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.

© 2011 Elsevier Inc. All rights reserved.

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

Download English Version:

https://daneshyari.com/en/article/10356114

Download Persian Version:

https://daneshyari.com/article/10356114

Daneshyari.com