Journal of Biomedical Informatics 44 (2011) S31–S38

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Enabling enrichment analysis with the Human Disease Ontology

Paea LePendu *, Mark A. Musen, Nigam H. Shah

Stanford Center for Biomedical Informatics Research, 251 Campus Drive, Medical School Office Building, Room X215, Mail Code 5479, Stanford University, Stanford, CA 94305-5479, USA

ARTICLE INFO

Article history:
Available online 29 April 2011

Keywords:
Enrichment analysis
Human disease
Ontology
Annotation
Information integration
Electronic Health Records

ABSTRACT

Advanced statistical methods used to analyze high-throughput data such as gene-expression assays result in long lists of "significant genes." One way to gain insight into the significance of altered expression levels is to determine whether Gene Ontology (GO) terms associated with a particular biological process, molecular function, or cellular component are over- or under-represented in the set of genes deemed significant. This process, referred to as enrichment analysis, profiles a gene set, and is widely used to make sense of the results of high-throughput experiments. Our goal is to develop and apply general enrichment analysis methods to profile other sets of interest, such as patient cohorts from the electronic medical record, using a variety of ontologies including SNOMED CT, MedDRA, RxNorm, and others.

Although it is possible to perform enrichment analysis using ontologies other than the GO, a key prerequisite is the availability of a background set of annotations to enable the enrichment calculation. In the case of the GO, this background set is provided by the Gene Ontology Annotations. In the current work, we describe: (i) a general method that uses hand-curated GO annotations as a starting point for creating background datasets for enrichment analysis using other ontologies; and (ii) a gene-disease background annotation set – that enables disease-based enrichment – to demonstrate feasibility of our method.

1. Introduction

One way to gain insight into the significance of a particular set of genes is to determine whether functional terms that are associated with each gene are over- or under-represented in the set of genes deemed significant. This process, referred to as enrichment analysis, profiles a gene set, and is widely used to make sense of the results of high-throughput experiments such as gene-expression assays. The canonical example of enrichment analysis is in the interpretation of a list of differentially expressed genes in some condition. The usual approach is to perform enrichment analysis with the Gene Ontology (GO). We can aggregate the annotating GO concepts associated with a particular biological process, molecular function, or cellular component for each gene in this list, and arrive at a profile of the biological processes or mechanisms affected by the condition under study [1]. There are currently over 400 publications on methods and tools for GO-based enrichment, but (to the best of our knowledge) only a single other tool, *Genes2-Mesh*, uses something besides the GO (i.e., the Medical Subject Headings or MeSH), to calculate enrichment [2]. Our goal is to develop and apply general enrichment analysis methods to profile other sets of interest, such as patient cohorts from the electronic medical record, using a variety of ontologies including SNOMED CT, MedDRA, RxNorm, and others.

While the GO has been the principal target for enrichment analysis, we can carry out the same sort of profiling using *any* ontology available in the biomedical domain. Tirrell et al. have developed a prototype tool [3] called RANSUM – Rich Annotation Summarizer – that performs generalized enrichment analysis using any ontology from the National Center for Biomedical Ontology's (NCBO) online repository of public ontologies called BioPortal [4].

By using a disease ontology in such analysis, we can enable translational questions: just as scientists can ask which *biological process* is over-represented in a set of differentially expressed genes, they can also ask which *disease* (or class of diseases) is over-represented in a set of genes or proteins that share a common characteristic. For example, by annotating known protein mutations with disease terms, Mort et al. identified a class of diseases—blood coagulation disorders—that are associated with a significant depletion in substitutions at O-linked glycosylation sites [5]. Similarly, by identifying other disease associations for the genes involved in a certain disease of interest we can gain insight into how the causation of seemingly unrelated diseases might be related, e.g., *Werner's syndrome*, *Cockayne syndrome*, *Burkitt's lymphoma*, and *Rothmund–Thomson Syndrome* [6–9]. We can also apply the enrichment analysis methodology to other sets of interest—such as patient cohorts. For example, enrichment analysis might detect specific co-morbidities that have an increased

* Corresponding author. Fax: +1 650 725 7944.
E-mail address: plependu@stanford.edu (P. LePendu).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.04.007

incidence in rheumatoid arthritis patients—a topic of recent discussion in the literature and considered essential to provide high quality care [10–12]. Enrichment analysis to identify common pairs of terms of different semantic types can identify combinations of drug classes and co-morbidities, or test risk-factors and co-morbidities that are common in this population; in fact Petri et al. recently identified co-morbidities in *rheumatoid arthritis* patients using relative risk analysis (which shares similarities with enrichment analysis) calculated from ICD9 codes in a retrospective cohort study using medical claims data [13].

Note that enrichment analysis as discussed in this paper and as performed by the majority of the tools listed online[1] by the GO Consortium is conceptually different from the similarly named Gene Set Enrichment Analysis (GSEA) method [20], where groups of genes that are known to share common biological function, chromosomal location, or regulation are tested collectively for significant difference in expression between two phenotypic conditions such as tumors that are sensitive versus resistant to a drug. The goal of GSEA is to determine whether members of a gene set S—as defined by common biological function, chromosomal location, or regulation—tend to occur toward the top (or bottom) of the list L (comprised of genes showing the largest difference in expression between the two phenotypic classes), in which case the gene set is deemed to be correlated with the phenotypic condition under study.

One key aspect of calculating functional enrichment (such as GO term enrichment) is the choice of a reference-term frequency since the calculation compares the term frequencies in the annotations of a set of interest against the annotations of a reference set. It is not clear what the appropriate reference-term frequency should be when calculating enrichment of ontology terms for which a "background set" is not defined. For example, in the case of Gene Ontology annotations, the background set is usually the GO annotations of the set of genes on which the data were collected on a microarray or the GO annotations of all the genes known in the genome for the species on which the data were collected. A natural background set is not available, however, when calculating enrichment using disease ontologies because these ontologies have not been used for manual annotation in a way the Gene Ontology has been used.

For situations lacking an obvious background set, there are two main options: As Tirrell et al. note, we can use the frequency of ontology terms in a large corpus, such as the NCBO Resource Index [14,15], MEDLINE abstracts or on Web pages indexed by Internet search engines such as Google. Using such an "off the shelf" reference set has the drawback of not being representative of the specific set of interest being analyzed, for example, in the case of analyzing patient cohorts. One alternative is to construct a reference annotation set using automated methods.

Our approach is to construct a reference set programmatically using manually created GO annotations as a starting point. We specifically choose GO annotations because they provide a reliable foundation—highly trained curators associate GO terms to gene products, based on exhaustive literature review. Building upon this foundation, we demonstrate how, with the availability of tools for automated annotation with terms from disease ontologies, it is possible to create reference annotation sets for enrichment analysis using ontologies other than the GO—for example, the Human Disease Ontology (DO).

Basically, a manually curated GO annotation associates a gene product with a PubMed article with high accuracy. We hypothesize that if a disease term is mentioned in the abstract of the article based on which a GO annotation is created for a gene product, then that disease term is likely to be associated with that gene product; and we can associate relevant disease terms to those gene products by analyzing the text in the title and abstract of the article. Unlike GO terms, which actually appear in the text with low frequency (see Section 4.1), or gene identifiers, which are ambiguous, disease terms are highly amenable to automated, term extraction techniques [16]. Therefore, using tools that recognize mentions of ontology terms in user submitted text such as the NCBO annotator [17], we can automatically recognize occurrences of disease terms from the DO in a given corpus of text; the key is to identify a reliable text source to recognize disease terms from, to associate with genes and gene products.

Therefore, by starting with curated gene associations we can reliably obtain gene-disease associations from biomedical literature. Researchers can then use these associations to automatically generate a gene-disease association file as a background set (or reference set) for disease-specific enrichment analysis. Moreover, researchers can reuse our method to examine annotations along other dimensions. For example, researchers can use the Pathway ontology to generate gene-pathway associations, or fragments of SNOMED CT to generate gene-anatomy associations.

What differentiates our method from other approaches that infer gene-disease associations—such as co-occurrence analysis or syntactic–semantic relationship extraction techniques, which might require difficult to obtain training sets for finding gene-disease associations [18]—is the reuse of publicly available GO annotations as a basis for identifying reliable gene-publication records that serve as the foundation for generating automated annotations. Furthermore, unlike dictionary-based approaches [18], we assign public ontology term identifiers (e.g., DO identifiers or DOIDs) during the annotation process, which can be reasoned over to aggregate, filter, and cross-reference associated disease terms. In a similar approach to ours, Osborne et al. argue that annotating GeneRIF descriptions with DO terms to infer gene-disease relationships offers greater signal-to-noise than mining 20 million MEDLINE articles directly, given the nature of curated GeneRIF descriptions [16]. In the results, we quantify the increased coverage of our approach.

In summary, our main contributions are: (i) a general method, which uses hand-curated GO annotations as a starting point for creating background datasets for enrichment analysis using other ontologies; and (ii) a gene-disease background annotation set—that enables disease-based enrichment analysis—to demonstrate feasibility of our method.

2. Methods

Fig. 1 summarizes our method. First, we start with GO annotations, which provide the PubMed identifiers of papers based on which gene products are associated with GO terms by a curator. The annotations essentially give us a link between gene identifiers and PubMed articles and only those PubMed articles that were deemed to be relevant for the process of creating GO annotations. Next, we recognize terms from an ontology of interest (e.g., DO) in the title and abstracts of those articles. Finally, we associate the recognized ontology terms with the gene identifiers to which the article analyzed was associated.

2.1. Obtaining gene-publication associations

We download GO annotation files[2] for human gene products from geneontology.org. These files are tab-delimited text files that contain, among other things, a list of gene identifiers, associated

[1] http://geneontology.org/GO.tools_by_type.term_enrichment.shtml

[2] http://www.geneontology.org/GO.downloads.annotations.shtml

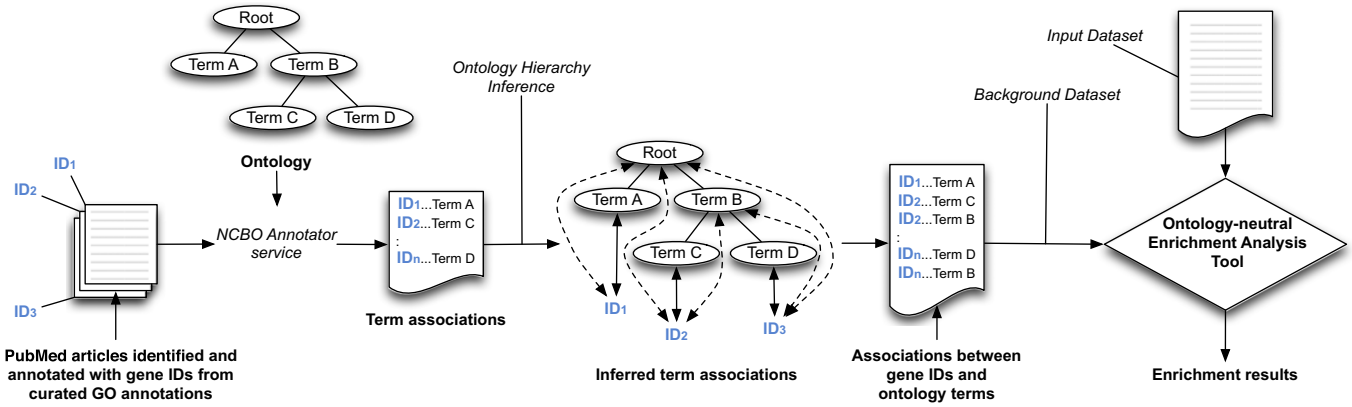

Fig. 1. Workflow for generating background annotation sets for enrichment analysis: First, we start with a corpus of PubMed articles identified in manually curated GO annotations. These curated annotations provide gene-to-article associations. Next, we annotate the titles and abstracts of each article with ontology terms using the NCBO Annotator service. Terms associations can be expanded based on inferred hierarchical relationships. Finally, the gene-to-article associations are linked with the curated article-to-term associations to obtain a list of gene-to-term associations. The resulting term frequencies provide a background set for enrichment analysis.

GO terms, and the publication source (a PubMed identifier) on the basis of which the GO annotation was created.

We remove all electronically inferred annotations (IEA) from the file because they are less reliable. We also remove all qualified annotations, such as negated (NOT) ones. As a result, we obtain a list of publications and the genes they describe, gene-publication tuples. In the next phase, we process the publications to obtain publication–disease tuples.

2.2. Parsing article titles and text

Using the PubMed identifiers obtained from the GO annotation files, we fetch each article's title and abstract using the National Library of Medicine eUtils.[3] We save each article's title and abstract as a file and process it using the disease ontology with the Annotator service.

2.3. Annotating text using terms from the Human Disease Ontology

For each PubMed article, we use the NCBO Annotator Web-service[4] to identify mentions of disease terms from the DO in the text. We configure the service to find the longest, whole-word matches for term labels and synonyms from the DO. Specifically, we set the following Annotator Web-service parameters for our study: *wholeWordOnly* = *true*, *scored* = *true*, *ontologiesToExpand* = 42986, *withDefaultStopWords* = *true*, *levelMax* = 0. The Human Disease Ontology is indicated by the BioPortal version id 42986. Other details on the Web-service parameters are documented in the online user guide.

For every matched term, we acquire the appropriate disease concept identifier resulting in a list of publication-to-disease concept tuples. In other words, the annotation process performs concept normalization. Having the concept identifier, we can invoke hierarchy and mapping Web-services from BioPortal to expand the associations to include parent or related terms, if desired. Multiple matches of a term within the text, and relative word ordering are not taken into account. Also, for this study, *levelMax* = 0, means that we do not expand terms based on hierarchies.

2.4. Obtaining the gene-disease background annotation set

Once we have the publication–disease tuples, we connect them to the gene-publication tuples resulting in gene-disease associations for 7316 human genes. Finally, we compute the disease concept frequencies from the gene-disease background annotation set by simply counting the number of distinct genes associated with each disease over all documents in the corpus. To compute the background frequency for each disease concept X, we add up the gene occurrence counts for disease concept X across all publications. We return this number (m) as well as the total number of genes in the background annotation set ($M = 7316$). The fraction m/M then represents the background frequency of the disease concept X in the annotated corpus. Using this frequency we can compute significant comparative over- or under-representation in an input dataset. See Fig. 2 for an illustration.

2.5. Calculating the gene-disease enrichment value for a set of interest

Using 261 genes known to be associated with aging (obtained from GenAge, see Section 3) as an example set of interest, we re-compute the frequency for each concept X by summing the gene occurrences for disease concept X this time only if the gene appears within this subset. We return this number (n) as well as the total number of genes in the set of interest ($N = 261$). The fraction n/N then represents the observed frequency of the concept X for the set of interest. See Fig. 2 for an illustration.

Finally, we assess how "surprising" it is to find n, given m, M and N by calculating the probability of observing a specific value given the background distribution. We use a simple binomial model to

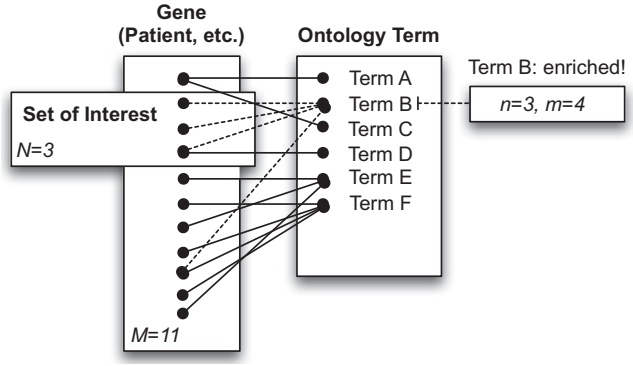

Fig. 2. Enrichment analysis: The background set consists of the entities under study (i.e., genes, patients, etc.). The background set depicted has $M = 11$ entities. The set of interest has $N = 3$ entities. Term B has three links to the set of interest ($n = 3$) and four links to the background set ($m = 4$). For p-values under 0.05, the binomial test shows that Term B is enriched for n/N versus m/M.

[3] http://eutils.ncbi.nlm.nih.gov/
[4] http://www.bioontology.org/wiki/index.php/Annotator_User_Guide

obtain these *p*-values, which assumes that the probability of picking a gene annotated with a disease concept is fixed and is equal to the proportion of genes annotated with that term in the reference set. Such an approximation is quite reasonable for large reference sets (e.g., the whole genome) because the probability of selecting a gene annotated with the term into the set of interest does not change significantly after each selection. An alternative for smaller reference sets would be to use a hypergeometric distribution, which models the selection process without replacement. The trade-offs between using the binomial versus the hypergeometric distribution for gene-centric enrichment tests are studied by McLean et al. [19].

2.6. Evaluation

In order to validate our background annotation set, we evaluate our gene-disease association dataset in several ways. First, we manually verify a handful of well-known genes, such as TP53, for known associations with diseases. Next, we examine a set of genes related specifically to aging from the GenAge database [23] for their coherence in terms of the assigned disease annotations. Finally, we perform disease-based enrichment analysis on the aging gene set using our newly created background set.

3. Results

The GO annotation files for *homo sapiens* reference 44,103 distinct PubMed articles and 11,125 distinct genes. This represents 44.5% of the 25,000 genes roughly estimated to exist currently. Of 25,000 genes, we are able to annotate 7316 (29.2%) with at least one disease concept from the DO.

3.1. Recapitulating known disease associations

On examining the automatically assigned disease annotations for a well-known gene such as TP53, we found that TP53 was annotated to *DNA Damage* 25 times based on 25 different abstracts, to *cancer* 16 times, *fibroepithelial neoplasm* nine times and was also annotated with specific diseases such as *colorectal cancer and Li-Fraumeni syndrome*. TP53 was also annotated, wrongly, to *Recruitment* four times. We discuss such mis-annotations and their effect on the enrichment analysis later in the paper. Similarly, BRCA1 was annotated with *hereditary breast ovarian cancer* 45 times, *malignant neoplasm of breast* 20 times, *DNA damage* 20 times, *malignant neoplasm of ovary* eight times and *retinoblastoma* two times (BRCA1 is known to bind the RB1, retinoblastoma 1, protein).

3.2. Summary of annotations of known aging genes

We used a set of 261 human genes known to be associated with aging, as provided by the GenAge database. For this subset, we pulled out the gene-disease associations from our automatically created disease annotation dataset. We were able to create annotations for 236 (91%) of the known aging genes in humans. By aggregating the number of genes per disease concept, we obtained frequencies for the top disease concepts listed in Table 1.

As we can see in Fig. 3, the concepts that annotate multiple genes in this aging-related gene set make biological sense. For example, *DNA Damage* is known to occur with aging. *Alzheimer's disease* and *Atherosclerosis* are also known to increase with age. There is also an obvious mis-annotation – *Recruitment*; this term is in the DO and is a synonym of *auditory recruitment* (DOID:12659) but does not have an asserted superclass, indicating a possible error in the ontology.

Table 1
The following table lists the top disease terms associated with aging-related genes. Most of the enriched terms are biologically meaningful as determined by medical experts; meaning that these terms are diseases or conditions that are understood to be associated with aging. Note that mis-annotated terms (such as Recruitment) and non-informative terms (such as Disease) are not deemed enriched by the analysis, as expected.

Term	Frequency	Enriched?	*p*-values
DNA damage	53	Yes	0.00002429743370435670
Alzheimer's disease	20	Yes	0.00210785976415961000
Recruitment	18		0.99999999719663000000
Atherosclerosis	13	Yes	0.00269862200185977000
Fibroepithelial neoplasm	10		1.00000000000000000000
Disease	9		1.00000000000000000000
Insulin resistance	7	Yes	0.01181595036983900000
Ataxia	6		0.36411632694226000000
Cockayne syndrome	6	Yes	0.00480121896261565000
Dependence	6		0.73713523946908800000
Burkitt's tumor or lymphoma	4	Yes	0.00975066042777684000
Dwarfism	2	Yes	0.00923404669915151000

Fig. 3. Aging-related diseases: Using the 261 human genes identified in the GenAge database and the gene-disease linkages discovered during the workflow we have defined, the tag cloud displays more frequently appearing terms using larger font sizes.

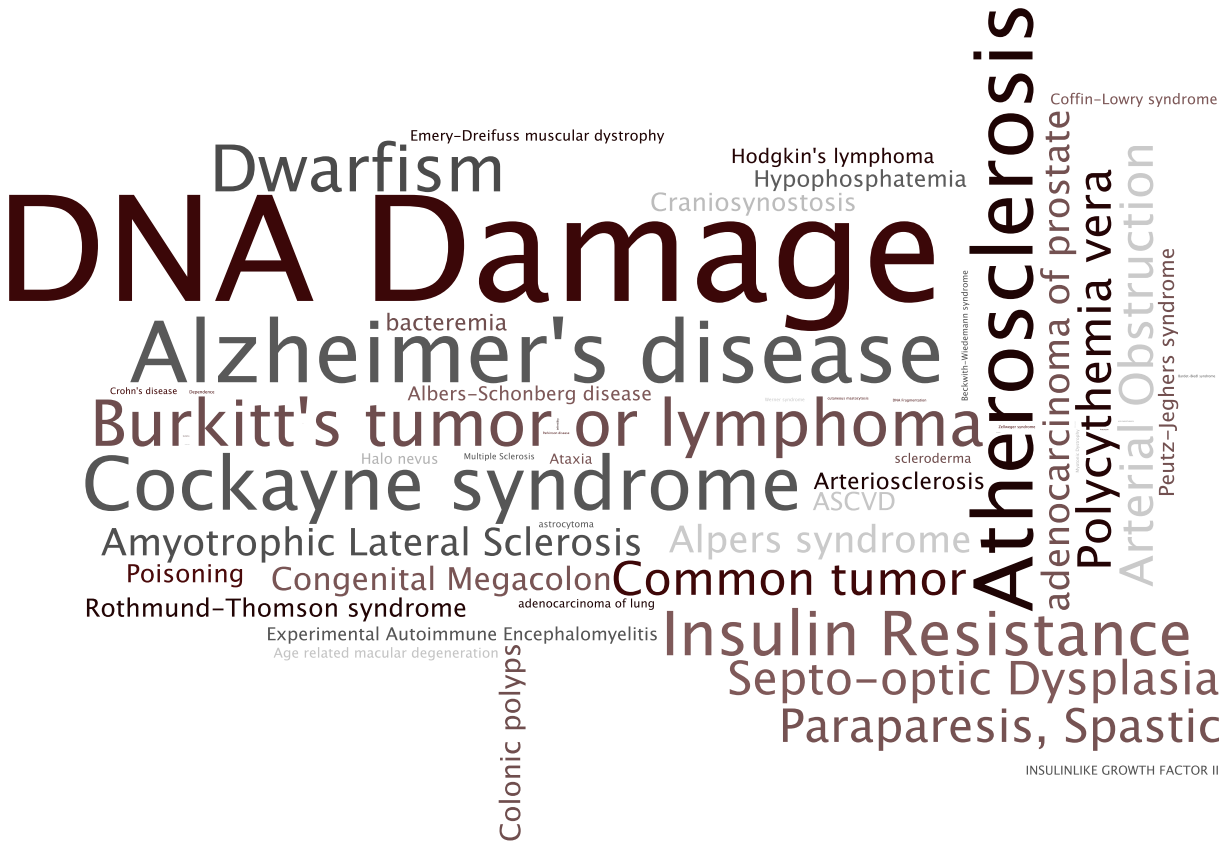

Fig. 4. Disease terms significantly enriched in annotations of aging-related genes: This tag cloud shows those disease terms in the annotations of the 261 aging-related genes that are statistically enriched given our gene-disease background annotation set. Terms that are significantly enriched appear larger.

3.3. Enrichment analysis for aging genes

As a further validation of the utility of our background annotation set, we identified disease concepts that are statistically enriched – in the annotations of the 261 aging-related genes – given our gene-disease background annotation dataset. As mentioned in our methods, we used a binomial test to detect enriched disease concepts in the aging-related gene set. Whether a particular concept is enriched or not is shown in Table 1 (enrichment column) and Fig. 4 provides a tag-cloud visual. Note that mis-annotated terms (such as *Recruitment*) and non-informative concepts (such as *Disease*) are not deemed enriched in the statistical analysis.

On examining the disease concepts found enriched, we found that in most cases the disease assigned to the aging gene was confirmed by a literature search. For example, *Cockayne syndrome*, *Burkitt's lymphoma*, *Spastic Paraparesis* and *Rothmund–Thomson Syndrome* are all specifically linked to the underlying gene that GeneAge declared as aging-related (and hence was in our set of 261 genes). In fact the relationship between accelerated aging syndromes (such as *Rothmund–Thomson Syndrome*) and natural aging is an emerging research area [6–9].

4. Discussion

From our results, the use of human curated annotations as a starting point for generating automated annotations using other ontologies seems promising-annotating 29.2% of genes. Previous methods that use advanced text mining have been able to annotate 4408 genes (17.7%) [21]. A study based on OMIM associated 1777 genes (7.1%) with diseases to create a human "diseasome" [22]. Based on a similar assumption that curated annotations will yield better results, Osborne et al. annotate GeneRIF descriptions to infer disease relationships to 5376 genes (21.5%) [16]. Because the number of human genes known at the time of these studies varies, we make the comparisons loosely.

However, there are some caveats to consider. First, not all ontologies are equally suited for creating automated annotations. Second, automated annotation depends highly on the quality of the text corpus. We discuss these issues below.

4.1. Using other ontologies

The Human Disease Ontology is a community-driven, open source ontology designed specifically to link disparate datasets through disease concepts under the principles of the Open Biomedical Ontologies Foundry. Osborne et al. note that the DO provides several advantages including good disease coverage and a useful hierarchical structure for mapping disease terms to a large text corpus [16]. For the purpose of the current discussion, and enrichment analysis in general, just about any disease ontology that provides a clear hierarchy of parent–child for diseases would be suitable for use—e.g., SNOMED CT, the National Cancer Institute thesaurus (NCIt), the International Classification of Diseases (ICD)—insofar as their vocabularies work well for entity extraction (see below). For concepts that are linked together (as in the UMLS Metathesaurus), the additional synonyms obtained from related concepts can be incorporated to expand the coverage of the concept recognition system. This feature, which is only partially available in the form of the Annotator

Table 2
The 24 (out of 261 aging-related) genes and gene products that were not associated with a disease term.

UniProt ID	Recommended name	Gene name
O00327	Aryl hydrocarbon receptor nuclear translocator-like protein 1	ARNTL
O15120	1-acyl-sn-glycerol-3-phosphate acyltransferase beta	AGPAT2
O15217	Glutathione S-transferase A4	GSTA4
O15243	Leptin receptor gene-related protein	LEPROT
O15516	Circadian locomoter output cycles protein kaput	CLOCK
O75844	CAAX prenyl protease 1 homolog	ZMPSTE24
O95985	DNA topoisomerase 3-beta-1	TOP3B
P00390	Glutathione reductase, mitochondrial	GSR
P00395	Cytochrome c oxidase subunit 1	MT-CO1
P09629	Homeobox protein Hox-B7	HOXB7
P13639	Elongation factor 2	EEF2
P20382	Pro-MCH	PMCH
P25874	Mitochondrial brown fat uncoupling protein 1	UCP1
P32745	Somatostatin receptor type 3	SSTR3
P36969	Phospholipid hydroperoxide glutathione peroxidase, mitochondrial	GPX4
P61278	Somatostatin	SST
P62987	Ubiquitin-60S ribosomal protein L40	UBA52
P78406	mRNA export factor	RAE1
P98177	Forkhead box protein O4	FOXO4
Q00613	Heat shock factor protein 1	HSF1
Q13219	Pappalysin-1	PAPPA
Q99643	Succinate dehydrogenase cytochrome b560 subunit, mitochondrial	SDHC
Q99807	Ubiquinone biosynthesis protein COQ7 homolog	COQ7
Q9UBI1	COMM domain-containing protein 3	COMMD3

mapping expansion component[5], is one of the aims of the NCBO but the effort is not complete yet.

Although we specifically focus on creating annotations with disease terminology, the method we have devised (Fig. 1) can create annotations with terms from *any* ontology. In our workflow, to obtain a background dataset for enrichment analysis using some ontology other than the DO, researchers would simply configure a parameter for the NCBO Annotator to use their ontology of choice from BioPortal. In fact, other researchers have used a similar annotation workflow to recognize morphological features in textual descriptions of fish species [24]. Moreover, researchers can use the annotation tool of choice (e.g., MetaMap).

However, not all ontologies are viable candidates for automatic annotation because the vocabulary used in texts is not always reflective of that in ontologies. For example, using term-frequency counts—for all terms from BioPortal ontologies—in MEDLINE abstracts [25], we calculated that disease terms are mentioned 46% more often than GO terms in MEDLINE abstracts. As another example, on comparing NCBO Annotator results using the GO to automatically annotate genes based on the PubMed articles provided as the basis of the GO annotation, we find that only 10% of the curated GO annotations can be detected directly in the paper abstract supporting a particular GO annotation.

Because disease terms are mentioned significantly more often than GO terms, the automated annotation process works better for annotating genes with disease ontology terms than it would for performing automatic GO annotation. Starting with curated gene-publication annotations ensures high accuracy.

4.2. Missing annotations

Out of the 261 aging-related genes in our evaluation subset, the NCBO Annotator left-out 24 genes (9%). Therefore, we have no disease terms associated with those genes in our gene-disease association dataset. The 24 genes (mentioned as UniProt IDs) that were not associated to a disease term are listed in Table 2. These missed annotations provide an opportunity for refining the annotation workflow to use sources beyond just the papers referenced in GO annotations, e.g., GeneRIF references.

4.3. Dealing with annotation errors

Some errors in annotation are inevitable in an automated process. For example, in the reference annotation set we created, TP53 was also annotated, wrongly, to *Recruitment*. Papers that were the basis of creating GO annotations for TP53 certainly mention the term *Recruitment*, however, that term is not a disease. Despite these kinds of errors, because they affect annotation of both the set of interest and the reference set equally, the errors will most likely cancel each other out when computing statistical enrichment (Fig. 4)—though that is not guaranteed. We can deal with such errors in a variety of ways. For example, we can add a list of terms that commonly lead to errors to our set of stop words (i.e., terms to ignore). We can also use more advanced text mining techniques to analyze the context in which a potential disease term is mentioned using MEDLINE term frequencies and part-of-speech tags [25] to detect false positives. The accuracy of the NCBO Annotator in recognizing disease names has been evaluated in prior work [17]. A repeat study of the sensitivity and specificity (or recall and precision) of the NCBO Annotator is outside the scope of this work.

4.4. Future work

Text mining can potentially provide checks against possible human errors. The *Recruitment* mis-annotation is one case in point. Additionally, for questionably enriched diseases, such as *Septo-optic dysplasia* (a developmental disorder believed to be linked to an aging gene), we plan to perform more thorough, manual validation. We would use the approach of review by two or more experts and a protocol for resolving disagreement, and we plan to do this work specifically in the context of disease enrichment in electronic patient records.

[5] http://www.bioontology.org/wiki/index.php/Annotator_User_Guide#The_mapping_expansion_component

Fig. 5. Disease terms significantly enriched in annotations of patient reports coded with ICD9 789.00 (Abdominal pain, unspecified), one of the top reasons for patients being admitted to emergency rooms. Patient reports were obtained from the University of Pittsburg NLP Repository (http://www.dbmi.pitt.edu/blulab/nlprepository.html) with IRB approval.

We acknowledge the fact that publications might mention negated findings. For this reason, we exclude the publications that were the basis of a negated GO annotation (i.e., qualified with NOT). Having negation detection functionality can certainly affect the outcome of enrichment analysis by altering the reference term frequencies. For this study, negations were not considered, but we have subsequently implemented the NegEx algorithm [26] within the framework of the Annotator and we are currently testing negation detection within the context of electronic health records.

Finally, we relied upon GO annotations to provide one reliable source for gene-publication associations. Likewise, GeneRIF descriptions and their referenced PubMed articles could offer another source of similar information. Compared to the 44,103 articles obtained from GO annotations in our case study, GeneRIF entries for *homo sapiens* reference a much larger corpus of 208,406[6] distinct PubMed articles. Other possible sources could include the CALBC[7] corpus, or Entrez eLink[8] publications. We have subsequently used GeneRIFs for finding gene-disease associations and are currently thinking of ways to compare the results with those from PubMed.

Although we demonstrate the benefits of enrichment analysis with ontologies other than the GO (especially those for which a reference annotation set does not exist) within the context of its most popular use (gene expression profiling), the method is general enough to benefit a host of other applications as well, such as profiling patient reports. For example, one of the most common reasons patients are admitted to the emergency room is for abdominal pain, which is often coded with ICD9 789.00 (Abdominal pain, unspecified). We can designate our set of interest as all patients coded with 789.00 and compare them against all patients on record using enrichment analysis over their clinical notes, to obtain a profile of this cohort (Fig. 5). We can use DO to obtain a disease profile, or we can use RxNorm to obtain a drug profile, and so on. We have subsequently obtained access to Stanford's entire clinical data warehouse, and have annotated over 9.5 million patient reports with 15 different ontologies. We are currently in the process of analyzing the results, which will include enrichment analysis along not only diseases and findings, but also along drugs and adverse events. The main challenge will be to develop methods of analyzing these aspects in combination.

5. Conclusion

Enrichment analysis using GO annotations is widely used to make sense of the results of high-throughput experiments. Our goal is to generalize enrichment analysis methods to use a variety of ontologies including SNOMED CT, Human Disease Ontology, and others. With the availability of automated ontology-based annotation with terms from biomedical ontologies, it is possible to

[6] http://www.ncbi.nlm.nih.gov/projects/GeneRIF/stats/
[7] http://www.calbc.eu/
[8] http://www.ncbi.nlm.nih.gov/corehtml/query/static/elink_help.html

perform enrichment analysis using ontologies other than the Gene Ontology. However, a key pre-requisite for such analysis is the availability of a background set of annotations to enable the enrichment calculation.

We have described a general method, which uses hand-curated GO annotations as a starting point, for creating background datasets for enrichment analysis using other ontologies—such as the Human Disease Ontology, for which hand-curated annotations are not available.

To demonstrate the feasibility and utility of our method, we have created a background set of annotations to enable enrichment analysis with the DO and validated that background set by using the created annotations to examine the coherence of known aging-related genes and by performing enrichment analysis on an aging-related gene set from the GeneAge database [23]. In future work, we plan to apply enrichment analysis methods to analyze patient cohorts from electronic health records.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Acknowledgments

We thank Sean D. Mooney at the Buck Institute for Age research for discussions. We acknowledge support from NIH grant U54 HG004028 for the National Center for Biomedical Ontology.

References

[1] Khatri P et al. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 2005;21(18):3587–95.
[2] Ade AS, et al. Genes2Mesh; 2007 [cited August 2010] <http://gene2mesh.ncibi.org>.
[3] Tirell R, et al. An ontology-neutral framework for enrichment analysis. In: AMIA annual symposium, Washington, DC; 2010.
[4] Noy NF et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nuc Acids Res 2009;37:W170–3.
[5] Mort M et al. In silico functional profiling of human disease-associated and polymorphic amino acid substitutions. Hum Mutat 2010;31(3):335–46.
[6] Puzianowska-Kuznicka M, Kuznicki J. Genetic alterations in accelerated ageing syndromes. Do they play a role in natural ageing? Int J Biochem Cell Biol 2005;37(5):947–60.
[7] Cox LS, Faragher RG. From old organisms to new molecules: integrative biology and therapeutic targets in accelerated human ageing. Cell Mol Life Sci 2007;64(19–20):2620–41.
[8] Ramírez CL et al. Human progeroid syndromes, aging and cancer: new genetic and epigenetic insights into old questions. Cell Mol Life Sci 2007;64(2):155–70.
[9] Ding SL, Shen CY. Model of human aging: recent findings on Werner's and Hutchinson-Gilford progeria syndromes. Clin Interv Aging 2008;3(3):431–44.
[10] Michaud K, Wolfe F. Comorbidities in rheumatoid arthritis. Best Pract Res Clin Rheumatol 2007;21(5):885–906.
[11] Turesson C et al. Cardiovascular co-morbidity in rheumatic diseases. Vasc Health Risk Manag 2008;4(3):605–14.
[12] John H et al. Cardiovascular co-morbidity in early rheumatoid arthritis. Best Pract Res Clin Rheumatol 2009;23(1):71–82.
[13] Petri H et al. Data-driven identification of co-morbidities associated with rheumatoid arthritis in a large US health plan claims database. BMC Musculoskelet Disord 2010;11:247.
[14] LePendu P et al. Optimize first, buy later: analyzing metrics to ramp-up very large knowledge bases. Int Sem Web Conf (ISWC) 2010;9:486–501.
[15] Jonquet C, et al. NCBO resource index: ontology-based search and mining of biomedical resources. Int Sem Web Challenge at ISWC (1st prize). <http://challenge.semanticweb.org/>; 2010.
[16] Osborne JD et al. Annotating the human genome with disease ontology. BMC Genomics 2009;10(Suppl 1):S6.
[17] Jonquet C, et al. The open biomedical annotator. AMIA TBI Summit; 2009.
[18] Krallinger M et al. Analysis of biological processes and diseases using text mining approaches. Methods Mol Biol 2010;593:341–82.
[19] McLean CY et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 2010;28(5):495–501.
[20] Subramanian A et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005;102(43):15545–50.
[21] Hoffmann R, Valencia A. Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics 2005;21(Suppl 2):ii252–8.
[22] Goh KI et al. The human disease network. Proc Natl Acad Sci USA 2007;104(21):8685–90.
[23] de Magalhaes JP et al. The human ageing genomic resources: online databases and tools for biogerontologists. Aging Cell 2009;8(1):65–72.
[24] Sarkar N. Using biomedical ontologies to enable morphology based phylogenetics: a feasibility study for fishes. At the bio-ontologies SIG at ISMB, Boston, MA; July 9–10th, 2010.
[25] Xu, R., et al. A Comprehensive analysis of UMLS metathesaurus terms using eighteen million MEDLINE abstracts. In: AMIA annual symposium, Washington, DC, 2010.
[26] Chapman W et al. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(5):301–10.