BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 12 2009, pages 1543–1549
doi:10.1093/bioinformatics/btp259

Data and text mining

Application and evaluation of automated semantic annotation of gene expression experiments

Leon French[1], Suzanne Lane[2], Tamryn Law[2], Lydia Xu[2] and Paul Pavlidis[2,3,*]

[1]Bioinformatics Graduate Program, [2]Department of Psychiatry and [3]Centre for High-Throughput Biology, University of British Columbia, Vancouver, BC, Canada

Received on December 16, 2008; revised on March 9, 2009; accepted on April 10, 2009

Advance Access publication April 17, 2009

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Many microarray datasets are available online with formalized standards describing the probe sequences and expression values. Unfortunately, the description, conditions and parameters of the experiments are less commonly formalized and often occur as natural language text. This hinders searching, high-throughput analysis, organization and integration of the datasets.

Results: We use the lexical resources and software tools from the Unified Medical Language System (UMLS) to extract concepts from text. We then link the UMLS concepts to classes in open biomedical ontologies. The result is accessible and clear semantic annotations of gene expression experiments. We applied the method to 595 expression experiments from Gemma, a resource for re-use and meta-analysis of gene expression profiling data. We evaluated and corrected all stages of the annotation process. The majority of missed annotations were due to a lack of cross-references. The most error-prone stage was the extraction of concepts from phrases. Final review of the annotations in context of the experiments revealed 89% precision. A naive system, lacking the phrase to concept corrections is 68% precise. We have integrated this annotation pipeline into Gemma.

Availability: The source code, documentation and Supplementary Materials are available at http://www.chibi.ubc.ca/GEOMMTX. The results of the manual evaluations are provided as Supplementary Material. Both manual and predicted annotations can be viewed and searched via the Gemma website at http://www.chibi.ubc.ca/Gemma. The complete set of predicted annotations is available as a machine readable resource description framework graph.

Contact: paul@chibi.ubc.ca

1 INTRODUCTION

A challenge in the utilization of genomics databases is in the automated retrieval of relevant data. For example, naive approaches to automatically retrieve gene expression studies about 'brain' will fail to find datasets that only mention 'cerebrum' in their descriptions, because free text-based retrieval algorithms are generally unable to make the inference that 'cerebrum' is part of 'brain'. In addition, using free text for information retrieval can produce false positives due to ambiguity and additional false negatives due to synonyms (Bhogal *et al*., 2007). For these reasons, it is valuable to use formal ontologies to describe genomics studies, where inference can be conducted using the structure of the ontology. However, tagging studies with terms from ontologies is currently done by human curators. Such manual curation efforts are costly and often lag behind the generation of new datasets. In this article, we describe efforts to use automated text analysis to assist in the process of accurately tagging genomics studies with terms from ontologies for later retrieval and analysis operations.

Semantically rich annotation is possible in the biomedical domain as there are now available formal ontologies for anatomy, phenotype, environment, cell types and many other areas (Smith *et al*., 2007). Using formal ontologies affords a number of advantages in addition to the information retrieval scenario described above. A concept in an ontology can have extensive semantic information beyond its textual representation, and allows computational integration with other resources that use the same ontologies (Rubin *et al*., 2008).

Of particular interest for biomedical resource annotation are the open biomedical ontologies (OBOs) (Smith *et al*., 2007) and the Unified Medical Language System (UMLS) (Bodenreider, 2004). Previous work on automated genomics experiment annotation has often linked textual names and synonyms to concepts from the UMLS (Butte and Kohane, 2006). However, OBO offers some advantages. First, the UMLS is very large and broad in scope, containing concepts from over 100 source vocabularies (Bodenreider, 2004), which makes it unwieldy. Second in contrast to the UMLS, which requires registration due to license restrictions on the source vocabularies, the OBOs are publicly accessible online (Smith *et al*., 2007). Third, UMLS is complex because its many source vocabularies provide differing relationships between concepts, in contrast to the more orthogonal nature of the OBOs. For example, two UMLS concepts may form a parent–child relationship in one source vocabulary and a sibling relationship in another. In OBO, those two concepts would exist in only one source because OBO enforces orthogonality, providing a single view of their relations. Finally, UMLS lacks tools for programmatically navigating its complex data structures (Srinivasan, 2008). Because we used ontologies defined in the Web Ontology Language (OWL) and represent our results in resource description framework (RDF), we were able to use general purpose semantic web tools. Importantly, mappings exist between many UMLS and OBO concepts.

*To whom correspondence should be addressed.

Here, we focus our attention on annotation of data in Gemma, a resource for re-use and meta-analysis of gene expression profiling data (Hamer,K. *et al*., submitted for publication, http://www.chibi.ubc.ca/Gemma). The majority of Gemma's datasets is downloaded from the Gene Expression Omnibus (GEO), which provides primarily free text descriptions of data and limited use of controlled vocabularies (Barrett *et al*., 2007). Over 500 gene expression experiments from mouse, human and rat have been manually annotated with ontology terms in Gemma, providing a useful resource for evaluating automated methods. The annotations are linked to the experiments using categories from the MGED Ontology (Whetzel *et al*., 2006). Furthermore, Gemma's search employs query expansion using ontology reasoners. For example, a search for 'brain' will return all experiments annotated with any part of the brain. While the ultimate goal of our annotation efforts is to formally describe every sample in Gemma, our current work focuses on providing useful 'high-level' descriptions of experiments. A typical use case is the retrieval of all cancer-related studies. The fact that this might retrieve studies that contain cancer as well as non-cancer samples is acceptable in this scenario. Therefore, we aim to automatically link experiments to concepts that identify the treatments, conditions or locations. A similar task is extraction of diagnosis from clinical documents and medical records or Gene Ontology terms from literature (Spasic *et al*., 2005; Zeng *et al*., 2006).

We evaluated a simple approach for automatic annotation of gene expression experiments using standard OBOs. We used natural language processing to link phrases to biomedical concepts from a large lexicon and then map them to OBOs. Importantly, we extensively evaluated the method, and show that it yields very high quality annotations with few false positives. Although we designed our system with the Gemma system in mind, the approach is general and should be applicable to other databases.

2 METHODS

2.1 MetaMap Transfer

To map text to concepts in UMLS, we used the MetaMap Transfer software (MMTx version) developed at the National Library of Medicine (http://mmtx.nlm.nih.gov/). MMTx retrieves UMLS concepts from input free text by matching terms derived from the concepts. Natural language processing is used to parse text into sentences, phrases, tokens and parts of speech (Smith *et al*., 2004). Phrases are the result of MMTx parsing the sentences for preposition phrases and noun phrases. These phrases are matched against the terms (textual realizations of a concept). The terms for a concept come from its main name, synonyms and abbreviations from UMLS. The terms are then expanded to produce spelling, derivational and inflectional variants. This results in a vast number of terms, primarily due to the size and sources of UMLS. For each term and phrase, MMTx scores the pair using measures of coverage, cohesiveness, centrality and variation (Aronson, 2001; Aronson, 2006). The final result is several scored concepts for each phrase. For a detailed example of how the system uses MMTx, view the Supplementary website.

We used MMTx to perform the main tasks of natural language processing and mapping to UMLS concepts. We employed version 2.4C using the default strict database derived from UMLS version 2006AA as provided by the MMTx website. Our initialization options for MMTx are 'an_derivational_variants' to allow adjective–noun derivational variation and 'no_acros_abbrs' to limit the use of acronym/abbreviation variants. For each phrase in the input text, we keep all final mappings with an MMTx score >850. Although the MMTx score ranges from 0 to 1000, we set a high limit to reduce processing time and the amount of evaluations. Although we record the UMLS string and concept identifiers, we only store those that have a mapping to one of the three ontologies described below.

2.2 Ontologies

For this study, we limited our analysis to three ontologies to represent concepts from the domains of neuroscience, anatomy and diseases: BIRNLex (Bug *et al*., 2008), Foundational Model of Anatomy (FMA; Rosse and Mejino, 2003) and Disease Ontology (DO) (http://diseaseontology.sourceforge.net/). Concept identifiers or codes from the original source are cited during curation into UMLS. Conversely, some ontologies are based on UMLS concepts and reference UMLS directly. Many other ontologies or terminologies have these references and could be used in our system such as the Gene Ontology (Ashburner *et al*., 2000), Medical Subject Headings (MeSHs), the NCI Thesaurus (Sioutos *et al*., 2007) and the NCBI taxonomy (Wheeler *et al*., 2001). Each is a UMLS source, allowing direct linking from MMTx results. While there are ontologies created specifically for the annotation of gene expression data (Kelso *et al*., 2003), we choose larger general purpose ontologies with UMLS cross-references, allowing us to leverage the UMLS software systems.

FMA is recognized as a high quality ontology and listed as a mature ontology in the OBO Foundry (Smith *et al*., 2007). We used the 'Lite version' which only contains the relationships for 'is_a', 'part_of' and 'has_part' (Rosse and Mejino, 2003). We converted 61 370 Digital Anatomist ID's provided by UMLS into URIs (uniform resource identifiers) referring to the same concepts in the FMA lite ontology.

The DO is primarily designed for medical coding purposes and is based primarily on the UMLS. Like FMA, it is considered a mature OBO Foundry ontology. The cross-references were extracted by retrieving all classes that had 'hasDbXref' (http://www.geneontology.org/formats/oboInOwl#) property to UMLS concept identifiers (CUI) prefixed with 'UMLS_CUI:'. The result is 17 776 UMLS concept references.

BIRNLex is an ontology-based lexicon developed to support the Biomedical Informatics Research Network (Bug *et al*., 2008). With a focus on neurodegenerative disease, it provides extensive concepts pertaining to sensation, behavior, cognition and neuroanatomy. It follows the OBO Foundry guidelines and was chosen to enhance Gemma's utility as a neuroinformatics resource. Four hundred and sixty-nine cross-references were extracted by retrieving all classes that had the 'UmlsCui' (http://purl.org/nbirn/birnlex/ontology/annotation/OBO_annotation_properties.owl#) property, which point to UMLS CUIs.

2.3 Gemma

For each microarray experiment, we used several sources of free text as input. At the top level, we used the main title and description. In some cases, the experiments are linked to journal articles, for which we processed the title and abstract. For each RNA sample, we processed its name and description. To reduce computation time we employed a memory cache that recalled the concepts from previously seen text fragments.

We used experiments performed on rat, mouse and human samples. Although the ontologies we used are not organism independent, we found that they are useful for annotation of the three mammalian experiment types.

For evaluation we extracted annotations previously applied by the Gemma curators before our method was applied. Although the annotations occur as both free text and concepts from several ontologies, we selected only the annotations that were present in one of the aforementioned ontologies.

2.4 OWL/RDF data

We used the Jena semantic web API throughout the project (http://jena.sourceforge.net/). Queries were written in the SPARQL language and executed using ARQ query engine. Tabulator (http://www.w3.org/2005/ajar/tab) and IsaViz (http://www.w3.org/2001/11/IsaViz/) tools were used to

visualize the generated RDF data. OWL versions of FMA Lite and DO were downloaded from the OBO Download matrix (http://www.berkeleybop.org/ontologies/).

2.5 Evaluation

We evaluated the process of converting free text to UMLS concepts, and separately the mapping of concepts to URIs. Additionally, we evaluated whether the final extracted term was appropriate for the original Gemma experiment. Each evaluation was performed by two human curators (SL, TL or LX) and final agreement was achieved through review.

To examine the extraction from phrase to CUI, we reviewed a list of phrases and their mappings. For each phrase and concept extracted by MMTx a UMLS string identifier is provided (SUI). The SUI allows filtering of specific synonyms linked to a concept. Because many phrases can result in the same SUI and CUI combination, we reviewed all phrases for each. If one phrase was deemed to be a false positive for that concept, then that CUI+SUI combination was rejected. Several guidelines were created for this evaluation:

(1) we rejected abbreviations unless the whole term contained a word (e.g. 'CA1 region').

(2) In some cases, the concept name does not fit the phrase—for example, 'gland' → 'Gland Structure'. In these cases we referred to UMLS for the definition of the concept.

(3) We accepted mappings to concepts even if they were considered uninformative or nonspecific, for example, 'Branch' or 'Genes'. Note that some of these terms are filtered out at later stages.

(4) We rejected general to specific mappings and accepted the reverse. For example, we would reject the mapping of the free text 'deficiency' to the UMLS concept of 'Malnutrition', but would accept the mapping of the free text 'malnutrition' to the concept 'Deficiency'.

For a given phrase MMTx provides many concepts, some of which are outside our scope. For our evaluation we only reviewed phrases and concepts that had a cross-reference to a FMA, DO or BIRNLex URI. We comprehensively reviewed this entire set of 7449 phrase-to-concept mappings for errors.

To validate that the mapping from UMLS concept to URI was correct, we manually reviewed 387 cross-references. This evaluation was done on only a subset of the data as the cross-references were expected to be more accurate than text processing, since they were created by expert curators (i.e. they are provided as part of the UMLS system or the ontologies). Specifically, we assumed as correct 609 cross-references where the UMLS concept name and the URI label matched (ignoring case).

Although the resolved concept might fit the phrase it was found in, it may not describe the experiment. A phrase extracted and successfully mapped to a UMLS concept may have a different meaning in the context of the experiment. For example, a study abstract that mentioned 'malnutrition' may have done so in passing, not in the context of describing the study itself. We evaluated the appropriateness of the automatically generated annotations in two ways. First, we compared predicted annotations to annotations drawn from the same ontologies that were previously added to Gemma by curators. Exact comparison was performed by matching the URIs. However, because the number of manual annotations was limited, this could only provide an accurate measure of false negatives and a lower bound of true positive predictions. To get a second measure of accuracy, we manually reviewed the predicted annotations for a random sample of 100 experiments, excluding concepts that were already manually annotated. Two curators manually reviewed each of the 100 experiments and marked each predicted annotation as correct or incorrect.

The guidelines for the manual evaluation of annotation quality performed on the 100 experiments focused on information retrieval. We accepted concepts that described an experiment even if a more precise concept was appropriate, for example we would accept a cancer annotation of a lung cancer dataset. After review, we decided to accept an annotation of concept C if, in the judgment of the evaluator, a researcher searching for datasets pertaining to C would want to retrieve the experiment. We rejected concepts that described the experimental method or technique used (e.g. 'Decapitation' in the preparation of mouse brain tissue).

3 RESULTS

Our approach uses the entire UMLS for the text mining stage, and then translates the UMLS concepts to three domain-specific ontologies (Fig. 1). By narrowing down to the specific ontologies, we reduce the number of concepts to the domains of neuroscience, anatomy and diseases. Using the open ontologies follows the efforts of other neuroinformatics resources that have provided data in semantic web formats (French and Pavlidis, 2007; Ruttenberg *et al*., 2007).

The results of running our mapping procedure are outlined in Tables 1 and 3. For all Gemma experiments, 58 030 text to URI mappings were found. Further processing reduced the predictions to 2740 URI to experiment pairings between the 782 URIs and 595 experiments. The predicted annotations are provided as a RDF graph.

3.1 Text mining evaluation

We manually evaluated all 7449 phrase-to-CUI MMTx mappings, yielding a rejection rate of 17%. Inter-evaluator agreement was 91%. We manually reviewed the conflicts for full agreement, resulting in 246 rejected SUI to CUI pairings. The main reasons for rejection were ambiguous terms and general to specific mappings.

As a result of the CUI to URI evaluation, we rejected only four pairings that corresponded to the UMLS to FMA pairings of Neurotransmitters→Biogenic amine; Lamina→Subintima; Cephalic→Rostral; and Branch→Leaf of cardiac valve. A further nine UMLS mappings pointed to non-existent FMA URI's and were removed.

To remove uninformative concepts from the results, we manually selected concepts from a list of most frequently found annotations. In addition, we designated concepts to be uninformative during the manual evaluation of annotation quality. Examples from the full list of 19 include 'Genes', 'RNA' and 'Cells'. The complete list is available as Supplementary Material.

3.2 Extraction

We ran our program on all 595 gene expression experiments stored in the Gemma system. It extracted 801 unique concepts from the text sources. Processing time averaged 46 s per experiment on a dual core 2.6 GHz processor. Most of this time was attributable to computations done by MMTx. Before filtering rejected mappings, MMTx found 58 030 mentions of concepts that had mappings to one or more of the three ontologies. Filtering for rejected and uninformative mappings reduced this amount to 26 525 mentions (Table 1).

Predicted annotations could come from multiple free text sources associated with the experiment. At the top level, 3075 mentions (12%) were extracted from the main title and description. The titles and abstracts of linked journal abstracts revealed 2384 mentions (9%). The names and descriptions of experimental samples (representing a single microarray run) resulted in 21 066

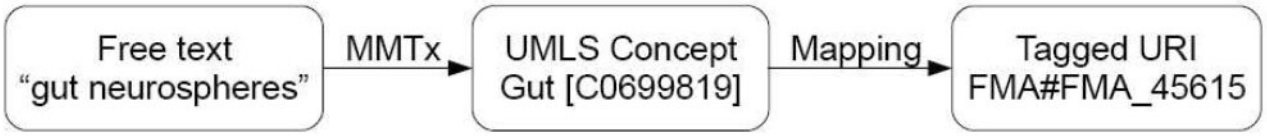

Fig. 1. Outline of the methods. The procedure starts with free text associated with a genomics study. The text is converted to UMLS concepts then mapped to a FMA ontology term. The ontology term can then be associated with the genomics study.

Table 1. Number and accuracy of mentions before and after filtering steps

Stage	Mentions	Annotations	Recall	Precision (min)
Unfiltered	58 030	5484	0.497	0.094
Filtered for rejected SUI + CUI, and CUI → URI pairs	39 155	3985	0.488	0.128
Filtered for uninformative concepts	26 525	2740	0.488	0.185

mentions (79%). The concepts found across these sources are not unique and many duplicates exist within and across the sources. These repetitions are demonstrated by our final list of 26 525 mentions that reduces to 2740 unique experiment–concept pairs.

Table 2 displays the predicted annotation concepts and their frequencies. The average number of predicted annotations per experiment was 4.6. Thirty-five experiments had no predicted annotations. At the maximum, an experiment surveying tissue-specific expression in mouse had 60 predicted annotations (48 are from the FMA).

3.3 Evaluation of annotation relevance

Because our goal is to generate meaningful annotations of expression studies, we performed an evaluation of the relevance of predicted annotations to the target experiments. As described in the methods, we did this in two ways: by comparing predicted annotations to manually generated ones, and by manually evaluating the quality of predicted annotations. These evaluations are covered in the next two sections.

3.4 Comparison to manual annotations

Due to the cost of manual annotation, curation in Gemma is incomplete. In other words, we expect that the automated procedure will generate 'correct' annotations which are, strictly speaking, false positives when compared with the existing manual annotations. Indeed, the manual annotation yielded only 1.8 concepts per experiment, while our method produced 4.6. Thus, the comparison to manual annotations provides an estimate of recall but only a lower bound on precision. The system automatically recalled approximately half of the 1042 existing manual annotations in Gemma. For 213 experiments, our method perfectly recalled all 298 of the existing annotations.

Table 3 shows how performance varied across the three ontologies. Predictions of FMA concepts were the most numerous and precise but recall was relatively low. BIRNLex annotations were low in number owing to its small scope and limited UMLS mappings. Manual inspection of the disease predictions revealed many related predictions for a single disorder, possibly explaining the low precision. An example is the experiment 'Cytotoxic activity of HTI-286 in prostate cancer' (GSE8325). Predicted annotations for this study were 'Malignant neoplasm of prostate', 'Malignant Neoplasms', 'Refractory Carcinoma' and finally 'Prostate carcinoma', which was the only annotation chosen

Table 2. Top 40 concepts mapped to experiments

Concept name	Count
Brain	119
Cerebral cortex	61
Spinal cord	56
Malignant neoplasms	55
cancer	49
Hippocampus	46
Spleen	42
Stem cell	35
Cerebellum	34
Heart	32
Liver	31
Muscle tissue	30
Kidney	28
Pair of lungs	27
Infection	25
Communicable diseases	24
Nervous system	21
Skeletal muscle tissue	21
Breast	21
Epithelial cell	19
Blood	18
Hypothalamus	17
Neurodegenerative disorders	17
Chromosome	16
Retina	16
Carcinoma	16
Prostate	16
Neoplasm metastasis	15
Frontal lobe	15
Bone marrow	15
Malignant neoplasm of breast	15
Breast carcinoma	15
Amygdala	14
Colon	14
Alzheimer's disease	13
Neuraxis	13
Mammary neoplasms	12
Primary tumor	12
Fibroblast	12
Epithelium	11

Table 3. Comparison to manual annotations, divided by ontology

Name	Existing	Predicted	Intersection	Recall	Precision (min)
FMA	682	1351	304	0.446	0.225
DO	217	1041	127	0.585	0.122
BIRNLex	143	348	77	0.538	0.221
All	1042	2740	508	0.488	0.185

Table 4. Recall of annotations with cross-references

Name	Existing annotations with mappings	Predicted annotations	Recall
FMA	404	1351	0.752
DO	217	1041	0.585
BIRNLex	100	348	0.770
All	721	2740	0.705

in manual curation. In this case, all the predicted annotations fit the experiment and provide more details except 'Malignant neoplasm of prostate', which can be inferred from its DO child term of 'Prostate carcinoma'. We found that over one-third of the DO predictions can be inferred in this way. None of the BIRNLex and 5.3% of the FMA predictions fit this case. Furthermore, expanding the DO predictions to all parent terms increases recall to 66.8%. For BIRNLex and FMA, the same procedures produce little or no increase in recall.

We tested if the coverage of the cross-references affected recall of the existing annotations. Using the mappings, we discovered 8 of the 34 BIRNLex URLs and 72 of the 184 FMA URIs used by the annotators did not have a mapping from a UMLS concept. All DO URIs had a UMLS mapping (the ontology is directly based on the UMLS). We removed the 43 BIRNLex and 278 FMA annotations that used the non-mapped URIs. By excluding these annotations, the recall increased by 30.7% and 23.2% for BIRNLex and FMA, respectively (Table 4). For BIRNLex and FMA, the lack of the cross-references explains the majority of annotations our method failed to recall.

3.5 Manual evaluation of annotation quality

In contrast to the comparison of manual annotations, our evaluation of the appropriateness of the annotations was expected to give a better estimate of precision (how many of the annotations are correct), but not of recall (the possibility of 'missing' annotations was not considered in this phase).

An initial review yielded an interannotator agreement of 71.2%. The annotators both rejected 10.5% of the annotations and accepted 60.7%. A review of the 117 annotations that were disagreed upon revealed ambiguity in the review guidelines provided to the annotators. We therefore extended the guidelines and allowed the curators to re-review the conflicts. The precision of predicted annotations did not differ between the 71.2% agreed annotations in the first evaluation and the complete set of decisions determined by this final evaluation.

As shown in Table 5, the evaluation of the 100 randomly chosen experiments indicate the software that yields high-precision annotations. Overall precision was 89% but varied among the ontologies. All 55 BIRNLex annotations were deemed correct, compared to 84% of the FMA annotations. During this evaluation we added nine terms to the list of uninformative concepts. In only one case did a negation cause an error: 'Non-Hodgkin lymphoma' was annotated as 'Hodgkin's lymphoma'. Most rejections were due to the annotation being only tangentially related to the experiment. For example, many were extracted from background material and findings introduced in the experiment description or associated article abstract. Thus while the annotation might correctly reflect something mentioned in the abstract, it was not a good annotation for the expression study.

Table 5. Manual evaluation of annotation quality

Name	Predicted	Accepted	Precision
FMA	213	179	0.840
DO	195	176	0.903
BIRNLex	55	55	1.000
All	463	410	0.886

We removed annotations derived from specific single text sources (e.g. experimental description versus abstract of the published article) to determine the effect they had on precision. It was found that the annotations extracted only from the abstract of the literature reference were the least precise. By removing these annotations, precision on the set of 100 increased to 90.8%, while recall decreased by 0.019 for all experiments. If all annotations found only in a single text source are removed, precision increases to 93% but recall on all experiments decreases to 30% from 49%.

4 DISCUSSION

We have presented a simple method for automatically converting text associated with gene expression studies into terms from formal ontologies. One of our contributions is a thorough evaluation of the results, which, in addition to providing useful performance measures, provides insights into the limitations of the approach and how it could be improved. While we focused on gene expression studies, the approach is quite general and could be applied to other types of data where free text descriptions are available.

It is difficult to compare our method to existing techniques as similar systems are tailored to specific databases and ontologies. Perhaps, the closest is GenoText, which mined each GEO expression experiment for UMLS concepts using MMTx (Butte and Kohane, 2006). GenoText extracted 4127 concepts from 448 GEO datasets, resulting in approximately twice as many concepts per experiment than our method. Although Butte and Kohane (2006) describe concept mapping errors, they did not provide accuracy rates. GenoText does not include the step of mapping to formal ontologies. Presumably for this reason, GenoText's results include many terms that are of limited use in information retrieval, such as 'total'. Similar automated mapping for cancer terms has been demonstrated on the Stanford Tissue Microarray database (Shah *et al*., 2007). By manually evaluating a sample of annotations, they estimated 86% precision and 95% recall. Another system, 'Whatizit' provides a web-based system for annotating free text

using several term sources, including disease terms (Rebholz-Schuhmann *et al*., 2008). Like us they employed MMTx to extract disease mentions, and found that MMTx performs at par to other techniques (Jimeno *et al*., 2008). A fifth related project is the Open Biomedical Annotator (OBA), from the National Center for Biomedical Ontology. It links OBO ontology concepts to resources from several biomedical resources, including the GEO (Jonquet *et al*., 2008). Although it spans many ontologies and databases, OBA does not provide the extensive evaluation that we present here. In general, formal evaluation of text-to-concept mapping methods is limited in the bioinformatics literature. A notable exception is BioCreative (Krallinger *et al*., 2008), a coordinated evaluation effort focused on methods for extracting information about genes. We believe that most of the BioCreative tasks are much more difficult than the one we pose.

The most important measure of the utility of our method is given by our manual evaluation of 100 datasets. At this final stage, we reached precision of 89% and in the case of the BIRNLex ontology all 55 predicted annotations were deemed correct. Thus, while errors are apparent, the large number of annotations generated and their general high quality suggests that the method can be used as it is. Indeed, we have integrated the method into Gemma and are including the annotations in the database. To differentiate annotations generated with our software from manually generated tags, we flag the annotations with the 'IEA' evidence code (Inferred by Electronic Annotation) taken from the Gene Ontology (Ashburner *et al*., 2000).

We found that the first step, from phrase to UMLS concept, was the most error prone. With 17% of the mappings rejected, it reduced the total number of mentions by 12.1%. If these mentions were carried into the final evaluation of 100 experiments, precision would drop to 68%. Since this evaluation can now be used to filter annotations in the final annotation pipeline, we believe that annotation accuracy of new experiments will be closer to 89%. Our results with the MMTx step are consistent with a past evaluation (Jimeno *et al*., 2008). In future experiments, we will attempt to generalize the specific rejections using different MMTx parameters and output.

At the UMLS concept to Ontology URI stage, we found the mappings to be accurate with only four rejected cross-references. Unfortunately, not all the concepts in the ontologies had links from UMLS. Removing the annotations with missing links increased recall to 71.3%. This suggests more complete UMLS mappings for FMA and BIRNLex will increase recall significantly. Indeed, preliminary experiments we have conducted with more recent mappings yielded an improvement in recall from 0.45 to 0.69. By expanding annotations using the semantic information provided by the ontologies, we were able to increase the recall of DO annotations to 66.8% from 59%.

In addition to providing a new source of annotations, our method uncovered some annotation errors in Gemma. In one case, the automated method predicted the annotation 'Macaque' for an experiment that had been annotated as being a study of human tissues. Upon investigation, we discovered the experiment assayed rhesus macaque tissue on a human genome-based microarray platform. This combination had escaped the notice of the Gemma curators, highlighting the utility of automated annotations in assisting human curators.

We designed our system to be general purpose, allowing addition of other ontologies and data sources such as the Gene Ontology (Ashburner *et al*., 2000), MeSH, NCI Thesaurus (Sioutos *et al*., 2007) and NCBI taxonomy (Wheeler *et al*., 2001). The annotation sources might also be extended to include full text articles, though our finding that abstracts produced the least precise annotations indicates that full text sources would require a more complex text analysis system than the one we propose. By using the relatively terse (but still free text) descriptions that are typically associated with genomics studies, we avoid challenges related to the size and complexity of other sources.

Although our framework provided good results, several limitations should be pointed out. One is the need for a cross-reference between UMLS and a formal ontology. This requirement allows leveraging of UMLS but can limit the concepts that our framework can extract. On the other hand, this also helps ensure that terms are relevant to the domain of interest. We also did not evaluate the utility of the ontology-based annotations compared with free text-based searches from the end-user standpoint. One expected advantage is the ability to leverage ontology structure. Anecdotally, this can be effective. For example, as mentioned in Section 1, searches for 'brain' can retrieve studies that are only associated with the term 'cerebrum' because Gemma can infer the relationship between the terms using simple reasoning on the FMA. Formally evaluating the quality of search results is a potential subject for future research.

In addition to their role in information retrieval, high-quality annotations can improve analysis of gene expression experiments. For example, many groups have applied biclustering methods to large sets of gene expression datasets, revealing smaller subsets of genes and experiments that show common patterns of expression (Prelic *et al*., 2006). These algorithms are primarily evaluated in the gene dimension as common themes can be found by using Gene Ontology enrichment testing. Currently, finding common themes in the experiment or sample dimension can only be revealed by manual review. Experiment annotations from formalized ontologies will allow more meaningful evaluations of these and other types of studies.

ACKNOWLEDGEMENTS

L.F. thanks Ben Good and Mark Wilkinson who provided introduction to biomedical ontologies and the life sciences semantic web.

Funding: Natural Sciences and Engineering Research Council of Canada (to L.F.); a career award from the Michael Smith Foundation for Health Research, a Canadian Institutes of Health Research (CIHR) New Investigator award and a Human Brain Project grant from the National Institutes of Health (GM076990) (to P.P.).

Conflict of Interest: none declared.

REFERENCES

Aronson,A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Annual Symposium.* pp. 17–21.

Aronson,A.R. (2006) MetaMap: mapping text to the UMLS Metathesaurus. Available at http://skr.nlm.nih.gov/papers/references/metamap06.pdf (last accessed date April 27, 2009)

Ashburner,M. *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Barrett,T. *et al*. (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.

Bhogal,J. *et al*. (2007) A review of ontology based query expansion. *Inform. Process. Manag.*, **43**, 866–886.

Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, D267–D270.

Bug,W.J. *et al*. (2008) The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics*, **6**, 175–194.

Butte,A.J. and Kohane,I.S. (2006) Creation and implications of a phenome-genome network. *Nat. Biotechnol.*, **24**, 55–62.

French,L. and Pavlidis,P. (2007) Informatics in neuroscience. *Brief. Bioinformatics*, **8**, 446–456.

Jimeno,A. *et al*. (2008) Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, **9 (Suppl. 3)**, S3.

Jonquet,C. *et al*. (2008) Help will be provided for this task: ontology-based annotator web service. *Technical Report*. (last accesses date on April 27, 2009).

Kelso,J. *et al*. (2003) eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res.*, **13**, 1222–1230.

Krallinger,M. *et al*. (2008) Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biol.*, **9** (Suppl. 2), S1.

Prelic,A. *et al*. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122–1129.

Rebholz-Schuhmann,D. *et al*. (2008) Text processing through web services: calling whatizit. *Bioinformatics*, **24**, 296–298.

Rosse,C. and Mejino,J.L. Jr (2003) A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *J. Biomed. Inform.*, **36**, 478–500.

Rubin,D.L. *et al*. (2008) Biomedical ontologies: a functional perspective. *Brief. Bioinformatics*, **9**, 75–90.

Ruttenberg,A. *et al*. (2007) Advancing translational research with the semantic web. *BMC Bioinformatics*, **8 (Suppl. 3)**, S2.

Shah,N.H. *et al*. (2007) Annotation and query of tissue microarray data using the NCI Thesaurus. *BMC Bioinformatics*, **8**, 296.

Sioutos,N. *et al*. (2007) NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Inform.*, **40**, 30–43.

Smith,B. *et al*. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.

Smith,L. *et al*. (2004) MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics*, **20**, 2320–2321.

Spasic,I. *et al*. (2005) Text mining and ontologies in biomedicine: making sense of raw text. *Brief. Bioinform.*, **6**, 239–251.

Srinivasan,S. (2008) Re: MetamorphoSys Tool. UMLS users discussion list. NIH Listserv. https://list.nih.gov/cgi-bin/wa?A2=ind0808&L=umlsusers-l&T=0&P=3026 (last accessed date April 27, 2009).

Wheeler,D.L. *et al*. (2001) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **29**, 11–16.

Whetzel,P.L. *et al*. (2006) The MGED Ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics*, **22**, 866–873.

Zeng,Q.T. *et al*. (2006) Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med. Inform. Decis. Mak.*, **6**, 30.