A Model for Assessing the Quality of Gene Ontology

Shuheng Wu
Florida State University
College of Communication and Information
sw09f@my.fsu.edu

Abstract

With the proliferation of bio-ontologies in the molecular biological community, concern remains about their quality. There are a number of frameworks and models have been developed to assess the quality of Gene Ontology, which is one of the most influential and widely used bio-ontologies in the community. However, without any theoretical guide for quality assessment, most of these frameworks and models are incomplete and intuitive, and cannot assess the quality of Gene Ontology consistently, systematically, and completely. This study uses a theoretical information quality assessment framework to guide the development of a quality evaluation model for the Gene Ontology, using both the conceptual and empirical approaches.

Keywords: data quality, quality evaluation, knowledge organization, ontologies, Gene Ontology

Introduction

Since the pubication of human genome in 2001, the world has entered into the "post-genome age" (Higgs & Attwood, 2005, p. 4). The vast growth in the amount of biological sequence data has led to the coexistence of heterogeneous data types, formats, and vocabularies delaying research process in biology and posing challenges for data management in biological repositories. There are urgent needs in scientific communities for knowledge organization (KO) systems (e.g., metadata schema, ontologies, Semantic Web) to provide access to and make sense of huge amount of scientific data (Gray, 2007; Gray et al., 2005). Scientific research has become increasingly multi-institutional, multinational, and interdisciplinary. This creates challenges for traditional KO systems to represent interdisciplinary data and improve data and metadata interoperability across disparate vocabularies and domains (Allard, 2012). It also raises the expectation on the quality of scientific data and its metadata (Anderson, 2004).

Due to the complexity of molecular biological entities (e.g., genes, proteins) and their relationships, there has been a trend towards the development and adoption of bio-ontologies in the biomedical and molecular biological communities. Among many current bio-ontologies, the Gene Ontology (GO) is one of the most influential in molecular biology and biomedicine, and has been widely used for text mining and information extraction (Blaschke, Hirschman, & Valencia, 2002). The difficulties of maintaining ontologies are in gaining community acceptance, integrating new knowledge, and reflecting established knowledge (Open Biological and Biomedical Ontologies, 2011). With the proliferation of bio-ontologies, concern remains about their quality.

Köhler, Munn, Rüegg, Skusa, and Smith (2006) proposed two automatic metrics—circularity and intelligibility—to assess the quality of term definitions in ontologies, and tested the metrics using empirical data collected from GO. Buza, McCarthy, Wang, Bridges, and Burgess (2008) developed a composite automatic quality metric—GO Annotation Quality (GAQ) score—to evaluate the quality of GO annotations, and tested it by measuring the annotations for chicken and mouse over a period of time in GO. The GAQ score is a product of the breath of annotation (i.e., the number of GO terms assigned to each gene product) and the evidence code (i.e., an indicator of the source of annotation) rank.

Acknowledgement: Many thanks to Dr. Besiki Stvilia for helpful conversations.
Wu, S. (2013). A model for assessing the quality of Gene Ontology. *iConference 2013 Proceedings* (pp.953-956). doi:10.9776/13492

Leonelli et al. (2011) collected empirical data from GO curators, and identified five quality problem sources: (a) mismatch between GO representation and reality; (b) the scope extension of GO; (c) divergence in how the GO terminology is used across user communities; (d) new discoveries that change the meaning of GO terminology and relationships; and (e) the addition of new relations. Defoin-Platel et al. (2011) proposed a framework of 12 quality metrics for assessing the quality of GO's functional annotations. Most of these frameworks and models are incomplete and intuitive, mainly based on individual perception of quality requirements. Without any theoretical guide for quality assessment, they are unable to assess the quality of GO consistently, systematically, and completely. With the increasing popularity and impact of GO in the molecular biological community, there is a great need for developing a comprehensive and systematic quality evaluation model for GO to inform KO system designers and curators and support users' decision making process.

Theoretical Framework

This study uses Stvilia's theoretical Information Quality (IQ) assessment framework (Stvilia, Gasser, Twidale, & Smith, 2007) to develop a quality evaluation model for GO. Stvilia' IQ assessment framework consists of a well-defined typology of IQ problem sources linked with affected information activities and a taxonomy of 22 IQ dimensions along with 41 generic IQ metrics. The reason for using Stvilia's framework is it provides consistent and complete logic to deal with context sensitivity, specifying methodologies to analyze an information activity system and identify users' IQ value structure and allowing rapid and inexpensive development of context-specific IQ measurement models. The generic IQ metrics have been successfully reused in different contexts (e.g., the English Wikipedia). Compared to previous IQ assessment frameworks, Stvilia's is more systematic, comprehensive, and reusable (Stvilia, 2006; Stvilia et al., 2007). Stvilia's framework has been operationalized in different settings (e.g., an online collaborative encyclopedia, an aggregated digital repository) and domains (e.g., biology, healthcare, and information science). Stvilia (2007) used his framework to construct a model for assessing the quality of biodiversity ontologies, and suggested future research to develop quality evaluation models for type-specific ontologies.

Methodology

Following Stvilia and Gasser's (2008) suggestion, this study uses both conceptual (top-down) and empirical (bottom-up) approaches, since a conceptual IQ model can guide the empirical analysis but may differ from a community's active IQ model and empirical data can reflect the community's actual IQ requirements. The conceptual approach is built upon an information entity's use scenarios to analyze its activity system context (Stvilia et al., 2007). This study uses activity theory (Leont'ev, 1978; Nardi, 1996; Vygotskii, 1978) and the findings of a related study (Wu, Stvilia, & Lee, 2012) to guide the conceptual analysis to identify the activities of using, developing, and maintaining GO, and the types of quality problems to which GO may be prone. The conceptualization of GO's activity system context and the suggested quality problem structure will be used to guide the empirical analysis to develop a quality evaluation model for GO.

The empirical approach involves qualitative and quantitative analyses on the Ontology and GO users' IQ evaluations, generating statistical profiles of GO and users' IQ value structure (Stvilia et al., 2007). GO has created the curator requests tracker to allow users to provide feedback to the Ontology, such as suggesting a new term or definition and reorganizing a section of the Ontology (Gene Ontology, 2012a; SourceForge, 2012). GO curators review users' requests, and implement edits where appropriate. GO also has a mailing list—GO Discuss—for users to report errors or omissions in GO annotations (Gene Ontology, 2012a, 2012c). Next in the empirical analysis, this study will collect GO users' requests and curators' comments from these sources to conduct content analyze to identify GO community's quality requirements. Finally, this study will relate the identified quality problems to the activities around and the measurable attributes of GO and GO annotations to develop a set of context-specific IQ metrics to evaluate the quality of GO.

Preliminary Findings

Go consists of three ontologies describing the cellular locations, molecular functions, and biological processes of genes and gene products in a species-neutral manner, and intends to provide each gene and gene product with a cellular context (Gene Ontology, 2012a). The development and maintenance of GO consist of three parts: (a) developing and maintaining GO terms and relationships among the terms; (b) annotating gene products, associating genes and gene products curating in collaborating databases (e.g., UniProt, WormBase) with GO terms; and (c) developing tools to facilitate the development, maintenance, and use of GO. Users can access GO terms and annotation data via a browser named AmiGO.

There are generally two types of annotation in GO: computational and manual (Gene Ontology, 2012b; Rogers & Ben-Hur, 2009). Computational annotation usually involves: (a) searching for a similar gene to a newly sequenced gene in BLAST; (b) searching AmiGo to find GO terms associated with the similar gene; and (c) assigning those GO terms to the new gene, assuming similar genes share similar cellular context. Biologists usually perform manual annotation through laboratory experiments to learn about the cellular context of genes and gene products. Manual annotation is accurate but time-consuming and labor-intensive. Compared to manual annotation, computational annotation is fast but less accurate and detailed. Each GO annotation contains an evidence code, indicating whether the annotation is inferred from primary research, literature, curators, computational analysis, or other databases.

Using three data use scenarios, a recent study identified and conceptualized the data quality problem sources in molecular biology as incomplete and inconsistent mapping, and dynamic quality problems caused by context changes, changes in the entity, and changes to the underlying entity (Wu et al., 2012). The data quality change model developed in that study may be applicable to GO as the concepts represented by GO are within the domain of molecular biology.

Based on the GO annotation activities, one may expect GO annotations have inaccurate, incomplete, and inconsistent mapping problems. An inaccurate/incomplete computational annotation is an instance of inaccurate/incomplete mapping between a gene product and GO terms. Inaccurate definition of a GO term may also cause inaccurate mapping between the GO term and gene products. An inconsistent mapping may occur when a GO collaborating database updates the cellular context of a gene product (e.g., new discovery of protein functions), but its GO annotation remains unchanged. One may also expect dynamic quality problems in GO terms and annotation data. Over a time period, there might be new discoveries that change the meaning of certain GO terms, and might change their relations to other GO terms and the annotation data. Similarly, there might be new discoveries of the attributes of gene products, which might change their GO annotations.

Conclusion

Guided by Stvilia's theoretical IQ assessment framework, this study aims to construct a model to assess the quality of GO using both conceptual and empirical approaches. The conceptual analysis of this study identifies the activities of creating, using, and maintaining GO, and suggests types of quality problems may occur in GO. This conceptualization will inform the empirical analysis, the next step of this study, to develop GO community's active IQ evaluation model.

References

Allard, S. (2012). DataONE: Facilitating eScience through collaboration. *Journal of eScience Librarianship, 1*, 4-17. doi:10.7191/jeslib.2012.1004

Anderson, W. L. (2004). Some challenges and issues in managing, and preserving access to, long-lived collections of digital scientific and technical data. *Data Science Journal, 3*, 191-202.

Blaschke, C., Hirschman, L., & Valencia, A. (2002). Information extraction in molecular biology. *Briefings in Bioinformatics, 3*, 154-165.

Buza, T. J., McCarthy, F. M., Wang, N., Bridges, S. M., & Burgess, S. C. (2008). Gene Ontology annotation quality analysis in model eukaryotes. *Nucleic Acids Research, 36*(2), e12. doi:10.1093/nar/gkm1167

Defoin-Platel, M., Hindle, M. M., Lysenko, A., Powers, S. J., Habash, D. Z., Rawlings, C. J., & Saqi, M. (2011). AIGO: Towards a unified framework for the analysis and the inter-comparison of GO functional annotations. *BMC Bioinformatics, 12*, 431. doi:10.1186/1471-2105-12-431

Gene Ontology. (2012a). *An introduction to the Gene Ontology*. Retrieved from http://www.geneontology.org/GO.doc.shtml

Gene Ontology. (2012b). *Annotation standard operating procedures*. Retrieved from http://www.geneontology.org/GO.annotation.SOP.shtml

Gene Ontology. (2012c). *GO mailing list*. Retrieved from http://www.geneontology.org/GO.mailing.lists.shtml

Gray, J. (2007). Jim Gray on eScience: A transformed scientific method. In: T. Hey, S. Tansley, & K. Tolle (Eds.), *The fourth paradigm: Data intensive scientific discovery* (pp. 5-12). Edmond, WA: Microsoft Research.

Gray, J., Liu, D. T., Nieto-Santisteban, M., Szalay, A., DeWitt, D. J., & Heber, G. (2005, February). Scientific data management in the coming decade. *CTWatch Quarterly, 1*(1). Retrieved from http://www.ctwatch.org/quarterly/articles/2005/02/scientific-data-management/

Higgs, P. G., & Attwood, T. K. (2005). *Bioinformatics and molecular evolution*. Malden, MA: Blackwell Publishing Company.

Köhler, J., Munn, K., Rüegg, A., Skusa, A., & Smith, B. (2006). Quality control for terms and definitions in ontologies and taxonomies. *BMC Bioinformatics, 7*, 212. doi:10.1186/1471-2105/7/212

Leonelli, S., Diehl, A. D., Christie, K. R., Harris, M. A., & Lomax, J. (2011). How the Gene Ontology evolves. *BMC Bioinformatics, 12*, 325. doi:10.1186/1471-2105-12-325

Leont'ev, A. (1978). *Activity, consciousness, personality*. Englewood Cliffs, NJ: Prentice–Hall.

Nardi, B. (1996). Studying context: A comparison of activity theory, situated action models, and distributed cognition. In B. Nardi (Ed.), *Context and consciousness: Activity theory and human–computer interaction* (pp. 35–52). Cambridge, MA: MIT Press.

Open Biological and Biomedical Ontologies. (2011). *Archive of original principles*. Retrieved from http://www.obofoundry.org/crit_2006.shtml

Rogers, M. F., & Ben-Hur, A. (2009). The use of gene ontology evidence codes in preventing classifier assessment bias. *Bioinformatics, 25*, 1173-1177. doi:10.1093/bioinformatics/btp122

SourceForge. (2012). *Gene Ontology: Tracker*. Retrieved from http://sourceforge.net/tracker/?group_id=36855

Stvilia, B. (2006). *Measuring information quality*. (Doctoral dissertation, University of Illinois at Urbana - Champaign). Retrieved from http://wwwlib.umi.com/dissertations/fullcit/3223727

Stvilia, B. (2007). A model for ontology quality evaluation. *First Monday, 12*(12). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2043/1905

Stvilia, B., & Gasser, L. (2008). Value-based metadata quality assessment. *Journal of Library and Information Science Research, 30*, 67-74. doi:10.1016/j.lisr.2007.06.006

Stvilia, B., Gasser, L., Twidale, M., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology, 58*, 1720-1733. doi:10.1002/asi.20652

Vygotskii, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wu, S., Stvilia, B., & Lee, D. J. (2012). Authority control for scientific data: The case of molecular biology. *Journal of Library Metadata, 12*, 61-83. doi:10.1080/19386389.2012.699822