Databases and ontologies

The Semantic Measures Library and Toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies

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

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. OBO, RDF). The project targets both designers and practitioners of semantic measures providing a JAVA library, as well as a command-line tool which can be used on personal computers or computer clusters.
Availability: Downloads, documentation, tutorials, evaluation and support are available at http://www.semantic-measures-library.org.
Contact: harispe.sebastien@gmail.com

1 INTRODUCTION
Biomedical ontologies provide well-structured and controlled vocabularies of specific domains, e.g. biological processes, 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). Indeed, open repositories such as the 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), Open Biomedical Ontology (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 in order 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 which are similar or related to the query. Such an imprecise search is based on information retrieval techniques which 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 analyse 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 which 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). Due 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; Yu et al., 2010; Li et al., 2011; McInnes et al., 2009). 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. This paper presents its benefits for the computation of semantic measures using bio-ontologies.

2 THE SEMANTIC MEASURES LIBRARY AND TOOLKIT
The Semantic Measures Library (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 programme which can be used by non-developers to easily compute semantic measures on personal computers or computer clusters. The SML and the toolkit are distribut-

*To whom correspondence should be addressed.
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-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) in order 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 (API) and framework to easily develop, test and evaluate new measures. Moreover, due to its generic underlying graph data model, semantic measures developed using the SML will benefit a large audience. Indeed, 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 multi-core processors. Table 1 presents a running time comparison between 3 libraries dedicated to the Gene Ontology and the SML (detailed protocol, associated source code and additional evaluations are provided at http://www.semantic-measures-library.com/sml/performance).

<table>
<thead>
<tr>
<th></th>
<th>1K</th>
<th>10K</th>
<th>1M</th>
<th>100M</th>
</tr>
</thead>
<tbody>
<tr>
<td>FastSemSim</td>
<td>0m13.36</td>
<td>0m16.79</td>
<td>7m8.14</td>
<td>X</td>
</tr>
<tr>
<td>GOsim</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>GOSemSim</td>
<td>27m02.66</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SML</td>
<td>0m10.01</td>
<td>0m11.18</td>
<td>1m38.87</td>
<td>133m27.44</td>
</tr>
<tr>
<td>SML parallel</td>
<td>0m9.80</td>
<td>0m10.24</td>
<td>0m47.62</td>
<td>58m</td>
</tr>
</tbody>
</table>

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 pre-processing (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 XML 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.

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. This work was partly supported by the French Life Sciences and Healthcare Alliance (AVIESAN).

REFERENCES

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

1 http://sourceforge.net/projects/fastsemsim/