Using computational predictions to improve literature-based Gene Ontology annotations

Julie Park, Ph.D.
Saccharomyces Genome Database • http://www.yeastgenome.org/
Department of Genetics • Stanford University School of Medicine
Attaining curation nirvana...

- Curation efficiency
- Annotation consistency
- Data accuracy
…is not easy!

Annotation errors

1. Mistakes in capturing the annotation
2. Outdated information
3. Missing annotations

How can you find these errors?
Flavors of GO annotations

1. Literature-based – “Manual”
   Individually assigned by biocurators based on the published literature

2. Computationally-predicted – “Computational”
   Automatically generated by in silico methods such as protein signatures or computational algorithms

Sources of computational predictions in SGD
   - InterPro
   - Swiss-Prot Keywords (SPKW)
   - YeastFunc
   - BioPixie

2. http://www.ebi.ac.uk/GOA/Swiss-ProtKeyword2GO.html
Flavors of GO annotations

1. Literature-based – “Manual”
   Individually assigned by biocurators based on the published literature

2. Computationally-predicted – “Computational”
   Automatically generated by \textit{in silico} methods such as protein signatures or computational algorithms

Sources of computational predictions in SGD
- InterPro
- Swiss-Prot Keywords (SPKW)
- YeastFunc
- BioPixie

\textit{Is it possible to take advantage of the strengths of computational predictions and leverage these annotations to improve manual ones?}
CvManGO: Computational vs. Manual GO Annotations
Do discrepancies between a literature-based annotation and a computational prediction indicate that the manual annotation needs to be updated?
CvManGO: Computational vs. Manual GO Annotations

OK

Discrepancies

No parent-child relationship between terms
CvManGO: Computational vs. Manual GO Annotations

M = C

4379 genes
Reviewed 336 genes

No parent-child relationship between terms

from October 2009 gene_association.sgd file
(6353 total genes)
Discrepancies can identify genes that need updating

Control

Genes flagged with discrepant annotation pairs

Percentage of total genes reviewed

- No change
- Updatable
Extrapolating to the entire genome

Still requires reviewing 4379/6353 genes—can we narrow this down further?
Factoring in the type of update

- Flagged for review resulting in potential improvement (add novel annotation): 20%
- Flagged for review resulting in potential improvement (refinement and removal only): 33%
- Not flagged for review: no discrepancies with computational predictions: 16%
- No computational prediction available: 15%
- Flagged for review: no change needed: 16%
Factoring in the type of update

- Flagged for review resulting in potential improvement (add novel annotation): 20%
- Flagged for review resulting in potential improvement (refinement and removal only): 33%
- Not flagged for review: no discrepancies with computational predictions: 16%
- No computational prediction available: 15%
- Not flagged for review: no change needed: 16%
Attributes of flagged genes

What are factors that enrich for genes missing annotations?

• Type of discrepancy
• GO aspect
• Amount of literature for a gene
• Source of the computational prediction
• Number of computational sources with discrepancies
Attributes of flagged genes

What are factors that enrich for genes missing annotations?

• Type of discrepancy

- GO aspect
- Amount of literature for a gene
- Source of the computational prediction
- Number of computational sources with discrepancies
Analysis by Class of Discrepancies

Shallow class

M → C

Mismatch class

M → C

M → C

M → C
Analysis by Class of Discrepancies

<table>
<thead>
<tr>
<th>Class</th>
<th>% updatable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shallow class</td>
<td>78.8%</td>
</tr>
<tr>
<td>Mismatch class</td>
<td>59.2%</td>
</tr>
</tbody>
</table>
Types of annotation updates by class

Number of genes needing:
- annotation refinement
- annotation removal
- novel annotation addition

Mismatch:
- 42 genes
- 18 require refinement
- 16 require removal
- 14 require novel annotation addition

Shallow:
- 138 genes
- 5 require refinement
- 1 require removal
- 16 require novel annotation addition

- 24 require refinement
- 20 require novel annotation addition
Summary & Conclusions

• Majority of *S. cerevisiae* literature-based GO annotations are good

• Comparing manual vs. computational prediction can identify genes whose annotations need updating

• Additional work needs to be done to pinpoint these annotations and genes
Summary & Conclusions

• Majority of *S. cerevisiae* literature-based GO annotations are good

• Comparing manual vs. computational prediction can identify genes whose annotations need updating

• Additional work needs to be done to pinpoint these annotations and genes

It works but there is still work to do!
Future plans

• Identify predictive features of genes that need updating
  - Are there specific GO terms used for manual curation more likely to be updated?
  - Do specific computational predictions indicate a GO term should be updated?
  - Examine node distance between GO terms used for computational and literature-based annotations
  - Examine contribution of annotation date and new publications
  - A combination of or all of the above?

• Evaluate the accuracy of computational predictions for *S. cerevisiae*

• Expand to evaluate annotations made based on orthology
  - Annotations from GOC PAINT project

• Develop a pipeline for curation prioritization at SGD

• Extend to other annotation projects
GO Consortium

http://www.geneontology.org/
Saccharomyces Genome Database staff

@yeastgenome  http://on.fb.me/ksgskb  sgd-helpdesk@lists.stanford.edu

Supported by NIH, National Human Genome Research Institute