), the literature annotation services of Europe PubMed Central (Europe PMC Consortium, 2015), the BioModels repository (Glont, et al., 2018), and the Immune Epitope Database (IEDB) (Dhanda, et al., 2019).

Acknowledgements

We thank Marco Pagni of the Vital-IT group of the SIB Swiss Institute of Bioinformatics for critical reading of the manuscript and helpful comments, and the Cheminformatics and Metabolism Team of EMBL-EBI for their work in maintaining and developing ChEBI.

Funding

UniProt is supported by the National Eye Institute (NEI), National Human Genome Research Institute (NHGRI), National Heart, Lung, and Blood Institute (NHLBI), National Institute on Aging (NIA), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of General Medical Sciences (NIGMS), and National Institute of Mental Health (NIMH) [U24HG007822]. UniProt activities at the SIB are also supported by the Swiss Federal Government through the State Secretariat for Education, Research and Innovation SERI. Additional support for the EMBL-EBI's involvement in UniProt comes from European Molecular Biology Laboratory (EMBL) core funds, the British Heart Foundation (BHF) [RG/13/5/30112], the Parkinson's disease United Kingdom (PDUK) [G-1307], the NHGRI [U41HG02273], the Biotechnology and Biological Sciences Research Council (BBSRC) [BB/M011674/1], and Open Targets. PIR's UniProt activities are also supported by the NIGMS [R01GM080646, G08LM010720, P20GM103446], and the National Science Foundation (NSF) [DBI-1062520]. Rhea is supported by the Swiss Federal Government through the State Secretariat for Education, Research and Innovation (SERI); SwissLipids project of the SystemsX.ch, the Swiss Initiative in Systems Biology (in part); EMBL; ELIXIR Implementation study on 'A microbial metabolism resource for Systems Biology' (in part).

References

Aimo, L., et al. The SwissLipids knowledgebase for lipid biology. Bioinformatics 2015;31(17):2860-2866.
Altenhoff, A.M., et al. The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucleic Acids Res 2018;46(D1):D477-D485.
Bairoch, A. The ENZYME database in 2000. Nucleic Acids Res 2000;28(1):304-305.
Breuza, L., et al. The UniProtKB guide to the human proteome. Database (Oxford) 2016;2016.
Canyelles, M., et al. Trimethylamine N-Oxide: A Link among Diet, Gut Microbiota, Gene Regulation of Liver and Intestine Cholesterol Homeostasis and HDL Function. Int J Mol Sci 2018;19(10).
Chhibber-Goel, J., et al. The Metabolite Trimethylamine-N-Oxide is an Emergent Biomarker of Human Health. Curr Med Chem 2017;24(36):3942-3953.
Cottret, L., et al. MetExplore: collaborative edition and exploration of metabolic networks. Nucleic Acids Res 2018;46(W1):W495-W502.
Dhanda, S.K., et al. IEDB-AR: immune epitope database-analysis resource in 2019. Nucleic Acids Res 2019;47(W1):W502-W506.
Djoumbou Feunang, Y., et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminform 2016;8:61.
Duigou, T., et al. RetroRules: a database of reaction rules for engineering biology. Nucleic Acids Res 2019;47(D1):D1229-D1235.
Europe PMC Consortium. Europe PMC: a full-text literature database for the life sciences and platform for innovation. Nucleic Acids Res 2015;43(Database issue):D1042-1048.

bioRxiv preprint doi: https://doi.org/10.1101/709899; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Fabregat, A., et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res 2018;46(D1):D649-D655.
Fahy, E., et al. Update of the LIPID MAPS comprehensive classification system for lipids. J Lipid Res 2009;50 Suppl:S9-14.
Gaulton, A., et al. The ChEMBL database in 2017. Nucleic Acids Res 2017;45(D1):D945-D954.
Glont, M., et al. BioModels: expanding horizons to include more modelling approaches and formats. Nucleic Acids Res 2018;46(D1):D1248-D1253.
Hastings, J., et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res 2016;44(D1):D1214-1219.
Jain, E., et al. Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics 2009;10:136.
Kale, N.S., et al. MetaboLights: An Open-Access Database Repository for Metabolomics Data. Curr Protoc Bioinformatics 2016;53:14 13 11-18.
King, Z.A., et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res 2016;44(D1):D515-522.
Kriventseva, E.V., et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res 2019;47(D1):D807-D811.
Lombardot, T., et al. Updates in Rhea: SPARQLing biochemical reaction data. Nucleic Acids Res 2019;47(D1):D596-D600.
McDonald, A.G., Boyce, S. and Tipton, K.F. ExplorEnz: the primary source of the IUBMB enzyme list. Nucleic Acids Res 2009;37(Database issue):D593-597.
McDonald, A.G. and Tipton, K.F. Fifty-five years of enzyme classification: advances and difficulties. FEBS J 2014;281(2):583-592.
Meldal, B.H., et al. The complex portal--an encyclopaedia of macromolecular complexes. Nucleic Acids Res 2015;43(Database issue):D479-484.
Moretti, S., et al. MetaNetX/MNXref--reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. Nucleic Acids Res 2016;44(D1):D523-526.
Morgat, A., et al. Updates in Rhea - an expert curated resource of biochemical reactions. Nucleic Acids Res 2017;45(D1):D415-D418.
Orchard, S., et al. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat Methods 2012;9(4):345-350.
Pedruzzi, I., et al. HAMAP in 2015: updates to the protein family classification and annotation system. Nucleic Acids Res 2015;43(Database issue):D1064-D1070.
Rahman, S.A., et al. Reaction Decoder Tool (RDT): extracting features from chemical reactions. Bioinformatics 2016;32(13):2065-2066.
Sigrist, C.J., et al. New and continuing developments at PROSITE. Nucleic Acids Res 2013;41(Database issue):D344-347.
Sud, M., et al. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Res 2016;44(D1):D463-470.
The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res 2019;47(D1):D330-D338.
The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47(D1):D506-D515.
Tyzack, J.D., et al. Exploring Enzyme Evolution from Changes in Sequence, Structure, and Function. Methods Mol Biol 2019;1851:263-275.
Wishart, D.S., et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res 2018;46(D1):D608-D617.
Zerbino, D.R., et al. Ensembl 2018. Nucleic Acids Res 2018;46(D1):D754-D761.