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]

[1] Hasselt University, Center for Statistics, Agoralaan Building D, 3590 Diepenbeek, Belgium
[2] Hasselt University and transnational University of Limburg, Theoretical Computer Science Research Group, Agoralaan Building D, 3590 Diepenbeek, Belgium

Introduction

Ji and Tan [2] presented an algorithm to extract clusters of co-regulated genes from gene expression data. However, they assumed no replications, and thus ignored the random variability in this type of data. This can lead to a high rate of false-positive findings. Therefore, we propose a modification of the algorithm, that accounts for the presence of random variability.

Material

We use a dataset of 500 genes for which the gene expression levels are measured under three conditions:

- 20 replicates for condition 1
- 15 replicates for condition 2
- 15 replicates for condition 3

This dataset is available from the SAM add-in in Excel. It is assumed that these data already have been normalized.

Methods

Ji and Tan Algorithm

The algorithm proposed by Ji and Tan [2] consists of three steps:

1. Transform the gene expression matrix into a larger "binned" matrix capturing the changing tendency between pairwise conditions (increase, decrease or no change)
2. Extract the semi-co-regulated gene clusters (semi-CGCs)
3. Generate positive and negative co-regulated gene clusters (PNCGCs)

The first step is performed as follows:

- The gene expression data can be represented as an $O = n \times m$ matrix ($n =$ number of genes, $m =$ number of conditions)
- Transform O into an $O'' = n \times [m \times (m-1)]/2$ matrix

$$O''_{i,kj} = \begin{cases} \frac{O_{i,j}-O_{i,k}}{|O_{i,k}|} & \text{if } O_{i,k} \neq 0 \\ 1 & \text{if } O_{i,k} = 0 \text{ and } O_{i,j} > 0 \\ -1 & \text{if } O_{i,k} = 0 \text{ and } O_{i,j} < 0 \\ 0 & \text{if } O_{i,k} = 0 \text{ and } O_{i,j} = 0 \end{cases}$$

- Obtain O' from O'', by setting a normalization threshold t ($t > 0$)

$$O'_{i,kj} = \begin{cases} 1 & \text{if } O''_{i,kj} \geq t \\ -1 & \text{if } O''_{i,kj} \leq -t \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

In the second step, Ji and Tan constructed semi-CGCs as follows:

1. Gene i + : + (Gene Set A)@(Condition pairs where Gene i increases)
2. Gene i + : – (Gene Set B)@(Condition pairs where Gene i increases)
3. Gene i – : – (Gene Set C)@(Condition pairs where Gene i decreases)
4. Gene i – : + (Gene Set D)@(Condition pairs where Gene i decreases)

Semi-CGCs deliver information such as 'when the expression of a certain gene increases or decreases, what changing tendency the other genes may display accordingly'.

In the third step, they combine these semi-CGCs into

- Positive co-regulated gene clusters (PCGCs)
  [Gene $i \cup$ (Gene Set A $\cap$ Gene Set C)] @ (Condition pairs)
- Negative co-regulated gene clusters (NCGCs)
  [Gene i : (Gene Set B $\cap$ Gene Set D)] @ (Condition pairs)

Disadvantages of Ji and Tan's Algorithm

- Does not incorporate the variability when replicated measurements are available
- Arbitrary normalization threshold

Our Adaptation

We adapt the first phase of Ji and Tan [2] such that

1. Random variability is taken into account
2. 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 Tusher, Tibshirani and Chu [3]:

- Use repeated permutations of the data
- Calculate the statistic

$$d_i = \frac{r_{i,jk}}{s_{i,jk} + s_0}$$

for each permutation, where

- $r_{i,jk} = \overline{O}_{i,j} - \overline{O}_{i,k}$ is the difference in average gene expression between conditions j and k
- $s_{i,jk} = \sqrt{s_p^2(\frac{1}{n_j} + \frac{1}{n_k})}$ is a standard deviation, with s_p^2 the pooled standard deviation
- s_0 is a fudge factor

- Determine whether the value of the statistic of any gene is significantly different from zero or whether this value is observed by chance

More details can be found in Chu *et al.* [1].

This way, we can

- Consider replicates, and take the variability into account
- Control the number of false-positive clusters
  - very important, since further investigating all clusters of co-regulated genes is very time-consuming
- Choose this false discovery rate in advance

Steps 2 and 3 of the Ji and Tan algorithm are left unchanged.

Results

- Compare the performance of our algorithm with that of Ji and Tan [2]
  - for the latter, summarize the replicates into one single observation per condition, e.g. the mean value
    ⇒ results in 388 PNCGCs
  - for our algorithm, allow a false discovery rate of 5%
    ⇒ results in 38 PNCGCs
  - same thresholds as in [2]
- Remarkable that 38 gene clusters are not at all a subset of 388 clusters; possible explanation
  - 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 statistically significant
  - or in no difference in gene expression level while this difference might be (highly) significant
- Investigation of the effect of the 3 thresholds in Ji and Tan [2]

1. Precision threshold
   lowest permitted conditional probability $P(\text{Gene } X = a | \text{Gene } i = b)$ with $a = \{-1, 1\}$ and $b = \{-1, 1\}$

normalization	frequency	precision	# semi	# PNCGC
0.3	0	0	700	0
0.3	0	0.1	1400	412
0.3	0	0.2	1400	412
0.3	0	0.3	1400	412
0.3	0	0.4	1400	412
0.3	0	0.5	1400	412
0.3	0	0.6	1400	388
0.3	0	0.7	1400	388
0.3	0	0.8	1400	388
0.3	0	0.9	1400	388
0.3	0	1	1400	388

2. Frequency threshold
   lowest permitted probability $P(\text{Gene } i = 1)$ or $P(\text{Gene } i = -1)$

normalization	frequency	precision	# semi	# PNCGC
0.3	0	0.8	1400	388
0.3	0.1	0.8	1400	388
0.3	0.2	0.8	1400	388
0.3	0.3	0.8	1400	388
0.3	0.4	0.8	932	0
0.3	0.5	0.8	932	0
0.3	0.6	0.8	932	0
0.3	0.7	0.8	386	0
0.3	0.8	0.8	386	0
0.3	0.9	0.8	386	0
0.3	1	0.8	386	0

3. Normalization threshold
   t from eq.(1)

normalization	frequency	precision	# semi	# PNCGC
0	0	0.8	1504	471
0.1	0	0.8	1470	441
0.2	0	0.8	1432	411
0.3	0	0.8	1400	388
0.4	0	0.8	1330	330
0.5	0	0.8	1256	280
0.6	0	0.8	1188	232
0.7	0	0.8	1124	182
0.8	0	0.8	1070	146
0.9	0	0.8	1032	73
1	0	0.8	984	58

- Note that
  - choice of thresholds has a large influence on the number of CGCs found
  - use and effect of precision and frequency threshold can be explained
  - normalization threshold, however, is completely arbitrary
  - this normalization threshold does not appear anymore in our algorithm

Conclusion and Discussion

It is clear from the results that

- Controlling the false-positive rate provides much less positive and negative co-regulated gene clusters
- Creation of the binned matrix is now based on statistical procedures that have been adequately assessed, and used in many other settings
- Properly accounting for the variability in the data does influence the conclusions

A next step is to set up a simulation study to investigate the performance of our algorithm in more detail.

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.

◇To whom correspondence should be addressed: ivy.jansen@uhasselt.be