The Rhea SPARQL endpoint supports integrated analyses of genome, proteome and metabolome

Thierry Lombardot[1[0000-0003-4157-0029]], Anne Morgat[1[0000-0002-1216-2969]], Sebastien Gehant[1[0000-0003-1608-9954]], Nicole Redaschi[1[0000-0001-8890-2268]] and Alan Bridge[1[0000-0003-2148-9135]]

[1] Swiss-Prot Group, SIB Swiss Institute of Bioinformatics, CMU, 1 rue Michel-Servet, CH-1211 Geneva 4, Switzerland
thierry.lombardot@sib.swiss, anne.morgat@sib.swiss, sebastien.gehant@sib.swiss, nicole.redaschi@sib.swiss, alan.bridge@sib.swiss

Abstract. 'Omics technologies enable comprehensive analyses of how genetic variation (genomics), gene expression patterns (transcriptomics), proteins and complexes (proteomics), and small molecules (metabolomics) interact in complex biological systems. RDF and SPARQL can connect these 'omics domains and help us realize the full potential of this valuable experimental data.

Rhea is a comprehensive and non-redundant resource of expert-curated biochemical reactions based on the ChEBI ontology. Here we demonstrate how to link genome (Ensembl), proteome (UniProt), and metabolome (ChEBI) data using federated queries that leverage the SPARQL endpoint of Rhea (https://sparql.rhea-db.org/sparql) and data from these other resources.

Keywords: Biochemical reactions, metabolites, RDF, Rhea, SPARQL.

1 Introduction

RDF and SPARQL, two core technologies of the Semantic Web, offer the possibility to perform complex queries across distinct data and knowledge resources, integrating distinct 'omics domains to generate new insights. Many core resources in the life sciences, including UniProt (https://sparql.uniprot.org), support these technologies.

Here we describe our work on the development of an RDF representation and SPARQL endpoint for the Rhea database of biochemical reactions [1, 2]. Rhea uses chemical entities from the ChEBI ontology (https://www.ebi.ac.uk/chebi) to represent reaction participants. We show how federated queries across the Rhea and UniProt SPARQL endpoints can further integrate knowledge of genome, proteome, and metabolome to improve our understanding of the interactions in complex biological systems.

2 Rhea RDF and SPARQL endpoint

A detailed description of the Rhea data model is available at our website (https://www.rhea-db.org/rhea_rdf_documentation.pdf). The data can be downloaded in RDF/XML format at ftp://ftp.expasy.org/databases/rhea/rdf/ or directly queried at

the Rhea SPARQL endpoint (https://sparql.rhea-db.org/sparql), which uses Virtuoso software (https://virtuoso.openlinksw.com) and is hosted at the Vital-IT Center for high-performance computing (https://www.vital-it.ch) of the SIB Swiss Institute of Bioinformatics.

The Rhea SPARQL endpoint can be used to answer a variety of biological questions, from building reaction networks for UniProt proteomes to inferring possible enzymes for unannotated metabolites and many more [1]. Below we provide a sample SPARQL query that returns a list of diseases linked to a specific class of metabolite - the glycosphingolipids (CHEBI:24402) - as well as the relevant genes (Ensembl) and proteins, using enzyme annotations from UniProtKB/Swiss-Prot.

```
PREFIX rh:<http://rdf.rhea-db.org/>
PREFIX taxon:<http://purl.uniprot.org/taxonomy/>
PREFIX ec:<http://purl.uniprot.org/enzyme/>
PREFIX up:<http://purl.uniprot.org/core/>
PREFIX chebi:<http://purl.obolibrary.org/obo/CHEBI_>
PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
SELECT distinct(?disease) ?reaction ?metabolite ?protein ?gene WHERE {
   ?reaction rdfs:subClassOf rh:Reaction ; rh:status rh:Approved ;
      rh:ec ?ec ;
      rh:side/rh:contains/rh:compound/rh:chebi ?metabolite .
   ?metabolite rdfs:subClassOf+ chebi:24402 .
   SERVICE <https://sparql.uniprot.org/sparql/> {
      ?protein a up:Protein ; up:reviewed true ; up:organism taxon:9606 ;
         (up:domain|up:component)?/up:enzyme ?ec ;
         up:annotation [ up:disease [ skos:prefLabel ?disease ] ] ;
         rdfs:seeAlso [ up:transcribedFrom ?gene ] . } }
```

This query provides a means to explore the relations between specific classes of metabolites and diseases, and could be enhanced with information on disease classifications and phenotype data. The Rhea SPARQL endpoint and our recent publication provide more examples of complex and federated queries for Rhea users. Current work on the incorporation of Rhea as an annotation vocabulary for UniProtKB will improve the depth and precision of enzyme annotation and further enhance the power of queries using both resources.

References

1. Lombardot, T., Morgat, A., Axelsen, K.B., Aimo, L., Hyka-Nouspikel. N., Niknejad, A., Ignatchenko, A., Xenarios, I., Coudert, E., Redaschi, N., Bridge, A.: Updates in Rhea: SPARQLing biochemical reaction data. Nucleic Acids Res., gky876 (2018).
2. Morgat, A.., Lombardot, T., Axelsen, K.B., Aimo, L., Niknejad, A., Hyka-Nouspikel, N., Coudert, E., Pozzato, M., Pagni, M., Moretti, S., et al.: Updates in Rhea - an expert curated resource of biochemical reactions. Nucleic Acids Res., 45, D415–D418 (2017).