Databases and ontologies

 ontoCAT: an R package for ontology traversal and search

Natalija Kurbatova1,∗, Tomasz Adamusiak1, Pavel Kurnosov1, Morris A. Swertz1,2 and Misha Kapushesky1.

1EMBL Outstation-Hinxton, European Bioinformatics Institute, Cambridge, UK and 2Genomics 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 programatically 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.


Contact: natalija@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, 2009) 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 use ontology(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.
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, term)` searches for the term in all ontologies in the batch;
- `searchTermInOls(batch, term)` searches for the term in OLS repository;
- `searchTermInBioportal(batch, term)` searches for the term in BioPortal repository; and
- `searchTermInAll(batch, term)` 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)` lists all ontologies in the batch as well as in OLS and BioPortal repositories.
- `listOntologies(batch)` and `listOlsOntologies(batch)` lists all ontologies and search in OLS and BioPortal.

When the sought terms are found and term-specific operations (parent/child/other relationships retrieval, etc.) are needed, the getOntology method is used to support relationships.

3 TECHNICAL DETAILS

The package is based primarily on the Ontology Common API Tasks Java library, on the OWLAPI and depends on the BioPortal, OLS: list LoadedOntologies and list OLS Ontologies.

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/

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


