1 Comparison with Ge, et al.

Ge et al. (2001) propose a method for evaluating the correlation between the transcriptome and the interactome using gene expression data and protein-protein interaction (PPI) data. We will examine in particular the set of yeast cell-cycle data reported in Cho et al. (1998), along with the literature and yeast-two-hybrid data sets of protein-protein interactions collected by Ge et al. (2001). The cell-cycle expression data was divided into 30 clusters using k-means clustering. The clusters are reported at http://arep.med.harvard.edu/network_discovery. The literature data set contains 1666 interaction pairs, 335 of which correspond to genes in the cell-cycle data. The yeast-two-hybrid data set contains 1709 interaction pairs, 347 of which could be mapped to genes in the cell-cycle data set. The investigators also looked at a combined list of the two protein-protein interaction data sets; this resulted in 3222 unique pairs, 670 of which mapped to genes in the cell-cycle data set.

Ge et al. (2001) concluded that there is a significant correlation between the transcriptome and the interactome by using what they call a protein interaction density (PID). They form a 30×30 matrix M in which each row and column represents a cluster of the cell-cycle expression data. The squares in the matrix represent the PID between two clusters; specifically, M_{ij} represents the PID between genes in cluster i and cluster j, $i = 1, ..., 30$, and $j = 1, ..., 30$. For each squares, they calculated the number of observed interacting pairs (IP) in which one protein was the member of cluster i and the other was a member of cluster j. They also measured the total number of possible pair-wise combinations of protein pairs (PP) between clusters i and j, and then let PID=IP/PP. The off-diagonal squares represent between-cluster interactions and the squares on the diagonal represent within-cluster interactions. It was observed that the mean PID for the on-diagonal squares was higher than the off-diagonal squares. This difference in PID was determined to be statistically significant according to a cumulative binomial distribution. For the combined PPI data sets, there were 670 interactions that were observed in the data set, 117 of which corresponded to intra-cluster interactions. They calculated an expected number of intra-cluster interactions by assuming that the interacting pairs were drawn at random, and calculating a probability p of drawing an intra-cluster pair. Specifically, they find

$$p = \frac{\sum_{k=1}^{K}[n_k(n_k+1)/2]}{T(T+1)/2},$$

where K is the total number of clusters (here $K = 30$), n_k is the number of genes in cluster k, and T is the total number of genes in all clusters. They then compute a cumulative binomial probability for observing the number of reported intra-cluster interactions, or a more extreme number of interacting within-cluster pairs by

$$P(i > i_0) = \sum_{i=i_0}^{I} p^i (1-p)^{I-i} [\frac{I!}{i!(I-i)!}],$$

where I is the number of protein interaction pairs sampled, i_0 is the number of protein interaction pairs in the intracluster region, and p is described above.

Note that the calculation of p allows for homodimer interactions; that is, a protein interacting with itself. In their initial paper, Ge et al. (2001) allowed for these interactions.

This biases the calculations, however, since a protein is necessarily assigned to the same cluster as itself. They did publish a revision in which the homodimer interactions were removed (Ge et al., 2003). They did not state whether or not p was adjusted accordingly, but we assume that it was.

The cumulative binomial distribution does indicate a significant number of observed intra-cluster interactions for the literature data set with homodimers removed (42 observed interactions) with a p-value of 1.1×10^{-12}. This setting is actually better modeled with a hypergeometric distribution. Consider the possible interacting pairs to be balls in an urn. Suppose the balls representing intra-cluster interactions are red, and the balls representing inter-cluster interactions are white. For the literature data set, this would give 162070 red balls, and 4172970 white balls, for a total of 4335040 balls in the urn. We then know that we will randomly draw 315 balls from this urn. The probability of drawing 42 or more red balls can be calculated using the hypergeometric distribution. Specifically,

$$P_{\text{hypergeometric}}(i > i_0) = \sum_{i=42}^{315} \frac{\begin{pmatrix} 162070 \\ i \end{pmatrix} \begin{pmatrix} 4172970 \\ 315 - i \end{pmatrix}}{\begin{pmatrix} 4335040 \\ 315 \end{pmatrix}}.$$

Calculated according to the hypergeometric distribution, we arrive at a p-value of $P_{\text{hypergeometric}}(i > i_0) = 1.567496 \times 10^{-12}$. The conclusions are qualitatively the same as those using the cumulative binomial.

2 Graph representation of Ge, et al. data

References

Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D. & Davis, R. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* **2**, 65–73.

Ge, H., Liu, Z., Church, G. & Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genetics* **29**, 482–486.

Ge, H., Liu, Z., Church, G. & Vidal, M. (2003). Does mapping reveal correlation between gene expression and protein-protein interaction? -In reply. *Nature Genetics* **33**, 16–17.