# Identification and exploration of gene product annotation instability and its impact on current usages

by

Adriana Estela Sedeño Cortés

B.Sc., National Autonomous University of Mexico (UNAM), 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

October 2014

# Abstract

Proteins are macromolecules responsible for a wide range of activities in the structure and function of cells. Their activities have been described in different contexts as a mean to elucidate their "function". These descriptions have been captured across biological databases in a standardized format called Gene Ontology Annotations (GOA), to disseminate the knowledge and extrapolate the information to other proteins whose function is still unknown. Furthermore, the annotations are used to analyse and interpret data from high-throughput studies and also as a benchmark for the assessment of protein function prediction algorithms. Constant changes occur in GOA that can potentially impact such usages, but only limited effort has been put into exploring their instability, or to assess the impact that these changes have on reproducibility or interpretation of previous analyses.

In the present work, I performed the most comprehensive analysis of the annotation instability for 14 representative model organisms (E.coli, fruit fly, mouse, etc.). The results showed important instability patterns that were species-specific. As such information would be of use to the community to trace the instability of annotations of their interest, a web-based visualization tool was built to track these changes on a protein, functional term and species specific basis.

Additionally, we identified artifacts on the annotation data that can be attributed to curation patterns. We propose such artifacts to be considered for a more accurate assessment of function prediction algorithms. Furthermore, the impact that changes in the annotations have on common settings like gene set enrichment analyses was also explored. In particular, 2,000

datasets were used to assess the robustness of enrichment results over time. On average, the results would display a  60% similarity after only 2 years. However, cases were found were the similarity will drop  80% within the same year, demonstrating the impact that the instability has on such applications. In conclusion, the results of this work will prove useful for those who use the annotations to interpret their studies to assess their reliability on a case-by-case scenario.

# Preface

The present work was elaborated at the Centre of High-Throughput Biology (CHiBi) in the UBC's Michael Smith Laboratories (MSL) under the supervision of Paul Pavlidis.

I am responsible for the data collection, design and code implementation done in this project to pre-process and analyse the data. My supervisor Paul Pavlidis contributed with the study design, supervision and editorial suggestions for all chapters.

Jesse Gillis contributed suggestions for the evaluation of the performance of gene function prediction algorithms.

# Table of Contents

# List of Tables

# List of Figures

# List of Scripts

# List of Abbreviations and Definitions

**AUROC** Area Under the Receiver Operating Characteristic Curve

**BP** Biological Process Ontology

**CAFA** Critical Assessment of Automated Function Prediction

**CC** Cellular Component Ontology

**ChEBI** Chemical Entities of Biological Interest

**DAG** Directed Acyclic Graph

**DDBJ** DNA Data Bank of Japan

**DE** Differentially Expressed

**EBI** The European Bioinformatics Institute

**EMBL** The European Molecular Biology Laboratory

**GAF** Gene Association File Format

**GO** Gene Ontology

**GOA** Gene Ontology Annotation

**GOC** Gene Ontology Consortium

**GPAD** Gene Product Association Data format

**GPI** Gene Product Information Format

**HPO** Human Phenotype Ontology

**MF** Molecular Function Ontology

**OBO** Open Bio-medical Ontologies

**SGD** Saccharomyces Genome Database

**SwissProt** Swiss Institute of Bioinformatics Database

**TrEMBL** Translated EMBL Nucleotide Database

**UBERON** Integrated Cross-species Ontology

**UniProtKB** The Uniprot Knowledgebase

# Acknowledgements

I would like to thank the Centre for High-Throughput Biology and the Michael Smith Laboratories at the University of British Columbia, where this research project was conducted. A space where thoughts and passion are shared among and beyond its community.

To my supervisor, Paul Pavlidis, Professor of the Department of Psychiatry and Associate Director of the UBC Graduate Program in Bioinformatics. I greatly appreciate all the support, accessibility, guidance, patience to explain things, constructive feedback and critical thinking that made this an invaluable learning experience to me.

To Jesse Gillis, Assistant Professor at the Cold Spring Harbor Laboratory, previously a post-doctoral researcher at the Pavlidis lab, for all his advice, insightful thoughts and guidance.

To Ryan Brinkman and Nobuhiko Tokuriki, members of the thesis committee for their constructive feedback.

To the NSERC and NIH as founding sources.

To the Gene Ontology Consortium (GOC), the European Bioinformatics Institute-European Molecular Biology Laboratory (EMBL-EBI) and organizers from the Critical Assessment of Function Annotation Experiment, sources of the data used for this research project.

# Dedication

To my family and friends for all your love and support. Being far from you has been difficult but you are always in my thoughts. Thank you for all those moments, memories and time spent together, for supporting me in every step of the way.

*"The important thing is not to stop questioning.*
*Curiosity has its own reason for existing.*
*One cannot help but be in awe when he contemplates*
*the mysteries of eternity, of life, of the marvelous*
*structure of reality. It is enough if one tries merely*
*to comprehend a little of this mystery every day".*

`Albert Einstein.` *From the memoirs of William Miller, an editor, quoted in Life magazine, May 2, 1955; Expanded, p.281*

# Chapter 1

# Introduction

Proteins are biological macro molecules responsible for a wide range of activities in the structure and function of cells. Research has focused on describing protein activity in different contexts as a mean to elucidate their "function". This information is being captured across biological databases in a standardized format, called Gene Ontology annotations (GOA). The primary reason to create the annotations is to disseminate this knowledge, compare the information across species and extrapolate the information to other similar proteins whose function is still unknown. GOA has become over time a key resource and is increasingly used to analyse or interpret the large amount of data generated from high-throughput studies and also as a benchmark for the assessment of protein function prediction algorithms.

However, the GO and the GOA are not complete nor perfect. Multiple changes occur in their structure to better reflect the current knowledge. The variability derived from such modifications are likely to affect the outcome of the current uses, specially for the interpretation of biological data, but critical evaluations on the limitations of GOA are limited in number and scope.

To properly assess the usefulness of the annotations to analyse or interpret biological data, it is crucial to understand first: 1)how these annotations are generated, 2) where annotations come from and 3)what factors influence their changes, if one aims to identify the limitations of GOAs in current applications. Furthermore, the historical information should be accessible for comparison purposes to the community. However, no tool has been developed and made available to conduct such evaluations.

In this thesis, I performed the most comprehensive analysis of the historical changes that have influenced GO and GO annotations; built a visualization tool to make this data accessible for exploration and assessed the impact that these changes have in current applications. Even though there is concern within the scientific community of such impacts, only a handful of studies evaluating the annotation quality have been published, all with some limitations that I attempt to overcome.

In this chapter, I introduce the background for my research, with an overview of the Gene Ontology and its annotations, properties, current usages and describe some of the previous assessments that have been done on this data.

## 1.1 The Gene Ontology

16 years ago the Gene Ontology project was created to integrate and facilitate the exploration of biological information behind different genomic and proteomic studies. The Gene Ontology Consortium (GOC) is a set of genome database organizations and communities that have joined efforts to develop and maintain the Gene Ontology (GO), currently considered the most important ontology within bioinformatics. Its original publication [1] has over 13,371 citations based on Google Scholar (as of September 29,2014). The GO describes gene attributes using a standardized vocabulary (terms) in the form of a directed acyclic graph (DAG) [1]. The terms are classified in three independent aspects or domains: 1) *Molecular Function Ontology* (MF): Activities of the gene product within the cell (e.g. binding, receptor, enzymatic or transporter activities); 2) *Biological Process Ontology* (BP): A series of activities or events that a gene product is involved in within the cell (e.g. cell-cell signalling, locomotion, cell death); and 3) *Cellular Component Ontology* (CC): Describes sub cellular locations and macro molecular complexes within the cell (e.g. membrane, pyruvate dehydrogenase complex, protein storage vacuole).

In each of those domains, terms are represented as nodes (with a name and an identifier or accession number) and are inter-connected with other parental terms (more general entities) and/or children terms (more detailed entities) by edges that represent different relationships:

- **is_a**: represent cases where the the children term **B** is a sub type of the parental term **A** (e.g. "enzyme regulator activity" *is_a* "molecular function"; "anoikis" *is_a* "apoptotic process").
- **part_of**: represent cases where the children term **B** implies the presence of the parental term **A**, but given **A** we cannot ensure that **B** exists (e.g."catalytic activity" *part_of* metabolic process"; "signal transduction" *part_of* cell communication").
- **has_part**: represent cases where the parental term **A** always has the children term **B** as a part; if **A** exists, **B** will always exist (e.g. "protein binding transcription factor activity" *has_part* protein binding"; "nitrogen utilization" *has_part* nitrogen compound metabolic process").
- **regulates**: represent cases where the children term **B** necessarily regulates the parental term **A**, but **A** may not always be regulated by **B**. The regulation of a process does not need to be part of the process itself. Two sub-relations exist to represent more specific forms of regulation (e.g."regulation of mesenchymal cell apoptotic process" *regulates* "mesenchymal cell apoptotic process"; "positive regulation of catalytic activity" *positively_regulates* "catalytic activity"; "negative regulation of M phase" *negatively_regulates* "cell cycle process").
- **occurs_in**: Used to link an occurring function or process to a location (process A necessarily occurs in component B) (e.g. "mitochondrial RNA processing" *occurs_in* "mitochondrion"; "COPII-coated vesicle budding" *occurs_in* "Golgi membrane").

The three ontologies of GO are each represented by a root term with no common parental node, but their terms can be inter-connected through

the *part_of* or *regulates* relationships. For example, "catalytic activity" *is_a* "molecular function", but is also *part_of* "metabolic process" which *is_a* "biological process" (**Figure** 1.1).

The GO structure is constantly revised and modified to cover missing links and incorporate new biological knowledge. Some modifications often found are:

1. **extensions**: Terms being added when missing attributes are identified;
2. **reductions**: Terms being deleted, when definitions are vague or do not accurately represent a biological aspect;
3. **revisions**: Terms being split, merged, substituted, moved on a different location within the graph;
4. **cross-products**: Terms combined through aggregating certain relations (includes terms from other ontologies such as the Cell Ontology, Plant Ontology, Uber Anatomy Ontology (UBERON) or ChEBI (chemical entities of biological interest)). Example: "DNA replication" + *"occurs in"* + "mitochondrion" = "mitochondrial DNA replication" [2, 3].

These modifications are included in each new release of GO. Daily and monthly versions can be found and different formats are available to use:

- Basic version: Includes *is_a*, *part_of* and *regulates (positively and negatively)* relationships and excludes those that inter-connect the ontologies. It is the recommended format for GO annotations. This version is often used for most of the GO-based annotation tools available.
- Core version: Available in two formats (OBO and OWL-RDF/XML). It is the non-filtered version and includes the *has_part* and *occurs_in* relationships, but excludes relationships to other ontologies. These relationships are recommended to be excluded for propagation, which

is important to note as many enrichment tools consider the propagated terms for their results.

- Plus version: Includes dependencies to other external ontologies and some inter-ontology relationships.
- GO Slim version: It is a subset of the ontology created to provide a broad view of the graph, the most granular terms are removed. This version is often used in many applications as it does not include species-specific terms.

Over time, many different tools have been developed to browse GO or its annotations, each one retrieving one of these different formats. For example, tools like "CateGOrizer"[4] or "GOSlimViewer" [5] use GO slim versions, whereas "GO::TermFinder" [6] use the basic version but consider only terms associated to gene products.

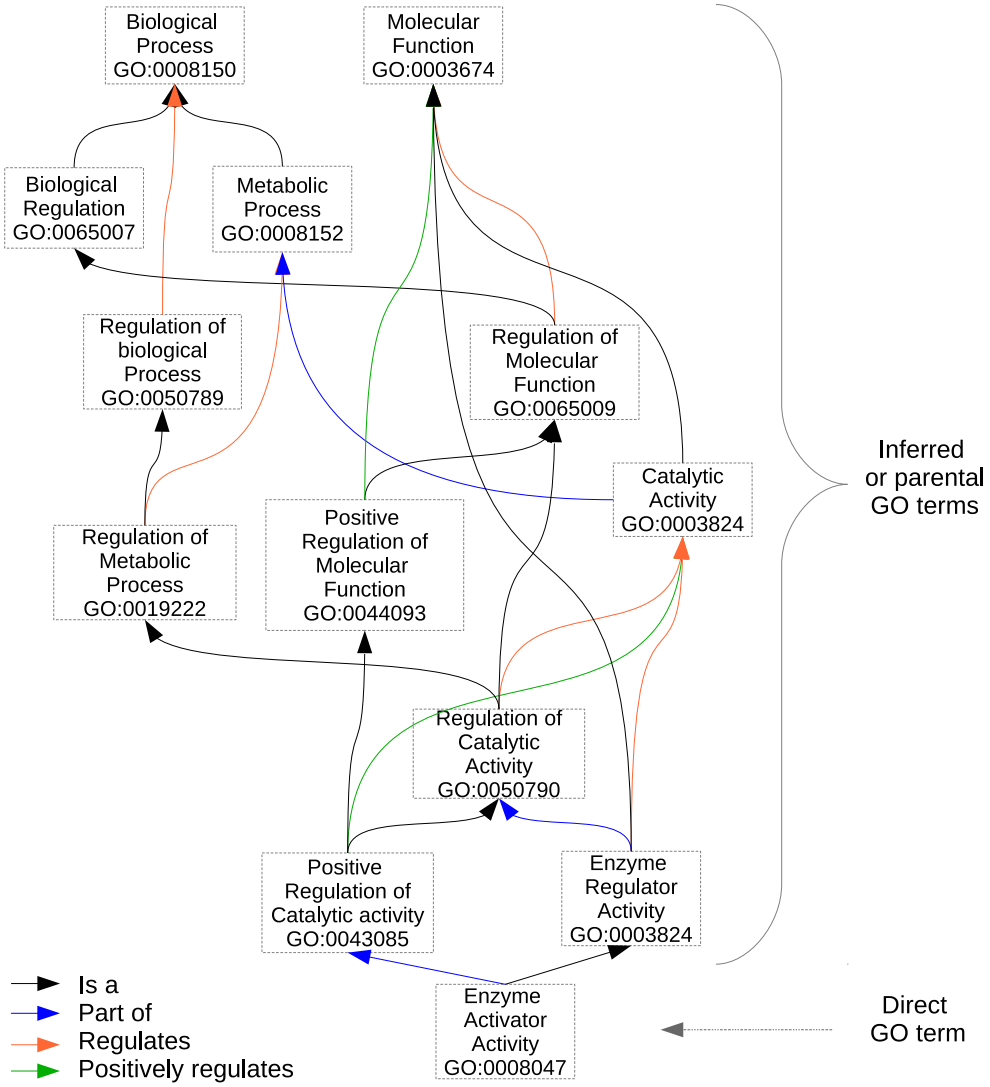



Figure 1.1: An illustration of the structure of the Gene Ontology graph. A term can have multiple parental and children terms and different relationships between them. The term that is often annotated to a gene product is called a "Direct GO term" and the terms that can be inferred by propagation to the root node are called "Inferred or parental GO terms".

## 1.2 GO Annotations

With the active collaboration of 36 groups, the GOC releases monthly versions of GO Annotation files (GOA) that capture the association between gene products and GO terms for different species. For a gene product to become annotated, an electronic or experimental evidence must indicate that such gene product possess an attribute, i.e. that it has a particular function; is involved in a certain process or is located on a cellular component. Then, the most appropriate GO term to reflect such attribute (from the most up-to-date version of the GO graph at the time) is assigned to the gene and annotated with the evidence supporting it (**Figure** 1.2).

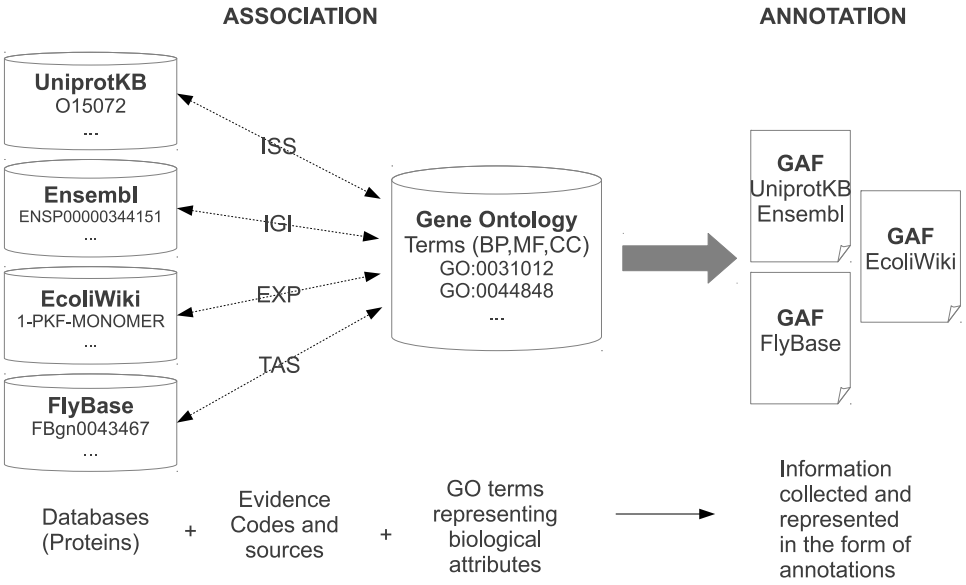



Figure 1.2: Schematic representation of the protocol followed to generate gene annotations. Different members of the GO Consortium link proteins stored in their databases with the GO terms that best reflect their biological attributes (based on certain evidence) and store the relationships in annotation files.

An evidence code is also incorporated into the annotation to indicate if the source is based on experimental or computational evidence or from a statement made by an author or curator(**Table** 1.1). Additionally, the

qualifiers "NOT", "colocalizes_with", or "contributes_to" can be added in the annotation to modify its interpretation.

Table 1.1: Evidence codes used in GOA files.

| **Evidence Codes** | | |
|---|---|---|
| Reviewed by a curator | **Experimental source** | |
| | EXP | Inferred from experiment |
| | IDA | Inferred from direct assay |
| | IPI | Inferred from physical interaction |
| | IMP | Inferred from mutant phenotype |
| | IGI | Inferred from genetic interaction |
| | IEP | Inferred from expression pattern |
| | **Computational source** | |
| | ISS | Inferred from sequence or structural similarity |
| | ISO | Inferred from sequence orthology |
| | ISA | Inferred from sequence alignment |
| | ISM | Inferred from sequence model |
| | IGC | Inferred from genomic context |
| | RCA | Inferred from reviewed computational analysis |
| | **Author statements** | |
| | TAS | Traceable author statement |
| | NAS | Non traceable author statement |
| | **Curator Statements** | |
| | IC | Inferred by curator |
| | ND | No biological data available |
| Obsolete | NR | Not recorded |
| | **Electronic source** | |
| Not reviewed | IEA | Inferred from electronic annotation |

Users can browse the annotations online through website tools provided by the GOC such as "AmiGO" [7], "QuickGO" [8], or retrieve the information from the annotation files that can be downloaded. Third-party tools and sources like "NCBI Gene" [9] are also used to retrieve the information, although these are not necessarily synchronized and updated with the most up-to-date version of GO or GOA.

The Gene Association File (GAF) is the primary format created by GOC and has had two versions: GAF1.0 (deprecated as of June 2010) and GAF2.0 (July 2010-current)[10]. A detailed description of each for-

mat and their differences can be found (**Table** 1.2). There are important differences between these two versions that most studies assessing annotation history do not address, specially in the protocol used to identify genes and gene products. This is crucial to interpret annotations for each gene / gene product. However, most assessments or tools only consider one of the two versions or do not to take into account the differences between such formats. While this thesis was being developed, in 2013, a new format was introduced: the Gene Product Association Data (GPAD) file format. This format is a simplified version that only contains annotation data without the information about the gene product (gene names or synonyms) and it was proposed as a "more normalized version" that can be used across databases (http://geneontology.org/page/gene-product-association-data-gpad-format). If one aims to collect the gene product information, other formats such as the Gene Product Information files (GPI) were created for this task.

Table 1.2: Attributes of a GO Annotation.

| Content | Description | GAF1.0 (2001-2010) | GAF2.0 (2010-current) | Entry Examples |
|---|---|---|---|---|
| 1. DB | Source of the Object ID | Pre-merge stage: Uniprot and Ensembl annotations incorporated in one GOA file | Mostly UniProtKB is used | UniProtKB, SGD,Ensembl |
| 2. DB Object ID | Unique identifier for a gene product. | Able to refer to particular protein isoforms or post-translationally cleaved or modified proteins | A top-level primary gene/gene product ID. Isoforms no longer valid. | O15072,S000038306, 1-PFK-MONOMER, FBgn0043467 |
| 3. DB Object Symbol | A symbol/ORF name to which the DB Object ID is matched. | present | present | ADAMTS3,FruK, COX1,064Ya ,14-3-3epsilon |
| 4. Qualifier | Flags that modify the interpretation of the annotation. | present | present | NOT, contributes to, co-localizes with |
| 5. GO ID | GO term ID attributed to the DB Object ID. | present | present | GO:0031012 |
| 6. DB Reference | Source of the attribution (literature,database or computational reference). | present | present | PMID:22261194, FB:FBrf0174215, SGD_REF:S000050955 |

Continued on next page

**Table 1.2 – continued from previous page**

| Content | Description | GAF1.0 | GAF2.0 | Entry Examples |
|---|---|---|---|---|
| 7. Evidence Code | Indicate how the annotation to the GO term is supported. | present | present | TAS, EXP,IGI |
| 8. With or From | Other gene products to which the annotated gene product is similar or interacts with. | present | present | UniProtKB-SubCell:SL-0039 |
| 9. Aspect | Refers to the Ontology to which the GO term ID belongs. | present | present | C,F,P |
| 10. DB Object Name | Name of the gene/gene product. | present | present | sonic hedgehog |
| 11. DB object synonym | Alternative gene symbols or previous gene product identifiers associated to the DB Object ID. | previous DB Object IDs would be gradually incorporated | many that were present in GAF1 editions were removed | DPS1_MOUSE\|Pdss1\|Dps1\|Sps1\|Tprt\|IPI00123984\|B8JJW9\|Q9WU69 |
| 12. DB Object type | Used to describe if the product is a gene, transcript, protein or functional RNA. | present | present | protein,gene |
| 13. Taxon | The taxonomic identifier of the organism encoding the gene product. | present | present | taxon:9606 |

Continued on next page

**Table 1.2 – continued from previous page**

| Content | Description | GAF1.0 | GAF2.0 | Entry Examples |
|---|---|---|---|---|
| 14. Date | The date on which the annotation was submitted into the database (not the date of the GOA file) | present | present | 20120228 |
| 15. Assigned by | The database that made the annotation. Can differ from DB (column 1) | present | present | BHF-UCL;MGI; UniProtKB; InterPro; RefGenome |
| 16. Annotation extension | Contains cross references to other ontologies (Cell Type Ontology), targets of processes/functions to indicate gene products/chemicals involved. | no | present | part_of (UBERON:0002084); acts_on_population_of(CL:0000100); has_regulation_target (MGI:MGI:107364); occurs_in(CL:0000057) |
| 17. Gene Product Form ID | Annotate specific variants of the gene product used at the DB Object ID(differential splicing, post-translational cleavage or post-translational modifications) | no | present | UniProtKB: A5YKK6-2 |

Most of the annotations in GOA files are derived from computational sources (**Figure** 1.3). Mostly, because the ratio of scientific discovery or publications available largely exceeds the amount of information that can be curated and annotated. To increase the "coverage" of gene products, many sources are constantly pooled together for this task. Many of those inferences are based on the assumption that a marked similarity exists between two proteins through evolution (duplication or speciation) from the same ancestral sequence (homology).

Features that are commonly used for this type of assignments include: 1) structure similarity (ISS); 2) sequence similarity (ISO, ISA, ISM, IKR); 3) protein profiles and phylogenetic relationships (IBA, IBD, IRD, IGC); 4) supervised machine learning algorithms (based on features from protein sequences (ISM)) or 5) high throughput studies (RCA). However, when the association is generated electronically but hasn't been reviewed by a curator, the evidence code IEA is assigned.

Computationally-inferred annotations are often considered to have limitations in their reliability compared to human-curated evidence [11]. Errors can arise when, for example, proteins have high sequence similarity but different functions or when they possess a similar function but their sequences are highly divergent. Some of these cases have been identified in the curation process and can be identified in the GOA files with the NOT qualifier and the evidence code IKR, which is characterized by the lack of key sequence residues. It is important to consider that some of such cases are likely to be present in inferred annotations but haven't been revised. Another problem arises in determining the level of GO term granularity that can be assigned to an annotation based on similarity alone [12]. Hence, it is common to find broad GO terms assigned in the annotations, which do not provide insight from a biological perspective, specially when root terms are assigned.

Despite these limitations, annotations are being computationally generated for more than 483,000 taxonomic groups (according to UniProtKB).

The GOC has grown considerably since its foundation [13] and is currently integrated by 32 institutions, specialist groups and major resources, all of which participate collectively in the evolution and implementation of GO and GOA.

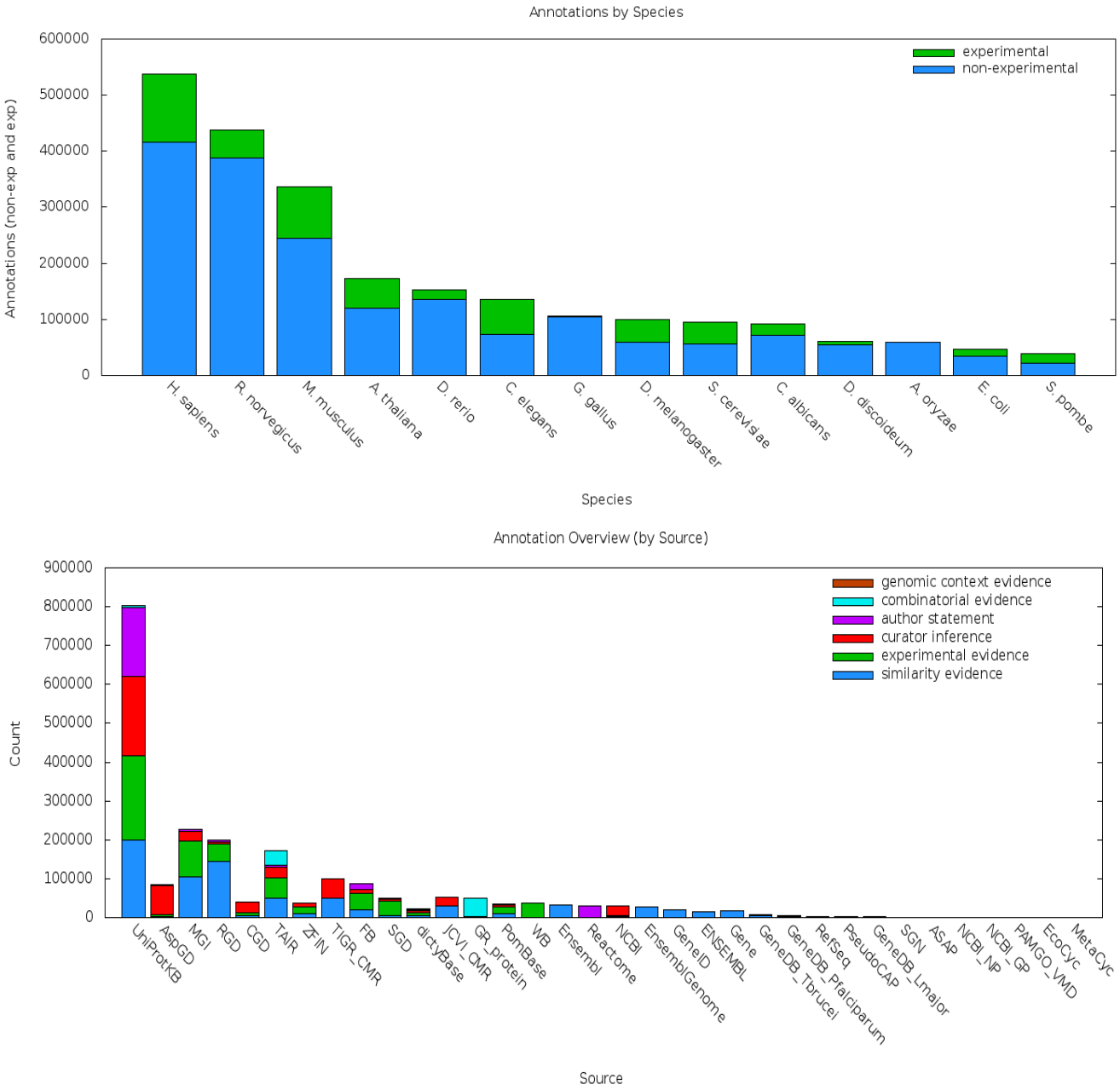



Figure 1.3: GO annotation overview. Figures highlight the number of annotations that are non-experimental compared to the number of experimental annotations across all species. UniProtKB is the largest source of GO annotations [1] (**Figure taken from**: http://geneontology.org/page/current-go-statistics. August, 2014.)

Each institution or resource generates species-specific annotations and is responsible to update them when a change is made to the annotation protocols or to the GO structure. However, there are cases where the resource that makes the annotation differs from the institution that provides support in the long-term. Such cases can be identified in the GAF files (with the DB and the Assigned By attributes). Likewise, when the research communities for certain model species do not have an established group that commits to the long-term maintenance, the annotations are done by collaborations through the UniProtKB-GO Annotation (UniProtKB-GOA) multi-species group.

Hence, some resources might have a larger or faster curation effort, or might have internal changes in their protocols that affect the annotations they handle. Together, such differences can influence species-specific annotation biases.

The GOA project is predominantly supported by the database UniProtKB, considered the largest source of protein knowledge, with over 80,370,243 entries of protein sequences (**Figure** 1.3).
These entries are derived from multiple sources and are classified in 2 sections:

- UniProtKB/SwissProt: This protein sequence database comprises high quality, manually reviewed and non-redundant entries and is continuously revised and updated. Each entry contains information about one or more protein sequences derived from the same gene to avoid redundancy. Often, entries that are present in the UniProtKB/TrEMBL database are revised and integrated into the corresponding UniProtKB/SwissProt entry. As of July 2014, 546,000 entries for 498,088 species can be found on this database. Most of those entries were inferred from homology(70%) or have evidence at the protein or transcript level (26%). The rest are classified as predicted or putative (http://www.uniprot.org/statistics/Swiss-Prot) [14] (**Figure** 1.4).

- UniProtKB/TrEMBL: This protein sequence database contains all the sequences that are not yet present in UniProtKB/SwissProt. These

sequences are derived from public databases, such as EMBL, GenBank or DDBJ but haven't been revised. Over 130 databases have also been cross-referenced. As of July 2014, 79,824,243 entries integrate this database. Most of them are bacterial (82%), a smaller proportion are eukaryotic (14%) and the rest are from archaeal or viral origin (5%). Almost 76% of these entries have been predicted or inferred by homology (23%). Only a small proportion has evidence at the transcript level (1.18%) or at the protein level (0.58%). For each and all of these entries, automatically inferred annotations are also assigned (http://www.uniprot.org/statistics/TrEMBL). As mentioned above, when UniProtKB/TrEMBL entries are revised, they are often merged to a matching UniProtKB/SwissProt entry [14].

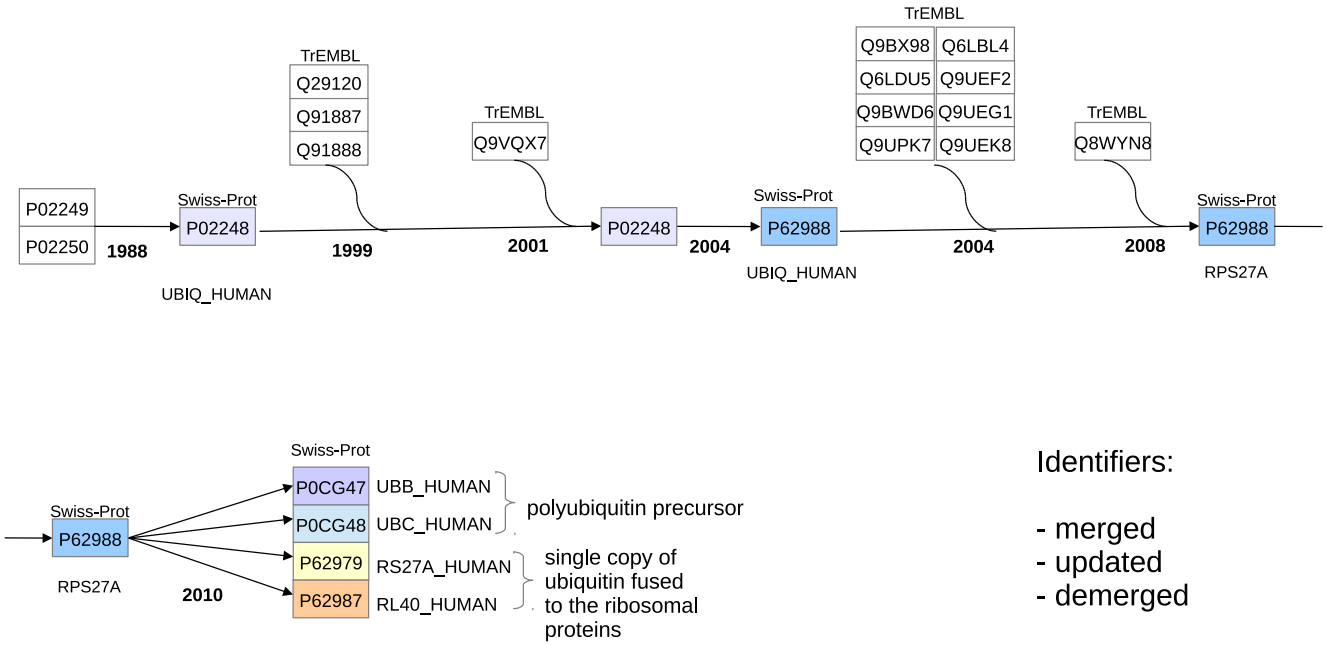



Figure 1.4: Illustration of changes in UniProtKB entries over time. The figure exemplifies a typical process of revision and upgrades in entries from the UniProtKB database. UniProtKB/TrEMBL entries that have been revised at particular time points are merged into a matching UniProtKB/SwissProt entry. Likewise, UniProtKB/SwissProt entries are revised and updated. In this example, two entries for the "ubiquitin" protein sequences were available back in 1988. The redundancy was eliminated and only one new UniProtKB/SwissProt ID remained. UniProtKB/TrEMBL entries whose sequences were derived from the same gene were gradually merged. In 2010, four protein sequences were identified to come from different genes, so the entry representing "ubiquitin" demerged into 4 new UniProtKB/SwissProt entries.

## 1.3 Uses, Challenges and Assessments of GO

The current GO and GOA structure do not aim to cover aspects relevant to mutants or diseases, attributes of sequences, protein-protein interactions, anatomical or histological information or any feature that is context-dependant (environmental). Also, the annotation format, reflects a functional "independence" between gene products [3], but in reality, the gene products can interact and participate collaboratively in different pathways. Additionally, the incompleteness of the annotations is a concern among the community [15]. Despite this, the usage of GO and GOA for the interpretation of biological data is continuously growing (as observed from querying the Gene Ontology using PubMed Discovery tools: Results by Year graph) [16].

The increase in the number of publications using GO is in part due to the challenge that scientists have had (ever since microarrays became available [17]) to interpret the large volume of data generated from high-throughput technologies. In a typical setting, researchers compare experimental conditions and generate a list of differentially expressed (DE) genes. To extract meaning from those long lists, features that are common among them are searched by using gene set enrichment analysis tools [1]. As such tools often base their results on the biological information captured in a particular GOA version, the quality of the annotations acquires even more relevance.

The first exploration was made by Lord *et al* in 2003 [18]. The authors looked at the quality of GO annotations indirectly by assessing the validity of using semantic similarity to compare proteins annotated in the SwissProt database at that time. The validation of their study was based on the hypothesis that proteins with a certain sequence similarity would have similar annotations and that the quality of the evidence codes assigned should be comparable. In their study, they found that some GO annotations were

[1] As of August 2014, running a PubMed query with the terms "enrichment analysis/-analyses" and "gene set/gene-set" would result in 2553 related publications.

incorrect or inconsistent and was thus reflected in a reduced semantic similarity score. After grouping sequences that were similar (based on BLAST searches) and comparing the corresponding similarity scores, they observed that annotations with a TAS evidence code assigned would tend to increase in "similarity" compared to others.

A year later, annotation quality was explored in bacterial and archaeal genomes. Several genome annotation inconsistencies were also found, challenging the common misconception from users that "reliable annotations" could be obtained from sources like EMBL or GenBank[19]. Afterwards, several other groups described similar inconsistencies for other organisms and databases, regardless of their origin (automatically generated or manually curated). This issue further highlighted the need for standardized annotation protocols between research groups [20–22].

Further more, an estimation by Baumgartner *et al* showed that the speed of manual curation at that time point was not sufficient to complete the annotation of even the most important model organisms [23], extending the problematic not only qualitatively, but also quantitatively.

The importance of assessing annotation quality was recognized and different groups started to propose metrics. For example, Buza *et al*, suggested a quality score based on two features: 1) the level of detail (depth) of the annotation, by considering the longest path from the term to its root node and 2)the evidence code used for the annotation, in which the authors assigned arbitrary rankings to the evidence codes. Then, for assessing the "overall quality" for a gene product, the authors proposed to sum all the individual scores for each one of its annotations [24]. The "quality score" proposed had the limitation that both parameters are quite subjective. The length of each path in the graph does not necessarily reflect its specificity and an arbitrary ranking of evidence codes does not necessarily reflect the quality of the source.

A second metric was proposed by Gross *et al* in 2009. They proposed

that the "quality score" for an annotation should be based on five parameters: 1) how many times the evidence codes assigned to the annotation changed across editions (quality); 2) how many editions have been created since the annotation first appeared (age of the annotation); 3) the number of editions where the annotation is present (existence); and 4) by considering previous editions (without the current one), the "stability" could be measured by the number of editions where the annotation remained with the same quality with respect to its existence. Finally, a "combined stability" would be assigned per annotation, which is basically the minimum score obtained in either the "existence" or the "quality" [25].

Gross *et al* do explore (although indirectly) the effects that changes in the ontology structure have on the annotations across editions. However, the authors did not make clear whether they considered the properties of the GOA files. In particular, a specific association (gene product-GO term) can be incorporated in multiple rows in just one GOA file, specially when multiple sources supporting the relationship exist. I raise this concern because they do not trace the source of the annotation, but only the evidence code, so the possibility of considering an annotation "unstable" by mistake is present. Furhermore, they are unable to assess if the evidence code changed for a "better option", as they only quantify the number of times it changed (and removed annotations that had the evidence codes ND or NR from the analyses).

Gross *et al* concluded that annotations derived from Ensembl were not paired with their corresponding GO releases, using often an older version and that in general, Ensembl annotations were more unstable than those derived from SwissProt. However, they did not explore the changes/updates that tend to occur in the accession numbers assigned to SwissProt entries, with the potential risk of losing track of the gene products. It is important to note that, back in 2009, Ensembl annotations could be distinguished from Uniprot ones even if they were both integrated in just one GOA file; but these databases are now merged in the UniProtKB, so the conclusions de-

rived from the "source of the data" cannot be re-explored in current versions.

Just after such assessments were made, the GOC introduced the GO reference Genome Annotation project, implementing more rigorous annotation protocols. Since then, existing anotations are being revised and replaced with more specific experimental codes. Thus, the GOC acknowledged that the changes in the evidence codes assigned should not be considered statements of the quality of the annotation, specially as some methods or references may have a higher confidence or specificity than others. For example, previous annotations would often be assigned with EXP, which is the parental code for IDA, IPI, IMP, IGI and IEP. However, curators were encouraged to revise old annotations with such code and replace them with children codes of higher specificity [26].

Changes in the GO structure took place as more emphasis was put on the assessment on the impact of changes of GO and GOA in applications like enrichment analysis. In particular, Alterovitz *et al* (2010) proposed modifications to the GO because they identified terms misplaced within the graph that affected the results of enrichment analyses. Such modifications were discussed with the GOC and incorporated afterwards [27]. The quality of computationally inferred annotations also seemed to improve after such changes were made [11].

Some members of the GOC also introduced annotation efforts focused towards prioritized gene sets, and the EMBL-EBI explored the impact that such prioritization had on gene set enrichment results. In particular, they observed that more GO groups where such genes belonged could be retrieved [28]. An independent assessment by Clarke *et al* (2013) also highlighted that changes in GOA versions had a larger impact (compared to the changes in GO structure alone) in the reproducibility of the results of enrichment analyses over time [29].

Changes in GO/GOA also affect widely used tools for enrichment anal-

yses, such as GSEA [30] or DAVID[31]. The tool developers have to keep up and follow the recommendation from the GOC to use the latest version of GOA available [32], but in many cases, the have not. Hence, users run their analyses on annotations that are considerably outdated. Even if users acknowledge the situation, they often forget to cite or check the version for interpreting their findings or future references [32].

A different set of tool-related problems arise when they fail to remove negative associations (those with a "NOT" qualifier) [33]; do not consider the same protocol to map gene identifiers, sources, type of relationships within GO, incorporate robust statistical analyses or correct for data artifacts [34]. For example, GSEA [30] considers "regulates" relationships in the GO structure within their analyses, whereas ErmineJ [35] only considers "is_a" or "part_of" relationships and also takes into account artifacts such as gene multifunctionality (i.e., genes which have multiple functions, reflected as the number of GO terms that have been assigned to them). This is particularly relevant when multifunctional genes are often retrieved from the results, but are not necessarily related to the question of interest [36].

Another missing gap in the assessment of GO annotation quality was partially filled in 2013 when Gillis and Pavlidis explored the stability of GO annotations by measuring how genes can lose their "functional identity". In particular, they expressed functional identity in terms of how semantically similar a gene's annotations were across editions. If a gene was most semantically similar to its previous incarnations in the GOA, compared to other genes, then it was considered to retain its "functional identity". Loss of functional identity is expected as annotations are added, but the rate of this loss had not been previously evaluated. They found that at least 20% of the genes can lose their identity after 2 years. They also characterized a circularity problem, where the same publications are used to support protein interaction databases and GO annotations, affecting the applicability of protein-protein interactions for gene function prediction [37].

Parallel to the usage of GOA for enrichment analyses and despite the challenges mentioned above, these have also been used in algorithms for gene function prediction and their assessment. For those genes whose attributes are not known or haven't been processed by curators, the challenge relies on predicting their "function", especially because the experimental investigation is limited and costly. However, the issues arising from using a gold standard that is incomplete, such as GO, often makes this task more challenging. Huttenhower *et al* (2009) highlighted some of these problems while assessing how the performance of machine learning algorithms are affected in this context, but also suggested that the methods were still able to make "useful predictions" out of incomplete standards [38].

To assess the performance of function prediction algorithms, the task has been set to predict GO terms for some target genes [39–41]. Such assessments often use, as a benchmark set, recently curated annotations from a subset of those targets. For example, in the CAFA assessment, a 6 to 12 month waiting period (after the submission) is considered for the accumulation of manual GO annotations. Then, a subset of those "new" annotations are selected for the evaluation, which in turn look at which GO terms were assigned to each target gene. However, Gillis and Pavlidis (2013) criticized such task, as predicting biologically meaningful gene functions may not be equivalent to predicting GO annotations. This is particularly relevant when considering that patterns that can be attributed to the curation process have been used to predict "gene function" since 2002. To give an example, commonly co-annotated GO terms have been proposed as predictive methods [42], and even popular tools like the "GeneMANIA prediction server" utilize such patterns to weight their predictions [43].

In fact, the results from the first Critical Assessment of Automated Function Prediction (CAFA) in 2013, showed that the top performing methods incorporated existing knowledge of GO or based their algorithms on sequence similarity. As many of the existing computational annotations (either IEA or manually revised) (**Table** 1.1) are based on sequence similarity and incor-

porated into the UniProtKB annotation pipeline, many inferred annotations that have not been curated (IEA) can be considered predictions. In their results, BLAST was outperformed by what was called the naïve method, a control that assigns all the target sequences the exact same predictions based on GO term prevalence from GOA data. This suggests that something was wrong with the metric. Having a control that performs better than a popular sequence similarity method, highlights the impact that artifacts attributed to the curation process have on the performance metrics.

Gillis and Pavlidis (2013) assessed the CAFA results independently using function-centered metrics,i.e., by asking the question "which genes should be assigned to a particular function? In fact, when considering a function-centric metric, BLAST was a top performing method and many of the manually curated annotations were derived from existing electronic annotations (IEA) [37]. This result could be interpreted as an attempt to predict which targets would be curated or upgraded from existing annotations without considering any biological attribute from the targets, which seems to fall out of the scope of the actual "function prediction" task.

In conclusion, given the increased usage of GO for different scenarios and the constant changes in GO annotations, it is of interest to make a comprehensive assessment of the annotation instability and assess the impact that these changes have on current usages.

# Chapter 2

# Objectives

In this thesis, I aim to:

- Run exploratory analyses to identify trends and the evolution of GOA over time for 14 different taxa.
- Apply different metrics that can be used to learn more about the instability of particular genes and annotations, including the degree to which GO assignments are distributed unequally for each gene over time; the functional identity of each gene compared to its current state and the "existence" of an annotation considering the source of the annotation.
- Assess the impact of these changes and artifacts that can be attributed to curation efforts in applications such as the Assessment of Protein Function Prediction Algorithms and Gene set Enrichment Analysis.
- Build a visualization tool to extend this information to GO users who want to explore the instability of genes and annotations of their interest.

# Chapter 3

# Methods

The current project involved several steps. To ensure the quality of the analysis, the first part involved a careful pre-processing of the raw information and the implementation of exploratory analyses to observe changes in the annotation data. The second part involved the design and creation of a database and a website to visualize and make this information available. The third part consisted of the exploration of the impact that changes in the annotation data have on common settings (**Figure** 3.1).

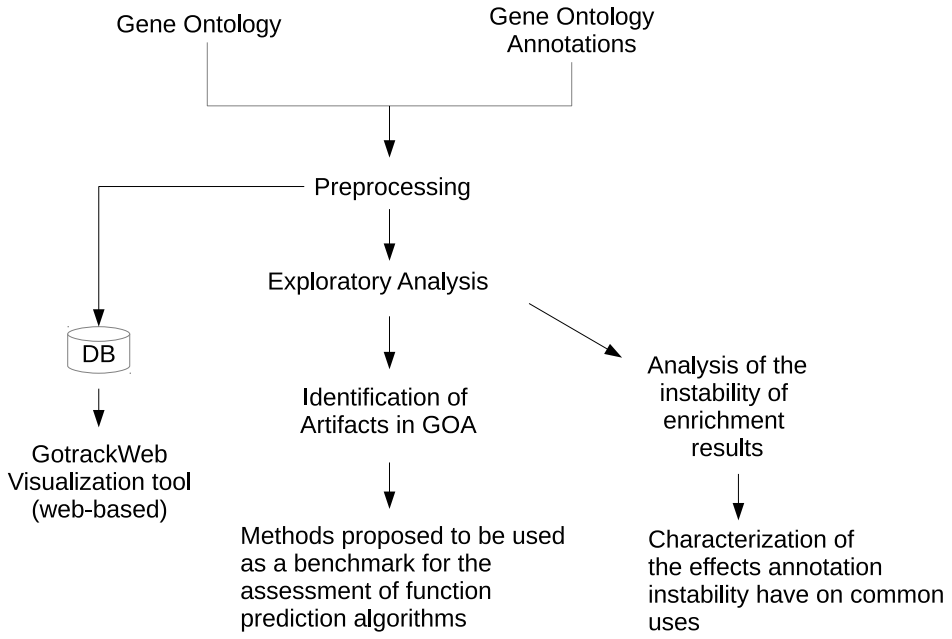



Figure 3.1: General overview of the methods and analyses done in the present study.

## 3.1 GOtrack: Pre-processing and Analysis

In this section, I will describe the methods used at each step, the database built to store this information and the web-based tool that was implemented to make the data accessible to other users.

### 3.1.1 Data Collection

To collect all the historical annotations available for each one of the 14 species considered, monthly releases of Gene Association Files ("GOA") in GAF1.0 and GAF2.0 formats were retrieved from the EMBL-EBI FTP website for: *Arabidopsis thaliana* (thale cress), *Gallus gallus* (chicken), *Bos taurus* (cow), *Dictyostelium discoideum* (slime mold), *Canis familiaris* (dog), *Drosophila melanogaster* (fruit fly), *Homo sapiens* (human), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Sus scrofa* (pig), *Danio rerio* (zebrafish), *Caenorhabditis elegans* (worm) and *Saccharomyces cerevisiae* (yeast) [44]. As the EBI repository only contains annotations for yeast and fruit fly with the date stamp from 2011 until now, earlier versions of GOA files for these organisms were retrieved from: FlyBase [45] (2006-2011), SGD [46] (2001-2004) and SGD [47] (2005-2010). The editions (or versions) were ordered and renumbered consecutively based on the release date. Data for *Escherichia coli* was solely retrieved from EcoCyc [48].

To match the annotations with the corresponding versions of the GO graph at each time point, monthly releases of the core version of the Gene Ontology database (termdb-xml files) were collected from the GO repository [49]. Each GOA file was paired with their respective GO version by using their release dates and the date embedded on each termdb file name. In cases where the GOA date did not have a matching termdb file, an earlier version of the termdb was considered. The purpose of such matching was to infer parental terms in the GO hierarchy for each GO annotation, considering only "is_a" and "part_of" relationships and excluding root and obsolete terms.

### 3.1.2 ID Mapping

As described in the introduction, each GOA file can be created and supported by different databases. In particular, those created with the GAF1.0 format had database-specific accession codes for each gene product (DB Object IDs). When the GAF2.0 format was introduced, some DB Object IDs remained the same, but some others were merged, demerged, deleted, replaced or mapped to their equivalents in the UniProtKB DB Object ID version. These changes were implemented at different time points for each species and some others, like SGD, had internal changes in their internal DB identifiers even before the GAF format changed. As these changes are of major importance to track the historical annotations of each gene product, I implemented a procedure attempting to map the identifiers in a robust manner was implemented (**Figure** 3.2 and **Script** 5.5).

The procedure considered the information retrieval and integration from different sources:

- "Mapping Files" provided by the UniProtKB which map a list of identifiers from external databases to UniProtKB accession IDs[50].
- Three custom dictionaries created with the information currently available in the Uniprot Documentation for E.coli, yeast and fruit fly. The dictionaries map Uniprot/SwissProt entries with gene designations, ordered locus names, SwissProt primary accession numbers, entry names and cross-reference accession numbers to the original accession IDs assigned from EcoliWiki, SGD and Flybase, respectively.
- A custom dictionary to track Uniprot accession numbers that were once "primary" accessions and latter became "secondary" because of a merging or demerging event. Before 2010, when the transition period from the GAF1.0 to GAF2.0 format, the secondary IDs would be normally incorporated as "synonyms" in the annotations. In subsequent GAF2.0 format files, these "synonyms" were removed from the annotations.

The last GOA edition for each species where these secondary accession numbers were found as synonyms were selected for the creation of another custon dictionary: (Human (edition 105); Arabidopsis (edition 56); Chicken (edition 53); Cow (edition 46); Mouse (edition 69); E.coli (edition 95); Fly (edition 30); Rat (edition 72); Dictyostelium (edition 25); Dog (edition 25); Zebrafish (edition 57); Worm (edition 25)). The information stored on "DB Object Synonym" was collected and mapped to its primary "DB Object ID".

- Automated queries to the UniProtKB website were also implemented to maintain the UniProtKB DB Object IDs (primary accession numbers) as updated as possible [2].

Some protein sequences and their corresponding accession numbers are deleted from UniProtKB and disappear in subsequent GOA files [3]. The deletions occur when entries correspond to open reading frames (ORFs) or pseudo genes wrongly predicted to code for proteins [14](`http://www.uniprot.org/help/deleted_accessions`).

A mapping procedure was also implemented to track "DB reference objects" (sources of annotation, mostly publications) from old GOA files (GAF1.0 format). Earlier versions were found to incorporate obsolete MEDLINE IDs that in subsequent editions were replaced for PubMed IDs. The mapping files used for this mapping process were retrieved through the National Library of Medicine [51] (**Script**: 5.6).

Finally, old DB Object IDs and DB Reference Objects were updated in the GOA files for further analyses.

[2] Changes between primary and secondary accession numbers can also be found on `ftp://ftp.uniprot.org/pub/databases/uniprot/knowledgebase/docs/sec_ac.txt`

[3] `www.uniprot.org/faq/11`

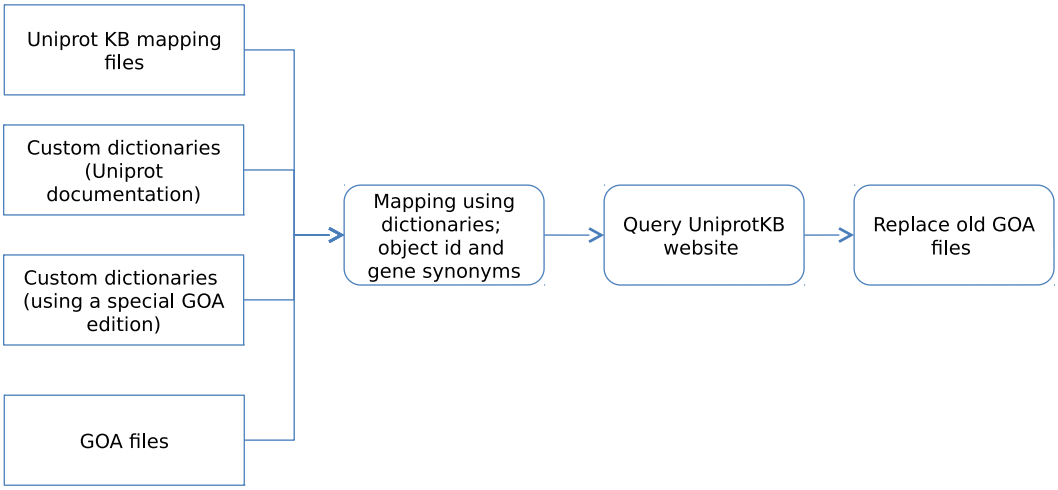



Figure 3.2: General schema of the mapping procedure.

The analysis by Gillis and Pavlidis (2013) only considered gene products that were consistently present in all the GOA file versions. No mapping procedure was implemented and the instability of the identifiers was not considered.

### 3.1.3 Exploratory Analyses

After mapping all the identifiers, I aimed to consider as many gene products as possible, but discarding those that were considered "mistakes", which were disappearing from the GOA over time. For this reason, a series of lists were generated to identify:

- **All Terms**: All the GO terms that have been used at least once across GOA editions,
- **Terms Always Present**: GO terms that were present across all GOA editions,
- **All Genes**: All the DB Object IDs that have been used at least once (after mapping) across GOA editions,

- **Genes Almost Always Present**: DB Object IDs representing "genes" that are almost always annotated across GOA editions (user defines threshold. Analyses were run to trace the annotations of gene products that are present in at least 85% of the GOA editions).

The implementation gave the option to focus on gene products always present, but including those that might not have been always made it even more flexible and powerful. Therefore, in the analysis I considered "genes products almost always present" and a threshold can be set up when running the analysis. In particular, I considered "gene products almost always present" if they were present at least in 85% of the GOA editions.

A series of metrics were implemented to assess the instability of the annotations for the gene products "almost always present" (**Figures** 3.3 and 3.4, **Scripts** 5.1 and 5.2):

- **Semantic similarity**: For each GOA edition, a hash table (an associative array called gomatrix) was implemented to trace the GO terms directly associated for those gene products "almost always present". Then, an assessment of how "functionally similar" each one of these gene products is to itself was conducted by comparing the gomatrices from previous editions vs. the current one (**Scripts** 5.3 and 5.4).

- **Multifunctionality**: Multifunctional genes in the last edition were identified and ranked per species. Likewise, a multifunctionality score for each gene product (using ErmineJ) was also calculated per GOA edition. Gene products that have been prioritized for annotation by the GOA tend to have more GO terms assigned. This metric does not reflect that certain gene products are more biologically relevant than others, but that they tend to be more studied or annotated. Difference among gene products in their multifunctionality have an important effect [36]. The score is the Area Under the Receiver Operating Characteristic Curve(AUROC, a comparison of the true positive rate and false positive rate at various threshold settings) obtained by

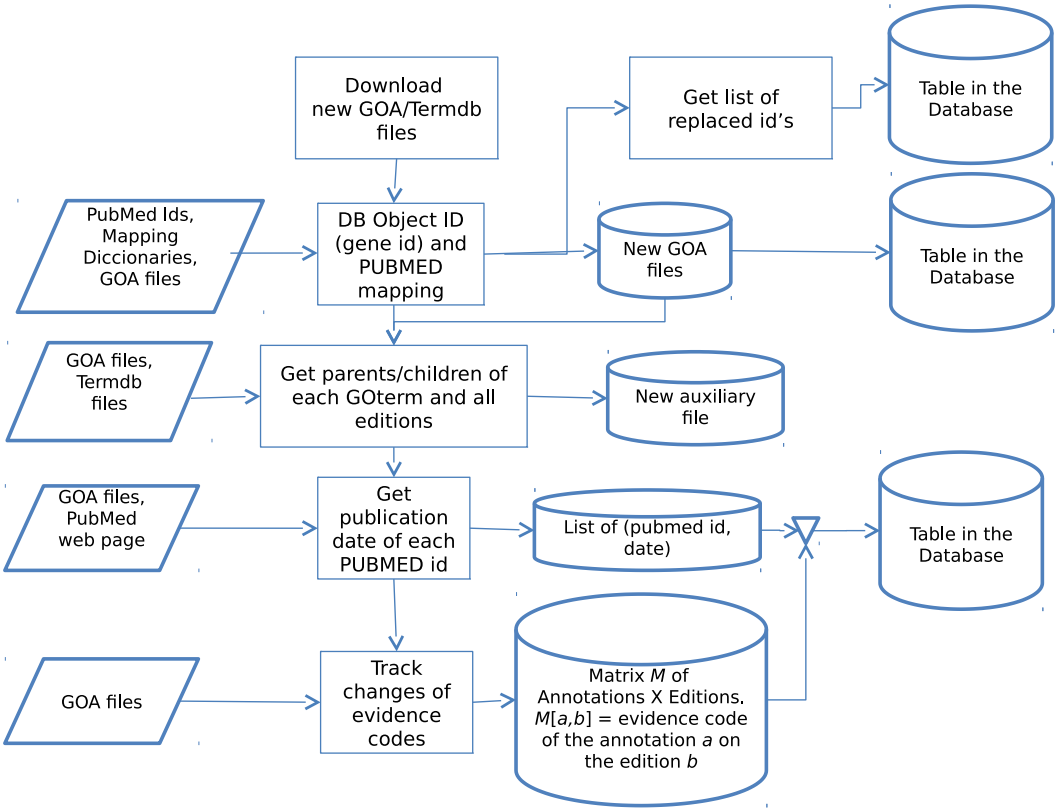



Figure 3.3: General GOtrack pipeline to pre-process the data for any species.

comparing the genes that are members of a GO group to the ranking provided by the "GO term membership".

- **GO term membership**: A way to assess GO term "popularity" is to assess how many gene products have been associated to each GO term over time. This metric can also be interpreted as how prevalent a GO term is on GOA at each time point. This quantitative measure can also indirectly reflect when the terms are incorporated or discarded from the graph or when the GOC decided the term was no longer suitable for annotations.

- **Source Instability**: If the sources of annotation are robust enough to

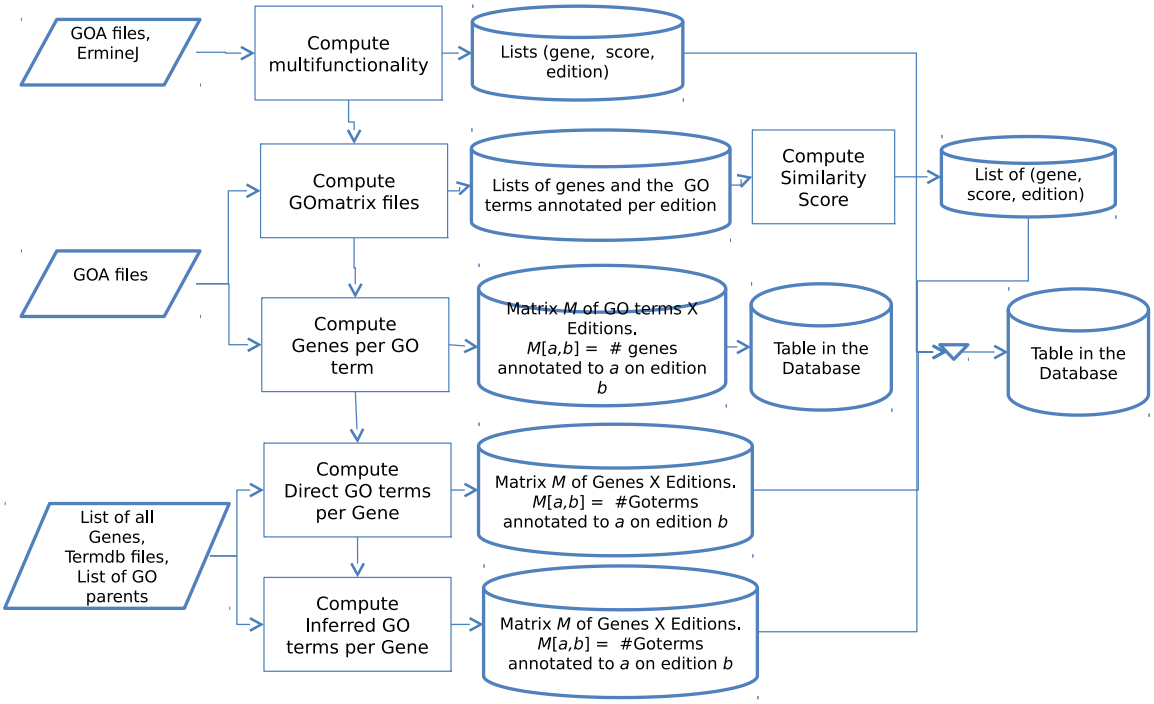



Figure 3.4: General GOtrack pipeline for exploratory analyses.

support an association (such as experimental publications) then, even if the GO terms assigned to reflect a determined "function" across annotations change, these sources should remain linked to each gene product across editions. To explore this hypothesis and assess if there is also an instability in terms of sources used, a "Publication history" analysis was made by linking the gene product with a publication ID (supporting at least one of its annotations), among with the release date of such paper and tracing their connection across all editions. Hence, one can observe when the source was first used and when it was discarded if that is the case. Tracing the date of publication is also useful to visualize the age of the sources supporting GOA on a global scale.

- **Evidence code Instability**: The usage of evidence codes to reflect

the source for an annotation has changed over time. Since the GO Reference Genome annotation effort was established, annotations have been revised to assign more specific experimental codes to annotations.The guide to best practices for GO manual annotation also suggests that annotations that had TAS codes should be replaced with those that reflect published experimental results [52]. Hence, evidence codes assigned to each annotation (gene product + GO term + PubMed ID) were traced across editions ("Evidence code history") to visualize such changes on a case-by-case basis.

- **Number of direct GO terms annotated per gene product**: The total count of GO terms directly annotated to each gene product was traced to assess quantitatively if it gains or lose terms over time.

- **Number of propagated GO terms**: The total count of GO terms that can be inferred from those directly annotated to each gene product was traced. This metric is used to assess if the gene product gains or lose terms over time because of changes in the GO structure. The propagation was made using ErmineJ and only "is_a" and "part_of" relationships were considered.

- **Number of promoted annotations**: When annotations are revised by curators, they can disappear, but one of the reasons is because the GO terms assigned are replaced with a more granular GO term that better reflects or supports the association. Annotations whose original GO terms were "promoted" to a more granular (children term) were identified and counted across editions (**Figure** 3.5).

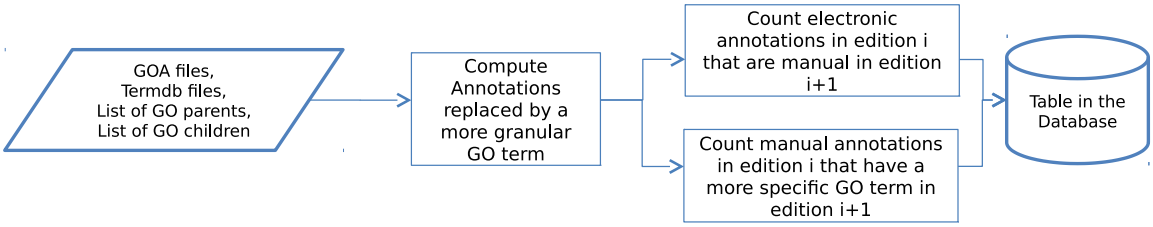



Figure 3.5: General overview to track commonly upgraded annotations.

### 3.1.4 Database Design, Creation and Management

A database was created to store annotation data and retrieve the information (**Figure** 3.6). There are tables created to store general information for all the species and tables that store specific information for each species (**Script** 5.7).
The tables that contain information for all the species are:

- **popularGenes**: Stores the query history of GOtrackWeb users across all species. It aims to provide us with an idea of the usage of GO and what genes are of popular interest.
- **edition_to_date**: Stores the release date of each GOA file per species.
- **species**: Stores a catalog of all the species analyzed.
- **GO_names**: Stores the GO term accession IDs and their corresponding GO names (human readable names) for each Termdb file over time. Therefore,if a GO term changed its name, previous names can be retrieved.
- **unique_go_functions**: Stores the relationship between the GO term accession ID with its most recent GO name.
- **avgAllSpeciesCount**: Stores for each species and edition: 1) the average number of GO terms directly annotated to the DB Object IDs (gene products); 2) the average multifunctionality score; 3) the average semantic similarity score of the DB Object IDs (gene products) in

that edition with respect to the current one; 4) the average number of parental GO terms that can be inferred from the annotations and 5) the total number of DB Object IDs (gene products) that are present in each edition.

- **annotAnalysisTab**: Stores for each DB Object IDs (gene products) and edition per species: 1) the total number of annotations that have been promoted from an IEA evidence code to a curated evidence code; 2) the total number of annotations that have been promoted to a more granular GO term; the average number of GO annotations that have negative a NOT qualifier with respect to the total number of DB Object IDs (gene products) that have at least one negative annotation.

The tables that contain information for each species are:

- **species_gene_annot**: Stores the information from the GOA files: DB Object IDs (gene products), GO term, evidence code, PubMed ID, taxon, DB Object symbol, GO term name, Ontology.

- **species_replaced_id**: Stores the relationships between the original DB Object IDs assigned to the annotations and the new DB Object IDs (gene products) that were replaced with during the mapping process.

- **species_evidence_code**: A simplified version of "gene_annot" where the PubMed identifiers have been cleaned. It was created to make queries faster. Stores the information: DB Object IDs (gene products) GO term, PubMed ID, evidence code, edition.

- **species_count**: Stores the pre-processed information for each gene and edition. Contains: DB Object symbol, total number of GO terms directly annotated, total number of GO terms inferred, the multifunctionality score, semantic similarity (Jaccard) score.

- **species_gene_per_go**: Stores information about how many DB Object IDs (gene products) belong to a particular GO term (GO term membership) per edition.

- **species_avg**: This table is created dynamically after all the information has been inserted into the database. It stores the average value of the columns present in species_count and annotAnalysisTab for each species.

- **species_unique_gene_symbol**: This table is created dynamically from "species_gene_annot" after the data for one species has been inserted. It contains the relationship of the DB Object symbols and the DB Object IDs (gene product accession IDs).

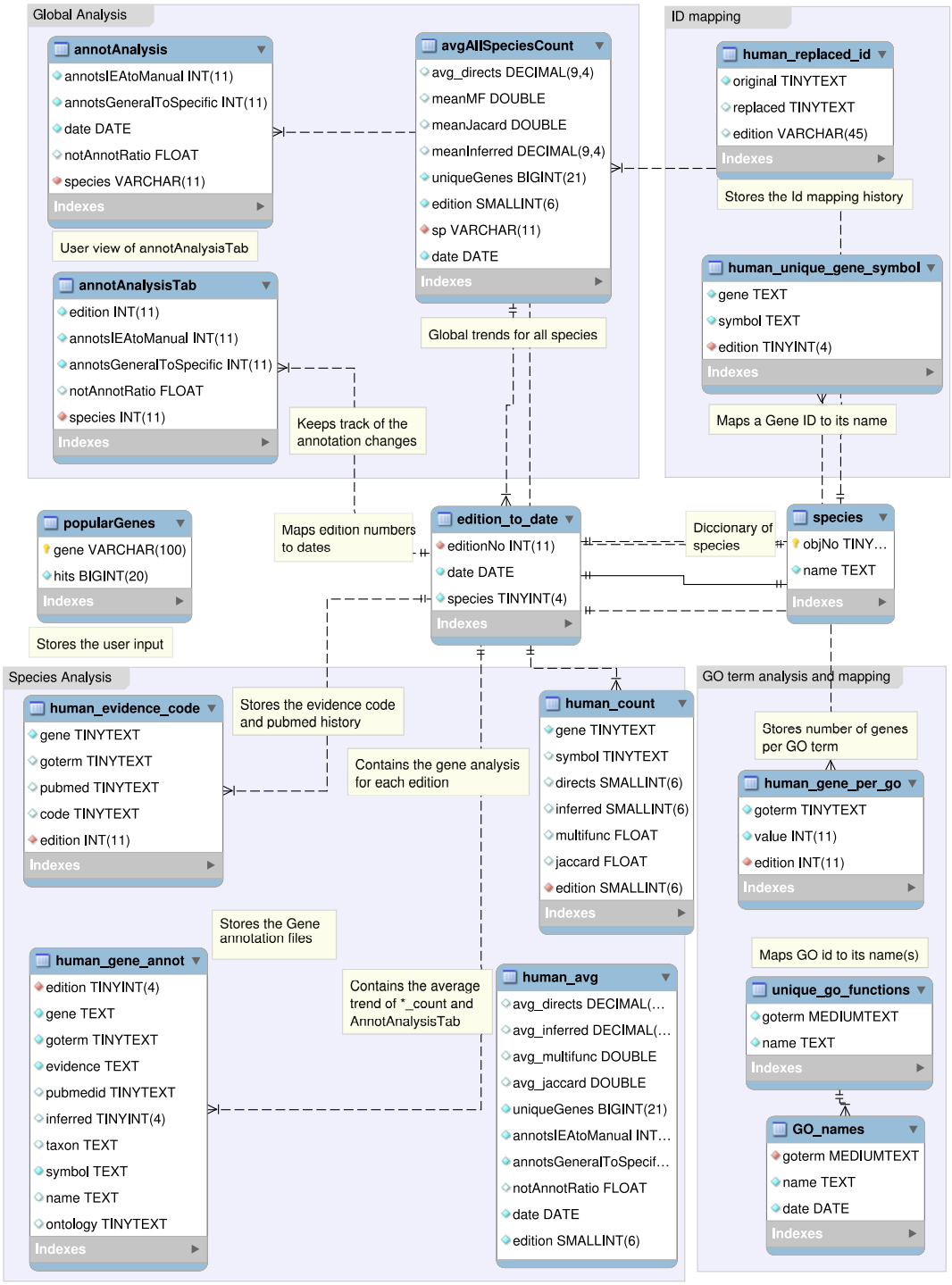



Figure 3.6: GOtrack database model. See main text for description.

### 3.1.5 Web-based Visualization Tool: GOtrackWeb

A website was designed an implemented (Figure 3.7) and is now available at: `www.chibi.ubc.ca/GOtrackWeb`.

In the main page, the top 10 queries made by the users and the top multifunctional gene products per species from the last edition available are displayed.

The main page was designed for users to query the historical information for a gene product of their interest. The query allows the use of UniProtKB accession number IDs, synonyms or or DB Object Symbols (gene symbol). If the user decides to use a Symbol, all the DB Object IDs (accession numbers) that match (whether these are UniProtKB/SwissProt or UniProtKB/TrEMBL) are retrieved. If the user queries a UniProt accession number, the specific information will be retrieved. Obsolete IDs can also be queried. If the annotations from that obsolete ID are assigned to a newer ID, the information associated to the most recent accession number is retrieved. To make the query species-specific, the user must select a particular species from the list and click on the Search button. The program on the back-end does the following:

1. Search first for the gene product (DB Object IDs) in the db table "species_unique_gene_symbol" (list1).
2. Search each element from list1 in the db table "species_repaced_id" to retrieve the most recent DB Object IDs available (list2).
3. Retrieves the data stored on table "species_count" for each element in list2.
4. Retrieves the data stored on table "species_avg" for each element in list2.
5. The DB Object IDs found for the queried gene product are displayed above linked to the Uniprot website for more information.

6. A dynamic plot using the retrieved data is generated in section "Count history", which includes the total number of inferred and direct GO terms annotated to each ID from list2 across editions, as well as its multifunctionality score and semantic similarity, which can also be compared with the average values for the queried species.

7. Update the table "popularGenes" inserting the query made by the user.

8. Retrieves the corresponding GO terms and GO names annotated to each DB Object ID found in the db tables "GO_names" and "species_gene _annot", separated by Ontology and displayed on the userRequest.xhtml web page.

The user can then explore the annotations associated to the queried gene. The user should change to the tab named "Functionality". This tab contains a browsable list of all the GO terms that have been ever assigned to that gene product,separated by Ontology. The user can select those that are of interest and click "continue". The page is redirected to functionality.xhtml with the tab "evidence code history", which displays the historical existence of the annotation. A dynamic plot is displayed, coloured by the type of evidence code used. Each row corresponds to one annotation (DB Object ID + GO term + evidence code + source (PubMed)). The user can also click on a table with link-outs to access those papers that were used to support the associations. Another tab named "GO term membership" is also incorporated. In this section, the total number of DB Object IDs (gene products) annotated to each of the GO terms selected are displayed on a dynamic plot. Alternatively, on the main website, the user can just search for a GO term ID, and only the plot with the GO term membership is displayed.

A different section on the website was created to provide a general panorama for each species. The section is called "Global Trends" and can be accessed through the top panel. Two tabs are displayed, one allowing the user to select two species for comparison. Two dynamic plots are generated.

The first one retrieves information from the table "avgAllSpeciesCount" and displays for each edition and species: the average number of direct GO terms; the average multifunctionality score; the average number of parental GO terms; the average semantic similarity score and the total number of gene products (measured by DB Object IDs found after the mapping process).

The second plot is generated from table "annotAnalysis" and displays for each edition and species: the total number of annotations that were replaced with a more specific granular GO term; the total number of electronic annotations that were revised (switching the evidence code IEA to another evidence code); and the average number of GO annotations that have a negative association (those with the NOT qualifier) relative to the total number of gene products that have at least one negative annotation in each edition.

To explore the overall historical data for one species, the information stored on "species_avg" is retrieved and displays per edition: average multifunctionality score, total number of annotations promoted from IEA to a manual evidence code and from a general to a more granular GO term, average number of direct GO terms and the total number of unique DB Object IDs (gene products). As there were limitations in the the visualization plot, to aid with the comparison of changes in the multifunctionality score with other parameters, this score was multiplied by 10,000,000.

The components used to develop GOtrack and the web-based tool were: Eclipse Juno, Maven 3.1.0, NetBeans 7.3.1, mysql-connector-java-5.1.18, Primefaces 4, JSF 2.2, Apache Tomcat and a Google Charts API (https://google-developers.appspot.com/chart/).

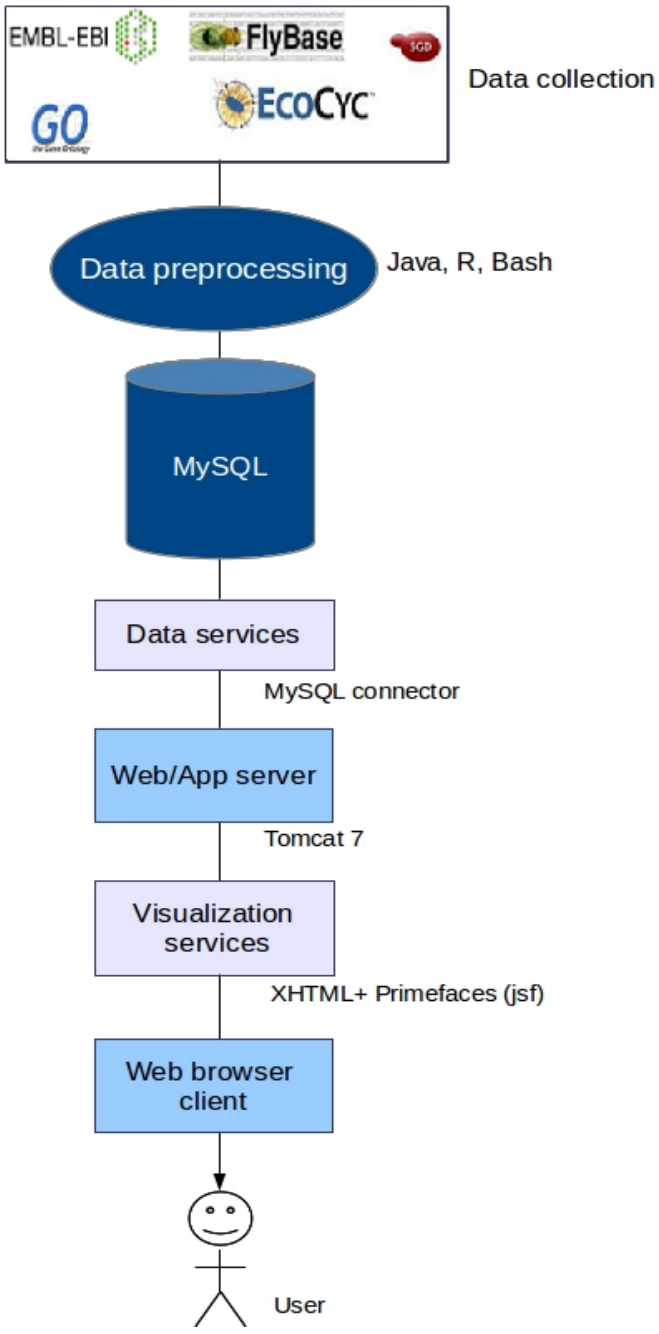



Figure 3.7: General overview of the GOtrackWeb implementation.

## 3.2 Proposing a Benchmark for the Assessment of Function Prediction Algorithms

As mentioned in the introduction, GO and GOA are used in the context of Gene function prediction and the evaluation of the performance of such algorithms. Currently, no reliable "gold standard" is available to use in this context, but GO annotations that get freshly curated are used as a benchmark set.

However, we have to consider the limitations that the annotations have for this task. The task has been defined to predict GO terms to a set of target genes. This task can be interpreted as an assessment of the participant's ability to predict the curation activity, specially as the functional information for some of the target gene products might already be published but just haven't been captured in the annotations (post-dictions). Even more, the "accumulation period" of only 6 months is a limiting step, as few actual functional discoveries would be made and annotated in the same evaluation time point. With this in mind, it is also important to consider that some curation patterns can be found in GOA, such as: GO terms frequently used, GO terms commonly upgraded or GO terms often co-annotated. Algorithms attempting to "predict" curation activity might use such patterns to increase their "performance". Gillis and Pavlidis (2013) observed that in the "state of the art" publication by CAFA, the participating algorithms do exploit this information [53]. However, such artifacts do not have any biological relevance and should be subtracted.

In the present study, we identify, use and propose these parameters as a baseline for a better assessment of function prediction algorithms. To test the actual "performance" of such artifacts, we submitted the "predictions" inferred from such patterns to the CAFA2 assessment, but as of September 2014, the results haven't been made available to the participants. In the meantime, I elaborated an independent analysis of the performance using target sequences and old predictions submitted by participating algorithms

of the CAFA1 assessment(**Script** 5.8).

### 3.2.1 Data Collection

Targets from the current (2013) and previous (2011) CAFA assessment were used as gene sets to study patterns associated to the GO curation process (**Tables** 3.1 and 3.2). GOA files were retrieved for the selected species for the CAFA (2013) assessment.

- CAFA 2013-2014 Target sequences were retrieved from: `http://biofunctionprediction.org/node/12`
- Annotations (GOA) for *A.thaliana, D.discoideum, H.sapiens, M.musculus, R.novergicus, S. cerevisiae, D. rerio, E.coli* were used from the GO-track analyses.
- *S. pombe* annotations were retrieved from: `ftp://ftp.ebi.ac.uk/pub/databases/pombase/pombe/Gene_ontology/`
- *X. laevis* annotations were retrieved from: `http://www.uniprot.org/uniprot/?query=taxonomy\%3a8355&format=*`
- *H. pylori, M. genitalium, S. enterica, P. syringae, P. putida, S. pneumonia, M. genitalium, B. subtilis* were retrieved from: `ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/`.
- *P. aeruginosa* was retrieved from: `http://cvsweb.geneontology.org/cgi-bin/cvsweb.cgi/go/gene-associations/gene_association.pseudocap.gz`
- *M. jannaschii, I. hospitalis, N. maritimus, H. salinarum, S. solfataricus, H. volcanii*: `ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/`
- *P. furiosus* from: `http://www.uniprot.org/uniprot/?query=taxonomy\%3a186497&format`

Table 3.1: Target sequences and species considered for the CAFA2 assessment.

| Species ID | Organism | Domain | No.Targets |
|---|---|---|---|
| 3702 | *Arabidopsis thaliana* | Eukarya | 12069 |
| 44689 | *Dictyostelium discoideum* | Eukarya | 4126 |
| 9606 | *Homo sapiens* | Eukarya | 20257 |
| 10090 | *Mus musculus* | Eukarya | 16613 |
| 10116 | *Rattus norvegicus* | Eukarya | 7854 |
| 559292 | *Saccharomyces cerevisiae* | Eukarya | 6621 |
| 8355 | *Xenopus laevis* | Eukarya | 3365 |
| 284812 | *Schizosaccaromyces pombe* | Eukarya | 5089 |
| 7955 | *Danio rerio* | Eukarya | 2885 |
| 7227 | *Drosophila melanogaster* | Eukarya | 3195 |
| 224308 | *Bacillus subtilis subsp. subtilis 168* | Bacteria | 4188 |
| 83333 | *Escherichia coli K12* | Bacteria | 4431 |
| 85962 | *Helicobacter pylori ATCC 700392* | Bacteria | 581 |
| 243273 | *Mycoplasma genitalium ATCC 33530* | Bacteria | 483 |
| 208964 | *Pseudomonas aeruginosa PA01* | Bacteria | 1245 |
| 160488 | *Pseudomonas putida KT2440* | Bacteria | 693 |
| 223283 | *Pseudomonas syringae pv.tomato str.DC3000* | Bacteria | 675 |
| 321314 | *Salmonella enterica* | Bacteria | 882 |
| 99287 | *Salmonella typhimurium* | Bacteria | 1771 |
| 170187 | *Streptococcus pneumoniae TIGR4* | Bacteria | 502 |
| 478009 | *Halobacterium salinarum R1* | Archaea | 267 |
| 309800 | *Haloferax volcanii DS2* | Archaea | 93 |
| 453591 | *Ignicoccus hospitalis KIN4/I* | Archaea | 125 |
| 243232 | *Methanocaldococcus jannaschii DSM 2661* | Archaea | 1787 |
| 186497 | *Pyrococcus furiosus DSM 3638* | Archaea | 480 |
| 273057 | *Sulfolobus solfataricus P2* | Archaea | 448 |
| 436308 | *Nitrosopumilus maritimus strain SCM1* | Archaea | 91 |

### 3.2.2 GO Term Prevalence in Annotation Data

The prevalence of the terms in the GOA are considered in both CAFA1 and in the present study as a baseline (naïve method or null). Some GO terms are noticeably more used than others, specially those that are "generic". Prevalence was computed using the last GOA edition available (before the CAFA2 or CAFA1 submission deadline). Annotations were propagated and the size of each GO group was determined. Prevalence was computed separately for each taxon (at the species level, except for bacteria and archaea organisms, within each of which annotations were pooled). The assessment allowed the assignment of up to 1500 predictions per target sequence. Hence,

the top 1500 most prevalent GO terms were assigned to each and all the target sequences after filtering those that were already assigned as manual annotations for each gene product.

Prevalence provided the initial scores for each target-GO term association. As simply predicting commonly used terms yields surprisingly strong performance according to Gillis and Pavlidis [53], this score would only be replaced by the IEA upgrade or the co-occurrence methods if they had a stronger prediction than prevalence.

### 3.2.3 Identification of Inferred Electronic Annotations Commonly Reviewed and Re-annotated by Curators in GOA

Electronic annotations for each species were identified within a two year interval (12-2008 to 12-2012). Posterior dates were then used to check for annotation upgrades, which could be the same GO term but with a manual evidence code assigned or updated to a children term with a manual evidence code. The data for each species was independently processed. The promotions were translated into probabilities by pooling the frequency for which term is upgraded across all the taxa.

### 3.2.4 Identification of GO Terms Frequently Co-annotated in GOA

The probability of co-occurrence of GO terms was calculated based on the conditional likelihood of getting a GO term "B" given that a gene already has GO term "A" assigned with a manual evidence code. These probabilities were calculated by pooling gene annotation data from all the taxa, considering an interval of two years before submission deadline.

Given the GO terms **A,B**, the correlation matrix **M** is defined as the matrix whose entries **M[A,B]** are the number of gene products(Uniprot accession IDs) that have annotations to **A,B** simultaneously (integrating GOA annotations for all species from 12-2008 until the submission deadline), i.e.

freq $A \cap B$.

Step 1: $freq(A \cup B) = M[A, A] + M[B, B] - M[A, B]$
Step 2: $P(A \cap B) = \frac{freq(A \cap B)}{freq(A \cup B)} = \frac{M[A,B]}{freq(A \cup B)}$
Step 3: $P(A|B) = \frac{P(A \cap B)}{P(A)}$

### 3.2.5 Evaluation of the Performance of Function Prediction Algorithms and the Proposed Benchmark

To assess the performance of the methods, predictions were generated using the same pipeline described above but simulating the predictions that would have been likely assigned if we were participating in the CAFA1 assessment. The results were later compared with those of other algorithms submitted in the CAFA1, which were provided to us anonymously.

The final list of predictions used as "gold standard" by the organizers of CAFA1 was also considered to assess the performance of our methods. The gold standard list included 866 targets from a total of 48,298 target sequences initially set for the assessment and 1876 annotations (which, after propagation,formed a total of 16,888 relations (gene-GO term). Only BP and MF ontologies were considered. A filtered "gold standard" list was also considered using only those GO terms that had 10 to 100 members in the true positive list (after propagation).

Table 3.2: Target sequences and species considered for the CAFA1 assessment.

| **Taxon ID** | **Organism** | **Number of targets** |
|---|---|---|
| 10090 | *Mus musculus* | 231 |
| 10116 | *Rattus norvegicus* | 45 |
| 3702 | *Arabidopsis thaliana* | 86 |
| 44689 | *Dictyostelium discoideum* | 2 |
| 8355 | *Xenopus laevis* | 16 |
| 9606 | *Homo sapiens* | 285 |
| 4932 | *Saccharomyces cerevisiae* | 5 |
| 1423 | *Bacillus subtilis* | 16 |
| 83333 | *Escherichia coli K12* | 153 |
| 287 | *Pseudomonas aeruginosa* | 2 |
| 1313 | *Streptococcus pneumoniae* | 25 |

The primary metric used for the CAFA1 assessment was gene-centric and is called the "CAFA score". For each predicted annotation from the algorithms and each annotation present in the "gold standard" list, terms were propagated to the root. Any overlap between the predicted annotations and the "gold standard" was considered a true positive. Precision, recall, thresholds and F-score were calculated using the ROCR package in R[54]. The average precision for each threshold $t$ was calculated across targets with respect to the number of targets for which at least one prediction was made above that threshold $t$. The average recall was calculated across all targets regardless of the threshold. The F-measure (harmonic mean) was also calculated as defined by the CAFA organizers [41].

As proposed by Gillis and Pavlidis [53], a different gene-centric approach was considered by using the Resnik measure to explore the semantic similarity between the actual and the predicted function. In particular, this metric was used to find how many predictions were more informative than the null, i.e., how many of those predictions had a higher score than what could be assigned by prevalence alone.

A function-centric measurement was also performed by calculating the

Area under the receiver operating characteristic curve (AUROC). Terms were propagated to the root and the scores assigned by the prediction methods were used. The package pROC [55] was used to calculate the area under the ROC curve.

## 3.3 Analysis of the Instability of Gene Set Enrichment Analysis Over Time

Gene Set Enrichment analysis are increasingly used to analyse and interpret biological information. For this reason, it is important to explore the impact that annotation instability has on the results of such analysis on a comprehensive scale.

In this study, more than 2,000 hit lists stored in GMT files from MolSigDB [56] (**collection C2**: curated gene sets from online pathway databases, publications in PubMed and knowledge of domain experts and **collection C3**: motif gene sets based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat and dog genomes) [30] were retrieved from the GSEA website:
(http://www.broadinstitute.org/gsea/msigdb/genesets.jsp). Only hit lists with more than 10 members were considered (**Figure** 3.8). After that initial filtration, a series of enrichment analyses were ran using yearly GOA (from May and November). For each gene set, the same score was assigned to all the genes (0.001). Enrichment analyses were done using the software ErmineJ, considering an over-representation(ORA) analysis and a FDR $\leq 0.1$ (**Figure** 3.9).

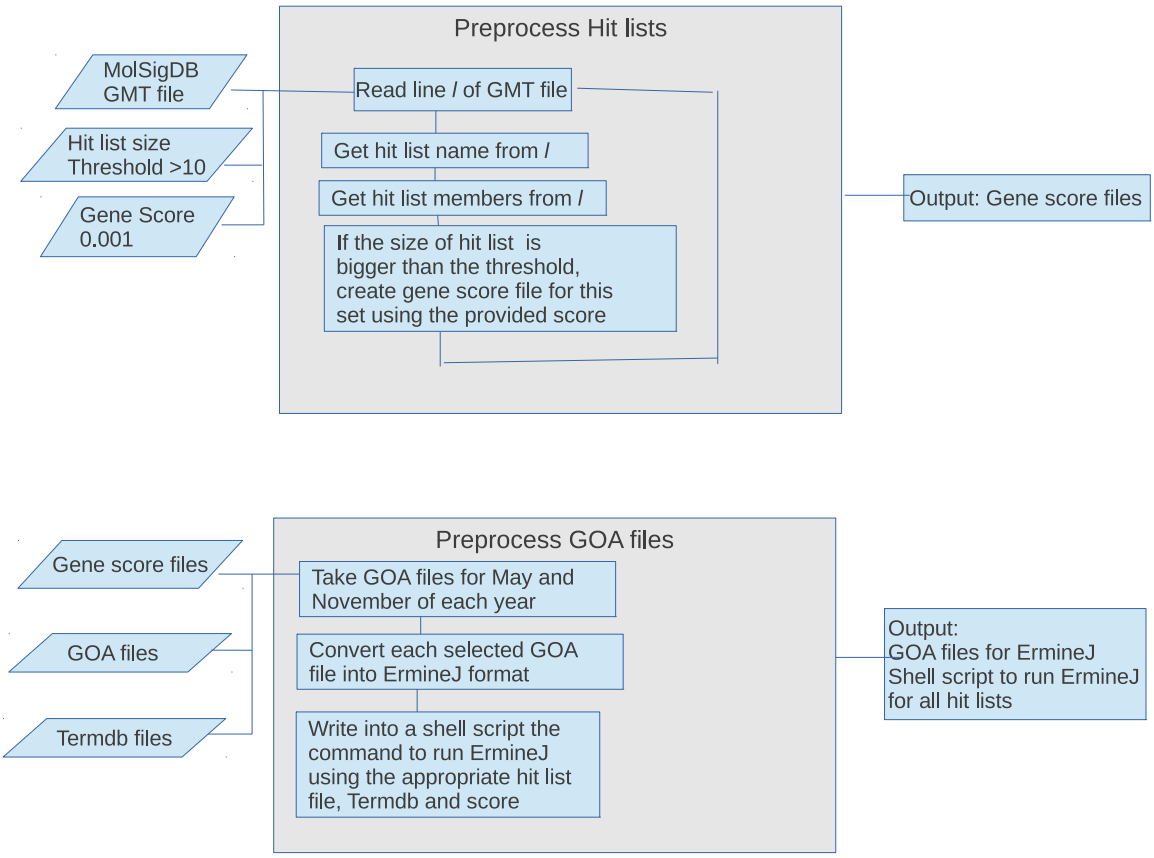



Figure 3.8: Pre-processing steps for enrichment analysis.

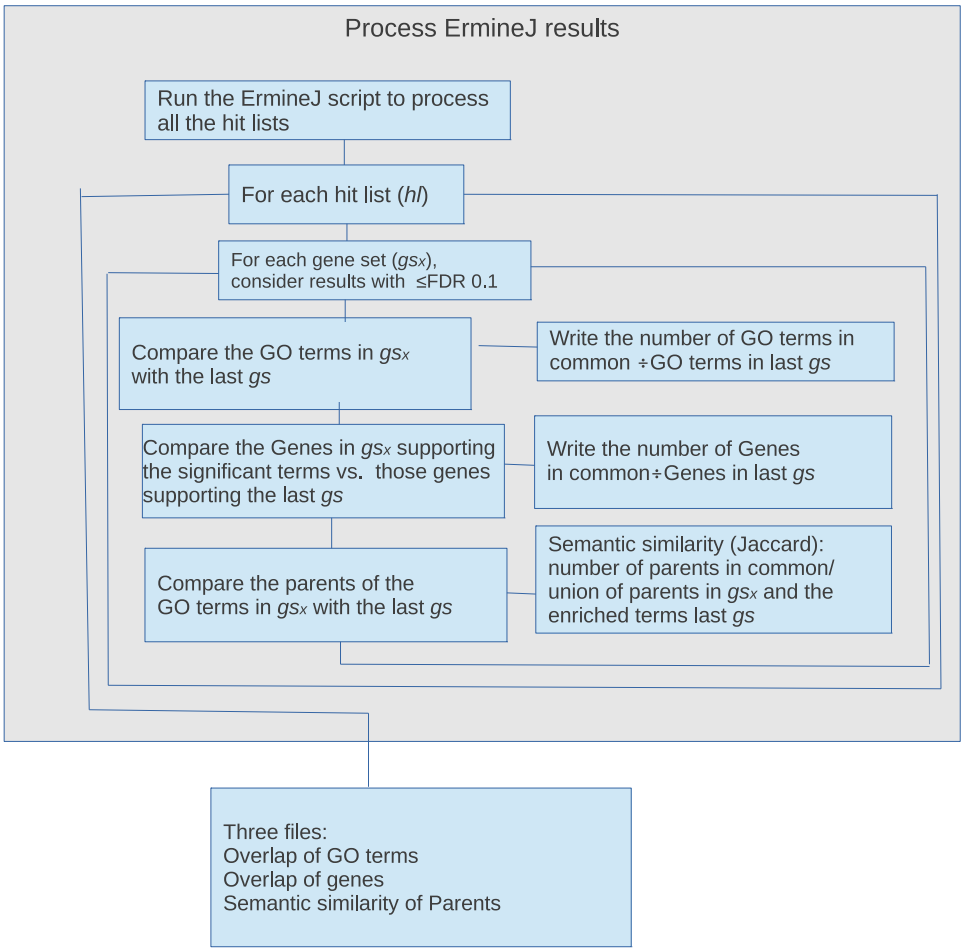



Figure 3.9: Pipeline to compare results of enrichment analyses over time.

# Chapter 4

# Results and Discussion

## 4.1 Exploratory Analyses

To get a general overview of the data, the first exploration was made towards exploring if the annotations were more prominent or supported for organisms with a smaller genome size or if the biases previously reported were most likely influenced by curation preference. By considering the the genome size of each species (considered as the total number of coding genes reported in the current assembly listed in the Ensembl Genome Browser), the number of existing annotations and the number of annotations supported by publications. The results showed that there is no association between of the number of annotations and the genome size. Within organisms that have less than 10,000 genes, yeast has more annotations and also more publications (circle size), followed by E. coli; while dictyostelium clearly lacked annotations and publication supporting those annotations. For those organisms that have a genome size between 15,000 and 20,000 genes, the fruit fly had considerably less annotations than chicken or dog, but more of them were supported by publications. This result is expected as chicken, dog or cow have most of their annotations are inferred by homology. For those organisms whose genome size is bigger than 20,000 genes, mouse was noticeably the organism with the highest number of annotations, followed by human and rat. These three organisms also had a considerable number of annotations supported by publications. However, other important model organisms like zebrafish, worm or Arabidopsis fall behind in the number of annotations available and those supported by publications (**Figure** 4.1).

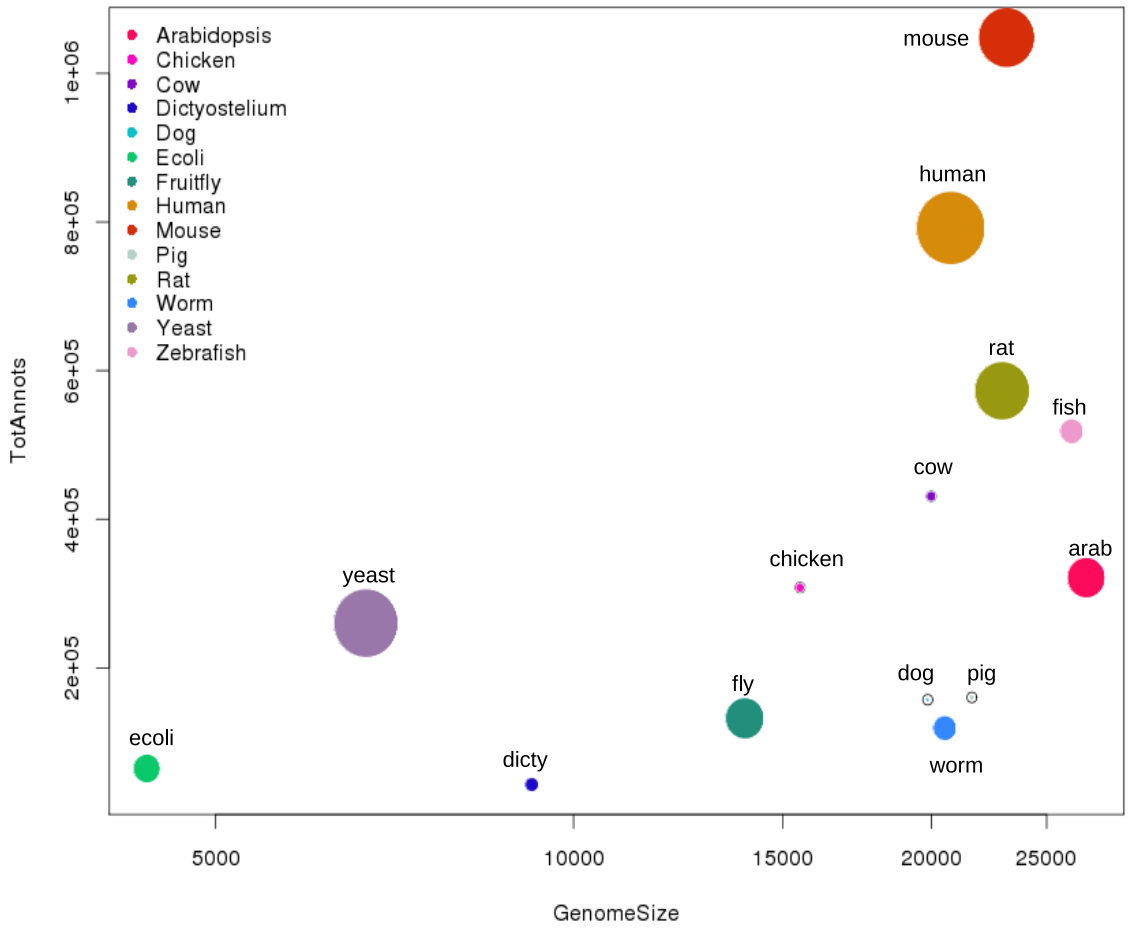



Figure 4.1: Overview of Species-specific biases on manual curation efforts. No relationship was found between the genome size (quantified by the number of protein-coding genes), the number of total annotations and the total number of publications supporting them (circle size).

All 14 organisms were processed for analysis. However, for descriptive purposes and an easier comparison, only 5 representative species are discussed: human, Arabidopsis, E.coli, yeast and fruit fly.

It is clear that there are more gene products than genes in the genome.

However, one would likely expect a constant or subtle increase in the number of those gene products annotated over time. Nevertheless, these numbers are highly dependant on the source and the database that provides that information. As mentioned in the introduction, DB Object IDs from the UniProtKB are reflecting gene products mapped to each gene. However, the instability of such database (and others) will likely impact in the number of DB Objects available.

Consistent with this hypothesis, the number of gene products (DB Object IDs) available at each GOA version over time had a gradual increase for organisms with larger genomes such as mouse, human or Arabidopsis. In contrast, smaller organisms like fly or yeast seemed more stable in terms of the number of gene products. However, exceptions were observed in a particular time points for human and E.coli, where clearly some gene products were lost and recovered intermittently over time.

Even if the GAF2.0 GOA format has the rule of assigning DB Object IDs to a top level primary gene or gene product ID, the total number of proteic entries are directly influenced by the source of the annotation files. In particular, the sudden decrease in the number of DB Object IDs observed for human data between 2009 and 2011 can be explained by the fact that, at the end of 2008, a draft of the complete human proteome was released specifically from UniProtKB/SwissProt. This release had approximately 20,000 putative human protein-coding genes manually reviewed. UniProtKB/TrEMBL products were also revised and 15,000 isoforms were merged with 40% of the UniProtKB/SwissProt entries, causing a large reduction in the number of annotated products. These numbers are consistent with the shifts observed (**Figure** 4.2).

The next increase for human was observed in 2011. The most likely reason is that a complementary pipeline was implemented to import predictions from UniProtKB/TrEMBL sequences, which are non-revised and potentially redundant (`http://www.uniprot.org/help/human_proteome`).

Similar shifts can occur when annotation pipelines are revised by other databases that are species-specific, as it may be the case for the changes observed in E.coli (EcoCyc).

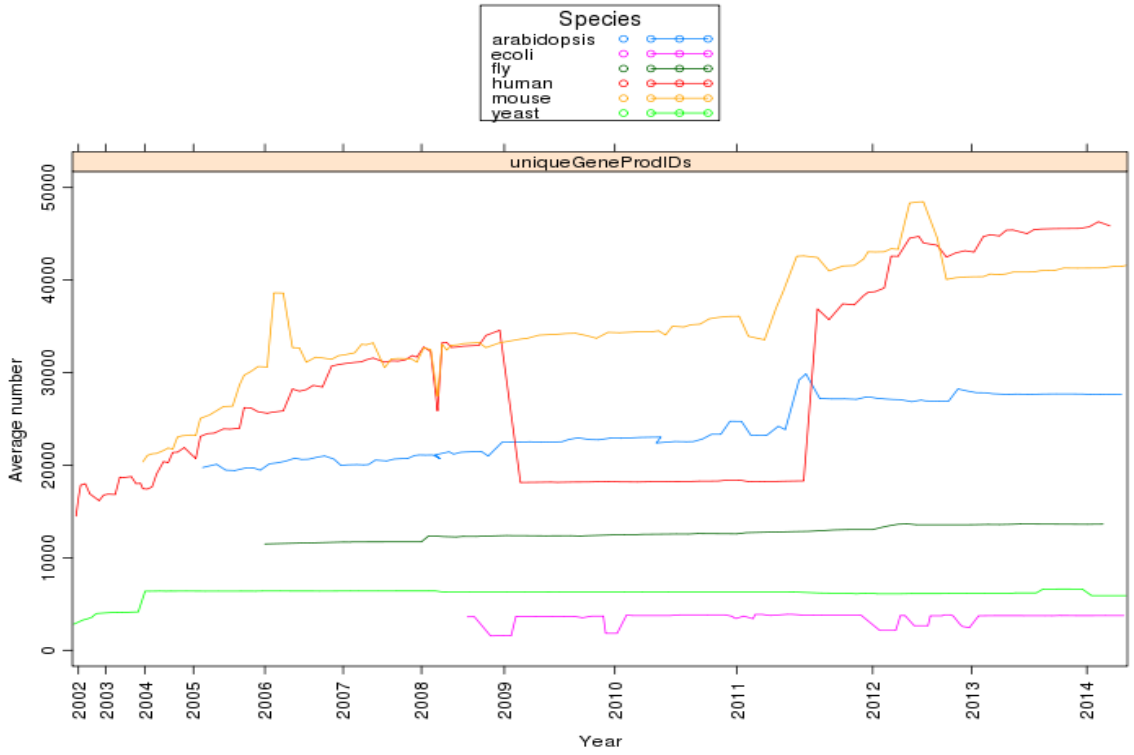



Figure 4.2: Total number of Gene Product Identifiers found for each species.

Additionally, a scientist would most likely be interested to know how stable are the annotations of each gene product over time. This can be explored by looking at the changes in the number of the GO terms directly annotated to each gene product over time.

To have a general overview of the number of "functions" one can expect for each gene product and species, the average number of GO terms assigned

to all the gene products was computed. The results showed a considerable variation within and between species, ranging from 3 to 10 GO terms per gene product. In this metric, no distinction was made for GO terms manually assigned or inferred electronically. However, in the case of human, the average count is clearly influenced by electronic annotations, specially when considering the changes discussed above (where UniProtKB/TrEMBL products were removed between 2009 and 2011). From this observation, it is clear that the gene products from UniProtKB/SwissProt (in that period of time) had considerably more GO terms assigned than their non-reviewed counterparts originated from UniProtKB/TrEMBL. But, as both sets are incorporated in-distinctively in the GOA files, the average value decreased when the new pipeline reincorporated UniProtKB/TrEMBL entries into the database.

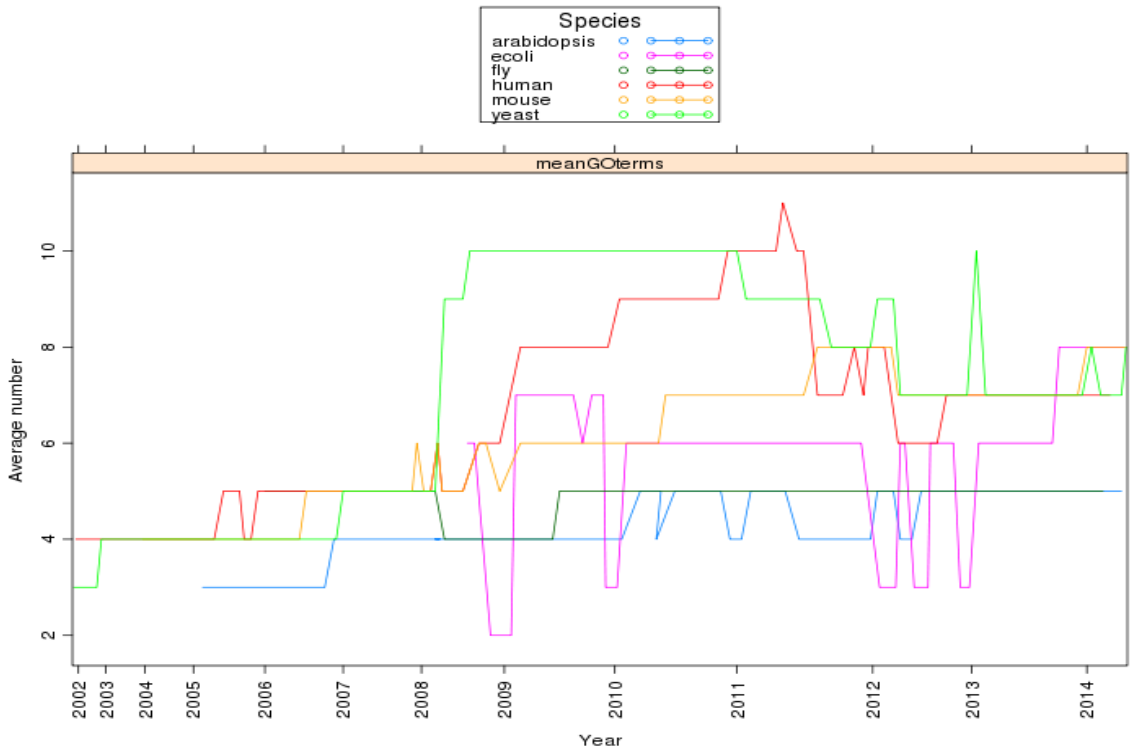



Figure 4.3: Average number of GO terms directly annotated to gene products across editions.

A closer look at the data showed a high variability in the number of GO terms assigned to each gene product over time (**Figure** 4.3). In some cases, the changes would be noticeable, whereas in others, no changes would be detected. An example of such changes is shown (**Figure** 4.4), where a random set of genes from human was taken to compare the difference between the number of GO terms assigned from one edition to the next one. A common "assumption" is that existent annotations remain stable and new ones would be incorporated over time. A gradual increase in the number of GO terms might be then expected. However, this figure showed that some notorious shifts can occur on small periods of time and that these differences occur in larger proportions for some gene products and in particular editions.

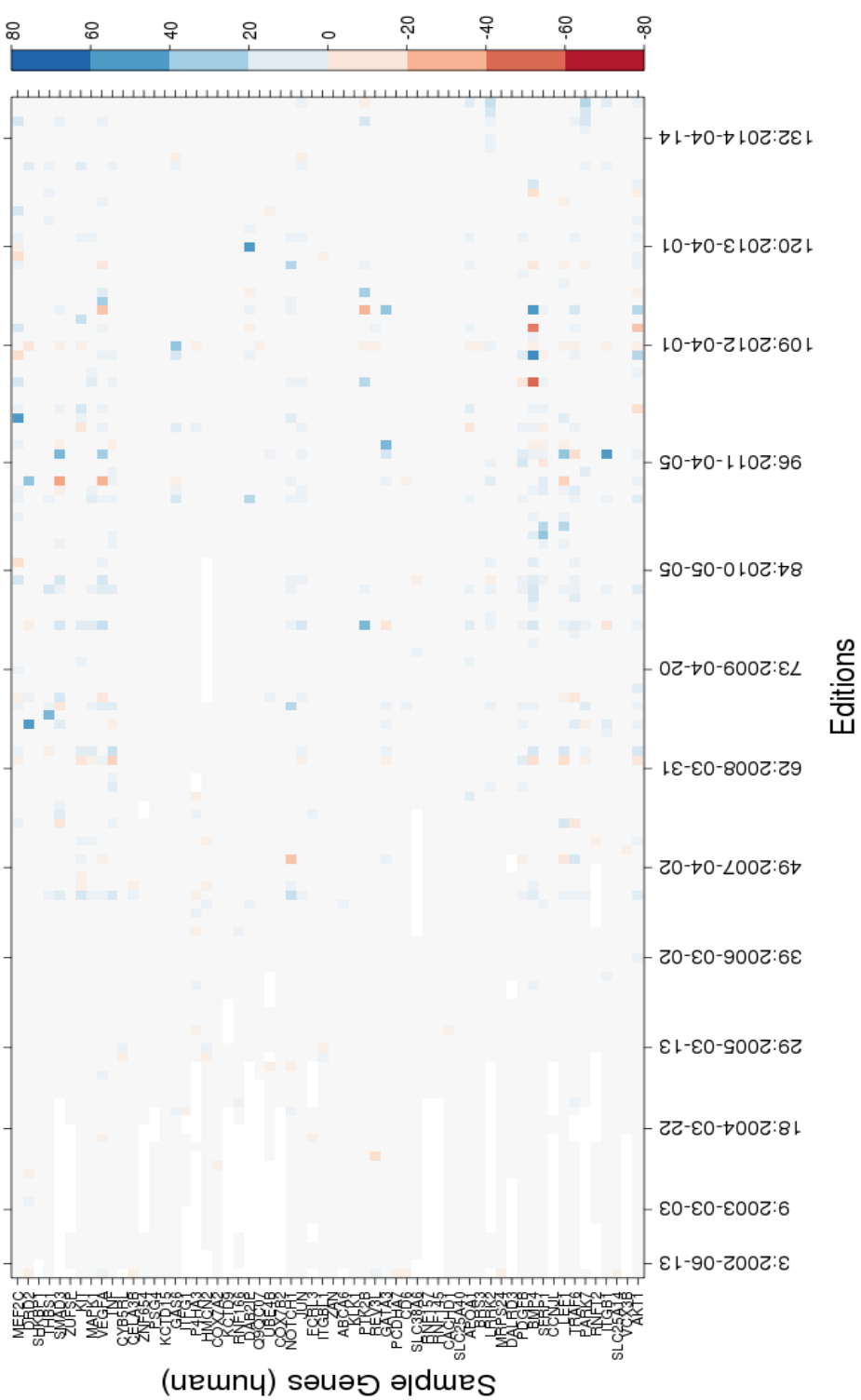



Figure 4.4: Contrasting shifts found in the number of GO terms assigned to a random set of gene products across editions. The gray colour indicate that no changes in the number of terms were observed compared to those of a previous edition. A red colour would indicate that the gene product lost more than 20 terms within that particular edition compared to the previous one and the blue colour would reflect that it gained more than 20 GO terms compared to the previous edition. All genes were explored, but only a random sample is shown. Genes that showed the more drastic changes in human data showed a magnitude of over 100 GO terms added or removed at particular time points compared to previous editions.

However, gene products that seemed to have a relatively constant number of GO terms are not necessarily "stable". In particular, terms can appear and disappear from one edition to the next. An example of such instability is shown on (**Figure** 4.5).

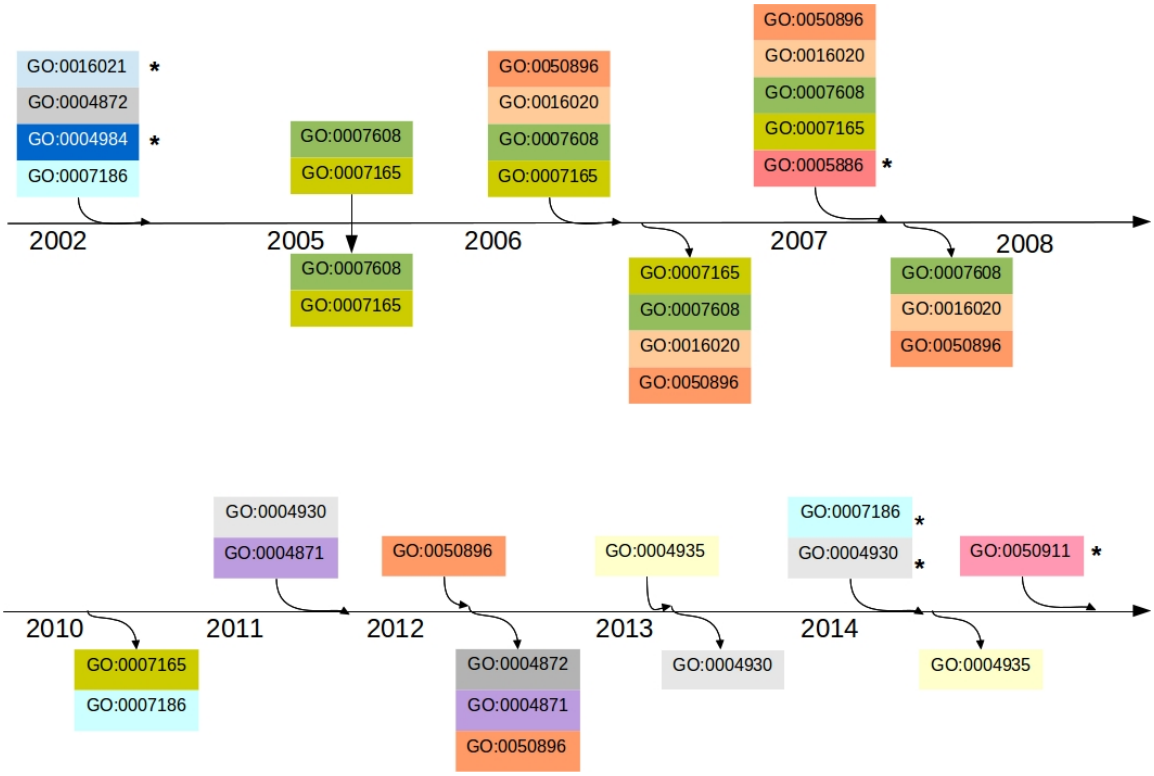



Figure 4.5: Example of the functional instability that gene products have over time. The gene product OR11H7 (Olfactory receptor 11H7) with the UniProtKB/SwissProt ID Q8NGC8 is used as an example of functional instability. The GO terms were traced and a constant adding and removal of the same terms across editions could be appreciated. The schema displays the historical changes, where each GO term directly annotated has been colour coded. Those marked with an asterisk (*) are GO terms that remained annotated to the gene product as of August 2014.

The latter example clearly reflects the problem of the instability of the GO terms. A way to explore such shifts on a wider scale is by looking at how semantically similar a gene product is to itself. This can be done by

using the Jaccard distance (**Figure** 4.6).

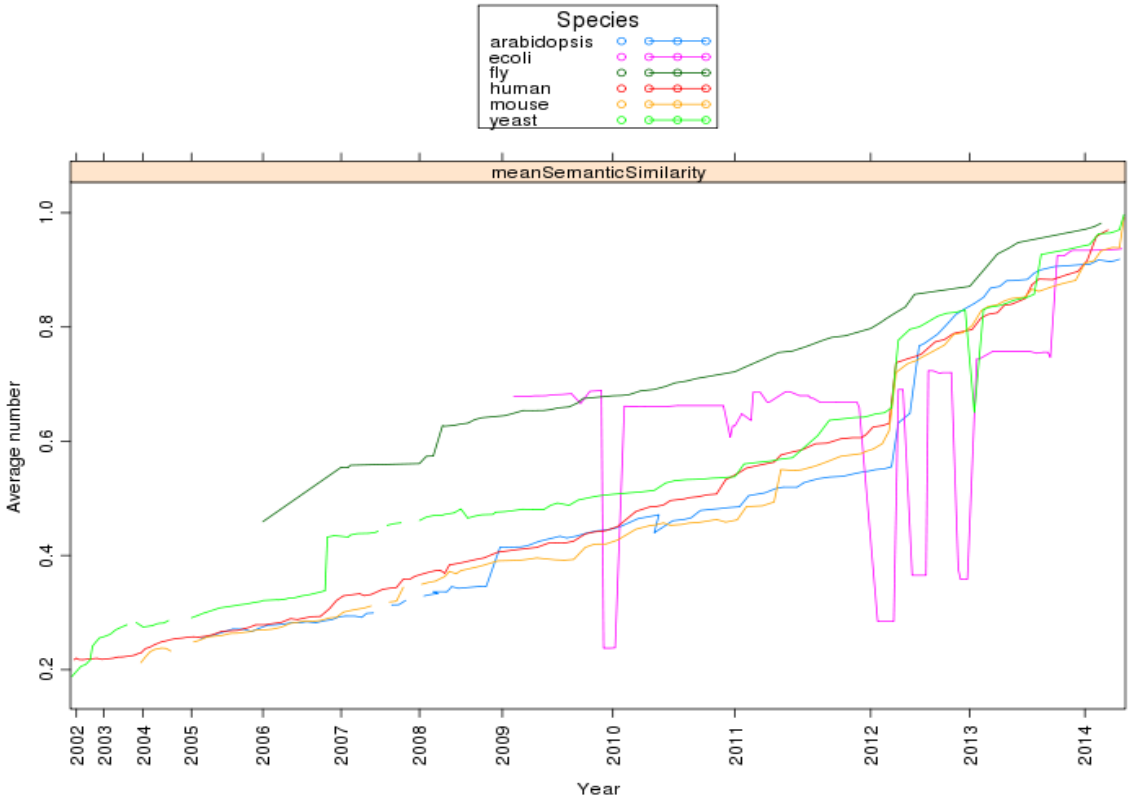



Figure 4.6: Average values of how semantically similar genes are across editions.

Noticeably, after only 2 years, yeast, human, Arabidopsis and mouse data showed only a semantic similarity of just 50% compared to the current annotations but increased to 80% within the same year. Interestingly, the fruit fly had a different behaviour and on average, its gene products retained a higher functional similarity. In contrast, E.coli data showed a considerably abnormal pattern of semantic similarity, which seemed correlated with the drops observed in the number of gene product IDs. The results are similar to those reported by Gillis and Pavlidis (2013) for "genes always present"

in human data [37].

Gene products that are multifunctional often tend to be prioritized for curation. In fact, when looking at the gene sets defined by different GOC projects, they seemed to rank high in terms of their multifunctionality score. This behaviour was particularly observed in the gene sets listed for cardiovascular processes and those derived from the reference genome project (**Figure** 4.7).

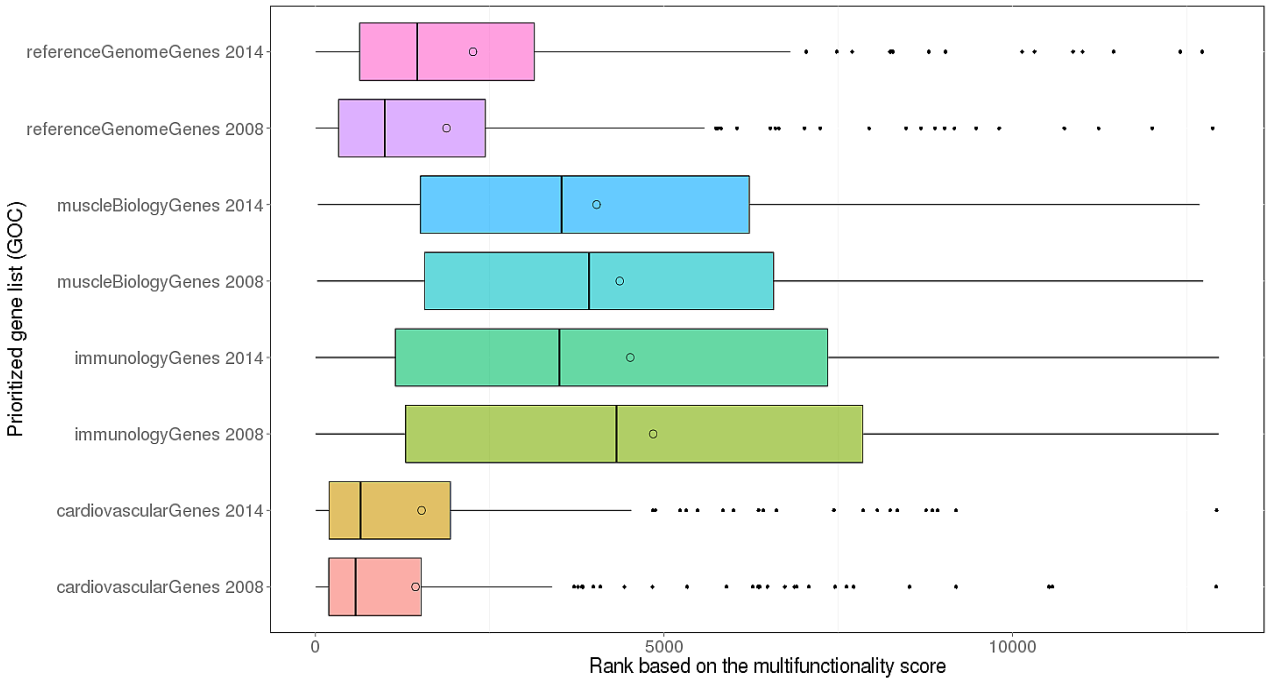



Figure 4.7: Exploring the association between prioritized gene sets for curation and multifunctionality. Genes that have been prioritized for curation seemed also to be highly multifunctional. However, some projects seemed to rank higher than others. Some genes are not in the top multifunctional ranking, seemingly because the curation project is still in progress.

However, from a general overview, the average multifunctionality score for each species showed a very slight gradual increase, and only yeast and

E.coli data had considerable shifts. This means that on average, the score of gene multifunctionality across genes do tend to increase over time, potentially due to the increase in the multifunctionality effect of the prioritized genes (**Figure** 4.8).

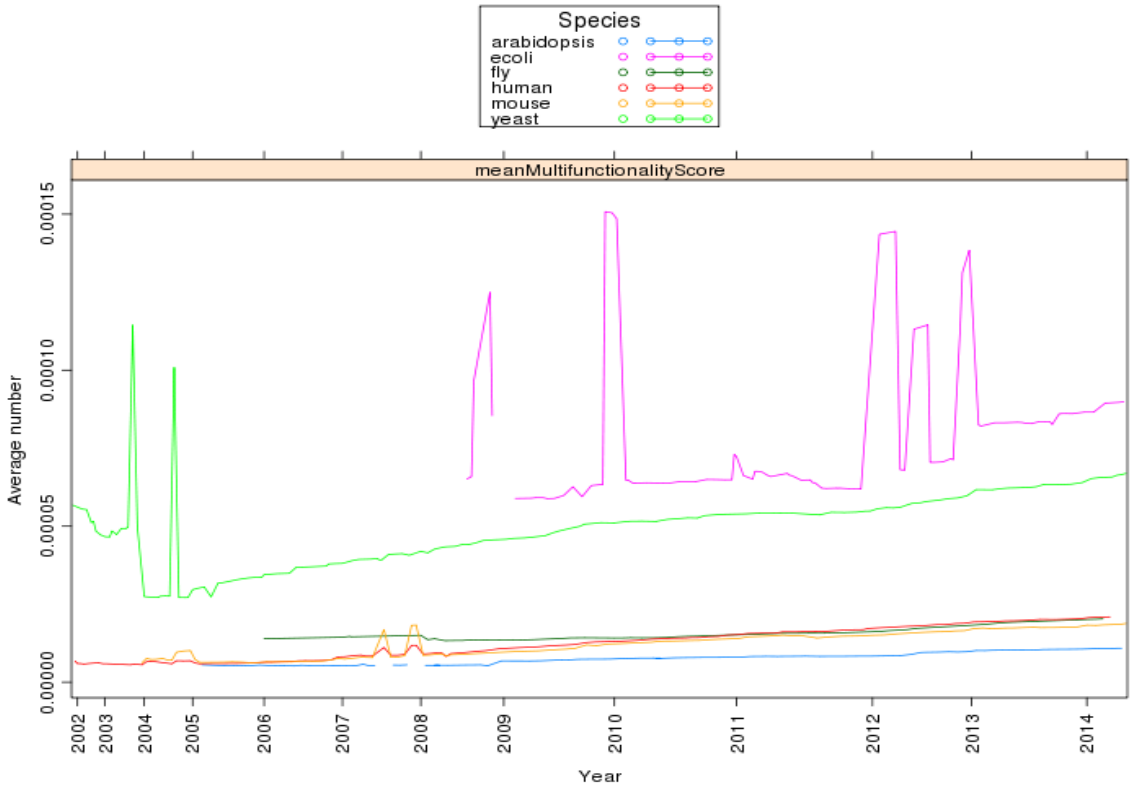



Figure 4.8: Average score of gene multifunctionality across editions.

In fact, most of the gene products present in GOA data are not considered multifunctional, specially when on average, each gene product has 3 to 10 GO terms. One particular problem that has been constantly observed and described is that most genes have very shallow GO terms assigned, specially because these were inferred computationally and in most cases, have not been curated.

One indirect form to verify this is by exploring the overall GO term membership across editions (i.e. the number of gene products belonging to a particular GO group). The results showed that only a few set of GO terms are directly annotated to support the largest number of gene products (**Table** 4.1).

In particular, the top GO terms used across editions were too shallow, reflecting the lack of a proper coverage in the GOA data, specially as the root terms were also the ones most commonly assigned: "biological process" (GO:0008150), "molecular function"(GO:0003674/ GO:0005554), "cellular component" (GO:0008372/GO:0005575), "cytoplasm"(GO:0005737), "nucleus" (GO:0005634), "translation" (GO:0006412), "plasma membrane" (GO:0005886), "membrane" (GO:0016020), "integral component of membrane" (GO:0016021), "protein binding" (GO:0005515) and "ATP binding" (GO:0005524).
On one hand, a person would assume that such shallow terms are only assigned by IEA annotations. However, when IEA annotations are removed, the top terms most commonly assigned (as direct GO terms) remain mostly the same, except for a few others like: "regulation of transcription, DNA-templated" (GO:0006355), "cellular response to DNA damage stimulus" (GO:0006974), "structural constituent of ribosome"( GO:0003735) or "cytosolic ribosome" (GO:0022626), which were particularly prevalent in Fly data.

Table 4.1: The GO terms most frequently used in GOA data.

| **Species** | **>5,000 genes** | **1,000-5,000 genes** | **100-1,000 genes** | **<100 genes** | **Total GO terms(GOA)** |
|---|---|---|---|---|---|
| Human | 1-5 | 7-33 | 96-436 | 3244-14034 | 3347-14507 |
| Arabidopsis | 1-2 | 9-22 | 85-265 | 1534-5223 | 1628-5509 |
| E.coli | 0 | 1-2 | 2-38 | 859-3146 | 861-3191 |
| Yeast | 0 | 5-8 | 6-85 | 1396-4925 | 1402-4965 |
| Fruit fly | 0 | 4-13 | 61-128 | 4320-6238 | 4410-6374 |

The numbers on the table reflect the range in number of GO terms that have a certain number of genes (GO membership) for each species. The last column of the table shows the range of GO terms that have been ever been used to support the annotations across GOA editions.

Another way to observe this is by looking at the number of inferred terms for each direct annotation (**Figure** 4.9). The average values obtained were consistent with previous observations. In particular, for human data, the number of propagated functions between 2009 and 2011 was considerably higher from those of other time points where UniProtKB/TrEMBL annotations (with shallow GO terms) are present. However, it is important to note that all the species seem to have a gradual increase in the number of propagated terms, which can lead to think that annotations are gaining "specificity". Nevertheless, it is important to remember that one of the properties of the GO structure is that the depth of a term in the branch does not necessarily reflect how specific a term can be.

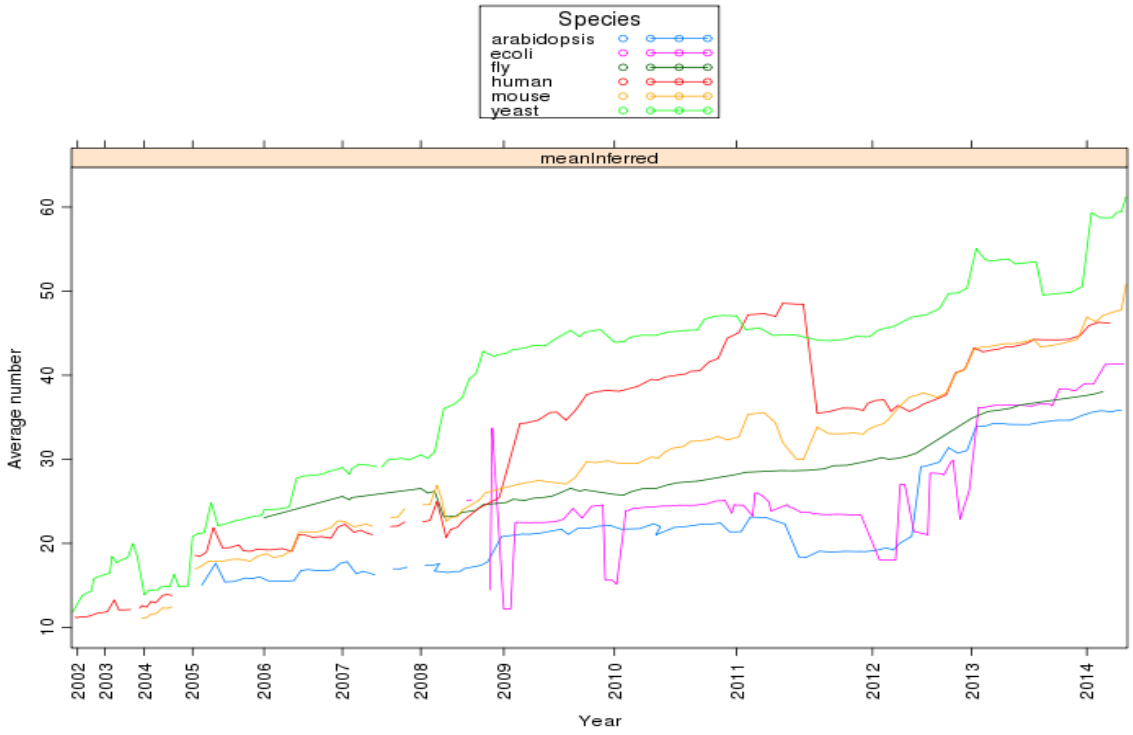



Figure 4.9: Average number of inferred terms over time.

However, an interesting way to interpret that subtle an gradual increase is by arguing that in the last couple of years, the Reference Genome Project

focused their efforts towards improving and revising existent GO annotations. The results reflect such efforts. Particularly, a large number of annotations are being curated from electronic inferences. In contrast, a small but still appreciable proportion of previous manual annotations have also been upgraded to more granular GO terms. However, these promotions were considerably different between species (**Figure** 4.10).

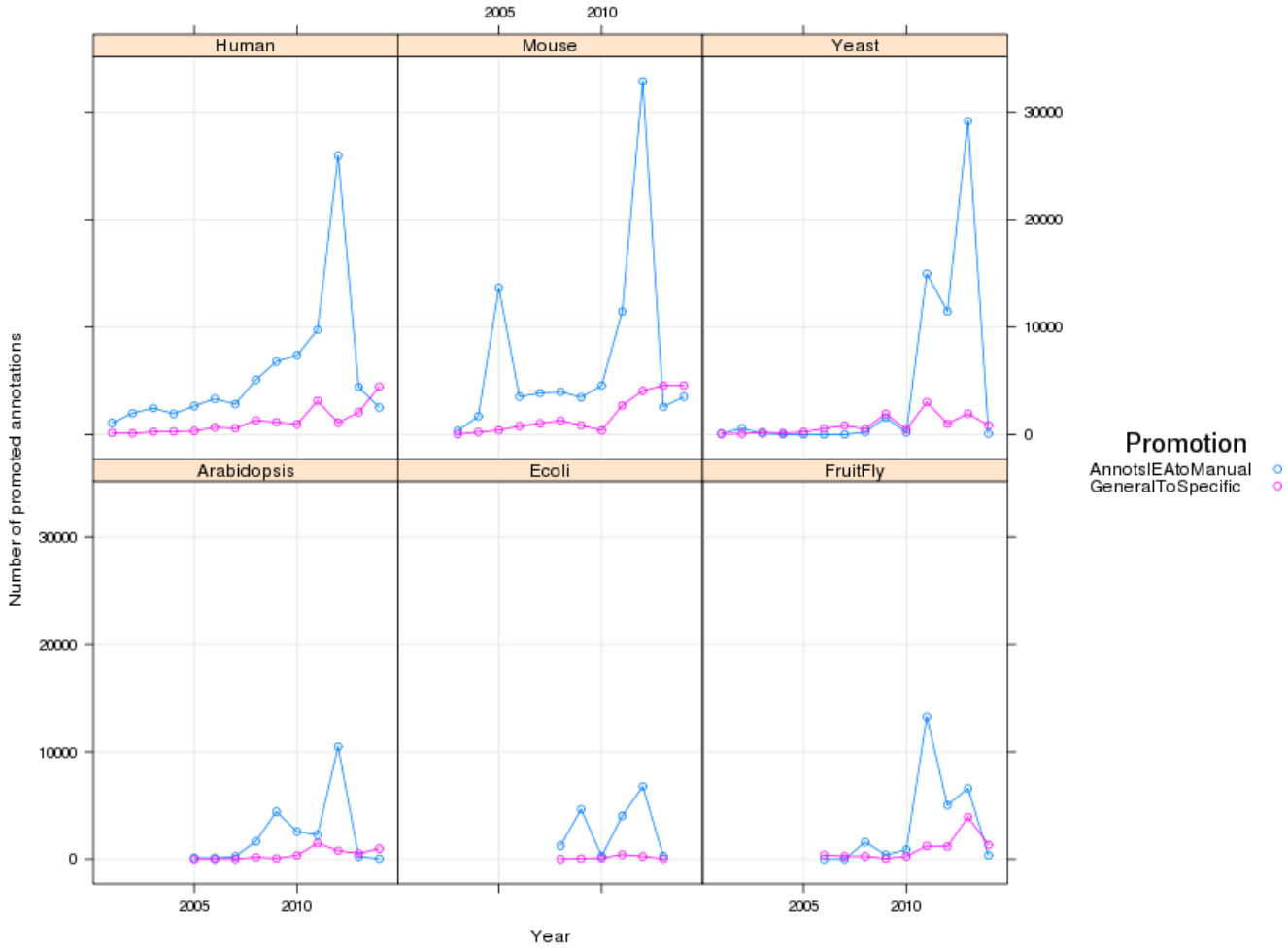



Figure 4.10: Total number of electronic annotations that are curated and total number of manual annotations that are promoted with a more granular GO term.

The effect of upgrading IEA annotations to manual revisions can also be explored by observing the changes in the number of annotations assigned to other manual evidence codes for each ontology. For example, in the cellular component ontology, recent annotations have been assigned to the computational codes IGC, IBA, IKR, and IRD. Some codes are only used for one

species: ISO and IGC in fruit fly and IKR in human. Others have been used for longer and have increased its usage, like RCA and ISM. It seemed that more increases in annotations from this ontology are found for those with experimental evidence codes, such as IPI, IDA and in a smaller proportion IGI. The usage of IEP was apparently discontinued and EXP evidence is only found currently in E.coli data. Interestingly, TAS annotations have remained stable, except for human data, where an important increase was found. The usage of other codes assigned by curators, such as IC or ND have remained stable, except for Arabidopsis, where a considerable increase of gene product annotations with the code ND (no biological data is available) was reported in 2011 (**Figure** 4.11).

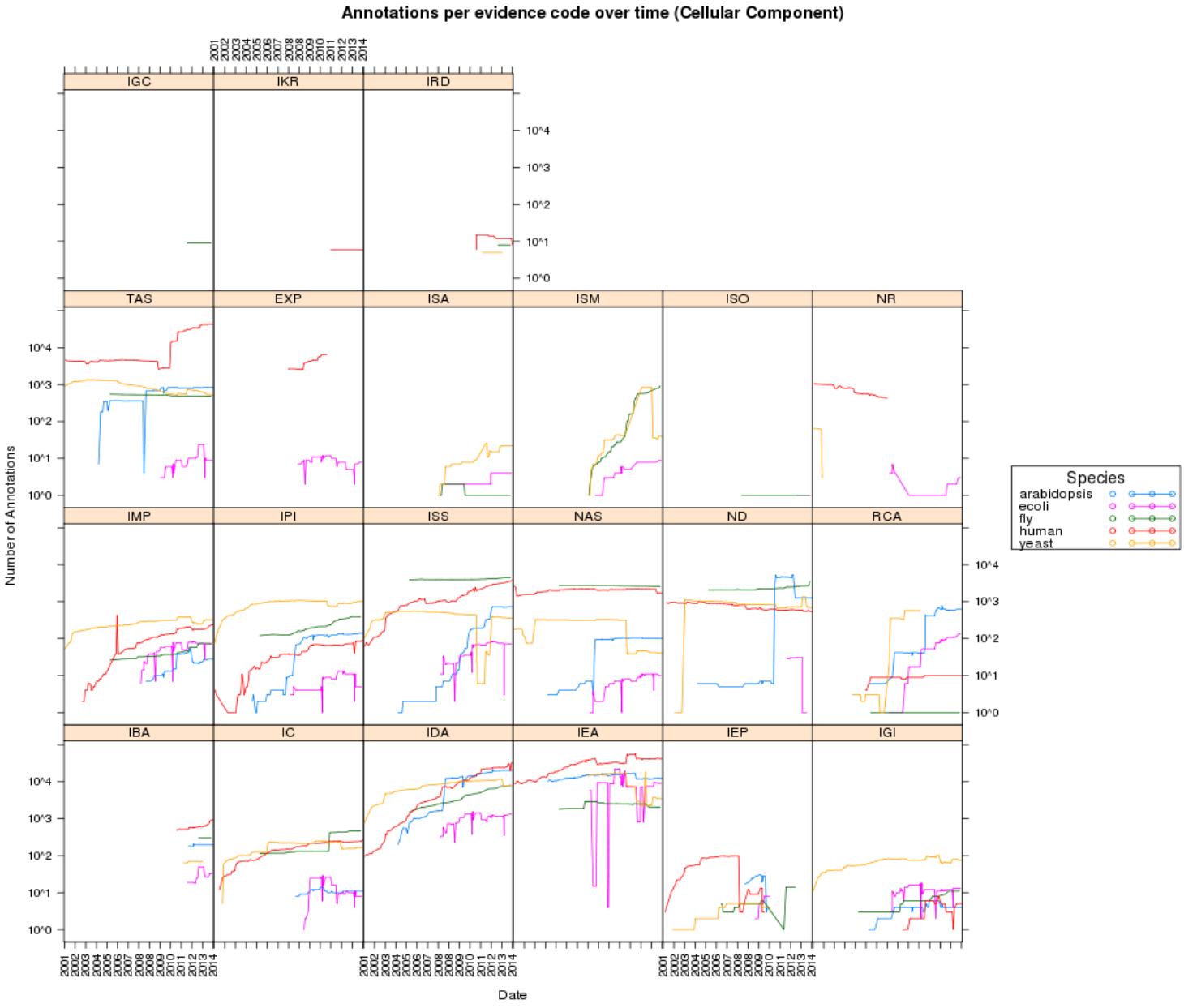



Figure 4.11: Changes in the usage of evidence codes for the Cellular component Ontology.

In the molecular function ontology, recent annotations were assigned to the computational codes ISO, ISA, IBA, IKR and IRD. Some increased its usage like ISS (although remained the same for fly or E.coli), ISM in fly (but

dropped in yeast), and RCA remained stable in all but yeast data, were its usage dropped. Experimental annotations have also increased slightly for IDA, IPI and IGI. The usage of IEP was also discontinued in this ontology. TAS annotations have remained stable, whereas IC have slightly increased. ND annotations have also remained stable, except for Arabidopsis again. Contrary to the cellular component, annotations assigned to ND were removed from E.coli data (**Figure** 4.12).

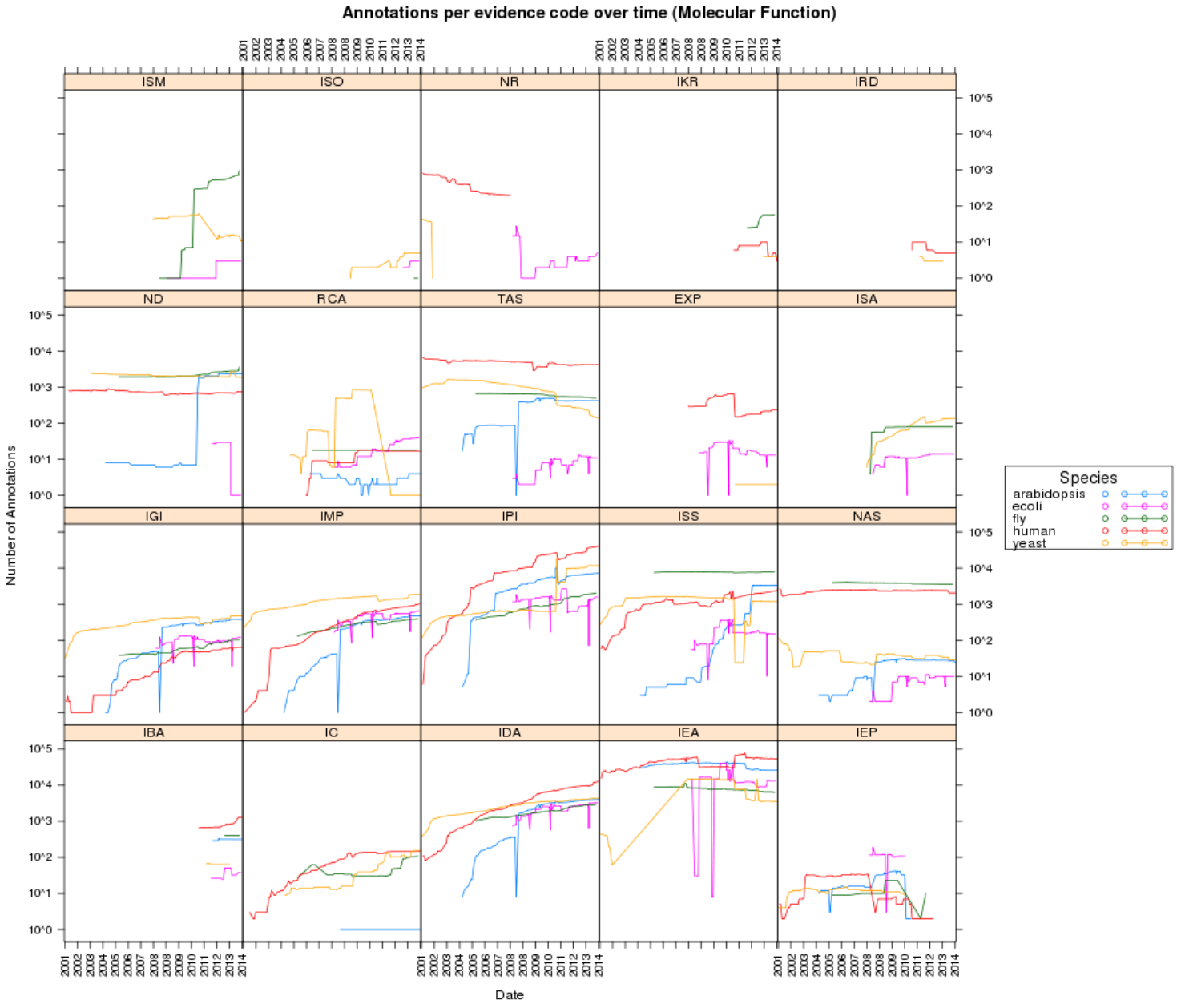



Figure 4.12: Changes in the usage of evidence codes for the Molecular Function Ontology.

In the biological process, many new annotations have been recently incorporated with the computational codes ISA, ISM, IBA, IRD, IKR and

RCA. However, they were mostly used for fly, yeast and human data. Experimental annotations are gradually increasing for IDA, IGI, IMP or IC codes, and has remained stable in TAS or ND annotations. It seems that NR data is still present in E.coli data, indicating that some genes haven't been characterized (**Figure** 4.13).

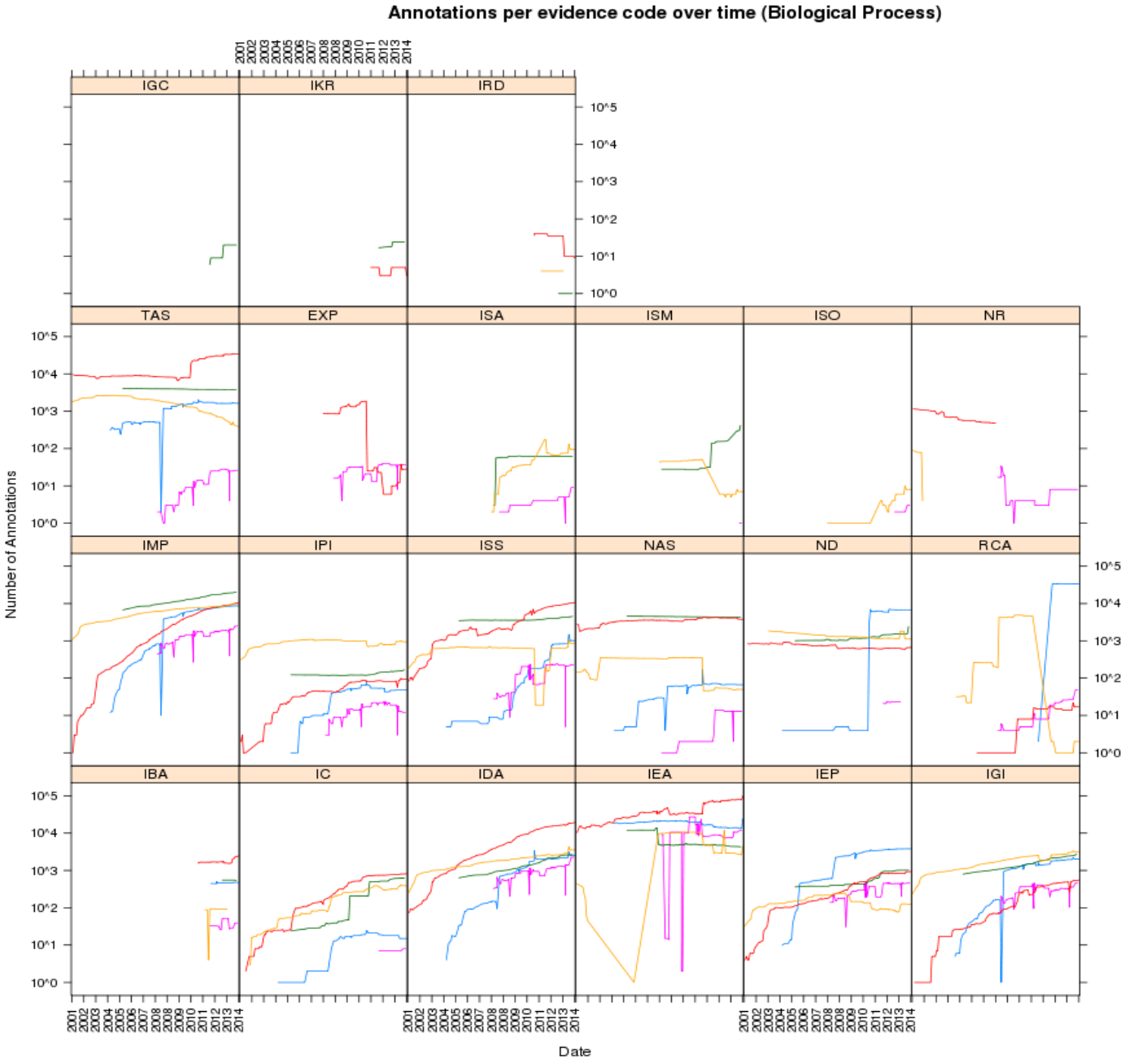



Figure 4.13: Changes in the usage of evidence codes for the Biological Process Ontology.

It is important to remember that annotations can change not only due to

GO term upgrades or changes in the evidence codes, but also due to inconsistencies in the databases. A way to visualize the shifts observed (**Figure** 4.5) on a practical way for users is by plotting the "existence" of the annotation at each time point.

Contrary to the previous methods discussed in the introduction, I considered not only the GO term or evidence code, but also the supporting publication to trace the annotation. This is arguably the most important factor to consider if one aims to properly trace the existence of the annotation and, even if the annotation disappeared, the association can still be validated if the source is accessible, specially for manual annotations.

Users often assume that manually curated annotations from revised gene product IDs (as its the case for those derived from the UniProtKB/SwissProt database) remain relatively present or stable in subsequent GOA editions. However, an exploration made with human data for the highly popular and multifunctional gene TP53 with a random selection of its annotations showed the opposite. ( **Figure** 4.14) shows that, even for highly studied genes such like this one, important changes occur from one edition to the next. Old annotations can be removed completely, others can remain relatively stable, disappear after just a few editions or exist in only one single GOA edition.

Another potential hypothesis of why manual annotations can disappear, is that maybe the source was not robust enough, was wrongly interpreted or was even derived from a retracted paper. An exploration of how many retracted papers are used to support annotations however, showed that the numbers are negligible (less than 5, data not shown).

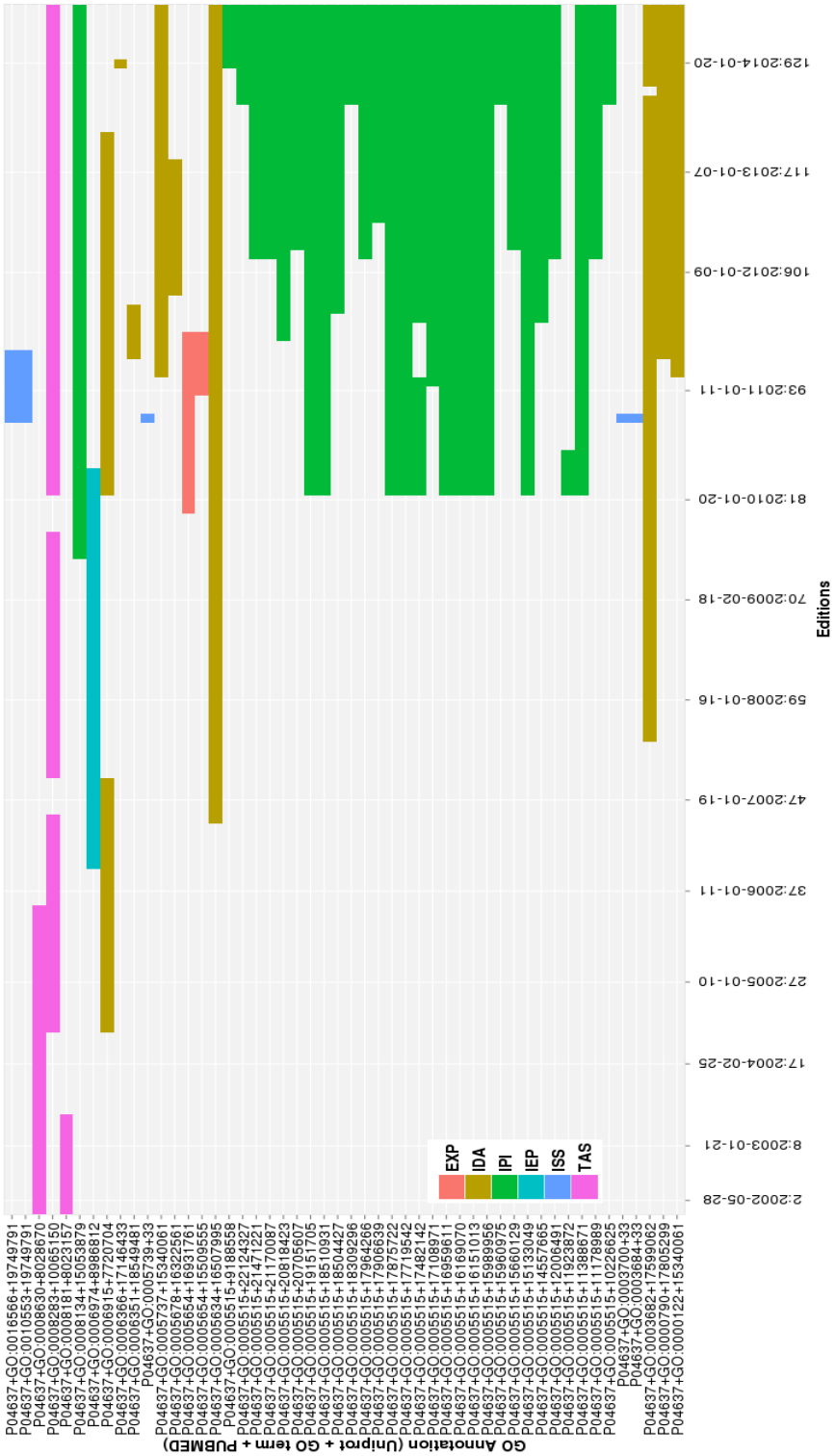



Figure 4.14: Manually curated annotations are also unstable. A common assumption is that manually curated annotations are stable. However, an exploration for 50 random annotations in human for the multifunctional gene TP53 highlighted that their annotations can be highly volatile over time.

## 4.2 Utility of Creating a Web-based Visualization Tool: GOtrackWeb

Clearly, there is still a lot to learn from exploratory data analyses in gene annotation data. The previous results highlighted the high variability of GO annotations. However, the importance of assessing the instability should also be translated into an application that users can use, and particularly, that allows them to explore how these factors impact genes and annotations of their interest. Such tool was not available until now.

A database and a web interface (GOtrackWeb) were built to study and extend this information to the community and designed to keep the information as updated as possible. This is a large contribution that can be useful not only for researchers, but also for GO curators.

On the website, users are able to explore all the UniProtKB/SwissProt or UniProtKB/TrEMBL identifiers mapped to particular gene symbols; compare parameters such as the number of direct and propagated terms for their gene product of interest over time, changes in their multifunctionality score or how semantically similar they were in previous versions as compared to the latest version on a dynamic plot. Similarly, they can retrieve which are the top multifunctional genes for each species in the latest edition (**Figure** 4.15).

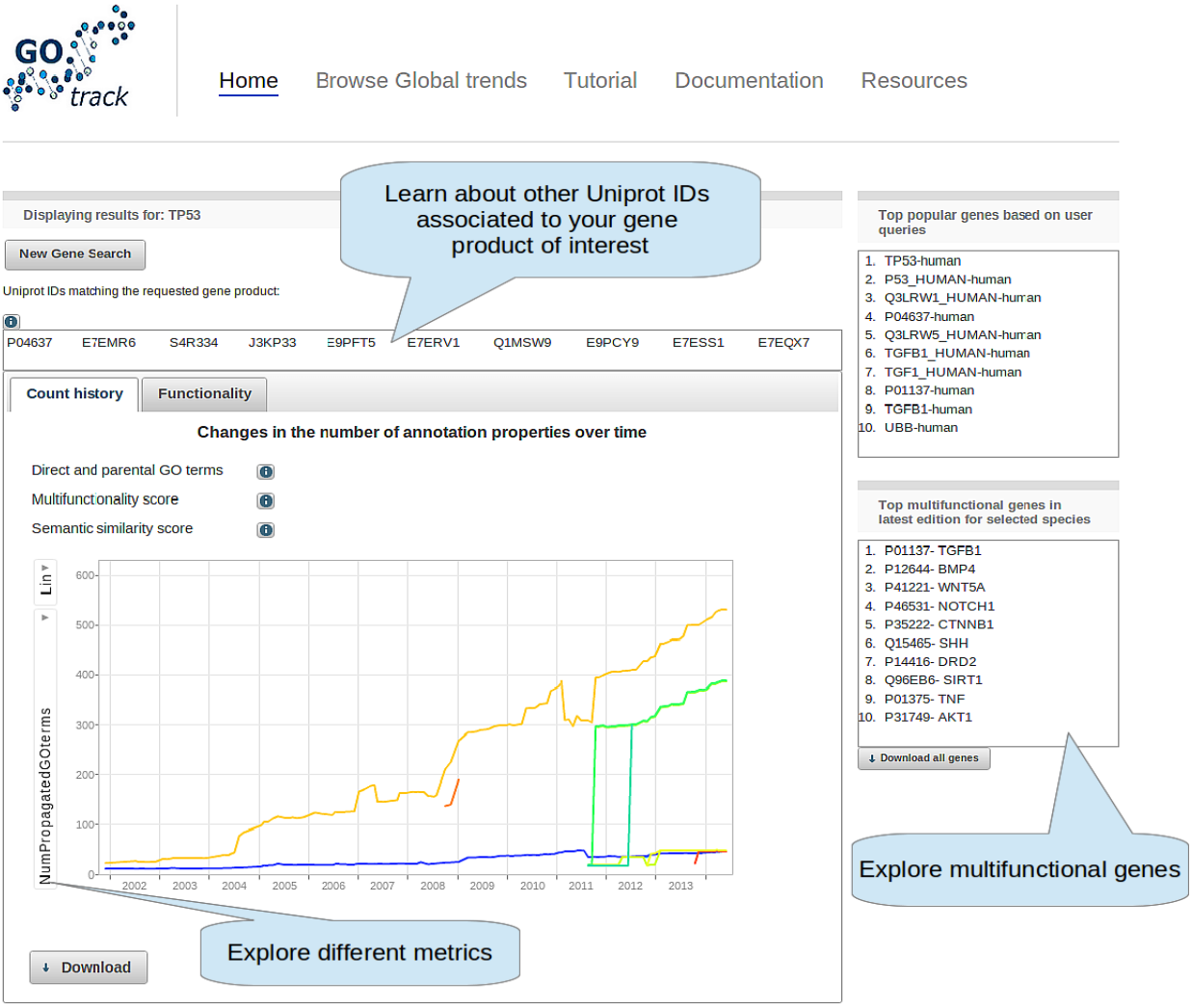
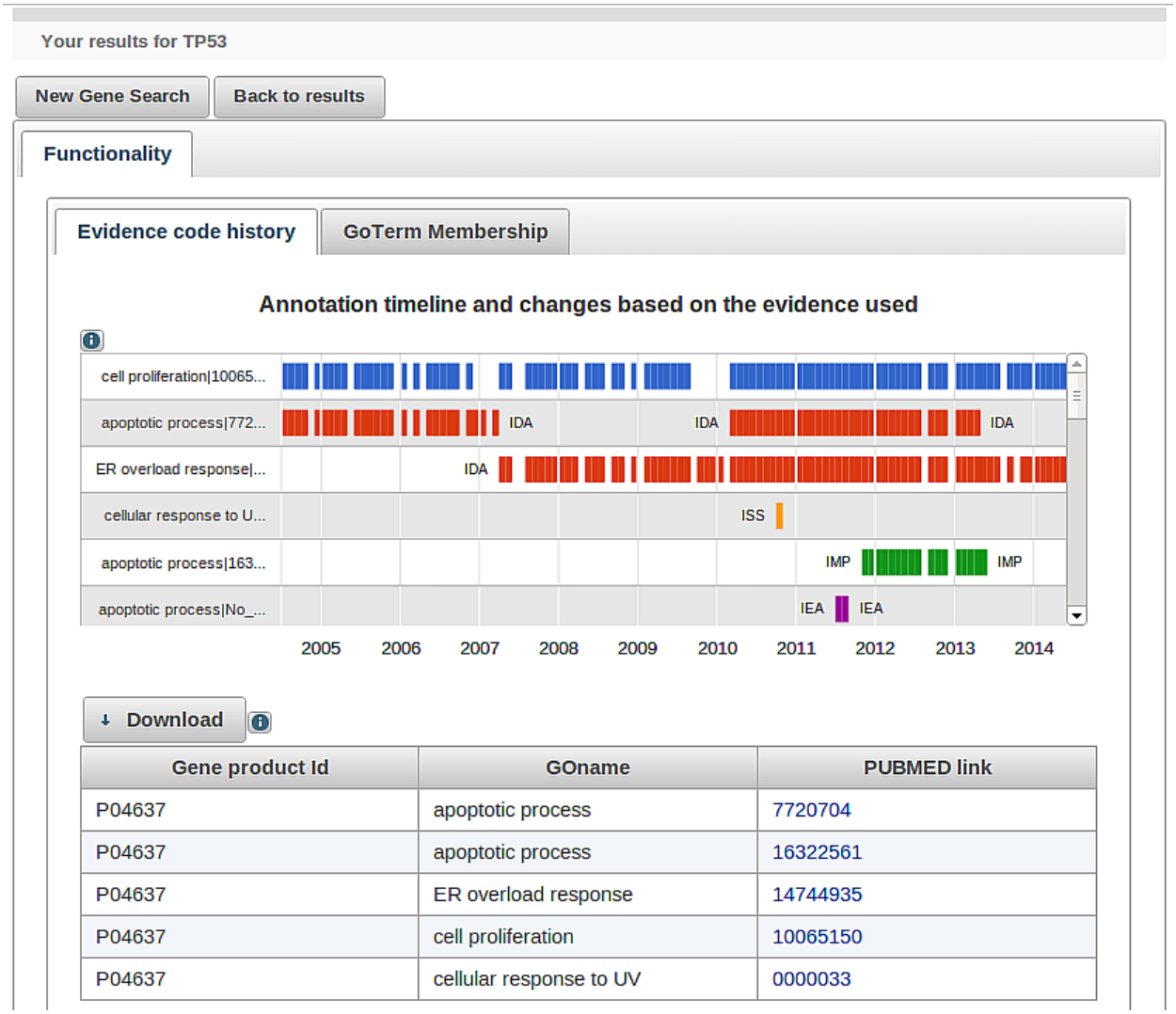



Figure 4.15: Explore different exploratory metrics for a gene product.

Additionally, the functionality tab is very handy as users are able to explore all the GO terms that have been ever annotated to the queried gene product and visualize the "existence" of each annotation across editions. A "time line" with monthly squares are displayed for users to visualize this existence. The colours of each monthly box represent the evidence code

that was used in that annotation at that time point. When the annotation was absent for a particular edition, a gap in the time line is shown. Even if the GO term is no longer used or if the annotation seems unstable, the user can have access to the sources used to support them by using the table displayed below (in case there is a PubMed ID). Likewise, users can download this information or even check the time line of how many gene products have been assigned to the respective GO terms at each time point (**Figure** 4.16).




| Gene product Id | GOname | PUBMED link |
|---|---|---|
| P04637 | apoptotic process | 7720704 |
| P04637 | apoptotic process | 16322561 |
| P04637 | ER overload response | 14744935 |
| P04637 | cell proliferation | 10065150 |
| P04637 | cellular response to UV | 0000033 |

Figure 4.16: Explore the historical existence of annotations associated to functions for a particular gene product of interest.

## 4.3 The Assessment of Gene Function Prediction Algorithms

As it was noted in the exploratory analyses, curation effort is skewed to certain model organisms and within each organism, prioritized gene sets seem to favor multifunctional genes. The results from the GO term membership also highlighted that mostly shallow GO terms are assigned to annotation data and that often IEA annotations are preferred in manual revisions.

As most of the genes considered in the CAFA assessment already have at least an electronically inferred function assigned to them (mostly by sequence alignment or other similar methods), it is likely that for the CAFA assessment, manual annotations that are likely to come up in the "accumulation period" (and most likely used for the assessment), will reflect IEA upgrades.

Taken all these factors into account, a submission was made for the CAFA2 assessment by considering the GO term prevalence as the null model, and assigning the most prevalent terms (excluding those already assigned manually) to every single target. Additionally, GO terms that are frequently updated or commonly co-occurred were also assigned as "predictions" if their probability score was higher than what could be assigned by prevalence alone.

To assess the performance of this set of methods, which basically can be attributed to annotation artifacts and do not consider any biological reality of the targets, I reproduced the predictions by using in this case the targets included in the gold standard set of the CAFA1 assessment.

In general, co-occurrence and IEA upgrade showed a small contribution on top of prevalence. Specifically, on average across all the evaluation targets, 15% of the predictions were derived from co-occurrence and 84% were derived from prevalence. Contrary to what was expected, only 1% of the

predictions were assigned by IEA upgrade, as their probabilities rarely improved those from prevalence. When using the gold standard set reported in the CAFA1 paper to assess how many of our predictions became true positives, an average of 10 (true positive) annotations and a maximum of 125 annotations were derived from prevalence, 0 on average and a maximum of 2 annotations were derived from IEA upgrades and 1 on average and 41 as the maximum number of annotations were derived from co-occurrence. These trends didn't differ when combining prevalence with only one of the two other methods. The apparent utility derived from IEA upgrades might in part be just a reflection of the closed world assumption of the evaluation, a limitation that has already been criticized [15] and recently explored [57]. In fact, many IEA annotations are likely to be accurate even if they are not directly upgraded by curators in the time frame used for the evaluation.

To further assess the relative "performance of these methods, I used the 18 sets of predictions that were provided to Jesse Gillis and Paul Pavlidis from the organizers of the CAFA1 assessment for an independent evaluation. They also provided the "predictions" from the prevalence set used as a control,as well as those derived from the GOtcha and BLAST methods. When exploring the performance of the proposed methods in a function-centric evaluation versus the others, it was clearly observed that regardless of the ontology, the combined method performed better than prevalence alone and was comparable to others. GOtcha and BLAST were the top performing methods. An alternative "gold standard" was used by including GO terms that have 10 to 100 genes assigned, but the results didn't show any significant differences between using this or the CAFA1 gold standard (**Figure** 4.17; **Table** 4.2).

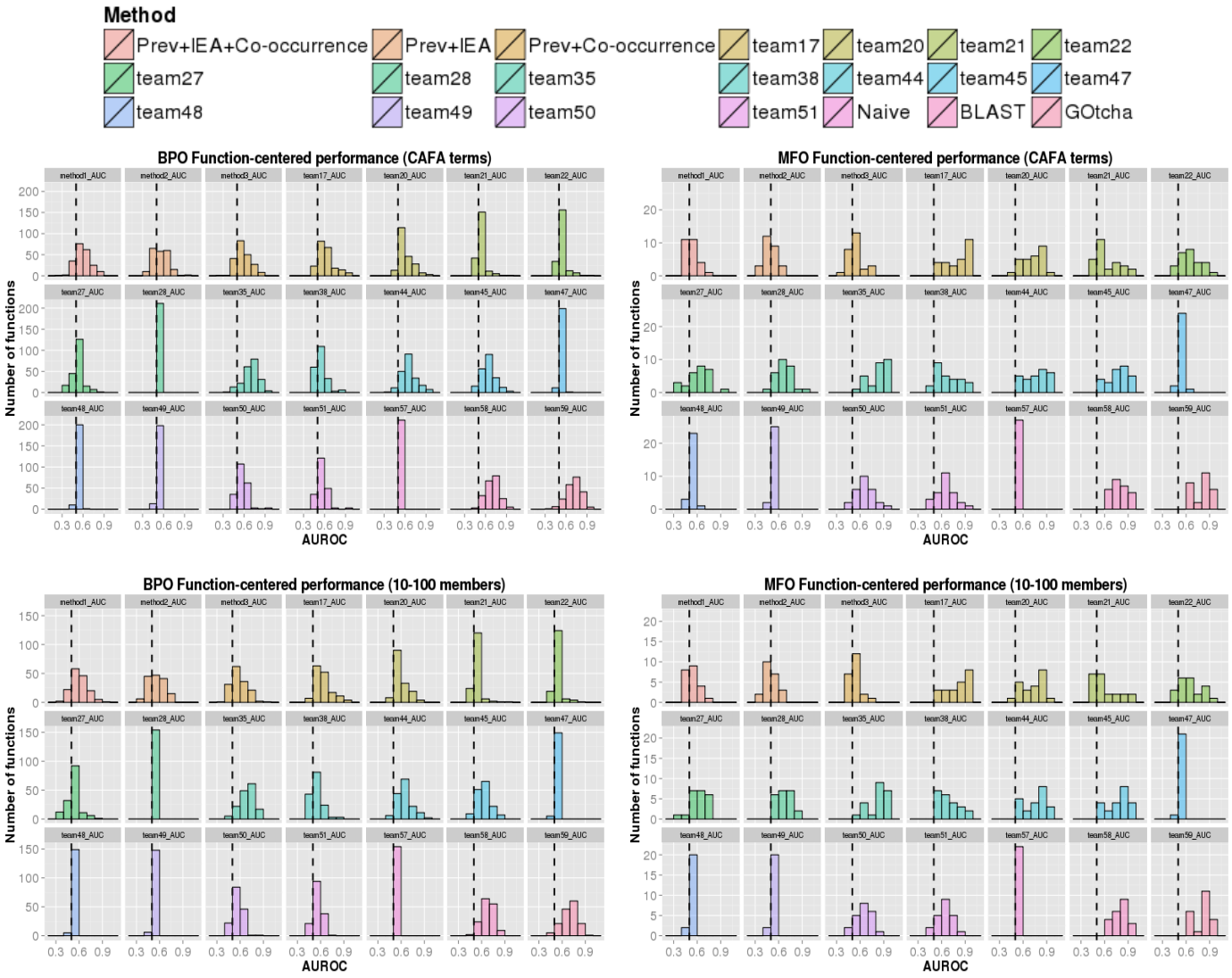



Figure 4.17: Results of the performance of function-prediction algorithms as measured by AUROC. A and B show performance using GO terms present in the gold standard list; C and D show performance using GO terms that had 10 to 100 genes assigned.

Table 4.2: Results of the function-centered performance as measured by AUROC.

| **Ontology** | **Number of GO** (terms out of 18974) | **Method** for **AUC analysis** | **Team-data** | **AUC value** (Prevalence filtered) |
|---|---|---|---|---|
| BP | 154 | GOTerms considered if they had 10-100 members | BLAST | 0.680 |
| | | | Gotcha | 0.698 |
| | | | Naive Method | 0.500 |
| | | | Prevalence + IEA + co-occurrence | 0.599 |
| | | | Prevalence + IEA | 0.557 |
| | | | Prevalence + co-occurrence | 0.588 |
| | | | Top score | 0.698 |
| | | | Lowest score | 0.500 |
| | | | Average (all teams) | 0.571 |
| | 211/233 | All GOterms considered in the gold standard | BLAST | 0.702 |
| | | | Gotcha | 0.715 |
| | | | Naive Method | 0.500 |
| | | | Prevalence + IEA + co-occurrence | 0.595 |
| | | | Prevalence + IEA | 0.557 |
| | | | Prevalence + co-occurrence | 0.590 |
| | | | Top score | 0.715 |
| | | | Lowest score | 0.500 |
| | | | Average (All teams) | 0.575 |
| MF | 22 | GO terms considered if they had 10-100 members | BLAST | 0.798 |
| | | | Gotcha | 0.815 |
| | | | Naive Method | 0.500 |
| | | | Prevalence + IEA + co-occurrence | 0.540 |
| | | | Prevalence + IEA | 0.508 |
| | | | Prevalence + co-occurrence | 0.538 |
| | | | Top score | 0.831 |
| | | | Lowest score | 0.500 |
| | | | Average (All terms) | 0.688 |
| | 27/28 | All GOTerms present in the gold standard | BLAST | 0.792 |
| | | | Gotcha | 0.813 |
| | | | Naive method | 0.500 |
| | | | Prevalence + IEA + co-occurrence | 0.530 |
| | | | Prevalence + IEA | 0.505 |
| | | | Prevalence + co-occurrence | 0.545 |
| | | | Top score | 0.831 |
| | | | Lower score | 0.500 |
| | | | Average (All terms) | 0.671 |

Regardless of what was considered the "gold standard", our methods had a stronger performance compared to prevalence alone. Performance was similar to the average performance across the other algorithms considered, but could not outperform BLAST or GOTcha.

I was not able to reproduce the results of the "CAFA score" used for the original CAFA1 assessment. The results computed were highly discordant with what was reported in the publication and the reason could be that I did not have all the results of the algorithms submitted, but just a subset. The descriptions to compute the metrics were also not entirely clear on the original publication. I took an alternative approach to evaluate the performance as described by Gillis and Pavlidis (2013). The evaluation is based on the semantic similarity between the predictions and the true assignments in the gold standard. This metric allows the exploration of how informative a prediction is, in particular, by looking at those predictions that had a higher probability score than what could be assigned by prevalence alone.

Using information content (IC) as a measure of term specificity is adequate for this purpose because of the shallow annotation problem. The metric proposed by Resnik was used to explore this. The results obtained from the molecular function ontology showed that, on average, 14.17% of the predictions were more informative than prevalence across all methods. The proposed controls yielded 16.08% informative predictions and from the datasets, the top performing method (labelled as team20) had 30% informative predictions. Contrary to the function-centered evaluation, BLAST and GOtcha were not the top most informative, but yielded between 14-15% informative predictions (**Table** 4.3).

Even though all the methods assessed here seemed to improve the baseline set with prevalence alone, some of them had a very low performance. The results showed that some -but not all- the algorithms can make correct and specific predictions. Such percentages, however, can be affected when some rarely used generic terms in GOA (and thus, not that prevalent) acquire a high IC score but do not necessarily reflect the specificity of the term [58].

Table 4.3: Results of the function-centered performance measured by information content (molecular function ontology).

| Method | Total Predictions | Total Informative predictions | Percentage |
|---|---|---|---|
| Prevalence+IEA+Co-Occurrence | 690 | 111 | 16.1 % |
| Prevalence+IEA | 690 | 79 | 11.4 % |
| Prevalence+Co-Occurrence | 690 | 110 | 15.9 % |
| GOtcha | 690 | 105 | 15.2 % |
| BLAST | 690 | 102 | 14.8 % |
| Naive | 690 | 0 | 0.0 % |
| Team_51 | 690 | 118 | 17.1 % |
| Team_50 | 690 | 117 | 17.0 % |
| Team_49 | 690 | 1 | 0.1 % |
| Team_48 | 690 | 13 | 1.9 % |
| Team_47 | 690 | 12 | 1.7 % |
| Team_45 | 690 | 134 | 19.4 % |
| Team_44 | 690 | 184 | 26.7 % |
| Team_38 | 690 | 97 | 14.1 % |
| Team_35 | 690 | 133 | 19.4 % |
| Team_28 | 690 | 116 | 16.8 % |
| Team_27 | 690 | 15 | 2.2 % |
| Team_22 | 690 | 147 | 21.3 % |
| Team_21 | 690 | 131 | 19.0 % |
| Team_20 | 690 | 208 | 30.1 % |
| Team_17 | 690 | 120 | 17.4 % |

In contrast, when looking at the results obtained for the biological process ontology, it was noted that on average, only 6.3% of the predictions were more informative than prevalence. Our methods, similar to MF, yielded 17% informative predictions and again, team20 was the most informative one, but only with 12.73%. In this case, GOtcha only yielded 6% and BLAST performed better with 12% (**Table** 4.4).

Table 4.4: Results of the function-centered performance as measured by information content (biological process ontology).

| Method | Total Predictions | Total Informative predictions | Percentage |
|---|---|---|---|
| Prevalence+IEA+Co-Occurrence | 1186 | 206 | 17.4 % |
| Prevalence+IEA | 1186 | 93 | 7.8 % |
| Prevalence+Co-Occurrence | 1186 | 203 | 17.1 % |
| GOtcha | 1186 | 73 | 6.2 % |
| BLAST | 1186 | 142 | 12.0% |
| Naive | 1186 | 0 | 0.0 % |
| Team_51 | 1186 | 35 | 3.0 % |
| Team_50 | 1186 | 33 | 2.8 % |
| Team_49 | 1186 | 0 | 0.0 % |
| Team_48 | 1186 | 2 | 0.2 % |
| Team_47 | 1186 | 5 | 0.4 % |
| Team_45 | 1186 | 76 | 6.4 % |
| Team_44 | 1186 | 139 | 11.7 % |
| Team_38 | 1186 | 49 | 4.1 % |
| Team_35 | 1186 | 91 | 7.7 % |
| Team_28 | 1186 | 0 | 0.0 % |
| Team_27 | 1186 | 4 | 0.3 % |
| Team_22 | 1186 | 59 | 5.0 % |
| Team_21 | 1186 | 78 | 6.6 % |
| Team_20 | 1186 | 151 | 12.7 % |
| Team_17 | 1186 | 89 | 7.5 % |

The results obtained are comparable to those reported by Gillis and Pavlidis (2013) and also highlight the large impact that prevalence has on the assessment. Similarly, other patterns that can be attributed to biases in the annotation process, such as gene multifunctionality, or commonly annotated terms should be considered in critical assessments such as CAFA.

However, as it was described earlier, such artifacts are now considered priors for prediction methods, when in fact, they do not take into account any meaningful biological information. In particular, the results from the IC evaluation showed that only scarce informative terms are assigned by the current function prediction algorithms and are equivalent to the performance obtained in the proposed benchmark, which is partially reflecting

the continuous problem of assigning shallow terms that might not answer the question of what a certain target gene do in a biological context. The question still remains on how to design an assessment not affected by such biases. However, the methods proposed in this thesis serve as a constructive baseline that any algorithm focused on function prediction should clearly outperform.

## 4.4 Instability of Gene set Enrichment Results

The last part of the study aimed to explore the impact that annotation instability has on the performance of gene set enrichment analyses. More than 2,000 experimentally-derived hit lists from MolSigDB were analysed for this matter. Previous studies have explored this variability on a small scale and suggested that changes in GO annotations have an important influence in the enrichment results [29]. However, to my knowledge, no analysis has been made on a large scale to identify their variability. An example of the enrichment results for one hit list from MolSigDB is given, at 3 different time points, to exemplify problems in reproducibility of the results, where the first time point indicated vesicle transport-related terms, whereas the third time point involved terms highly associated with organ development. Such a difference in the top enriched terms may impact the interpretation of results (**Figure** 4.18).

A biologist would prefer to consider terms that are "robust" regardless of the changes in the annotations. Members of the GOC recommend to use the latest version of GO/GOA (if possible) for analysis[32]. To explore the "robustness" of the data, I considered for each gene set, how many enriched GO terms reaching statistical significance (with a FDR$\leq$ 0.1) will overlap (at each time point) with those present in the "last edition" (July 2014).

The results showed a considerable variability in the number of significant GO terms in the "last edition" across data sets (from 1 to more than a 100). To facilitate the exploration, the sets were classified in arbitrary groups to explore the variability of those gene sets that contain less than 50 significant terms vs. those that have more than 150 significant terms (**Table** 4.5).

It is of interest to find the proportion of enrichment results that are likely to show some stability for a certain period of time. It is also expected to identify shifts in the results at certain time points that would correlate with reported changes in GO/GOA.

The variability in "motif gene sets" (C3 collection) between May vs. the

HIT LIST EXAMPLE

| | 11-2005 (ed. 36) | 05-2009 (ed 74) | 08-2014 (ed.136) |
|---|---|---|---|
| 1 | Clathrin-coated vesicle | heat shock protein binding | loop of Henle development |
| 2 | Trans-Golgi network transport vesicle | Structure-specific DNA binding | ATPase activator activity |
| 3 | coated vesicle | Clathrin-coated vesicle | regulation of inclusion body assembly |
| 4 | transport vesicle | in utero embryonic development | core promoter proximal region sequence-specific DNA binding |
| 5 | Golgi vesicle | secretion by cell | core promoter proximal region DNA binding |
| 6 | cytoplasmic membrane-bound vesicle | growth cone | ATPase regulator activity |
| 7 | cytoplasmic vesicle | site of polarized growth | regulation of tranforming growth factor beta production |
| 8 | NAD(P)+-protein-arginine ADP-ribosyltransferase activity | coated vesicle | ADP binding |
| 9 | endothelial cell differentiation | gene silencing | positive regulation of ATPase activity |
| 10 | axon cargo transport | steroid hormone receptor activity | nephron tubule development |
| 11 | Notch binding | Single-stranded DNA binding | kidney epithelium development |
| 12 | transcription coactivator activity | viral reproductive process | renal tubule development |
| 13 | keratinocyte differentiation | neuron projection | regulation of ATPase activity |
| 14 | response to metal ion | Ligand-dependent nuclear receptor activity | Chaperone-mediated protein folding |
| 15 | adenylate cyclase activity | myosin complex | cardiac septum morphogenesis |
| 16 | response to inorganic substance | regulation of gene expression, epigenetic | Ligand-activated sequence-specific DNA binding RNA polymerase II transcription factor acitivy |
| 17 | cell fate determination | morphogenesis of an epithelial sheet | cardiac cell development |
| 18 | myoblast differentiation | NAD(P)+-protein-arginine ADP-ribosyltransferase activity | direct ligand regulated sequence-specific DNA binding transcription factor activity |
| 19 | chromatin silencing complex | Serine-type carboxypeptidase activity | Trans-Golgi network |
| 20 | cAMP biosynthesis | bile acid biosynthetic process | energy reserve metabolic process |
| 21 | muscle cell differentiation | Substrate-bound cell migration, cell extension | neural crest cell development |
| 22 | cAMP metabolism | viral assembly, maturation, egress and release | RNA polymerase II core promoter proximal region sequence-specific DNA binding |
| 23 | protein import into nucleus, docking | transforming growth factor beta receptor, pathway-specific cytoplasmic mediator activity | mesenchyme development |
| 24 | chromatin silencing complex | cleavage furrow | steroid hormone receptor activity |
| 25 | heterochromatin formation | establishment of nucleus localization | histone lysine methylation |
| 26 | clathrin coat of trans-Golgi network vesicle | depyrimidination | negative regulation of protein catabolic process |
| 27 | cell fate commitment | positive regulation of B cell differentiation | glycogen metabolic process |
| 28 | negative regulation of gene expression, epigenetic | nerve growth factor binding | steroid hormone mediated signaling pathway |
| 29 | skeletal muscle development | protein phosphatase 2A binding | cellular glucan metabolic process |
| 30 | skeletal muscle fiber development | ventricular cardiac muscle cell development | actomyosin structure organization |
| 31 | muscle fiber development | Serine-type exopeptidase activity | glucan metabolic process |
| 32 | cytokinesis | viral reproduction | neural crest cell differentiation |
| 33 | subtilase activity | axon | actomyosin |
| 34 | regulation of cell migration | muscle cell differentiation | nephron epithelium development |
| 35 | regulation of locomotion | regulation of endopeptidase activity | myosin complex |
| 36 | regulation of behavior | 3-beta-hydroxy-delta5-steroid dehydrogenase activity | transcription initiation from RNA polymerase II promoter |
| 37 | regulation of cell motility | gene silencing by miRNA, production of miRNAs | core promoter sequence-specific DNA binding |

Figure 4.18: Example of a motif gene set at different time points showing problems in reproducibility and interpretation of results. The data corresponds to with genes with promoter regions around a Transcription Start Site containing the motif YTTCCNNNGGAMR. The motif does not match any known transcription factor. The top 37 enriched GO terms are shown and unstable terms are coloured for comparison.

last edition showed that 17% of the gene sets had less than 80% overlap in their enriched terms. In fact,in a few cases the overlap barely reached 40%. This raised the concern that some results can considerably in short periods of time (a couple of months). The pre-defined groups A and B,

Table 4.5: Classification of gene sets by the number of significant GO terms.

| **C2 collection (1870 total hit lists)** | | | **C3 collection (636 total hit lists)** | | |
|---|---|---|---|---|---|
| Group | # sigGOterms | # gene sets | Group | # sigGOterms | # gene sets |
| Group A | $\leq 10$ | 160(28) | Group A | $\leq 10$ | 163(40) |
| Group B | 11-50 | 386(26) | Group B | 11-50 | 178(48) |
| Group C | 51-100 | 377(9) | Group C | 51-100 | 139(14) |
| Group D | >100 | 947(1) | Group D | >100 | 156 (7) |

The table shows an arbitrary classification of the gene sets by considering the number of statistically significant GO terms in the "last edition". For example, gene sets that had 11 to 50 significant GO terms belong to Group B. For reference, the numbers in parenthesis show how many gene sets in each group had less than 80% overlap in May 2014 with those of July 2014.

which have less than 50 significant GO terms, showed the largest variability. Most gene sets (regardless of the group where they were classified) only showed a very small overlap before 2009, which matches the period of time where the Reference Genome Project started to revise and improve the quality of GO/GOA. A large variation was also observed between 2010 and 2011, which also coincides with major changes in GO annotations for human data, as described earlier. In general, most results will show more than 50% similarity after 2012. However, interesting outliers were also observed (**Figure** 4.19). A similar trend was observed in curated gene sets from online pathway databases (C2 collection), although these gene sets showed a higher percentage of overlap with the last edition, compared to C3 (**Figure** 4.20).

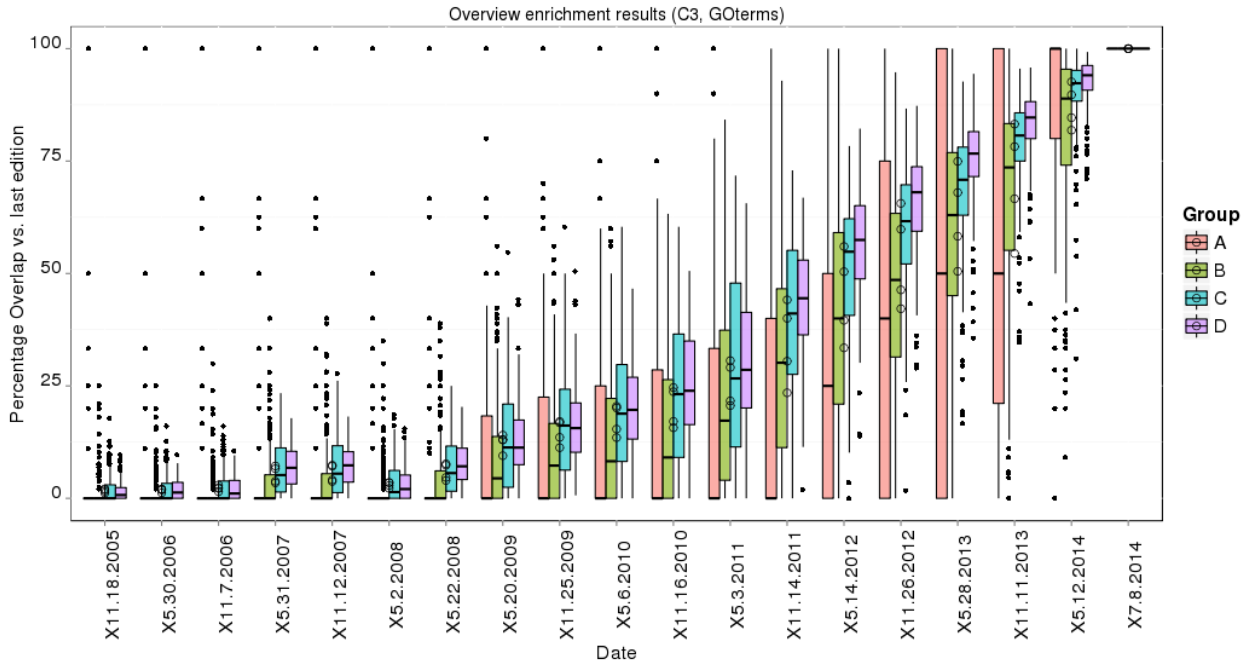



Figure 4.19: Variability of GO term overlap in the gene sets (C3). Figure shows that, gene sets with less than 50 significant terms, tend to show more variability after comparing different time points vs. the "last edition". Outliers were identified which showed almost no overlap after just a couple of months (May vs. July 2014).

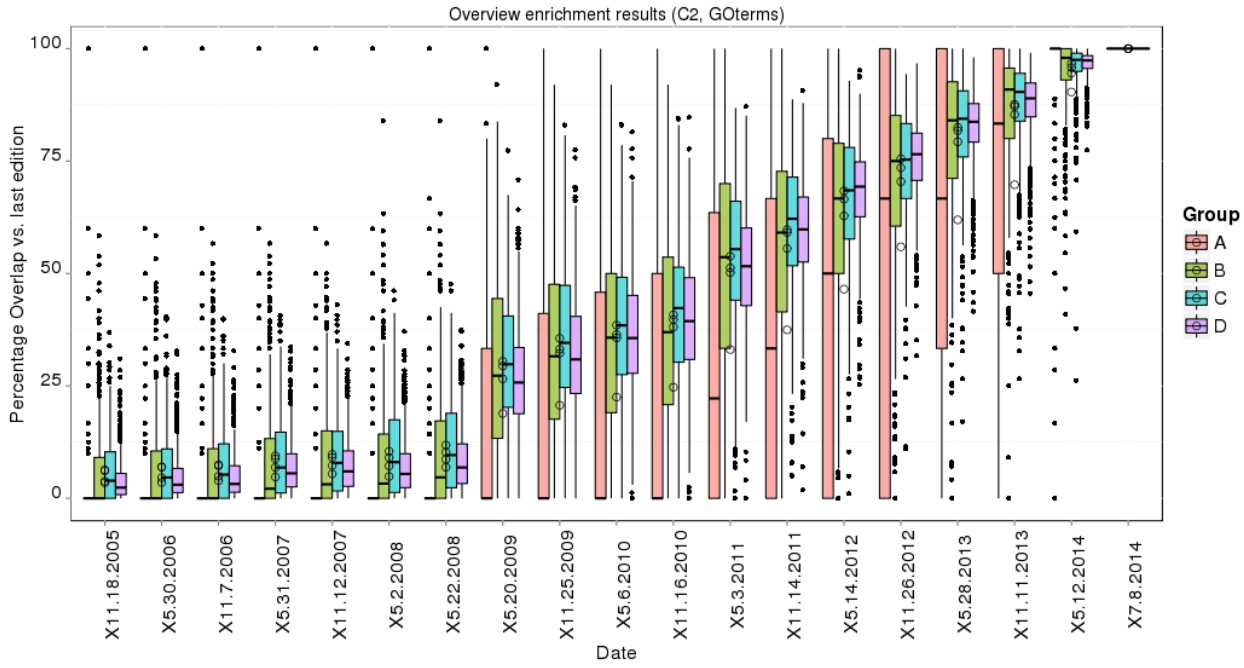



Figure 4.20: Variability of GO term overlap in the gene sets (C2).Figure shows that, gene sets with less than 50 significant terms(Group A and B), tend to show a small variability after comparing different time points vs. the "last edition". Outliers were identified with almost no overlap after just a couple of months (May vs. July 2014).

The previous results showed a considerable variation in the enrichment results for group A in C2 vs. C3, while the rest of the groups seemed to have a consistent variation. In general, some variability is expected, although the extent of this variation has not been explored on a larger scale until now.

The differences between the overlap of significant terms at different time points might not be substantial if the changes reflect an "improvement" or upgrade in the annotations. Hence, the gene sets are likely to contain more specific GO terms, which in turn, will share many parental terms with previous results. As long as the semantic similarity of the results is maintained,

the results could be considered consistent.

To further explore how similar the results are, the significant results were compared in terms of how semantically similar the significant GO terms are vs. the "last edition" (after propagation). One limitation of this method is that the changes in the GO structure are not considered. However, according to the results of Clarke *et al* [29] and Gross *et al* [59], changes in the GO structure have a smaller influence in the instability of the results compared to the effect of changes in GOA. For comparative purposes, the gene sets were also grouped by an arbitrary classification, defined by the number of parents belonging to the significant terms from the "last edition" (**Table** 4.6).

Table 4.6: Classification of gene sets by the number of parental terms.

| **C2 collection (1870 total hit lists)** | | | **C3 collection (636 total hit lists)** | | |
|---|---|---|---|---|---|
| Group | # Parental terms | # gene sets | Group | # Parental terms | # gene sets |
| Group E | $\leq 50$ | 249(114) | Group E | $\leq 50$ | 167(79) |
| Group F | 51-150 | 406(140) | Group F | 51-150 | 158(86) |
| Group G | 151-250 | 335(87) | Group G | 151-250 | 129(53) |
| Group H | >250 | 1026(82) | Group H | >250 | 182(29) |

The table shows an arbitrary classification of the enrichment results by considering, for each gene set, the number of parental terms linked to the significant GO terms in the "last edition". The numbers in parenthesis show how many of the enrichment results in each group had a semantic similarity value of 80% or less between May 2014 and July 2014.

The results showed a notable difference between the similarity of the results from 2013 vs. the "last edition" in the motif gene sets collection (C3). For the curated gene sets (C2), the overall similarity remained slighly higher at each time point (**Figures** 4.21 and 4.22). The small groups (E and F) had the largest variability across gene sets. The outliers observed on the bottom side of May 2014 also demonstrate that, in some cases, the results are highly discordant between editions and thus, we might be getting a completely different result. On the contrary, the outliers observed in the upper side of the groups from 2009-2010 show that, in some cases, the gene sets are highly similar, extending the possibility of having a "robust result" after 5 years. Changes in the GO structure are also likely to influence these

measurements.

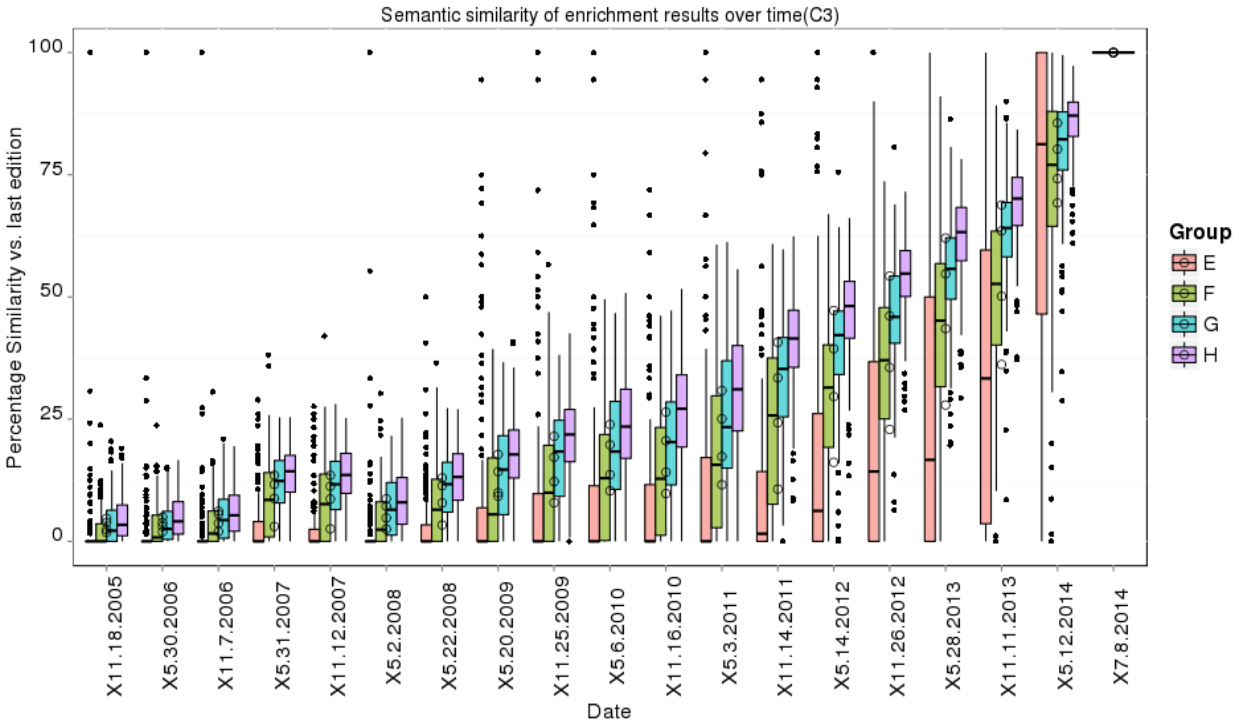



Figure 4.21: Semantic similarity of enriched gene sets by group (C3).Figure shows that gene sets with more than 150 parental terms tend to show less variability in their semantic similarity. Outliers were detected showing almost no overlap after just a couple of months (May vs. July 2014).

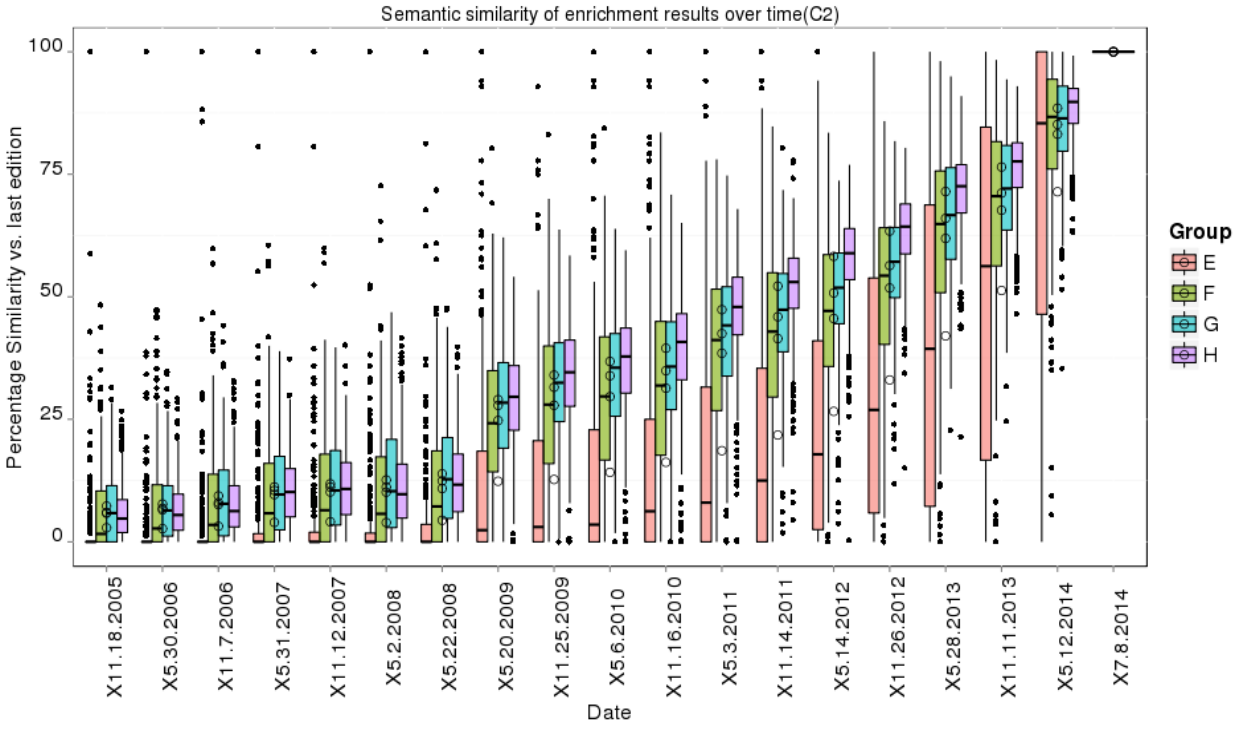



Figure 4.22: Semantic similarity of enriched gene sets by group (C2).Figure shows that gene sets with more than 150 parental terms tend to show less variability in their semantic similarity. Outliers were detected showing almost no overlap after just a couple of months (May vs. July 2014).

Even if the gene sets change, a biologist might be more interested in looking at the genes responsible to support those results and formulate further hypotheses. It is then relevant to assess whether the same genes are actually supporting the significant results, even if the GO terms change. If the overlap is high, then the groups can be considered highly similar (however, they might likely also be supported by multifunctional genes). If the overlap is low, the actual functional result has changed, and this can be likely be due to changes in the size of the GO groups in GOA.

The results showed a similar trend to the other two metrics. In particular, the variability observed for gene sets with less than a thousand genes ranged from 0 to a 100% overlap in the years 2012 and 2013. Compared to the results from May 2014, most of the results showed a considerable overlap,but outliers would also cover the entire range. Even for the years 2005-2010, outlier gene sets were found to have a high overlap, but most of them seemed to be completely different results. Those gene sets supported by more than 4,000 genes showed a high percentage of overlap and a noticeably smaller variability. This reflects that GO groups with a high number of genes are likely to contain those present in the hit lists and often show in the results of the enrichment analysis, although most of the results are derived from GO groups with 1000 or less genes (**Figures** 4.23 and 4.24).

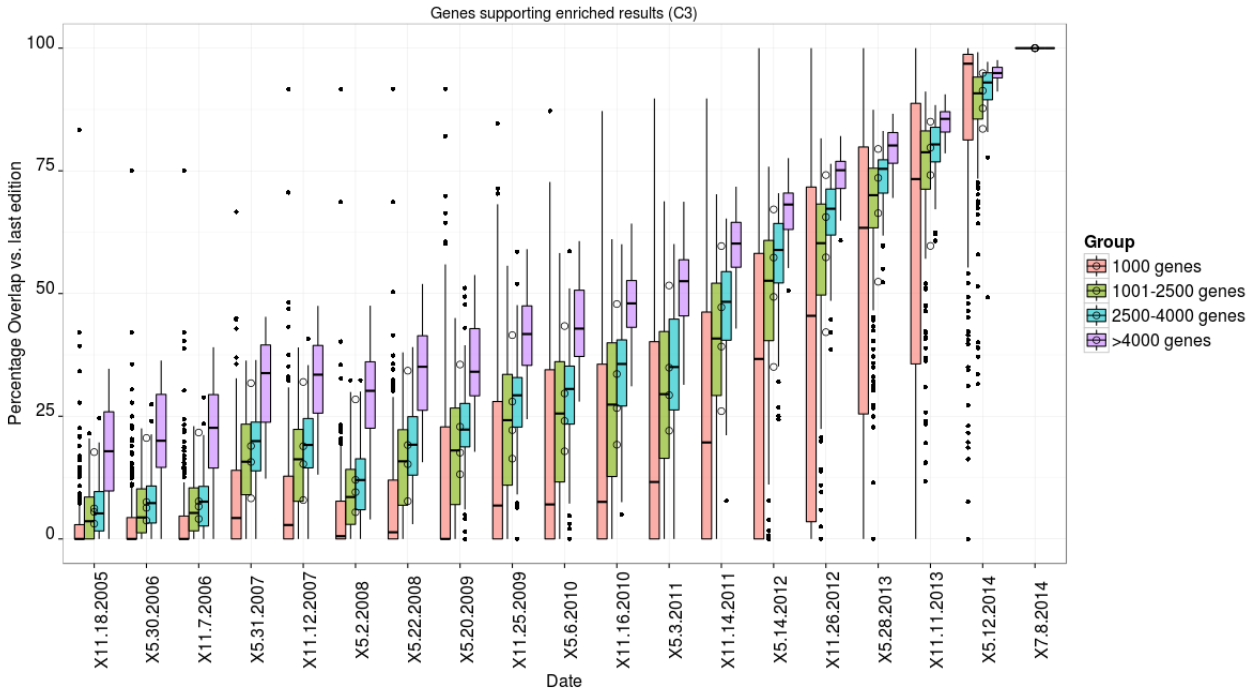



Figure 4.23: Percentage of overlapped genes in the gene sets from C3. Gene sets were grouped by number of genes supporting their significant terms in the last edition.

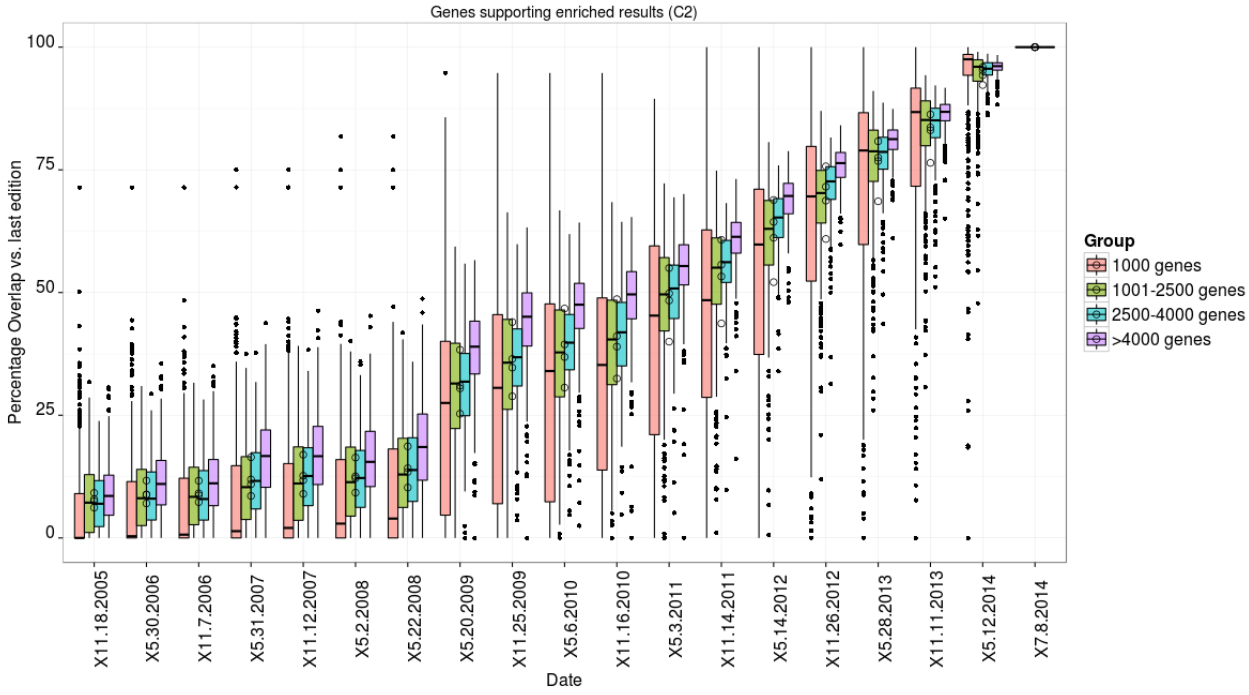



Figure 4.24: Percentage of overlapped genes in the gene sets from C2. Gene sets were grouped by number of genes supporting their significant terms in the last edition.

Taken together, these results show a considerable number of gene sets with a high degree of variability, even in small time frames. Some terms might disappear in future analyses, influenced by the effect of changes in the annotations, which was clearly correlated with the changes observed in these results at particular time points. The interpretation of all the results derived from gene set enrichment analyses should be considered with caution and used as a complementary exploration rather than a "conclusive result", specially as the reproducibility of results for studies that are older than 5 years seemed to be jeopardized in most cases. Enrichment tools that use GO annotations older than 2009 (like DAVID) might then display completely different results to what could be obtained using current GO annotations.

# Chapter 5

# Future Directions

There is clearly a great deal of interest in better understanding GO and GOA both among biologists who use it and even among GOC itself. In this thesis I have presented results obtained from the integration of historical GO Annotation data for 14 different organisms. I showed differences in the annotation patterns for different species and built a tool to track and extend this analyses to the research community. The assessment of GO Annotation instability is still challenging, but the work presented here provides an overall panorama of how annotation data is evolving. By nature, each change is dependent on decisions made by the GOC and a constant evolution of the databases they rely on, such as UniProtKB; which in turn, limits the feasibility of assigning predictive scores for future annotation instability. Such decisions also impact the traceability of protein annotations, specially when previous identifiers are removed or de-merged or new annotation pipelines are implemented. However, this study has addressed changes that had occurred in the 14 year history of GOA and filling important gaps in the assessment of annotation quality and instability. Different metrics were implemented and incorporated into a web-based tool, along with a baseline method that could be employed for the assessment of function prediction algorithms. While the web-based tool does reflect the existence of an annotation, it does not reflect time points where annotations are promoted. Likewise, the influence of annotation extensions and cross-references to other ontologies on GOA instability wasn't addressed. Future work in this regard would answer the question of whether such additions do contribute to the interpretability of GOA annotations or adds in a detrimental way to the problem of annotation instability. Likewise, incorporating these metrics and historical information into actual applications such as enrichment tools

will definitely contribute to the interpretability of the shared functions for further analyses. In particular, the final aim is that any user can submit a set of genes and obtain enrichment results over time for such dataset. The most stable or relevant GO terms and genes for their study could be then prioritized for further interpretation and analyses. Likewise, the possibility of a parallel exploration of the instability of the parental terms within the enrichment analyses should be considered.

In the meantime, this work altogether has been presented to the community of interest at the fourth annual CHiBi/GSAT retreat (UBC's Loon Lake Research and Education Centre, October 3-4,2013),the 3rd Annual Canadian Human and Statistical Genetics Meeting (Fairmont Empress Hotel Victoria B.C.,May 3-6, 2014), at the Bio-ontologies Special Interest Group and the Automated Function Prediction Interest Group in the SIG-Meetings and Intelligent Systems for Molecular Biology (ISMB) conference (Boston MA, July 11-15, 2014), receiving a positive feedback from the GOC, UniProtKB, users of gene-set enrichment tools and by participants from the CAFA assessment.

# Bibliography

[1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, pp. 25–29, May 2000.

[2] C. J. Mungall, M. Bada, T. Z. Berardini, J. Deegan, A. Ireland, M. A. Harris, D. P. Hill, and J. Lomax, "Cross-product extensions of the gene ontology," *Journal of Biomedical Informatics*, vol. 44, pp. 80–86, Feb. 2011.

[3] R. P. Huntley, M. A. Harris, Y. Alam-Faruque, J. A. Blake, S. Carbon, H. Dietze, E. C. Dimmer, R. E. Foulger, D. P. Hill, V. K. Khodiyar, A. Lock, J. Lomax, R. C. Lovering, P. Mutowo-Meullenet, T. Sawford, K. V. Auken, V. Wood, and C. J. Mungall, "A method for increasing expressivity of gene ontology annotations using a compositional approach," *BMC Bioinformatics*, vol. 15, p. 155, May 2014.

[4] H. Zhi-Liang, J. Bao, and J. Reecy, "Categorizer: a web-based program to batch analyze gene ontology classification categories," *Online J Bioinformatics*, vol. 9, pp. 108–112, 2008.

[5] F. M. McCarthy, N. Wang, G. B. Magee, B. Nanduri, M. L. Lawrence, E. B. Camon, D. G. Barrell, D. P. Hill, M. E. Dolan, W. P. Williams, *et al.*, "Agbase: a functional genomics resource for agriculture," *BMC genomics*, vol. 7, no. 1, p. 229, 2006.

[6] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock, "Go:: Termfinderopen source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes," *Bioinformatics*, vol. 20, no. 18, pp. 3710–3715, 2004.

[7] S. Carbon, A. Ireland, C. J. Mungall, S. Shu, B. Marshall, S. Lewis, *et al.*, "Amigo: online access to ontology and annotation data," *Bioinformatics*, vol. 25, no. 2, pp. 288–289, 2009.

[8] D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'Donovan, and R. Apweiler, "Quickgo: a web-based tool for gene ontology searching," *Bioinformatics*, vol. 25, no. 22, pp. 3045–3046, 2009.

[9] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez gene: gene-centered information at ncbi," *Nucleic acids research*, vol. 33, no. suppl 1, pp. D54–D58, 2005.

[10] G. O. Consortium *et al.*, "The gene ontology: enhancements for 2011," *Nucleic acids research*, vol. 40, no. D1, pp. D559–D564, 2012.

[11] N. Skunca, A. Altenhoff, and C. Dessimoz, "Quality of computationally inferred gene ontology annotations," *PLoS Comput Biol*, vol. 8, p. e1002533, May 2012.

[12] L. d. Plessis, N. kunca, and C. Dessimoz, "The what, where, how and why of gene ontologya primer for bioinformaticians," *Briefings in Bioinformatics*, vol. 12, pp. 723–735, Nov. 2011.

[13] R. P. Huntley, T. Sawford, M. J. Martin, and C. O'Donovan, "Understanding how and why the gene ontology and its annotations evolve: the GO within UniProt," *GigaScience*, vol. 3, no. 1, p. 4, 2014.

[14] U. Consortium *et al.*, "Update on activities at the universal protein resource (uniprot) in 2013," *Nucleic acids research*, vol. 41, no. D1, pp. D43–D47, 2013.

[15] C. Dessimoz, N. kunca, and P. D. Thomas, "CAFA and the open world of protein function predictions," *Trends in Genetics*.

[16] K. Canese, "Pubmed discovery tools," *NLM Tech Bull*, vol. 386, no. May-June, 2012.

[17] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, vol. 270, pp. 467–470, Oct. 1995.

[18] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble, "Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation," *Bioinformatics*, vol. 19, pp. 1275–1283, July 2003.

[19] D. W. Ussery and P. F. Hallin, "Genome update: annotation quality in sequenced microbial genomes," *Microbiology*, vol. 150, pp. 2015–2017, July 2004.

[20] M. E. Dolan, L. Ni, E. Camon, and J. A. Blake, "A procedure for assessing GO annotation consistency," *Bioinformatics*, vol. 21, pp. i136–i143, June 2005.

[21] C. Andorf, D. Dobbs, and V. Honavar, "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach," *BMC Bioinformatics*, vol. 8, p. 284, Aug. 2007.

[22] C. E. Jones, A. L. Brown, and U. Baumann, "Estimating the annotation error rate of curated GO database sequence annotations," *BMC Bioinformatics*, vol. 8, p. 170, May 2007.

[23] W. A. Baumgartner, K. B. Cohen, L. M. Fox, G. Acquaah-Mensah, and L. Hunter, "Manual curation is not sufficient for annotation of genomic databases," *Bioinformatics*, vol. 23, pp. i41–i48, July 2007.

[24] T. J. Buza, F. M. McCarthy, N. Wang, S. M. Bridges, and S. C. Burgess, "Gene ontology annotation quality analysis in model eukaryotes," *Nucleic Acids Research*, vol. 36, pp. e12–e12, Feb. 2008.

[25] A. Gross, M. Hartung, T. Kirsten, and E. Rahm, "Estimating the quality of ontology-based annotations by considering evolutionary changes," in *Data Integration in the Life Sciences* (N. W. Paton, P. Missier, and C. Hedeler, eds.), no. 5647 in Lecture Notes in Computer Science, pp. 71–87, Springer Berlin Heidelberg, Jan. 2009.

[26] The Reference Genome Group of the Gene Ontology Consortium, "The gene ontology's reference genome project: A unified framework for functional annotation across species," *PLoS Comput Biol*, vol. 5, p. e1000431, July 2009.

[27] G. Alterovitz, M. Xiang, D. P. Hill, J. Lomax, J. Liu, M. Cherkassky, J. Dreyfuss, C. Mungall, M. A. Harris, M. E. Dolan, J. A. Blake, and M. F. Ramoni, "Ontology engineering," *Nature Biotechnology*, vol. 28, pp. 128–130, Feb. 2010.

[28] Y. Alam-Faruque, R. P. Huntley, V. K. Khodiyar, E. B. Camon, E. C. Dimmer, T. Sawford, M. J. Martin, C. O'Donovan, P. J. Talmud, P. Scambler, R. Apweiler, and R. C. Lovering, "The impact of focused gene ontology curation of specific mammalian systems," *PloS One*, vol. 6, no. 12, p. e27541, 2011.

[29] E. L. Clarke, S. Loguercio, B. M. Good, and A. I. Su, "A task-based approach for gene ontology evaluation," *Journal of Biomedical Semantics*, vol. 4, p. S4, Apr. 2013.

[30] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 15545–15550, Oct. 2005.

[31] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic acids research*, vol. 37, pp. 1–13, Jan. 2009.

[32] J. A. Blake, "Ten quick tips for using the gene ontology," *PLoS Comput Biol*, vol. 9, p. e1003343, 11 2013.

[33] S. Y. Rhee, V. Wood, K. Dolinski, and S. Draghici, "Use and misuse of the gene ontology annotations," *Nature Reviews. Genetics*, vol. 9, pp. 509–515, July 2008.

[34] P. Khatri and S. Drghici, "Ontological analysis of gene expression data: current tools, limitations, and open problems," *Bioinformatics*, vol. 21, pp. 3587–3595, Sept. 2005.

[35] J. Gillis, M. Mistry, and P. Pavlidis, "Gene function analysis in complex data sets using ErmineJ," *Nature Protocols*, vol. 5, pp. 1148–1159, June 2010.

[36] J. Gillis and P. Pavlidis, "The impact of multifunctional genes on "guilt by association" analysis," *PLoS ONE*, vol. 6, p. e17258, Feb. 2011.

[37] J. Gillis and P. Pavlidis, "Assessing identity, redundancy and confounds in gene ontology annotations over time," *Bioinformatics (Oxford, England)*, vol. 29, pp. 476–482, Feb. 2013.

[38] C. Huttenhower, M. A. Hibbs, C. L. Myers, A. A. Caudy, D. C. Hess, and O. G. Troyanskaya, "The impact of incomplete knowledge on evaluation: an experimental benchmark for protein function prediction," *Bioinformatics (Oxford, England)*, vol. 25, pp. 2404–2410, Sept. 2009.

[39] W. K. Kim, C. Krumpelman, and E. M. Marcotte, "Inferring mouse gene functions from genomic-scale data using a combined functional network/classification strategy," *Genome Biology*, vol. 9, no. Suppl 1, p. S5, 2008.

[40] L. Pea-Castillo, M. Tasan, C. L. Myers, H. Lee, T. Joshi, C. Zhang, Y. Guan, M. Leone, A. Pagnani, W. K. Kim, C. Krumpelman, W. Tian, G. Obozinski, Y. Qi, S. Mostafavi, G. N. Lin, G. F. Berriz, F. D. Gibbons, G. Lanckriet, J. Qiu, C. Grant, Z. Barutcuoglu, D. P. Hill,

D. Warde-Farley, C. Grouios, D. Ray, J. A. Blake, M. Deng, M. I. Jordan, W. S. Noble, Q. Morris, J. Klein-Seetharaman, Z. Bar-Joseph, T. Chen, F. Sun, O. G. Troyanskaya, E. M. Marcotte, D. Xu, T. R. Hughes, and F. P. Roth, "A critical assessment of mus musculus gene function prediction using integrated genomic evidence," *Genome Biology*, vol. 9, p. S2, June 2008.

[41] P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, G. Pandey, J. M. Yunes, A. S. Talwalkar, S. Repo, M. L. Souza, D. Piovesan, R. Casadio, Z. Wang, J. Cheng, H. Fang, J. Gough, P. Koskinen, P. Trnen, J. Nokso-Koivisto, L. Holm, D. Cozzetto, D. W. A. Buchan, K. Bryson, D. T. Jones, B. Limaye, H. Inamdar, A. Datta, S. K. Manjari, R. Joshi, M. Chitale, D. Kihara, A. M. Lisewski, S. Erdin, E. Venner, O. Lichtarge, R. Rentzsch, H. Yang, A. E. Romero, P. Bhat, A. Paccanaro, T. Hamp, R. Kaner, S. Seemayer, E. Vicedo, C. Schaefer, D. Achten, F. Auer, A. Boehm, T. Braun, M. Hecht, M. Heron, P. Hnigschmid, T. A. Hopf, S. Kaufmann, M. Kiening, D. Krompass, C. Landerer, Y. Mahlich, M. Roos, J. Bjrne, T. Salakoski, A. Wong, H. Shatkay, F. Gatzmann, I. Sommer, M. N. Wass, M. J. E. Sternberg, N. kunca, F. Supek, M. Bonjak, P. Panov, S. Deroski, T. muc, Y. A. I. Kourmpetis, A. D. J. van Dijk, C. J. F. ter Braak, Y. Zhou, Q. Gong, X. Dong, W. Tian, M. Falda, P. Fontana, E. Lavezzo, B. Di Camillo, S. Toppo, L. Lan, N. Djuric, Y. Guo, S. Vucetic, A. Bairoch, M. Linial, P. C. Babbitt, S. E. Brenner, C. Orengo, B. Rost, S. D. Mooney, and I. Friedberg, "A large-scale evaluation of computational protein function prediction," *Nature Methods*, vol. 10, pp. 221–227, Mar. 2013.

[42] O. D. King, R. E. Foulger, S. S. Dwight, J. V. White, and F. P. Roth, "Predicting gene function from patterns of annotation," *Genome Research*, vol. 13, pp. 896–904, May 2003.

[43] D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, A. Maitland,

S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G. D. Bader, and Q. Morris, "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function," *Nucleic Acids Research*, vol. 38, pp. W214–W220, July 2010.

[44] EMBL-EBI, "European bioinformatics institute." `ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/old/`, 2014. 2002-2014. Fly and Yeast (07/2011-current).Last accessed: 2014-07-10.

[45] FlyBase, "Flybase repository." `ftp://ftp.flybase.org/releases/`, 2014. Files under precomputed files folder. Retrieved until 2011-06. Last accessed: 2014-07.

[46] SGD, "Sgd repository." `http://downloads.yeastgenome.org/curation/literature/archive/`, 2004. Files only until 2004. Last accessed: 2014-07.

[47] SGD, "Gene ontology repository." `http://cvsweb.geneontology.org/cgi-bin/cvsweb.cgi/go/gene-associations/gene_association.sgd.gz`, 2014. Files from 2004-01 until 2011-06. Last accessed: 2014-07.

[48] EcoCyc, "Gene ontology repository." `http://cvsweb.geneontology.org/cgi-bin/cvsweb.cgi/go/gene-associations/gene_association.ecocyc.gz`, 2014. 2008-2014.Last accessed: 2014-07-10.

[49] G. O. db, "Gene ontology repository." `ftp://ftp.geneontology.org/pub/go/godatabase/archive/full/`, 2014. termdb rdf xml files. Last accessed: 2014-07.

[50] U. Consortium, "Id mapping files by organism." `ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/`, 2014. 2002-2014. Last accessed: 2014-06-01.

[51] NLM, "Medline baseline repository." `http://mbr.nlm.nih.gov/Download/MUIDtoPMID/index.shtml`, 2013. 2004-2013. Last accessed: 2013-09-18.

[52] R. Balakrishnan, M. A. Harris, R. Huntley, K. Van Auken, and J. M. Cherry, "A guide to best practices for gene ontology (GO) manual annotation," *Database*, vol. 2013, pp. bat054–bat054, July 2013.

[53] J. Gillis and P. Pavlidis, "Characterizing the state of the art in the computational assignment of gene function: lessons from the first critical assessment of functional annotation (CAFA)," *BMC Bioinformatics*, vol. 14, p. S15, Apr. 2013.

[54] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "Rocr: visualizing classifier performance in r," *Bioinformatics*, vol. 21, no. 20, pp. 3940–3941, 2005.

[55] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, "proc: an open-source package for r and s+ to analyze and compare roc curves," *BMC bioinformatics*, vol. 12, no. 1, p. 77, 2011.

[56] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (msigdb) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 2011.

[57] Y. Jiang, W. T. Clark, I. Friedberg, and P. Radivojac, "The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective," *Bioinformatics*, vol. 30, pp. i609–i616, Sept. 2014.

[58] M. Mistry and P. Pavlidis, "Gene ontology term overlap as a measure of gene functional similarity," *BMC Bioinformatics*, vol. 9, p. 327, Aug. 2008.

[59] A. Gross, M. Hartung, K. Prfer, J. Kelso, and E. Rahm, "Impact

of ontology evolution on functional analyses," *Bioinformatics*, vol. 28, pp. 2671–2677, Oct. 2012.

# Appendix

I provide a list of the algorithms used and implemented for the elaboration of this thesis.

**General definitions**

Let $M$ a hash table, and $k$ a key of the hash map. We define $M(k)$ as the function that get all the elements mapped to $k$

Let $M$ be a Hash Table of type $< S, E >$, $k$ a key of type $S$ and e an element of type $E$ We define $Put(M, k, e)$ the function that creates the relationship $k- > e$ in $M$

**Program 5.1** This is the main structure of GOtrack, built to pre-process and analyse historical GO annotations.

1. Read user arguments.
2. Download new GOA files and TermDB files from the repository, collect the release date of each file.Output: edition2dates.txt.
3. Update DB Object IDs to current version and old MEDLINE IDs to PubMed IDs (mapping).
4. Create a list of all the GO terms and GO terms always present in the GO graph.
5. Create a file listing all DB Object IDs and DB Object IDs almost always present across editions with the user threshold. For each GOA file, get parental GO terms for each direct GO term annotated.
6. Retrieve the PubMed date of each publication. If possible, use the pre-computed file per species or query on website.
7. Track changes in the evidence codes assigned to each annotation. Output: evidencecodehistory.txt
8. Compute parameters using the EDITIONANALYSIS algorithm.
9. Count the number of GO terms annotated to each DB Object ID per GOA file. Output: countGenseperGoTerm.txt
10. Count the number of direct GO terms assigned to each DB Object ID per GOA file. Output: countDirectTermspergene.txt
11. Count the number of parental GO terms with "is a" and "part of" relationships to the direct GO term assigned to each DB Object ID per GOA file. Output: countInferredTermspergene.txt
12. Count the number of parental GO terms with "is a" and "part of" relationships to the direct GO term assigned to each DB Object Symbol per GOA file. Output: countInferredTermsperSymbol.txt
13. Generate a file with 4 columns: the DB Object ID, number of direct GO terms per DB Object ID, number of inferred GO terms per DB Object ID and the edition. This file will be loaded into the database.
14. Compute the JACCARD algorithm for all GOA editions.
15. Count the number of GO annotations that are replaced in future editions with a more granular GO term.

**Program 5.2** An algorithm run by GOtrack to compute one single edition.

1. Get the parental GO terms of each direct GO term in the GOA file.
2. Get a list of DB Object IDs in the GOA file. Output: genes.+edition+.txt
3. Compute the multifunctionality score per DB Object ID.
4. Create gomatrix file (GOMATRIX algorithm).

**Program 5.3** Program to create GOMatrix files. They list genes and the GO terms they are associated to on a particular edition.

**Data**: GenesAlmostAlwaysPresent.txt ($GAAP$): List of genes.
**Data**: Gene Association File ($GOA$): File with GO Annotations.
$hashGenesTerms \leftarrow$ a hash table that maps a gene $g$ to the set of GO terms annotated to it.
**for** *DB Object ID* $g \in GAAP$ **do**
    **if** $g \in GOA$ **then**
        **for** $term \in hashGenesTerms(g)$ **do**
            print DB Object Id + GOterm
    **else**
        print DB Object ID + "-1" ;
        (used to indicate that the DB Object ID is not in GOA)
Output: gomatrix.*.txt

**Program 5.4** An algorithm to compute semantic similarity based on Jaccard distance

```
Data: Gene Association Files (GOA): File with GO Annotations.
Data: gomatrix.*.txt files
for edition i ∈ 1... n do
    Let edA be the edition i;
    Let edB be the edition n;
    genesA ← all DB Object IDs in edA;
    genesB ← all DB Object IDs in edB(done only once);
    for each DB Object ID ∈ genesA do
        if g ∉ edB then
            sim ← −1//gene not in the last edition
        else
            goA ← GOterms associated to g in edA (gomatrix file);
            goB ← GOterms associated to g in edB (gomatrix file);
            jaccardScore ← (|goA ∩ goB|)/(|goA ∪ goB|);
            //jaccardScore is the similarity score for gene g in edition i
Output: jaccardpergeneovertime.txt
```

**Program 5.5** An algorithm to map old DB Object IDs to the most current version.

```
Data: IdMap: DB Object ID mapping for not Uniprot Ids
Data: Gene Association File (GOA): File with annotations.
Data: List of genes (allgenes.txt): File with all the DB object IDs
      annotated across all editions of GOA.
Data: Genes Always Present (GAP): List of DB Object IDs present
      in all GOA.
Data: Genes Last Edition (GLE): List of DB Object IDs present in
      the current GOA.
Result: GOA files with updated DB Object IDs
//Create the most updated version of the IDs
for gene ∈ allgenes.txt do
    IdWebMap ← value returned by the Uniprot Website for gene
//Build a dictionary using a special edition of the GOA files
s ← pre-selected edition for the current species
for annot ∈ GOA_s do
    customDic ← hash table that maps the DB Object ID to the DB
    Object Symbol and the synonyms present in annot
//Update IdMap and customDic using IdWebMap
for match ∈ IdMap ∪ customDic do
    if match points to a different id in IdWebMap then
        update match
for GOA_i ∈ GOA do
    for annot ∈ GOA_i do
        Get DB Object ID and synonyms;
        //For Ecoli the symbol is used ;
        if DB Object ID is already an Uniprot ID then
            if DB Object Id ∈ GAPorGLE then
                Leave current DB Object ID;
            else if DB Object Id ∈ customDic then
                Replace old DB Object ID with new one;
            else
                Search synonyms in annot and replace DB Object ID
                with the most prevalent candidate;
        else if DB Object ID is in IdMap then
            Replace old DB Object ID with matching DB Object ID;
        else
            Search synonyms in annot and replace DB Object ID with
            the most prevalent candidate;
        if DB Object Id IdWebMap then
            Update the If with the latest version in idWebMap;
        Write to *syn file the latest version of annot
    Rename the *syn files to *gz
```

**Program 5.6** An algorithm to map old MEDLINE IDs to current PubMed IDs

**Data**: Dictionaries (MuID-PmID.ids.gz): Files with conversion information from older IDs used in MEDLINE to new PubMed IDs.
**Data**: Gene Association Files (GOA): Files with GO Annotations
**Result**: GOA files with updated PubMed IDs.
Procedure;
Read Dictionaries.
**for** $GOA_i \in GOA$ **do**
  **for** $annot \in GOA_i$ **do**
    $PubMedIdgets$ PubMed id annotated in *annot*;
    **if** *PubMedId is in Dictionaries* **then**
      Replace DB Reference from *annot* with new DB Reference in a new copy (*syn) of the GOA file;
  Rename file *syn to *gz;

**Program 5.7** An algorithm to load the information to the database

**Data**: Go Tree files (termdb):
**Data**: Gene Association Files (GOA): Files with GO Annotations
**Data**: countGenesPerGoTerm.txt: Contains the information about the number of genes that are annotated to a go term
Procedure;

1. Load the GO term names in the termdb files
2. Load the ¡species¿_count table
3. Load number of genes per GO term
4. Load evidence code history
5. Load the relationship of GOA file and the publication date
6. Load the DB Object Ids that were updated
7. Load the GOA files
8. Load the analysis of annotations
9. Execute post load procedures

**Program 5.8** CAFA main algorithm

**Data**: go.*.txt : Prevalence GO terms
$ids \leftarrow$ hashset that maps $partialId$ to $CafaId$ ;
$annot \leftarrow$ read annotation file, create hash map that associates a $CafaId \in ids$ to arrayList of annotations ;
$annot \leftarrow$ read children and parents for each annotation ;
$topGoterms \leftarrow$ read prevalence goterms $go.*.txt$ ;
**for** $gene \in annot$ **do**
    $predictionsForThisGene \leftarrow$ will save all predictions for this target ;
    $predictionsForThisGene \leftarrow$ Call $predictGoTerms$ if method $Cooccurrence$ is active ;
    $predictionsForThisGene \leftarrow$ add all items in $topGoterms$ if method $prevalence$ is active ;
    $predictionsForThisGene \leftarrow$ Call $predictGoTerms$ if method $IEAupgrade$ is active ;
    $predictionsForThisGene \leftarrow$ remove dups, if any, also order predictions by score.;
    **if** *two or more methods predict the same GO term* **then**
        take the one with highest score
    **if** *A prediction is already manually annotated to gene* **then**
        Don't take a prediction
    print only the first 1500 predictions in $predictionsForThisGene$