Quantitative summarization of gene annotations using Disease Ontology

Pan Du, Ph.D.

Biomedical Informatics Center
Northwestern University
Outline

- Background
- Method of quantitative summarization of gene annotations
- Evaluation of the method
- Applications
- Conclusion
4070 genes associated with 1851 DO terms
Lots of genes associate with tens, even over one hundred DO terms (multiple GeneRIF records may map to the same DO term).
  - Over 1100 genes have > 5 disease related GeneRIF records
  - Over 600 genes have > 10 disease related GeneRIF records
  - Gene EGFR (Entrez ID 1956) has 458 disease related GeneRIF records
An Example Gene Annotation using DO (Gene TDP-43)

GeneRIFs: Gene References Into Function

1. tdp-43 deposits have been associated with neurodegenerative diseases--{review}
2. several mutations of the TDP-43 gene were identified as the causative gene of autosomal-dominant familial ALS (review)
3. TDP-43-positive inclusions within neurons and oligodendroglia were found in brains from patients with Alzheimer’s disease (AD) and dementia with Lewy bodies (DLB)
5. TDP-43 is the newest member of the growing list of neurodegenerative proteinopathies, but unique in that it lacks features of brain amyloidosis.
6. TDP-43 may not necessarily be the key disease protein in ALS and indicate that the major target(s) of ubiquitination remain to be identified.
7. The presence of TDP-43 inclusions in PID suggests that TDP-43 accumulation may be an important component of many neurodegenerative diseases, and that its presence in only some cases of PID may indicate different pathways of disease development.
8. distinct TDP-43 profiles may affect clinical phenotypes differentially in patients with frontotemporal lobar degeneration with ubiquitin-positive inclusions
9. TDP-43 proteinopathies are distinct from most other neurodegenerative disorders--{review}

Submit: New GeneRIF  Correction

Map the GeneRIFs to Disease Ontology terms:

"Alzheimer's disease" (2), "Amyloidosis", "Amyotrophic Lateral Sclerosis" (2), "dementia", "Motor Neuron disease", "Neurodegenerative disorders" (2), "Pelvic Inflammatory disease"
How to identify the major annotation of a gene?

Overview of annotation summarization method

Annotation evidences (GeneRIFs) → Mapping → Associated ontology terms
  - Alzheimer's disease (2)
  - Amyloidosis
  - Amyotrophic Lateral Sclerosis (2)
  - ...

Testing the enrichment of ontology terms

Ontology terms ranked by annotation scores
  - Neurodegenerative disorders (8.7)
  - Degenerative disease (8.6)
  - ...

Pruning based on scores and topology
  - Neurodegenerative disorders (8.7)
  - Pelvic Inflammatory disease (1)
  - Amyloidosis (1)

Producing annotation plot

GeneRIFcompendiate

NORTHWESTERN UNIVERSITY
Gene Annotation Summarization using Enrichment Tests

- Evidences (GeneRIFs) + Prior knowledge (ontology) → Gene annotation summarization
  - To test whether a specific annotation (ontology terms) is over-represented among all annotations of a gene
  - The test is similar with the functional enrichment analysis based on a gene list
  - The enrichment test will be performed for every ontology with GeneRIF association (direct / indirect)
Estimate the Statistics Significance of Annotation Enrichment

- If we have a gene \( \text{Gene}_i \), it has \( k \) DO mappable GeneRIF records, \( q \) of them are associated with an DO term \( \text{DO}_j \), which has \( m \) associated GeneRIFs (from different genes) in NCBI database. We want to test the enrichment of function \( \text{DO}_j \) for this \( \text{Gene}_i \).

- P-value: The probability of randomly picking \( k \) GeneRIFs from all \( N \) DO-mappable GeneRIFs in NCBI database, and \( q \) GeneRIFs will be associated with the \( \text{DO}_j \) by chance.

- Define an annotation score:

\[
\text{Annotation Score} = \begin{cases} 
\max\{1, -\log_{10}(p - \text{value})\}; & \text{with annotation evidences} \\
0; & \text{without annotation evidence}
\end{cases}
\]
Estimating annotation score of “Neurodegenerative disorders”:

\[ N = 44019 \]
\[ |Evd_{onto}| = 2236 \]
\[ |Evd_{gene}| = 10 \]
\[ |Evd_{gene} \cap Evd_{onto}| = 8 \]

p-value = 1.995262e-09

Annotation score of “Neurodegenerative disorders” is 8.7
Motivation:
- Lots of DO terms could be significant after enrichment tests
- Need to define a DO subgraph to best represent the gene annotation

Assumption:
- The DO term with highest score (larger than a certain threshold) can best summarize all GeneRIFs associated with this DO term
- Other DO terms with the same or subset of associated GeneRIFs can be represented by this DO term.
miniSet:
“Neurodegenerative disorders” (8.7),
“Pelvic Inflammatory disease” (1),
“Amyloidosis” (1)
Evaluation

- Test the robustness of function summarization method by adding random annotation
- Compare top annotation in the miniSet with publication records
- Apply the miniSet annotation in the functional analysis based on a gene list
Evaluation Results by Adding Random Annotations

![Graph showing evaluation results by adding random annotations. The x-axis represents the Annotation Score, and the y-axis represents the Percentage of correct matches (%). Different lines and markers represent different percentages of random annotations added: 5%, 10%, 20%, 40%, and pure random annotations.]
Compare top annotation in miniSet with publication records

![Graph showing the percentage of direct GeneRIF matches against the annotation score of the top DO term.](image)
A benchmark microarray data set of pancreatic cancer study (Antonov, et al., 2008) was previously utilized to test GO-based annotations. A list of 125 genes identified in that study.

Functional analysis based on 125 genes

Compare DOLite (using direct annotation mapping of each gene), DOLite (using miniSet mapping of each gene)
## Compare the Evaluation Results

<table>
<thead>
<tr>
<th>Rank</th>
<th>DOLite-direct</th>
<th>DOLite-miniSets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cancer</td>
<td>Cancer</td>
</tr>
<tr>
<td>2</td>
<td>Breast cancer</td>
<td>Breast cancer</td>
</tr>
<tr>
<td>3</td>
<td>Embryoma</td>
<td>Colon cancer</td>
</tr>
<tr>
<td>4</td>
<td>Colon cancer</td>
<td>Embryoma</td>
</tr>
<tr>
<td>5</td>
<td>Lung cancer</td>
<td>Lung cancer</td>
</tr>
<tr>
<td>6</td>
<td>Neoplasm metastasis</td>
<td>Pancreas cancer</td>
</tr>
<tr>
<td>7</td>
<td>Stomach cancer</td>
<td>Diabetes mellitus</td>
</tr>
<tr>
<td>8</td>
<td>Squamous cell cancer</td>
<td>Neoplasm metastasis</td>
</tr>
<tr>
<td>9</td>
<td>Prostate cancer</td>
<td>Prostate cancer</td>
</tr>
<tr>
<td>10</td>
<td>Pancreas cancer</td>
<td>Ovary cancer</td>
</tr>
</tbody>
</table>
Relations of top 10 category (DOLite using direct mapping)
Relations of top 10 category (DOLite using miniSet mapping)
Applications

- Identify the major annotations of a gene, and provide good guidance for researchers and database curators
- Functional analysis based on a gene list
- Estimate the functional similarity between genes
- Evaluate the ontology structure (or sub-structure)
Conclusion and Future Work

■ Conclusion
  – The first method provides quantitative annotation information
  – The method is robust for random mapping (annotation errors)
  – The method is effective: most top summarized functions match with publication records.
  – Applying miniSet annotations to the functional enrichment analysis provides more concise and biologically relevant analysis

■ Future Work
  – Using the annotation scores in the functional analysis
  – Further expert curations based on computational results
  – Using similar strategies for other ontologies, like GO
Acknowledgements

- NUgene Team at Northwestern University
- Rex Chisholm, founder of Disease Ontology
- Disease Ontology PIs – Warren Kibbe, NU, and Lynn Schriml, UMB
- John Osborne for MMTx and UMLS mining
- NUBIC: Simon Lin, Gang Feng, Jared Flatow
- NIH Grants: 1R01RR025342-01 (Disease Ontology) and 5U01HG004609-03 (NUgene eMERGE consortium)
5. “GeneRIFcompendiate : Quantitative gene annotation using collective GeneRIF associations and ontology terms” (submitted)