An Algorithm to Find Co-Regulated Gene Clusters: Adjustment for Random Variability of Gene Expression Data.

Ivy Jansen[1], Kerstin Koch[2], Natalia Kwasnikowska[2], Tomasz Burzykowski[1]

Keywords: Gene expression data, Multiple testing, Significance Analysis of Microarray data

1 Introduction.

In literature, many methods can be found to extract clusters of co-regulated genes from gene expression data. However, all of them are restricted to situations without replications, and thus ignoring the random variability in this type of data. Since people are getting convinced of the presence of variability and start repeating their experiments several times, the existing methods are not satisfactory anymore. Therefore we propose an alternative to such an algorithm ([2]), accounting for the presence of random variability in gene expression data.

2 Material and Methods.

In a first phase, Ji and Tan [2] transform the gene expression matrix into a larger "binned" matrix capturing the changing tendency between pairwise conditions (increase, decrease or no change). However, this transformation uses a threshold that can be chosen completely arbitrary, and is not able to incorporate the variability when replicated measurements are available.

We adapt this first phase of the Ji and Tan algorithm [2] in such a way that the random variability is taken into account, and the threshold is based on a prespecified significance level, correcting for the multiple testing problem that is very common in analyses of gene expression data (one test for every gene and for every pairwise comparison). This correction is based on the technique called SAM (Significance Analysis of Microarray data) by [3], and basically uses the statistic

$$d_i = \frac{r_i}{\sqrt{s_i + s_0}}$$

with $r_i = \overline{x}_{i2} - \overline{x}_{i1}$ the difference in average gene expression between condition 1 and 2, s_i a standard deviation, and s_0 a fudge factor. Details can be found in [1]. This way, we can guarantee that much less false-positive clusters will be found, which is very important, since further investigating all clusters of co-regulated genes is very time-consuming.

Phases 2 and 3 are left unchanged.

We use a dataset of 500 genes for which the gene expression levels are measured under 3 conditions (20 replicates for condition 1, 15 replicates for conditions 2 and 3). This dataset is available from the SAM add-in in Excel.

[1]Hasselt University, Center for Statistics, Agoralaan Building D, 3590 Diepenbeek, Belgium. E-mail: `firstname.lastname@uhasselt.be`

[2]transnational University Limburg, School for Information Technology, Agoralaan Building D, 3590 Diepenbeek, Belgium. E-mail: `firstname.lastname@uhasselt.be`

3 Results.

To be able to compare the performance of our algorithm with the Ji and Tan algorithm [2], we need to summarize the replicates into one single observation per condition for the original algorithm. Doing so, and therefore using the mean value, results in 388 positive and negative co-regulated gene clusters (assuming threshold values as they are proposed in [2]). Inputting the data in our algorithm, allowing a false discovery rate of 5%, results in 38 gene clusters.

A remarkable fact is that the 38 gene clusters are not at all a subset of the 388 clusters found by the Ji and Tan algorithm. This is due to the fact that the multiple gene expressions were summarized into a single value, not taking into account the variability in those values, resulting in either a conclusion of difference in gene expression level while it is not, or in no difference in gene expression level while this difference might be (highly) significant.

4 Conclusion and Discussion.

It is clear from the results that our algorithm provides much less positive and negative co-regulated gene clusters. This is a real advantage for the gene network researchers, since this limits the work that needs to be done afterwards (screening again all genes that appear in a cluster). Also, the creation of the binned matrix is now based on statistical procedures that have been proven to be successful, and used in many other settings. Accounting for or ignoring the variability might result in completely different conclusions.

A next step is to set up a simulation study in which we include some known gene clusters, to see whether our algorithm finds more of those clusters than the original algorithm does. Work is still in progress, but looks very promising.

References

[1] Chu, G., Narasimhan, B., Tibshirani, R. and Tusher, V.G. 2001. SAM "Significance Analysis of Microarrays" Users guide and technical document. *Technical report, Stanford University.*

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

[3] 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.