BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 5 2014, pages 740–742
doi:10.1093/bioinformatics/btt581

Databases and ontologies Advance Access publication October 9, 2013

The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies

Sébastien Harispe*, Sylvie Ranwez, Stefan Janaqi and Jacky Montmain
LGI2P/EMA Research Centre, Site EERIE, Parc Scientifique G. Besse, 30035 Nîmes cedex 1, France

Associate Editor: Martin Bishop

ABSTRACT

Summary: The semantic measures library and toolkit are robust open-source and easy to use software solutions dedicated to semantic measures. They can be used for large-scale computations and analyses of semantic similarities between terms/concepts defined in terminologies and ontologies. The comparison of entities (e.g. genes) annotated by concepts is also supported. A large collection of measures is available. Not limited to a specific application context, the library and the toolkit can be used with various controlled vocabularies and ontology specifications (e.g. Open Biomedical Ontology, Resource Description Framework). The project targets both designers and practitioners of semantic measures providing a JAVA library, as well as a command-line tool that can be used on personal computers or computer clusters.

Availability and implementation: Downloads, documentation, tutorials, evaluation and support are available at http://www.semantic-measures-library.org.

Contact: harispe.sebastien@gmail.com

Received on July 30, 2013; revised on September 23, 2013; accepted on October 4, 2013

1 INTRODUCTION

Biomedical ontologies provide well-structured and controlled vocabularies of specific domains, e.g. biological processes and clinical healthcare terminology. They are increasingly used to drive data integration, information retrieval, data annotations and decision support, to cite a few (Stevens *et al*., 2000).Open repositories such as the Open Biomedical Ontology (OBO) Foundry or BioPortal (Smith *et al*., 2007; Whetzel *et al*., 2011) provide access to hundreds of biomedical ontologies expressed in various formats, e.g. Resource Description Framework (RDF), OBO, Web Ontology Language (OWL). These structured vocabularies are used to characterize entities through conceptual annotations. For instance, genes (products) can be annotated by Gene Ontology (GO) terms to define their molecular functions, their cellular locations or the biological processes in which they are involved (Ashburner *et al*., 2000). Those unambiguous annotations can, therefore, be used to query large collections of data taking into account the knowledge defined in the ontology, i.e. practitioners searching for genes annotated to 'nucleoside binding' will also retrieve genes annotated to 'ATP binding', as the ontology specifies that 'ATP binding' is a specific type of 'nucleoside binding'. However, in some cases, exact searches are too constraining, and we search for entities that are similar or related to the query. Such an imprecise search is based on information retrieval techniques that require a function to estimate whether or not two entities are similar or related with regards to their conceptual annotations. Therefore, to exploit ontologies and corresponding annotations, semantic measures are required. They aim to compare concepts by taking into account the semantic space in which they are defined. They can, therefore, be used to assess the degree of likeness of concepts defined in ontologies or between entities annotated by those concepts (Pesquita *et al*., 2009).

An increasing number of algorithms rely on semantic measures, for instance to analyze genes based on their molecular functions (Sy *et al*., 2012) or related diseases (Li *et al*., 2011). Semantic measures can also assist in comparisons of patient records, chemical compounds, diseases or any entity that can be characterized by unambiguous terms or concepts defined in ontologies or thesauri.

Numerous communities are involved in the study of semantic measures (e.g. bioinformatics, Natural Language Processing, artificial intelligence and Semantic Web). Owing to their popularity, many measures have been designed for different ontologies and treatments (e.g. gene analysis, information retrieval): a recent survey distinguished tens of measures dedicated to the GO alone (Guzzi *et al*., 2012). However, communities focusing on other types of annotated entities (e.g. patient records) also benefit theoretical findings made by studying measures in other specific domains such as molecular biology and *vice versa*. Nevertheless, most software solutions related to semantic measures are developed for a specific terminology/ontology and only focus on a limited set of measures (Fröhlich *et al*., 2007; Li *et al*., 2011; McInnes *et al*., 2009; Yu *et al*., 2010). To federate efforts related to the design and analysis of semantic measures and to respond to the need for a generic software tool dedicated to them, we developed the semantic measures library (SML). This article presents its benefits for the computation of semantic measures using bio-ontologies.

2 THE SML AND TOOLKIT

The SML is an extensive, efficient and generic open-source library dedicated to the computation, development and analysis of semantic measures. Numerous functionalities provided by the SML are also available within the SML-Toolkit, a command-line program that can be used by non-developers to easily compute semantic measures on personal computers or computer clusters. The SML and the toolkit are distributed under the open-source

*To whom correspondence should be addressed.

CeCILL license (compatible with the widely used GNU General Public License).

The SML uses cross-platform JAVA programing language version 1.7, which is available for most operating systems. It can be used to compute semantic similarities of concepts/terms defined in structured terminologies and ontologies. It can also be used to assess the semantic similarity of pairs of entities annotated by concepts, e.g. patient records annotated by groups of concepts, genes annotated by GO terms, PubMed articles annotated by MeSH descriptors. Considering a pair of terms/entities, the library computes a similarity score. Developers can, therefore, easily embed source code referring to the library to compute measures in their own algorithms and applications.

The library supports various ontology formats and specifications (e.g. OBO, RDF, OWL). Specific ontology loaders are also provided to handle widely used biomedical terminologies such as MeSH and SNOMED Clinical Terms (SNOMED CT). Custom knowledge representation loaders can also be added to the SML. In addition, low-level access to the library enables developers to finely control the underlying graph model (ontology) to apply specific treatments sometimes required for the computation of semantic measures (e.g. transitive reduction to remove taxonomical redundancies).

A large collection of semantic measures is provided out-of-the-box—version 0.7 supports about 50 measures relying on different strategies. Thanks to the fine-grained control provided by the library, this leads to about 1500 specific measure configurations that can be specified for context-specific applications. In addition, the algorithms developed in the SML provide the designers of semantic measures an extensive Application Programming Interface and framework to easily develop, test and evaluate new measures. Moreover, because of its generic underlying graph data model, semantic measures developed using the SML will benefit a large audience. Those measures are not restricted to a specific ontology, which is the case with existing software solutions, and can, therefore, be used with the various knowledge representations supported by the library. Furthermore, the SML relies on a graph model compatible with the Linked Data paradigm. This enables SML users to take advantage of the growing number of datasets published according to Linked Data and Semantic Web visions, e.g. see Bio2RDF initiative (Belleau *et al*., 2008).

The SML enables large-scale computations and analyses of semantic measures. It supports multi-threaded processes for fast parallel computation on multicore processors. Table 1 presents a running time comparison between three libraries dedicated to the GO and the SML (detailed protocol, associated source code and additional evaluations are provided at http://www.semantic-measures-library.com/sml/performance).

Based on the SML, an open-source toolkit enables non-developers to benefit from functionalities provided by the library through easy to use command-line software. The SML-Toolkit is highly tuneable and enables context-specific configurations to be specified depending on the experiment performed: knowledge base to use (ontologies, annotations), required data preprocessing (e.g. the removal of taxonomic redundancies), measure constraints (e.g. algorithmic complexity, information to take into account), set of queries to perform (i.e. concept or entity identifiers) and other (optional) parameters (e.g. output file, computer resources allocated). Detailed configurations can be specified using an extensible mark-up language file. Specific command-line interfaces, called *profiles*, are also developed to ease the use of the SML-Toolkit in specific use cases, e.g. to estimate the similarity of genes regarding their GO term annotations. Such profiles can be used to hide the advanced capabilities of the library, and therefore improve the experience for users interested only in computing semantic measures in a specific context of use (e.g. gene or disease analysis). Related source code and issue trackers are available from the public dedicated repository. Community support is also provided to facilitate usage and ensure improvements of both the library and the toolkit.

Table 1. Running times of the generic SML and three tools dedicated to the GO

Tools	1K	10K	1M	100M
FastSemSim[a]	0m13.36	0m16.79	7m8.14	X
GOSim	X	X	X	X
GOSemSim	27m02.66	X	X	X
SML	0m10.01	0m11.18	1m38.87	133m27.44
SML parallel	0m9.80	0m10.24	0m47.62	58m

[a] http://sourceforge.net/projects/fastsemsim/.

Note: Four tests have been performed considering random samples of gene pairs with fixed sizes (see columns, K = 103, M = 106). SML parallel corresponds to the SML configured with four threads. 'X' specifies that the process required >6Go of RAM or took >4 h.

Open source, generic, efficient and highly tuneable, the SML and the toolkit are not limited to a specific ontology and can, therefore, be used in a broad field of application, (scientific) projects and software solutions [e.g. Harispe *et al*. (2013), Sy *et al*. (2012)].

ACKNOWLEDGEMENT

The authors would like to thank SML users and developers for their contributions to the project.

Funding: French Life Sciences and Healthcare Alliance (AVIESAN).

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Belleau,F. *et al*. (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inf*., **41**, 706–716.

Fröhlich,H. *et al*. (2007) GOSim–an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics*, **8**, 166.

Guzzi,P.H. *et al*. (2012) Semantic similarity analysis of protein data: assessment with biological features and issues. *Briefings Bioinformatics*, **13**, 569–585.

Harispe,S. *et al*. (2013) Semantic measures based on RDF projections: application to content-based recommendation systems. In: *On the Move to Meaningful Internet Systems: OTM 2013 Conferences*. Springer Berlin Heidelberg, Graz (Austria), pp. 606–615.

Li,J. *et al*. (2011) DOSim: An R package for similarity between diseases based on Disease Ontology. *BMC Bioinformatics*, **12**, 266.

McInnes,B.T. *et al.* (2009) UMLS-Interface and UMLS-Similarity: open source software for measuring paths and semantic similarity. *AMIA Annu. Symp. Proc.*, **2009**, 431–435.

Pesquita,C. *et al.* (2009) Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.*, **5**, 12.

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

Stevens,R. *et al.* (2000) Ontology-based knowledge representation for bioinformatics. *Briefings Bioinformatics*, **1**, 398–414.

Sy,M.-F. *et al.* (2012) User centered and ontology based information retrieval system for life sciences. *BMC Bioinformatics*, **13** (**Suppl. 1**), S4.

Whetzel,P.L. *et al.* (2011) BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.*, **39**, W541–W545.

Yu,G. *et al.* (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, **26**, 976–978.