BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 17 2011, pages 2468–2470
doi:10.1093/bioinformatics/btr375

Databases and ontologies Advance Access publication June 22, 2011

ontoCAT: an R package for ontology traversal and search

Natalja Kurbatova[1,*], Tomasz Adamusiak[1], Pavel Kurnosov[1], Morris A. Swertz[1,2] and Misha Kapushesky[1]

[1]EMBL Outstation-Hinxton, European Bioinformatics Institute, Cambridge, UK and [2]Genomics Coordination Center, University Medical Center Groningen and University of Groningen, Groningen, The Netherlands

Associate Editor: Alex Bateman

ABSTRACT

Motivation: There exist few simple and easily accessible methods to integrate ontologies programmatically in the R environment. We present ontoCAT—an R package to access ontologies in widely used standard formats, stored locally in the filesystem or available online. The ontoCAT package supports a number of traversal and search functions on a single ontology, as well as searching for ontology terms across multiple ontologies and in major ontology repositories.

Availability: The package and sources are freely available in Bioconductor starting from version 2.8: http://bioconductor.org/help/bioc-views/release/bioc/html/ontoCAT.html or via the OntoCAT website http://www.ontocat.org/wiki/r.

Contact: natalja@ebi.ac.uk

Received on January 13, 2011; revised on June 8, 2011; accepted on June 17, 2011

1 INTRODUCTION

The R package ontoCAT was created to support basic operations on ontologies: traversal and search, uniform access to ontologies in OWL (Lacy, 2005) and OBO (Barry *et al*., 2007) formats and to provide R access to major ontology repositories OLS and BioPortal.

Several hundreds of public ontologies and numerous private ontologies for describing biological data exist today. Using ontologies in R (Gentleman, 2008; R Development Core Team, 2006) is difficult due to the lack of uniform package support. At the same time numerous Java-based ontology projects are available. ontoCAT takes advantage of a standard Java library with the same name 'ontoCAT' (Adamusiak *et al*., 2010, 2011) to implement its functionality.

The ontoCAT package:

- gives unified, format-independent access to ontology terms and the ontology hierarchy represented in OWL and OBO formats;
- provides basic methods for ontology traversal, such as searching for terms, listing a specific term's relations, showing paths to the term from the root element of the ontology, showing flattened-tree representations of the ontology hierarchy; and
- supports working with groups of ontologies and with major public ontology repositories: searching for terms across ontologies, listing available ontologies and loading ontologies for further analysis as necessary.

No other package with similar functionality exists at the moment in the R environment.

The integration of the above functionality into R allows combining and automating ontology-related tasks. Different examples of ontology-related tasks that can be accomplished with the help of the ontoCAT package are given in the package documentation (http://www.ontocat.org/wiki/OntocatGuide): gene enrichment test and grouping of results, search and re-annotation of free-text to ontology and operations with relationships.

ontoCAT has been included into Bioconductor, the main R open source project in bioinformatics. ontoCAT has been downloaded 347 times since its first release in December 2010.

There is a large research community already using R to work with Gene Ontology (GO). Working with other ontologies is not as well–developed and ontoCAT helps to fill the gap.

2 METHODS

The ontoCAT R package consists of two main parts, grouping similar methods:

- Single ontology traversal methods.
- Methods to work across multiple ontologies.

2.1 Single ontology traversal methods

The ontoCAT package can load an ontology in OWL or OBO format from a local file or on-the-fly from a URI.

Reasoning over ontologies and extracting relationships is supported by using HermiT (Motik *et al*., 2009) reasoner. OBO ontologies are translated by OWL API (Horridge and Bechhofer, 2009) into valid OWL format that can be reasoned over.

Ontologies can also be loaded from ontology repositories. Two public repositories are supported: BioPortal for accessing and sharing biomedical ontologies (Noy *et al*., 2009), currently hosting 241 ontologies and the Ontology Lookup Service (OLS) for querying multiple ontologies (Cote *et al*., 2008), currently hosting 81 ontologies.

To load an ontology `getOntology(path/accession)` method of the *Ontology* class is available. It takes a single argument, specifying the local filesystem path, the full URI for the ontology file, or its OLS/BioPortal accession.

The reference ontology supported by ontoCAT is Experimental Factor Ontology (EFO) (Malone *et al*., 2010), developed for applications in functional genomics. The latest version of EFO can be loaded by using the method `getEFO()`.

When an ontology is loaded, other ontoCAT methods become available.

*To whom correspondence should be addressed.

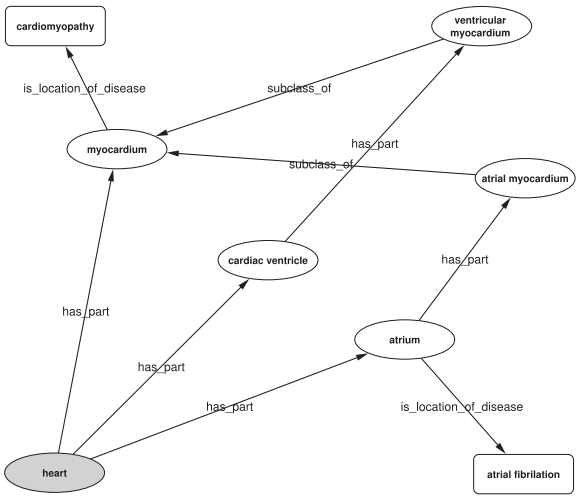

Fig. 1. An example of a directed acyclic graph representing all the relationships in an ontology for a particular EFO ontology term 'EFO_0000815' (heart). Edges are labeled according to the relationship. Organism part classes are represented as ellipses and disease classes are shown as rectangles. The `ontoCAT` package was used to compute the relationships which were later processed in Cytoscape (Cline *et al.*, 2007).

In `ontoCAT` the subsumption 'subclass/superclass' is supported in a user friendly form of 'child – parent' relationship.

No distinction is made between universals (classes) and particulars (instances) as they are both treated as ontology terms with parent–child relationship: class is treated as parent, instances are children of the class.

The advantage of using a reasoner in `ontoCAT` is the ability to work with different relationships in addition to subsumption. However, if reasoning over ontology is not desirable `getOntologyNoReasoning` method should be used instead of `getOntology` method described above.

An example of relationships that can be retrieved by `ontoCAT` starting from a particular EFO ontology term 'EFO_0000815' (heart) is given in Figure 1.

Below there are several examples of `ontoCAT` methods:

- `searchTerm(Ontology, 'myocardium')` returns a list of terms where 'myocardium' is mentioned;
- `getTermParentsById(Ontology, 'EFO_0003087')` lists parents of the term 'EFO_0003087' (atrial myocardium): 'EFO_0000819' (myocardium) (see Fig. 1);
- `getTermById(Ontology, 'EFO_0000819')` returns ontology term by its accession: 'EFO_0000819' (myocardium);
- `getTermChildren(Ontology, term)` lists children of the term 'EFO_0000819' (myocardium): 'EFO_0003087' (atrial myocardium) and 'EFO_0003088' (ventricular myocardium) (see Fig. 1);
- `showHierarchyDownToTerm(Ontology, 'EFO_0000819')` prints out a flattened-tree representation of the ontology from the root term down to 'EFO_0000819' (myocardium) by using parent–child relationships;
- `getTermRelationsById(Ontology, 'EFO_0000815', 'has_part')` returns terms in relation 'has_part' with 'EFO_0000815' (heart): 'EFO_0000277' (atrium), 'EFO_0000819' (myocardium) and 'EFO_0000317' (cardiac ventricle) (see Fig. 1);
- `getTermSynonyms(Ontology, term)` returns a list of synonyms for the term 'EFO_0000819' (myocardium): 'muscle of heart', 'cardiac muscle', 'heart muscle'; and
- `getRootTerms(Ontology)` returns a list of terms without parents in ontology of interest.

A number of self-descriptive methods like `showPathsToTerm(Ontology, term)`, `isRoot(Ontology, term)` are also available. The full list of methods together with descriptions is provided in online documentation (http://www.ontocat.org/wiki/r) and is included into the package.

2.2 Operations on multiple ontologies

`ontoCAT` provides methods to work with groups or 'batches' of ontologies, local or web-based. Users can search for terms across such resources and load specific individual ontologies by accession.

To create a local batch of ontologies the `getOntologyBatch(path)` method of the *batch* class is provided, taking a single argument: the path to the local directory containing ontology files.

By default, a call to `getOntologyBatch()` without any arguments will load the EFO ontology. Ontologies can be added to an existing batch as needed via the `addOntology()` method.

After a batch of ontologies is created, various methods become available, including:

- `searchTerm(batch, 'heart')` searches for the term in all ontologies in the batch;
- `searchTermInOLS('heart')` searches for the term in OLS repository;
- `searchTermInBioPortal('heart')` searches for the term in BioPortal repository; and
- `searchTermInAll(batch,'heart')` searches for the term in all ontologies in the batch as well as in OLS and BioPortal repositories.

The `ontoCAT` package also provides methods to list ontologies in batches and to list ontologies available in BioPortal and OLS: `listLoadedOntologies(batch)`, `listOLSOntologies(batch)` and `listBioportalOntologies(batch)`.

When the sought terms are found and term-specific operations (parent/child/other relationships retrieval, etc.) are needed, the `getOntology(batch,accession)` returns the ontology object for the concrete ontology with all single-ontology methods available.

3 TECHNICAL DETAILS

The package is based primarily on the Ontology Common API Tasks Java library, on the OWL API and depends on `rJava` R package. HermiT reasoner is used to support relationships.

`ontoCAT` R package is open-source and is available under the Apache License Version 2.0.

We provide two versions of `ontoCAT`:

- Light-weight `ontoCAT` package version is available in Bioconductor (http://bioconductor.org) starting from release 2.7, and includes all single-ontology functionality except for methods to work with multiple ontologies and search in OLS and BioPortal.
- Full version includes batch methods and due to package size limitations are available only from the project website.

The package sources and full documentation are available at http://www.ontocat.org/wiki/r.

4 CONCLUSION

The `ontoCAT` R package consists of convenient methods for working with ontologies in the R environment. The package has been successfully used in a number of projects in the Functional

Genomics Group at the European Bioinformatics Institute and this is its first public release.

The package provides basic operations on ontologies represented in standard formats and enables searches in online ontology repositories: OLS and BioPortal.

ACKNOWLEDGEMENT

Ontology Common API Tasks development team and the EBI Gene Expression Atlas development team.

Funding: European Community's Seventh Framework Programme projects GEN2PHEN (grant number 200754); SYBARIS (grant number 242220); NWO/Rubicon (grant number 825.09.008).

Conflict of Interest: none declared.

REFERENCES

Adamusiak,T. *et al*. (2010) OntoCAT – a simpler way to access ontology resources. *Nature Precedings*, [Epub ahead of print; doi:10.1038/npre.2010.4666.1].

Adamusiak,T. *et al*. (2011) OntoCAT – simple ontology search and integration in Java, R and REST\JavaScript. *BMC Bioinformatics*, **12**, 218.

Barry,S. *et al*. (2007) The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.

Cline,M.S. *et al*. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protocol.*, **2**, 2366–2382.

Cote,R.G. *et al*. (2008) The ontology lookup service: more data and better tools for controlled vocabulary queries. *Nucleic Acids Res.*, **36**(Suppl. 2), W372–W376.

Gentleman,R. (2008) *R Programming for Bioinformatics*. Chapman & Hall\CRC.

Horridge,M. and Bechhofer,S. (2009) The OWL API: a Java API for working with OWL 2 ontologies. *OWLED 2009, 6th OWL Experienced and Directions Workshop*. Chantilly, Virginia.

Lacy,L.W. (2005) *OWL: representing Information Using the Web Ontology Language*. Trafford Publishing, Victoria, BC.

Motik,B. *et al*. (2009) Hypertableau reasoning for description logics. *J. Art. Intell. Res.*, **36**, 165–228.

Noy,N.F. *et al*. (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.*, **37**, W170–W173.

Malone,J. *et al*. (2010) Modeling sample variables with an experimental factor ontology. *Bioinformatics*, **26**, 1112–1118.

R Development Core Team (2006) R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. Vienna, Austria.