Auditing Subtype Inconsistencies among Gene Ontology Concepts

Rashmie Abeysinghe*, Eugene W. Hinderer III†, Hunter N.B. Moseley†‡§¶, Licong Cui*‡
*Department of Computer Science
†Department of Molecular and Cellular Biochemistry
‡Institute for Biomedical Informatics
§Markey Cancer Center
¶Center for Environmental and Systems Biochemistry
University of Kentucky, Lexington, Kentucky, USA

***Abstract*—Gene Ontology (GO) provides a controlled vocabulary for describing genes and related gene products. Quality assurance of Gene ontology (GO) is a vital aspect of the terminology management lifecycle. In this paper, we introduce a lexical-based inference approach to detecting subtype (or is-a) inconsistencies among GO terms (i.e., biological concepts). We first model the name of each concept as a set of words. Then, we generate hierarchically linked and unlinked pairs of concepts (A, B), where A and B have the same number of words, and contain common words as well as a single different word. Each linked concept-pair infers a linked term-pair, and each unlinked concept-pair infers an unlinked term-pair. A term-pair appearing as both linked and unlinked is considered a potential inconsistency, which may represent a subtype inconsistency between the original linked and unlinked concept-pair. Applying this approach to the 03/28/2017 release of GO, a total of 3,715 potential subtype inconsistencies were obtained. Evaluation of a random sample of potential inconsistencies revealed two types of potential errors: missing subtype relations and incorrect subtype relations in GO, and achieved an accuracy of 56.33% for detecting such errors. This indicates that this lexical-based inference approach using the set-of-words model is a promising way to facilitate quality improvement of GO.**

I. Introduction

Biomedical ontologies, such as the Gene Ontology (GO) [1] and SNOMED CT [2], provide essential domain knowledge to drive data annotation, data integration, information extraction, and decision support in biomedicine [3], [4], [5], [6], [7]. In particular, GO is recognized as a tool for the unification of biology [8], and has been widely adopted for codifying, managing, and sharing biological knowledge.

GO provides a controlled vocabulary of terms for describing gene and gene product characteristics and related annotation data from GO Consortium members [1]. Since GO is constantly evolving to keep pace with the rapid discovery of biological knowledge, inconsistencies or errors may be introduced during the terminology management lifecycle. Quality assurance or auditing of GO is an important task, since quality issues may affect all GO-driven downstream applications. However, auditing GO becomes more challenging due to the ever-increasing size and structural complexity of GO. It is time-consuming and labor-intensive to manually uncover potential quality issues in the ontology. Thus, there is an urgent need to develop automated and effective approaches to detecting potential quality issues in GO.

In this paper, we introduce a lexical-based inference approach to auditing GO by automatically deriving inconsistencies in hierarchically linked and unlinked concept-pairs. Such inconsistencies may indicate missing subtype (i.e., is-a) relations or incorrect subtype relations in GO. Domain experts reviewed a random sample of inconsistencies to evaluate the effectiveness of this approach.

This work was supported by the National Science Foundation through grants 1657306 and 1252893, and by the National Institutes of Health through grant UL1TR001998-01. Correspondence: licong.cui@uky.edu

II. Background

A. Gene Ontology (GO)

GO was constructed as a collaborative effort to address the need for consistent description of genes and their related gene products across databases [1], [9]. It contains three sub-hierarchies which describe gene products in terms of their associated *Biological Processes*, *Cellular Components*, and *Molecular Functions* [9]. It contains over 40,000 biological concepts, which are constantly revised to reflect latest discoveries and current biological knowledge.

B. Quality Assurance of GO

Various approaches have been introduced [10] for quality assurance of GO. Ochs et al. [11] have applied abstraction network (AbN) based methods to audit GO. They have developed two kinds of AbNs: area taxonomy and partial-area taxonomy, for GO hierarchies and derived specifically for the biological process sub-hierarchy of GO. Xing et al. [12] have developed an effective dynamic-programming-based algorithm to detect redundant hierarchical relations, and applied it to two biomedical ontologies including GO. Bodenreider et al.[13] have introduced three non-lexical approaches: computing similarity in a vector space model, statistical analysis of co-occurrence of GO terms in annotation databases, and association rule mining, to identify associative relations across hierarchies in GO. Mougin [14] has studied a method to identify redundant and missing relations in GO. Reasoning over relationships has been exploited to identify redundant relations between

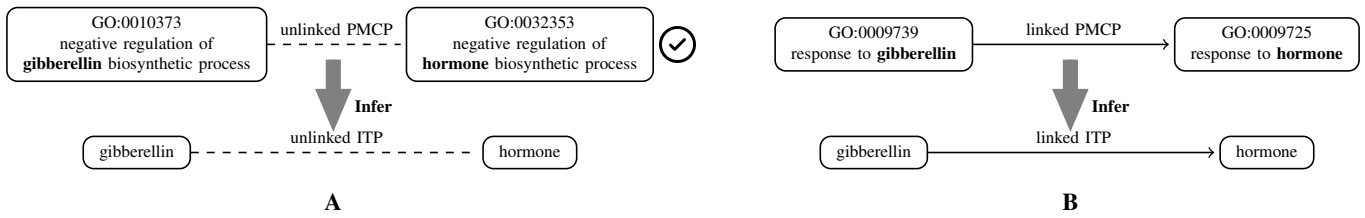

Fig. 1. **A**: Unlinked PMCP and its unlinked ITP; **B**: Linked PMCP and its linked ITP. This example reveals a potentially **missing subtype relation** in **A**, that is, GO:0010373 (*negative regulation of gibberellin biosynthetic process*) is-a GO:0032353 (*negative regulation of gibberellin hormone process*).

concepts. Compositional structure of the preferred names of GO concepts has been used to detect missing necessary and sufficient conditions. Ceusters [15] has shown how realism-based principles for ontology evolution can be used for terminology auditing in GO.

III. Methods

In this work, we introduce a lexical-based inference approach to identify potentially missing subtype relations and incorrect subtype relations. Firstly, we represent the name of each GO concept as a set of words. Then, we generate partial matching pairs of concepts that are hierarchically linked or unlinked. We further derive linked and unlinked term-pairs from the concept-pairs. Then, we detect potential subtype inconsistencies between concept-pairs that share the same term-pair. Finally, domain experts evaluate a random sample of detected potential inconsistencies and suggest the potential errors indicated by those inconsistencies.

A. Modeling GO Concept Names

For each concept in GO, we represent the name of the concept as a set of words. For example, the name of the concept GO:0009785 (the unique identifier) is *blue light signaling pathway*; and its unordered set-of-words representation is $\{$*blue, light, signaling, pathway*$\}$.

B. Generating Partial Matching Concept Pairs (PMCPs)

We define a pair of concepts as a *partial matching concept pair (PMCP)*, if the names of the two concepts have the same number of words, and contain a single different word and at least one word in common. For instance, GO:0009739 (*response to gibberellin*) and GO:0009725 (*response to hormone*) is a PMCP.

We further define two categories of PMCPs as follows:

- *Linked PMCP*: If the two concepts in a PMCP are connected through a subtype relation (either direct or indirect), then we say that this pair of concepts is a linked PMCP.
- *Unlinked PMCP*: If the two concepts in a PMCP are not connected through a subtype relation, then we say that this pair of concepts is an unlinked PMCP.

For example, Fig. 1A shows an example of an unlinked PMCP, where the two concepts GO:0010373 (*negative regulation of gibberellin biosynthetic process*) and GO:0032353 (*negative regulation of hormone biosynthetic process*) differ in a single word – *gibberellin* versus *hormone*. Fig. 1B shows an example of a linked PMCP, where the two concepts GO:0009739 (*response to gibberellin*) and GO:0009725 (*response to hormone*) also differ in a single word – *gibberellin* versus *hormone*.

Note that we leverage the pre-computed transitive closure of the subtype relation (i.e., direct and indirect is-a relations) to determine if a PMCP is linked or unlinked. That is, if a PMCP is in the transitive closure, then it is linked; otherwise, it is unlinked. For instance, the PMCP (GO:0009739, GO:0009725) in Fig. 1B is in the transitive closure; thus it is linked. However, the PMCP (GO:0010373, GO:0032353) in Fig. 1A is not in the transitive closure; thus it is unlinked.

C. Deriving Inferred Term Pairs (ITPs)

For each PMCP (C_1, C_2), we use the different words between the names of C_1 and C_2 to derive an *Inferred Term Pair (ITP)*. We further define two categories of ITPs based on the corresponding PMCPs:

- *Linked ITP*: If an ITP is derived from a linked PMCP, then we say it is a linked ITP.
- *Unlinked ITP*: If an ITP is derived from an unlinked PMCP, then we say it is an unlinked ITP.

Take Fig. 1A as an example, the unlinked concepts GO:0010373 (*negative regulation of gibberellin biosynthetic process*) and GO:0032353 (*negative regulation of hormone biosynthetic process*) derives an unlinked ITP (*gibberellin*, *hormone*).

D. Detecting Potential Inconsistencies

If an unlinked PMCP and a linked PMCP derive the same ITP, we consider these two PMCPs a potential subtype inconsistency. For instance, the unlinked PMCP (GO:0010373, GO:0032353) in Fig. 1A and the linked PMCP (GO:0009739, GO:0009725) in Fig. 1B is considered a potential inconsistency, since they derive the same ITP (*gibberellin*, *hormone*).

E. Evaluating Detected Inconsistencies

For the evaluation of the potential subtype inconsistencies detected above, we classify them into three categories: missing subtype relations, incorrect existing subtype relations, and false positives. We describe each category in detail as follows. Given an inconsistency I consisting of an unlinked PMCP (U_1, U_2) and a linked PMCP (L_1, L_2).

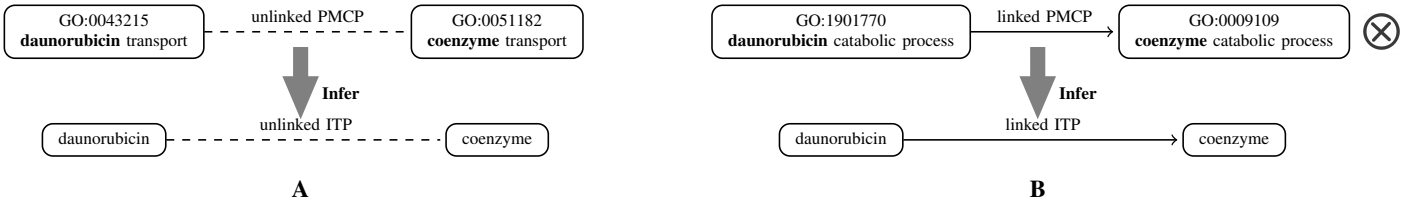

Fig. 2. **A**: Unlinked PMCP and its unlinked ITP; **B**: Linked PMCP and its linked ITP. This example reveals a potentially **incorrect existing subtype relation** in **B**, that is, GO:1901770 (*daunorubicin catabolic process*) is not a subtype of GO:0009109 (*coenzyme catabolic process*).

TABLE I
EXAMPLES OF THE SUBTYPE INCONSISTENCIES FOUND.

ITP	Unlinked PMCP	Linked PMCP	Inconsistency type
(telencephalon, forebrain)	GO:0021537: telencephalon development GO:0030900: forebrain development	GO:0022029: telencephalon cell migration GO:0021885: forebrain cell migration	Incorrect relation
(oxidase, dehydrogenase)	GO:0003884: D-amino-acid oxidase activity GO:0008718: D-amino-acid dehydrogenase activity	GO:0004158: dihydroorotate oxidase activity GO:0004152: dihydroorotate dehydrogenase activity	Incorrect relation
(methotrexate, drug)	GO:0031427: response to methotrexate GO:0042493: response to drug	GO:0051870: methotrexate binding GO:0008144: drug binding	Missing relation
(ethanolamine, peptide)	GO:0046336: ethanolamine catabolic process GO:0044248: cellular catabolic process	GO:0006580: ethanolamine metabolic process GO:0044237: cellular metabolic process	Missing relation
(cortisol, hormone)	GO:0034651: cortisol biosynthetic process GO:0042446: hormone biosynthetic process	GO:0043400: cortisol secretion GO:0046879: hormone secretion	Missing relation

1) Missing subtype relations: If the unlinked PMCP (U_1, U_2) indeed forms a valid subtype relation, then it is regarded as a missing subtype relation in GO (i.e., U_1 should be a subtype of U_2).

2) Incorrect existing subtype relations: If the linked PMCP (L_1, L_2) is found to be an invalid subtype relation, then it is regarded as an incorrect existing subtype relation (i.e., L_1 should not be a subtype of L_2).

3) False positives: If the linked PMCP (L_1, L_2) is indeed a valid subtype relation and the unlinked PMCP (U_1, U_2) is found to be an invalid subtype relation, then I is regarded as a false positive that is identified by our approach.

A random sample of potential inconsistencies was selected and evaluated by two domain experts (authors EWH and HNBM), to assess the effectiveness of our lexical-based inference approach in detecting inconsistencies. The two domain experts reviewed the samples independently and then discussed the samples together to resolve disagreements.

IV. RESULTS

A. Summary Results

Using the 03/28/2017 release of GO, a total of 33,463 linked PMCPs were obtained, and derived 15,299 (distinct) linked ITPs. A total of 4,293,953 unlinked PMCPs were obtained, and derived 2,763,106 (distinct) unlinked ITPs. The ITPs derived include (*telencephalon*, *forebrain*), (*ethanolamine*, *peptide*), (*methotrexate*, *drug*), (*ethanolamine*, *peptide*), and (*cortisol*, *hormone*). A total of 3,715 potential inconsistencies were found.

B. Evaluation

Each detected inconsistency indicates a potentially missing subtype relation or an incorrect existing subtype relation in GO (a valid inconsistency), or is a falsely identified inconsistency (an invalid inconsistency).

A random sample of 158 detected inconsistencies was reviewed by the domain experts, and 89 were found to be valid inconsistencies. Among these, 62 were missing subtype relations and 27 were incorrect existing subtype relations. Therefore, the overall accuracy of our method is 56.33% (= 89/158).

Fig. 2 shows an example of incorrect existing subtype relation in GO, where the linked PMCP (GO:1901770, GO:0009109) in Fig. 2B is found to be an invalid subtype relation, because daunorubicin is not a coenzyme, it is a small molecule intercalating agent that inserts directly into the structure of DNA. That is, GO:1901770 (*daunorubicin catabolic process*) should not be a subtype of GO:0009109 (*coenzyme catabolic process*).

Table I lists 5 examples of the valid inconsistencies confirmed by the domain experts. Each example consists of the ITP, unlinked PMCP, linked PMCP, and inconsistency type (i.e., missing subtype relation or incorrect subtype relation).

V. DISCUSSION

In this paper, we investigated a lexical-based inference method to audit GO by detecting potential inconsistencies between linked and unlinked inferred term pairs. This method is not only applicable to GO, but also applicable to other terminologies for quality assurance analysis.

A. Distinction with Related Work

In [16], Agrawal et al. introduced the notion of Positional Similarity Set (PSS) to aid in the process of auditing SNOMED CT. PSSs are defined as lexically similar concepts having only one different word at the same position of their names. PSSs in combination with the concepts' structural definitions were used to identify unjustified modeling inconsistencies in SNOMED CT. While they studied modeling inconsistencies within a PSS, our work is focusing on detecting subtype defects in GO by leveraging the inconsistent ITPs derived across linked and unlinked PMCPs.

In [17], we investigated a structural-lexical approach to auditing the NCI Thesaurus, where one of the lexical patterns leveraged inferred terms in non-lattice subgraphs to suggest potentially missing *is-a* relations. In this work, although we share a similar idea of using lexical-based inference, we exhaustively consider all the linked and unlinked PMCPs for investigating potential inconsistencies in GO without limiting to any substructure. Furthermore, this work identifies potentially incorrect existing *is-a* relations in addition to missing *is-a* relations.

B. Limitations and Future Work

A limitation of this work is that we only considered hierarchical *is-a* relations in this work. Since some of the unlinked concept pairs identified by our method have already been linked through *part-of* relations in GO rather than *is-a*, we plan to incorporate *part-of* relations to our work in the future.

VI. Conclusion

In this paper, we investigated a lexical-based inference approach to audit Gene Ontology based on the inconsistencies of inferred term-pairs derived from linked and unlinked concept-pairs. This approach is found to be an effective way to detect subtype inconsistencies, which may indicate missing subtype relations as well as incorrect subtype relations in GO. This approach is also applicable to other biomedical terminologies for quality assurance analysis.

References

[1] G. O. Consortium *et al.*, "The gene ontology (GO) project in 2006," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D322–D326, 2006.

[2] K. Donnelly, "SNOMED-CT: The advanced terminology and coding system for eHealth," *Studies in health technology and informatics*, vol. 121, p. 279, 2006.

[3] D. Lee, N. de Keizer, F. Lau, and R. Cornet, "Literature review of SNOMED CT use," *Journal of the American Medical Informatics Association*, vol. 21, no. e1, pp. e11–e19, 2013.

[4] O. Bodenreider, "Biomedical ontologies in action: role in knowledge management, data integration and decision support," *Yearbook of medical informatics*, p. 67, 2008.

[5] G. L. Holliday, R. Davidson, E. Akiva, and P. C. Babbitt, "Evaluating functional annotations of enzymes using the gene ontology," *The Gene Ontology Handbook*, pp. 111–132, 2017.

[6] G.-Q. Zhang, L. Cui, S. Lhatoo, S. U. Schuele, and S. S. Sahoo, "MED-CIS: multi-modality epilepsy data capture and integration system," in *AMIA Annual Symposium Proceedings*, vol. 2014. American Medical Informatics Association, 2014, p. 1248.

[7] L. Cui, A. Bozorgi, S. D. Lhatoo, G.-Q. Zhang, and S. S. Sahoo, "EpiDEA: extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification," in *AMIA Annual Symposium Proceedings*, vol. 2012. American Medical Informatics Association, 2012, p. 1191.

[8] 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 *et al.*, "Gene Ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, p. 25, 2000.

[9] (2017) Gene ontology consortium - documentation. [Online]. Available: http://www.geneontology.org/page/documentation

[10] X. Zhu, J.-W. Fan, D. M. Baorto, C. Weng, and J. J. Cimino, "A review of auditing methods applied to the content of controlled biomedical terminologies," *Journal of biomedical informatics*, vol. 42, no. 3, pp. 413–425, 2009.

[11] C. Ochs, Y. Perl, M. Halper, J. Geller, and J. Lomax, "Quality assurance of the gene ontology using abstraction networks," *Journal of bioinformatics and computational biology*, vol. 14, no. 03, p. 1642001, 2016.

[12] G. Xing, G.-Q. Zhang, and L. Cui, "FEDRR: fast, exhaustive detection of redundant hierarchical relations for quality improvement of large biomedical ontologies," *BioData mining*, vol. 9, no. 1, p. 31, 2016.

[13] O. Bodenreider, M. Aubry, and A. Burgun, "Non-lexical approaches to identifying associative relations in the gene ontology," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. NIH Public Access, 2005, p. 91.

[14] F. Mougin, "Identifying redundant and missing relations in the gene ontology." in *MIE*, 2015, pp. 195–199.

[15] W. Ceusters, "Applying evolutionary terminology auditing to the gene ontology," *Journal of biomedical informatics*, vol. 42, no. 3, pp. 518–529, 2009.

[16] A. Agrawal, Y. Perl, C. Ochs, and G. Elhanan, "Algorithmic detection of inconsistent modeling among SNOMED CT concepts by combining lexical and structural indicators," in *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*. IEEE, 2015, pp. 476–483.

[17] R. Abeysinghe, M. Brooks, J. Talbert, and L. Cui, "Quality assurance of NCI Thesaurus by mining structural-lexical patterns," in *AMIA 2017 Annual Symposium Proceedings*. American Medical Informatics Association, In Press.