BIOINFORMATICS ORIGINAL PAPER
Vol. 28 no. 20 2012, pages 2671–2677
doi:10.1093/bioinformatics/bts498

Databases and ontologies
Advance Access publication September 6, 2012

Impact of ontology evolution on functional analyses

Anika Groß[1,*], Michael Hartung[1], Kay Prüfer[2], Janet Kelso[2] and Erhard Rahm[1]

[1]Department of Computer Science, University of Leipzig, P.O.Box 100920, 04009 Leipzig and [2]Department of Evolutionary Genetics, Max-Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Ontologies are used in the annotation and analysis of biological data. As knowledge accumulates, ontologies and annotation undergo constant modifications to reflect this new knowledge. These modifications may influence the results of statistical applications such as functional enrichment analyses that describe experimental data in terms of ontological groupings. Here, we investigate to what degree modifications of the Gene Ontology (GO) impact these statistical analyses for both experimental and simulated data. The analysis is based on new measures for the stability of result sets and considers different ontology and annotation changes.

Results: Our results show that past changes in the GO are non-uniformly distributed over different branches of the ontology. Considering the semantic relatedness of significant categories in analysis results allows a more realistic stability assessment for functional enrichment studies. We observe that the results of term-enrichment analyses tend to be surprisingly stable despite changes in ontology and annotation.

Contact: gross@informatik.uni-leipzig.de

Supplementary information: Supplementary Data are available at *Bioinformatics* online.

Received on April 10, 2012; revised on July 3, 2012; accepted on August 8, 2012

1 INTRODUCTION

Ontologies are increasingly used in the life sciences (Bodenreider and Stevens, 2006; Lambrix *et al*., 2007). They provide a uniform vocabulary to describe and structure a domain of interest. As a prime example, the Gene Ontology (GO) Consortium (2008) contains knowledge about biological processes (BPs), molecular functions (MFs) and cellular components (CCs). The categories of GO are presented as a directed acyclic graph (DAG), with the edges representing relationships between the categories. Higher-level categories represent more abstract descriptions and encompass all-child categories. These categories are used to semantically describe genes and gene products (Thomas *et al*. 2007). Genes are therefore described by the particular category to which they are annotated and, by extension, by all-parent categories that give a more abstract description. Applications that perform functional enrichment analysis of gene sets take advantage of this property to identify more general categories that contain significantly more signal than expected at random (Tilford and Siemers, 2009).

Ongoing scientific research provides new domain knowledge that needs to be incorporated into ontologies and annotations. In the case of GO, these changes are incorporated on a regular basis with regular public releases (Gene Ontology Consortium, 2008; Leonelli *et al*., 2011). This ontology evolution has been previously analyzed (Hartung *et al*., 2008; Park *et al*., 2008; Pesquita and Couto, 2011) and yielded insights into the differences between ontology versions (Noy and Musen, 2002; Hartung *et al*., 2012b). Typical ontology changes include the addition of new categories and relationships as well as the revision of the existing structure (Hartung *et al*., 2012b, 2008). These ontological modifications can trigger changes in the annotation (Gross *et al*., 2009), e.g. when a category is removed, the annotations need to be moved or deleted. Further, annotations may be edited to reflect new knowledge or to eliminate inconsistencies (Dolan *et al*. 2005).

The GO has evolved substantially since its inception in 2000. Its three sub-ontologies evolved at different rates and in different ways. Between 2007 and 2010, BP increased by about 70%, compared with CC (≈40%) and MF (≈20%) (see Supplementary Table S1). Applying the method described in Hartung *et al*. (2010), we can identify how much different parts of GO have evolved by aggregating the change intensity in subtrees. Figure 1 illustrates that different subsections of the GO-MF evolve differently. The non-uniform distribution of changes may have an effect on functional enrichment analyses.

The aforementioned studies only considered changes in the ontology while neglecting the potential effects on downstream analyses; for example, how these changes may lead to different results in functional enrichment analyses. While it is rather obvious that the high degree of occurred changes in GO and its annotations will impact analysis results, it is still unknown whether earlier findings are significantly affected or even invalidated. The impact of ontology changes on functional enrichment analyses may depend on where the changes are located in the ontology and what kind of changes dominate. For example, additions of categories at the leaf level might be less critical than structural revisions within the ontology.

We provide a method to test to what degree changes of GO and GO annotations (GOAs) may affect functional enrichment analyses. We demonstrate the applicability and usefulness of our approach by analyzing two real-world experimental datasets as well as 50 random datasets. For the experimental datasets, we provide an in-depth study of the underlying changes and their impact on the analysis results. The presented analysis is informative for both ontology curators and users of functional enrichment methods.

*To whom correspondence should be addressed.

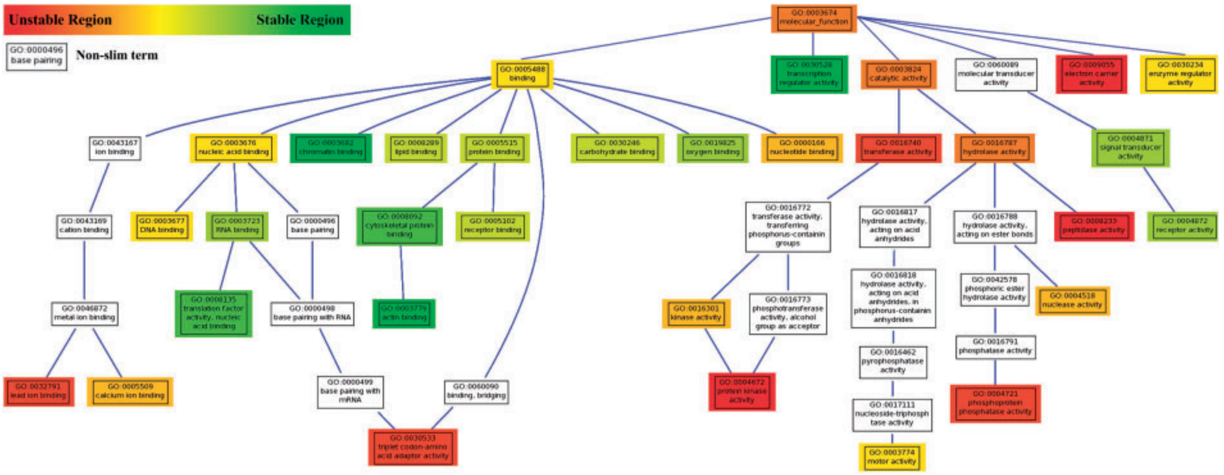

Unstable Region
Stable Region
GO:0000496 base pairing
Non-slim term
GO:0003674 molecular_function
GO:0005488 binding
GO:0030528 transcription regulator activity
GO:0003824 catalytic activity
GO:0060089 molecular transducer activity
GO:0009055 electron carrier activity
GO:0030234 enzyme regulator activity
GO:0043167 ion binding
GO:0003676 nucleic acid binding
GO:0003682 chromatin binding
GO:0008289 lipid binding
GO:0005515 protein binding
GO:0030246 carbohydrate binding
GO:0019825 oxygen binding
GO:0000166 nucleotide binding
GO:0016740 transferase activity
GO:0016787 hydrolase activity
GO:0004871 signal transducer activity
GO:0043169 cation binding
GO:0003677 DNA binding
GO:0003723 RNA binding
GO:0000496 base pairing
GO:0008092 cytoskeletal protein binding
GO:0005102 receptor binding
GO:0016772 transferase activity, transferring phosphorus-containin groups
GO:0016817 hydrolase activity, acting on acid anhydrides
GO:0016788 hydrolase activity, acting on ester bonds
GO:0008233 peptidase activity
GO:0004872 receptor activity
GO:0046872 metal ion binding
GO:0008135 translation factor activity, nucleic acid binding
GO:0000498 base pairing with RNA
GO:0003779 actin binding
GO:0016818 hydrolase activity, acting on acid anhydrides, in phosphorus-containin anhydrides
GO:0042578 phosphoric ester hydrolase activity
GO:0004518 nuclease activity
GO:0016301 kinase activity
GO:0016773 phosphotransferase activity, alcohol group as acceptor
GO:0016791 phosphatase activity
GO:0032791 lead ion binding
GO:0005509 calcium ion binding
GO:0000499 base pairing with mRNA
GO:0060090 binding, bridging
GO:0004672 protein kinase activity
GO:0016462 pyrophosphatase activity
GO:0004721 phosphoprotein phosphatase activity
GO:0017111 nucleoside-triphosph tase activity
GO:0030533 triplet codon-amino acid adaptor activity
GO:0003774 motor activity

Fig. 1. Evolution of slim terms in GO molecular functions between 2007 and 2010 (only partially shown, see Supplementary Figures 1–3 for BP, CC and MF visualization). We computed evolution intensities for slim terms by propagating changes from slim-term subtrees (Hartung *et al.* 2010). Colors denote evolution intensities of the slim terms. Non-slim term categories on the path to the root are coloured white. The graphic was produced using the visualization tool of Amigo (Carbon *et al.*, 2009)

2 METHODS

2.1 Ontology and annotation model

For our study, we use the GO, which is represented as a DAG, in which categories are linked by directed edges representing the relationships. Each category c has a unique identifier (e.g. 'blood coagulation' has the identifier GO:0007596). Various annotation sets A are associated with the GO. We use the GOA (Barrell *et al.*, 2009). Annotations to one category ($A(c)$) include annotations to c and to all its descendant categories (subgraph of c). The evidence supporting the annotation of a gene to a category is represented as an evidence code, which can be used to assess the origin and likely the quality of an annotation, e.g. experimentally verified or automatically generated.

Our evaluation requires consideration of multiple versions of ontologies and annotation sets. New versions always supersede the information of older versions. Furthermore, an annotation set is always associated with a particular ontology version. We denote an ontology version as O_v and an annotation version as A_v, where v stands for the date of release.

2.2 Ontology change detection

To understand the evolution of ontologies, we use a previously published *diff* algorithm (available in the CODEX web tool (Hartung *et al.*, 2012a)) to identify the changes that occurred between two versions of an ontology. The algorithm (Hartung *et al.*, 2012b) uses a set of rules to first identify basic changes (insert/update/delete) that are then aggregated into a smaller set of more complex (semantic) changes, such as merge, split or changes of entire subgraphs. The following ontology changes are relevant for our study:

- *addC* – addition of a category
- *toObsolete* – mark a category as obsolete
- *merge* – merge of two or more categories into one category
- *split* – split one category into two or more categories
- *substitute* – substitute one category by another
- *addR/delR* – addition/deletion of a relationship between two categories
- *move* – move a category from one parent to another parent category.

2.3 Term enrichment using FUNC

We use the program FUNC (Prüfer *et al.*, 2007) to carry out the functional enrichment analysis. FUNC tests each category of the input ontology for significance and then carries out randomizations to correct for multiple testing. For our further analysis, we consider the family-wise error rate corrected P-values associated with each category.

2.4 Stability measures

We propose two kinds of stability measures to assess the impact of ontology and annotation changes on the experimental results set of a functional enrichment analysis. For this purpose, we consider a fixed set of genes, and we compute experimental result set (ER) for different points in time with freely chosen ontology and annotation versions. We will use the example result sets displayed in Figure 2 to illustrate our stability measures.

2.4.1 Basic stability measure For our measures, we use the following cardinalities for result sets ER_i and ER_j produced with different ontology or annotation versions:

$|ER_i|, |ER_j|$ – *number of categories in* ER_i *and* ER_j

$|ER_i \cap ER_j|$ – *number of overlapping categories between* ER_i *and* ER_j

$|ER_i \setminus ER_j|$ – *number of categories only in* ER_i *but not in* ER_j

$|ER_j \setminus ER_i|$ – *number of categories only in* ER_j *but not in* ER_i.

Note that categories are different if their unique identifier (accession) differs. For determining the overlap, we count categories with identical accessions in both ER_i and ER_j.

The key idea for assessing the basic stability of a result set is the following. A result set is considered stable in comparison with an older set if both sets share all categories and no set has unique categories. With fewer categories shared, the measure decreases to indicate instability. Based on the common set similarity measure dice, we can compute the basic result set stability as follows:

$$RSstab_{basic}(ER_i, ER_j) = \frac{2 \cdot |ER_i \cap ER_j|}{|ER_i| + |ER_j|}.$$

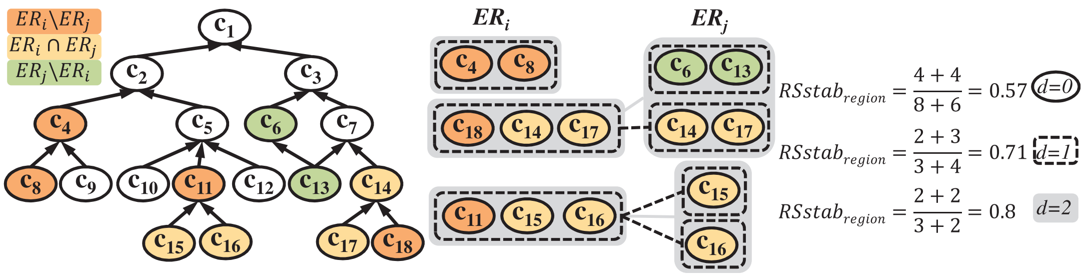


Fig. 2. Stability measures. Colored nodes denote significant result set categories. We consider two versions *i* and *j* of the result set *ER* (for clarity with stable ontology structure). Yellow categories are significant in both result sets ER_i and ER_j, red (green) categories are only significant in ER_i (ER_j). Stability is computed for *CRs* grouped using distance $d = 0\ldots2$. For $d = 0$, each concept is considered as a region (yellow concepts overlap). For $d = 1$ ($d = 2$), overlapping *CRs* are connected by a dashed black (continuous grey) line

The stability returns a value between 0 and 1, whereby 0 denotes complete instability between the two result sets (no overlapping categories). A stability of 1 means that two considered result sets completely overlap. For instance, the example shown in Figure 2 results in a stability of 0.57 since ER_i (ER_j) contains eight (six) categories of which four overlap.

We use an analogous measure to determine the stability between annotation versions (*Astab*). For this, we consider the whole annotation set *A* or annotations to a specific category $A(c)$. We calculate the set of annotations that are identical between the two versions. This value is then normalized by the total number of annotations. The calculation of the stability measure for *A* (and analogous for $A(c)$) is thus identical to the stability between result sets:

$$Astab(A_i, A_j) = \frac{2 \cdot |A_i \cap A_j|}{|A_i| + |A_j|}.$$

While there are alternative set similarity measures, such as Cosine or Jaccard, we consider dice as a suitable approach. It corresponds to the harmonic mean, and we compare sets of similar size. According to Manning *et al.* (1999), the Jaccard and Cosine measures produce slightly different values than dice, especially in case of low set overlap. Lin (1998) proposes a more general measure based on information theory, taking probabilities instead of set overlap into account. This approach is particularly suited for the evaluation of term-enrichment tools, and we consider this a topic for future work.

2.4.2 Region stability measure The basic stability measure evaluates the overlap of categories in the result sets without considering the semantic relatedness of the categories, i.e. it treats them as independent from each other. This may result in an apparent instability even when differing categories are semantically related. We therefore generalize our model to include structural similarity. We first enhance the result sets by grouping together semantically related categories within so-called category regions (*CRs*). We then define the stability with respect to the *CRs* of the result sets. The basic model without grouping remains valid as a special case of the region-based approach.

Semantic grouping of ER categories: We base the semantic grouping of categories in the experimental result sets on the distance within the ontology. This takes advantage of the fact that related categories tend to be in close proximity, i.e. they are either directly connected by an ontology relationship or only a few relationship 'edges' apart. We control the grouping by a distance parameter *d*; the base case without grouping corresponds to $d = 0$. For $d > 0$, we recursively group together all categories that are connected by $\leq d$ edges (see Supplementary Algorithm 1). For the example shown in Figure 2, we obtain for $d = 1$ three regions in ER_i and four regions in ER_j, e.g. in ER_j, we group c_6 and c_{13} into one region. For $d = 2$, the number of regions is reduced to two in ER_i and three in ER_j, e.g. c_{14} and c_{17} are grouped together with c_6 and c_{13} in ER_j.

Alternative methods exist to group categories in the result sets by applying different semantic similarity measures (e.g. Pesquita et al. 2009; Wang et al. 2007) or classification algorithms (e.g. Pandey et al. 2009). Our approach makes use of semantic similarity based on ontology structure and is simple to apply. The approach is also in agreement with term-enrichment approaches such as FUNC that determine the significance of categories based on the ontological structure that try to restrict significant categories to few areas within an ontology.

Determining region stability: The semantic grouping of categories allows us to determine the stability of results sets based on their *CRs* and to assume stability as long as the same or at least overlapping regions are retained in the result sets. We therefore determine the number of regions in ER_i having an overlap with regions in ER_j and vice versa (see Supplementary Algorithm 3). We consider two regions as overlapping if they share at least one category. We denote the overlapping regions in ER_i with CR_i^o and in ER_j with CR_j^o, respectively.

We can now use the information about overlapping regions to compute the region stability as follows:

$$RSstab_{region}(CR_i, CR_j) = \frac{|CR_i^o| + |CR_j^o|}{|CR_i| + |CR_j|}.$$

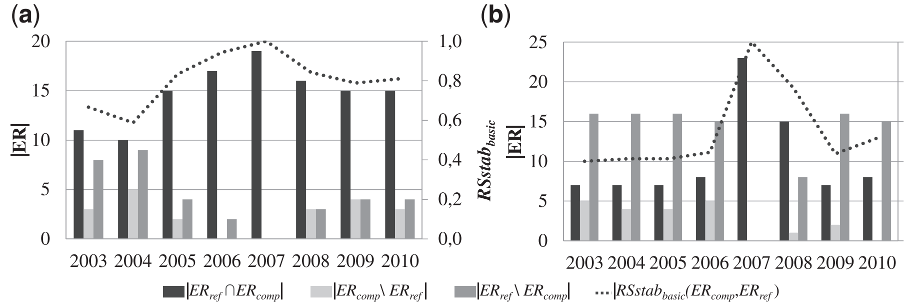
The stability values are, as before, distributed between 0 and 1; for $d = 0$ (no grouping), the region stability equals the basic stability. In Figure 2, the region stability is 0.57 for $d = 0$, 0.71 for $d = 1$ and even 0.8 for $d = 2$. Increasing the distance leads to fewer but larger regions that more likely overlap with the regions of updated result sets. Non-overlapping regions indicate the addition or removal of larger areas in the ontology and, thus, more significant changes than individual category changes quantified with the basic stability measure.

2.5 Datasets

We re-analyzed two datasets that were first tested in 2007 (Kosiol *et al.*, 2008). The authors performed functional enrichment analyses of genes that show signals of positive selection in primates and rodents. We repeat the analyses using newer versions of ontology and annotation.

We consider yearly versions between 2003 and 2010, i.e. eight GOA versions (8, 17, 27, 38, 47, 59, 70, 81) and the corresponding GO versions (01-2003, 02-2004, 01-2005, 01-2006, 01-2007, 01-2008, 02-2009, 01-2010). All newer and older versions are compared with the version used in the original publication (GOA_{47} and $GO_{01-2007}$). For testing, we used the Wilcoxon rank test with 10 000 random sets and a cutoff of at least 20 genes per category. A significant category has to have a *P*-value that does not exceed a value of 0.05.

To test whether the observations from the real datasets apply generally, we generated 50 datasets, randomly seeding significant categories. For this, a set of categories is chosen using a fixed first ontology version and marked as significant by choosing a higher ratio of genes that are



(c)

Change	*RSstab* $_{basic}$ *(ER* $_{2007}$*,ER* $_{2010}$*)*	
	primates	***rodents***
O_v , A_w	0.81	0.52
O_v	0.84	0.7
A_w	0.97	0.5

Fig. 3. Evolution of *ERs* between 2003 and 2010 for primates **(a)** and rodents **(b)**. We compared the *ERs* for each version ER_{comp} against the reference version ER_{ref} from 2007: overlapping categories (black bar), categories in ER_{comp} but not in ER_{ref} (light gray bar), categories in ER_{ref} but not in ER_{comp} (dark grey bar). **(c)** Basic result set stability $RSstab_{basic}$ between 2007 and 2010. Different input versions: evolution of ontology and annotation versions (change O_v and A_w), only changing the ontology version (change O_v) and only changing the annotation version (change A_w)

marked as significant. Then, this simulated dataset can be tested using a different GO version. We generate datasets with the GO versions from 2007 and 2010. We then test the 2007 dataset with the 2010 version (Task A) and vice versa (Task B).

3 RESULTS AND DISCUSSION

Between 2003 and 2010, GO grew by a factor of 2.4. Similarly, the number of input annotations increased by factor 2.7 for mouse and human (see Supplementary Table S2). However, some annotations were also removed. A comparison of whole annotation sets of 2007 and 2010 shows substantial instabilities [$Astab(A_{2007}, A_{2010}) = 0.7$], i.e. every third annotation was affected by a change. To understand the impact of ontology and annotation evolution on term-enrichment results, we measure the stability for two real-world datasets (primate and rodent dataset) at different time points. We quantify changes in result sets using our basic stability measure and analyze causes for changes using *Astab* and *diff*. To identify crucial changes in the results, we applied our region-stability measure. We analyzed 50 random datasets to test whether our observations are generalizable.

3.1 Primate and rodent datasets

3.1.1 Basic stability We computed result sets for yearly versions between 2003 and 2010 for the primate (Fig. 3a) and rodent datasets (Fig. 3b). We compared the result sets of each version (ER_{comp}) against the reference version ER_{ref} (from 2007) by computing the overlap and difference of significant categories. The primate result set (Fig. 3a) contains 19 significant categories in 2007. In general, ontology versions that are closer in time tend to share a higher fraction of significant categories. Since 2008, the results are more stable. A substantial fraction of significant categories are detected in 2007 and 2010, as evidenced by a stability measure ($RSstab_{basic}$) of 0.81.

The rodent result set (Fig. 3b) contains 23 significant categories in 2007. The preceding (2006) and successive (2008) version overlap by only 8 and 15 categories, respectively. Using the 2010 version, only eight categories of the reference result set remained, and no new categories were detected as significant. Overall, the results set stability (0.52) is lower compared with the primate dataset.

Figure 3c shows a comparison of the result set stability $RSstab_{basic}$ between 2007 and 2010 for both the datasets. To identify the main cause for the changes observed between significant categories found in different years, we tested the results changing ontology and annotation independently. Changing only the ontology affected both datasets (rodents 0.84, primates 0.7). Changing the annotation version (A_w) but fixing the ontology only marginally affected the primate result set (0.97) while it substantially reduced the stability of the rodent result set (0.5). This shows that both ontology and annotation evolution have an impact on the results of term-enrichment analyses.

We explore the causes for the differences in significant categories between 2007 and 2010 (Fig. 4). We used *diff* to identify ontology changes that have caused changes in *ER*. Moreover, we analyzed the annotation stability (*Astab*) of all significant result categories *c* to see whether changes in annotations predominate in some cases.

First, there were three new significant categories for primates in 2010. Two categories ('molecular transducer activity' GO:0060089, 'system process' GO:0003008) were added between 2007 and 2010 such that they were additionally detected in the functional enrichment analysis. Another category ('cognition' GO:0050890) has been included in the result set since it received additional annotation. On the other hand, no new categories were detected as significant in rodents.

Several categories were no longer significant in both rodent and primate datasets. Three of the 15 non-significant categories in the rodent dataset are directly affected by an ontology evolution operation (*merge*) while most other categories that became non-significant show a strongly reduced annotation stability of less than 0.7. For instance, for 'regulation of immune system process' ('GO:0002682') only 22 out of 143 annotations in 2010 overlap to the annotation set of 2007 ($Astab \leq 0.3$). This is in contrast to the primate dataset, where three of the categories were affected by ontology evolution operations (*merge*, *substitute*, *toObsolete*) while only one category shows a large change in annotation.

We further observed strong structural changes in the direct semantic context of significant categories: $|addR| = 31(48)$, $|delR| = 10(11)$, $|move| = 102(57)$ for primates (rodents) (see Supplementary Fig. S5). Such structural

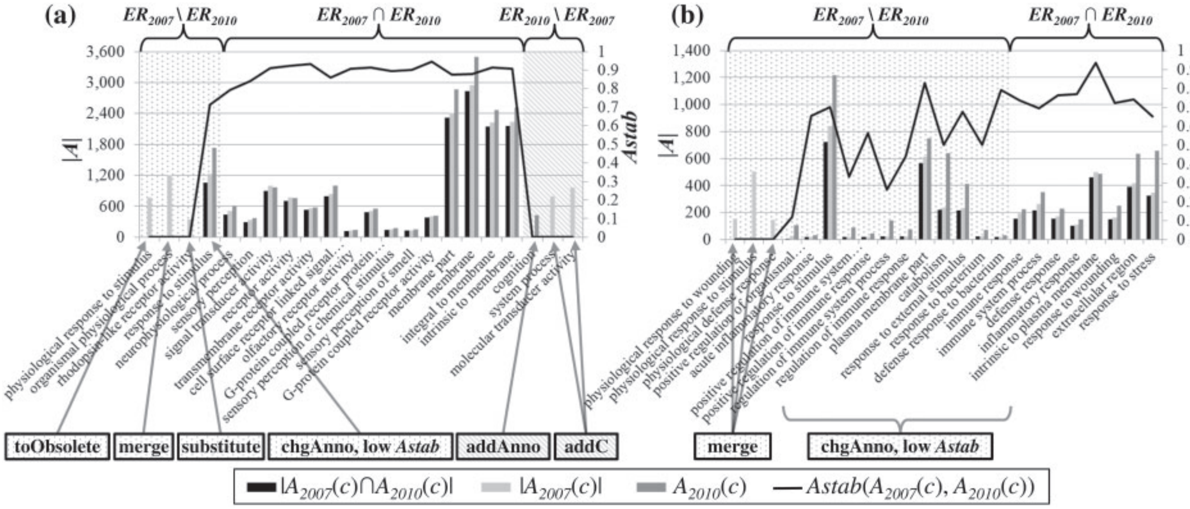


Fig. 4. Annotation stability for result set categories in ER_{2007} and ER_{2010} for (a) primates and (b) rodents. The diagrams show all significant categories computed in 2007 and/or 2010 (x-axis). For each category, the overlap of annotations between both versions ($|A_{2007}(c) \cap A_{2010}(c)|$, black bars, y-axis), the number of annotations in 2007 ($|A_{2007}(c)|$, light gray bars, y-axis) and 2010 ($|A_{2010}(c)|$, dark gray bars, y-axis) as well as the resulting annotation stability $Atab(A_{2007}(c), A_{2010}(c))$ (black curve, z-axis) are shown. Categories in the dotted (striped) areas were only present in the analysis result of 2007 (2010). All other categories are present in both version sets. The lower boxes highlight evolution operations that happened to categories (dotted box - information reducing/revising, striped box - added information). Result categories (including GO accessions) for both species and years are shown in Supplementary Figure 4

modifications may influence annotation propagation. Adding or deleting relationships to an upper category c leads to a changed annotation set $A(c)$ and, thus, to a reduced annotation stability *Astab* and possibly to a changed significance of c. Overall, most unaffected significant categories showed a higher annotation stability than categories that gained or lost significance.

Electronic annotation that is based on automated methods to infer the function of genes may produce less reliable annotations than manually curated entries. To test this hypothesis, we repeated our analysis of the rodent and primate datasets using only manually curated annotation. However, because as much as 60% of the annotation is derived from automated approaches, the dataset is too small to draw reliable conclusions (primates $|ER_{2007}| = 4$, rodents $|ER_{2007}| = 9$; see Supplementary Figs S6 and S7 for details).

3.1.2 Region stability Applying the *CR* stability measure (distance $d = 1$) summarizes significant categories into semantically related regions (Fig. 5). For primates, we identify four *CRs*. All four regions overlap, and the contents of the regions changed only slightly from 2007 to 2010. Considering regions instead of single categories, we obtain perfect stability ($RSstab_{region} = 1$ instead of $RSstab_{basic} = 0.81$). For rodents, there were four significant *CRs* in 2007 (CR_1, CR_2, CR_3, CR_4) and 2010 ($CR_2, CR_3, CR_{4a}, CR_{4b}$) but only three of them overlapped. One region (CR_1 'catabolism') lost significance, due to strong annotation evolution (see Fig. 4). Moreover, there was one very large region (CR_4 'response to stimulus'), where 13 categories were no longer significant in 2010. Six remaining categories were split into two regions (CR_{4a} 'response to stress', CR_4b 'immune system process'). The rodent dataset has a *CR* stability of $RSstab_{region} = (3 + 4)/(4 + 4) = 0.875$ (instead of $RSstab_{basic} = 0.52$).

Comparing *CRs* incorporates semantic information, which leads to higher stability values since larger regions are considered. If the *CR* stability is reduced, we argue that there is a meaningful difference in results.

With the datasets analyzed here, we see that ontology evolution operations and annotation changes can have effects on term-enrichment analyses. There is a substantial variability between the primate and rodent datasets. While the primate results proved to be relatively stable, we see meaningful changes for the rodent data. The primate dataset was more influenced by ontology changes while annotation changes had a higher impact on the rodent dataset. Term-enrichment analysis will often yield semantically related categories reducing the influence of changes in semantically related categories. The majority of changes have no effect on the semantic interpretation of functional enrichment analysis, although we found some instances in which the interpretation may change. We conclude that term-enrichment results are relatively robust to ontology and annotation evolution.

3.2 Simulated datasets

We observe that several categories in the real datasets are affected by evolution of ontology and annotation, and that these changes lead to differences in the enrichment analysis results. To test whether the patterns we observe are generalizable, we generated 50 datasets, randomly seeding significant categories. Repeating this generation of datasets for two versions of the GO (2007, 2010) enables us to test to what extent differences between these two ontology versions influence enrichment results. To distinguish semantically related categories that are likely equally affected by changes, from semantically distinct categories, we apply the *CR* stability measure with distance $d = 1$.

Table 1 shows average values over all 50 random experiments (details in Supplementary Tables S4 and S5). Note that the

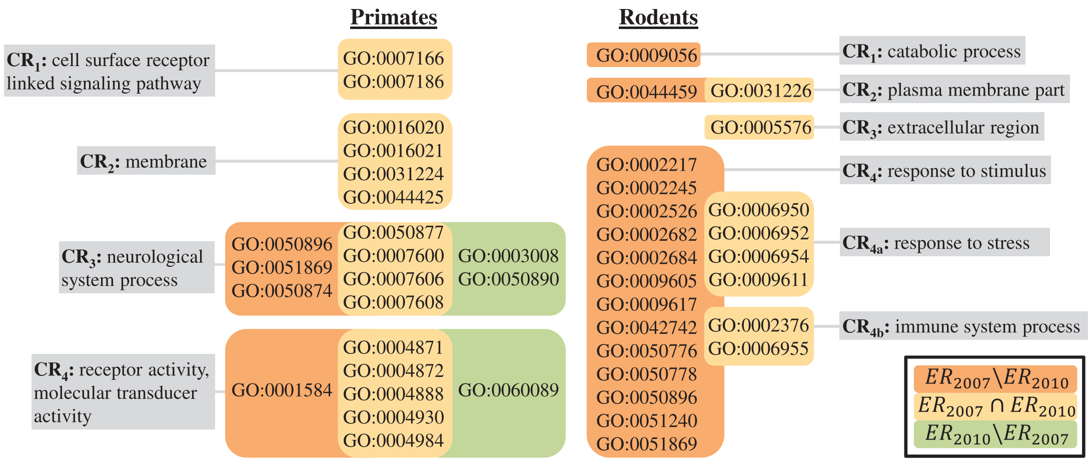


Fig. 5. *Category regions* for primate and rodent dataset, $d = 1$

Table 1. Average results over 50 random experiments for tasks A and B

	Task A	Task B
$avg(\lvert ER_{only2007} \rvert)$	129.3	62.6
$avg(\lvert CR_{only2007} \rvert)$	8.9	4.2
$avg(\lvert ER_{only2010} \rvert)$	440.7	856.8
$avg(\lvert CR_{only2010} \rvert)$	18.9	21.4
$avg(RSstab_{basic})$	0.625	0.519
$avg(RSstab_{region})$	0.719	0.707

Average number of significant result categories ($|ER|$) and more compact *CRs* ($|CR|$) only in 2007 (missing), only in 2010 (new), $avg(RSstab_{basic})$ – average of basic stability, $avg(RSstab_{\text{region}})$ – average of *CR* stability.

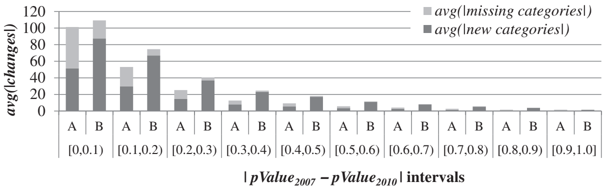


Fig. 6. Average number of changed categories grouped by absolute *P*-value difference ($|Pvalue_{2007} - Pvalue_{2010}|$) for tasks A and B

random datasets are larger and thus cover larger ontology parts, possibly leading to a generally lower stability. When testing changes from 2007 to 2010 (Task A), we observe on an average nine *CRs* that lose significance while 19 newly significant groups are identified. Changes from 2010 to 2007 (Task B) yield to four *CRs* with lost significance and 21 newly significant regions. Note that comparing single result categories instead of regions leads to less compact results (e.g. 129.3 categories versus 8.9 *CRs* for Task A in 2007). Moreover, using $avg(RSstab_{region})$ results in higher values (≈0.7) than $avg(RSstab_{basic})$ (≈0.5 to ≈0.6) since changes of semantically related categories are not considered as really new/missing significant regions. These observations affirm the results of the two real-world datasets.

Because we define categories as significant based on a fixed *P*-value cutoff, some of the categories that lose or gain significance may do so due to small fluctuations in *P*-value. To test for the relative contribution of this effect to the observed changes, we computed the differences of *P*-values between 2007 and 2010 for each significant category. Figure 6 shows the distribution of *P*-value differences for lost and gained significant categories. Most of the categories reveal only a relatively small change in *P*-value (< 0.1), showing that most category gains and losses are driven by small fluctuations in significance. However, some categories show substantial *P*-value differences, suggesting that real structural changes in the ontology or annotation are the basis for the change in *P*-value.

4 CONCLUSIONS

Enrichment analyses use ontology and annotation to detect significantly enriched categories of genes. We studied the impact of ontology and annotation evolution on term-enrichment analyses by comparing term-enrichment results over different ontology and annotation versions. We proposed different measures to assess the stability of result sets and applied them to analyze the impact of evolution for two real-world and 50 random datasets.

The GO undergoes continuous changes in its structure and the annotation due to the ongoing incorporation of new knowledge. These changes are unequally distributed and can cluster in regions representing specific topics. At the level of individual categories, the results of term-enrichment analyses can be significantly affected by ontology and annotation evolution. However, these changes do not necessarily change the interpretation of the result since these terms are often semantically related. This effect is captured by our CR stability measure. The experimental evaluation showed that term-enrichment analyses are generally robust to ontology and annotation evolution.

Using our measures, users can identify categories that tend to change their level of significance due to structural ontology changes or heavily changed annotation sets. The two following audiences can benefit from our methods and results:

(i) Ontology curators: For these users, it is important to determine whether planned changes in the ontology or the annotations result in a semantic change. We show here that structural ontology changes do not necessarily imply a semantic change of the results, and we provide stability measures that allow testing of proposed changes in the GO and the implications for functional enrichment analyses.

(ii) Biologists using the GO for functional enrichment analyses: For these users, it is interesting to know that enrichment results may change over time due to changes in the ontology and annotations and that interpretation of their own results should be made with this in mind.

We already provide a tool for ontology evolution evaluation (Hartung *et al.* 2012a). We plan to extend this tool to detect annotation evolution and to incorporate the here-presented stability measures. The tool may be of particular interest for curators of ontologies to judge the potential impact of changes for users.

Funding: This work was supported by the German Research Foundation, grant RA 497/18-1 (‘Evolution of Ontologies and Mappings’) and the Max Planck Society.

Conflict of Interest: none declared.

REFERENCES

Carbon,S., Ireland,A., Mungall,C., Shu,S., Marshall,B., Lewis,S. *et al.* (2009) *Amigo: online access to ontology and annotation data*. *Bioinformatics*, **25**, 288–289.

Barrell,D. *et al.* (2009) The GOA database in 2009–an integrated gene ontology annotation resource. *Nucleic Acids Res.*, **37** (Suppl. 1), D396–D403.

Bodenreider,O. and Stevens,R. (2006) Bio-ontologies: current trends and future directions. *Brief. Bioinformatics*, **7**, 256–274.

Dolan,M. *et al.* (2005) A procedure for assessing go annotation consistency. *Bioinformatics*, **21** (Suppl. 1), i136–i143.

Gene Ontology Consortium(2008) The gene ontology project in 2008. *Nucleic Acids Res.*, **36** (Database Issue), D440–D444.

Gross,A. *et al.* (2009) Estimating the quality of ontology-based annotations by considering evolutionary changes. In *Proceedings of the 6th Int. Workshop on Data Integration in the Life Sciences*. Springer, Heidelberg, pp. 71–87.

Hartung,M. *et al.* (2008) Analyzing the evolution of life science ontologies and mappings. In *Proceedings of the 5th Int. Workshop on Data Integration in the Life Sciences*. Springer, Heidelberg, pp. 11–27.

Hartung,M. *et al.* (2010) Discovering evolving regions in life science ontologies. In *Proceedings of the 7th Int. Conference on Data Integration in the Life Sciences*. Springer, Heidelberg, pp. 19–34.

Hartung,M. *et al.* (2012a) CODEX: exploration of semantic changes between ontology versions. *Bioinformatics*, **28**, 895–896.

Hartung,M. *et al.* (2012b) Conto–diff: generation of complex evolution mappings for life science ontologies. *J. Biomed. Informat.* (in press).

Kosiol,C. *et al.* (2008) Patterns of positive selection in six mammalian genomes. *PLoS Genet.*, **4**, e1000144.

Lambrix,P. *et al.* (2007) Biological ontologies. *Semantic Web*, 85–99.

Leonelli,S. *et al.* (2011) How the gene ontology evolves. *BMC Bioinformatics*, **12**, 1–7.

Lin,D (1998) An information-theoretic definition of similarity. In: *Proceedings of the 15th international conference on Machine Learning, ICML '98*. Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, pp. 296–304.

Manning,C. *et al.* (1999) *Foundations of Statistical Natural Language Processing*. Vol. 999, MIT Press, Cambridge, Mass.

Noy,N. and Musen,M (2002) Promptdiff: a fixed-point algorithm for comparing ontology versions. In: *Proceedings of the National Conference on Artificial Intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press, pp. 744–750.

Pandey,G. *et al.* (2009) Incorporating functional inter-relationships into protein function prediction algorithms. *BMC Bioinformatics*, **10**, 1–22.

Park,J. *et al.* (2008) Monitoring the evolutionary aspect of the gene ontology to enhance predictability and usability. *BMC Bioinformatics*, **9** (Suppl. 3), S7.

Pesquita,C. and Couto,F.M. (2011) Where GO is going and what it means for ontology extension. In *Proceedings of the 2nd International Conference on Biomedical Ontology*, vol. 833. Buffalo, NY, USA, pp. 3–9. CEURWorkshop Proceedings, ISSN 1613-0073.

Pesquita,C. *et al.* (2009) Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.*, **5**, e1000443.

Prüfer,K. *et al.* (2007) FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics*, **8**, 1–10.

Thomas,P. *et al.* (2007) Ontology annotation: mapping genomic regions to biological function. *Curr. Opin. Chem. Biol.*, **11**, 4–11.

Tilford,C. and Siemers,N. (2009) Gene set enrichment analysis. *Methods Mol. Biol.*, **563**, 99–121.

Wang,J. *et al.* (2007) A new method to measure the semantic similarity of go terms. *Bioinformatics*, **23**, 1274–1281.