Automated analytics for the evolution of the Gene Ontology

Miguel-Angel Sicilia

Ms.C. in Bioinformatics and Biostatistics

Consultor: Alexandre Sánchez Pla

February 2017

FICHA DEL TRABAJO FINAL

Title:	Automated analytics for the evolution of the Gene Ontology
Author:	Miguel-Angel Sicilia
Nombre del consultor:	Alexandre Sánchez Pla
Fecha de entrega (mm/aaaa):	01/2017
Area:	Ad-hoc
Program:	*Ms.C. in Bioinformatics and Biostatistics*

Resumen del Trabajo:

La Gene Ontology (GO) es un recurso clave en la investigación bioinformática, dado que el conocimiento biológico que incorpora se utiliza en diversas herramientas de amplia difusión que comparan o clasifican productos genéticos. Por tanto, los hallazgos que se incluyen en muchos informes de investigación están soportados por los resultados de diferentes herramientas basadas en la GO.

No obstante lo anterior, la GO no es un recurso estático sino un proyecto en constante evolución, y esto suscita la cuestión del grado en que las conclusiones extraídas de una versión de la ontología son aún válidas cuando esa estructura cambia, bien en el nivel de la propia ontología y/o en el de la gran base de datos de anotaciones asociadas a la misma. Esa cuestión se hace aún más acuciante por la razón de que las versiones concretas de la GO o las herramientas utilizadas no se proporcionan en todos los estudios publicados, aun debiendo ser esto un requisito desde la perspectiva del paradigma de la investigación reproducible.

El trabajo que aquí se presenta describe el diseño y desarrollo de una herramienta que permite evaluar cambios entre las versione de la GO, permitiendo dar un primer paso hacia una comprensión más profunda de la evolución de la GO y su impacto en las conclusiones soportadas por su uso. Se describe el diseño y uso de la biblioteca `pygoa`, y se acompaña de un análisis de versiones mensuales de la GO. El análisis se hace a nivel de la ontología (análisis terminológico), incluyendo métricas de ontologías, y también al nivel de la base de datos de anotaciones, e incluye la evolución de ciertas medidas de similaridad.

Los resultados muestran que la GO es un recurso en evolución, con un crecimiento global dominado por la subontología de procesos biológicos. La base de datos de anotaciones amalgama las contribuciones de diferentes subproyectos con patrones cambiantes y se basa fundamentalmente en anotaciones obtenidas de forma automática. Se ha encontrado que el cambio en la estructura de la red de relaciones entre versiones tiene un impacto solo en una pequeña proporción de las métricas de contenido de información, que a su vez impactan en métricas de similaridad. Estos hallazgos apuntan a la necesidad de una exploración futura acerca del grado en que esos cambios podrían afectar a la salida de las herramientas basadas en la GO cuando utilizan diferentes versiones que a su vez podrían impactar en las conclusiones biológicas obtenidas de los estudios que las utilizan.

Abstract (in English, 250 words or less):

The Gene Ontology (GO) is a key resource in bioinformatics research as its embedded biological knowledge is used as the basis of various tools that are broadly used in comparing or classifying gene products. Consequently, the findings reported in many research are supported by the outcomes of different GO-based tools.

However, the GO is not a static resource but an evolving project, and this raises the question on the extent to which conclusions drawn from a version of the ontology are still valid as it changes structure, both in the ontology itself and/or in the large database of associated annotations. That question becomes even more relevant since the concrete versions of the GO or the tools used are not reported in all the studies published, which should be required from the perspective of the reproducible research paradigm.

The work presented here reports on the design and development of a tool that provides the means to assess changes in the GO across versions, which represents a first step towards a deeper understanding of the evolution of the GO and its impact in the conclusions supported by its use. The design and use of the `pygoa` library is reported along with an analysis across monthly snapshots of the GO. The analysis is done both at the level of the ontology (terminological database), including ontology metrics, and at the level of the annotation database, including the evolution of several similarity measures.

Results show that the GO is an evolving resource, with an overall growth driven by the biological process subontology. The associations database amalgamates inputs from several contributors with changing patterns and relies fundamentally on computationally derived annotations. Changes in the network structure of relations across versions have been found to impact on a small fraction of information content metrics, and subsequently on similarity metrics. These deserve further exploration to assess the extent to which these changes may affect the output of GO tools using different versions of the ontology that might in turn affect the biological conclusions found.

Keywords (entre 4 y 8):

Gene ontology, metrics, similarity, ontology evolution.

Índice

Figure index

1. Introduction

1.1 Motivation

The Gene Ontology (GO) is a collaborative project that attempts to address the need for consistent descriptions of gene products across different databases. The project was founded in 1998, as a collaboration between three model organism databases, FlyBase (Drosophila), the Saccharomyces Genome Database (SGD) and the Mouse Genome Database (MGD), and has grown to a much larger coverage of organizations called the Gene Ontology Consortium (GOC).

Typically, studies report the use of concrete GO tools, that in turn internally use the GO itself or its related annotation databases. However, the GO is an evolving resource, and reporting the use of a tool in some cases does not give full information on the version of the GO used, or it is difficult to trace and find out.

The GO is not formally subject to release control, at least as it is made available in its Web site and associated repositories. Nonetheless, different timestamped versions or snapshots of the GO can be considered by comparing the different historical copies (typically made available monthly) or if required, by inspecting its CVS repository.

The above described situation has an impact in both the reproducibility of studies and the intrinsically provisional nature of conclusions drawn from GO tools. Particularly, there are several levels or dimensions that need to be accounted for when assessing changes in the GO:

I. *Terminological level*. The biological knowledge incorporated in the ontology itself changes with time, by addition or deprecation of terms or re-structuring of relations resulting from those changes.
II. *Annotation level*. The database of annotations is constantly growing. In addition, the evidence levels associated with the annotations change as part of the manual curation process (du Plessis, Škunca, & Dessimoz, 2011).

Since the GO is an ever-evolving resource, there is a need to understand and analyse the potential impact of its evolution in the findings and conclusions derived from tools that use them internally. This evolution may put into question the validity of conclusions drawn from previous versions of the ontology. For example, changes in the ontology structure affect topological similarity measures used to compute similarities, and those similarities in turn are may be used to

assess, for example, the biological processes affected by an experimental condition in a high throughput study. A second concern is that the evolution of the GO makes the resource richer in some sense, and this opens new opportunities for attaining new insights from the same data, simply by using updated versions of the GO.

Both concerns are the departure assumptions for the work presented here. However, the assessment of the impact of the GO evolution in concrete studies is out of the reach of the current work, as it would entail the re-execution of the pre-processing, analysis and interpretation workflow of concrete studies, and the current level of automation of reproducible studies is not consistent and lacks platforms for its execution. This entails that reproducing a study entails mostly manual work due to several missing resources in current studies, including:

- Concrete information on versions of tools used, including common libraries and statistical tools (e.g. R libraries for statistical analysis) but also GO-specific tools.
- Detailed information on configuration parameters used in the workflow of the study.
- Lack of executable workflow artefacts, that would ease reproducibility.

The above makes difficult the actual assessment of the impact of GO evolution in concrete studies, but it is still possible to get insights and estimates of that potential impact, as a first step towards a future practice of re-execution of studies that only change the resources of the GO and automatically contrasts the provisional conclusions published with potentially new or updated insights. The impact of that future practice is large, from two perspectives:

- Re-evaluating past studies as a form to productively search for new biological conclusions.
- Changing the practices and habits of the use of the GO, so that re-execution of studies can be triggered by changes in measures related to the GO that are considered as potentially relevant in that they may change the results provided by GO tools.

While the actual impact of the evolution of the GO in the conclusions of studies is out of the scope of this work, here we report on a tool that is necessary towards that end. Concretely, this thesis reports the design and implementation of a Python library named `pygoa` that allows for the comparison of GO versions, and provides metrics and analytic functions to get insights on the potential impact of the evolution of the GO. The main aim of that library is that of easing and automating the analytics of GO versions, covering the first step towards the abovementioned long-term goals.

1.2 Objectives

The overall aims of the thesis are the following:

- O1: Design and develop a library that allows extracting features and metrics from the GO that can be used from broadly used data science stacks.
- O2: Report on an analysis of those features and metrics across versions of the GO, analyzing its temporal evolution, trends and relation with known problems of the GO.

These overall aims are further specified in the following specific objectives.

- O1.1. Design and develop a framework to extract features and metrics from the GO compatible with SciPy, the scientific stack built around the Python ecosystem.
- O1.2. Design and develop tools for the analysis of internal relationships inside the GO.
- O2.1. Develop software for the analysis of GO versions along time including its annotation database.
- O2.2. Evaluating potential known GO problems according to the analysis done, as for example:
    - the impact of relations across GO sub-ontologies.
    - potential biases related with evidence codes or other annotation related features.
    - differences across annotation databases (different projects).

The two major languages and projects for data science that are built around open source communities are currently those of R and Python. Python was selected for the following reasons:

- having available more mature ontology-processing libraries.
- having better direct integration with medium and big-data frameworks than R.
- possibility of direct and seamless use on top of hosted data science frameworks for bioinformatics as Galaxy[1].

Nonetheless, these are practical reasons and the library could have also be implemented on top of a R stack of libraries.

1.3 Approach and methods

The main outcome of the work reported is that of the software library and the associated analysis done with it. In consequence, the main steps for reaching

[1] https://galaxyproject.org/

that goal starts with an iterative approach to developing the library that drives the rest of the process.

The main steps can be summarized as follows:

- Study the structure, format and sharing means of a part of the GO.
- Design and develop the base software for getting that data.
- Reviewing the metrics, features or analytic tasks for that part that are potentially useful or applicable.
- Including those metrics or features in the library.
- Testing and analytics phase using the features implemented.
-

The selected platfrom is the Python scientific stack (SciPy[2]). Other GO Access libraries have been considered as candidates for ideas on how to structure the API. Concretely, BioPython was discarded as the GO access libraries do not seem to be in active development. Other libraries used in common GO tools are task-specific and do not support the retrieval of GO snapshots directly, so that they have also been discarded.

The library has been made openly available under a MIT license to maximize possibilities of reuse and extension, which is critical for a longer-term impact of the kind of analysis done.

1.4 Outcomes

The main outcome of the project is the `pygoa` library itself. It is shared as open source under a permissive MIT license and can be found here:

https://github.com/msicilia/pygoa

The results of the analysis of versions across time that is reported in this document can be reproduced using the library. Notebooks are provided as examples in the Github repo.

1.5 Structure of this document

The rest of this document is structured as follows.

- Chapter 2, **Background,** provides background information on the GO, its updating processes and how it is shared as an open resource.
- Chapter 3, **Data acquisition**, describes the overall design principles for the library developed, discussing the strategies adopted for dealing with large files.

---

[2] https://www.scipy.org/

- Chapter 4, **Ontology evolution**, reports on the overall growth pattern of the GO terminology and then discusses findings related to ontology metrics. Then, it reports on the exploration of the use of graph models and the use of the clustering coefficient as an initial account of structural change in the ontology across versions. Finally, it reports on the study of changes in terminology-based information content and similarity measures.
- Chapter 5, **Annotation database evolution**, discusses annotation-based information content metrics across time, complementing the analysis in the previous chapter. Then, it discussed findings on the distribution of the contributions of different sub-projects to the annotation database, and patterns in evidence codes.
- Chapter 6, **Conclusions and outlook**, summarize the main findings and assesses the results with respect to the original aims. It also provides a short discussion on potential extensions to the work presented in the rest of the document.
- Finally, references and glossary sections are provided at the end od the document.

Code artefacts, including tets and examples, are provided in the associated Github repository (https://github.com/msicilia/pygoa), so that this document does not include the details on the interfaces and design of the library that are more likely to change in the future evolution of the code can be founde there.

2. Background

2.1. The Gene Ontology as the result of a process

The GO project develops and curates three structured ontologies that describe gene products in terms of their associated biological processes, cellular components and molecular functions. However, the actual work done encompasses three different sets of activities:

1. the development and maintenance of the ontologies themselves.
2. the annotation of gene products, i.e. recording associations between ontology elements and the genes and gene products in the collaborating databases
3. the development of tools that facilitate the creation, maintenance and use of ontologies.

While many other ontology engineering efforts cover both (1) and (2), annotation of non-ontology resources is not typically part of them.

It should be noted that from the perspective of the GO consortium, the ontology itself (not the annotations) is considered as a common query and drill-down language across a system of decentralize databases.

Groups or organizations external to the GOC can contribute to the GO by providing either proposals for changes in the ontology or annotations. All these are reviewed by editors of the ontology.

2.2. Annotations and evidence in the GO

Annotations are included in the GO by aggregating inputs from the different organizations or projects contributing to the GO[3] as members of the GO Consortium. However, research groups in general may also contribute to the GO,

[3] http://www.geneontology.org/page/go-consortium-contributors-list

as documented in the contribution pages of the GO[4]. Edits proposed are reviewed by ontology editors and implemented where appropriate.

The process of annotating results in the assignment of GO terms to gene products. It follows a number of policies and guidelines[5] aimed at attaining homogeneous results. The assignment of a term need to be classified with an evidence code that records the criterion or method used. Out of all the evidence codes available, only Inferred from Electronic Annotation (IEA) is not assigned by a curator but by automated means. Manually-assigned evidence codes fall into four general categories: experimental, computational analysis, author statements, and curatorial statements. Obviously, IEA annotations should be given different consideration to manual ones, as they are subject to a different kind of potential error. However, the evidence code is not in general an statement of quality, since it is possible to have high quality IEA annotation and maybe for example some bad quality annotations originated from other sources of uncertainty.

2.3. Similarity of terms in the GO

The GO is used as a source of biological knowledge in different kind of studies that involve identifying relevant genes or gene products in the results of experiments. This requires some form of measuring the semantic relations among GO elements based on the structure of the ontology and its annotations.

This has led to several proposals for semantic similarity measures that have been included in different kind of tools, resulting in a number of related and overlapping efforts. These tools and proposals have been recently surveyed and classified by Mazandu, Chimusa and Mudler (2016) to provide a guidance on the current state of the art for users and researchers. That survey in turn is based on previous review studies by Pesquita et al. (2009) and Guzzi et al. (2012) so that we take it as the most comprehensive to date. The approach presented in that article is that of classifying similarity approaches based on the strategies used to compute the scores. This results in a classification in three groups of measures:

- Information Content (IC)
- Term semantic similarity
- Functional similarity.

Measures of IC are typically used as part of the computation of semantic similarity scores, and functional similarity builds on both.

---

[4] http://www.geneontology.org/page/contributing-go

[5] http://www.geneontology.org/page/annotation

2.4. GO tools

There are a number of tools that are developed and maintained by the GOC, and many others that are external.

2.5. Versions and usage of the GO in the tools

The GO is not formally subject to release control. However, there are several ways of acquiring old versions of the ontology and its annotations.

The most recent version of the ontology can be downloaded from direct links in the download page[6], in OBO and OWL formats, and with two variants:

- filtered or basic version, that omits the relationship that cross the three subontologies.
- fully axiomatized version (in OWL), including import of additional external ontologies.

Application or organism-specific subsets ("slims") are also provided as direct downloads.

When looking for older versions, the download page provides "non-recommended" legacy downloads, that are maintained to support tools using those older versions.

Older versions can be accessed via two resources:

- An FTP archive of ontology file snapshots, deposited monthly[7].
- A Concurrent Version System (CVS) repository (including a Web interface for recent files, updated every thirty minutes).

The CVS interface can be used to retrieve all revisions made by editors, often with day-to-day changes.

Annotations are provided separately, with the most recent versions directly accessible via the downloads page. For older versions, the help system mentions several formats:

1. Database dumps, in relational format, in the archive site[8].
2. CVS and SVN repositories for individual gene association files.

---

[6] http://www.geneontology.org/page/download-ontology

[7] ftp://ftp.geneontology.org/go/ontology-archive/

[8] http://archive.geneontology.org/full/

3. FTP archive of snapshots of databases, provided by EBI[9] and directly from GO pages[10].

In the FTP archive, the database is provided as three different resources:

- Termdb: containing only the information on the GO terms and relationships.

- Assocdb: containing both the GO vocabulary and associations between GO terms and gene products. This database is a superset of termdb.

- Seqdb: containing the two above, plus the sequences associated with the annotated gene products.

Association information is provided in MySQL dump or SQL sentences formats. Also, the Termdb is provided in OBO and OWL formats.

---

[9] ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/old/

[10] ftp://ftp.geneontology.org/go/godatabase/archive/

3. Data acquisition

3.1. Data sources

From all the possible sources of copies of the terminological part of the GO, we have selected the FTP archive as a source. This is due to several practical reasons:

- The level of time granularity is sufficient for the kind of use cases considered in this work. Monthly updates provide a good account of the changes in GO contents (terms or annotations).
- The space of snapshots or versions considered is equally distributed in time. In the case of versions that can be obtained from the control version system, there is not a periodicity in the updates, which would make harder the analysis.
- Getting the data does not require any library or pre-processing as in the case of using for example CVS interfaces.
- Processing the data in the case of terminology files is straightforward as it is provided in OBO and OWL formats, for which there are parsing libraries available.
- It is possible also to obtain monthly snapshots of the annotation database, matching thus the terminological and annotation view.

In the case of the annotation database, the database format split into tables provided in the FTP archive has been used as source. The other alternative available were SQL dumps, but it was discarded as they require re-building a relational database first, which is a non-necessary step due to the kind of processing done. As explained below, the processing of the annotation database has been done using the Apache Hadoop framework, this is another reason why the flat table format was preferred over a reconstruction of the relational database.

The main limitation of the current implementation is that the snapshots available before April 2004 are currently not processed, since they are not provided in the FTP in a format compatible with the OBO format parsed by the selected libraries (see below). This represents three years of missing data, as the older snapshots available are dated January 2001 (they are also provided in separate files for the three sub-ontologies). Future versions may extend till that date by parsing the older OBO format.

3.2. Selected technologies

The library has been built on top of Python libraries, and particularly, providing the required translation mechanisms to the scientific python (SciPy) stack so that the features and metrics could be further processed using diverse SciPy libraries and interactive analytic environments as Jupyter[11].

3.2. Framework design

3.2.1. Terminological summaries

The main ideas behind the framework design are described in the following diagram, boxes in grey are objects reused from other frameworks. Dashed arroes represent dependencies.

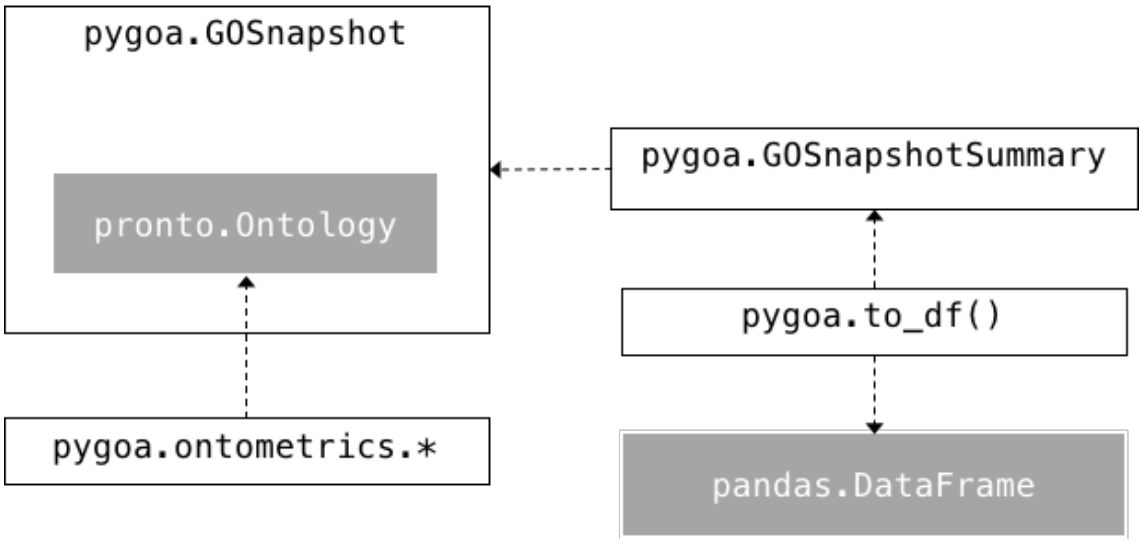


Fig. 3.1. Main design elements of the `pygoa` library for the terminological part of the GO.

Basically, the versions (called "snapshots" as they do not follow a release rationale but are states of the GO at some given points of time) are represented via `GOSnapshot` objects. These essentially extend `Ontology` objects (from the pronto library) with the specifics of getting automatically GO files. These are large objects, which makes impractical having large collections of them simultaneously in memory (e.g. a collection of snapshots across several years). Then, the `GOSnapshotSummary` object is simply a summary of features or counts coming from a snapshot. Summaries are efficient in memory and are thus used to deal with them for statistical purposes across time.

[11] http://jupyter.org/

The utility function `to_df()` convert the summaries into a pandas `DataFrame` object, that can then be used with the SciPy stack as any other dataset, for analytic purposes. This simple machinery allows for moving the needed data to a common analytic format. Libraries extracting metrics or features from snapshots may deal directly with ontology objects, with snapshots, or with collections or summaries. For example, ontology metric functions in the `ontometrics` package deal directly with ontology objects, as they are intended to obtain regular ontology metrics, not specific to the GO.

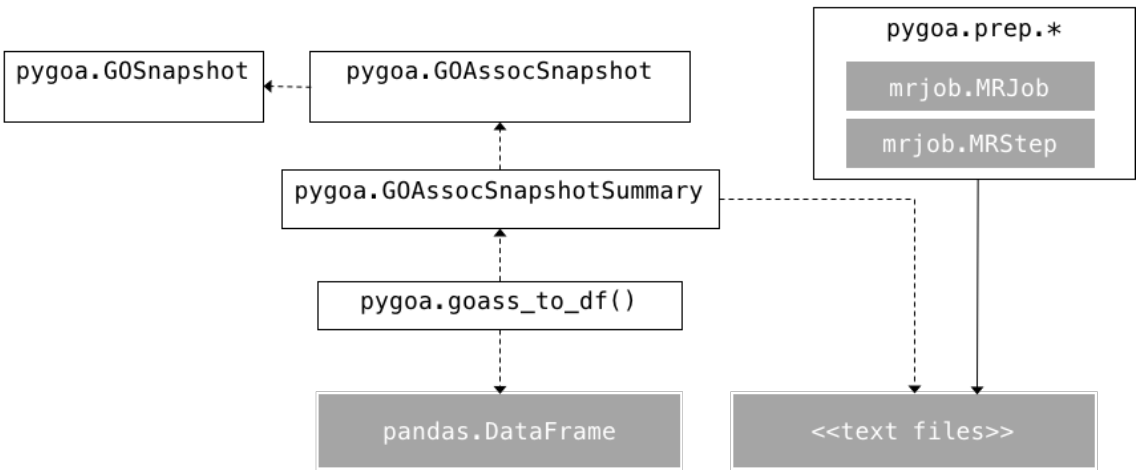
3.2.2. Annotation summaries

The GO annotation database is a large database of matches of gene products with GO terms. Processing full snapshots of this database in memory is not practical, and this is the reason why two methods have been devised for getting information from it, namely:

- Database summaries, which are concise quantitative summaries of the GO that are provided directly as text files in the FTP site of the GO.
- Processing of the full database contents. For data not available in the summaries, the processing is done out-of-core. Concretely, the map-reduce distributed computing framework provided by Apache Hadoop[12] is used to pre-process the data.

The use of Hadoop provides scalability for the current volumes of annotation data and also accommodates its future growth to any scale, as the pre-processing can be distributed over a cluster of computers easily. The drawback is that it requires an extra step to produce intermediate files that are of a size that can be handled in memory.

The following Figure gives an overview of how this is integrated in the `pygoa` library.



[12] http://hadoop.apache.org/

Fig. 3.2. Main design elements of the `pygoa` library for the associations database of the GO.

For consistency with the class design of the terminological part, there are `GOAssocSnapshot` and `GOAssocSnaphotSummary` classes, indicating reference to the database and extracted summary information respectively. However, in the current design, the association database is never brought into memory as it is impractical. As in the previous case, a function `goass_to_df()` is responsible to bridge summary information to the SciPy stack, again in the form of a `DataFrame` object.

The main difference in this case is that all the pre-processing is done separately in modules in the `pygoa.prep` package. This includes utilities to download fragments of the database, but it notably implements the map-reduce tasks on top of the Apache Hadoop framework. The `mrjob` library[13](maintained by Yelp) interfacing Python and Hadoop is used for that task. It should be noted that these processes can be executed equally in a local computer or in a cluster of computers, including commonly used cloud systems as Amazon EMR[14].

3.3. Summary of the extraction

The extraction in the case of the terminological database (non-including annotations) proceeds by simple download and parsing of the entire ontology file. The growth rate of the ontology, as described later, does not appear to pose problems that call for solutions using out-of-core computation.

However, in the case of the annotation database, growth rates follow a different rate and the current volumes require out-of-core computation and some form of parallel processing for the speed up of the acquisition and to guarantee future scalability as the database grows.

In the remainder of this section, the overall measures on the ontology are presented, providing a ground for the rest of the results presented in the following chapters.

3.3.1. Overall ontology evolution

The following Figure depicts the growth of the GO terminology, measured as number of terms (`nterms`) and relations (`nrelations`). The number of terms are including obsolete terms, so that a corrected measure for terms is also provided (`nterms-c`). Deprecations of terms appear to be constant over time and not affecting significantly the growth pattern. It is also apparent that the number of relations grow at a higher rate, as it could be expected since typically the addition of a term results in adding more than one relation, and the

[13] https://pythonhosted.org/mrjob/

[14] https://aws.amazon.com/es/emr/

connectedness of the relations graph if constant would result in a higher growing rate.

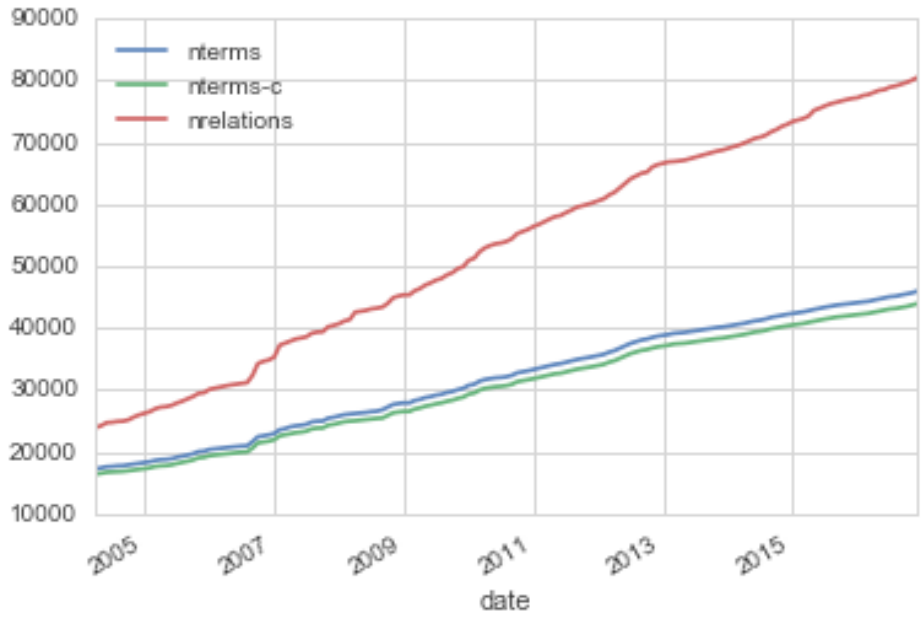


Fig. 3.2. Growth of the GO terminology across time (2004-2016)

It should be noted that there are a few gaps in certain months, for which no ontology snapshot is available. Concretely, January 2012 is missing for the terminology files.

If we look at the different sub-ontologies as depicted in the following plot, we can see that the biological process sub-ontology is accounting for most of the overall grow in number of terms.

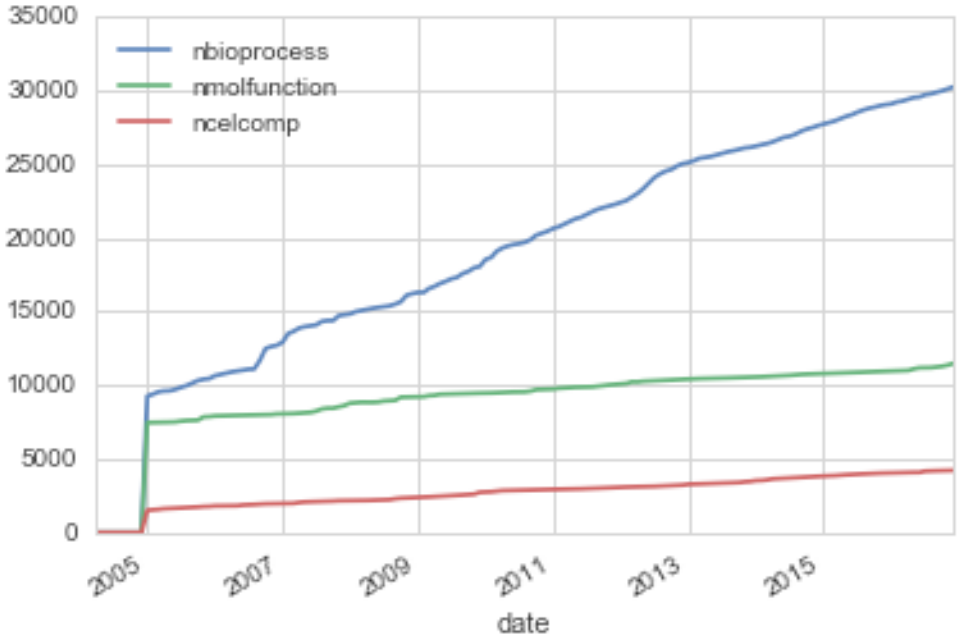


Fig. 3.3. Growth of the GO terminology across time (2004-2016), split by sub-ontology

This is unsurprising as the terminology for cell components and molecular functions may reasonably be expected to be more static, as discovering or refining new of their elements is rather uncommon, especially when compared with the discovery and refinement of biological processes.

3.3.2. Overall annotation database evolution

The following plot depicts the data from annotation summaries in the GO FTP site. Data is not available for several dates, concretely: July, October and December 2004, October 2006, July 2007, February, May and September and November 2008, May, June and July 2009, February and December 2011, August and October 2013, March and June 2014 and January and all months from September 2016. However, the data is still spread enough for a meaningful analysis.

The annotations database grows at a higher rate than the terms in the ontology, which is no surprise, as it encodes cases and not general knowledge. The following Figure depicts the relation of sizes of the annotation database and the number of terms in the ontology (which growth pattern has been discussed above).

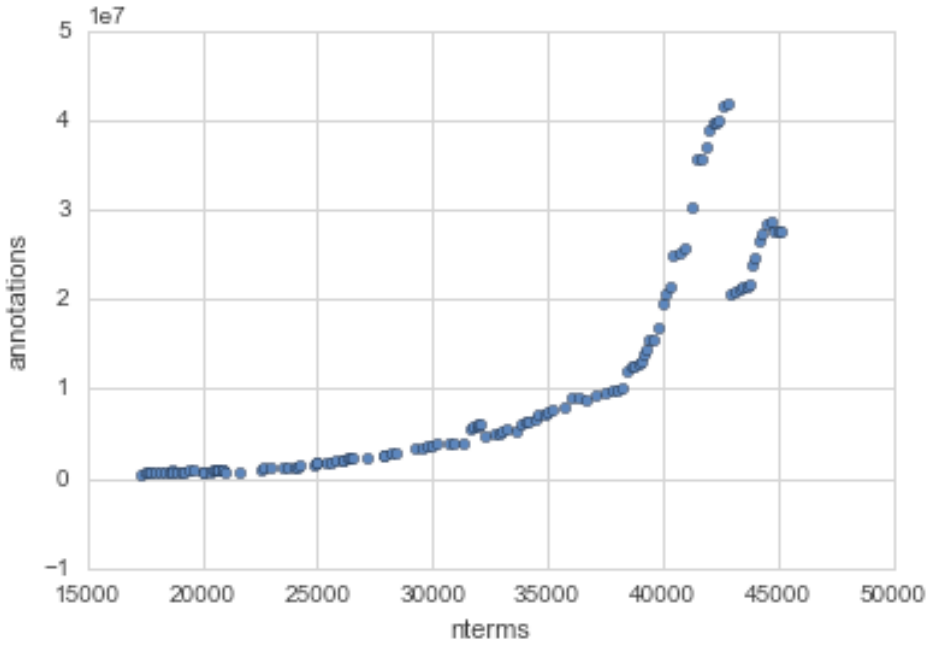


Fig. 3.5. Relative sizes of the GO association database and number of terms across time.

The clear majority of GO annotations correspond to the code IEA (Inferred from Electronic Annotation), which corresponds to automated annotations not involving any form of curatorial intervention. The following Figure provides the percentage of IEA annotations with respect to the overall number. It is overall above 90% of annotations. This is an important fact, since the quality of the annotations thus depend on the algorithmic means to produce them.

Fig. 3.5. Proportion of IEA annotations to the overall number of annotations.

It is also interesting to look at the different categories of annotations other than IEA. The following Figure depicts the number of aggregated experimental annotations, curatorial annotations, annotations taken from computational analysis and those coming from author statements.

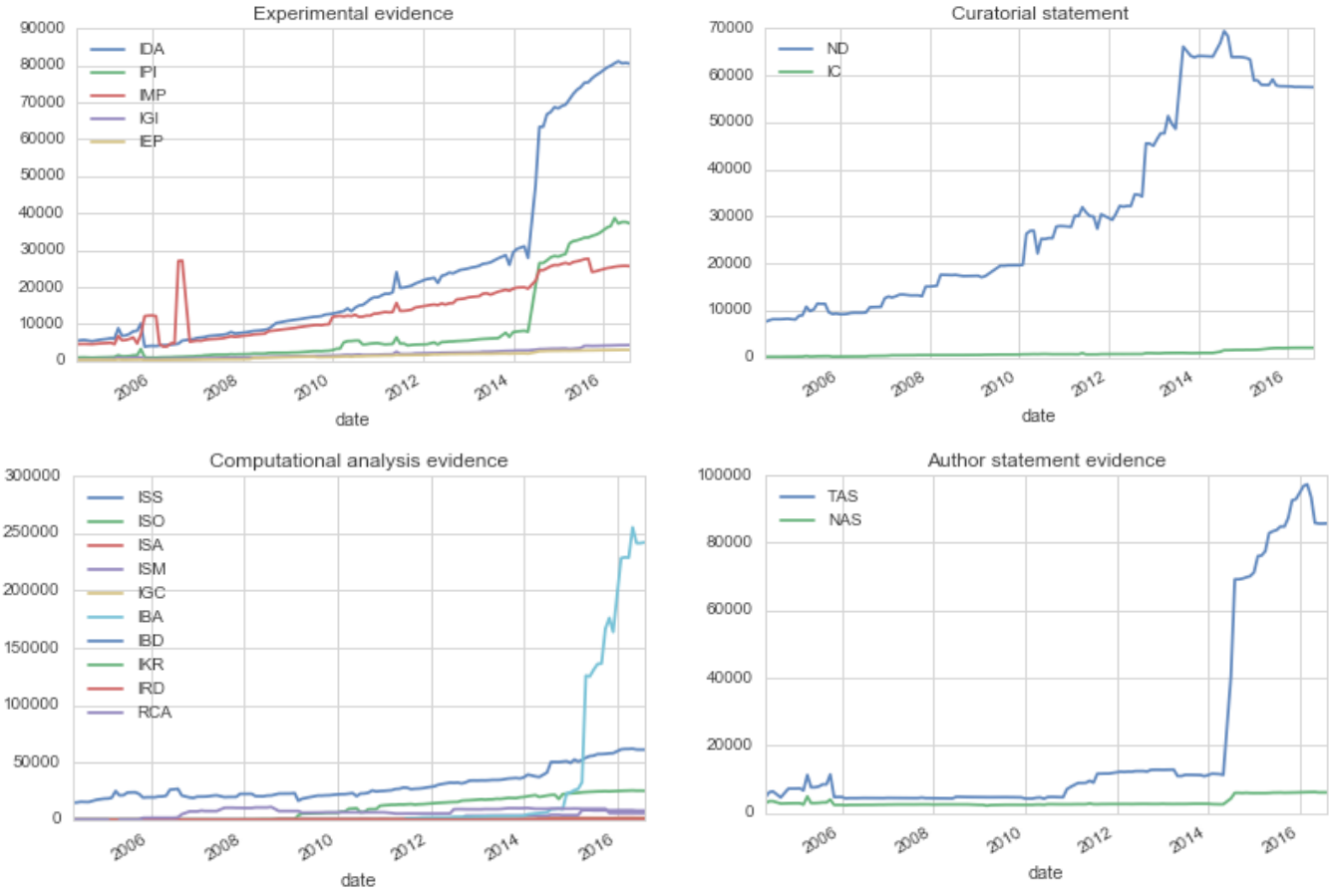


Fig. 3.6. Proportion of annotations by category across time.

As can be evidenced from the plots, there are kinds of evidence in each category that dominate the group. In the case of author statement evidences, Traceable Author Statements (TAS) dominate. This is sensible, as non-traceable author evidence is used when no traceable evidence is found, which is in most cases available. In the case of curatorial statements, most of the annotations are Inferred by Curator (IC) and only a small proportion accounts for “non-biological data available” (ND). This again is a logical consequence of the effort in curating the annotations and having lack of evidence only as exceptional.

The case of computational analysis evidence is interesting, as most of the annotations belong to the “Inferred from Biological aspect of Ancestor (IBA)” code. The second largest group is “Inferred from Sequence or Structural Similarity” (ISS), however this is a generic group, that has three more specific codes: ISA, ISO or ISM, respectively representing inference from alignments, orthologues or sequence modelling methods.

Finally, in the experimental evidence category, the larger group is Inferred from Genetic Interaction (IGI), with Inferred from Physical Interaction (IPI) and Inferred from Mutant Phenotype (IMP) as second and third. Notably, IPI has only in recent years become larger than IMP which may reflect a change in focus of experimental methods.

The discussion so far provides an overall account of the size of the GO and its growth pattern. In the following a more detailed exploration of its features is discussed.

4. Ontology evolution

4.1. Overall metrics

The GO is different to many other application-oriented ontologies in its highly focused scope and its design as a means towards the end of annotating gene products. This entails that not every ontology metric of the many proposed in the literature are relevant or significant, and comparative studies (Sicilia et al., 2012) are not necessarily meaningful.

No instance metrics has been included as the GO does not have instances in a strict sense (no "`Instance`" entries in the OBO files). The annotated gene product may in some non-strict sense be regarded as instances (metrics are provided in the next section), or leaf terms in the hierarchy of relationships could be considered as instances (this can be examined via the "number of leaf classes metric"). However, the latter is controversial, as there is not a strict consideration of what is an instance and a term in the GO. In consequence, no instance metrics have been included.

All the metrics have been examined across time, to find relevant patterns in the evolution of the ontology.

4.1.1. Number of classes and overall shape of the tree

The following Figure depicts the number of classes (`noc`) in contrast with the number of root classes (`norc`) and number of leaf classes (`nolc`).

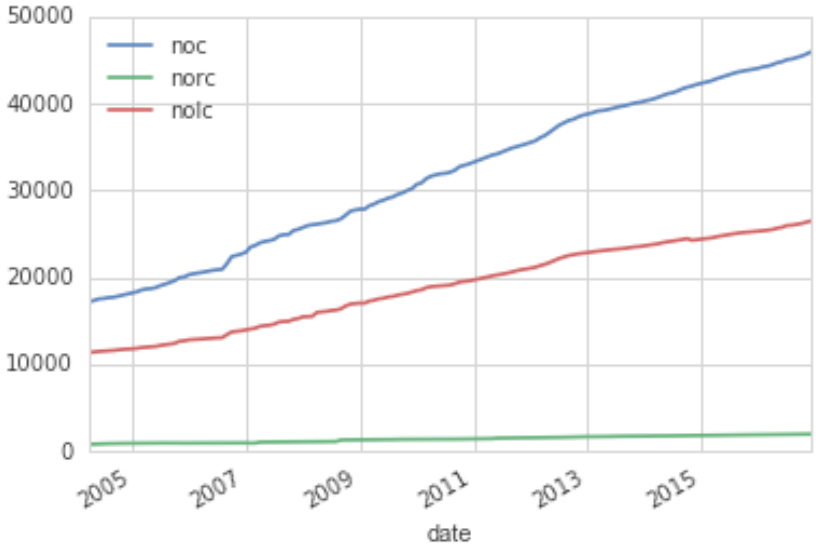


Fig. 4.1. Number of (root, leaf) classes across time.

Classes here have been interpreted as “terms”. It should be noted that, differently from the interpretation in other studies, here we consider that “root classes are so for every relation with a “bottomup” interpretation. This in our case includes both subsumption (“`is_a`”) and composition (“`part_of`”).

The evolution of the GO shows a broadening of the tree in the leaves with the number of root classes growing at a much lower rate. This can be interpreted as a further specialization of the GO at a smooth rate over time. As mentioned above, this is mainly the result of an increase in biological process terms.

4.1.2. Property-related metrics

Another category of important metrics relates to properties and their relations. Here we have selected the following subset of metrics:

- depth of subsumption (`dos`).
- relationship richness (`rr`). Defined as the ratio of the number of properties divided by the sum of the number of subclasses plus number of properties.
- inheritance richness (`ir`). Average number of subclasses per class.

These metrics give another view to the distribution of properties across terms. The depth of subsumption has `grown` moderately, from 14 to 16, which supports the idea that the grow of the GO is done in breadth.

The following Figure depicts `rr` and `ir` across time.

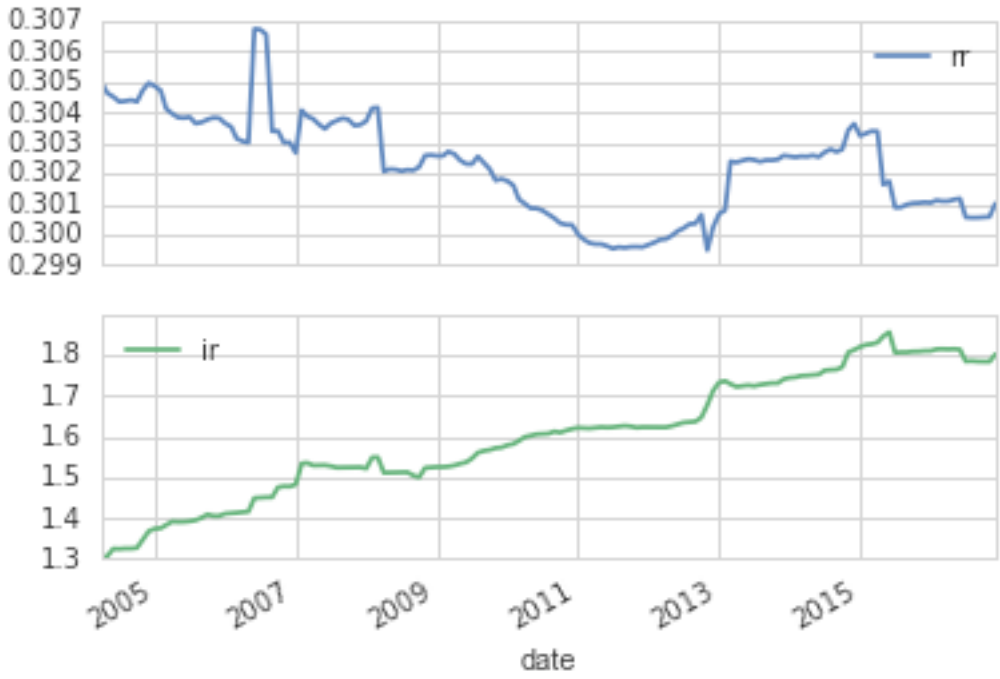


Fig. 4.2. Properties, depth of subsumption, relationship and inheritance richness and across time.

Inheritance richness has grown consistently with the increase in breadth of the hierarchy. However, relationship richness has declined over time, this represents an increase of the dominance of `is-a` relationship over other kind of relations in the ontology.

4.2. Relational term analysis

In addition to the metrics and overall quantitative insights, we have carried out an analysis based on network models of the GO terminology. The techniques and tools used are those coming from Social Network Analysis (SNA) methods (Wasserman and Faust, 1994).

When examining connectivity, it is important first to note that the three sub-ontologies present very different connectivity patterns, and they preserve along time. This can be appreciated in the following Figure that compares the snapshots of May 2004 and 2016. The apparent patterns appear similar, and this is also the case when examining snapshots in between these dates.

The subgraph on the left corresponds to the biological process sub-ontology (the one with the largest growing rate), the one in the centre to the molecular function and the smaller one is the depiction of the cellular component sub-ontology. As it can be appreciated, there is a different pattern of interconnectedness for molecular functions, appearing to have a more clustered structure.

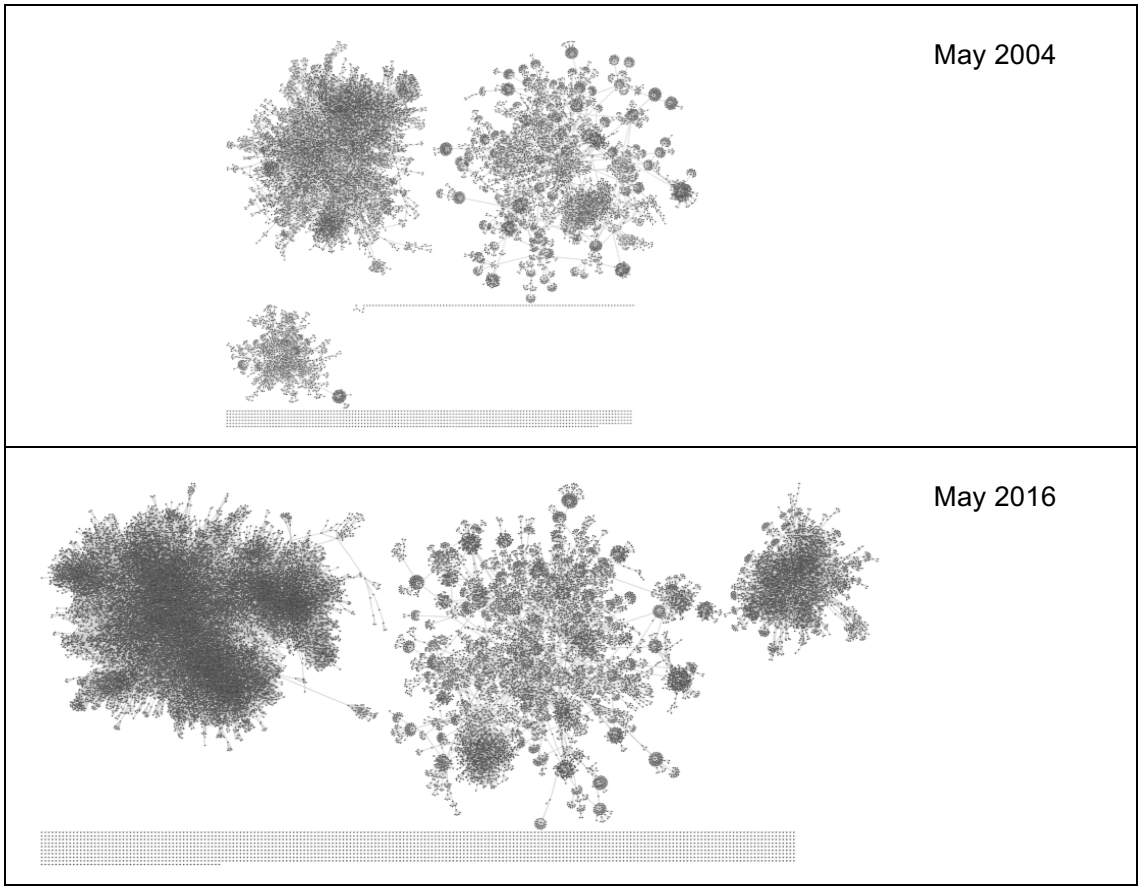


Fig. 4.3. Overall depictions of the GO three sub-ontologies in 2004 and 2016.

The graphics were generated using Cytoscape 3.4.0[15] from the directed graph of all the relations. Isolated points correspond to deprecated terms. The rendering was done using a force-directed layout. That layout positions graph elements based on a physics simulation of interacting forces, in which nodes repel each other, edges act as springs, and drag forces (similar to air resistance) are applied.

4.2.1. Structural change: clustering across versions.

A key question on the evolution of the GO that may impact similarity measures is that of the patterns of connectedness in the network. In SNA, this can be analysed via the clustering coefficient of terms (nodes). Essentially, that coefficient measures the number of "triangles" among terms. More formally, we use the following definition of a clustering coefficient for a node u in an undirected graph.

$$c_u = \frac{2T(u)}{deg(u)(deg(u)-1)},$$

where $T(u)$ is the number of triangles, and $deg(u)$ is the degree of the node.

[15] http://www.cytoscape.org/

Typically, distributions of clustering coefficients show an extremely uneven pattern, with most of the nodes having very low coefficients, and only a few of them having higher ones, as shown in the following Figure.

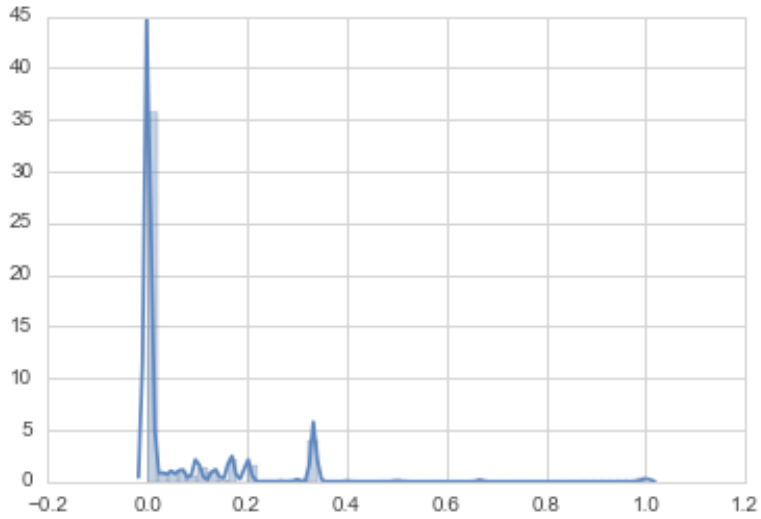


Fig. 4.4. Example distribution of clustering coefficients for the September 2016 snapshot of the GO.

Computing average clustering coefficients may not be detailed enough to trigger potential changes in different versions. For that reason, we observe the distribution of clustering coefficients in the graph. Concretely, we measure the clustering coefficients, then systematically assess differences in their distributions across snapshots.

The hypothesis is that the distribution of the coefficients, i.e. the overall structure of the ontology, does not change significantly across temporal versions. As the distribution of clustering coefficients is unknown, we have used a Kolmogorov-Smirnov test for two samples. The null hypothesis is that the two independent samples are drawn from the same continuous distribution. The rationale for this is that the ontology is a representation of reality with some topological structure and the different versions can be considered samples. While the idea that two versions are independent samples may be controversial, still the test helps us measure the difference.

The procedure for testing was that of using as samples the ontology snapshots as times t_i and t_{i+1}. This results in significant differences to be signals of changes that entail some important structural change. The following Figure depicts the resulting KS statistic and associated p-value. Also, the number of terms added between temporal snapshots and the increase in deprecated (obsolete) terms is also showed.

Fig. 4.5. Statistical differences of the distribution of clustering coefficients, and terms added or deprecated between dates.

The plots do not appear to show a systematic relation of the number of updates (new terms or terms discarded) with significant differences in the clustering coefficient distribution. In consequence, it can be hypothesized that there are changes that have a larger impact in the topological relations in the ontology than others.

4.2.2. Topological information content analysis

The main method for assessing the relevance of gene products to terms in the GO (e.g. indicating a biological process that may be of interest) is that of using similarity measures. These typically use some form of measure of the semantic relatedness of terms in the GO, in some cases also involving the accumulated evidence in the association database. These similarities in many cases entail the computation of a subgraph of the ontology and then applying some quantifying function that can be considered as a topological measure.

In any case, the differences in the network structure that have been described previously call for a more detailed analysis of the changes across versions of the similarity measures, as these are affected by the network structure of relations.

Quantifying that change is important towards the end of considering the extent to which the use of a GO-based tool may have significantly different outcomes when using a different version. Evidently, this is not conclusive of that impact, as it depends on many factors, including:

- Metric related: the extent to which the similarity measure used is sensitive or robust to changes in the network structure of the GO.
- Tool related: the extent to which the way the tool presents the results or enable filtering of most important associations does impact in the final conclusions of a study.
- Study related: not every subset of the GO related to a biological object of interest has been historically subject to changes in the GO. This entails that not every study is a priori affected by changes between versions.

Despite the complexity of the analysis, an initial account of the extent to which similarity measures may impact similarity measures is needed. Here we approach that analysis by considering topological measures only, as for now we are restricted to analysis using the terminological database. More concretely, we have implemented a generic form of similarity measure that enables extension to broaden this study to other proposed measures. The following Figure depicts how the main elements of the library that account for this.

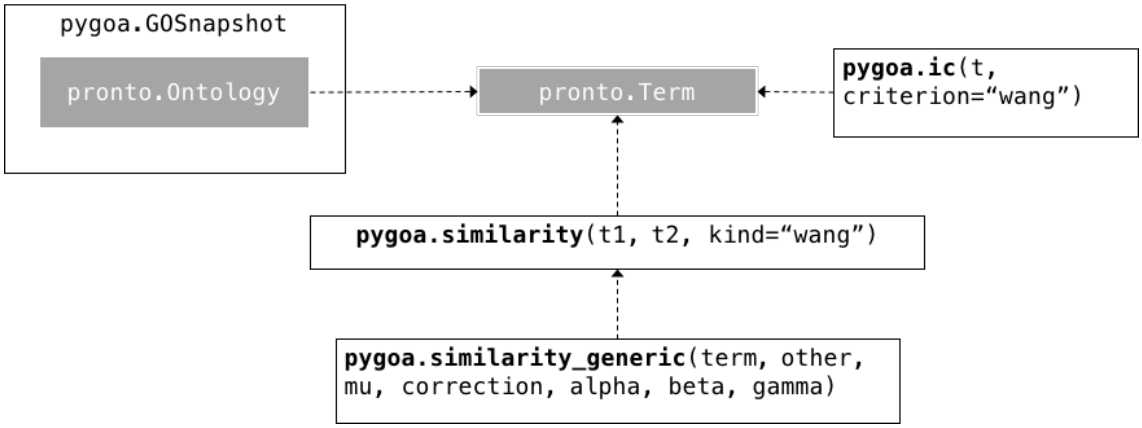


Fig. 4.6. Basic classes for defining semantic specificity and similarity measures

The different measures just depend on the `pronto.Term` class, that provides the required ontology context and method for computing the metrics based on the structure of the ontology. The `ic` and `similarity` functions just dispatch the types of measures defined for a single term and two-term similarity computations. The `similarity_generic` function implements the generic IC-based parametrizable formula discussed above, easing the implementation of variants of the measures by simply redefining the parameters.

As a measure of a topology-based information content (IC) we have examined the S-metric by Wang et al. (2007):

$$\begin{cases} S_A(A) = 1 \\ S_A(t) = \max\{w_e * S_A(t') | t' \in \text{childrenof}(t)\} \text{ if } t \neq A \end{cases}$$

Where a GO term A can be represented as $DAG_A = (A, T_A, E_A)$ where T_A is the set of GO terms in DAG_A, including term A and all of its ancestor terms in the GO graph, and E_A is the set of edges (semantic relations) connecting the GO terms in DAG_A. This metrics allows weighting the different relations and we used the weights w_e proposed in the original example of the paper, i.e., "is a" with value 0.8 and "part of" with value 0.6. From the above definition, the semantic value of GO term A, SV(A) is defined as:

$$SV(A) = \sum_{t \in T_A} S_A(t)$$

The computation of the SV metric to all the terms in the ontology across versions was used to examine the impact of changes in the GO in this semantic measure.

We have computed SV for each term and snapshot from 2004 to 2016. The analysis showed that 1870 out of 20532 terms analyzed had changes in their SV. This accounts for only a 9% of the terms, which is a small fraction, but still relevant enough to deserve further attention. It should be noted that this is a measure of semantic specificity, and it is used in turn to compute similarities between pairs of terms.
When looking at deprecated terms (the "`is_obsolete`" descriptor in GO OBO files) we found that around 58% of the terms that have changed their IC are marked as obsolete. The remaining 781 terms are the ones that persist in the ontology and changed their IC.

When examining the position in the taxonomy of the elements that changed their IC, an interesting observation arises. Concretely, the average number of descendants (defined recursively and including `is_a` and `part_of` relations) varies significantly for that subset. Interestingly, the pattern is also dependent on the sub-ontology considered. The following Figures depict the differences.

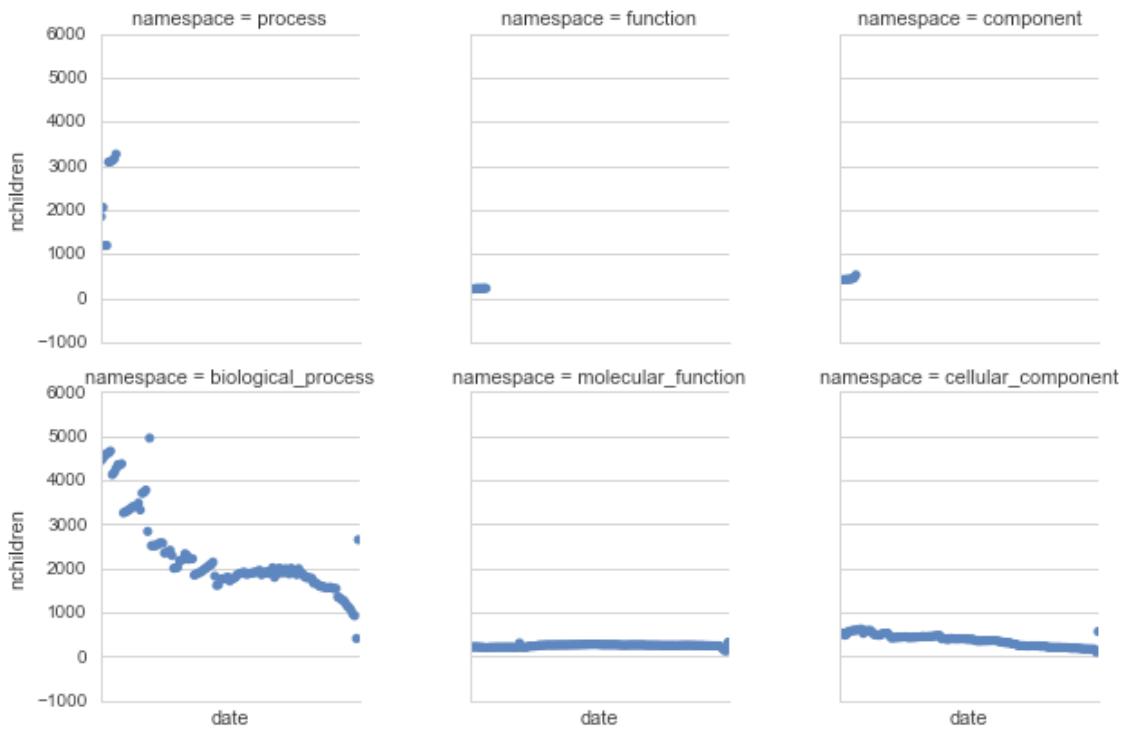


Fig. 4.7. Number of descendants for non-obsolete terms that changed their IC.

In both plots, the three plots in the first row correspond simply to a renaming in the files provided by the GO in older files.

Fig. 4.8. Number of descendants for non-obsolete terms that did not change their IC.

It appears as evident that number of children is a determinant of IC change. A Kruskal-Wallis H-test for null hypothesis that the population median of both groups are equal gives a p-value of zero, confirming the difference.

The intuitive explanation for this is that the terms changing their IC across versions tend to be those up in the hierarchy, as they get affected by additions in the hierarchical tree. This is in principle an argument in favour of considering that changes in IC overall do not affect the outcomes of studies, as tools tend to attempt to find the most specific terms and annotation guidelines are also guided by maximum specificity. Also, this difference affects mainly the biological process sub-ontology, that accounts for most the changes in the ontology.

4.2.3. Topological information similarity analysis

A second aspect of interest is the analysis of the change in similarities across versions. In doing the analysis, we have used and computed systematically across the versions 2004-2016 the following similarity measures.

Formula	Function
$S_{GO}(A,B) = \frac{\sum_{t \in T_A \cap T_B} (S_A(t) + S_B(t))}{SV(A) + SV(B)}$	`similarity_wang`
$S(x,y) = \frac{2 \max [IC(t) for\ t\ in\ Ax \cap Ay]}{IC(x) + IC(y)}$	`similarity_lin`

The formula by Wang uses the previously described IC measure. $S_A(t)$ is the S-value of GO term *t* related to term A and $S_B(t)$ is the S-value of GO term *t* related to term B. The formula by Lin is implemented over the generic function described above (Lin, 1998) where A_x represents a set with the term plus all its ancestors (the *subsumer* of x).

These metrics do not account for the annotation database, but are purely based on the network structure of the ontology.

Examining the complete space of similarity pairs for the different versions of the GO was discarded for practical reasons of the volume of computation. For example, generating the full combination of similarity pairs only for a single 2007 snapshot yields more than 134K million records, even when excluding the terms that have been discarded (made obsolete) along the time. Consequently, we have explored here the impact considering only the terms that changed their IC across versions (extracted from the analysis reported above) in relation to their neighbourhood, i.e. the terms that are adjacent to those in the graph of relations. While this does not give a complete picture of the impact, at least allows for having an initial impression of the impact of the changes in similarity measures.

The procedure then was done as follows:

1. Identify the terms that changed their topological IC metric across versions (extracted from the previous analysis), excluding terms made obsolete (that may bias the analysis as they are expected to have high variation as they become isolated in the graph once discarded).
2. Compute the subgraph of adjacent nodes for those terms.
3. Systematically compute the similarities using the different models for each of the pairs from the previous steps.

The results of the analysis when we take intervals of three years (2007, 2010, 2013, 2016) yields a large proportion of change of the 3832 similarities considered. Concretely, it accounts for around 71% in the case of Wang's metric, and near 75% in the case of Lin's metric. Around 85% of the similarities that change do that with both metrics (remember that the Lin's measures computed are based in the same IC account as Wang's for this analysis). Unsurprisingly, a large proportion of related terms is affected by the change, and this in turn may propagate to a lesser extent to terms that are reachable via more than one of the steps to one of the terms that changed IC.

The interpretation of the changes requires further analysis. It is important to note that the overall distribution of similarities in the sample analyzed do not appear to follow a unimodal distribution, as showed in the following plot for Lin's measures (std=0.3).

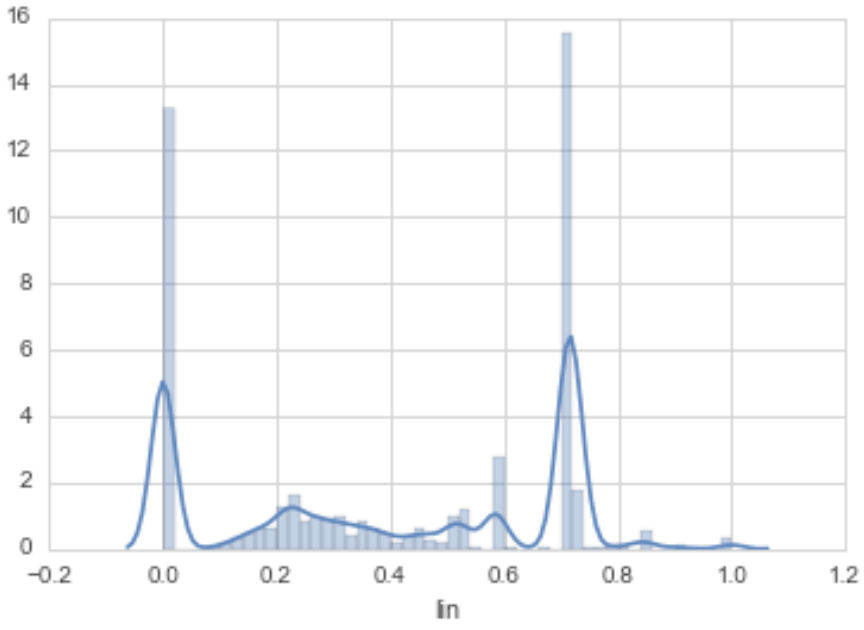


Fig. 4.9. Distribution of Lin's similarities between terms changed and their adjacent (intervals of three years, 2007-2016).

If we look at the trends per term for the case of Lin's measure, we find that the variation of similarities is concentrated in a few terms, only 49 of all the term pairs (1.7%) have a standard deviation above 0.1.

If we focus on those terms that account for most of the variation and build a graph model with all the relations, it results in a graph with 168 weakly connected

components, which can be interpreted as different independent significant changes. As can be appreciated in the following network plot[16], there are many independent terms that are isolated from the rest, then a connected subnetwork which appear to feature several local subnetworks.

The network diameter is 14, so considering that the minimum and maximum depth of subsumption (dosh) was between 13 and 16 along time, points out to a broad distribution of the changes across the graph. Density of the graph is low (0,001) and average clustering coefficient is 0,057.

Fig. 4.10. Graph of terms that changed Lin's similarity (intervals of three years, 2007-2016).

The network analysis up to this point does not provide clear insights on the locality of the changes in the network. However, if we focus attention on those nodes with standard deviation above 0.1 (discussed before), we can obtain the plots in the following Figure.

[16] Generated using Gephi's Force Atlas algorithm.

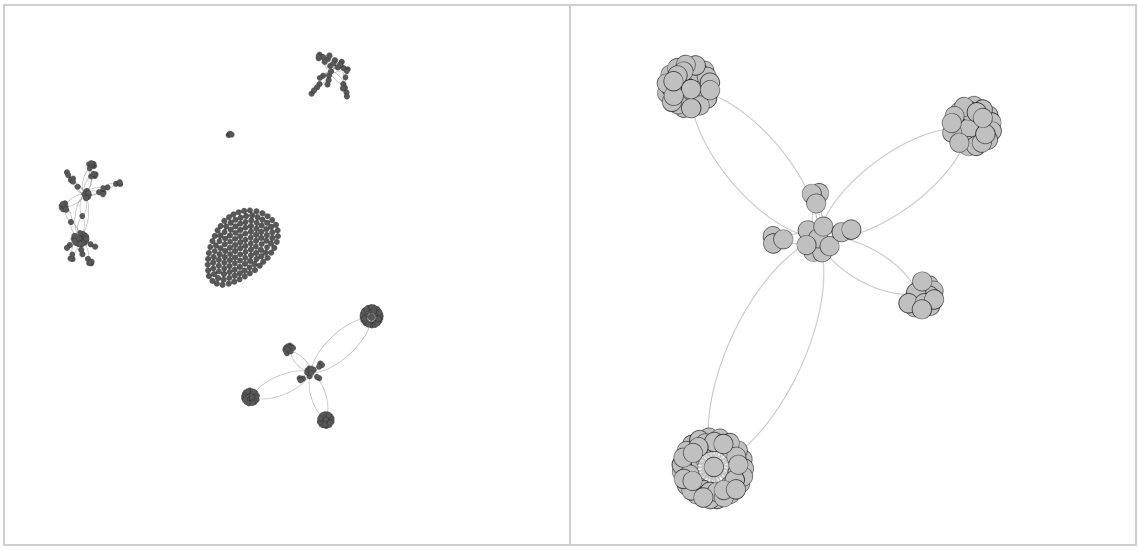


Fig. 4.11. Graph of terms that changed Lin's similarity with std>0.1 (intervals of three years, 2007-2016). Overall view (left) and detailed subgraph (right).

The element in the right of the previous one starts in the centre with GO:00003674, that is, "molecular function", i.e. the root of the subontology of the same name. Then three of the four clusters have centres using the Force Atlas visualization. They correspond to the following terms:

- GO:0003824 is "catalytic activity"
- GO:0005488 is "binding"
- GO:0005198 is "structural molecule activity"

As it can be appreciated, these terms cluster different forms of functions into groups. The other larger subgraphs account for the changes in the other subontologies.

The analysis done suggests that it is possible to localize the terms with the largest changes in the structure and relate them to particular kinds of studies or biological conclusions.

5. Annotation database evolution

In the previous sections, we have analysed the temporal evolution of the GO as a terminological knowledge base. However, it is in the associations database where the GO accumulates empirical evidence of high biological value. In this chapter, we turn our attention to that database, its evolution and composition.

5.1. Overall analysis

The overall temporal growth pattern was already discussed in Section 3. Here we focus on analysing how similarities derived from annotations vary over time. In doing so, we implemented a basic IC measure as follows:

$$IC(A) = -\ln\left(p(A)\right)$$

Where p(A) is the relative frequency of the term A for the annotation-based family. As we are here not differentiating among different databases contributing to the GO, we take the relative frequencies of the complete database.

Obtaining frequencies from annotation databases requires the off-line pre-processing of the large association database files that can be obtained from the GO consortium. The resulting files provide information on the number of annotations per term and per evidence code for each of the GO snapshots considered. The period 2005-2013 has been used in this analysis, but it could be extended to later years using the same methods.

The first important finding is that a total of 35,809 terms out of 39,264 have associated annotations in the period (considering the figures of the last year in the period), accounting for an 91% of the terms. However, the distribution of annotations across terms is unbalanced. The following plot depicts the distribution of the decimal logarithm of the total annotations for 2012 (similar empirical distributions can be appreciated for the rest of the years in the period).

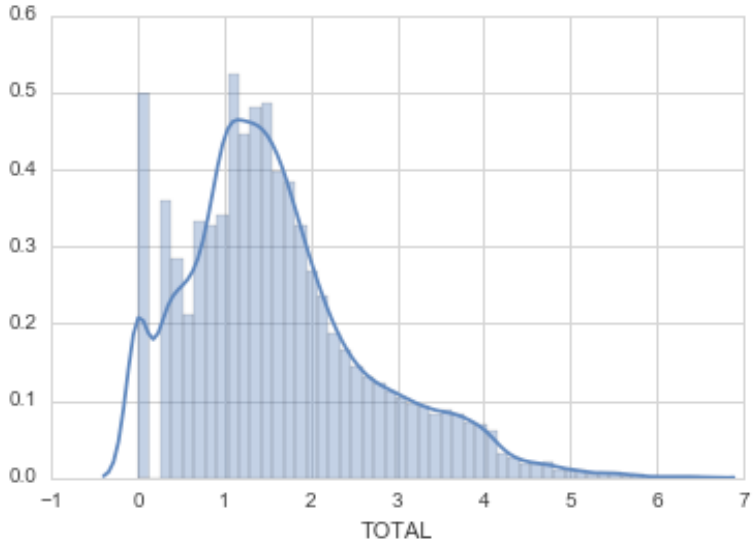


Fig. 5.1. Empirical distribution of number of annotations per term, in logarithmic scale (2012).

As the distribution appears stable, we can observe for a single snapshot the relation of annotations to other topological metrics in the ontology. However, there is not an apparent linear relationship between the number of annotations and measures of specificity as number of ancestors or descendants, as shown in the following Figure for the overall data and for the data per sub-ontology (obsolete terms removed).

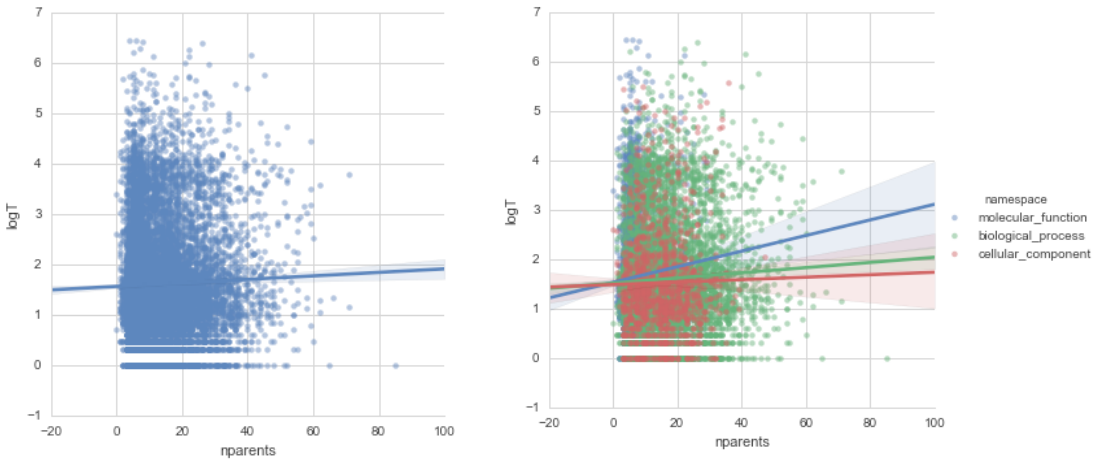


Fig. 5.2. Relationship of decimal logarithm of total annotations and number of ancestors of the term (2012). Total on the left, by sub-ontology on the right (linear model included in the plot)

The linear relationship appears to be different in the case of the molecular function sub-ontology and it is positive (as expected since annotation ideally should be associated with the most specific terms possible), but it is nonetheless unclear when considering other year's plots. The guidelines of the GO indicate that some terms are not intended to be associated to annotations, however, it is difficult so far to establish a relation.

To evaluate the changes in association-related IC measures, the first step was that of finding the impact of changes across versions in the period. A total of 5063 terms had changes in the frequencies, accounting for 14% of the total. However, unlike in purely topological ICs, here it does not seem reasonable to count all as changes, since a small increase in annotations may change the overall frequency.

Mean IC increases along versions, and more rapidly if we consider only terms that have their association-based IC changed. The following Figure depicts that increase.

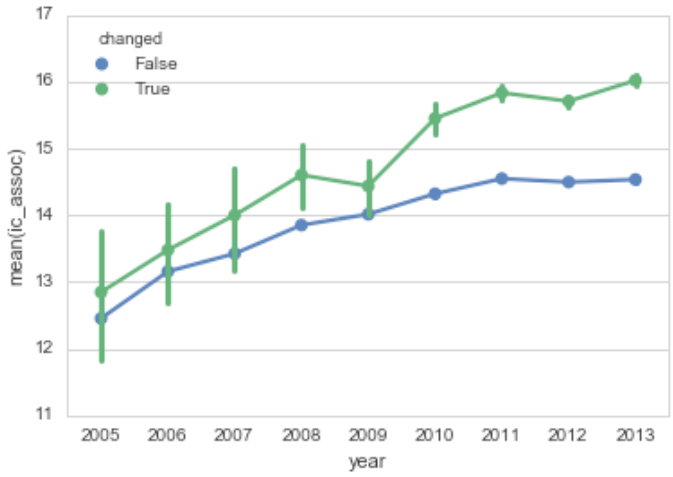


Fig. 5.3. Increase in mean association-based IC (2005-2013), split in groups of terms that change their IC in the period or not.

Another important aspect in this analysis is the extent to which these association based measures overlap with the IC measures studied in previous sections. In principle, these can be considered as addressing two different aspects of information content, namely, the connection with actual empirical products and on the other hand, the relationship to encoded knowledge in the terminology. Interestingly, the overlap of terms that changed its annotation-based IC in the term considered and those previously studied that changed their topological IC overlap only in 23 cases, which accounts to a 0,3% only of the terms[17].

5.2. Contrast between databases

The GO as a consortium has a decentralized approach to collecting associations. This entails that the association database is an aggregation of the contributions of different projects that specialize on an organism or have a concrete focus. This may be a source also of differences across the different contributing projects and here we provide the analysis of the contributions across time.

[17] This may not be accurate as the period considered for topological ICs is broader, but in general this accounts for a small proportion of the terms.

There is a total of 57 databases that have contributed to the GO in some moment. However, of those only 25 remain “active” in the sense that they appear in the latest version. This in some cases may be due to merging of some of them, but this also entails that tools should be aware of the underlying databases used and report them, as maybe in future versions they are no longer available as independent contributors.

The following Figure depicts the contribution of the 25 active databases. As it can be appreciated, there is not a growing pattern in some of them. On the contrary, the contribution patterns over time are highly heterogeneous thus there is not a characterization of the evolution of these contributions.

Also, it is noticeable that the contribution of databases is very uneven, with some of them being less than one hundred in size as `NCBI_GP`, and others like `UniProtKB` providing the bulk of the annotations.

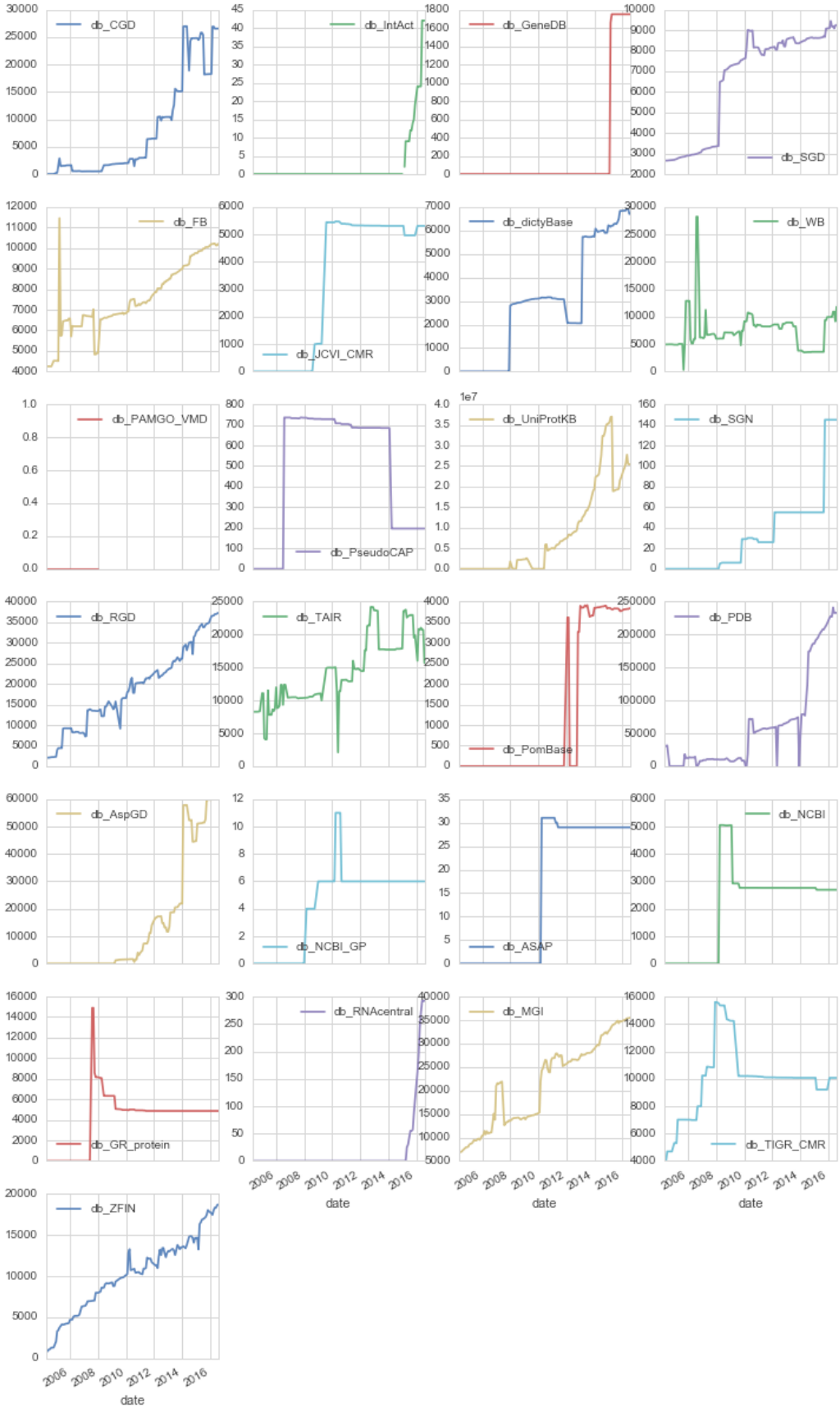


Fig. 5.4. Contribution over time of the 25 databases that are listed in the summaries up to august 2016.

5.3. Evidence level analysis

In Section 3, it was showed that IEA evidence codes dominated the overall database of annotations. Also, it was shown that the distribution of evidence codes between categories was also dominated by some kinds of evidence.

The following diagram shows the empirical distributions of the dominant categories for experimental and author categories (curatorial are omitted as their relative frequency is much lower).

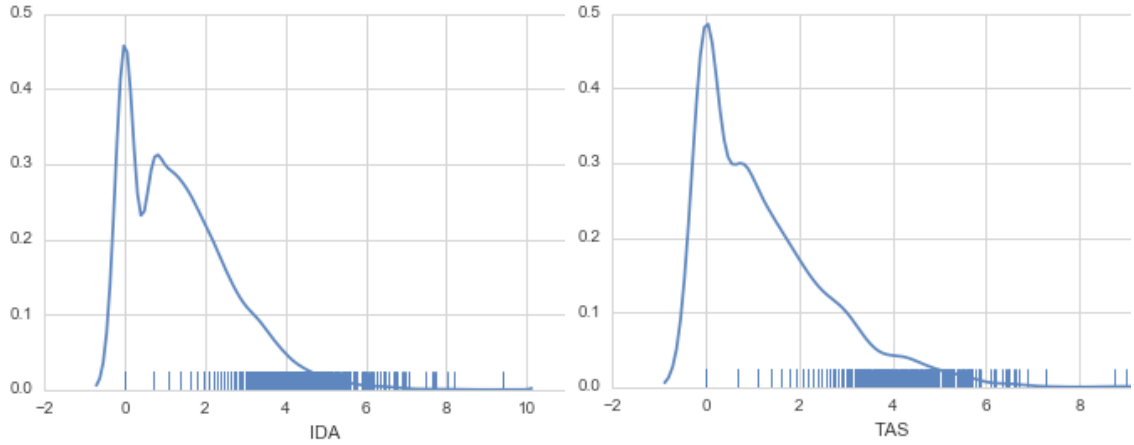


Fig. 5.5. Empirical distributions for the dominant categories of evidence codes for year 2013.

Both IDA and TAS show similar distributions. The case of the computational analysis codes is different, as up to 2013 there was a relative distribution differing significantly from the most recent period. The following Figure shows their distributions.

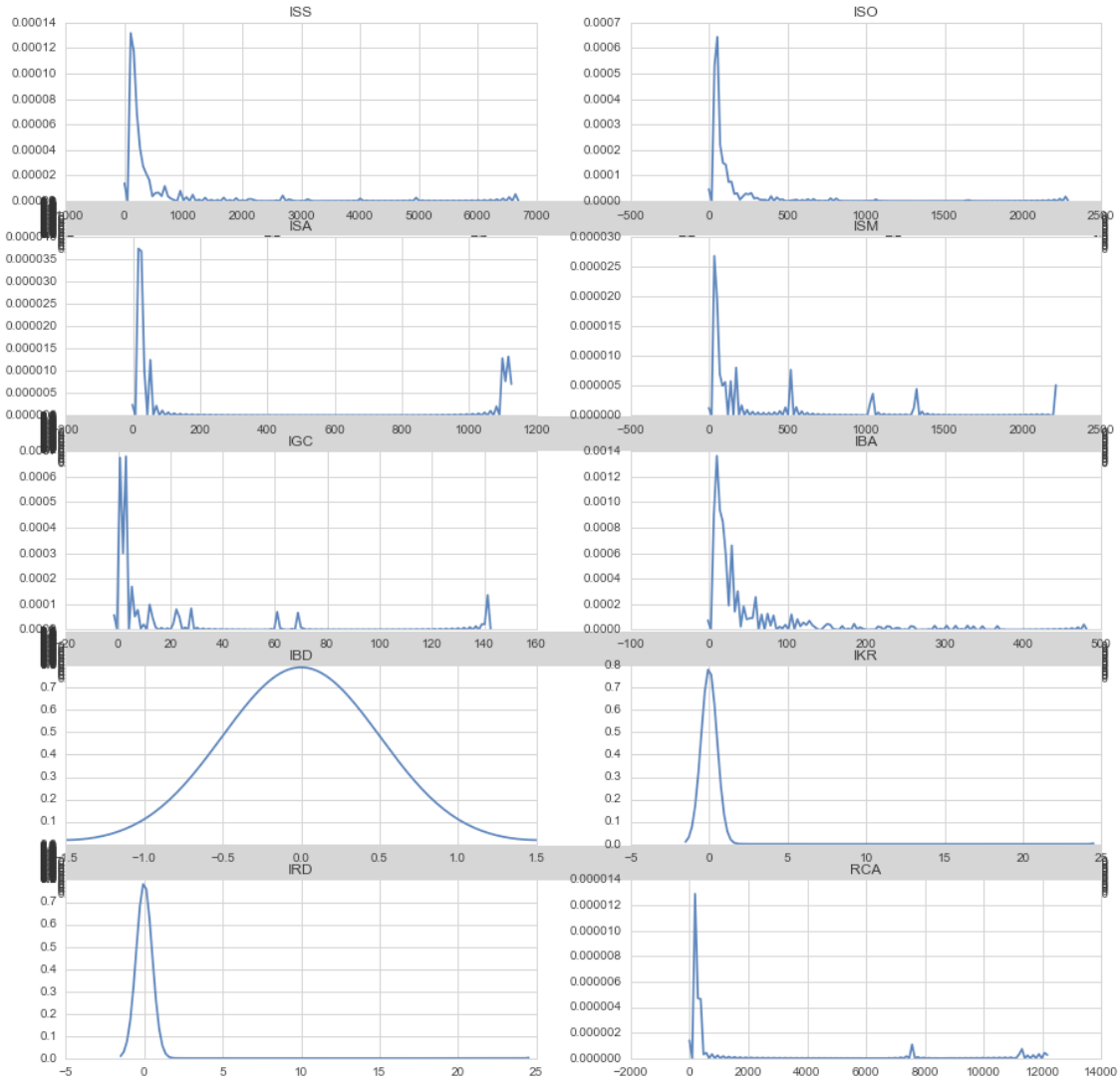


Fig. 5.5. Empirical distributions for the computational analysis code category of evidence codes for year 2013.

With the exception of IBD, all of them show a kind of unbalanced distribution.

Another interesting aspect is the potential relation of evidence kinds across different kind of terms, based on their position in the ontology. However, no regression model has been found that accounts for that relation.

6. Conclusions

This section summarizes the main conclusions and outcomes from the work described in the rest of the document and assess their relation with the original objectives posed for the thesis work. Then, a discussion of future direction to continue the effort reported in this document is provided as outlook.

6.1. Main outcomes

6.1.1. Overall findings

The analysis of the GO provides a picture of the ontology as a constantly evolving and growing resource. The terminology appears to follow a linear pattern of growth with an unsurprising higher rate in the case of relations when compared with the number of terms. Also, the biological process sub-ontology shows a more accelerated growth rate, than the other two. The associations database has been shown to be an amalgamation of contributions from different sub-projects that change their contribution proportion differently over time.

Evidence codes are dominated by electronically produced annotations across time, and considering the remaining relatively small portion there are dominant codes in each of the categories. Noticeably, the proportion of annotations of the IBA type has grown from 2015 on at a much higher rate, becoming in that short period the most frequent category.

From the analysis of ontology metrics, it is noticeable that relationship richness has declined over time, representing an increase of the dominance of `is-a` relationship over other kind of relations in the ontology.

A graph based analysis of the relations in the ontology across time shows different network structures among sub-ontologies. Considering the distribution of clustering coefficients as an indicator of relational structure, it has been shown how there are statistically significant changes between consecutive monthly versions of the ontology at some particular points. This could be further explored as an indicator of important structural changes.

The examination of terms for which topological information content (IC) measures change over time suggest that changes occur in higher levels of the hierarchy. Assuming this is true, that would be an argument in favor of the stability of such measures for the terms that are more specific. The changes in IC propagate to changes in the similarity of those terms with other ones. An analysis of terms

adjacent to those has shown that impact. However, when looking at the effect, it becomes apparent that only a few terms account for the majority of the variation, and some of them are related in the structure of the ontology. This opens opportunities to elaborate accounts of "evolution impact on terms" on particular sub-trees of terms. That would enable a practical account to signal potential needs of reinterpretation of studies that were based on terms that have significantly changed.

The analysis of annotation-based ICs has shown that they account for a different kind of changes than that of topological ICs. Further, that specificity grows over time, which may be interpreted as an increasing discriminatory power of the ontology when considering the annotations.

6.1.1. Processing tool

The `pygoa` library provides a practical approach to obtaining snapshots of the GO and transforming them into formats widely used in the Python scientific stack. For example, summary ontology data (both for terminology and associations) can be easily transformed into `pandas DataFrame` objects, or they can be exported to the digraph format of the `NetworkX` library. This maximizes the reuse of existing analytic tools and allows data analyst to use the more convenient tool for each analytic need.

The library has included pre-processing code for the case of the annotation database, which due to its size and growth pattern, falls into the requirement space of parallelizable computation. Concretely, the Apache Hadoop Map-Reduce paradigm has been used to get the frequencies.

6.2. Relation to original aims and lessons learned

The following Table summarizes the main outcomes reported in relation to the originally stated objectives.

Objective	**Where in the document**	**Main outcomes**
O1.1. Design and develop a framework to extract features and metrics from the GO compatible with SciPy, the scientific stack built around the Python ecosystem.	Described in chapters 4 and 5 in its major usage for analytics. Code available in Github.	The `pygoa` library, registered and made available open source. Use with SciPy libraries `numpy`, `pandas` and `scipy.stats`, and also with other libraries as `networkx` and `mrjob` (for preprocessing).
O1.2. Design and develop tools for the analysis of internal relationships inside the GO.	Described in chapters 4 and 5.	Libraries for ontology metrics, conversion to graph models, extracting term-related metrics and similarity and IC measures.

O2.1. Develop software for the analysis of GO versions along time including its annotation database.	Usage reported in chapters 4 and 5.	Tools for preprocessing GO files and systematically downloading and caching them.
O2.2. Evaluating potential known GO problems according to the analysis done.	Usage reported in chapters 4 and 5.	An analysis of the impact of changes in sub-ontologies along relations has been described in 4.2.3. Evidence codes have been examined in 5.3 and differences across databases in 5.2.

While all the objectives have been explicitly covered in the work done, many alternate paths or additional possibilities have been left unexplored, and some new have been suggested by the findings. In the next section, an acoount of some of them is provided.

6.3. Outlook

6.3.1. Difficulties found in data wrangling

Of the options available, monthly GO snapshots have been chosen as a solution balancing granularity and ease of retrieval. The alternative of using the CVS interface may have given finer temporal granularity at the expense of increased processing cost. However, the changes in the GO accumulate over time and the versions in the CVS are no releases semantically marking significant changes, but a series of small routine updates. This led us to assume that a monthly periodicity is enough for a realistic account of the changes in the GO.

As with any project spanning across many years, there are some changes in formats, schemes and procedures that affect data acquisition. We have found and reported problems in obtaining earlier versions of the GO (e.g. due to using older versions of the OBO format), and small changes in the policy of naming files. Also, the monthly snapshots repository has some missing months.

A way of caching versions was needed to avoid constant network retrieval of large files for the terminology of the GO. This has the drawback for the user of the library of consuming some additional disk space, reason why it was made optional. While for the terminological database, the size growth of the files does not appear to pose problems in the future, it is clearly a problem in the case of the annotation database.

The volume and growth of the annotation database clearly required an approach for offline computation that could scale to clusters to cope with future changes. Here we devised a simple pre-processing workflow exploiting the broad availability of cloud services and software built on top of the Apache Hadoop framework. This has the drawback of requiring the development of custom

workflows for particular data requirements. For example, here we have implemented the workflow for getting annotation frequencies per term, but other possible analysis may for example attempt to extract references to sequences, and this would require custom workflows and an additional level of scaling.

Another category of problems comes from the computational power required to do extensive longitudinal analysis of GO metrics. The clearest case is that of the computation of similarity changes across versions of the GO. The volume of computation required again requires moving to parallel processing paradigms. Here we have approached that using simple custom parallelism with IPython parallel[18] and ad hoc strategies to get the information required for the analysis presented, but a more generic analytic workbench may be more optimal for comprehensive analytics. Possible frameworks for that may be Apache Spark[19] or Apache Flink[20], but these have been considered out of the scope of our current presentation.

6.3.2. New forms of release management for the GO and its implications

Following the emerging paradigm of replicable experiments, GO-based studies should be made computationally replicable when published, so that relevant changes in the ontology or the associations database could trigger re-execution of the experiments and eventually provide new insights or test the significance or relevant of previously obtained ones.

The current release system of the GO is not systematic and a change into a more stable and consistent release effort may help in tool and paper authors document the concrete versions used.

Ideally, a form of meaningful release management in the GO would provide the benefit of signalling potential changes in the ontology that may affect previous conclusions or may open new opportunities for biological discovery. However, determining what may be "significant" as a new release and what not is difficult and in need of additional research. The metrics and comparisons reported in this work could be used as a starting point to study potential measures of ontology evolution that when combined with reproducible research, could lead to a "re-execution" of studies in an automated way, making the studies more valuable by being able of incorporating new knowledge embedded in the tools without a need of redoing the full path or workflow. It should be noted that re-execution may be done off-line, and only when some significant change is detected then the authors or curators of the results may enter into action. For example, in tools that provide GO-based processes most related to under-expressed genes in an experiment, that list could just be compared with the new outcomes.

---

[18] https://ipython.org/ipython-doc/3/parallel/

[19] http://spark.apache.org/

[20] https://flink.apache.org/

6.3.3. Beyond the analytics presented here

The analytics presented here have been obtained in a process of exploratory study. As such, they have served only as a way of obtaining additional findings for further studies. These findings that are suggestive of further work include but are not limited to the following:

1. Differences in the impact of changes in different similarity or IC measures point out to the need for a comparative robustness of different of such measures.
2. Differences in number of children in terms changed across versions of the ontology point to studying the locality of changes relative to the graph structure as a promising direction.
3. The identification of subgraphs in terms impacted by changes suggest studying the propagation of changes of similarity measures in particular subtrees of the ontology.

Also, a number of predictive tasks have been suggested by the results in this work, including but not limited to the following:

1. Is it possible to predict the GO terms that are in risk of being made obsolete? This could be done from the history of terms discarded, observing its structural position (knowledge-based view) or its associated annotations (use view).
2. Is it possible to predict the growth of the GO in particular branches or subtrees? As the biological process sub-ontology accounts for most of the additions, a model of which subtrees are more likely to grow via “is a” or “part of” specialization may be derived.

The second of the questions is related to the broader question of the extent to which the GO will at some point stop growing significantly in its terminological part. As the GO codifies biological knowledge, it can be hypothesized that there is a point in which our creation of subcategories for sub-parts or sub-processes may reach a limit, which is that of the material processes being described. Identifying that point is of paramount importance to research as it would mark the point in which the GO will become an overly static resource.

7. Glossary

7.1. Acronyms

- GO: Gene Ontology.
- OBO: Open Biomedical Ontologies.

7.2. Terms

- clustering coefficient: in network analysis, a measure of the connectedness of a node in the graph with pairs of other nodes in the same graph.
- information content: in our context, a measure of a term's specificity in the context of an ontology.
- ontology: a shared explicit specification of a conceptualization[21].
- ontology metric: a measure of an ontology elements that allows comparing ontologies and uses one or several features obtained from its distinctive characteristics: terms, relations, axioms.
- similarity measure: in our context, a measure of the level of semantic relatedness of two terms in a ontology.

[21] Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge acquisition, 5(2), 199-220.

8. References

du Plessis, L., Škunca, N., & Dessimoz, C. (2011). The what, where, how and why of gene ontology—a primer for bioinformaticians. Briefings in bioinformatics, bbr002.

Guzzi, P. H., Mina, M., Guerra, C., & Cannataro, M. (2012). Semantic similarity analysis of protein data: assessment with biological features and issues. Briefings in bioinformatics, 13(5), 569-585.

Lin, D. (1998). An information-theoretic definition of similarity. In ICML (Vol. 98, pp. 296-304).

Mazandu, G. K., Chimusa, E. R., & Mulder, N. J. (2016). Gene Ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery. *Briefings in Bioinformatics*, bbw067.

Pesquita, C., Faria, D., Falcao, A. O., Lord, P., & Couto, F. M. (2009). Semantic similarity in biomedical ontologies. PLoS comput biol, 5(7), e1000443.

Sicilia, M. A., Rodríguez, D., García-Barriocanal, E., & Sánchez-Alonso, S. (2012). Empirical findings on ontology metrics. Expert Systems with Applications, 39(8), 6706-6711.

Wang, J. Z., Du, Z., Payattakool, R., Philip, S. Y., & Chen, C. F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics*, *23*(10), 1274-1281.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications (Vol. 8). Cambridge university press.