UCL

UNIL | Université de Lausanne

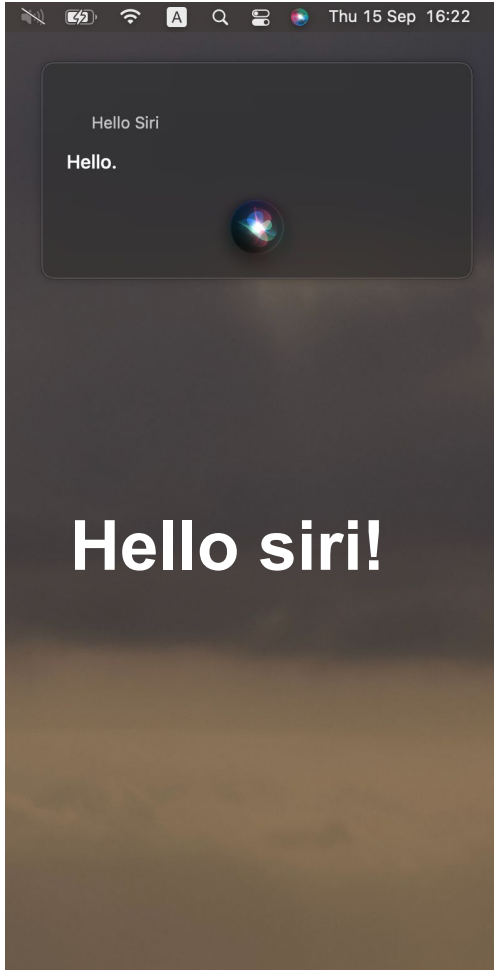
SIB Swiss Institute of Bioinformatics

BIOQA: toward a benchmark dataset of biological questions/answers involving orthology, gene expression, and complementary omics data

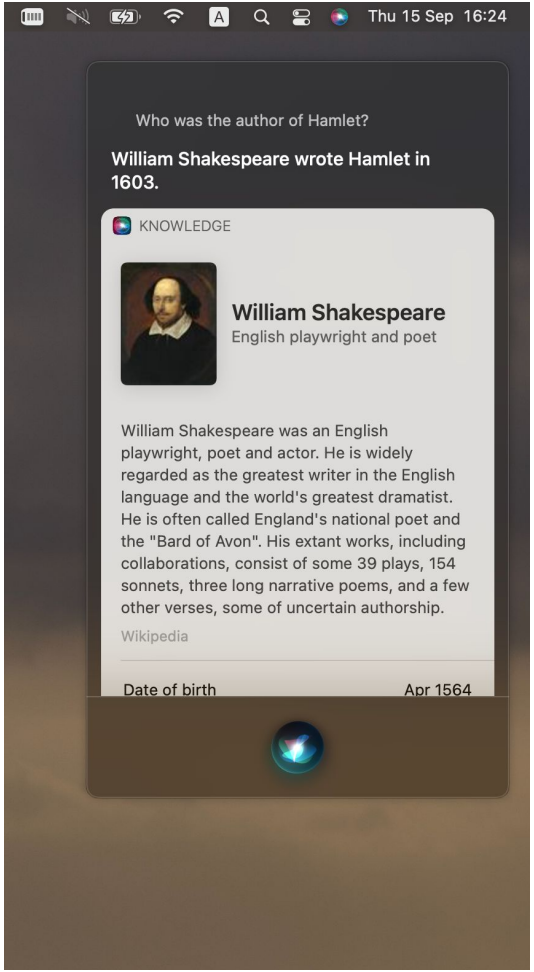
Borbala Banfalvi, Petros Liakopoulos, Xinyi Wang, Christophe Dessimoz,
Sina Majidian, Ana Claudia Sima

DESSIMOZ LAB
UNIL
UCL

@DessimozLab
@SinaMajidian


Thu 15 Sep 16:22
Hello Siri
Hello.
Hello siri!


Thu 15 Sep 16:25
Who was the author of Hamlet?


Thu 15 Sep 16:24
Who was the author of Hamlet?
William Shakespeare wrote Hamlet in 1603.
KNOWLEDGE
William Shakespeare
English playwright and poet
William Shakespeare was an English playwright, poet and actor. He is widely regarded as the greatest writer in the English language and the world's greatest dramatist. He is often called England's national poet and the "Bard of Avon". His extant works, including collaborations, consist of some 39 plays, 154 sonnets, three long narrative poems, and a few other verses, some of uncertain authorship.
Wikipedia
Date of birth
Apr 1564

Thu 15 Sep 11:18

what are the homologs of Hemoglobin subunit beta?

Here's what I found.

WEBSITES

3043 - Gene ResultHBB hemoglobin subunit beta [(human)] - NCBI
ncbi.nlm.nih.gov

HBB gene: MedlinePlus Genetics
medlineplus.gov

HBB Gene - GeneCards | HBB Protein | HBB Antibody
genecards.org

Hemoglobin subunit beta | DrugBank Online
go.drugbank.com

Entry - *141900 - HEMOGLOBIN--BETA LOCUS; HBB - OMIM
omim.org

See more in Safari...

WolframAlpha computational intelligence

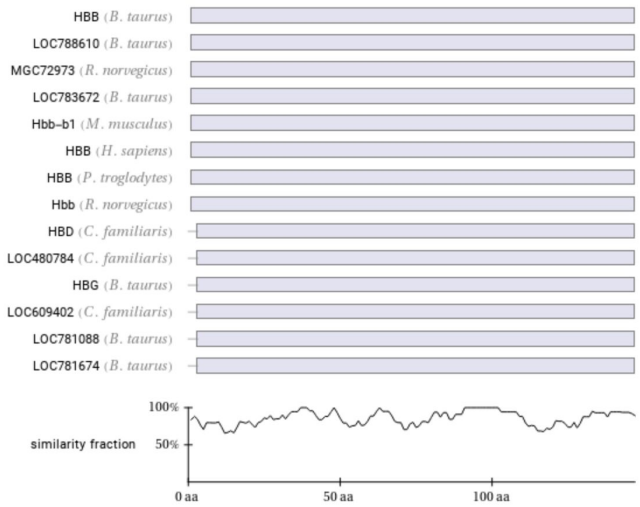
what are the homologs of Hemoglobin subunit beta

NATURAL LANGUAGE | MATH INPUT | EXTENDED KEYBOARD | EXAMPLES | UPLOAD | RANDOM

Input interpretation

HBB (cow gene) | homologs across organisms

Result | Show legend

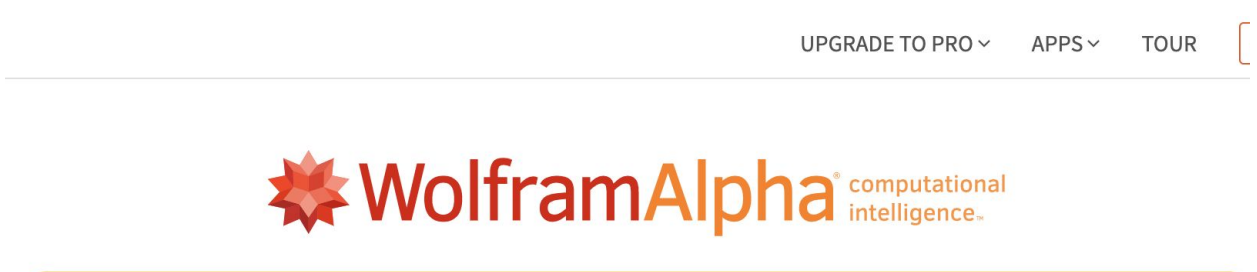


(shading indicates regions with similar amino acids)

(no homologs found in *A. thaliana, O. sativa, S. pombe, S. cerevisiae, N. crassa, P. falciparum, C. elegans, A. gambiae, D. melanogaster, D. rerio, G. gallus, K. lactis, E. gossypii,* or *M. grisea*)

Sources | Download Page

POWERED BY THE **WOLFRAM LANGUAGE**


UPGRADE TO PRO
APPS
TOUR
WolframAlpha computational intelligence
Which are the mouse's genes expressed in the lung and are orthologous to human's TAL1 gene?
NATURAL LANGUAGE
MATH INPUT
EXTENDED KEYBOARD
EXAMPLES
UPLOAD
RANDOM
Interpreting as: human's TAL1 gene
Input interpretation
TAL1 (human gene)
Standard name
T–cell acute lymphocytic leukemia 1
Alternate names
More
SCL | TCL5 | tal–1 | bHLHa17 | …
Location
genome build 37
locus
chromosome 1 | p32
strand
minus
coordinates
47 681 963 to 47 695 443

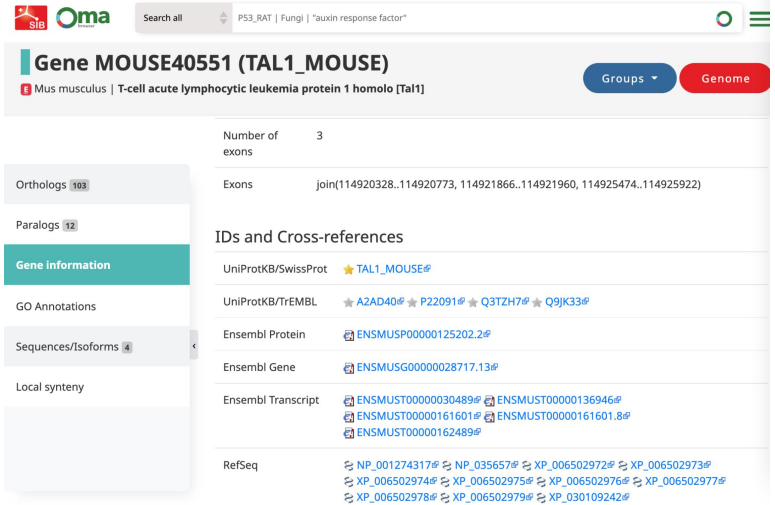
```
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX genex: <http://purl.org/genex#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX orth: <http://purl.org/net/orth#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX lscr: <http://purl.org/lscr#>
SELECT ?name1 ?protein1 ?name2 ?protein2 ?OMA_link2 ?anatomicalEr
    SELECT DISTINCT * {
        SERVICE <https://bgee.org/sparql/> {
            ?taxon up:commonName 'human' ;
                up:commonName ?name1 .
            ?taxon2 up:commonName 'mouse' ;
                up:commonName ?name2 .
        }
        SERVICE <https://sparql.omabrowser.org/sparql/> {
            ?cluster a orth:OrthologsCluster .
            ?cluster orth:hasHomologousMember ?node1 .
            ?cluster orth:hasHomologousMember ?node2 .
            ?node2 orth:hasHomologousMember* ?protein2 .
            ?node1 orth:hasHomologousMember* ?protein1 .
            ?protein1 a orth:Protein .
            ?protein1 rdfs:label 'TAL1' ;
                orth:organism/obo:RO_0002162 ?taxon .
            ?protein2 a orth:Protein ;
                sio:SIO_010079 ?gene ; #is encoded by
                orth:organism/obo:RO_0002162 ?taxon2 .
            ?gene lscr:xrefEnsemblGene ?geneEns .
            ?protein2 rdfs:seeAlso ?OMA_link2 .
            FILTER ( ?node1 != ?node2 )
        }
        SERVICE <https://bgee.org/sparql/> {
            ?geneB a orth:Gene .
                ?geneB genex:isExpressedIn ?cond .
                ?cond genex:hasAnatomicalEntity ?anat .
                ?geneB lscr:xrefEnsemblGene ?geneEns .
            ?anat rdfs:label 'lung' ;
                rdfs:label ?anatomicalEntity .
            ?geneB orth:organism ?o .
            ?o obo:RO_0002162 ?taxon2 .
        }
    }
    LIMIT 10
}
LIMIT 10
```

BIOSQDA
Biological Search over Databases

Which are the mouse's genes expressed in the lung and are orthologous to human's TAL1 gene?

Oma browser

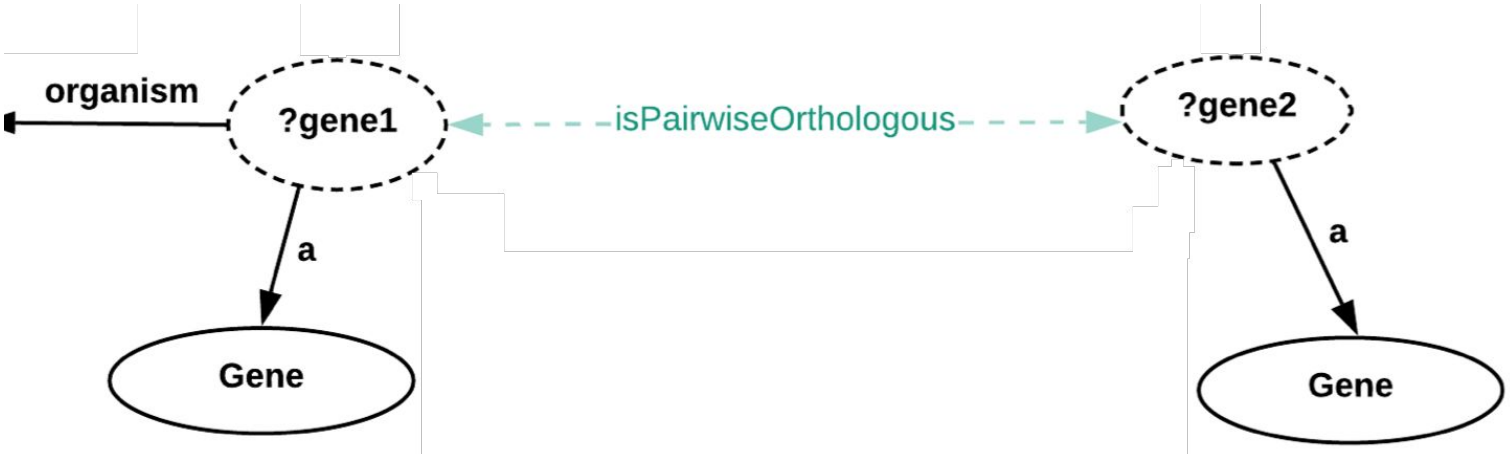
Bgee



What is SPARQL?

a programming language for retrieving information

- from a graph database
- everything as *triples* (*Subject, Predicate, Object).*



f1000research.com/articles/8-1822

Long-term goal of the project

Natural language  SPARQL query

- Benchmark dataset
- Training dataset

How did we start?

Identification of putative producers of rhamnolipids/glycolipids and their transporters using genome mining

Meryam Magri [a], Ahmad M. Abdel-Mawgoud [a,b,*]

[a] Institute of Integrative Biology and Systems, Laval University, 1030 Ave. de la Médecine, Quebec, QC G1V 0A6, Canada
[b] Department of Biochemistry, Microbiology and Bioinformatics, Faculty of Science and Engineering, Laval University, 1045 Ave. de la Médecine, Quebec, QC G1V 0A6, Canada

ARTICLE INFO

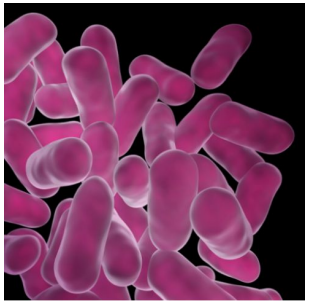
Keywords:
Rhamnolipids
RhlABC orthologs
Genome mining
Thin layer chromatography
Rhamnolipid transport

ABSTRACT

Rhamnolipids (RLs) are microbial glycolipids (GLs) with interesting structure-dependent bioactivities and physicochemical properties making them suitable for diverse medical and industrial applications. The discovery of RLs with more interesting bioactivities and properties has relied on laborious screening of new RL producers isolated from the environment, and has resulted in the redundant identification of already known RL producers and structures. Here, we present a genome mining approach that enabled the identification of 80 RL-producing species (including the two reference species), 71 of which were previously unreported. Distance trees of two of their RL biosynthetic enzymes, RhlAB, allowed for the identification of 11 distinct clades. Preliminary experimental validation with thin layer chromatography on one non-pathogenic RL/GL producer, *Nevskia soli*, confirmed its putative production of RLs. Additionally, this study led to the discovery of the putative RL transport mechanism involving three transmembrane proteins whose coding genes are highly conserved and clustered with one of the RL biosynthetic gene clusters in most RL/GL producers identified in this study.

What is the paper about?

- RhamnoLipids (RL) are a class of microbial glycolipids.
- produced by certain bacterial species (Pseudomonas and Burkholderia)
- interesting bioactivities and physicochemical properties.
- suitable for diverse medical and industrial application.
- RhlA & RhlB genes encode rhamnolipid biosynthetic enzymes.



How orthology is used?

- Identification of RL producing species

- finding orthologs of RhlA genes of OMA.
- 71 new putative species discovered
- Identified RhlA orthologs categorised into 11 phylogenetic clades.

Extracted questions and answers

Question	Answer
How many RhlA, RhlB and RhlC orthologs are harboured by rhamnolipid producers?	Rhamnolipid producers harbour approximately 40 RhlA, 370 RhlB, and 640 RhlC orthologs.
How many strains of *P. aeruginosa* and *B. thailandensis* harbour RhlAB and RhlABC orthologs?	4 strains of *P. aeruginosa* harbour RhlAB orthologs. 10 strains of *P. aeruginosa* harbour RhlABC orthologs. 15 strains of *B. thailandensis* harbour RhlABC orthologs.
How many rhamnolipid producers harbour RhlAB orthologs?	4 rhamnolipid producers were identified that harboured RhlAB orthologs.
How many rhamnolipid producers harbour RhlABC orthologs?	15 rhamnolipid producers were identified that harboured RhlABC orthologs.

BIOQA pipeline (1)

Surveying the literature

(Gene expression, Orthology, ...)

Searching for relevant papers

Europe PMC

Exporting papers

Scoring relevance of papers

python™

Sorting papers and selected a subset

Title	Doi	No of matches
BgeeDB, an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests	10.12688/f1000research.9973.2	20
The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals	10.1093/nar/gkaa793	14
Accessing scientific data through knowledge graphs with Ontop	10.1016/j.patter.2021.100346	9
A gene expression resource generated by genome-wide lacZ profiling in the mouse	10.1242/dmm.021238	7
Selective Constraints on Coding Sequences of Nervous System Genes Are a Major Determinant of Duplicate Gene Retention in Vertebrates	10.1093/molbev/msx199	6
Comparative modular analysis of gene expression in vertebrate organs	10.1186/1471-2164-13-124	6
Comparative analysis of human and mouse expression data illuminates tissue-specific evolutionary patterns of miRNAs	10.1093/nar/gks279	4
New and continuing developments at PROSITE	10.1093/nar/gks1067	3
Molecular signaling in zebrafish development and the vertebrate phylotypic period	10.1111/j.1525-142X.2010.00400.x	3
Pigment Epithelium-Derived Factor (PEDF) Interacts with Transportin SR2, and Active Nuclear Import Is Facilitated by a Novel Nuclear Localization Motif	10.1371/journal.pone.0026234	3
A two-level model for the role of complex and young genes in the formation of organism complexity and new insights into the relationship between evolution and development	10.1186/s13227-018-0111-4	3
Autoantibodies Recognizing the Amino Terminal 1-17 Segment of CENP-A Display Unique Specificities in Systemic Sclerosis	10.1371/journal.pone.0061453	3

Mannually summarise papers and extracting relevant questions/ answers

BIOQA pipeline (2)

Dataset of Questions & answers

Manually classifying questions into high and low level (Sparql-able)

Categorising low-level questions + frequency analysis

Database assignment

Oma browser

Bgee

UniProt

Manually converting questions to SPARQL queries

```
PREFIX oma: <http://omabrowser.org/ontology/oma#>
PREFIX orth: <http://purl.org/net/orth#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX lscr: <http://purl.org/lscr#>
PREFIX up: <http://purl.uniprot.org/core/>
SELECT DISTINCT ?organism  { {
VALUES(?protein1){(<http://purl.uniprot.org/uniprot/A0A8K8P8E7>)

      ?cluster1 a orth:OrthologsCluster.
      ?cluster1 orth:hasHomologousMember ?node1.
      ?cluster1 orth:hasHomologousMember ?node2.
      ?node1 orth:hasHomologousMember* ?protein_OMA_1.
      ?node2 orth:hasHomologousMember* ?ortholog_OMA_1.
           ?protein_OMA_1 lscr:xrefUniprot ?protein1.
      ?ortholog_OMA_1 lscr:xrefUniprot ?ortholog1.
           ?ortholog_OMA_1 orth:organism ?organism.

FILTER(?node1 != ?node2)) } group by ?organism
```

The BIOQA dataset

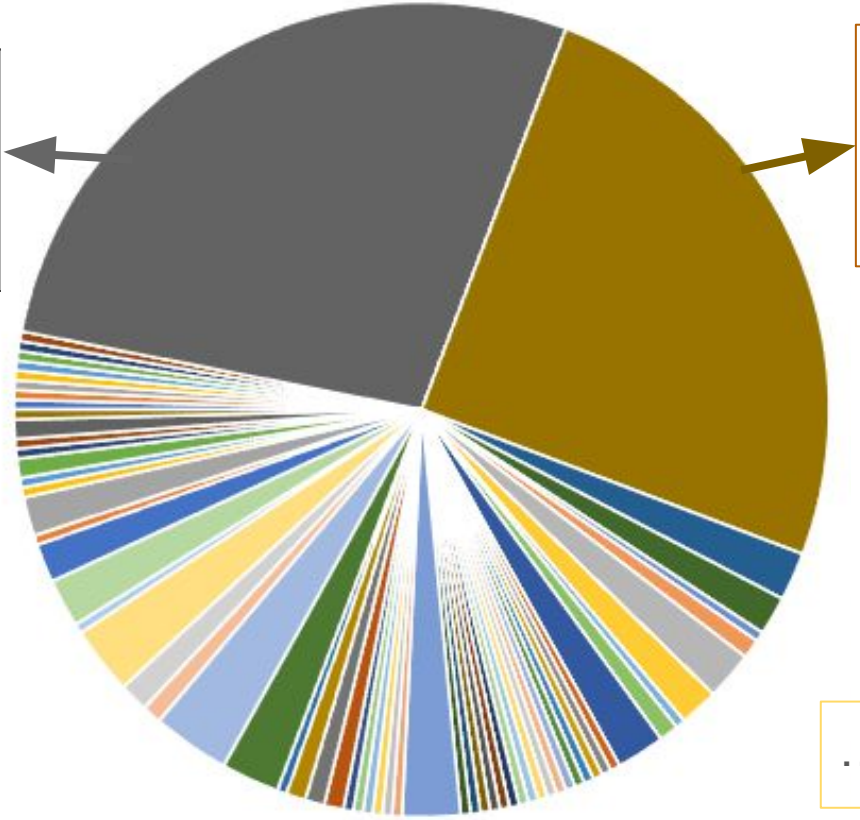
- Question and answer dataset
- SPARQL queries
- Question frequency stats

Dataset so far!

Surveyed papers	345
Relevant papers	150
Extracted questions/answers	267
Question categories	58

Frequency of questions

What is the ortholog of the geneX_speciesY in species Z?



What is the ortholog of the geneX_speciesY expressed in species Z?

… in tissue X

… in pathway X

10.1002/dvg.23331
10.1016/j.bbrc.2017.10.015
10.1128/mSphere.00012-16
10.1016/j.gene.2022.146263
10.1007/s13205-017-0656-2
10.1016/j.devcel.2019.09.017
10.1128/JVI.02065-17
10.1016/j.cub.2018.04.008
10.1186/s12862-017-0923-1
10.1007/s00436-018-5842-6
10.1038/s41598-019-50423-6
10.1016/j.bbalip.2015.12.016
10.1371/journal.pone.0171506
10.1242/jcs.258733
10.1016/j.anaerobe.2018.06.013
10.1530/JOE-16-0040
10.1111/mmi.13492
10.1111/jipb.12467
10.1534/g3.119.400903
10.1038/srep37306
10.1111/1744-7917.12501
10.1111/jth.15365
10.1128/JB.01001-15
10.1093/pcp/pcv130
10.1016/j.redox.2019.101323
10.3390/cells8040343

Question categories

Types of question	Question ID	doi	question	
A Genomes				
	C1	10.1371/journal.|	What are genomes of all strains in genus Acinetobacter?	
	C2	10.3389/fgene.2	What are the gene sequences of selected human genes and its orthologous ge	gene list provided
		10.1109/CSB.20	What are full-length cDNA sequences of mouse and human genomes?	
B orthologous groups				
	C3	10.1371/journal.|	How many orthologous groups do 232 strains of Acinetobacter in Set-R have?	species list provided
		10.1016/j.cub.20	How many orthologous groups are present in 67 species including 9 Xenacoel	species list provided
	C4	10.1093/nar/gka	How many orthologous groups have at least one gene from these species?	human, mouse, rat and pig
		10.1093/bib/bbr(	How to detect all orthology and paralogy relationships for human? (used the phylogenies given above)	
	C5	10.1073/pnas.21	How to assign protein sequences to orthogroups? (species used are in table S2)	
C orthologous gene pairs				
	C6	10.1186/s13059-	How to get orthologous gene pairs between species x and species y?	
		10.3390/ijms202	What are Mus musculus, Ratus norvegicus, Caenorhabditis elegans, Saccharomyces cerevisiae, and Droso	
		10.48550/arXiv.1	What are the orthologs of these brain specific genes?	genes found in Allen Gene E
		10.1177/117793:	What are ortholog gene collections between A. thaliana and S. bicolor?	
		10.1109/CSB.20	How to identify orthologous genes between human and mouse?	
		10.1073/pnas.17	How to find orthologous genes for human GPCRs and RAMPs in other species?	

Potential benefits for our community

- Enabling application, evaluation, and improvement of QA systems
  - in the context of orthology data
  - increase the use and impact of orthology databases

- How people are using orthology databases?
- Finding new biological questions

- As a measure of reproducibility of a paper
  - SPARQL query of the question of a paper
  - Retrieve the answer from database
  - Compare it with paper’s answer

Challenges

- Finding the relevant papers
- Summarising the right questions
- Finding a diverse range of questions
- Automating the process
- Converting English questions to database queries
- Interface of dataset Q/A

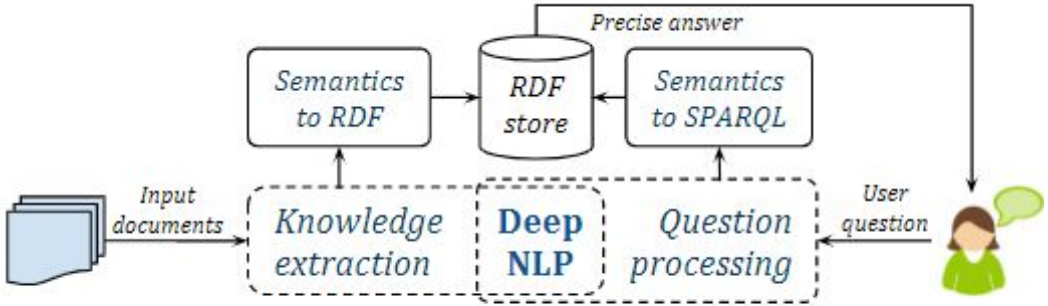
Acknowledgment

Thanks for your attention!

FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

QA system

- build systems to answer questions posed by humans in natural language.



- supervised learning methods require large training datasets

LC-QuAD 2.0
Largescale Complex Question Answering Dataset

DBpedia

- Few QA sets are available for biological data.

BioA2Q

Available databases in SPARQL at

Swiss Institute of Bioinformatics

SIB COVID-19 Integrated Knowledgebase
Integrated data relevant for research on SARS-CoV-2

OMA SPARQL endpoint
Orthology inference among complete genomes

HAMAP SPARQL endpoint
Use HAMAP + SPARQL to generate portable annotation pipelines

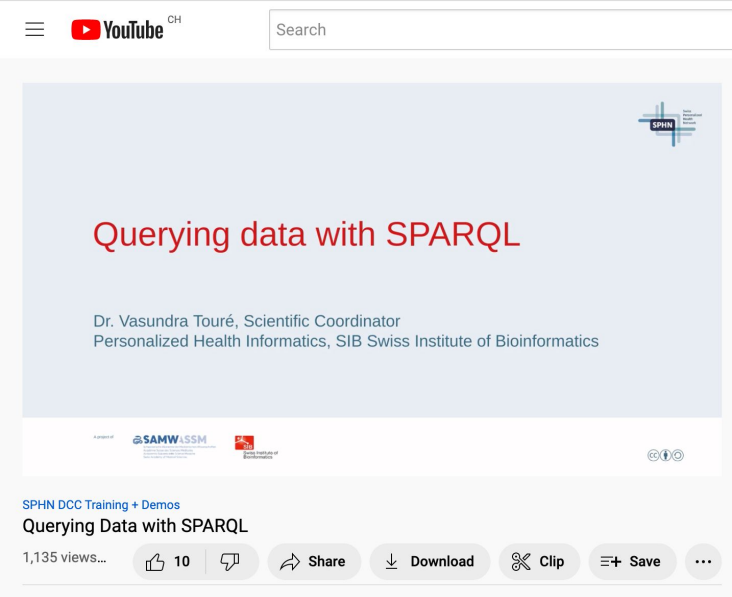
UniProt SPARQL endpoint
Query UniProt + related data in a powerful SQL-like language

Bgee SPARQL endpoint
Gene expression expertise

OrthoDB SPARQL endpoint
Evolutionary and functional annotations of orthologs

neXtProt SPARQL endpoint
SPARQL endpoint for the neXtProt Human protein knowledgebase

Rhea SPARQL endpoint
SPARQL access to the Rhea DB


YouTube CH
Search
Querying data with SPARQL
Dr. Vasundra Touré, Scientific Coordinator
Personalized Health Informatics, SIB Swiss Institute of Bioinformatics
SAMW SSM
SPHN DCC Training + Demos
Querying Data with SPARQL
1,135 views...
10
Share
Download
Clip
Save


f1000research.com/articles/8-1822
F1000Research
Search
BROWSE
GATEWAYS & COLLECTIONS
HOW TO PUBLISH
ABOUT
BLOG
Home » Browse » A hands-on introduction to querying evolutionary relationships across...
Check for updates
METHOD ARTICLE
REVISED
A hands-on introduction to querying evolutionary relationships across multiple data sources using SPARQL [version 2; peer review: 3 approved]
Ana Claudia Sima[1-3], Christophe Dessimoz[2-6], Kurt Stockinger[1], Monique Zahn-Zabal[2,3], Tarcisio Mendes de Farias[2-4,7]
Author details
This article is included in the The OMA collection collection.
ALL METRICS
1729
VIEWS
172
DOWNLOADS
Get PDF
Get XML
Cite
Export
Track
Email
Share
Abstract
The increasing use of Semantic Web technologies in the life sciences, in particular the use of the Resource Description Framework (RDF) and the RDF query language SPARQL, opens the path for novel integrative analyses, combining information from multiple data sources. However, analyzing evolutionary data in RDF is not trivial, due to the steep learning curve required to understand both the data models adopted by different RDF data sources, as well as the equivalent SPARQL constructs required to benefit from this data – in particular, recursive property paths. In this article, we provide a hands-on introduction to querying evolutionary data across several data sources that publish orthology information in RDF, namely: The Orthologous MAtrix (OMA), the European Bioinformatics Institute (EBI) RDF platform, the Database of Orthologous Groups (OrthoDB) and the Microbial Genome Database (MBGD). We present four protocols in increasing order of complexity. In these protocols, we demonstrate through SPARQL queries how to retrieve pairwise orthologs, homologous groups, and hierarchical orthologous groups. Finally, we show how orthology information in different data sources can be compared, through the use of federated SPARQL queries.

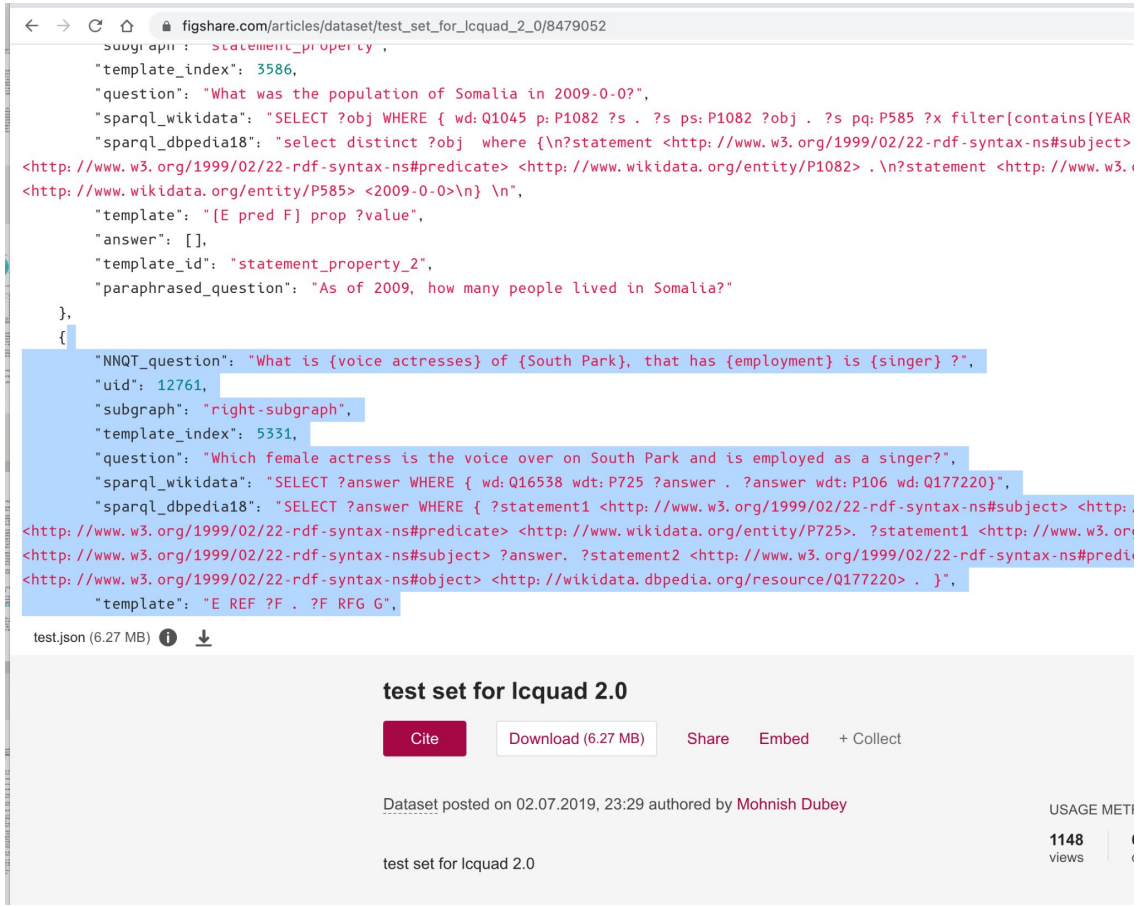
OMA SPARQL Endpoint Explorer

Enter SPARQL Query

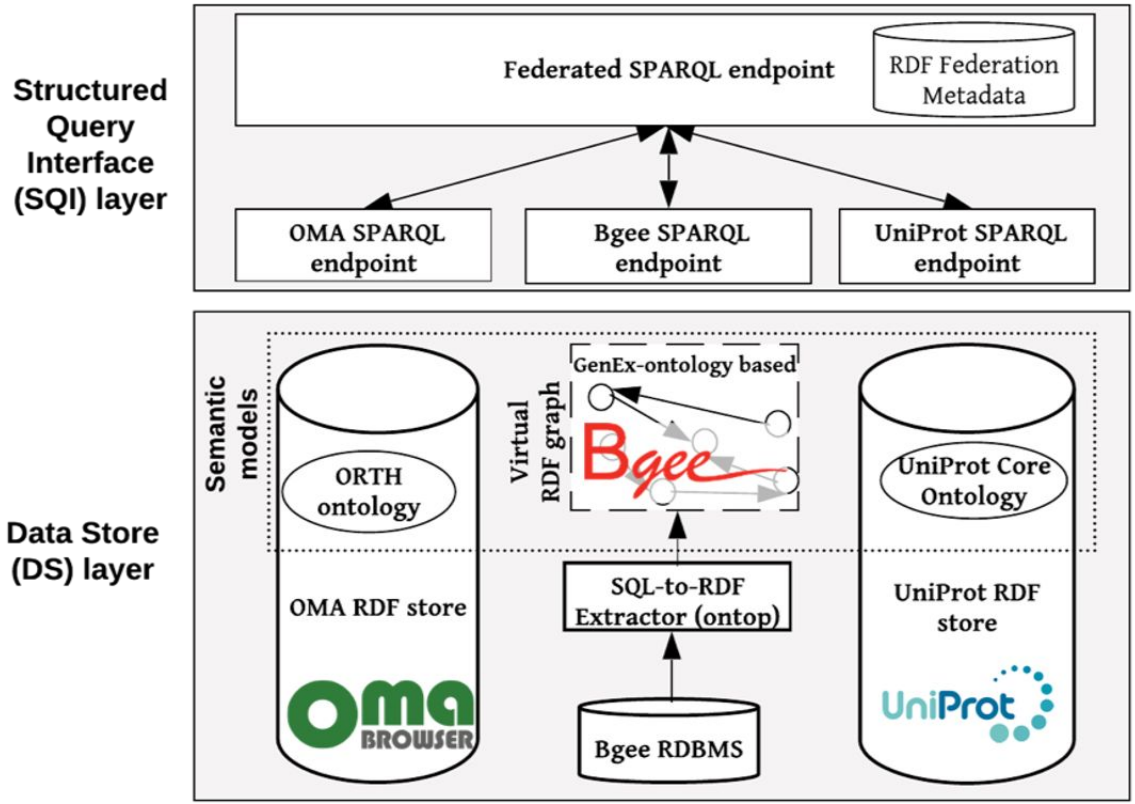
```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX ensembl: <http://rdf.ebi.ac.uk/resource/ensembl/>
PREFIX oma: <http://omabrowser.org/ontology/oma#>
PREFIX orth: <http://purl.org/net/orth#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX taxon: <http://purl.uniprot.org/taxonomy/>
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX void: <http://rdfs.org/ns/void#>
PREFIX lscr: <http://purl.org/lscr#>

select ?protein2 ?OMA_LINK
where {
#The three that contains paralogs. The leafs are proteins.
#This graph pattern defines the relationship protein1 is paralogous t
?cluster a orth:ParalogsCluster.
?cluster orth:hasHomologousMember ?node1.
?cluster orth:hasHomologousMember ?node2.
?node2 orth:hasHomologousMember* ?protein2.
?node1 orth:hasHomologousMember* ?protein1.
########

#Specify the protein to look for its paralogs
?protein1 sio:SIO_010079/lscr:xrefEnsemblGene  ensembl:ENSG0000024473
########

#The OMA link to the second protein
?protein2 rdfs:seeAlso ?OMA_LINK.
########

filter(?node1 != ?node2)
}
```

Example Queries

- Query 1: Find all Rattus norvegicus' proteins present in OMA RDF database.
- Query 2: Which species are available on OMA database and their scientific names?
- Query 3: Retrieve all proteins in OMA that is encoded by the INS gene and their mnemonics and evidence types from Uniprot database (federated query).
- Query 4: Retrieve all genes that are orthologous to ENSLACG00000002497 Ensembl gene (identifier).
- Query 5: Retrieve all genes that are paralogous to ENSG00000244734 Ensembl gene (identifier).
- Query 6: Retrieve all genes that are paralogous to HUMAN00529 OMA protein (identifier) and their cross-reference links to OMA and Uniprot.
- Query 7: Retrieve all genes that are orthologous to HUMAN22169 OMA protein (identifier) and their cross-reference links to OMA and Uniprot.
- Query 8: Retrieve all genes per species that are orthologous to Rabbit's APOCI or APOC1 gene and their cross-reference links to OMA and Uniprot including the corresponding Ensembl gene identifier.
- Query 9: Retrieve all Rabbit's proteins encoded by genes that are orthologous to Mouses's hemoglobin Y gene and their cross-reference links to Uniprot.


figshare.com/articles/dataset/test_set_for_lcquad_2_0/8479052
"subgraph": "statement_property",
"template_index": 3586,
"question": "What was the population of Somalia in 2009-0-0?",
"sparql_wikidata": "SELECT ?obj WHERE { wd: Q1045 p: P1082 ?s . ?s ps: P1082 ?obj . ?s pq: P585 ?x filter(contains(YEAR
"sparql_dbpedia18": "select distinct ?obj where {\n?statement <http://www.w3.org/1999/02/22-rdf-syntax-ns#subject>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#predicate> <http://www.wikidata.org/entity/P1082> .\n?statement <http://www.w3.
<http://www.wikidata.org/entity/P585> <2009-0-0>\n} \n",
"template": "[E pred F] prop ?value",
"answer": [],
"template_id": "statement_property_2",
"paraphrased_question": "As of 2009, how many people lived in Somalia?"
},
{
"NNQT_question": "What is {voice actresses} of {South Park}, that has {employment} is {singer} ?",
"uid": 12761,
"subgraph": "right-subgraph",
"template_index": 5331,
"question": "Which female actress is the voice over on South Park and is employed as a singer?",
"sparql_wikidata": "SELECT ?answer WHERE { wd: Q16538 wdt: P725 ?answer . ?answer wdt: P106 wd: Q177220}",
"sparql_dbpedia18": "SELECT ?answer WHERE { ?statement1 <http://www.w3.org/1999/02/22-rdf-syntax-ns#subject> <http:
<http://www.w3.org/1999/02/22-rdf-syntax-ns#predicate> <http://www.wikidata.org/entity/P725>. ?statement1 <http://www.w3.or
<http://www.w3.org/1999/02/22-rdf-syntax-ns#subject> ?answer. ?statement2 <http://www.w3.org/1999/02/22-rdf-syntax-ns#predi
<http://www.w3.org/1999/02/22-rdf-syntax-ns#object> <http://wikidata.dbpedia.org/resource/Q177220> . }",
"template": "E REF ?F . ?F RFG G",
test.json (6.27 MB)
test set for lcquad 2.0
Cite
Download (6.27 MB)
Share
Embed
+ Collect
Dataset posted on 02.07.2019, 23:29 authored by Mohnish Dubey
test set for lcquad 2.0
USAGE METI
1148
views

Federated queries



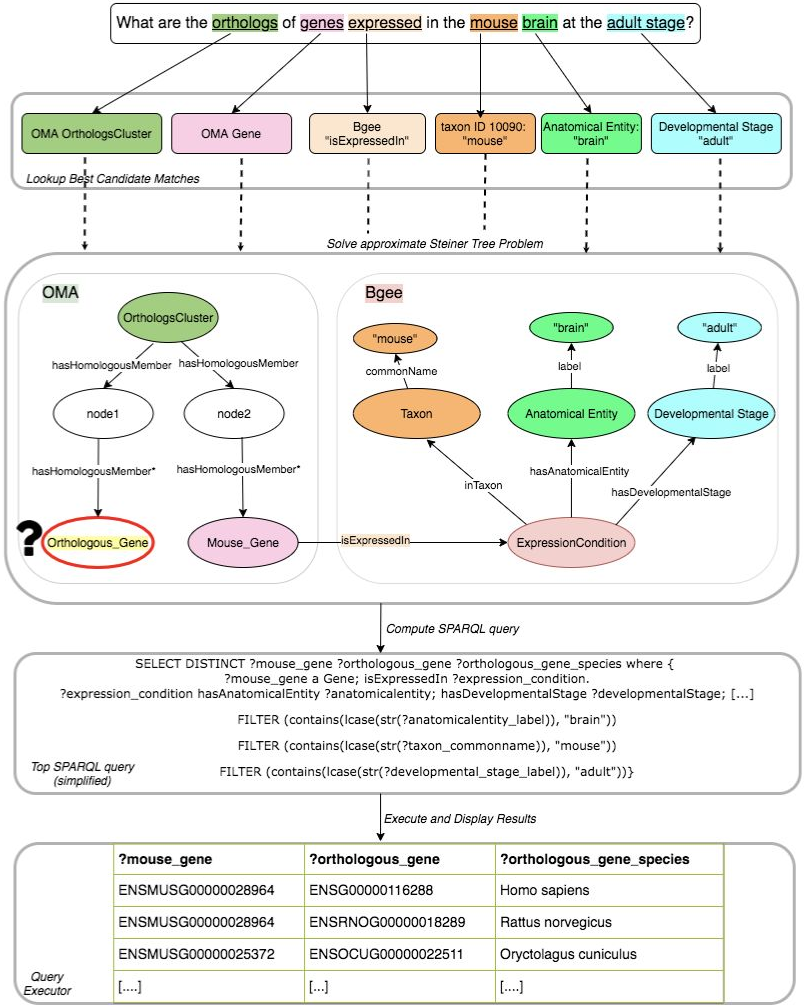
Federated queries - example

```
SELECT * WHERE {

SERVICE <https://bgee.org/sparql/>  {

SELECT DISTINCT ?gene ?anatEntity ?anatName where

{ ?gene a orth:Gene;

        orth:organism ?organism ;

        rdfs:label ?geneName .

        ?organism obo:RO_0002162 <http://purl.uniprot.org/taxonomy/10116>
.

        ?gene genex:isExpressedIn ?anatEntity.

        ?anatEntity a genex:AnatomicalEntity .

        ?anatEntity rdfs:label ?anatName .

FILTER (LCASE(?geneName) = LCASE('APOC1')) } }
```

```
SERVICE <https://sparql.omabrowser.org/sparql/> {
select ?protein1  ?protein2  ?gene ?geneName2

where {

?cluster a orth:OrthologsCluster.

?cluster orth:hasHomologousMember ?node1.

?cluster orth:hasHomologousMember ?node2.

?node2 orth:hasHomologousMember* ?protein2.

?node1 orth:hasHomologousMember* ?protein1.

?protein1 a orth:Protein;

orth:organism/obo:RO_0002162/up:scientificName ‘Homo sapiens’.

protein2 sio:SIO_010079 ?gene . }
```

Where is APOC1 expressed in the rat and what are its orthologs in human?

Bio-SODA:
Semantic Search
over Bioinformatics
RDF Databases



1. Identify and rank candidate matches

2. Compute minimal subgraph connecting candidate matches

3. Derive SPARQL query, execute and present results

Federated queries - even more examples