Extracting Between Pathway Models From E-MAP Interactions Using Probabilistic Graph Summarization
David Kelley and Carl Kingsford
Dept. CS, CBCB, U. Maryland College Park

Methods: PGS

PGS takes the graph compression framework of GS and implements it over the distribution of graphs, defined by the edge probabilities. The general goal will be to merge nodes that share similar edges together into modules. Consider modules m and n and the distribution of the number of edges between them:

When there are many edges between m and n, they can be summarized by a single edge that represents a biclique. Then any edges missing from the bicliques are added as removal corrections, and any edges that are not summarized are added as addition corrections. The expected cost for a pair of modules m and n is:

\[ cost(m, n) = \sum_{e=0}^{\min(n_{mn}, e+1)} P(e) \min(n_{mn} - e, 1) \]

where \( n_{mn} \) is the number of possible edges between m and n. The global cost is computed by summing the edge costs over every pair of modules. A greedy algorithm that at every iteration merges the two supernodes that would most reduce the graph cost is used to minimize the cost function.

Results: BPM

We created a centroid expression vector for each module by taking the average of the expression values for genes in that module across all experiments. If the centroid vectors for a pair of modules designated as a BPM are correlated, we consider the BPM validated.

<table>
<thead>
<tr>
<th>BPMs</th>
<th>% Correlated</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGS</td>
<td>.129</td>
</tr>
<tr>
<td>Ullitsky et al.</td>
<td>153</td>
</tr>
<tr>
<td>Brady et al.</td>
<td>555</td>
</tr>
<tr>
<td>Bandyopadhyay et al.</td>
<td>208</td>
</tr>
</tbody>
</table>

Negative values in the E-MAP mean that the actual fitness of the double knockout was lower than expected and suggest aggravating interactions. Positive values mean that the actual fitness was higher and suggest alleviating interactions. Because these two interaction types indicate very different relationships between the genes, they are considered separately but simultaneously in the graph.

Values from both types of edges are mapped to the interval [0,1] using a logistic function so that significant interactions lead to high probabilities.

"1/3 of the E-MAP experiments fail, leaving many untested pairs of genes in the network. We account for this by reducing the term \( n_{mn} \) in PGS to count only tested pairs as having the potential for an edge. Because we also expect proteins in pathways to interact with each other, we only consider a merge between two supernodes if there is a physical interaction between some protein in each supernode.

To assess the biological plausibility of the modules, we report the number of GO terms enriched in ≥1 module and % of modules annotated with ≥1 term. We also used a set of gene expression measurements, aggregated from 132 studies, to test for correlation of expression among genes in a module.

Conclusions

We introduced a novel method for graph partitioning in weighted graphs and applied it to an exciting new type of genetic interaction data with the goal of identifying compensatory pathways in yeast. Our results are comparable in quality with the best previous attempts at this problem and return a greater quantity of BPMs, including many which have not previously been detected.

References