Summarize gene annotations based on collective ontology annotations

Pan Du[‡*], Gang Feng[‡†], Warren A. Kibbe[‡‡], Simon Lin[‡§]

October 18, 2010

[‡]The Biomedical Informatics Center
Northwestern University, Chicago, IL, 60611, USA

Contents

1 Introduction

As computational and high throughput analyses have been widely used in interpreting gene functions, the number of gene annotations and resultant metadata describing the conditions for each annotation has increased dramatically. To standardize these annotations, genes are usually annotated by associating with standard ontology terms. Since the number of these annotations have increased, interpreting the major biological roles of a given gene and gene product based on these ontology terms has become increasingly complex.

We proposed a statistic test to estimate the enrichment of ontology terms associated with a gene. These ontology terms are then ranked by annotation scores defined based on enrichment p-values. A miniSet of ontology terms is finally created to summarize the major functions associated with these ranked

[*]dupan@northwestern.edu
[†]g-feng@northwestern.edu
[‡]wakibbe@northwestern.edu
[§]s-lin2@northwestern.edu

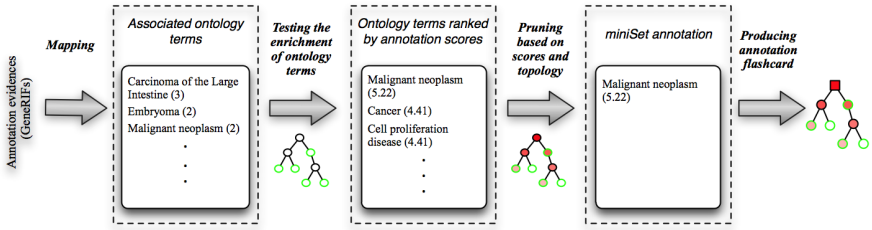

Figure 1: An example of the input data format

ontology terms, and this miniSet can be graphed as an annotation flashcard. We use Disease Ontology (DO) as the example to show the effectiveness of the functions.

Our evaluation results show that this method is robust (adding random assignment of 40% of the overall annotations does not significant perturb the result set with high annotation scores) and accurate (on average about 80% of summarized top miniSet annotations match with the existing publication records). The summarized annotations are much easier for researchers and curators to interpret and curate. Applying miniSet annotations to the functional enrichment analysis of a public gene list results in a more concise and biologically relevant analysis. This quantitative annotation method can be extended to any well-constructed ontology. Please check the reference paper [1] for more details.

2 Methods

Figure 1 shows the steps of gene function summarization starting from the annotation evidence (GeneRIF statements) to the final annotation flashcard. We have used the gene PEBP1 as an example. As described in [2], we first mapped GeneRIFs statements to Disease Ontology terms using the MMTx program [3] developed by NLM to build the association statements between a gene and a disease ontology term. Then we tested the enrichment of individual ontology terms by mapping GeneRIF associated ontology terms to more general terms (walking 'up the graph' of ontology terms) based on the ontology hierarchical structure. Annotation scores were calculated based on the enrichment p-values. To summarize the significantly enriched ontology terms, a miniSet annotation was further built based on the annotation scores and ontology structure. These summarized annotations are graphed in an annotation flashcard.

2.1 Major functions of gene function summarization

The *GeneAnswers* package implemented functions shown in Figure 1. Function `geneFunSummarize` summarize gene functions (annotations) based collective annotation evidences associated with ontology terms. Basically, it tests the enrichment (using hypergeometric test) of all associated ontology terms of the gene and ranks these related ontology terms based on their statistical significance (starting from the most significant one). Function `simplifyGeneFunSummary` simplifies the significant ontology terms to a mini-set, which includes the non-overlapping most significant terms and some other ontology terms, which

have direct gene mapping but not included in the significant ontology terms. Function `plotGeneFunSummary` plots ontology graphs of the summarized gene annotation (ontologies), which is return by function `geneFunSummarize`. For the convenience of customize the plot of ontology DAG, we added another function `plotGraph` to plot and render a graphNEL object. If the user processes the function summarization of many genes in batch. The results can be saved as a tab-seperated text file using function `saveGeneFunSummary`.

3 Example dataset

The *GeneAnswers* package includes a Disease Ontology related data file "DO.rda", which includes five datasets:

DO.graph.gene: a graphNEL object, which shows the ontology relations of DO

DO.graph.closure.gene: a graphNEL object, whose edges represent the link between a DO term and its offspring ontology terms. Only the DO terms with gene mappings were included.

DO2gene.map: a list show the mapping from DOIDs to genes

gene2DO.map: a list show the mapping from genes to DOIDs

DO.terms: a named character vector. Its names are DOIDs and elements are DO.terms

```
> rm(list=ls())
> library(GeneAnswers)
> ## load the DO data file, which includes several data sets.
> data(DO)
> ## show the datasets included in DO.rda file
> ls()
```

```
[1] "DO"                    "DO.graph.closure.gene" "DO.graph.gene"
[4] "DO.terms"              "DO2gene.map"           "gene2DO.map"
```

4 Examples of summarizing gene annotation and plot the annotation flashcard

Here we shows a simple example of summarizing the annotations of a particular gene, PEBP1 (Entrez Gene ID: 5037). The gene should be specified using "Entrez Gene ID". Figure 2 shows the summarized gene annotation based on Disease Ontology.

```
> geneSummary <- geneFunSummarize('5037', gene2DO.map, DO.graph.closure.gene)
> # simplify the summarized annotations to miniSet
> geneSummary.sim <- simplifyGeneFunSummary(geneSummary, DO.graph.closure.gene, p.value.th
> # print the miniSet
> geneSummary.sim
```

```
$`5037`
$`5037`$keptSigOntoID
[1] "DOID:462"
```

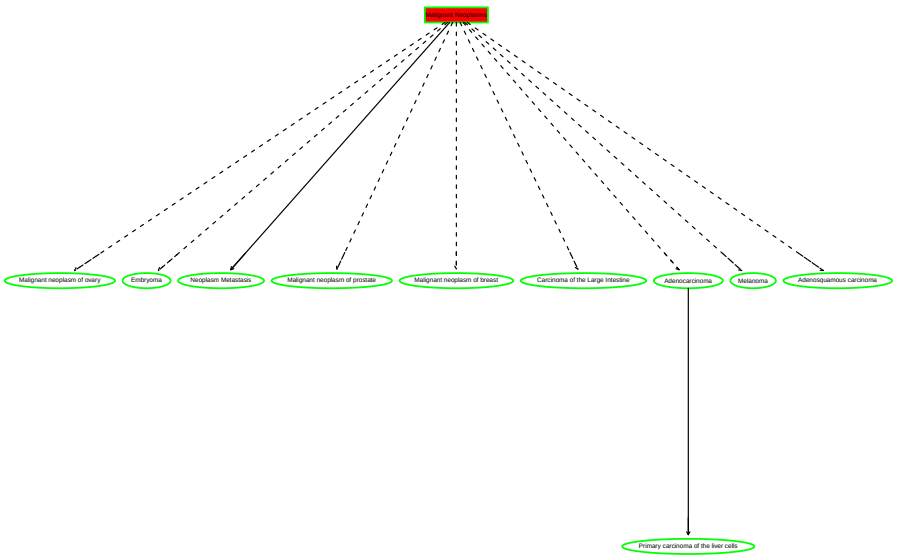

Figure 2: Plot of summarized annotation of gene PEBP1

```
$`5037`$keptEvidences
character(0)

$`5037`$scores
DOID:462
     5.2


attr(,"pValueT")
[1] 1e-05
```

Users can also process the gene function summarization in batch. The following code processes all genes in DO database.

```
> allGenes <- names(gene2DO.map)
> length(allGenes)
> # summarize all genes in a batch
> geneSummary.all <- geneFunSummarize(allGenes, gene2DO.map, DO.graph.closure.gene, fdr.ad
> # simplify the summarized annotation as the miniSet
> sim.geneSummary.d.all <- simplifyGeneFunSummary(geneSummary.all, DO.graph.closure.gene,
> # save the summarized annotations in a tab-separated text file.
> saveGeneFunSummary(geneSummary.all, simplifyInfo=sim.geneSummary.d.all, ID2Name=DO.terms
```

5 Session Info

```
> toLatex(sessionInfo())
```

- R version 2.12.0 (2010-10-15), x86_64-unknown-linux-gnu
- Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=en_US.UTF-8, LC_COLLATE=C, LC_MONETARY=C, LC_MESSAGES=en_US.UTF-8, LC_PAPER=en_US.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, grDevices, graphics, grid, methods, stats, utils
- Other packages: AnnotationDbi 1.12.0, Biobase 2.10.0, DBI 0.2-5, GO.db 2.4.5, GeneAnswers 1.6.0, Heatplus 1.20.0, KEGG.db 2.4.5, MASS 7.3-8, RColorBrewer 1.0-2, RCurl 1.4-3, RSQLite 0.9-2, XML 3.2-0, annotate 1.28.0, bitops 1.0-4.1, igraph 0.5.4-1, org.Hs.eg.db 2.4.6, org.Mm.eg.db 2.4.6, rgl 0.92.794
- Loaded via a namespace (and not attached): RBGL 1.26.0, Rgraphviz 1.28.0, graph 1.28.0, tools 2.12.0, xtable 1.5-6

6 References

1. Pan Du, Simon M. Lin, Gang Feng, Warren A. Kibbe, "GeneRIFcompendiate: Ranked gene annotations using collective GeneRIF associations and ontology terms", (under review)

2. Osborne, J.D., Flatow, J., Holko, M., Lin, S.M., Kibbe, W.A., Zhu, L.J., Danila, M.I., Feng, G. and Chisholm, R.L. (2009) Annotating the human genome with Disease Ontology, BMC Genomics, 10 Suppl 1, S6.

3. Aronson, A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesau-rus: the MetaMap program, Proc AMIA Symp, 17-21.