Quantitative cross-species comparison of GO annotations: advantages and limitations of semantic similarity measure

Olivier Dameron, Charles Bettembourg, Léa Joret

INSERM U936, Université de Rennes1, France
olivier.dameron@univ-rennes1.fr

1 Context

We propose a method supporting quantitative cross-species comparison of a gene product's Gene Ontology Annotations.

2 Material and methods

We highlight the various kinds of difficulties arising when comparing two sets of GO annotations using either the intersection of sets of annotations or semantic similarity measures.

We illustrate our approach by comparing the GO annotations of Apolipoprotein A-5 (apoa5) and Apolipoprotein A-1 (apoa1) respectively between human (hsa) and mice (mmu). Apoa5 is involved in similar biological processes in the two species [1], whereas Apoa1 is known to be significantly different [2].

For each gene product, we retrieved the annotations for each species using the GOA database from the EBI[1]. We also retrieved the evidence codes and modifiers of these annotations in order to take negation into account [3, 4].

The Gene Ontology provided the hierarchy between the annotations. We used the daily version[2].

3 Results

Due to space limitations, the version of this article including tabvles with numerical values is available online[3].

3.1 Straightforward intersection of annotation sets

Computing the intersection of the human and murine annotation sets yields different results for apoa1 and apoa5.

[1] ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/gene_association.goa_uniprot.gz

[2] http://archive.geneontology.org/latest-termdb/go_daily-termdb-data.gz

[3] http://www.ea3888.univ-rennes1.fr/dameron/protege2009/dameron2009protege-long.pdf

The results show that 19% of the annotations are common between human and mice for apoa1 , whereas 74% of the annotations are common for apoa5 . This is consistent with the biological evidence. Moreover:

- Negation has to be taken into account (e.g. the proportion of common cellular components annotations for apoa5 is of 67% when ignoring negation, versus 80% when considering negation);
- Annotations should be compared separately for biological process, cellular components and molecular functions.

However, this approach has two major limitations:

- For each species, the set of annotations contains *redundant annotations* (i.e. annotations that are ancestors of another annotation in the set). Such annotations are legitimate because they can have different evidence codes [3]. However, they should not be taken into account in the comparison, as they can artificially increase the size of species-specific annotations.
- Some annotation for one species can be an ancestor of some other annotation for the other species. This reflects some *difference of granularity* during the gene product annotation between the two species. Each annotation is then counted as species-specific, whereas the ancestor should be counted as common.

Both limitations arise because the semantics of the annotations (represented here by the taxonomy of Gene Ontology) was ignored. Sections 3.2 and 3.3 present two approaches for overcoming them.

3.2 Intersection of annotations sets restricted to leaves

We removed all the redundant annotations from each set of annotations. For computing the comparison, the potential differences of granularities between the annotations from the two species were taken into account.

The results are notably different from those of the straightforward approach while remaining consistent with the biological evidence .

However, when taking the difference of granularity into account, the number of intermediate annotations is ignored. An annotations linked by a direct *is_a* relation on the one hand, and annotations separated by several ancestors on the other hand lead to the same intersection, whereas intuitively, the species-specificity should be emphasized in the second situation.

3.3 Intersection of annotations sets extended with ancestors

We enriched each set of annotations with all the ancestors of each initial annotations. This solves the redundancy issue, as well as the difference of granularity issue identified in section 3.1. Moreover, the species-specificity identified in section 3.2 is also addressed.

The results are notably different from those of the previous sections while remaining consistent with the biological evidence .

However, by retrieving all the ancestors, this approach artifically promotes the propertion of common annotations. Indeed, the annotations from the higher levels of the hierarchy are rather general (e.g. Protein binding) and are likely to be common to the two species.

We believe this approach to address all the limitations identified in the previous sections but to suffer from an inner limitation promoting the proportion of common annotations.

3.4 Semantic similarity

Semantic similarity measures [5] tackle the problem identified in section 3.3 by attributing a higher weight to the more informative annotations (i.e. the lowest nodes in the hierarchy). Some measures have been specially adapted to Gene Ontology [6]. The measure proposed by Wang [7] is so far the better.

The results show that Wang semantic similarity measure is of 0.48 for apoa1, and of 0.90 for apoa5, which is consistent with the biological evidence .

However, several points remain to be addressed:

- negation is not taken into account, which is a serious limitation [4];
- the measure takes the *part_of* relation into account, but fails to reflect its inheritance by subclasses;
- there is no difference between a situation where 30% of the annotations are species1-specific, 40% are common and the remaining 30% are species2-specific on the one hand, and a situation where 50% of the annotations are species1-specific, 40% are common and the remaining 10% are species2-specific on the other hand. This calls for some semantic-specificity measure in addition to semantic similarity.

4 Discussion

We have presented different strategies for comparing the GO annotations of a gene product for two species.

Among them, computing the intersection of the sets of annotations enriched with the ancestors takes negation into account but artifically promotes the proportion of common annotations. The GO-specific semantic similarity measure proposed by Wang addressed this limitation but fails to consider negation and ignores the inheritance of the *part_of* relation.

Future work will try to reconcile both approaches.

References

[1] Len A Pennacchio and Edward M Rubin. Apolipoprotein a5, a newly identified gene that affects plasma triglyceride levels in humans and mice. *Arteriosclerosis, thrombosis, and vascular biology*, 23(4):529–534, 2002.

[2] Hafid Mezdour, Guilhem Larigauderie, Graciela Castro, Gerard Torpier, Jamila Fruchart, Maxime Nowak, Jean-Charles Fruchart, Mustapha Rouis,

and Nobuyo Maeda. Characterization of a new mouse model for human apolipoprotein A-I/C-III/A-IV deficiency. *Journal of lipid research*, 47(5):912–920, 2006.

[3] David P Hill, Barry Smith, Monica S McAndrews-Hill, and Judith A Blake. Gene ontology annotations: what they mean and where they come from. *BMC bioinformatics*, 9 Suppl 5:S2, 2008.

[4] Seung Yon Rhee, Valerie Wood, Kara Dolinski, and Sorin Draghici. Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*, 9(7):509–515, 2008.

[5] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[6] 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*, 19(10):1275–1283, 2003.

[7] James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip S Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics (Oxford, England)*, 23(10):1274–1281, 2007.