A Knowledge-Based Clustering Algorithm Driven by Gene Ontology

Jill Cheng

Affymetrix, Inc.

Jan 15, 2004
The DAG structure of Gene Ontology

One-stop-shopping for biological information

Digraphs are computable
Goal
The closer a node is to the root, the more general its biological classification, thus a greater amount of information is conveyed by higher level edges.

The more common parent nodes shared the higher the degree of similarity.
Pair-wise similarity score between GO terms

\[ W_p = \sum_{n=0}^{p} (wt)^n, \quad p > 0; \quad W_0 = 0 \]

A weighting factor \((wt)\) was assigned to each edge as a function of the depth \((n)\) in the digraph, I chose a value of 0.815 to maximize \((wt6 - wt3)\).

\[ C = \sum_{n=0}^{\max-1} (wt)^n \]

Determining the longest partial path shared by two nodes, \(W_p\) is the sum of weights for edges from root to level \(p\).

\[ Nf_p = \frac{W'_p}{W_p} \]

A partial normalization scheme was applied to factor in the unevenness of the GO digraph.

\[ W_m = Nf_p \sum_{n=0}^{m} (wt)^n, \quad m > 0 \]

Calculate the average length for all paths that go through the shared partial path \((p)\), followed by the weight for a hypothetical path with \(p\) edges \((Wp)\).

\(Wp\) is transformed to \(W'p\), the mean of \(Wp\) and \(C\).

The normalization factor \((Nfp)\) is the ratio of \(W'p\) and \(Wp\).

The value for a partial path with \(m\) edges \((Wm)\) is normalized by applying \(Nfp\).
Annotation database schema
Spike-in experiment

Five related GO nodes with GOids 5381, 8490, 15344, 15620, and 15621; labeled red; were spiked into a randomly selected pool of 20 nodes and subjected to GO clustering. The similarity analysis successfully re-created the set of related GO nodes. Column 1 and 2 in the table shows a pair of GO nodes and column 3 shows the pair-wise similarity scores. Nodes colored pink (15342, 15359) are from the randomly selected 20 Go nodes and were clustered with the spiked GO nodes. Green circle indicates the cluster root (15291), which is the lowest level common ancestor node.
Transgenic Myeloid Progenitor (MPRO) cells transgenic for the dominant negative Retinoic Acid (RA) receptor were induced to differentiate into Neutrophils with high doses of RA.

Gene expression at 0, 1, 2, 4, and 8 hours post RA induction was analyzed with Affymetrix U74Av2 mouse microarray.

Genes showing significant changes in their expression level across a series of time points are modulated by retinoic acid stimulation and cell differentiation.

We arbitrarily took the top 80 genes based on the F-score ranking.
GO clustering

Clique Finding
GO clustering on Leukocyte differentiation time-series experiment

<table>
<thead>
<tr>
<th>Rank</th>
<th>Title</th>
<th>Score</th>
<th>probe sets</th>
<th>Genes</th>
<th>X1</th>
<th>n1</th>
<th>X2</th>
<th>n2</th>
<th>Bootstrap p-val (10000) bootstrap</th>
<th>Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>defense response</td>
<td>3.403</td>
<td>93454_at, 102424_at, 92286_g_at, 102401_at, 103033_at, 102745_at</td>
<td>lymphocyte antigen 66, small inducible cytokine A3, interleukin 4, interferon regulatory factor 1, complement component 4, T-cell receptor gamma</td>
<td>6</td>
<td>29</td>
<td>221</td>
<td>3163</td>
<td>0.0148</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>transcription regulation</td>
<td>4.04</td>
<td>102996_at, 103259_at, 102401_at, 100156_at, 92644_s_at, 94325_at, 93264_at, 07904_at, 101502_at</td>
<td>eleven-nineteen lysine-rich leukemia gene growth factor independent 1, interferon regulatory factor 1, mini chromosome maintenance deficient 5, myeloblastosis oncogene, pre B-cell leukemia transcription factor 1, sterol regulatory element binding factor 1, transcription factor 7, T-cell specific TG interacting factor</td>
<td>9</td>
<td>29</td>
<td>486</td>
<td>3163</td>
<td>0.0287</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>steroid metabolism, steroid biosynthesis</td>
<td>3.662</td>
<td>93264_at, 94325_at</td>
<td>sterol regulatory element binding factor 1, pre B-cell leukemia transcription factor 1</td>
<td>2</td>
<td>29</td>
<td>38</td>
<td>3163</td>
<td>0.0478</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>cell cycle control</td>
<td>3.05</td>
<td>99066_at, 92644_s_at</td>
<td>kit oncogene, myeloblastosis oncogene</td>
<td>2</td>
<td>29</td>
<td>113</td>
<td>3163</td>
<td>0.2757</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>cytoskeleton organization and biogenesis</td>
<td>3.592</td>
<td>103212_at, 98747_at</td>
<td>dynein, axon, heavy chain 11, Wnt responsive Cdc42 homolog</td>
<td>2</td>
<td>29</td>
<td>119</td>
<td>3163</td>
<td>0.2972</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>protein modification</td>
<td>3.265</td>
<td>93274_at, 99956_at, 104598_at</td>
<td>CDC-like kinase, kit oncogene, PTP non-receptor type 16</td>
<td>3</td>
<td>29</td>
<td>282</td>
<td>3163</td>
<td>0.4871</td>
<td>No</td>
</tr>
</tbody>
</table>
GO-guided expression clustering
Hierarchical clustering

Hierarchical clustering

GO-guided hierarchical clustering

Gene clusters where correlations between biological function and expression profile are both evident were identified by GO guided clustering.
Acknowledgements

John Martin
Melissa Cline
David Finkelstein
Tarif Awad
Michael Stewart
Michael Siani-Rose
David Kulp

Thank you!