Lecture 4

Data Mining: Visualization Techniques

Combined Graphs and Pies

Customizing graphs
Data Mining on the DAG

- When working with large datasets, annotation results need to be summarized.
- The DAG provides visualization of annotation data within its biological context.
- In Blast2GO --&gt; Combined Graph Function.
Combined Graph

Each term has a number of sequences associated

Nodes can be coloured to indicate relevance

Each term is displayed around its biological context

Node shape to differentiate between direct and indirect annotation
Combined Graph

Different GO branches
Reduces nodes by number of annotate sequences
Node data to be displayed
Criterion for highlighting and filtering nodes
Let's paint the DAG of the dataset analyzed yesterday (1000 sequences)

Too many nodes!!!

Need way to find relevant information
Node information content

Accumulated by node
(Sequence Count)

Incomming information
(Node Score)

\[
\sum seq(g) \alpha^{dist(g, g')} \quad g \in desc(g')
\]
Node score

We compute a node score that reflects the amount of direct information at the node

\[
\text{score}(\hat{g}) = \sum_{g \in \text{desc}(\hat{g})} \text{gp}(g) \times \alpha^{\text{dist}(\hat{g}, g)}
\]

where:

- \text{desc}(x): all the descendent terms for a given GO term \(x\)
- \text{dist}(x, y): the number of edges between GO term \(x\) and GO term \(y\)
- \(g\): an element of the GO where GO is the overall set of all GO terms
- \(\text{gp}(x)\): the number of assigned gene product to a given GO term \(x\)
Node score

NodeScore (GO1) = $1 \times 0.6^0 = 1$

NodeScore (GO2) = $3 \times 0.6^0 = 3$

NodeScore (GO3) = $1 \times 0.6^1 + 3 \times 0.6^1 = 0.6 + 1.8 = 2.4$

NodeScore (GO4) = $1 \times 0.6^2 + 3 \times 0.6^2 + 1 \times 0.6^0 = 0.36 + 1.08 + 1 = 2.5$

$$\sum_{g \subseteq \text{desc}(g')} \text{seq}(g) \times \alpha^{\text{dist}(g, g')}$$

$\alpha = 0.6$
Node score vs Annotation score

Annotation Score:
- In annotation context
- Relates to Blast results of ONE sequence

Node Score:
- In data-mining context
- Relates to analysis of a GROUP of sequences

AS = max(\%sim * Ecw) + (#InheAnn-1) * GOw

DO NOT MIX-UP !!!!!
Filtered Graph

- **Filtered Nodes**
- **Transition nodes**
- **Direct annotations**
Compacting Graphs by GOSlim
Show node content
Saving Options

Save as picture and as txt

<table>
<thead>
<tr>
<th>Level</th>
<th>GO ID</th>
<th>Term</th>
<th>Type</th>
<th>#Seqs</th>
<th>Graph</th>
<th>Score</th>
<th>Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>GO:0016052</td>
<td>carbohydrate catabolic process</td>
<td>biological_process</td>
<td>36</td>
<td></td>
<td>4.67</td>
<td>C02009A12, C04019G12,</td>
</tr>
<tr>
<td>7</td>
<td>GO:0043687</td>
<td>post-translational protein modification</td>
<td>biological_process</td>
<td>62</td>
<td></td>
<td>34.82</td>
<td>C04016C08, C08010E08,</td>
</tr>
<tr>
<td>3</td>
<td>GO:0016043</td>
<td>cellular component organization and</td>
<td>biological_process</td>
<td>242</td>
<td></td>
<td>51.93</td>
<td>C04018F11, C18004G08,</td>
</tr>
<tr>
<td>5</td>
<td>GO:0051252</td>
<td>regulation of RNA metabolic process</td>
<td>biological_process</td>
<td>25</td>
<td></td>
<td>13.77</td>
<td>C04016E04, C04013G11,</td>
</tr>
<tr>
<td>4</td>
<td>GO:0006725</td>
<td>aromatic compound metabolic process</td>
<td>biological_process</td>
<td>44</td>
<td></td>
<td>26.64</td>
<td>C08012A08, C02016F08,</td>
</tr>
<tr>
<td>4</td>
<td>GO:0046907</td>
<td>intracellular transport</td>
<td>biological_process</td>
<td>38</td>
<td></td>
<td>24.57</td>
<td>C18004G08, C02009C02,</td>
</tr>
<tr>
<td>8</td>
<td>GO:0006094</td>
<td>gluconeogenesis</td>
<td>biological_process</td>
<td>21</td>
<td></td>
<td>21</td>
<td>C04013C12, C07009E02,</td>
</tr>
<tr>
<td>4</td>
<td>GO:0006519</td>
<td>amino acid and derivative metabolic process</td>
<td>biological_process</td>
<td>78</td>
<td></td>
<td>34.82</td>
<td>C02016F08, C04013E11,</td>
</tr>
<tr>
<td>3</td>
<td>GO:0009719</td>
<td>response to endogenous stimulus</td>
<td>biological_process</td>
<td>48</td>
<td></td>
<td>20.37</td>
<td>C18004D02, C18002H02,</td>
</tr>
<tr>
<td>5</td>
<td>GO:0007047</td>
<td>cell wall organization and biogenesis</td>
<td>biological_process</td>
<td>15</td>
<td></td>
<td>5.75</td>
<td>C02015B04, C04018D06,</td>
</tr>
<tr>
<td>4</td>
<td>GO:0044248</td>
<td>cellular catabolic process</td>
<td>biological_process</td>
<td>75</td>
<td></td>
<td>23.13</td>
<td>C08012A08, C08011C08,</td>
</tr>
</tbody>
</table>
Graph Charts
Graph Charts

- Sequence Distribution/GO as Bar-Chart
- Sequence Distribution/GO as Level-Pie (level selection)
- Sequence Distribution/GO as Multilevel-Pie (#score or #seq cutoff)
M-Pies

# Seq filter = 20

Handy to summarize functional content

GOSlim
Example: analysis a specific function

Find a function on the graph
Example: analysis a specific function

Select all sequences annotated to this function and its descendents.
Example: analysis a specific function

Locate these sequences
Example: analysis a specific function

Exporting sequence table you see sequences
Annotated to the function

Explore annotation diversity within
The function
Conclusions

✓ DAGs are interesting for browsing functional annotation but can be too large
✓ With filtering and pruning options you can create more navigable DAGs
✓ Pies are good to compact information: try out levels
✓ GOSlim compacts to more equivalent terms than filtering the GO
✓ You can create your own colouring schema