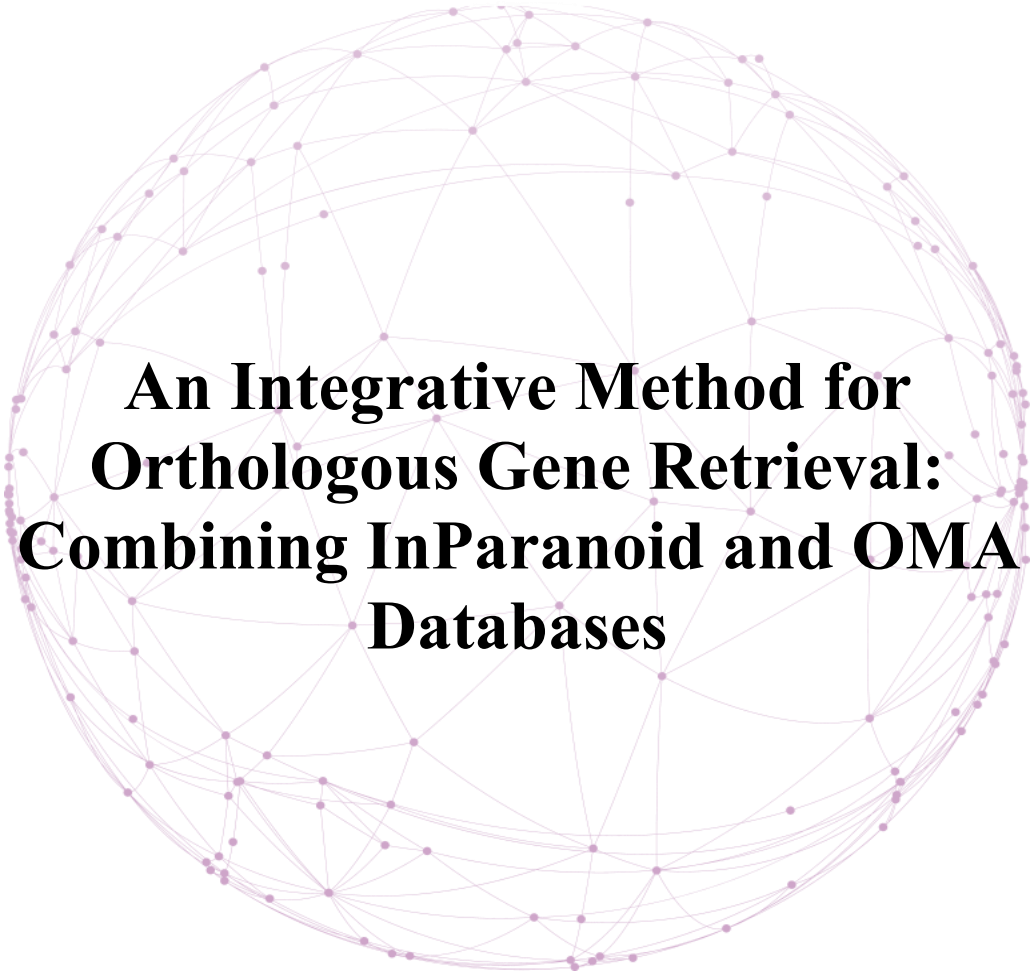
Master's Thesis
Master in Bioinformatics

Faculty of Biology – University of Murcia



An Integrative Method for Orthologous Gene Retrieval: Combining InParanoid and OMA Databases

Author: Asunción Turpín Gómez

Tutor: Jesualdo Tomás Fernández Breis

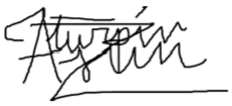
Academic Year: 2023-2024

Declaration of Authorship

I, Asunción Turpín Gómez, with ID number 49306526T, a student of the Master's in Bioinformatics at the Faculty of Biology of the University of Murcia, **declare**:

That the Master's Thesis I present for its exhibition and defence titled: "An Integrative Method for Orthologous Gene Retrieval: Combining InParanoid and OMA Databases", and whose supervisor is Dr. Jesualdo Tomás Fernández Breis, **is original and that all the sources used for its realization have been duly cited therein.**

Murcia, on July 14, 2024.

Signature 

Contribution to Sustainable Development Goals (SDGs)

This Master's thesis develops a computational method for retrieving orthologous genes across species. By integrating data from InParanoid and OMA and introducing a reliability scoring system, the method enhances the accessibility and usability of orthologous gene information. Its relationship with SDGs is clear:

- **Good Health and Well-being:** By facilitating advancements in biomedical research. Improved access to orthologous gene data aids in understanding gene functions related to diseases, potentially leading to new treatments and diagnostics.
- **Industry, Innovation and Infrastructure:** The innovative computational method advances bioinformatics and data integration infrastructure. By standardizing orthologous information and automating data queries, the project promotes technological progress and strengthens research capabilities.
- **Reduced Inequalities:** The method improves global access to reliable genomic data, reducing disparities in research resources. This contributes to more equitable opportunities for scientists and researchers, supporting broader and fairer scientific advancements.

INDEX

1. Abstract

This Master's thesis focuses on developing a computational method to facilitate the retrieval of orthologous genes across species. Orthologous genes, which are crucial for comparative genomics and biomedical research, are dispersed across heterogeneous databases, making interoperability a challenge. This project aligns with the efforts of the Quest for Orthologs consortium, which aims to standardize orthologous information using the ORTH ontology.

The method integrates data from two significant orthology databases, InParanoid and OMA, using SPARQL queries on RDF repositories. By automating the query process and introducing a reliability scoring system based on data curation status and occurrences count across the databases, this method enhances the accessibility and usability of orthologous data for researchers.

The results, exemplified by the retrieval of orthologous proteins to P53 gene of *Homo sapiens* from *Mus musculus,* among other IDs*,* illustrate the method's efficacy in integrating data from InParanoid and OMA. The method organizes this information into a dataframe that includes the UniProt IDs, protein curation status, occurrence counts, and reliability scores.

2. Resumen

Este Trabajo Fin de Máster se centra en el desarrollo de un método computacional para facilitar la recuperación de genes ortólogos entre especies. Los genes ortólogos, que son cruciales para la genómica comparativa y la investigación biomédica, están dispersos en diversas bases de datos heterogéneas, lo que dificulta la interoperabilidad. Este proyecto se alinea con los esfuerzos del consorcio Quest for Orthologs, que tiene como objetivo estandarizar la información ortóloga utilizando la ontología ORTH.

El método integra datos de dos importantes bases de datos de ortología, InParanoid y OMA, utilizando consultas SPARQL en repositorios RDF. Al automatizar el proceso de consulta e introducir un sistema de puntuación de fiabilidad basado en el estado de curación de los datos y la presencia en varias bases de datos, este método mejora la accesibilidad y usabilidad de los datos ortólogos para los investigadores.

Los resultados, ejemplificados por la recuperación de proteínas ortólogas al gen P53 de *Homo sapiens* a partir de *Mus musculus*, entre otros ID, ilustran la eficacia del método para integrar datos de InParanoid y OMA. El método organiza esta información en un dataframe que incluye los ID de UniProt, el estado de curación de las proteínas, los conteos de ocurrencias para cada proteína y las puntuaciones de fiabilida

3. Introduction

3.1. Homology and its relationship with computation

Evolution is a fundamental principle of biology, giving rise to the concept of homology, which describes relationships between genes that share a common ancestry. From a general homologous relationship, pairs of genes are classified in various sub-groups of homologs, including ortholog and paralog, among others (Zahn-Zabal et al., 2020). Ortholog genes are pair of genes that started diverging via evolutionary speciation, while paralog ones started diverging via duplication (Altenhoff et al., 2012, p. 159). This phenomenon is illustrated by a phylogenetic tree in Figure 1.

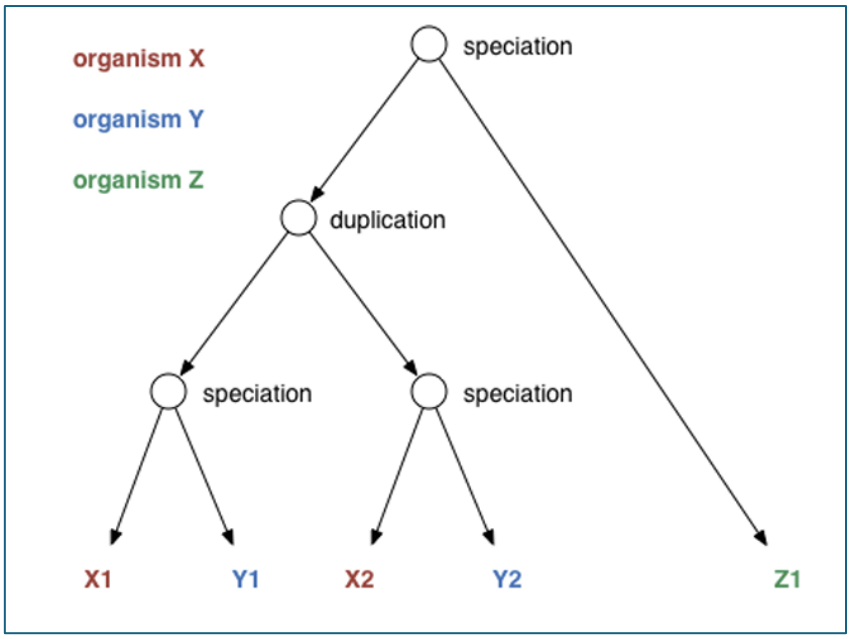


Figure 1. Sample of phylogenetic tree. Obtained from Fernández-Breis et al. (2015).

This phylogenetic tree represents the history of the five genes represented at the bottom of the image: X1, Y1, X2, Y2 and Z1. The different colours of the genes represent the species: red organism X, blue organism Y and green organism Z. The root node of the tree represents the common ancestor of the different genes, who has a speciation event associated (Baum, 2008), giving rise to two distinct species. One species has maintained the same exact gene (Z1), while the other lineage at the root underwent subsequent events of gene duplication and further speciation, eventually leading to the formation of the remaining genes (X1, X2, Y1, Y2) in their respective species.

Once it is clear what the tree represents the existing relationships between each gene can be obtained. For instance, X1 gene has two ancestral nodes associated with speciation events: the common ancestor of X1 gene and Y1 gene, and the common ancestor of X1 gene and Z1 gene. Hence, Y1 and Z1 are orthologs of X1 (Fernández-Breis et al., 2015).

Searching for paralogy genes an ancestral node associated with duplication must exist between them. Therefore, X1 has one ancestral node associated with duplication event: the common ancestor of X1 gene and X2 gene and the common ancestor of X1 gene and Y2 gene. Thus, X2 and Y2 are paralogs of X1 (Fernández-Breis et al., 2015).

Besides, as the number of sequenced genomes has increased in recent years, leading to the knowledge not only of reference organism genomes, but also of individual genomes, computational comparative analysis is playing a crucial part (Fernández-Breis et al., 2015) by inferring gene prediction via orthology and paralogy. Orthologs are likely to have similar biological functions between different species more than paralogs, as duplications are often followed by functional divergence. Consequently, orthologs are the ones of interest for inferring function computationally (Altenhoff et al., 2012, p. 160).

Predicting orthology from biological sequences remains a challenging problem as new genomes are in constant release (Mukherjee et al. 2019). In concrete, orthology information is fundamental for establishing evolutionary relations among genes from multiple organism's methods. Because of this, there is an urgent need to standardize and interoperate orthology resources, so the Quest for Orthologs (QfO) consortium was founded, and semantic web technologies as OrthoXML was created, which permits the comparison and integration of orthology data from different resources (Fernández-Breis et al., 2016).

3.2. Semantic Web Technologies

In recent years, semantic web formats as RDF and SPARQL have become more important within the bioinformatics community. RDF is a general proposition language for the web, unifying data from diverse sources using URI (Uniform Resource Identifier), and SPARQL a query language for RDF that joins data from different databases (World Wide Web Consortium [W3C], 2007).

A knowledge graph or semantic network represents a network of real-world entities (objects, events, situations or concepts) and illustrates the relationship between them. It is made up of three main components: nodes, edges and labels. Edges define the relationship between the nodes (IBM, n.d.). RDF is used to represent this information in the form of triples: subjects, predicates and objects, and URIs act as unique identifiers for nodes and relationships within this graph. A URI can be a URL, but they do not necessarily enable access to the resource they describe so, in most cases, they do not represent actual web pages (Ontotext, 2024, p. 581).

SPARQL allows users to retrieve specific information, make inferences, and connect distributed data across different RDF resources. This capability is essential for semantic integration and interoperability within knowledge graphs. SPARQL provides efficient navigation and extraction of meaningful information from structured RDF-based knowledge graphs (Ali et al., 2021).

To fully understand what a graph is made of, an example of one triplet can be seen in Figure 2. This triplet represents p53 is a gene. The subject consists of the URI http://example.org/gene/p53, which identifies uniquely p53 gene. The predicate uses http://example.org/property/type URI to indicate "type" property, and the object is the value associated to the property of the subject, which indicates that p53 is a gene within the URI http://example.org/type/Gene.

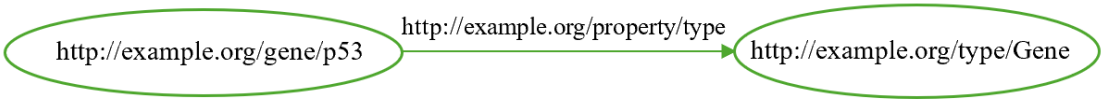


Figure 2. Example of RDF triplet which explains p53 is a gene through the use of subject, predicate and object within URIs.

Additionally, ontologies are frequently mentioned in the context of knowledge graphs, and they serve to create a formal representation of the entities within the graph (IBM, n.d.). To enable such a description, it is need to formally specify components such as individuals, classes, attributes and relations. As a result, ontologies do not only introduce sharable and reusable knowledge representation, but can also add new knowledge about the domain (Ontotext, n.d.).

The Web Ontology Language (OWL) is an example of a widely adopted ontology, supported by the World Wide Web Consortium (W3C), an international community that champions open standards for the longevity of the internet. This organization of knowledge is underpinned by technological infrastructure such as databases, APIs, and machine learning algorithms, which exist to help people and services access and process information more efficiently (IBM, n.d.).

The report of the 2013 QfO meeting (Sonnhammer et al., 2014) concluded that the orthology community should use ontologies in order to facilitate data sharing, so the Orthology Ontology (ORTH) was developed and implemented in OWL, available at https://github.com/qfo/OrthologyOntology. The availability of ORTH enables the possibility of using it for the generation of RDF datasets (Fernández-Breis et al., 2015).

Since the first version of the ontology in 2015, a second version of the ORTH ontology was released in 2017 to formalize new homology-related concepts. This new version included additional properties to describe information from more than 40 orthology databases including OMA, InParanoid. These databases structure the data in one or more different manners (Farias et al., 2017):

OMA can identify three types of ortholog genes: Hierarchical Orthologous Groups (HOGs), OMA Groups, and pairwise orthologs. The OMA algorithm follows a multi-step process to infer these orthologs. Initially, it performs all-against-all Smith-Waterman alignments to determine pairwise orthologs. Then, it uses evolutionary distances to find the closest homologs, defining the

best hit hypothesis orthologs. Finally, the identified ortholog genes are clustered into HOGs and OMA Groups (Zahn-Zabal et al., 2020).

As mentioned, pairwise orthologs are based on the sequence similarity of genes between genomes. Additionally, an example of Hierarchical Orthologs Groups (HOGs), is illustrated in Figure 3. There is a main HOG that englobes every gene from every species due to the first event of speciation. Subsequently, as a second event of speciation occurs within mammals, two other sub-groups of HOGs are formed within the main HOG. Finally, in OMA Groups all genes are connected to each other by pairwise orthologous relations, specific to OMA (Zahn-Zabal et al., 2020).

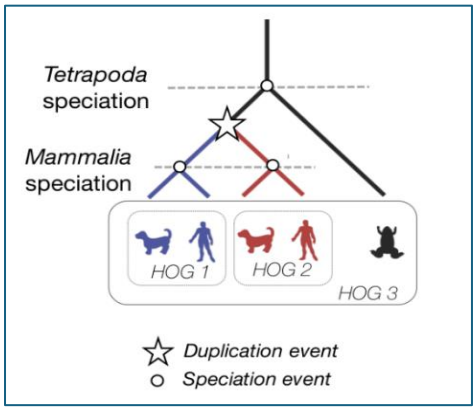


Figure 3. Example of the hierarchical nature of HOGs. The root node corresponds to the common ancestor of mammals and tetrapods, who diverged from Tetrapoda from an event of speciation. Obtained from Zahn-Zabal et al. (2020).

In this database, model organisms are frequently updated in each re-annotation or added by user requests. Besides, OMA provides synteny data and domain annotations for each gene. For those reasons, OMA is considered to hold a rich visual interface which provides a bunch of information and harbours many species (De Boissier & Habermann, 2020).

InParanoid database uses reciprocal BLASTp searches to identify orthologs via RBH (Reciprocal Best Hit) method (De Boissier & Habermann, 2020). It infers pairwise relations to structure their data. The pairwise orthologous relationships among these proteins are carried out using BLAST, followed by a clustering step using the InParanoid algorithm. An InParanoid cluster is seeded by a reciprocally best-matching orthologous pair, around which inparalogs are gathered independently (Bawono & Heringa, 2014).

This database offers a moderate range of organisms and is limited to well-conserved orthologs that can be found by BLASTp (De Boissier & Habermann, 2020). InParanoid host dasta for 273 organisms, while OMA 2851, as will be contrasted in the next section. These differences highlight the distinct methods each database uses to infer orthologs in the two databases, given the different algorithms they employ.

3.3. Objective

The main objective of this work is to develop a computational method that integrates results from different orthologous gene prediction databases that use the ORTH ontology, specifically OMA and InParanoid. On the one hand, the results from the different databases will be retrieved using SPARQL queries, while on the other hand, to combine these results, a reliability score will be

created based on two criteria. This approach aims to streamline the compilation of orthologous gene information and provide a quality assessment through a comprehensive scoring system.

4. Materials and Methods

In order to explain how this project has been developed, in this section the used tools and the work flow will be explained.

4.1. The Orthology Ontology

The Orthology Ontology forms the foundation of this work. Therefore, understanding its basis is essential before delving into its development. The design of the ontology incorporates existing ontologies to enhance interoperability between biomedical domains. Additionally, it employs a modelling approach based on gene members clustered as homologs, orthologs, or paralogs. This allows for the inference of pairwise relationships by analyzing the structure and content of the dataset (Fernández-Breis et al., 2015).

The ontologies used for the development of the Orthology Ontology are:

1. **CDAO** (Comparative Data Analysis Ontology): Provides a framework for understanding data in evolutionary context. Used for representing terms as phylogenetic terms, OTUs (Operational Taxonomic Units) and molecular characters (Prosdocimi et al., 2009).
2. **RO** (Relations Ontology): First reference of properties included in ORTH.
3. **SIO** (Semanticsience Integrated Ontology): Provides a simple, integrated ontology for objects, processes and their attributes (Dumontier et al., 2014).
4. **HO** (Homology Ontology): For representing terms related to homology (Roux & Robinson-Rechavi, 2010).
5. **NCIt** (Nation Cancer Institute Theasaurus): It covers vocabulary for clinical care, translational and basic research (Sioutos et al., 2007).
6. **COG** (Cluster of Orthologous Groups) **Analysis Ontology**: Supports the COG enrichment study (Lin et al., 2011).

4.1.1. Description of the Orthology Ontology

In Figure 4 an excerpt of this ontology is represented. There, boxes represent classes and arrows represent properties, which form the ORTH nucleus. Within its structure three different parts can be identified:

1. Cladogenetic changes: Evolutionary process in which an ancestral lineage splits into sibling lineages that evolve independently and acquire their own repertoire of evolutionary changes. This part of the structure reuses CDAO and defines evolutionary processes such as speciation (cdao:speciation) and duplication (cdao:duplication).

2. Clustering structure: Structure of homolog clusters (HomologsCluster) is modelled by differentiating orthologs cluster (OrthologsCluster) from paralogs cluster (ParalogsCluster) by cdao:has property, which connects every cluster to its corresponding evolutionary event. Moreover, as clusters are represented as trees, everyone belongs to GeneTreeNode class, which inherits cdao:Node to provide interoperability with other resources. Additionally, cluster membership is expressed by means of hasHomologous property, which inherits sio:has_part property, so this structure makes it possible to traverse the homologs tree to find pairwise relationship between genes.

3. Biological information: Gene, subgene and protein classes are present in this part. They belong to SequenceUnit class and subsequently, to cdao:TU (Taxonomic Unit). To connect this information within clusters, rdfs:subClassOf is used between SequenceUnit and GeneTreeNode. Finally, for genes and protein a connection to NCBIT organisms has been made using ro:in_taxon property.

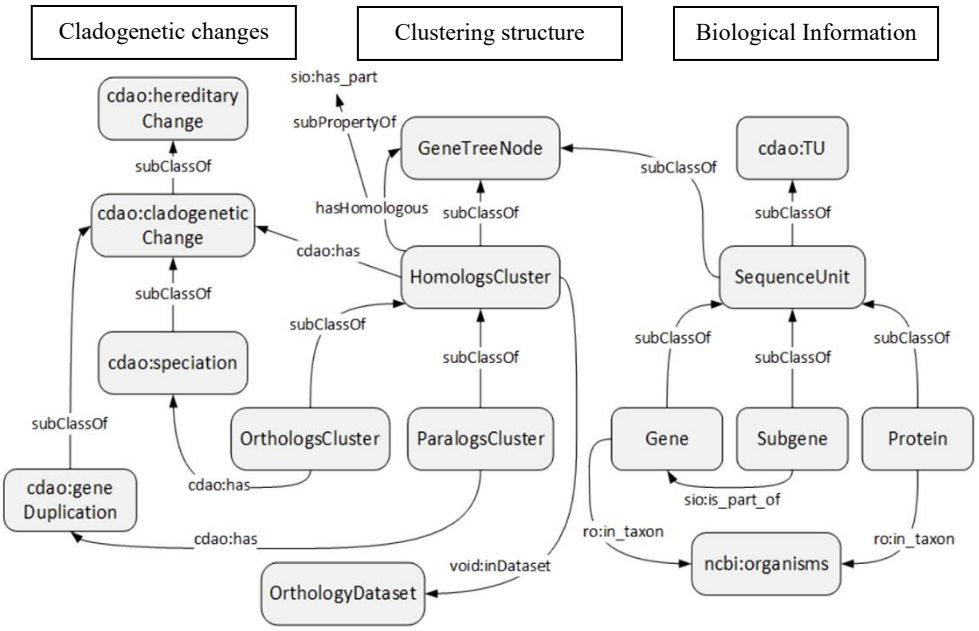


Figure 4. Excerpt of the Orthology Ontology. The entities without prefix are defined in the orthology ontology. Obtained from Fernández-Breis et al. (2015).

4.2. Tools and Equipment

GraphDB

GraphDB is a highly efficient and robust graph database with RDF and SPARQL support. It streamlines the load and use of linked data cloud datasets, as well as own resources, and can perform semantic inferencing at scale. It is used for storing, managing and querying structure data with respect to ontologies, so it acts as a semantic repository (Ontotext, 2024, 597).

The workbench is the web-based administration interface to GraphDB, and it consists of two main areas. On the one hand, the navigation area located on the left side facilitates tasks such as importing, exploring and querying data, along with other functionalities. On the other hand, the home page provides access to different actions such as creating a repository and finding a resource.

In this work, GraphDB is used to upload RDF data from InParanoid database. Initially, this data needed to be transformed from OrthoXML format. This transformation allows for the integration of the datasets from these three biological databases into a unified repository, facilitating querying via SPARQL through its workbench. The endpoint for our GraphDB instance is https://semantics.inf.um.es:7200. Once the data is uploaded, each graph within GraphDB has its own endpoint, which points to the specific data from that graph:

- InParanoid Graph: The endpoint of this graph is http://semantics.inf.um.es/inparanoid. Executing the query, shown in Figure 5, allows the retrieval of the organisms hosted by the database. The result is 273 organisms.

```
PREFIX orthology: <http://purl.org/net/orth#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX up: <http://purl.uniprot.org/core/>

select (COUNT(DISTINCT ?s) as ?organisms)
where {
  GRAPH <http://semantics.inf.um.es/inparanoid> {
    ?s rdf:type up:Taxon
  }
}
```

Figure 5. SPARQL query from GraphDB executed to retrieve organisms hosted by InParanoid database.

As the primary objective is to create a method to identify orthologous genes from two different species collected in the database, p53 gene in *Homo sapiens* is used as an example to determine its orthologs in the *Mus musculus* species. For this, a SPARQL query that follows ORTH structure needs to be developed, and it is shown in Figure 6. In addition, it is necessary to know InParanoid database uses UniProt IDs to identify every protein, so P04637 is the ID that corresponds to p53 *Homo sapiens* protein.

```
PREFIX orthology: <http://purl.org/net/orth#>
PREFIX orth: <http://purl.org/net/orth#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX resource: <http://semanticscience.org/resource/>

select DISTINCT ?iduniprot1 ?iduniprot2 ?species1 ?species2
where {
  GRAPH <http://semantics.inf.um.es/inparanoid> {
    ?cluster a orth:OrthologsCluster ;
             orthology:hasHomologousMember ?node1 ;
             orthology:hasHomologousMember ?node2 .
    ?node1 rdfs:label ?geneid1 ;
           resource:SIO_010078 ?uniprot1 .
    ?uniprot1 a orthology:Protein ;
              rdfs:label ?iduniprot1 ;
              obo:RO_0002162 ?taxid1 .
    ?taxid1 rdfs:label ?species1 .
    ?node2 rdfs:label ?geneid2 ;
           resource:SIO_010078 ?uniprot2 .
    ?uniprot2 a orthology:Protein ;
              rdfs:label ?iduniprot2 ;
              obo:RO_0002162 ?taxid2 .
    ?taxid2 rdfs:label ?species2 .
    filter (?node1 != ?node2 && ?species1 != ?species2)
    values (?iduniprot1 ?species1 ?species2 ) {("P04637" "Homo sapiens" "Mus musculus")}
  }
}
```

Figure 6. SPARQL query to obtain orthologous p53 Homo sapiens genes of Mus musculus from InParanoid graph in GraphDB.

OMA

OMA database provides its own endpoint explorer, reachable by using SPARQL interface through https://sparql.omabrowser.org/lode/sparql, updated in sync with the OMA Browser. In this database the number of different organisms is 2851. The query used is shown in Figure 7.

```
PREFIX obo:<http://purl.obolibrary.org/obo/>

select (COUNT(DISTINCT ?taxid) as ?organisms)
where {
    ?s obo:RO_0002162 ?taxid
}
```

Figure 7. SPARQL query from OMA Endpoint Explorer executed to retrieve organisms hosted by OMA database.

In this case, federated queries can be developed to jointly retrieve data from multiple resources in one single query, while in GraphDB this is not possible as it provides accession to a single repository with multiple graphs. An example of a federated query would involve the use of different SPARQL endpoints from different databases.

OMA provides a user-friendly web interface (OMA Browser) where researchers can explore genes and genomes, find HOGs and OMA Groups, visualize evolutionary relationships and access detailed gene annotations (Zahn-Zabal et al., 2020). Additionally, each gene in OMA has an OMA

identifier, consisting of the five-letter UniProtKB species code and a unique 5-digit number, besides UniProt and Ensembl IDs.

To retrieve orthologous p53 genes between *Homo sapiens* and *Mus musculus* the query shown in Figure 8 was launched into its endpoint.

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

SELECT DISTINCT ?name_prot1 ?name_species1 ?name_prot2 ?name_species2
WHERE {
    ?cluster a orth:OrthologsCluster ;
       orth:hasHomologousMember ?node1 ;
       orth:hasHomologousMember ?node2 .
    ?node1 orth:hasHomologousMember* ?protein1 .
    ?node2 orth:hasHomologousMember* ?protein2 .
    ?protein1 a orth:Protein ;
        orth:organism ?taxid1 ;
        lscr:xrefUniprot ?iduniprot1 .
    ?taxid1 obo:RO_0002162 ?taxon1 .
    ?taxon1 up:scientificName ?name_species1 .
    ?iduniprot1 dct:identifier ?name_prot1 .
    ?protein2 a orth:Protein ;
        orth:organism ?taxid2 ;
        lscr:xrefUniprot ?iduniprot2 .
    ?taxid2 obo:RO_0002162 ?taxon2 .
    ?taxon2 up:scientificName ?name_species2 .
    ?iduniprot2 dct:identifier ?name_prot2 .
    FILTER (?node1 != ?node2 && ?taxid1 != ?taxid2)
    VALUES (?name_species1 ?name_species2 ?name_prot1) {
        ("Homo sapiens" "Mus musculus" "P04637")
    }
}
```

Figure 8. SPARQL query to obtain orthologous p53 genes between Homo sapiens and Mus musculus from OMA Endpoint Explorer.

Jupyter Notebook – Python 3.6

To a better understand of this part of the work, the developed Jupyter Notebooks are available in a GitHub repository, accessible through this link https://github.com/asunturpin/MSc-Thesis.

To unify the results from the three queries shown earlier, Python 3.6 is used through Jupyter Notebook interface. Jupyter Notebook is an open-source web-based interactive development environment for notebooks, code, and data. It provides a flexible and powerful user interface for managing data in a single workspace. Using the Python programming language in Jupyter Notebook creates a robust environment for data analysis, where Python code can be written in an interactive notebook that allows immediate visualization of data and results (Project Jupyter, 2024).

Two notebooks are developed to connect to the endpoints from either GraphDB and OMA, where identical queries were executed. This section presents the libraries and custom functions implemented in the script. Detailed results will be presented in the subsequent section.

To expand the capabilities of the Python programming language, various libraries were imported, as outlined in Table 1:

Library name	Function
Certifi 2024.06.02	Used to ensure HTTP/HTTPS connections are secure by validating the servers' SSL certificates against a list of trusted root certificates (Reitz, 2024)
SPARQLWrapper 2.0.0	Allows to create an instance that handles communication with SPARQL endpoint (Herman et al., 2022)
Os	Needed to establish the environment variables necessary to ensure a secure HTTP/HTTPS connection.
Requests 2.27.1	To make easy HTTP request. Useful to interact with APIs (Reitz, 2022).
Pandas 1.1.5	Used to work with structured data and manipulate them easily (Pandas development team, 2022).

Table 1. Description of the implemented libraries in Python.

In addition, eight functions were defined in order to develop the method to combine the different queries:

1. **Configure_sparql.** This function is used to configure a SPARQL endpoint specified by 'endpoint_url' parameter which can be used for GraphDB or OMA. Optional parameters 'user' and 'password' are only necessary for connecting to GraphDB and are not required for OMA. The function sets the output format to JSON for query results and returns the configured 'sparql' object initialized with SPARQLWrapper.
2. **Execute_query.** This function executes a SPARQL query using the provided 'sparql' object. It sets the 'query' to be executed, retrieves and converts the results, and returns them.
3. **Query_InParanoid.** This function is used to execute a SPARQL query against the InParanoid database using the provided parameters 'uniprot_id', 'species1' and 'species2'. It retrieves orthologous proteins pairs and constructs a Pandas dataframe.
4. **Query_OMA.** This function is used to execute a SPARQL query against the OMA database using the same parameters 'uniprot_id', 'species1' and 'species2'. It also retrieves orthologous proteins pairs and constructs a Pandas dataframe.
5. **Is_curated.** This function checks if a UniProt protein entry specified by 'protein_id' is curated (or reviewed) in the Swiss-Prot database. It constructs a URL to access the UniProt REST API endpoint for the specified 'protein_id', and sends a GET request to

the constructed URL using Request library. It returns True if the UniProt protein entry is curated, otherwise returns False.

6. **Check_species.** This function is used to check if the input 'species' (species1 and species2) string is in the correct format. It will be used once the user enter the species he/she want to retrieve orthologous from. The entry must have two parts: the genus with a first uppercase, letter and species name consisting entirely of lowercase letter. If the input species format is incorrect, the function prints an error message.
7. **Get_human_uniprot_ids:** This function is needed to connect to the UniProt API to retrieve a list of UniProt IDs for proteins from *Homo sapiens* species. This is needed for validating the reliability of orthologous gene retrieval method with new examples in Results section.

'Assign_score' function

The 'assign_score' function plays a relevant role in the development of the method by calculating a reliability score for each protein result. This score is based on two main criteria, in order to provide a combined measure of reliability from the different databases and the curation status of proteins.

Criteria 1: Occurrences in Databases ('score_count'):

- This part of the score is determined by counting the number of occurrences of each protein across the different databases.
- The count is divided by the total number of databases considered (two in this case), resulting in a percentage (0 to 100).

Criteria 2: Curation Status ('score_curated'):

- This other part of the score is based on whether the protein is curated in the Swiss-Prot database (function 'is_curated').
- If the protein is curated (True), it is assigned a score of 50.
- If the protein is not curated (False) it is assigned a score of 0

The total score ('total_score') calculation is a combination of the two criteria: 'score_count' contributes 50 % to the total score, and 'score_curated' contributes the remaining 50 %. This ensures both the reliability from multiple databases and the curation status are equally weighted in the final score.

The libraries and developed functions are imported into the notebooks in order to avoid code duplication. The file which contains this data is named 'libraries_functions'.

5. Results

The main results of this work were obtained using the same gene example as the one mentioned earlier: retrieving orthologous genes for the p53 gene from *Homo sapiens* in *Mus musculus*.

As mentioned before, the developed notebooks of this section, which corresponds to the unified method, are available in the repository.

5.1. InParanoid results

The result from the SPARQL query executed to the InParanoid graph in GraphDB repository is shown below in Table 2:

Iduniprot1	Iduniprot2	Species1	Species2
P04637	P02340	*Homo sapiens*	*Mus musculus*

Table 2. Result from the SPARQL query executed to InParanoid graph in GraphDB.

As it can be seen, there is just one orthologous protein to P04637 of *Homo sapiens* in *Mus musculus,* which is collected under the name 'Iduniprot2', P02340.

5.2. OMA results

The results from the SPARQL query executed in OMA endpoint is shown in Table 3:

Iduniprot1	Iduniprot2	Species1	Species2
P04637	A0A158SIS7	*Homo sapiens*	*Mus musculus*
P04637	O70366	*Homo sapiens*	*Mus musculus*
P04637	P02340	*Homo sapiens*	*Mus musculus*
P04637	P53_MOUSE	*Homo sapiens*	*Mus musculus*
P04637	Q549C9	*Homo sapiens*	*Mus musculus*
P04637	Q91XH8	*Homo sapiens*	*Mus musculus*

Table 3. Results from the SPARQL query executed in OMA endpoint.

This query retrieves 6 orthologous proteins to P04637 of *Homo sapiens* in *Mus musculus,* which are collected under the name 'Iduniprot2', and have different UniProt IDs: A0A158SIS7, O90366, P02340, P53_MOUSE, Q549C9 and Q91XH8.

5.3. Unified method

The developed notebook includes an input section where these three parameters are required from the user: the UniProt ID of the gene from the first species, the name of the first species, and the second species from which the orthologous genes are to be retrieved. If the inputs are not provided correctly, an error message will be displayed. For this example, the inputs are as follows: for the

first species, *Homo sapiens*; for the UniProt ID, P04637 (UniProt ID for p53 gene from *Homo sapiens*); and for the second species, Mus musculus.

Using the mentioned functions 'query_InParanoid' and 'query_OMA' with those parameters, a first dataframe containing the results for each database is created. It includes the three parameters introduced by the user, plus the UniProt ID from the second species and the database they came from. This is shown in Table 4.

UniProt ID 1	Species 1	UniProt Id 2	Species 2	Source
P04637	*Homo sapiens*	A0A158SIS7	*Mus musculus*	OMA
P04637	*Homo sapiens*	O70366	*Mus musculus*	OMA
P04637	*Homo sapiens*	P02340	*Mus musculus*	OMA
P04637	*Homo sapiens*	P53_MOUSE	*Mus musculus*	OMA
P04637	*Homo sapiens*	Q549C9	*Mus musculus*	OMA
P04637	*Homo sapiens*	Q91XH8	*Mus musculus*	OMA
P04637	*Homo sapiens*	P02340	*Mus musculus*	InParanoid

Table 4. First dataframe containing UniProt ID from the first species, as well as the name of the first species itself, the UniProt IDs from the second species and the name of it, and the database they come from.

Next, the occurrences of the resultant protein IDs are counted, and the 'is_curated' function is used. This process results in a dataframe that contains the UniProt IDs retrieved from the second species, the number of occurrences for each one, and an indication of whether the protein is curated or not. The second dataframe is shown below in Table 5:

UniProt 2	Count	Curated
P02340	2	True
A0A158SIS7	1	False
O70366	1	False
Q549C9	1	False
Q91XH8	1	False
P53_MOUSE	1	True

Table 5. Second dataframe containing the UniProt ID from the second species, the occurrences count for each protein, and whether that protein is curated or not.

From this second dataframe, a fourth column named 'Reliability score' is added by applying the 'assign_score' function, and a third dataframe with the results sorted on descending order is created (Table 6):

UniProt 2	Count	Curated	Reliability score
P02340	2	True	100.0
P53_MOUSE	1	True	75.0
A0A158SIS7	1	False	25.0
Q549C9	1	False	25.0
Q91XH8	1	False	25.0
O70366	1	False	25.0

Table 6. Third dataframe including the reliability score for each protein result.

In this example, it can be observed that the UniProt ID P02340 is present in both databases and is curated (True), resulting in a reliability score of 100 %. The UniProt ID P53_MOUSE is only contained in one database (OMA) and has a reliability score of 75 %. As shown in previous dataframes, the rest of the IDs are contained in the OMA database and refers to non-curated proteins, resulting in the lowest reliability scores (25 %).

5.3.1. Additional example

In order to prove this method, a similar second notebook is developed. A sample retrieval of UniProt IDs from *Homo sapiens* (Q9MYI0, A0AB7M9U8, B9EF68, D3VVN2, A0A1W2PQB1, E5RH42, A0A5H2UYS1, Q8WYC4, A0A7T0Q6E8 and A0A6Q8PHP0 IDs) has been obtained to get their ortholog genes from *Mus musculus.* An extract of 10 orthologous proteins from the original dataframe (which has 19 results), is shown in Table 7.

UniProt ID 1	UniProt ID 2	Count	Curated	Reliabilty score
E5RH42	G3BP1_MOUSE	1	True	75
A0A6Q8PHP0	Q3V1L4	1	True	75
A0A6Q8PHP0	5NTC_MOUSE	1	True	75
E5RH42	P97855	1	True	75
B9EF68	P03921	1	True	75
B9EF68	NU5M_MOUSE	1	True	75
B9EF68	Q7GIP3	1	False	25
B9EF68	A0A141CM38	1	False	25
B9EF68	Q3TRR5	1	False	25
A0A6Q8PHP0	A0A494BBM7	1	False	25

Table 7. Orthologous proteins between Homo sapiens and Mus musculus obtained through the unified method, coming from the list of UniProt IDs of Homo sapiens,

6. Discussion

Recalling the objective of this work, a computational method has been developed to allow users to exploit data from multiple databases in a combined and straightforward manner, enabling the retrieval of orthologous genes form different species. To clearly understand how this method works and how it enhances both the quality of the results and ease of databases use, the results will be thoroughly analysed and compared.

6.1. InParanoid

As mentioned earlier, InParanoid collects its protein name through UniProt IDs. As shown in the Results section, there is only one orthologous gene to P04637 from *Homo sapiens* in *Mus musculus* coming from this database: P02340, which refers to Tp53 of *Mus musculus* (UniProt Consortium, 2024b).

6.2. OMA

OMA database provides a major number of orthologous genes than those in InParanoid, this is because OMA is distinguished from other databases because of its high specificity of its inference pipeline (Altenhoff et al., 2019). Besides P02350 gene ID, it provides five other gene IDs among which P53_MOUSE refers to the same protein as the previous one (UniProt Consortium, 2024b). The rest of the IDs belong to non-curated proteins (A0A158SIS7, Q91XH8, O70366 and Q549C9) present in TrEMBL, a database with non-curated proteins (Bairoch & Apweiler, 1999).

The results for InParanoid are lower than those in OMA, as InParanoid is based entirely on Swiss-Prot-TrEMBL, so the result proteins correspond to curated ones. This provides a solid quality and quantity of information.

6.3. Unified method

This method combines results from both databases, resulting in a total of 6 entries of orthologous proteins from *Mus musculus*. The UniProt ID P02340 appears twice with a reliability score of 100 %. As mentioned, the ID P53_MOUSE refers to the same protein as P02340 (Tp53), but receives a lower score (75 %) as it is only present in the OMA database.

While P53_MOUSE gene can be more visual and user-friendly for humans, according to UniProt (2022), the primary accession number (P02340) facilitate the unique identification of proteins in biological databases. It remains stable over time and is more suitable for data processing. Therefore, it is logical for both databases (InParanoid and OMA) to use this identifier, resulting in a higher reliability score.

This method is validated using new UniProt IDs from *Homo sapiens*, and a subset of the retrieved orthologous proteins between this species and *Mus musculus* is presented. Since these proteins are only listed in one database, their reliability scores vary and never reach 100 %. The highest scores are 75 % for curated proteins, while others only achieve about 25 %, which proves the capability of this method to correctly classify retrieved proteins.

The creation of this method not only facilitates the search for orthologous genes across species for the research community, as it combines two of the most important databases of orthologous proteins (Fernández-Breis et al., 2016), but also makes it more available by automating SPARQL queries, requiring only the correct species name and the UniProt ID. Moreover, the development of a reliability score based on the protein's presence in other databases and its curation status ensures the reliance of the results, which is crucial in fields such as comparative genomics (Trachana et al., 2011).

6.4. Strengths and limitations

Strengths:

- Automation of data retrieval processes, requiring minimal input from users (species names and UniProt ID), which enhances accessibility and usability through the use of SPARQL queries from different databases.
- Development of a reliability scoring system based on cross-database presence and curation status of proteins, ensuring robustness of the results.
- Integration of multiple orthologous gene databases that adhere to the ORTH ontology, facilitated by initiative like the Quest for Orthologs consortium.
- OMA and InParanoid are robust databases widely recognized for their comprehensive coverage and accuracy in ortholog data, providing reliable information for comparative genomic studies.

Limitations:

- GraphDB repository was initially set up with inference, causing significant complexity in handling complex queries, and requiring frequent repository restarts.
- Initially, the OMA database was uploaded to the repository but it was an older and reduced version, so the decision was made to use the OMA endpoint directly.
- TreeFam did not properly transform OrthoXML to RDF, and lacked connections between species and proteins in the repository, preventing ortholog search by species, thus it couldn't be integrated into this work.

- Due to the previous issue, only two databases (InParanoid and OMA) were unified, while the original intention was to include TreeFam as well, which couldn't be achieved within the timeframe.

6.5. Future approximations

As key future work, several areas for enhancement should be addressed to further develop this method. Resolving technical challenges with GraphDB to fully integrate TreeFam and other databases adopting the ORTH ontology will enrich the method's capability. The inclusion of additional databases would not only expand the breadth of orthology data accessible but also enhance the completeness of the reliability score by considering additional parameters.

Besides, collaboration with initiatives like Quest for Orthologs to promote standardization and interoperability will further advance the accessibility and utility of orthology databases. These efforts will not only refine the method presented here but also contribute to broader advancements in comparative genomics and biological research.

7. Conclusion

Orthologous information serves as a cornerstone in comparative genomics and biomedical research, playing a crucial role in understanding evolutionary relationships and biological functions across species. However, the current landscape of orthology databases is marked by significant heterogeneity, posing challenges to data interoperability. Addressing these challenges, the Quest for Orthologs consortium has been fundamental in advocating for standardized representations of ortholog information. The adoption of the ORTH ontology represents a main step towards semantic unification, facilitating the integration of diverse orthology databases such as InParanoid and OMA within a unified RDF repository.

To keep on contributing to the unification of the orthology research, this work has developed and demonstrated a computational method for retrieving orthologous genes across species using SPARQL queries on several databases which have adopted the ORTH ontology. The method successfully combines data from InParanoid and OMA, leveraging their strengths in orthology analysis. By automating SPARQL queries and implementing a reliability scoring system, this method enhances accessibility and usability for researchers seeking reliable orthology data.

In conclusion, ongoing efforts to enhance the method developed in this work will focus on several fronts. Technical refinements are necessary to resolve challenges with GraphDB, enabling seamless integration of additional orthology resources like TreeFam. Expanding the scope of integrated databases will further enrich the method's utility and reliability, accommodating a broader range of research needs in comparative genomics. Collaboration with initiatives promoting data standardization and interoperability will remain essential in advancing the accessibility and usefulness of orthology databases globally.

This work marks an initial stride towards consolidating existing resources, laying the groundwork for improved accessibility and streamlined queries across web-based orthology databases. It aims to provide the scientific community with a fast and comprehensive service for obtaining information from multiple orthologous databases.

8. References

The bibliographic reference style used throughout this work is American Psychological Associtation (APA), according to its seventh edition.

Ali, W., Saleem, M., Yao, B., et al. (2022). A survey of RDF stores & SPARQL engines for querying knowledge graphs. The VLDB Journal, 31, 1–26. https://doi.org/10.1007/s00778-021-00711-3

Altenhoff, A. M., Glover, N. M., & Dessimoz, C. (2012). Inferring orthology and paralogy. En M. Anisimova (Ed.), Evolutionary Genomics: Statistical and Computational Methods (pp. 159-186). Springer.

Altenhoff, A. M., Gil, M., Gonnet, G. H., & Dessimoz, C. (2013). Inferring Hierarchical Orthologous Groups from Orthologous Gene Pairs. PloS One, 8(1), e53786. https://doi.org/10.1371/journal.pone.0053786

Altenhoff, A. M., Levy, J., Zarowiecki, M., Tomiczek, B., Warwick Vesztrocy, A., Dalquen, D. A., Müller, S., Telford, M. J., Glover, N. M., Dylus, D., & Dessimoz, C. (2019). OMA standalone: orthology inference among public and custom genomes and transcriptomes. Genome Res., 29(7), 1152-1163. https://doi.org/10.1101/gr.243212.118

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Research, 27(1), 49-54. https://doi.org/10.1093/nar/27.1.49

Baum, D. (2008) Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups. Nature Education 1(1):190

Bawono, P., & Heringa, J. (2014). InParanoid. In A. Brahme (Ed.), Comprehensive Biomedical Physics (Vol. 6, pp. 93-110). Elsevier.

De Boissier, P., & Habermann, B. H. (2020). A Practical Guide to Orthology Resources. En Springer eBooks (pp. 41-77). https://doi.org/10.1007/978-3-030-57246-4_3

Dumontier, M., Baker, C. J., Baran, J., et al. (2014). The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. J Biomed Semant, 5, 14. https://doi.org/10.1186/2041-1480-5-14

Fernández-Breis, J. T., Chiba, H., Del Carmen Legaz-García, M., & Uchiyama, I. (2016). The Orthology Ontology: development and applications. Journal Of Biomedical Semantics, 7(1). https://doi.org/10.1186/s13326-016-0077-x

Fernández-Breis, J. T., Del Carmen Legaz-García, M., Chiba, H., & Uchiyama, I. (2015). Towards the Semantic Standardization of Orthology Content. SWAT4LS, 74-83. http://ceur-ws.org/Vol-1546/paper_19.pdf

Farias, T. M., Chiba, H., & Fernández-Breis, J. T. (2017). Leveraging Logical Rules for Efficacious Representation of Large Orthology Datasets. SWAT4LS. http://ceur-ws.org/Vol-2042/paper36.pdf

Herman, I., Fernández, S., Tejo Alonso, C., & Zakhlestin, A. (2022, March 14). SPARQLWrapper Documentation. Retrieved from https://rdflib.github.io/sparqlwrapper/

IBM. (n.d.). What is a knowledge graph? IBM. Retrieved June 15, 2024, from https://www.ibm.com/topics/knowledge-graph

Lin, Y., Xiang, Z., He, Y. (2011). Towards a semantic web application: Ontology-driven ortholog clustering analysis. In: ICBO.

Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Katta, H. Y., Mojica, A., Chen, I.-M. A., Kyrpides, N. C., & Reddy, T. (2019). Genomes OnLine database (GOLD) v.7: updates and new features. Nucleic Acids Res., 47(D1), D649–D659.

Ontotext. (n.d.). What are ontologies? Ontotext. Retrieved June 16, 2024, from https://www.ontotext.com/knowledgehub/fundamentals/what-are-ontologies/

Ontotext. (2024a). Chapter 8: References. In GraphDB SE Documentation Release 9.11.0 (Version 9.11.0, p. 581). Ontotext.

Ontotext. (2024b). Chapter 8: References. In GraphDB SE Documentation Release 9.11.0 (Version 9.11.0, p. 597). Ontotext.

Pandas development team. (2022). Pandas (Version 1.5.1). Retrieved July 2, 2024 from https://pypi.org/project/pandas/

Project Jupyter. (2024). JupyterLab documentation. Retrieved July 1, 2024, from https://jupyterlab.readthedocs.io/en/stable/

Prosdocimi, F., Chisham, B., Pontelli, E., Thompson, J. D., & Stoltzfus, A. (2009). Initial Implementation of a Comparative Data Analysis Ontology. Evolutionary Bioinformatics Online, 5, EBO.S2320. https://doi.org/10.4137/ebo.s2320

Reitz, K. (2022). Requests: HTTP for Humans (Version 2.27.1). Retrieved July 2, 2024, from https://pypi.org/project/requests/

Reitz, K. (2024). Certifi Documentation (Version 2024.6.2). Retrieved July 1, 2024, from https://pypi.org/project/certifi/

Roux, J., & Robinson-Rechavi, M. (2010). An ontology to clarify homology-related concepts. Trends In Genetics, 26(3), 99-102. https://doi.org/10.1016/j.tig.2009.12.012

Sioutos, N., de Coronado, S., Haber, M. W., Hartel, F. W., Shaiu, W. L., & Wright, L. W. (2007). Nci thesaurus: a semantic model integrating cancer-related clinical and molecular information. Journal of biomedical informatics, 40(1), 30-43.

Sonnhammer, E. L., Gabaldon, T., da Silva, A. W. S., Martin, M., Robinson-Rechavi, M., Boeckmann, B., Thomas, P. D., & Dessimoz, C., et al. (2014). Big data and other challenges in the quest for orthologs. Bioinformatics, btu492.

Trachana, K., Larsson, T. A., Powell, S., Chen, W. H., Doerks, T., Muller, J., & Bork, P. (2011). Orthology prediction methods: a quality assessment using curated protein families. Bioessays, 33(10), 769-780. https://doi.org/10.1002/bies.201100062

UniProt Consortium. (2024a). P04637 Cellular tumor antigen p54 Homo sapiens. UniProt. https://www.uniprot.org/uniprotkb/P04637

UniProt Consortium. (2024b). P02340 Cellular tumor antigen p53 Mus musculus. UniProt. https://www.uniprot.org/uniprotkb/P02340

UniProt. (2022, October 14). Accession numbers. Retrieved July 10, 2024, from https://www.uniprot.org/help/accession_numbers

World Wide Web Consortium (W3C). (2007). VLDB 2007. Retrieved from https://www.w3.org/2007/03/VLDB/

Zahn-Zabal, M., Dessimoz, C., & Glover, N. M. (2020). Identifying orthologs with OMA: A primer. F1000Res, 9, 27. https://doi.org/10.12688/f1000research.21508.1