Measuring coherence between electronic and manual annotations in biological databases

Catia Pesquita
cpesquita@xldb.di.fc.ul.pt
University of Lisbon

Daniel Faria
dfaria@xldb.di.fc.ul.pt
University of Lisbon

Francisco M. Couto
fcouto@di.fc.ul.pt
University of Lisbon

ABSTRACT

The use of controlled structured vocabularies for annotation purposes, such as the Gene Ontology (GO) is currently one of the strategies to cope with the increasingly cumbersome task of genome annotation. The Gene Ontology Annotation Database (GOA) uses GO to annotate gene products through curated literature analysis and uncurated electronic methods. Although electronic annotations constitute the large majority of annotations (over 95%), most researchers are reluctant to use them in their studies, since they are regarded as having a lower quality than curated ones. Assessing the quality of electronic annotations may help clarify the advantages and disadvantages of their use.

This paper proposes a preliminary measure of electronic annotation quality based on the coherence between electronic and manual annotations. Coherence is analysed both at the gene product and at the annotation level, based on semantic similarity of Gene Ontology terms. We have found that average annotation coherence values are around 60%, but can be as high as 81% for a less granular analysis. Based on this analysis we propose meaningful coherence thresholds for electronic annotation selection and filtering, and for highlighting gene products for annotation revision.

Categories and Subject Descriptors

H.2.8 [**DATABASE MANAGEMENT**]: Database Applications—*Scientific databases*

Keywords

Annotation agreement, BioOntologies, Semantic Similarity

1. INTRODUCTION

Genome annotation is currently one of the most important challenges in biology, and it has been greatly assisted by the development of controlled structured vocabularies (commonly referred to as BioOntologies) specifically designed for annotation, such as the Gene Ontology (GO).

SAC'09 March 8-12, 2009, Honolulu, Hawaii, U.S.A.

Annotations of gene products using GO terms are mainly generated by the Gene Ontology Annotation (GOA) project. GOA annotations are given an evidence code that registers the type of information that supported it, for instance IEA (inferred by electronic annotation) is the code for annotations generated by uncurated electronic methods. Since most IEA annotations are based on information retrieved from manually curated resources, GOA is confident that its electronic annotation is of a high standard, nonetheless proteins assigned this code are routinely disregarded in most applications of GO annotations, both due to their lower specificity and to the common assumption that they contain a higher portion of false positives. Ignoring electronic annotations is however, not without consequence, given that the large majority of annotations are of this type (over 95%) due to the expertise and time involved in manual curation of annotations.

In this work, we propose a preliminary measure for electronic annotation quality based on the coherence between electronic and manual annotations that is given by semantic similarity. This measure was developed with a double purpose: 1) to analyse the full GOA in order to give an overall evaluation of electronic annotations and 2) to provide the individual researchers with an intuitive measure to aid them in deciding which, if any, electronic annotations they should consider in their particular research.

2. ANNOTATION COHERENCE

We agree with [1] that electronic annotations that exactly match manual annotations or their ancestors can be assumed to be correct. However, in [1] annotations with higher granularity but within the same lineage or in a new lineage are simply considered potentially correct or incorrect. Hence, we propose a measure that weighs the annotations that cannot be assumed to be correct by the maximum semantic similarity between the electronically assigned term and a manually assigned one. This is defined as follows, for an electronic annotation e to a gene product g:

$$Ca_{e_g} = \begin{cases} 1, k_{e_g} \in K_{M_g} \\ max(sim(k_{e_g}, k_{m_g}), e_g \notin K_{M_g} \end{cases} \quad (1)$$

where k_{e_g} and k_{m_g} are the terms used in the electronic and manual annotations of gene product g respectively, and K_{M_g} is the set of terms used in g's manual annotations and their ancestors. So, annotations can be split into four categories: exact match, ancestor match, descendant match and new lineage; and semantic similarity is only calculated for the last two. Semantic similarity is calculated according to Resnik

[10], which has been shown to be the best information content based measure for GO [6, 9]. Information content was calculated as in [8] To arrive at a final coherence measure for each GO type, we take the arithmetic mean of the coherence value for all annotations.
The electronic annotation coherence for a given gene product is calculated as the ratio between the sum of all coherence values for each electronic annotation and the total number of electronic annotations the protein has:

$$Cg = \frac{1}{n_{e_g}} \times \sum_{e \in E_g} Ca_e \quad (2)$$

Considering that for many applications, researchers are not interested in deep annotations and just wish to have a global visualisation of GO annotation, we used the GOslim subset to re-calculate coherence values, assigning $C_a = 1$ to electronic annotations whose terms match a GOSlim of a manual annotation.
The dataset (from the December 2007 release of ProteInOn database [3]) includes all gene products (and their annotations) that have at least one IEA annotation and at least one non-IEA annotation. A second annotation dataset was similarly obtained using the GOSlim general.

3. RESULTS AND DISCUSSION

We calculated the coherence values for all annotations and gene products achieving an average C_a of 0.59 and an average C_g of 0.60. With the GOSlim dataset we observed and increase of the average C_g (0.81), as expected, since belonging to the same GOSlim was classified a match. Considering that values above 0.8 are generally considered as "perfect agreement" in medical research [5], IEA annotations can be considered as having good coherence to manual ones, for higher level GO applications. As for the more granular analysis, 0.6 falls in the upper range of "moderate agreement".
The distribution of coherence values for the "new lineage" and "descendant" annotations, where $0 < C_a < 1$, (data not shown) is similar for the three GO types, with a prevalence of values below 0.1, reflecting the higher proportion of "new lineage" annotations. These annotations are not necessarily wrong, but the great discrepancy between electronic and manual annotations can be used as an indicator that the annotations need to be revised.
We also calculated Cohen's κ [2], a widely used measure of inter-annotator agreement(0.40). The difference between κ and C_g values can be due to the fact that κ only considers exact matches, takes into consideration chance agreement and does not consider the possibility of an item being classified into several categories. Other strategies have been proposed to deal with taxonomies [7, 4], however they are not applicable to GO which is a DAG without uniform distribution of annotations and where edge length is variable.
The coherence measure for proteins is intended to help researchers select which if any electronic annotations they wish to use. We propose the usage of a coherence threshold that gives an equal weight to coherence and number of annotated proteins. This threshold is calculated solely based on coherence values that are not exactly one or zero and the number of proteins they correspond to, in order to maximise the captured information. However, different applications can benefit from different weightings for instance, to allow for a larger dataset or guarantee a high level of coherence.

4. CONCLUSIONS

Considering the diversity of applications of GO annotations it is particularly important for researchers to understand electronic annotations, since they make up the large majority of them.
This work proposes a preliminary measure to evaluate electronic annotations based on their coherence to manually curated ones. The measure is based on semantic similarity, and has two variants: one for single annotations, and an adaptation for gene products. Both measures are suitable to application with the Gene Ontology, in contrast to previously proposed inter-annotator agreement measures. Application of the proposed measures to GO annotated gene products revealed an average coherence value of 0.61, that is raised to 0.81 in a less granular analysis.
In future work we propose to improve this preliminary measure by considering relevant issues such as chance agreement.

5. ACKNOWLEDGEMENTS

This work was supported by the Portuguese Fundacao para a Ciencia e Tecnologia through the Multiannual Funding Programme, and the grants refs. SFRH/BD/42481/2007, SFRH/BD/29797/2006 and SFRH/BD/29150/2006.

6. ADDITIONAL AUTHORS

Rui Lopes, University of Lisbon.

7. REFERENCES

[1] E. B. Camon, D. G. Barrell, E. C. Dimmer, V. Lee, M. Magrane, J. Maslen, D. Binns, and R. Apweiler. An evaluation of go annotation retrieval for biocreative and goa. *BMC Bioinformatics*, 6 Suppl 1, 2005.

[2] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[3] D. Faria, C. Pesquita, F. Couto, and A. Falcão. Proteinon: A web tool for protein semantic similarity. Technical report, Department of Informatics, University of Lisbon, 2007.

[4] J. Geertzen and H. Bunt. Measuring annotator agreement in a complex hierarchical dialogue act annotation scheme. In *Proc. of the 7th SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, 2006.

[5] K. G. Landis JR. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.

[6] P. Lord, R. Stevens, A. Brass, and C. Goble. Semantic similarity measures as tools for exploring the gene ontology. In *Proc. of the 8th Pacific Symposium on Biocomputing*, 2003.

[7] I. Melamed and P. Resnik. Tagger evaluation given hierarchical tag sets. *Computers and the Humanities*, 2000.

[8] C. Pesquita, D. Faria, H. Bastos, A. O. Falcao, and F. Couto. Evaluating go-based semantic similarity measures. In *ISMB/ECCB 2007 SIG Meeting Program Materials*. ISCB, 2007.

[9] C. Pesquita, D. Faria, H. Bastos, A. Ferreira, A. Falcao, and F. Couto. Metrics for go based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9(Suppl 5):S4, April 2008.

[10] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of the 14th International Joint Conference on Artificial Intelligence*, 1995.