# Towards the semantic standardization of orthology content

Jesualdo Tomás Fernández-Breis<sup>1</sup>, María del Carmen Legaz-García<sup>1</sup>, Hirokazu Chiba<sup>2</sup>, Ikuo Uchiyama<sup>2</sup>

Facultad de Informática, Universidad de Murcia, IMIB-Arrixaca, Murcia, Spain National Institute for Basic Biology, National Institute of Natural Sciences, Okazaki, Japan

Abstract. The amount of resources and data about orthology, together with the increasing interest in orthologs in biomedical research has triggered the need for the shareability of the content generated by the different tools and stored in the different databases. In recent years OrthoXML permitted to advance in the standardization of the content for exchange within the orthology community, but the interest in exchanging content with other communities suggested the research on the use of Semantic Web languages like RDF and to use ontologies for making the meaning of the content explicit. This possibility was reinforced by the existence of initial efforts for using Semantic Web technologies for representing orthology content. In this work, we describe the advances done with the objective of the semantic standardization of orthology content. The need for a common ontology has permitted to obtain a draft of an orthology ontology built by reusing existing ontologies and following best practices in ontology construction. This ontology should serve as knowledge framework for the semantic standardization of orthology content. A mapping between OrthoXML and this orthology ontology has been designed, specified and applied to sample OrthoXML datasets. For this purpose, the Semantic Web Integration Tool (SWIT) has been used, which provides support for doing the previously described actions and permit to create an integrated repository from multiple orthology content sources.

#### 1 Introduction

In recent years, the number of genome sequences determined has significantly increased and many on-going research projects will permit to know not only the reference genome of many organisms but also the genome of individuals. In this new era, being able to perform computational comparative analysis might bring many opportunities to biomedical research. Homology information can play a central role in integrating and comparing multiple genomes, because homology permit establishing evolutionary relations between genes from multiple species. Three basic concepts, namely, homologs, orthologs, and paralogs need to be distinguished in this context[5]: (1) homologous genes have diverged from an ancestral gene; homology may hold for genes of the same species or from different

species; (2) orthologous genes have diverged by speciation from an ancestral gene, and their biological functions are usually conserved; orthology may hold between genes from different species; (3) paralogous genes have diverged by duplication from an ancestral gene; paralogy may hold between genes of the same species or from different species.

Despite homology relations are usually obtained in particular studies from a pairwise perspective, they are usually calculated and represented as clusters, that is, groups of genes holding homologous relationship. Homolog clusters are further divided into two types: ortholog groups consist of genes derived from a speciation event and paralog groups consist of genes derived from a duplication event. Thus, ortholog and paralog groups can be represented in a form of nested hierarchy according to their evolutionary history, where each orthologous group is associated with a taxonomic range that corresponds to a set of organisms derived from a specific speciation event.

Ortholog information is a useful resource to link the corresponding genes of different species and transfer the biological knowledge of model organisms to organisms with newly sequenced genomes. In addition, ortholog groups are a vital resource for the comparative analysis of multiple genomes, and they provide a basis for the analysis of phylogenetic profiles [11].

The Quest for Orthologs (QfO) Consortium has identified more than forty resources about orthology (http://questfororthologs.org/orthology\_databases), which represent different needs of information management in the orthology field. Many of these resources store information about prediction of gene evolutionary relations and there is a diversity of objectives for these databases. There is heterogeneity in how data are stored and shared by these resources. For example, Inparanoid [13] stores orthology relations between two species, whereas OrthoMCL [1] and MBGD [20] stores orthology groups among multiple genomes. OMA [3] provides various types of orthology relations including pairwise orthologs and hierarchical ortholog groups. Many resources use their own representation format based on tabular files, despite this community has developed the OrthoXML and SeqXML [15] formats to standardize the representation of orthology data. OrthoXML permits the comparison and integration of orthology data from different resources within the orthology community.

In recent years, semantic web formats have been used for representing orthology data. OGO [8] was created with the purpose of providing an integrated resource of information about genetic human diseases and orthologous genes. OGO integrated information from orthology databases such as Inparanoid [13], OrthoMCL [1], and OMIM [7]. This resource developed an OWL ontology for representing the domain knowledge. More recently, RDF has been used by sharing the content of the Microbial Genome Database for Comparative Analysis (MBGD) [2] database. This resource also developed an OWL ontology for representing the domain knowledge, called OrthO, which had similar concepts to the OGO one, despite being developed independently.

The report of the 2013 QfO meeting [19] identified a series of aspects about semantics that have been the key drivers of our activities: (1) the orthology

community should use shared ontologies to facilitate data sharing; (2) exploiting automated reasoning should be beneficial for the QfO consortium. In this work, we describe the development of a draft on an orthology ontology, built from existing related ontologies and we explain how we are approaching the transformation and exploitation of existing orthology datasets using semantic web technologies. We believe this work provides a step forward towards the standardization in the orthology community.

# 2 The Orthology Ontology

In this section, we describe the development of the orthology ontology. The design of the ontology has followed two basic principles: (1) reusing content from existing ontologies to facilitate interoperability across biomedical domains: (2) modelling based on membership of genes to clusters of homologs, orthologs or paralogs, so pairwise relations are obtained by analysing the structure and content of the dataset.

## 2.1 Related ontologies

We studied how orthology related concepts and properties were covered in potentially related ontologies. Our initial selection contained the following ontologies: Comparative Data Analysis Ontology (CDAO) [12], Relations Ontology (RO) [18], Semanticscience Integrated Ontology (SIO) [4], Homology Ontology (HO) [14], National Cancer Institute Thesaurus (NCIt) [17], and Clusters of Orthologous Groups (COG) Analysis Ontology [6]. It should be noted that these ontologies present some overlaps and, in such cases, we inspected the axioms and the textual content in order to make our decisions. For example, HO, NCIt and COG were not finally reused, since we found the other ontologies more appropriate. The CDAO provides a series of classes of interest that we decided to reuse for phylogenetic purposes. It contains a taxonomic module that defines types of Trees, which we found useful for defining phylogenetic trees. It also provides a taxonomic module of hereditary changes, which we can use in order to define concepts like orthologs or paralogs. The RO and the SIO include a series of properties of interest for our domain like has part, is part of, in taxon and many evolutionary relations. It seems that the recent versions of the RO include most classes of the HO as properties. The SIO also defines types of genes and concepts like databases in a way that can be effectively re-used for our purpose, so we also selected it. We also reused the NCBI taxonomy for representing the species.

## 2.2 Description of the orthology ontology

The orthology ontology has been implemented in OWL using Protégé and it is currently available at https://github.com/qfo/OrthologyOntology. Figure 1 represents an excerpt of this ontology. There, classes are represented as boxes and

properties as arrows. The entities without prefix are defined in our orthology ontology. The prefixes cdao, sio and ro represent entities reused from the corresponding ontologies. We can see HomologsCluster is a subclass of GeneTreeNode, and it has two descendants, namely, OrthologsCluster and ParalogsCluster, which are associated with cdao:speciation and cdao:geneDuplication respectively. GeneTreeNode is a subclass of cdao:Node, which is not shown in the figure. The members of HomologsCluster are instances of GeneTreeNode, which means that they can be clusters of homologs or sequence units. The class SequenceUnit has three subclasses, Gene, Subgene and Protein. The membership property instead of two hasOrthologous and hasParalogous, because the pairwise relations are obtained by analysing the tree. Finally, genes and proteins are linked to ncbi:organisms through the property  $ro:in\_taxon$ .

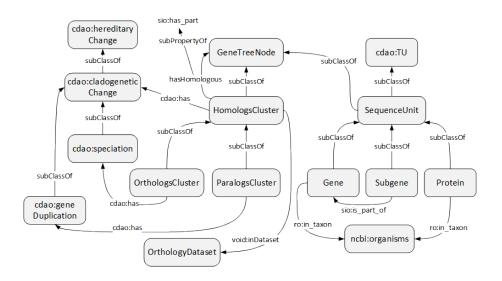


Fig. 1. Excerpt of the orthology ontology

## 3 Semantic Transformation of Orthology Data

The availability of the orthology ontology enables the possibility of using it as schema for the generation of RDF datasets from existing orthology resources. As it has been previously mentioned, OrthoXML has been a format proposed for the exchange of orthology content. Hence, we have defined a mapping between OrthoXML and the orthology ontology, whose main entities can be seen in Table 1. The first four rows show mappings corresponding to entities in the ontology. The rest are examples of mappings to properties in the ontology. For instance orthologGroup/paralogGroup means that a group of paralogs has been defined in an group of orthologs and in the ontology this is represented

through the property *hasHomologous*. The complete mapping file is available at https://github.com/qfo/OrthologyOntology.

Table 1. Main mappings between OrthoXML and the Orthology Ontology

OrthoXML	Orthology Ontology	
orthologGroup	OrthologsCluster	
paralogGroup	ParalogsCluster	
gene	Gene	
NCBITaxId	organisms	
orthologGroup/orthologGroup	<orthologscluster, hashomologous,="" orthologscluster=""></orthologscluster,>	
orthologGroup/paralogGroup	<orthologscluster,hashomologous, paralogscluster=""></orthologscluster,hashomologous,>	
orthoXML/species/database	<orthologydataset, database="" hassource,=""></orthologydataset,>	

One of the technical objectives of the work is to be able to use tools available for supporting the different processes, so the managers of orthology resources do not need to create their own transformation scripts into semantic formats. To this end, we have used the SWIT tool (http://sele.inf.um.es/swit), which was used in our research group to develop and maintain the OGO Linked Open Dataset [9]. SWIT provides a web interface through which the user is guided to perform all the steps of the process:

- 1. Inputs and outputs: the user has to provide the OrthoXML schema and the orthology ontology, as well as the corresponding datasets to be transformed. SWIT is able to generate the dataset in OWL or RDF formats, which can be downloaded or directly stored into a triple store like Virtuoso. This tool also enables the user to define the structure of the URI for the RDF/OWL instances (by default, the ontology base URI).
- 2. Definition of the mappings: the user can define the mappings between OrthoXML and the orthology ontology. Alternatively, mappings created in previous sessions can be uploaded and re-used. In this case, the file containing the mappings between OrthoXML and the orthology ontology described in the previous section was uploaded.
- 3. Once the mappings have been defined, they can be executed to generate the corresponding repository.

SWIT applies the mapping rules to the data source to generate the semantic content. Briefly speaking, the mapping rules associate entities and attributes of the OrthoXML schema with owl:Class, owl:DatatypeProperty and owl:ObjectProperty defined in the ontology. SWIT also generates integrated repositories from multiple sources. For supporting the integration process, SWIT permits the definition of identity rules, which aim to avoid the creation of redundant individuals in the semantic repository. Automated reasoning ensures that only logically consistent content is transformed. To this end, SWIT is currently using Hermit [16] as reasoner.

## 4 Calculating Pairwise Relationships

In this section we focus on how we can obtain pairwise relationships between genes in a given tree. Figure 2 will be used to illustrate some concepts. In the next subsections, we will describe how we can obtain them from the RDF repository.

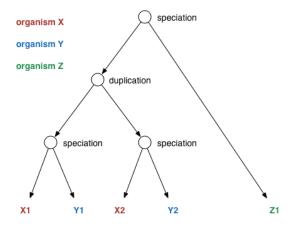


Fig. 2. Sample phylogenetic tree

#### 4.1 Pairwise orthologs

We can define pairwise orthologs as pairs of genes from different species whose common ancestor has a speciation event associated. If we inspect Figure 2, X1 has two ancestral nodes associated with speciation events: one is the common ancestor of gene X1 and gene X1, and the other is the common ancestor of gene X1 and gene Z1. Therefore, we can say that Y1 and Z1 are orthologs of X1, so the pairs (Y1, X1) and (Z1, X1). Table 2 shows the query for that purpose. It can be explained as the search for ?common\_ancestor, which is a cluster of orthologs (so it has a speciation event associated) and is the common ancestor of two nodes of trees (?tree\_node1 and ?tree\_node2) to which two genes (?gene1 and ?gene2) from different species belong. We have omitted the taxonomic range in the query for simplicity.

## 4.2 Pairwise paralogs

We can define pairwise paralogs as pairs of genes whose common ancestor has a duplication event associated. This means that the genes may be found in the same species or in different ones. If we inspect Figure 2, X2 and Y2 are paralogs to X1, since their common ancestor have a duplication event associated. Therefore, a similar query to the previous one for orthologs can work for this purpose by introducing two changes: use of *ParalogsCluster* instead of *OrthologsCluster* 

and not using the condition for species in the *FILTER* clause. This part is not shown. However, this query cannot work for Inparanoid in which in-paralogs are included in *OrthologsCluster* and not defined as *ParalogsCluster*. Thus, in this case, another query (see Table 3) is necessary to search for members of *OrthologsCluster* from the same species. In fact, many orthology resources use flat structure of ortholog clusters where all member genes are directly associated with the top-level speciation node. In such cases, although the same query as in Table 3 can be used to list all pairwise intra-species inparalogy relationships, it is impossible to distinguish between orthologs (e.g. X1 and Y1) and inter-species inparalogs (e.g. X1 and Y2).

**Table 2.** The SPARQL query for pairwise orthologs

```
SELECT ?gene1 ?species1 ?gene2 ?species2 ?common_ancestor
WHERE {
    ?common_ancestor a orthology:OrthologsCluster .
    ?common_ancestor orthology:hasHomologous ?tree_node1 .
    ?common_ancestor orthology:hasHomologous ?tree_node2 .
    ?tree_node1 orthology:hasHomologous* ?gene1 .
    ?tree_node2 orthology:hasHomologous* ?gene2 .
    ?gene1 a orthology:Gene .
    ?gene2 a orthology:Gene .
    ?gene2 a orthology:Gene .
    ?gene1 obo:RO_0002162 ?species1 .
    ?gene2 obo:RO_0002162 ?species2 .
    FILTER (?tree_node1 != ?tree_node2 && ?species1 != ?species2) }
```

Table 3. The SPARQL query for second condition for pairwise paralogs

## 5 Integrated Data Exploitation

The mapping defined between OrthoXML and the orthology ontology can be applied over specific datasets. In this work, we have transformed data from two orthology resources, Inparanoid 8 [10] and OMA Sept 2014 hierarchical orthologous groups [3]. These two resources provide data in OrthoXML format, but they

structure orthology data in a different manner, so they constitute an interesting exploratory use case. For example, OMA uses clusters of orthologs and clusters of paralogs in a hierarchical manner but Inparanoid only considers clusters of orthologs, which may contain paralogs generated by duplications after the speciation of the two target species (termed in-paralogs [13]). This means that all paralogy relations have to be inferred from Inparanoid datasets. Besides, OMA stores the taxonomic range associates with a certain cluster, but Inparanoid does not. Provided that both resources use OrthoXML we are able to reuse the same mapping file for transforming the data. SWIT is able to execute the rules that can be applied to each dataset, which permits to skip in the case of Inparanoid, for instance, the rule for transforming the taxonomic range. The transformed RDF contains 8798758 genes from OMA, 1713180 genes for Homo sapiens orthologs and 1367940 genes for Mus musculus orthologs from Inparanoid. Table 4 shows the SPARQL query for getting the orthologs of the human gene OR4D2 in Mus Musculus, which would include the results from both Inparanoid and OMA. The results of this query are shown in Table 5. In this use case, we have used Virtuoso 7 as the triple store. Some sample queries exploiting this integrated dataset are available at https://github.com/qfo/OrthologyOntology.

**Table 4.** Orthologs of the human gene OR4D2 in *Mus Musculus*. 9606 and 10090 stand for the taxonomic identifier for *Homo sapiens* and *Mus Musculus*, respectively

```
SELECT ?gene ?species ?database
WHERE {
    ? common\_ancestor\ a\ orthology: Orthologs Cluster\ .
    ? common\_ancestor\ orthology: has Homologous\ ? tree\_node1\ .
    ?common_ancestor orthology:hasHomologous ?tree_node2 .
    ?common_ancestor void:inDataset ?dataset .
    ?dataset orthology:hasSource ?database .
    ?tree_node1 orthology:hasHomologous* ?gene1 .
    ?tree_node2 orthology:hasHomologous* ?gene2 .
    ?gene1 a orthology:Gene .
    ?gene2 a orthology:Gene .
    ?gene1 obo:RO_0002162 ?species1.
    ?gene2 obo:RO_0002162 ?species2.
    ?gene1 dcterms:identifier ?id .
    ?gene2 dcterms:identifier ?gene .
    ?species2 rdfs:label ?species .
    FILTER (?tree_node1 != ?tree_node2 && ?species1 != ?species2
       && regex(?species1, "_9606$") && regex(?species2, "_10090$"))
    VALUES (?id ) {("OR4D2")} }
```

#### 6 Discussion and Conclusion

In this paper we have described how we can approach the standardization of orthology content using semantic web technologies. We have been able to build a draft of the orthology ontology by reusing existing ones. We have been able

**Table 5.** The results of the query shown in Table 4

gene	species	database
	"Mus musculus"	
"Olfr462"	"Mus musculus"	"InParanoid"
	"Mus musculus"	
	"Mus musculus"	
"MOUSE03762"	"Mus musculus"	"OMA"

to use state-of-the-art tools for all the processes involved: construction of the ontology, definition of mappings, transformation and exploitation of the data. Our effort would permit the systematic application of the process to any OrthoXML database to generate an integrated knowledge base and to carry out evaluation and testing processes in both usefulness and performance. There are also remaining tasks and challenges. First, the draft of the ontology needs to be improved in terms of classes, properties and documentation of what has already been produced. We have started by defining queries of pairwise paralogs and orthologs, but other relations and concepts like in-paralogs, out-paralogs, etc. could be defined and encoded as queries. So far, there is no friendly way to query the transformed content, only through SPARQL endpoints, which are for machines rather than for humans. Efforts in that sense should be made, taking into account that there will be distributed SPARQL endpoints of orthology-related content. The future work also includes the development of a Linked Data API over the triple store that has method calls enabling the exploitation of the datasets.

## Acknowledgements

This work was supported by the Ministerio de Economía y Competitividad and the FEDER programme through grant TIN2014-53749-C2-2-R2, the Fundación Séneca through grant 15295/PI/10, and the National Bioscience Database Center, Japan Science Technology Agency. We would like to thank the organizers of the BioHackathon meeting (http://www.biohackathon.org) for providing us an invaluable opportunity to accomplish this work.

## References

- Chen, F., Mackey, A.J., Stoeckert, C.J., Roos, D.S.: Orthomcl-db: querying a comprehensive multi-species collection of ortholog groups. Nucleic acids research 34(suppl 1), D363–D368 (2006)
- 2. Chiba, H., Nishide, H., Uchiyama, I.: Construction of an ortholog database using the semantic web technology for integrative analysis of genomic data. PloS one 10(4) (2015)
- 3. Dessimoz, C., Cannarozzi, G., Gil, M., Margadant, D., Roth, A., Schneider, A., Gonnet, G.H.: Oma, a comprehensive, automated project for the identification of orthologs from complete genome data: introduction and first achievements. In: Comparative Genomics, pp. 61–72. Springer (2005)

- Dumontier, M., Baker, C.J., Baran, J., Callahan, A., Chepelev, L.L., Cruz-Toledo, J., Nicholas, R., Rio, D., Duck, G., Furlong, L.I., et al.: The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery. J. Biomedical Semantics 5, 14 (2014)
- Koonin, E.V.: Orthologs, paralogs, and evolutionary genomics 1. Annu. Rev. Genet. 39, 309–338 (2005)
- 6. Lin, Y., Xiang, Z., He, Y.: Towards a semantic web application: Ontology-driven ortholog clustering analysis. In: ICBO (2011)
- McKusick, V.A.: Mendelian inheritance in man: a catalog of human genes and genetic disorders, vol. 1. JHU Press (1998)
- 8. Miñarro-Gimenez, J.A., Madrid, M., Fernandez-Breis, J.T.: Ogo: an ontological approach for integrating knowledge about orthology. BMC bioinformatics 10(Suppl 10), S13 (2009)
- 9. Miñarro-Giménez, J.A., Egaña Aranguren, M., Villazón-Terrazas, B., Fernández Breis, J.T.: Translational research combining orthologous genes and human diseases with the ogolod dataset. Semantic Web 5(2), 145–149 (2014)
- O'Brien, K.P., Remm, M., Sonnhammer, E.L.: Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic acids research 33(suppl 1), D476–D480 (2005)
- 11. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O.: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proceedings of the National Academy of Sciences 96(8), 4285–4288 (1999)
- 12. Prosdocimi, F., Chisham, B., Pontelli, E., Thompson, J.D., Stoltzfus, A.: Initial implementation of a comparative data analysis ontology. Evolutionary bioinformatics online 5, 47 (2009)
- 13. Remm, M., Storm, C.E., Sonnhammer, E.L.: Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. Journal of molecular biology 314(5), 1041–1052 (2001)
- Roux, J., Robinson-Rechavi, M.: An ontology to clarify homology-related concepts.
   Trends in Genetics 26(3), 99–102 (2010)
- 15. Schmitt, T., Messina, D.N., Schreiber, F., Sonnhammer, E.L.: Seqxml and orthoxml: standards for sequence and orthology information. Briefings in bioinformatics p. bbr025 (2011)
- Shearer, R., Motik, B., Horrocks, I.: HermiT: A highly-efficient OWL reasoner. In: Proceedings of the 5th International Workshop on OWL: Experiences and Directions (OWLED 2008). pp. 26–27 (2008)
- 17. Sioutos, N., de Coronado, S., Haber, M.W., Hartel, F.W., Shaiu, W.L., Wright, L.W.: Nci thesaurus: a semantic model integrating cancer-related clinical and molecular information. Journal of biomedical informatics 40(1), 30–43 (2007)
- 18. Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., Rosse, C.: Relations in biomedical ontologies. Genome biology 6(5), R46 (2005)
- Sonnhammer, E.L., Gabaldón, T., da Silva, A.W.S., Martin, M., Robinson-Rechavi, M., Boeckmann, B., Thomas, P.D., Dessimoz, C., et al.: Big data and other challenges in the quest for orthologs. Bioinformatics p. btu492 (2014)
- 20. Uchiyama, I., Mihara, M., Nishide, H., Chiba, H.: Mbgd update 2015: microbial genome database for flexible ortholog analysis utilizing a diverse set of genomic data. Nucleic acids research 43(D1), D270–D276 (2015)