

ICI: A new approach to explore between-cluster relationships with applications to gene expression data

G. Dyson
Allergan
Irvine, CA 92612, U.S.A.
dyson_greg@allergan.com

C. F. J. Wu
School of Industrial and Systems Engineering,
Georgia Institute of Technology
Atlanta, GA 30332, U.S.A.
jeffwu@isye.gatech.edu

1 Abstract

Motivation: Clustering methods are used to group objects with similar patterns into one of k sets. In the case of gene expression data, genes clustered into the same set tend to have similar expression profiles. Spellman *et al.* (1998), Getz *et al.* (2000), and Horimoto and Toh (2001) applied different clustering techniques to gene expression. However these (and other) methods lack a way to quantify the between-cluster relationship. The proposed ICI method will ascertain the between-cluster relationship and the between-gene relationship for given pairs of clusters.

Results: We propose the Inter-Cluster Investigator (ICI) to establish the degree of co-regulation between clustered sets of genes. This will yield an alternative way of characterizing the between-gene relationship that builds upon the existing structure obtained from clustering. A method is developed to allow the identification of negatively and positively correlated sets of objects, which may indicate repression or regulation in gene expression data.

Availability: The ICI technique was developed in R (<http://cran.r-project.org/>). Source code is available from the first author upon request.

2 Introduction

The objective of this paper is to systematically determine the pairwise inter-cluster relationship. Specifically, are there any groups of objects that exhibit a positive or negative association with another group of objects but is not detectable by standard clustering? This type of analysis is readily interpretable for microarray data as the co-regulation of genes in clusters. Positively co-regulated genes tend to be in the same clusters. However, there is no current way to discern which cluster pairs are positively or negatively co-regulated. The methodology discussed in this paper will allow pairs of clusters of objects (i.e., genes for microarray data) to be deemed "associated". This explanation is most relevant when discussing data where such a relationship is expected, such as gene expression data. Additionally, the ICI procedure can determine if a subgroup of objects in one cluster is significantly related to a subgroup of objects in another cluster.

3 Methods

We start with a discussion of the measure used to determine the between-cluster relationship.

3.1 A robust measure of correlation

The main reason to choose a robust measure of correlation over the standard Pearson correlation is to avoid situations in which a few outlying values can grossly influence the correlation estimate. The robust measure is less efficient, but will give a more reasonable estimate when outliers are anticipated. One such robust measure is the Minimum Covariance Determinant (MCD) correlation estimator, introduced by Rousseeuw (1984). The objective of the MCD is to find $l = \lfloor \gamma n \rfloor$ observations whose covariance matrix has the smallest determinant. The γ value is typically between 0.60 and 0.80. In the case of two variables, this objective reduces

to minimizing $\sigma_X^2\sigma_Y^2 - \rho_{XY}^2$, in effect, to maximize the covariance between X and Y for l out of n observations. All possible groups of l points will be tested using a fast MCD algorithm (Rousseeuw and Van Driessen, 1999) if the number of replicates per group is small. The fitting algorithm can be found in the original paper or source code available at the R homepage, <http://www.r-project.org>.

Since implementation of the MCD in gene expression data is only two-dimensional, the paper will discuss attributes of the MCD in this context. As the MCD is location-scale invariant, the l points that minimize the determinant of the covariance matrix of two vectors X and Y are the same points that minimize the determinant of the covariance matrix of $Z = \frac{X-\mu_X}{\sigma_X}$ and $W = \frac{Y-\mu_Y}{\sigma_Y}$. Having all the objects on the same scale is preferable for cluster analysis since the weight assigned to each input object will be the same. Consequently, differences in mean or variance will not grossly affect the clustering result.

3.2 Measures of association between clusters

Here is an overview of the algorithm to determine the co-regulation between two clusters of objects. Since there will be a large number of observations in some of the groups, use of a single number to measure the similarity of two clusters will not be appropriate. Instead we determine the *distribution* of all pairwise MCD correlations between two clusters. Specifically, calculate

$$\text{distr}\{MCD(x_{im}, x_{jl})\}, \quad 1 \leq l < m \leq k, \quad 1 \leq i \leq N(m), 1 \leq j \leq N(l), \quad (1)$$

where k is the total number of clusters and $N(p)$ is the number of objects in cluster p . We calculate (1) as follows. Cluster all n objects into k groups. Then compute all possible pairwise correlations between objects in clusters i and j . Derive the empirical quantiles (for predetermined percentiles) from the distribution of pairwise correlation between clusters i and j . These percentiles are plotted against the resulting empirical quantiles to obtain a graph of the distribution function of the correlations between clusters i and j . Do this for all $\binom{k}{2}$ pairs of clusters.

Now that the distribution of the MCD correlations between each pair of clusters has been calculated, we will need to determine if the observed difference in quantiles is statistically significant as compared to the baseline. *By comparing the baseline to the observed distribution, we determine which clusters of objects are positively or negatively associated.* The baseline, F_0 is defined as the distribution of the correlation for all possible pairs of objects (genes). Therefore for gene expression data, we are comparing the distribution of the correlations between pairs of genes within a pair of clusters and the distribution of the correlations between all possible pairs of genes. This baseline will allow us to make substantive claims about the between-cluster correlation effect.

3.2.1 Analysis with percentage measure

The analysis with percentages involves calculating the proportion of correlations between two clusters that fall above and below given thresholds. The 5th and 95th percentiles of the baseline, F_0 , are used as the lower and upper threshold values, respectively. For each pair of clusters, we will compare the percentage of correlations that fall below the 5th percentile and above the 95th percentile with the expected percentage under the baseline, 5% for both. To call a pair of clusters *negatively associated*, it is necessary that the percentage of the extremely negative correlations be high (e.g. > 10%) and the percentage of the extremely positive correlations be low (e.g. 0%). *Positively associated* cluster pairs will have a high percentage of extremely positive correlations (e.g. > 10%) and a low percentage of extremely negative correlations (e.g. 0%).

3.3 Analysis outline of Inter-Cluster Investigator

Once it is determined which of the pairs of clusters are positively or negatively associated, it will be informative to know which objects (e.g., genes) are driving that relationship. A new diagnostic tool for cluster analysis, *Inter-Cluster Investigator* (ICI), is developed to ascertain the validity of a cluster partition. In addition, this tool will be used to determine whether a cluster partition is improvable. The ICI technique will focus on the between-cluster

relationship, specifically co-regulation of objects (genes) in two clusters. This explanation is most feasible when discussing data where such a relationship is expected, such as gene expression data. However, unexpected cluster relationships emerge in other data (see Dyson, 2004). Additionally, the ICI procedure will determine if a subgroup of objects in one cluster is significantly related to a subgroup of objects in another cluster.

Suppose each object is associated with a vector of observations and there is a clustering partition with $i = 1, \dots, n$ groups. The proposed method, dubbed *Inter-Cluster Investigator* (ICI), works as follows.

- 1 Compute the correlation between each object in cluster i and each object in cluster j .
- 2 Produce the *Map of Association Measures (MAM)*, a plot of the correlation values from step 1 that conveys the strength of the between-object relationship. This will involve assigning different colors to visualize the individual object-by-object correlations.
- 3 Reorder the objects in i based on their correlation with objects in j , using one of the methods described below in Section 3.3.1
- 4 Repeat step 3 for objects in j based on their correlation with objects in i .
- 5 Re-plot the figure from step 2 based on the re-orderings from steps 3 and 4. This is referred to as an *enhanced MAM plot*.
- 6 Examine the plot for any interesting patterns.

The correlation measure employed in step 1 is the MCD, but any other reasonable measure may be used. The example given below will illustrate in detail each of the above steps.

3.3.1 Object re-ordering schema

Here are several methods that can visually display the relationship between two clusters, denoted as i and j .

- *Mean Adjustment Plot*: For each object in cluster i , calculate the average correlation with the objects in cluster j . For every object in cluster j , calculate the average correlation with the objects in cluster i . The reorderings will be based on the ranks of these averages.
- *Unsigned Mean Adjustment Plot*: For each object in cluster i , calculate the average unsigned correlation with the objects in cluster j . For every object in cluster j , calculate the average unsigned correlation with the objects in cluster i . The reorderings will be based on the ranks of these averages.
- *Positive Mean Adjustment Plot*: For each object in cluster i , calculate the average positive correlation with the objects in cluster j . That is, consider only the correlations that are greater than 0. For every object in cluster j , calculate the average positive correlation with the objects in cluster i . The reorderings will be based on the ranks of these averages.
- *Negative Adjustment Plot*: It is similar to the Positive Mean Adjustment Plot, except that it considers only the correlations that are less than 0.

The mean adjustment plot is the optimal enhanced MAM plot since it will separate the positive and negative association regions. The unsigned adjustment plot allows one to compare the strength between objects with a positive relationship and those with a negative relationship. However, it is less powerful in finding significant subsets of objects in the two clusters that drive the (positive or negative) relationship. The positive and negative adjustment plots focus only on half of the associations. See Dyson (2004) for details on the comparison of these plots.

4 Example

The ICI method will be applied to a microarray time course data. Since the yeast genome has been well studied, there is a vast amount of available biological information. This will

aid in interpretation and identification of clusters and cluster sub-groups. We examine a subset of genes chosen because of their cyclic behavior that act as a positive control.

Spellman *et al.* (1998) published an online data base <http://genome-www.stanford.edu/cellcycle> of 73 microarray time-series experiments of the yeast genome synchronized under four different experimental conditions (cdc15, cdc28, alpha factor, elutriation Elu). The genes are synchronized in the same state by introducing external substances, changing environmental conditions, or selecting cells of the same size (Filkov *et al.*, 2001). From these experiments, Spellman *et al.* (1998) chose a set of 800 genes which they called cell-cycle regulators, based on their periodicity and correlation algorithms. These 800 genes exhibited the most cyclic behavior across all the experiments. The analysis will focus on these 800 genes, which allows comparisons to be made with the original analysis.

4.1 Cluster identities for yeast data

We will use a simple method to assign each cluster a genetic function, based on a web-based yeast cluster interpreter, developed by Robinson *et al.* (2002). Given a set of yeast genes, their algorithm produces a summary of functional classes, cellular localizations, protein complexes, etc. that are enriched in the list, using existing annotations from the Munich Information Center for Protein Sequences (MIPS) and Gene Ontology Consortium (GO) databases (<http://funspec.med.utoronto.ca/>). A cluster of genes is said to be *functionally enriched* for an attribute if the proportion of genes within the cluster known to have that attribute exceeds the number expected from chance.

4.2 Data filtering

It is necessary, as with most gene expression data, to clean the data into a useable form. This usually involves background adjustment, normalization, and imputation. For the 800 genes, the following steps were undertaken to ensure the validity of the data. The data was already normalized by the authors. However, there is a large number of missing values. Any gene that had more than two consecutive missing values in any of the four experiments was

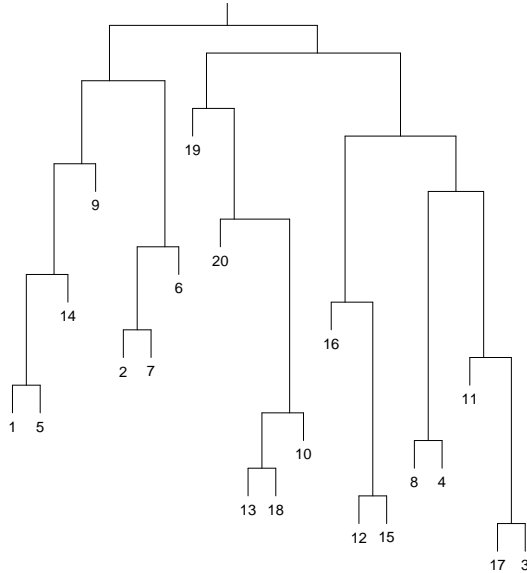


Figure 1: Schematic dendrogram of 624 gene yeast data

eliminated from consideration. The rest of the missing data for each gene were imputed as the mean of the previous and next observations in time cycle for that experiment. The 0 time point was also eliminated in the *cdc28*, alpha factor, and elutriation Elu experiments. Then we use hierarchical clustering (HC) to produce a 20-group HC using Euclidean distance and complete linkage. Eliminate the genes in any groups with less than 4 members and redo the clustering. Continue the process until no more genes are eliminated. After the filtering process, there are 624 out of the original 800 genes that are suitable for analysis. The final cluster partition, with 20 groups using 624 genes, is used in the later stages of the analysis and displayed in Figure 1. This subset of the original data will be referred to as the yeast-624 data.

Using the approach outlined in Section 4.1, each of the 20 clusters can be roughly categorized into functional groups, using the GO and MIPS biological process indices. The cluster functions are listed in Table 1. If there is no discernable functionality for a group, it is listed as NA. Additionally, the N in Table 1 refers to the size of the cluster, while the known genes column gives the number of genes with a known function. The later analysis will focus on

Table 1: Cluster functions for the yeast-624 data

Cluster	Function	N	Known Genes	Major Genes
1	Transporter	163	108	FUN26 ATR1 ANT1
2	Amino acid metabolism	47	40	MET3,6,14,17,28 STR3
3	DNA processing	94	67	CDC2,45 POL1 DUN1 SPO16
4	DNA recombination and repair	55	32	POL2 PRI2 CLB5 RAD5,51 RFA1
5	Drug transporters	27	19	SIT1 ARN1 ENB1
6	Histones	10	10	HTB1,2 HTA1,2 HHF1,2 HHT1,2 HHO1 PSA1
7	NA	14	12	PET9 ZRT1 PDR5 SIM1
8	DNA synthesis and replication	17	13	POL12 POL30 CLB6 RNR3 CLN2
9	NA	7	5	CLB1 CDC5 CHS2 HOF1 TPO3
10	Pre-replication complex	14	12	MCM3 CDC6 CDC46
11	Fungal cell differentiation	83	61	AFR1 EXG2 SLT2 YCK1
12	Pheromone response	12	10	HST4 STE2 AGA2 MFA2
13	NA	4	4	CST13 BUD9 RME1 PRY3
14	Amino acid metabolism	12	11	AGP1 ARO9 GAP1 PUT1 GCV2
15	Pheromone response	8	8	FUS1 KAR4 SST2 MDG1
16	Cytoplasm	5	4	GLK1 TSL1 ARG1 AGA1
17	C-compound and carbohydrate utilization	38	28	PMT5 PMT1 CWH41 PGM1 FKS1 WSC2
18	NA	5	4	PST1 FAR1 DSE3 NIS1
19	Cytokinesis, completion of separation	4	4	SCW11 DSE2 CTS1
20	Cell growth / morphogenesis	4	3	ASH1 EGT2 PIR1

the 10 (out of 20) groups that consist of at least 14 genes. In general, it is difficult to find a significant group function when there is a small number of genes in a group.

Most clusters have more than one significant functional group. The most prevalent and specific ones are listed in Table 1. The other significant functions are usually highly related to those listed in Table 1. This happens because these functions are hierarchical. For example any set of genes with the function of amino acid metabolism automatically has the functions of metabolism, physiological processes and biological processes (cluster 2). Some clusters are extremely tight, like the histone group (cluster 6); while others are more diverse (clusters 1 and 11). For illustrative purposes, the time plots for each cluster for the *cdc15* experiment

in Figure 2 are examined. Clusters with smaller number of genes are more likely to have similar time paths. Except for clusters 1 and 11, each cluster has a distinct cycle. Unlike the ICI method, examining these plots alone will not yield any insight into the between-cluster relationship.

The MCD correlation, discussed in Section 3.1, using 90% of the available data is employed for this analysis. Therefore, for the analysis of the yeast-624 data, the MCD will select 63 out of 70 observations to estimate the correlation between a pair of genes.

4.3 Determination of the pairwise cluster relationships

For this small data set, we proceed with the analysis described in Section 3.3 and determine which pairs of clusters are positively and negatively associated. Figure 3 shows the plot of the quantile functions of each of the $\binom{10}{2} = 45$ cluster pairs. The solid black curve indicates the baseline. It is clear that there are examples of both positive and negative relationships between cluster pairs. Cluster pairs corresponding to curves on the left-hand side of the graph are negatively related, while those on the right-hand side are positively related.

The objective is to determine which pairs of clusters are significantly different. The results using percentages measure is displayed in Table 2. Section 3.2.1 describes the components of the table and the methodology for selection of negative and positive associated clusters. Negatively associated or dissimilar cluster pairs found using percentages are $\{1, 4\}$, $\{1, 8\}$, $\{2, 10\}$, $\{4, 5\}$, $\{4, 7\}$, $\{5, 8\}$, $\{5, 17\}$, $\{7, 8\}$, $\{7, 10\}$. Positively associated or similar cluster pairs found using percentages are $\{3, 4\}$, $\{3, 8\}$, $\{3, 17\}$, $\{4, 8\}$, $\{4, 17\}$, $\{8, 17\}$. Positively associated cluster pairs are indicated with a "+" in Table 2, while negatively associated cluster pairs are indicated with a "-".

Note that cluster 8 appears most often in the lists. This can be attributed to the high number of known cell cycle regulatory genes in that group, including POL12, POL30, CLB6, CLN1, and CLN2. These analysis also confirm the assumption underlying the use of hierarchical clustering, namely, two terminal cluster nodes with the same immediate parent are

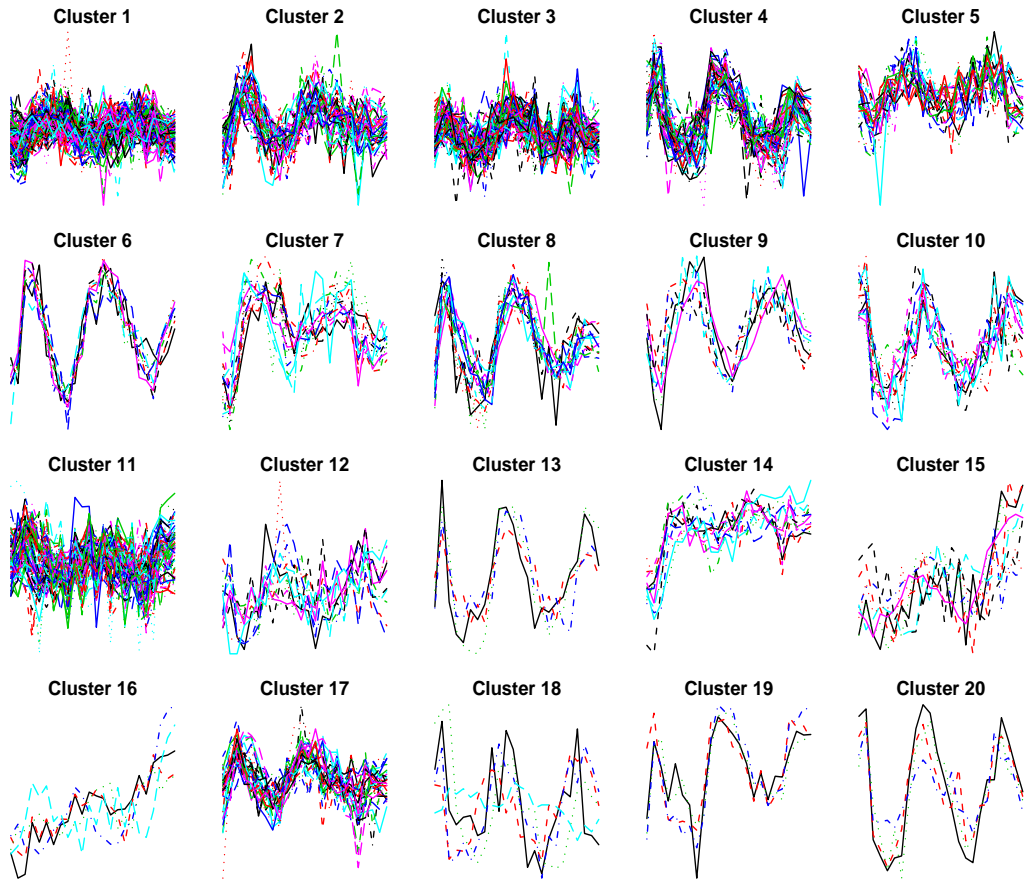


Figure 2: Time plots of yeast *cdc15* data for each cluster

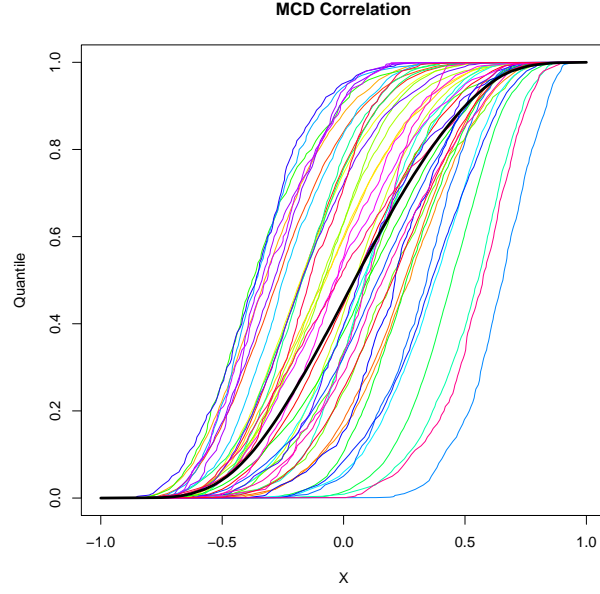


Figure 3: MCD quantile functions of all 45 cluster pairs

Table 2: Percentages of low and high extremes for cluster pairs for the yeast-624 data

Clus. Pair	Lower	Upper	Clus. Pair	Lower	Upper	Clus. Pair	Lower	Upper
1 2	0.027	0.024	2 11	0.038	0.048	5 7	0.000	0.045
1 3	0.079	0.000	2 17	0.000	0.058	-5 8	0.303	0.000
-1 4	0.193	0.000	+3 4	0.000	0.193	5 10	0.011	0.034
1 5	0.003	0.049	3 5	0.097	0.000	5 11	0.079	0.002
1 7	0.000	0.079	3 7	0.052	0.000	-5 17	0.204	0.000
-1 8	0.253	0.000	+3 8	0.000	0.385	-7 8	0.197	0.000
1 10	0.073	0.007	3 10	0.006	0.007	-7 10	0.240	0.000
1 11	0.053	0.008	3 11	0.011	0.013	7 11	0.055	0.028
1 17	0.091	0.001	+3 17	0.000	0.122	7 17	0.017	0.008
2 3	0.013	0.011	-4 5	0.121	0.000	8 10	0.008	0.000
2 4	0.053	0.011	-4 7	0.242	0.000	8 11	0.013	0.075
2 5	0.065	0.000	+4 8	0.000	0.631	+8 17	0.000	0.438
2 7	0.008	0.096	4 10	0.000	0.062	10 11	0.082	0.037
2 8	0.008	0.041	4 11	0.021	0.048	10 17	0.056	0.000
-2 10	0.302	0.000	+4 17	0.000	0.151	11 17	0.004	0.056

highly positively associated. Cluster pairs that fit this description (which are included in the analysis) from the dendrogram in Figure 1 are $\{3, 17\}$, and $\{4, 8\}$. In addition, clusters 3, 4, 8, and 17 are all positively associated with each other. The functions of these clusters are related to DNA processing. Thus it can be concluded that clusters 3, 4, 8 and 17 form a *class* within the HC framework. Notice that cluster 5 is negatively associated with nearly all of the clusters in the class. It can be inferred that genes with function transporter are negatively associated with the genes with function DNA processing.

Once pairs of clusters are declared significant, then the analysis proceeds by following the outline in Section 3.3. For illustrative purposes, we will focus on the relationship between clusters 5 and 8. Figure 4(a) displays the MAM plot. There is, as expected, no apparent pattern in this plot. Each square in the plot represents an MCD correlation between a gene from cluster 5 and a gene from cluster 8. The colors represent quantile ranges from the empirical distribution of the MCD correlations for 63 out of 70 observations. These two clusters are negatively associated, as evidenced by the large number of red and maroon points (less than the 5th percentile). Black and dark gray points represent extremely positive correlations. Other colors represent correlations in the middle 90% of the distribution.

Using the methods in Section 3.3.1, enhanced MAM plots are produced. The output should be simple to use and understand, especially for the target audience of biologists. The plot in Figure 4(b) displays the enhanced MAM plot using the unsigned mean adjustment method. It is easily seen that the genes that gravitate towards the upper right-hand corner (high values for the row and column indices) are mostly negative pairs of genes. This plot shows that the strength of the positive associations do not approach the strength of the negative associations. Figure 4(c) shows the enhanced MAM using the average mean adjustment method. In general, the lower left-hand corner will contain the extremely negatively associated gene pairs, while the upper right-hand corner will contain the extremely positively associated gene pairs. The analysis will proceed using the mean adjustment since the average correlation provides a visually clear result. However, the unsigned mean adjustment is valuable because

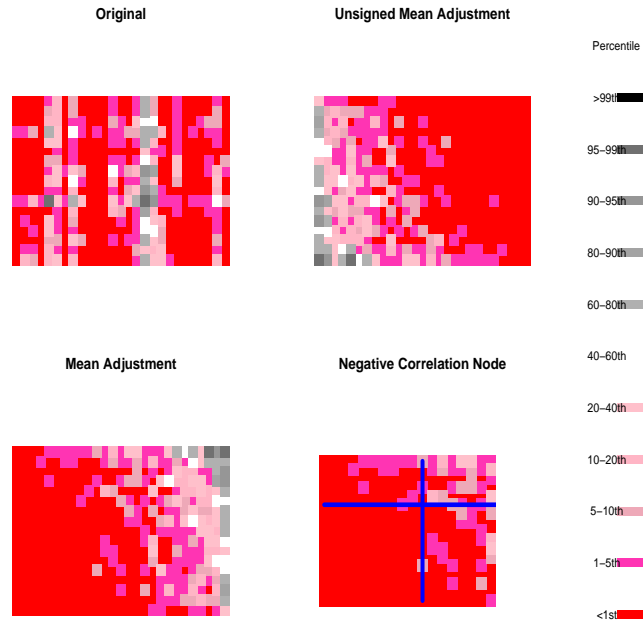


Figure 4: Analysis of negatively associated cluster pair for 624 gene yeast data

it allows the relative strength between mostly positive and negative rows and columns to be seen. For example, the strength of correlations of the cluster 5 gene (column) that is positively associated (black) in Figure 4 to many cluster 8 genes is much less when compared to the associations of the other genes in cluster 5. This can be readily seen by examining the enhanced MAM plot using the unsigned mean adjustment. The final plot Figure 4(d) is a zoom-in look at the lower left-hand corner of the enhanced MAM using mean adjustment.

A byproduct of this diagnostic tool is that interesting sub-nodes in the plot can be identified. A *sub-node* refers to subsets of genes from the two clusters that exhibit a strong positive or negative association between each other. For example, Figure 4(d) shows a typical negative sub-node enclosed by the black lines. All of the 10 genes from cluster 5 are negatively correlated with the 11 genes from cluster 8. Out of the 10 genes in cluster 5 (x-axis), 7 have a known function. The major genes in this group are SIT1, ARN1 and ENB1, which play a role in siderochrome transport. CDC47 (cell division control protein) is also included in

this subgroup, leading one to believe that it is somehow regulating siderochrome transport. Of the 11 genes in cluster 8 (y-axis), 8 have a known function. The major regulating genes in this group are CLB6, CLN1 and CLN2, which play a role in regulation of CDK (cyclin-dependent protein kinase) activity. CLN1 and CLN2 are believed to activate CLB6 (see <http://www.geneontology.org/>).

The Gramene project (<http://www.gramene.org>) defines siderochrome transport as the directed movement of siderochromes, low molecular weight Fe(III)-chelating substances, into, out of, or within a cell. In particular ARN1 is "transcriptionally activated under conditions of iron deprivation" (Yun *et al.* 2000). The results would also seem to imply that CDC47 is regulating these iron transports. In addition, CLN1 and CLN2 might repress these transports when CDK activity is needed. The assumption here is that these major genes are the driving force behind the results of the cluster analysis. In addition, genes RSB1 and PMA1, from cluster 8, have the function of ATPase activity, coupled to transmembrane movement of ions. These genes directly drive the transport of ions (including siderochrome iron) across a membrane and into the cell (<http://db.yeastgenome.org/cgi-bin/SGD/GO/go.pl?goid=42625>). Thus it seems apparent that in order for CDK activity (DNA synthesis) to proceed, siderochrome iron needs to be present.

Next, a pair of clusters (4 and 8) that are positively associated is examined. From Table 1, the function of cluster 4 is DNA recombination and repair while the function of cluster 8 is DNA synthesis and replication. *A priori*, we would expect these two clusters to be highly positively associated, which is validated by the analysis. Figure 5(a) displays the MAM plot. Clearly, most pairs of genes have an extremely positive relationship. The enhanced MAM plot using the unsigned mean and mean adjustments are shown in Figure 5(b) and (c). Again, the analysis will focus on the enhanced MAM plot using the mean adjustment. The positive sub-node is enclosed by the blue lines in Figure 5 (d). All 17 genes from cluster 8 and 36 genes from cluster 4 are included within this sub-node. As stated above, the 13 of 17 known genes in cluster 8 are enriched for the function DNA synthesis and repair. Among

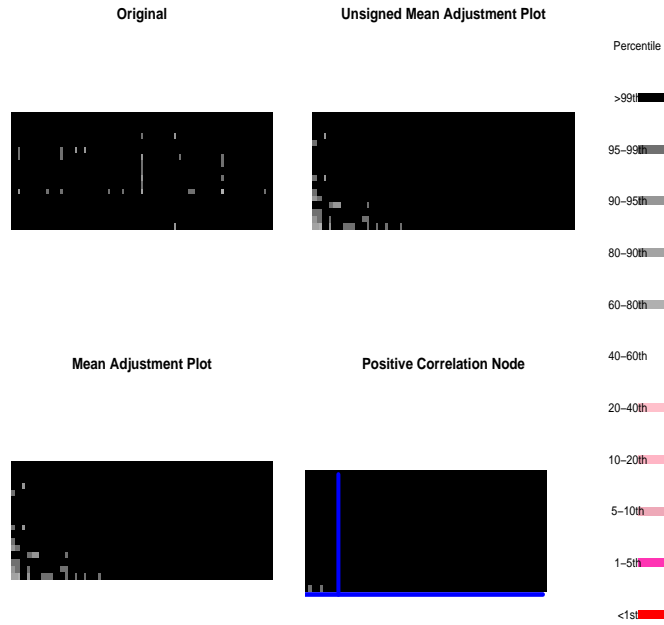


Figure 5: Analysis of positively associated cluster pair for 624 gene yeast data

the 36 genes from cluster 4, 26 have an unknown molecular function. However, we can assume that they are active in the S phase of mitotic cell cycle, since most known genes in that cluster are active in that phase. When more information is available on these unknown genes, some substantive determination on the sub-node relationship can be made. Therefore, we conclude that clusters 4 and 8 should not have been split and would join them for any further analysis, because of the extremely positive between-cluster relationship.

5 Concluding Remarks

In this paper a new technique is developed to determine and quantify the relationship between pairs of clusters. It gives a cluster-level summary of gene-to-gene associations. Furthermore, the method identifies negatively and positively associated pairs of clusters of genes. The ICI procedure will then identify subgroups of genes that drive the association between two clusters. In many cases, it is not possible to assign a function or an identifier for a cluster.

When this is possible, checking the performance of the ICI algorithm is feasible. For most genetic data, this type of "control data" is not available. However, the yeast genome and cell cycle is probably the most studied culture. Thus, some known negative relationships between cluster of genes should be found by the ICI procedure (e.g., DNA synthesis and mitosis). However, other interesting relationships that have not been studied or thought of can also appear (e.g., DNA synthesis and iron transport). The ICI method will shed meaningful insights into complex systems that may otherwise be overlooked.

6 Acknowledgements

This research was supported in part by NSF grant DMS 0305996.

References

- [1] Dyson, G. (2004) *New Techniques in Clustering and Microarray Data Analysis*. Ph.D Thesis, University of Michigan.
- [2] Filkov, V., Skiena, S., Zhi, J. (2001), Analysis techniques for microarray time-series data. *RECOMB 2001*, 124-131.
- [3] Robinson, M. D., Grigull, J., Mohammad, N., Hughes, T. R.,(2002) FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, **3**, 35.
- [4] Getz, G., Levine, E., Domany, E., Zhang, M. Q., (2000) Super-paramagnetic clustering of yeast gene expression profiles. *Physica A*, 457-464.
- [5] Horimoto K. and Toh H. (2001), "Statistical estimation of cluster boundaries in gene expressions profile data." *Bioinformatics*, **17**, 1143-1151.
- [6] Rousseeuw, P. J., (1984) Least median of squares regression. *Journal of the American Statistical Association*, **79**, 871-880.

- [7] Rousseeuw, P. J., Van Driessen, K.,(1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212-223.
- [8] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Bolstein, D., Futcher, B., (1998) Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Molecular Biology of the Cell* **9**, 3273-3297.
- [9] Yun, C-W., Tiedeman, J. S., Moore, R. E., Philpott, C. C., (2000) Siderophore-Iron Uptake in *Saccharomyces cerevisiae*. *Journal of Biological Chemistry*, **275**, 16354-16359.