Occurrence of Gene Ontology, Protein Ontology, and NCBI Taxonomy Concepts in Text toward Automatic Gene Ontology Annotation of Genes and Gene Products

Michael Bada[1], Dmitry Sitnikov[2], Judith A. Blake[2], and Lawrence E. Hunter[1]*

[1]*University of Colorado Anschutz Medical Campus, Aurora, CO, USA*

[2]*Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME, USA*

ABSTRACT

Annotations of genes and gene products in model-organism databases with Gene Ontology (GO) terms have become an important knowledge resource in biomedical research, which has spurred many efforts at automating this labor-intensive manual curatorial activity, including many text-mining approaches. In an effort to provide some guidance on these text-mining efforts, we have used a gold-standard manually annotated corpus to conduct an evaluation of the occurrence of three types of fundamental GO-annotation concepts in 34 journal articles that were the evidential bases of approximately 220 GO annotations largely created by the Mouse Genome Informatics (MGI) group.

In addition to an analysis of the occurrence of the GO concepts of the curated GO annotations associated with these articles in the corpus, we have analyzed the occurrence of NCBI Taxonomy (NCBITAXON) and Protein Ontology (PRO) concepts corresponding to the species-specific genes/gene products of these curated GO annotations. The GO, NCBITAXON, and PRO concepts corresponding to the curated GO annotations were analyzed both in the full-text versions of these articles as well as in only those sentences of the articles providing the strongest evidence for the GO annotations, as specified by an official MGI GO curator. Though this sample set may not necessarily be representative of all GO annotations, our analysis suggests that full-text articles mention substantial fractions of the GO concepts at least once; however, the mentions of these GO concepts constitute very low percentages of the mentions of all GO concepts in these articles. Nearly all PRO concepts corresponding to GO annotations are mentioned at least once in the full articles, and these PRO mentions constitute a substantial fraction of the mentions of all PRO concepts in these articles. *Mus musculus* is seldom mentioned, though mice (strictly corresponding to the genus *Mus*) are mentioned at least once in the full articles, and these *Mus* mentions also constitute a substantial fraction of the mentions of all NCBITAXON concepts in these articles. For all of the ontologies, counts of annotated concepts corresponding to the curated GO annotations in only the strongly evidential sentences are comparatively very low, amounting to several mentions or fewer per article. However, for most of the ontologies, concepts corresponding to the curated GO annotations appear overrepresented, though this must be viewed cautiously given that this is based on very low counts. Thus, it remains to be further examined whether this overrepresentation overrides the very low mention frequency and thus whether it would be beneficial for automatic GO-annotation systems to focus on these evidential sentences.

1 INTRODUCTION

Annotations of genes and gene products in model-organism databases with Gene Ontology (GO) terms have become an important knowledge resource in biomedical research (The Gene Ontology Consortium, 2000; Camon *et al.*, 2004; Lee *et al.*, 2005). This has spurred many efforts at automating this labor-intensive manual curatorial activity, including text-mining approaches (Camon *et al.*, 2005; Winnenburg *et al.*, 2008). In an effort to provide some guidance on these text-mining efforts, we have conducted an evaluation of the occurrence of three types of fundamental GO-annotation concepts in articles that were the evidential bases of GO annotations largely created by the Mouse Genome Informatics (MGI) group, who curate a wide range of data for the primary international database resource for the laboratory mouse (Drabkin and Blake, 2012).

For this effort, we have employed the Colorado Richly Annotated Full-Text (CRAFT) Corpus, a gold-standard corpus of journal articles whose full-text versions have been manually marked up with ~140,000 concept annotations, relying on nearly all of the concepts of eight prominent biomedical ontologies; it has also been manually marked up in a variety of other ways, including syntactic, coreferential, and discourse annotation (though these other types of annotation were not analyzed in this study) (Bada *et al.*, 2012; Verspoor *et al.*, 2012). In addition to an examination of the occurrence of the GO concepts of the curated GO annotations associated with these articles in the corpus, we have

* To whom correspondence should be addressed.

analyzed the occurrence in the corpus of NCBI Taxonomy (NCBITAXON) (Sayers *et al.*, 2009) and Protein Ontology (PRO) (Natale *et al.*, 2011) concepts corresponding to the species-specific genes/gene products of these curated GO annotations. The GO, NCBITAXON, and PRO concepts corresponding to the curated GO annotations were analyzed both in the full-text versions of these articles as well as in only those sentences of the articles providing the strongest evidence for the GO annotations, as specified by an official MGI GO curator.

It is important to note that we are not examining the occurrence of these concepts merely in terms of their mentions as exact string matches to the concepts' primary labels or synonyms in the text of these articles. Rather, every mention semantically equivalent to a concept in one of these ontologies (with rare exception) has been annotated with the corresponding concept according to the CRAFT concept-annotation guidelines (Bada *et al.*, 2010), whether or not the textual mention matches the concept label or one of its synonyms. Thus, ours is an analysis of the potential for recognizing the species, genes/gene products, and biological functionalities of GO annotations by looking for these concepts in text rather than an examination of what is possible with current text-mining technology, which would very likely miss some of these gold-standard concept annotations and incorrectly annotate other spans of text.

2 METHODS

We omit from this paper discussion of the markup of the concept annotations of the CRAFT Corpus (including the NCBITAXON, PRO, and GO concept annotations, which are analyzed in this study), as it has been extensively described elsewhere (Bada *et al.*, 2012).

The markup of the sentences of the articles of the CRAFT Corpus providing strong evidence upon which curators most relied for their GO annotations of genes/gene products associated with these articles was performed within Knowtator (Ogren, 2006), a tool for ontology-based annotation of text implemented as a tab plugin for Protégé-Frames (Gennari *et al.*, 2003); this is the same tool that was used to perform all of the concept annotation of the CRAFT Corpus. For the annotation of these evidential sentences, a simple ontology was manually constructed, including one class representing GO annotations and another representing evidence annotations. For the former, properties were defined for the Entrez Gene ID of the annotated gene, the GO term ID and primary label used for the annotation, the GO evidence code of the annotation (which specifies the type of evidence supporting the annotation), qualifier(s) of the annotation (such as one indicating negation), and for the sentences supporting the given GO annotation. A new Protégé-Knowtator project was created based on this ontology, and the GO-annotation class was programmatically populated with instances of curated GO annotations, with the appropriate values of all properties (except for that for the evidential sentences) filled in. The annotation of the evidential sentences was performed by an official MGI GO curator (DS), who, for each curated GO annotation instance, created instances of evidence annotations by selecting appropriate sentences and added them as property values for the appropriate GO-annotation instances. This markup was periodically reviewed by the project lead (MB) to check that the evidential sentences were being consistently marked up.

Though the full CRAFT Corpus consists of 97 articles, only 67 of the articles have been included in the 1.0 public release. (The other 30 are being temporarily reserved for use in future text-mining competitions, after which these too will be released.) The articles of the CRAFT Corpus were partly selected based on their serving as evidential bases for curated annotations of genes/gene products with terms from the GO and/or the Mammalian Phenotype Ontology (MPO) (Smith and Eppig, 2010). Since in this study we are analyzing only the GO annotations associated with these articles, we narrowed down the 67 publicly released articles to the 36 articles associated with one or more curated GO annotations. One of these articles (PMID:14611657) is an outlier in that it is associated with 4,524 curated GO annotations; this very large number of annotations for this paper (which correspond to a large set of olfactory receptor genes identified through a screening of a mouse olfactory epithelium cDNA library) would have completely eclipsed all of the other annotations in this study, and so we excluded this paper and its annotations. Another paper (PMID:16870721) was associated with one curated GO annotation, but during the course of this project, it was discovered to be an erroneous annotation; it has since been removed by MGI from its database, and so this paper and its annotation were excluded from this study as well. Excluding these two papers results in 34 papers with 254 curated GO annotations. An additional 28 curated GO annotations were excluded from this study since no evidential sentences were selected from the corresponding articles for these annotations by the MGI GO curator. (The large majority of these annotations were based on sequence or structure similarity of the annotated genes/gene products to homologous sequences (GO evidence code ISS) that presumably were studied in the corresponding articles.) This resulted in 34 papers with 226 curated GO annotations.

This study includes an analysis of the occurrence of NCBITAXON and PRO concepts in these articles; however, the curated GO annotations were originally specified for genes/gene products by their Entrez Gene IDs. Therefore, we mapped these Entrez Gene IDs (which refer to species-specific genes) to their corresponding NCBITAXON and PRO concepts, designating species and species-nonspecific genes/gene products, respectively. Properties for PRO and NCBITAXON IDs were created in the Protégé-Knowtator

project for the class of GO annotations, and the values for these IDs were manually entered for all of the GO-annotation instances. Thus, in the Protégé-Knowtator project of curated GO annotations, each GO annotation instance is formally associated with its corresponding GO, NCBITAXON, and PRO concepts as well as with the evidential sentences supporting the given annotation, all of which can be programmatically queried.

We subsequently noticed that for eight of the articles, there was a pair of curated GO annotations with identical NCBITAXON-PRO-GO triples; these are the result of the two annotations of a pair having only different GO evidence codes (*i.e.*, based on different types of evidence) or of one of the two annotations of a pair having an additional qualifier specifying that the gene/gene product contributes to a functionality rather than possessing the functionality itself. For the full-text article analysis, since we are only analyzing the occurrence of NCBITAXON, PRO, and GO concepts in the full-text versions of the articles and not these other aspects, we removed one of each of these pairs so as not to double-count them within their respective articles, resulting in 218 curated GO annotations for the full-text analysis. For the analysis of the concepts only within the evidence annotations, four of the eight pairs of duplicate NCBITAXON-PRO-GO annotations have different evidence annotations and so the concept annotations to be analyzed are in different "documents", *i.e.*, pieces of text; thus, for these four pairs, one annotation of each of the pairs was added back, resulting in 222 curated GO annotations for the evidence-annotation analysis.

Our analysis was implemented in a Java program. First, the Protégé-Knowtator project of curated GO annotations was queried via the Knowtator Java API for the NCBITAXON, PRO, and GO IDs and the start and end character positions of the span(s) of the evidence annotations for each GO annotation, and a mapping of the articles to their corresponding GO annotations was dynamically created. Then, for each category of concept annotation, the concept annotations for each article were retrieved. For the analysis of the occurrence of these concepts in the full-text versions of the articles, for each article each of its concept annotations was queried via the Protégé Java API to determine if this concept was an exact match to the concept used in each of the GO annotations associated with the given article, a subclass, a superclass, or none of these. For the analysis of the occurrence of these concepts in only the evidence annotations, for each article each of its concept annotations was first queried for its start and end character positions of its span(s), and these spans were compared to the spans of the evidence annotations of each of the GO annotations associated with the given article; only if the span(s) of the given concept annotation were found to be within the span(s) of any of the evidence annotations of a given associated GO annotation was the concept used in the concept annotation compared to the concept used in the GO annotation. These analyses were done once each for the NCBITAXON and PRO concepts and for each of the three branches of the GO, *i.e.*, biological processes (BP), molecular functions (MF), and cellular components (CC).

3 RESULTS

Table 1 displays statistics for the percentages and fractions of the curated GO annotations for which there is at least one annotated mention of the associated NCBITAXON concept, both in the full-text versions of the articles and only in the evidential sentences identified for the annotations. As shown in the table, there are few articles explicitly mentioning species exactly corresponding to the curated GO annotations: The species of only 7.3% and 0.9% of the curated GO annotations is annotated at least once in the corresponding full-text articles and the evidential sentences, respectively.

analysis	exact	superclasses	*Mus*
full-text articles	7.3% (16/218)	100% (218/218)	99% (215/218)
evidence only	0.9% (2/222)	47% (102/222)	44% (96/222)

Table 1. Percentages and fractions of curated GO annotations for which there is at least one annotated mention of the associated NCBITAXON concept, either in the full-text articles or only in the evidential sentences. The second, third, and fourth columns respectively show statistics for exact NCBITAXON concept matches, for all superclasses of the exact species concept, and for only the genus *Mus* (the taxon of all mice).

In addition to an analysis of the occurrence of annotated mentions of concepts exactly corresponding to the curated GO annotations, we have included analysis of all superclasses (*i.e.*, ancestors) of the directly corresponding concepts (*e.g.*, for the species *Mus musculus*, a mention of the genus *Mus*, the subfamily *Murinae*, the family *Muridae*, the order *Rodentia*, etc.). We have included analysis of such superclass annotations throughout our study because even though these are not mentions of concepts exactly corresponding to the curated GO annotations, at least some of these may be coreferential mentions of the exact species or may be mentioned within assertions pertaining to the ancestor concepts that also hold true for the directly corresponding concepts; therefore, they are potentially useful as well. As Table 1 shows, there is at least one mention of an ancestor NCBITAXON concept for 100% and 47% of the curated GO annotations in the corresponding full-text articles and in the evidential sentences, respectively. (Though there are concepts more specific than species in the NCBI Taxonomy, (*e.g.*, subspecies, strains), there are no such annotated mentions of such subclasses of organisms associated with the GO annotations in these articles.)

The small fractions of articles with annotated mentions of species exactly corresponding to the curated GO annotations

is somewhat deceptive in that mentions of species are very often referred to as higher-level taxa; for example, the most common species of laboratory fruit fly, *Drosophila melanogaster*, is often referred to as "*Drosophila*" (indicating its genus, which contains more than 1,500 species), "fruit fly" (a common name that can refer to this genus), or even "fly" (a common name than can also refer to the order *Diptera*, the higher-level taxon of (true) flies, which contains an estimated 240,000 species). As MGI focuses on compiling data and knowledge for the laboratory mouse, a large majority of the GO annotations examined in this study pertain to the most common species of laboratory mouse, *Mus musculus*, which is analogously commonly referred to as "mouse". For all of the concept annotations of the CRAFT Corpus, we consistently sought to annotate mentions with the closest semantic match to the selected text, even in cases in which a more specific class is known from context; we have found that such a strategy avoids a great amount of labor in many cases and reduces error overall. Therefore, mentions of mice (*i.e.*, mice not explicitly mentioned as a specific species) are annotated with the NCBITAXON term *Mus*, the genus of mice, and not with *Mus musculus*, the species colloquially known as the house mouse, even in cases where it is known to be referring to the house mouse. As most mentions of mice in these articles (which are annotated with *Mus*) very likely refer to *Mus musculus*, we also specifically examined the occurrence of *Mus* annotations for GO annotations of *Mus musculus* genes/gene products in our study. Table 1 shows that *Mus* is mentioned at least once in the corresponding full-text articles and evidential sentences of 99% and 44% of the curated GO annotations, respectively.

As for occurrence of PRO concepts, we found that the corresponding PRO concepts of 98% (213/218) and 83% (180/222) of the curated GO annotations are mentioned at least once in the corresponding full-text articles and evidential sentences, respectively. The hierarchical structure of the Protein Ontology is relatively flat, with many protein concepts as children of the ontology's basic protein concept, and the overwhelming majority of the PRO concept annotations were made using classes from this level. Thus, we did not examine the occurrence of superclasses and subclasses of the corresponding PRO concepts of the GO annotations.

As for the concepts from the three branches of the GO exactly matching the GO concepts used in the curated GO annotations, we found that CC concepts were most often mentioned at least once (with 59% and 55% of the GO CC annotations mentioned in the full-text articles and evidential sentences, respectively), followed by GO MF (with 39% and 27%, respectively) and GO BP (with 33% and 16%, respectively). There are also mentions of concepts more specific than the GO concepts used in the GO annotations; as these are subclasses of the latter, they deductively infer the former: Such subclasses are mentioned at least once in the full-text articles and evidential sentences for 7.4% and 0% of the curated GO BP annotations, respectively; for 45% and 12% of the GO MF annotations, respectively; and for 12% and 4% of the GO CC annotations, respectively. Finally, superclasses of the GO concepts used in GO annotations are mentioned at least once in the full-text articles and evidential sentences for 88% and 43% of the GO BP annotations, respectively; for 82% and 42% of the GO MF annotations, respectively; and for 69% and 45% of the GO CC annotations, respectively. These data for GO concepts are shown in Table 2.

analysis	exact	subclasses	superclasses
BP full-text articles	33% (45/136)	7.4% (10/136)	88% (119/136)
BP evidence only	16% (22/140)	0% (0/140)	43% (60/140)
MF full-text articles	39% (13/33)	45% (15/33)	82% (27/33)
MF evidence only	27% (9/33)	12% (4/33)	42% (14/33)
CC full-text articles	59% (29/49)	12% (6/49)	69% (34/49)
CC evidence only	55% (27/49)	4% (2/49)	45% (22/49)

Table 2. Percentages and fractions of curated GO annotations for which there is at least one annotated mention of the associated GO BP, MF, or CC concept, either in the full-text articles or only in the evidential sentences. The second, third, and fourth columns respectively show statistics for the exact GO concepts, for all subclasses of the exact GO concepts, and for all superclasses of the exact GO concepts.

ontology	total annotations	average/median annotations per article	min/max annotations per article
NCBITAXON	3,566	105 / 95	19 / 378
PRO	8,437	248 / 238	61 / 625
GO BP	8,366	246 / 218	25 / 781
GO MF	2,106	62 / 50	4 / 235
GO CC	4,102	121 / 116	10 / 347

Table 3. Total counts of NCBITAXON, PRO, GO BP, GO MF, and GO CC annotations in the 34 journal articles analyzed in this study, with average, median, minimum, and maximum counts of annotations per article.

To provide context of the occurrence of the NCBITAXON, PRO, and GO concepts associated with the curated GO annotations relative to the occurrence of all NCBITAXON, PRO, and GO concepts in these articles, Tables 3 and 4 show counts for all NCBITAXON, PRO, and GO annotations in the journal articles analyzed in this study, along with counts of unique mentions of all of these concepts. In Table 3, it can be seen that the average/median annotation counts for all concepts of these ontologies per article range from 62/50 GO MF annotations per article to 248/238 PRO annotations per article; note, however, the very wide range of these counts in the minimum and maximum annotations per article for all of the ontologies. In Table 4, it can be seen that the average/median counts of unique concepts mentioned at least once per article range from 12/10 unique NCBITAXON concepts mentioned per

article to 44/44 unique GO BP concepts mentioned per article; as with the annotation counts, there is a very wide range of counts of unique concepts mentioned, as seen in the minimum and maximum counts for all of the ontologies. (Corresponding counts for all 67 articles of the 1.0 release of the corpus have been previously published (Bada *et al.*, 2012).)

ontology	total unique concepts	average/median unique concepts per article	min/max unique concepts per article
NCBITAXON	113	12 / 10	3 / 49
PRO	523	19 / 20	5 / 40
GO BP	488	44 / 44	11 / 95
GO MF	187	13 / 12	1 / 26
GO CC	158	13 / 11	1 / 33

Table 4. Total counts of unique NCBITAXON, PRO, GO BP, GO MF, and GO CC concepts annotated in the 34 journal articles analyzed in this study, with average, median, minimum, and maximum counts of unique concepts annotated per article.

analysis	exact	superclasses	*Mus*
full-text articles	1.1 (1%)	66 (63%)	47 (45%)
evidence only	0.009 (0.6%)	1.7 (119%)	1.4 (97%)

Table 5. Counts of annotated NCBITAXON concept mentions associated with curated GO annotations, averaged over these GO annotations. These are also expressed as percentages relative to the average counts of all NCBITAXON annotations either throughout the entire articles (105, shown in Table 3) or only in the evidential sentences (1.4, data not shown). The second, third, and fourth columns respectively hold data for the exact concepts, for superclasses of the exact species concept, and for the genus *Mus* (the taxon of all mice).

Efforts at automatic GO annotation that attempt to find relevant GO concepts in text must not only be able to accurately identify concept mentions but also to choose the concept mentions that are relevant for GO annotation of genes/gene products from all of the identified concept mentions. For the PRO, there was found to be an average of 94 annotated mentions of the exact PRO concepts associated with the curated GO annotation in the full-text articles, amounting to 38% of the average total annotated PRO mentions per article of 248 (shown in Table 3); there was found to be an average of only 4.9 annotated mentions of the exactly matching PRO concepts in only the evidential sentences, but amounting to 92% of the average annotated mentions of all PRO concepts in the evidential sentences of 5.3 (data not shown). For the NCBI Taxonomy, there were found to be averages of 1.1 mention of the exact species and 47 mentions of *Mus*, respectively amounting to 1% and 45% of the average total NCBITAXON annotations per article of 105 (shown in Table 3); the NCBITAXON data are shown in Table 5. Table 6 shows corresponding data for the GO concepts associated with curated GO annotations, all of which were found to have very low mention numbers averaged over the GO annotations.

analysis	exact	subclasses	superclasses
BP full-text articles	5 (2%)	0.4 (0.1%)	25 (10%)
BP evidence only	0.2 (7%)	0 (0%)	2 (28%)
MF full-text articles	3 (5%)	5 (8%)	22 (35%)
MF evidence only	0.7 (19%)	0.7 (20%)	2 (59%)
CC full-text articles	8 (7%)	0.4 (0.3%)	10 (8%)
CC evidence only	2 (39%)	0.06 (1%)	1 (19%)

Table 6. Counts of annotated GO BP, MF, and CC concepts associated with curated GO annotations, averaged over these GO annotations. These counts are also expressed as percentages relative to the average counts of all GO BP, MF, or CC annotations either throughout the entire articles (246, 62, and 121, respectively, shown in Table 3) or only in the evidential sentences (5.4, 3.5, and 5.7 respectively, data not shown). The second, third, and fourth columns respectively hold data for the exact concepts, for subclasses of the exact concepts, and for superclasses of the exact concepts.

4 DISCUSSION

We have employed the 1.0 public version of the CRAFT Corpus to undertake an analysis of the occurrence in the corpus articles of annotations of NCBITAXON, PRO, and GO concepts corresponding to a set of official GO annotations largely curated by the Mouse Genome Informatics group. We have specifically relied on the concept annotations of the corpus, in which every mention of (nearly) every explicitly represented concept of eight prominent biomedical ontologies has been annotated with its corresponding ontological concept according to the CRAFT concept-annotations guidelines.

We have taken advantage of the fact that the articles of the CRAFT Corpus were selected partly based on their serving as evidential sources for curated annotations of genes/gene products with GO and/or MPO classes. With the complete concept annotation of these articles with the aforementioned ontologies, we are provided with an opportunity to evaluate the occurrence of NCBITAXON, PRO, and GO concepts corresponding to three fundamental elements of GO annotations—species, genes/gene products, and biological functionalities. The markup of the sentences in these articles supplying the strongest evidence for these GO annotations, as specified by an official MGI curator, provides us an additional opportunity to also examine the occurrence of these concepts in these sentences.

Though our relatively small sample set of GO annotations and their corresponding articles may not be representative of all GO annotations, this is the first study of a comparison of these components of GO annotations with a completely annotated gold-standard corpus of corresponding articles of which we are aware. As GO annotations have become an important knowledge resource for biomedical research, au-

tomatic methods of GO annotation have become of interest, including a number of text-mining approaches. Since such approaches often look for mentions of relevant concepts in articles, it is of interest to determine the degree of potential of finding the species, gene/gene products, and aspects of biological functionality directly in the text.

In the full-text articles, we have found that the GO BP, MF, and CC concepts exactly matching the GO concepts of these curated GO annotations are mentioned at least once in substantial fractions (33%, 39%, and 59%, respectively) of the annotations' corresponding articles; this coheres with empirical studies of occurrence of GO terms in biomedical text in which relatively low percentages of GO BP terms and higher percentages of GO CC terms are found due to their respective higher and lower complexities (*e.g.*, McCray *et al.*, 2002). Additionally, concepts that are more specific than (*i.e.*, subclasses of) the GO concepts of the GO annotations are mentioned at least once in corresponding articles in substantially smaller fractions of the BP and CC annotations (7.4% and 12%, respectively) and in a slightly higher fraction of the MF annotations (45%) (though the difference in absolute numbers of annotations is very small). In addition to the exactly matching concepts, occurrence of such subclasses is of interest because an annotation with a subclass deductively infers an annotation with each of the subclass's superclasses—including the exactly matching class—according to the GO true-path rule (Camon *et al.*, 2012). Concepts more general than (*i.e.*, superclasses of) of the GO concepts used in the curated GO annotations are mentioned at least once in larger fractions of the annotations' corresponding articles, which is intuitive since more general classes are more likely to be mentioned than specific ones. Even though these superclass annotations do not exactly correspond to the GO concepts of the curated GO annotations (nor do they deductively infer the exact GO concepts, as do the subclass annotations), they are potentially useful, since at least some of them may be coreferential mentions of the exact GO concepts or may be mentioned within assertions pertaining to the ancestor concepts that also hold true for the exact concepts.

NCBITAXON concepts representing the species of the curated GO annotations are very infrequently mentioned explicitly. However, mice (which in the corpus is annotated to the genus *Mus*) are very frequently mentioned in the articles corresponding to the GO annotations, and we expect such frequent mentions of higher-level common names for other species. These mentions of higher-level taxa are not guaranteed to refer to the most common laboratory species, of course; for example, there are species of mice other than *Mus musculus* that are used in laboratory experiments. However, such mentions could be leveraged as very reasonable abductive inferences. PRO concepts are also mentioned in very high fractions of the articles corresponding to the GO annotations, and the PRO concepts associated with the curated GO annotations constitute substantial fractions of the total mentions of all PRO concepts in the articles.

An interesting trend that can be seen for all of the ontologies concern the counts of annotated mentions of the concepts corresponding to the curated GO annotations, averaged over these GO annotations, in the full-text articles versus only in the evidential sentences. For all of the ontologies, the absolute counts of the associated ontological concepts in only the evidential sentences as expected are much lower than those in the full-text articles. At the same time, the annotated ontological concepts associated with the curated GO annotations seem to be overrepresented in the evidential sentences in that the ratios of counts of the ontological concepts associated with the GO annotations to the counts of all of the ontological concepts are higher within the evidential sentences than throughout the full-text articles. However, the counts within the evidential sentences are very low, this appearance of overrepresentation should be treated cautiously. Identification of these types of sentences seems a difficult task, as many of the evidential sentences annotated by the MGI curator describe low-level data (*e.g.*, phenotypes of animals, biochemical assays), and the functionalities of the genes/gene products are often not straightforwardly mentioned but are inferrable from the experimental results by domain experts. It remains to be seen whether a system could reliably identify such passages relevant for GO annotations, and if so, whether it might be beneficial to do so for the task of automatic GO annotation.

A caveat that should be stated is that not all of these annotated concept mentions will be relevant to the curated GO annotations associated with the articles. Therefore, we regard the statistics presented as upper bounds toward the inference of the GO annotations from the direct mention of these component concepts in text. A rigorous gold-standard investigation of which of these mentions are relevant to GO annotations extracted from the text would require an additional layer of annotation in which these concept annotations are appropriately relationally linked with each other. These annotated assertions, which would include links among mentions of species, genes/gene products, and aspects of biological functionality, could then be analogously analyzed to identify GO-annotation information. We do plan on performing such assertional annotation in the future. However, as was the case with the concept annotation of the CRAFT Corpus, this will almost certainly be a labor-intensive, multiyear effort. Furthermore, even annotation of direct assertions may not be sufficient, as GO-annotation information may require the use of coreferential information as well as other types of inference.

5 CONCLUSIONS

We have evaluated the potential for programmatic extraction of GO annotations of genes/gene products by examining the occurrence of direct mentions of NCBITAXON, PRO, and GO concepts corresponding to official curated GO annotations in a gold-standard manually annotated corpus of full-text articles partly selected as the basis for such GO annotations. GO concepts exactly matching the GO concepts of these GO annotations are mentioned in substantial fractions of the annotations' corresponding articles at least once; however, the mentions of these GO concepts constitute very low percentages of the mentions of all GO concepts in these articles. Nearly all PRO concepts corresponding to GO annotations are mentioned at least once, and these PRO mentions constitute a substantial fraction of the mentions of all PRO concepts in these articles. *Mus musculus* is seldom mentioned, though mice (strictly corresponding to the genus *Mus*) are very frequently mentioned, and as for the PRO, these *Mus* mentions constitute a substantial fraction of the mentions of all NCBITAXON mentions in these articles. This suggests that automatic textual recognition of PRO and NCBITAXON concepts relevant to GO annotations appear quite tractable relative to textual recognition of associated GO concepts. For all of the ontologies, counts of annotated concepts corresponding to the curated GO annotations in only the strongly evidential sentences are comparatively very low, amounting to several mentions or fewer per article. However, for most of the ontologies, concepts corresponding to the curated GO annotations appear overrepresented, though this must be viewed cautiously given that this is based on very low counts. Thus, it remains to be further examined whether this overrepresentation overrides the very low mention frequency and thus whether it would be beneficial for automatic GO-annotation systems to focus on these evidential sentences.

ACKNOWLEDGEMENTS

We thank William Baumgartner for programmatically importing GO-annotation information into the Protégé-Knowtator project of curated GO annotations. We also gratefully acknowledge support from NIH 5R01 LM008111, 2R01 LM009254, and 5T15 LM009451

REFERENCES

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W.A, Cohen, K.B., Verspoor, K., Blake, J.A., and Hunter, L.E. (2012) Concept annotation in the CRAFT Corpus. *BMC Bioinform*, **13**:161.

Bada, M., Eckert, M., Palmer, M., and Hunter, L.E. (2010) An overview of the CRAFT concept annotation guidelines. *Proc 4th Ling Annot Wkshp*, 207-211.

Camon, E., Barrell, D.G., Dimmer, E.C., Lee, V., Magrane, M., Maslen, J., Binns, D., and Apweiler, R. (2005) An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinform*, **6**(Suppl 1):S17.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucl Acids Res*, **32** (Database Issue), D262-D266.

Drabkin, H. and Blake, J.A. (2012) Manual Gene Ontology annotation workflow at the Mouse Genome Informatics Database. *Database*, **2012**.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nat Genet*, **25**, 25-29.

Gennari, J.H., Musen, M.A., Fergerson, R.W., Grosso, W.E., Crubézy, M., Eriksson, H., Noy, N.F., and Tu, S.W. (2003) The Evolution of Protégé: An Environment for Knowledge-Based Systems Development. *Internat J Human-Comp Studies*, **58**(1), 89-123.

Lee, V., Camon, E., Dimmer, E., Barrell, D., and Apweiler, R. (2005) Who Tangos with GOA? – Use of Gene Ontology Annotation (GOA) for Biological Interpretation of '-omics' Data and for Validation of Automatic Annotation Tools. *In Silico Biol*, **5**, 5-8.

McCray, A., Browne, A.C., and Bodenreider, O. (2002) The Lexical Properties of the Gene Ontology (GO) *Proc Am Med Inform Assoc Symp.*

Natale, D.A., Arighi, C.N., Barker, W.C., Blake, J.A., Bult, C.J., Caudy, M., Drabkin, H.J., D'Eustachio, P.D., Evsikov, A.V., Huang, H., Nchoutcmboube, J., Roberts, N.V., Smith, B., Zhang, J., and Wu, C.H. (2011) The Protein Ontology: a structured representation of protein forms and complexes. *Nucl Acids Res*, **39**(Database Issue), D539-D545.

Ogren, P.V. (2006) Knowtator: A plug-in for creating training and evaluation data sets for Biomedical Natural Language systems. *Proc 9th Internat Protégé Conf.*

Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D.J., Madden, T.L, Maglott, D.R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Souvarov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Yaschenko, E., and Ye, J. (2009) Database resources of the National Center for Biotechnology Information. *Nucl Acids Res*, **37** (Database Issue):D5-D15.

Smith, C.L. and Eppig, J.T. (2010) The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdisp Rev Syst Biol Med*, **1**(3), 390-399.

Verspoor, K., Cohen, K.B., Lanfranchi, A., Warner, C., Johnson, H.L., Roeder, C., Choi, J.D., Funk, C., Malenkiy, Y., Baumgartner, W.A., Jr., Ogren, P.V., Bada, M., Palmer, M., and Hunter, L.E. (2012) A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinform*, **13**:207.

Winnenburg, R., Wächter, T., Plake, C., Doms, A., and Schroeder, M. (2008) Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Brief Bioinform*, **9**(6), 466-478.