Journal of Biomedical Informatics 42 (2009) 824–830

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# Text-mining approach to evaluate terms for ontology development

Lam C. Tsoi [a,1], Ravi Patel [b,1], Wenle Zhao [b], W. Jim Zheng [b,*]

[a] Bioinformatics Graduate Program, Department of Biostatistics, Bioinformatics & Epidemiology, Medical University of South Carolina, 135 Cannon Street, Suite 303, Charleston, SC 29424, USA

[b] Department of Biostatistics, Bioinformatics & Epidemiology, Medical University of South Carolina, Charleston, SC, 29464, USA

ARTICLE INFO

*Article history:*
Received 11 April 2008
Available online 24 March 2009

*Keywords:*
Ontology development
Hypergeometric test
PubMed
Text mining

ABSTRACT

Developing ontologies to account for the complexity of biological systems requires the time intensive collaboration of many participants with expertise in various fields. While each participant may contribute to construct a list of terms for ontology development, no objective methods have been developed to evaluate how relevant each of these terms is to the intended domain. We have developed a computational method based on a hypergeometric enrichment test to evaluate the relevance of such terms to the intended domain. The proposed method uses the PubMed literature database to evaluate whether each potential term for ontology development is overrepresented in the abstracts that discuss the particular domain. This evaluation provides an objective approach to assess terms and prioritize them for ontology development.



## 1. Introduction

High-quality ontologies such as the Gene Ontology (GO) [1] have been instrumental in analyzing data generated from microarray experiments [2–13]. However, developing such high-quality ontologies still poses significant challenges, as a wide range of literature and domain experts need to be involved. To aid ontology development, numerous methods have been developed to extract terms from literature automatically. Daille proposed combined techniques to extract terms automatically from corpora by combining linguistic filters and statistical methods [14]. Frantzi et al. developed a *C*-value/NC-value method to extract multi-word terms automatically [15]. By taking advantage of semantic relations encoded between terms, Grabar and Zweigenbaum developed a two-step approach to collect semantically related terms and to align morphologically linked work forms for term extraction [16]. A "weirdness metric" was proposed by Ahmad and Rogers to evaluate terms overrepresented in the domain-specific corpus for ontology development [17]. Savova et al. developed a data-driven approach to extract the "most specific term" for ontology development using an algorithm combining statistical and linguistic approaches [18]. Another tool developed to extract terms for ontology development was Text2Onto, which was built upon the Probabilistic Ontology Model [19]. In addition, Smith et al. proposed a machine learning approach to retrieve definitional content for ontology development [20]. Concept maps have also been used by Castro et al. in the ontology development process [21]. Alexopoulou et al. developed two additional methods, one based on the relative frequency of a term in the corpus and the other using the document frequency derived from all phrases contained in PubMed abstract database, to extract terms for ontology development [22]. While these methods focus on extracting terms from the published literature, two other studies also proposed to extract terms from web resources for concept and ontology development [23,24]. Despite these efforts, it is still widely recognized that manual curation is the most reliable method for ontology development [25], and these automatic term extraction methods are rarely used as a mainstream approach in the current biomedical ontology development process.

In the manual curation process, curators read as much scientific literature as possible for a particular biological domain in order to identify corresponding ontology terms and to classify their relationships to the domain and to other terms within it (such as "is_a" and "has_a" relationships). One significant challenge during this process is to determine which terms should be used as the basic building blocks from which to develop an ontology for a particular domain. This challenge is compounded by the fact that many curators with diverse backgrounds may be involved in developing the ontology for a given domain; diversity in their backgrounds can result in the selection of a wide variety of terms to be compiled within the ontology. This term selection process relies on the expertise of individual curators, without either a preliminary or confirmatory test using some objective method and measure. A quantitative approach to evaluate whether terms are appropriate to develop an ontology for a particular domain would provide this objective method and measure, improve the utility of the resulting ontology, and reduce the amount of work imposed on curators.

* Corresponding author. Fax: +1 843 876 1126.
*E-mail address:* zhengw@musc.edu (W.J. Zheng).

[1] These two authors made equal contribution to this work.

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.03.009

While the above mentioned term extraction algorithms or their underlying metrics have the potential to be used to evaluate terms assembled by experts during manual curation for ontology development, several limitations exist for such an application. First, most of these algorithms are developed to extract terms from a corpus of selected literatures already identified by experts (in many cases involving manual selection) as relevant to a specific domain. Lacking such a pre-defined corpus, as is the case for many ontology development projects, the effectiveness of these methods is not clear. Second, the volume of existing literature databases like PubMed is significantly larger than a corpus related to a specific domain; this corpus size difference raises questions about the performance of these algorithms if applied to evaluate terms using PubMed abstracts in the absence of a domain-specific corpus. Furthermore, these methods have not been widely tested against manually assembled ontology terms. New approaches capable of dealing with large databases and confirmed through comparison to manually curated ontologies need to be developed.

One objective criterion to evaluate a term's suitability for incorporation in an ontology for a particular concept domain is to quantify the term's relevance to the domain within published biomedical literature. If a term occurs at high frequency in PubMed abstracts relevant to the concept domain, then it should be more suitable for ontology development than other terms occurring at low frequencies. Because ontology development aims to describe particular biological domains, we hypothesized that the terms used within an already existing ontology for a particular domain would be overrepresented in the PubMed abstracts relevant to that domain (Domain PubMed Abstract, DPA). We further hypothesized that the degree of overrepresentation could be detected by employing a hypergeometric enrichment test.

Testing based on hypergeometric distribution has been applied in the analysis of GO overrepresentation on biologically-interesting gene sets (review see [26]). Hypergeometric testing can measure the association between a term and the domain by calculating the probability of observing the term within the DPA as long as both categories are sampled without replacement from a finite population. Such a probability can be used as a direct measure of how relevant a term is to the domain: the higher probability we observe a term in the DPA, the more overrepresented this term is in the DPA, and the more relevant this term is. The degree of overrepresentation of terms relevant to a domain in the DPA can indicate the usefulness of the terms for developing this domain's ontology. Experts from diverse fields could use the information gleaned through such a hypergeometric evaluation to narrow candidate terms for a given ontology. The test could significantly reduce the manual effort involved in ontology development.

In this study, we first used GO [1] as a control to evaluate whether the proposed text-mining approach could detect the overrepresentation of ontology terms in the corresponding DPA. We demonstrated that the hypergeometric test could capture the relevant terms in the DPA and reflect their relative importance by their overrepresentation. We then demonstrated that this approach could be used to evaluate putative ontology terms generated by different experts for the development of a Clinical Trial Ontology/Ontology for Clinical Investigation [27]. Our results indicated that such a computational algorithm can provide an objective measure for the selection of putative ontology terms to facilitate ontology development.

## 2. Methods

### 2.1. PubMed database preparation

Fig. 1 illustrates a condensed version of our process. The entire PubMed database (2007) in XML format was downloaded from NCBI. The database was processed to extract all abstracts. Necessary formatting, such as capitalizing all the abstracts and removing special characters, was performed (box 'preprocessing'). All the software was implemented in C++. PubMed stopwords were also downloaded from NCBI.

### 2.2. Collection of GO and other terms

*Terms for evaluation by hypergeometric enrichment test:* In order to test whether terms relevant to a domain were overrepresented in the corresponding DPA, we identified test sets where each set had a domain term and a list of terms known to be relevant to that domain. To generate these test sets, we took advantage of the hierarchical structure of GO. In an ontology such as GO, terms with specific meaning are children of terms that are more general, thus comprising an "is_a" relationship. We viewed a parent term as a domain and the child term as a term relevant to that domain. This approach can be extrapolated to multi-level hierarchies such that an ontology term at a high level of the hierarchy can be viewed as a domain term, and all of its child terms can be viewed as terms that are relevant to this domain.

GO was downloaded from the GO Consortium website (http://www.geneontology.org, June, 2007). Two terms were selected from GO as domain terms for our study. The criteria for selection were: (1) each term had more than 50 child terms under an "is_a" relationship to yield a significant number of child terms relevant to the domain; (2) the selected domain term was not a child term of another domain term. Two domain terms, Monosaccharide Metabolic Process (GO:0005996) and Protein Kinase Activity (GO:0004672), were selected for this purpose. Monosaccharide Metabolic Process is categorized as a biological process within GO, and Protein Kinase Activity falls within the molecular function category. Descendents of these two terms were also collected. Since we focus on the terms that describe biological systems, common words such as "activity" or "process" were removed from these descendent terms (henceforth, such terms are shown in brackets). This practice was limited to the GO terms used for validation purposes and was not applied to the putative ontology terms evaluated in subsequent analyses. The number of unique terms after removing these common words was 56 and 97 for Monosaccharide [Metabolic Process] and Protein Kinase [Activity], respectively. The degree of overrepresentation of these descendent terms in the DPA was tested as described below.

We also selected additional GO terms as controls. These terms were randomly chosen from a pool of GO terms that had no descendant-ancestor relationships with Monosaccharide [Metabolic Process] or Protein Kinase [Activity]. The control terms iden-

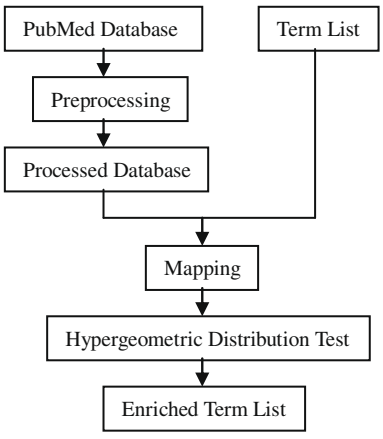


**Fig. 1.** Workflow of enrichment test to evaluate terms for ontology development.