Poster: Biological Databases – BD8

FastSemSim: fast and easy evaluation of semantic similarity measures on biomedical ontologies

Mina M(1), Sanavia T(2)

(1) MPBA, Fondazione Bruno Kessler, Trento (2) Department of Information Engineering, University of Padova, Padova

Motivation: Although biological databases addressed the problem of providing consistent and formal descriptions of genes and proteins using different data structures, the quantification of the functional similarity between genes or proteins exploiting these data is still a challenge. Several semantic similarity measures (almost 50) have been proposed to compare either the terms from biomedical ontologies or the genes/proteins annotated with them, using different approaches to consider the relationships among the annotations (Guzzi et al., 2012). Their implementations can be distinguished in two categories: on-line and off-line. For on-line analyses, G-SESAME (Zhidian et al., 2009), FunSimMat (Schlicker and Albrecht, 2010 and ProteInOn (Pesquita et al., 2009) are the most recently proposed and provide, beyond the classical information-theoretic based metrics, new approaches based on term specificity. On the other hand, almost all off-line tools are developed in R language: GOSemSim (Yu et al., 2010) and csbl.go (Ovaska et al., 2008), are among the most used. However, current applications are not specifically designed to manage huge amounts of data in order to support genome/proteome-wide analyses. Recently, new integrative approaches in the analysis of high-throughput data has proven that the integration of prior knowledge from biomedical ontologies is a useful resource to improve the identification of expression patterns (Di Camillo et al., 2012) and to provide more stable biomarker lists in classification problems (Sanavia et al., 2012). Therefore, more efforts are required to develop applications which are able to be both scalable across genome/proteome-wide data and enough flexible to provide a user-friendly platform to calculate these similarities and to integrate them within new computational pipelines.

Methods: FastSemSim is both a Python library and an end-user application, featuring an intuitive graphical user interface (GUI). As input data, the library requires the ontology graph and the gene/protein annotations. The current version of FastSemSim handles both obo and obo-xml daily updated files and supports any multi-rooted ontology (as long as it is acyclic). The library was implemented with a modular architecture: a core component includes all the routines for parsing the ontology and the annotations to extract common features for the measures (e.g. Information Content), whereas all the semantic similarity measures were developed on top of the core library as independent modules. The library currently supports 16 different measures, both pairwise and groupwise. While groupwise measures can directly evaluate gene/protein similarities considering the corresponding sets of terms, pairwise measures need a “mixing strategy” (Guzzi et al., 2012). The three most used strategies were implemented: the average (avg) and the maximum (max) of all term pairwise similarities, and the average of similarities between best matching terms (Best Match Average). FastSemSim supports inter-ontology and inter-category (e.g. Molecular Functions and Biological Process in Gene Ontology) relationships, and provides several filtering functions which allow the user to perform organism-specific analysis or to work only with specific evidence codes. In addition, a GUI is provided to easily calculate the similarity measures, characterized by a user-friendly front-end to load the ontology and the annotation files, to input the query and to select the output parameters. Both the library and the interface are compatible with Python 2.x and were tested on Microsoft Windows, OS X and different Linux distributions.

Results: Compared with the most used off-line and on-line available tools for semantic similarities, FastSemSim shows the highest coverage of implemented semantic similarity measures, enabling the systematic evaluation of different measures (Cho et al., 2013). Scalability was tested on Gene Ontology (GO) annotations for the categories Biological Process and Molecular Function across the proteomes of several organisms (Human, Mouse, Fly and Yeast). Resnik measure (Resnik, 1999) combined with the max mixing strategy, conventionally used for protein-based studies, was applied. FastSemSim was able to accomplish the analysis for all the proteomes, ranging between 2 minutes for the Yeast proteome (6380 GO annotated proteins) and 5 hours and 18 minutes for Human proteome (45576 GO annotated proteins). Available R applications do not provide an efficient implementation able to deal with more than 1000 genes/proteins efficiently. FastSemSim proved to meet the requirements of handling huge amounts of data. Moreover, the Python implementation and the modular architecture of the library can be easily exploited to both integrate semantic similarity within computational pipelines and extend the library with new measures.

Contact email: mamina@fbk.eu