An Algorithm to Find Co-Regulated Gene Clusters: Important Improvements

Ivy Jansen[1], Kerstin Koch[2] and Tomasz Burzykowski[1]

[1] Hasselt University, Center for Statistics, Belgium [2] University Paris-Sud, IBBMC, France

Detecting partial positive and negative coregulated gene clusters is an important task in microarray data analysis to get insight in the metabolism of organisms. Genes are clustered together if they show similar expression patterns under a number of conditions, assuming that they then are under the control of the same transcription factor and are related to a similar function in the cell.
To find interesting coregulated genes, the gene expression matrix is transformed into a binned matrix showing the pairwise changing tendencies between condition pairs (increase, decrease or no change) [1]. We propose a new approach to discretize the expression data, based on the SAM method [2], which is mainly used to analyse microarray experiments and detect significant genes. SAM can handle replicates for the different conditions, meanwhile accounting for the random variability present in gene expression data, and automatically corrects for multiple testing, since many genes and many conditions are involved in this process. Both issues are ignored in existing methods [1]. SAM also avoids the usage of a threshold arbitrarily chosen by the user to decide if a gene is differentially expressed or not. It only needs a prespecified false discovery rate (usually 5%).
Another important issue to consider, is the way the gene clusters are constructed from the binned matrix. Ji and Tan [1] define positive (similar behaviour under a number of condition pairs) and negative (opposite behaviour) coregulated gene clusters. However, their definitions are not symmetrical, so the clusters depend on the order in which the genes are processed. The output is also gene-centered, leading to multiple appearances of positive clusters containing more than one gene. An alternative is to construct a coregulation graph, in which genes with a similar gene expression pattern are clustered together at a vertex of the graph and negatively coexpressed clusters are connected by edges. Because genes may be coregulated by different transcription factors under different environmental conditions, our algorithm allows the same gene to fall into different clusters. Overlapping gene clusters are also allowed because coregulation normally takes place in only a fraction of the investigated condition pairs.
The new binning and clustering techniques are applied to the freely available Gene expression dataset GDS1804, containing 16 microarray experiments with expression levels from *E. coli* K12 cells at different time points after inducing an alternative sigma factor which plays a role in transcriptional regulation. Replicates are available. Several very interesting relationship were found, all of which can be interpreted biologically. Using the binning of [1] on these data (normalisation threshold 0.3) leads to different clusters, in which none of our biologically interpretable relationships appear. This underlines that coregulated genes can be found using the combination of our new binning and clustering techniques, where competing methods fail.

References

[1] Ji, L. and Tan, K. -L. (2004) Mining gene expression data for positive and negative co-regulated gene clusters. *Bioinformatics* **20**, 2711–2718.

[2] Tusher, V. G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences USA* **98**, 5116–5121.