

Order Restricted Clustering for Dose-response Trends in Microarray Experiments

Adetayo Kasim¹, Dan Lin¹, Ziv Shkedy¹,
An De Bondt², Willem Talloen²,
Hinrich Goehlmann² and Luc Bijmens²

1- Hasselt University, Center for Statistics, Biostatistics, Universitaire Campus, Building D,
B 3590 Diepenbeek, Belgium

2- Johnson & Johnson PR&D, Turnhoutseweg 30, 2340, Beerse, Belgium,

Abstract

Dose-response microarray experiments are designed to monitor expression profiles of thousands of genes with respect to increasing dose of certain treatments. It is primarily done to establish a dose-response relationship between gene expression and a compound concentration. The main interests are determination of minimum effective dose and the identification of the shape of the dose-response curve. There exist several methods to test for significantly monotonic trends (Lin *et al.*, 2007), and there is also an information based algorithm to classify the dose-response relationship (Lin *et al.*, 2008). In this paper, we propose an order restricted clustering approach to find clusters of genes with similar dose-response profiles. The method offers potentials to gain insight into hidden structure in a dose response microarray data and can provide an effective exploratory and fast tool for visualization of dose-response microarray data. The proposed method is motivated by the δ - biclustering introduced by Cheng and Church (2000) and can be seen as an extension of the δ - biclustering in the setting the parameters are assumed to be ordered.

Keywords: Dose-response curve; Microarray; δ - biclustering; isotonic regression.

1 Introduction

Dose-response studies are common experiments in the early drug development, as they provide important information of the biological activity of a chemical compound. In such studies the response of primary interest is measured at several increasing dose levels, with the first dose level being typically a control group with zero dose. In recent years, dose-response studies were extended to the microarray setting, in which the arrays are administered to measure intensities of thousands of genes. The goal of the such an experiment is to identify genes whose expression is affected by dose.

Recently, Lin *et al.* (2007) discussed several testing procedures, namely Williams' (1971, 1972), Marcus' (1976), the likelihood ratio test (Barlow *et al.* 1972), the M test (Hu *et al.* 2005), and the modified M (Lin *et al.* 2007) that can be used to identify genes with a monotonic relationship between gene expression and doses. To gain insight into the nature or the shape of the dose-response relationship, Lin *et al.* (2008) proposed to classify each gene into a finite set of possible dose-response trends using model selection procedure based on information theory proposed by Burnham and Anderson (2002). Assuming a monotonic relationship, the dose-response curve could be either linear, nonlinear, concave or convex. Furthermore, for an experiment with K dose levels, there is a fixed number of monotonic models, which can be fitted. For instance, in a dose-response experiment with 4 dose levels, upon the establishment of a monotonic relationship between gene expression and doses, there is a set of 7 models, that can be fitted for an increase trend. These models are shown in Table 1 and Figure 1. For a given set of candidate models the Order Restricted Information Criterion (*ORIC*), the Akaike Information Criterion (*AIC*), and the Bayesian Information Criterion (*BIC*) were used by Lin *et al.* (2007) to calculate the posterior probability of each model in the set. The model with the highest posterior probability is selected. The information criteria take into account both the goodness-of-fit and model complexity. Hence, for each gene, the selected dose-response curve represent the best compromise between goodness-of-fit and model complexity.

Figure 1 about here.

In this paper, we focus on a clustering method for dose-response microarray data in order to find clusters of genes with a similar dose-response relationship under monotonic constraints. Our proposal is motivated by the δ -biclustering algorithm proposed by Cheng and Church (2000). A biclustering is an algorithm that simultaneously cluster both genes and condition of a microarray data. A biclustering aims at unfolding the local structures or patterns in a microarray data. We refer to Madeira and Arlindo (2004) for a review of the biclustering methods for microarray data. Cheng and Church (2000) defined a bicluster as a subset of genes and a subset of conditions with a high similarity score. Similarity is a measure of coherence of the genes and conditions in a bicluster, using mean squared residual score. A bicluster is therefore, a subset of genes and a subset of conditions whose mean squared residue score is less than a pre-specified value - δ .

Within the dose-response microarray setting, the aim is to cluster genes into non-overlapping cluster while focusing on local structures that may be present in such data. As a result, we proposed a node deletion algorithm, called order restricted clustering for dose-response microarray experiments (ORCME) that find cluster of genes with a similar dose-response relationship using mean squared residue score as a measure of similarity. Note that in contrast with the δ -biclustering method of Cheng and Church (2000), which aim to find a subset of genes and conditions which form a cluster, with the dose-response microarray setting the aim is to find subset of genes when the conditions (dose levels) are held fixed.

The contents of the paper is organized as follow, in section 2, a description of data acquisition is given. In Section 3 we briefly discuss the global likelihood test for monotonic trend proposed by Barlow *et al.* (1972) and used in Lin *et al.* (2007) to identify genes for which there is a monotone relationship of gene expression and dose. After the initial inference step, in Section 4 we discuss the methods for the order restricted clustering; specifically, in subsection 4.1 we give a brief description of the δ -biclustering as proposed by Cheng and

Church (2000) , while in subsection 4.2, we adapt the method specifically for the dose-response microarray setting. We apply the proposed methods to our case study in Section 5. Section 6 concludes the paper with a discussion.

2 Data Acquisition

Human epidermal squamous carcinoma cell line A431 was grown in Dulbecco's modified Eagle's medium, supplemented with Lglutamine (20 mM), Gentamycin (5 mg/ml) and 10% fetal bovine serum. The cells were stimulated with EGF (R&D Systems, 236-EG) at different concentrations (0 ng/ml, 1 ng/ml, 10 ng/ml and 100 ng/ml) for 24h. RNA was harvested using RLT buffer (Qiagen). All microarray related steps including the amplification of total RNAs, labeling, hybridization and scanning were carried out as described in the GeneChip Expression Analysis Technical Manual, Rev.4 (Affymetrix 2004). Biotin-labeled target samples were hybridized to human genome arrays U133 A 2.0 containing probe sets interrogation approximately 22,000 transcripts from the UniGene database (Build 133). Hybridization was performed using 15 μ g of cRNA for 16 h at 45^oC under continuous rotation at 60 rpm. Arrays were stained in Affymetrix Fluidics stations using streptavidin/phycoerythrin staining. Thereafter, arrays were scanned with the Affymetrix scanner 3000, and images were analyzed using the GeneChip Operating System v1.1 (GCOS, Affymetrix). The collected data were quantile normalized in two steps: first within each sample group, and then across all sample groups obtained (Bolstad *et al.* 2002). The resulting data set consists of 12 samples (for three arrays at four dose levels) with 16,998 probe sets. For simplicity, we refer to probe sets as genes through our paper (Hubbell *et al.* 2002).

3 Initial Filtering: Testing No Dose Effect Against Ordered Alternatives Using the Likelihood Ratio Test

Our main interest is to identify clusters among genes for which monotone trend can be detected. Therefore, we carried out an initial filtering using likelihood ratio test. We consider a dose-response microarray experiments described in Section 2, in which the first dose level of the experiment is a control (zero dose). We formulate a gene specific linear model of the form

$$Y_{jk} = \mu(d_j) + \varepsilon_{jk}, \quad \varepsilon_{jk} \sim N(0, \sigma^2), \quad j = 0, 1, 2, 3, \quad k = 1, 2, 3. \quad (1)$$

Here Y_{jk} is the k th gene expression at the j th dose level for ; d_0, d_1, d_2, d_3 are the four dose levels; and $\mu(d_j)$ is the mean gene expression at dose level d_j . We further assume that gene expression increases or decreases with doses, although not necessarily in a linear fashion. The null hypothesis of no dose effect is given by

$$H_0 : \mu(d_0) = \mu(d_1) = \mu(d_2) = \mu(d_3), \quad (2)$$

and the alternative, under the assumption of a monotonic increasing trend, is

$$H_1^{Up} : \mu(d_0) \leq \mu(d_1) \leq \mu(d_2) \leq \mu(d_3), \quad (3)$$

with at least one strict inequality. Note that the direction of the trend is unknown in advance. Hence, for a monotonic decreasing trend the alternative is

$$H_1^{Down} : \mu(d_0) \geq \mu(d_1) \geq \mu(d_2) \geq \mu(d_3), \quad (4)$$

with at least one strict inequality. Both Barlow *et al.* (1972) and Robertson *et al.* (1998) discussed a procedure to test the equality of the ordered means. Let $\hat{\mu}^* = (\hat{\mu}_0^*, \hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\mu}_3^*)$ be the maximum likelihood estimates for the means under the ordered alternatives. Barlow *et al.* (1972) and Robertson *et al.* (1998) showed that $\hat{\mu}^*$: the non parametric maximum likelihood estimator under order constrains, can be estimated by isotonic regression. The likelihood ratio test statistic is given by:

$$\Lambda_{01}^{\frac{2}{N}} = \frac{\hat{\sigma}_{H_1}^2}{\hat{\sigma}_{H_0}^2}, \quad (5)$$

where

$$\hat{\sigma}_{H_0}^2 = \frac{1}{N} \sum_j \sum_k (y_{jk} - \hat{\mu})^2 \quad \text{and} \quad \hat{\sigma}_{H_1}^2 = \frac{1}{N} \sum_j \sum_k (y_{jk} - \hat{\mu}_j^*)^2.$$

Here $\hat{\mu}$ is the overall mean, N is the total number of arrays, $\hat{\sigma}_{H_0}$ and $\hat{\sigma}_{H_1}$ are the parameter estimates for variance under the null and the alternative hypotheses, respectively. The null hypothesis is rejected for a “small” value of $\Lambda_{01}^{\frac{2}{N}}$. Or, equivalently, H_0 is rejected for a large value of \bar{E}_{01}^2 , where

$$\bar{E}_{01}^2 = 1 - \Lambda_{01}^{\frac{2}{N}} = \frac{\sum_{jk} (y_{jk} - \hat{\mu})^2 - \sum_{jk} (y_{jk} - \hat{\mu}_j^*)^2}{\sum_{ij} (y_{jk} - \hat{\mu})^2}. \quad (6)$$

In order to estimate the parameters using isotonic regression one needs to know the direction of the trend. In practice, one can maximize the likelihood twice: for a monotonic decreasing trend and for a monotonic increasing trend, and choose the trend with a higher likelihood. In practice, thus, we can calculate \bar{E}_{01}^2 for each direction and choose the higher value of \bar{E}_{01}^2 (Barlow *et al.*, 1972). For an elaborate discussion on other testing procedures we refer to Lin *et al.* (2007).

3.1 Application to The Data

Following Lin *et al.* (2007) the likelihood ratio test statistics \bar{E}_{01}^2 was applied to the data. Raw p-values for the genes were calculated using resampling based techniques. Multiplicity adjustment was applied by using the Benjamini-Hochberg (BH) procedure for controlling the False Discovery Rate (FDR, Benjamini and Hochberg, 1995). Genes for which the adjusted p -values are smaller than 0.05 are declared differentially expressed (Ge *et al.*, 2003). In total the null hypothesis was rejected for 3,499 out of the 16,998 genes that were tested. These genes were used in the second stage of the analysis in which the primary interest is to identify clusters of genes with similar monotonic dose-response curve shape. We present in Figure 2 examples of genes with a significant monotonic dose-response relationship. The solid lines are the isotonic means, while the scatter points are the observed data for each dose.

4 Order Restricted Curve Clustering

Clustering methods have been used extensively in the analysis of gene expression data with main objective of identifying clusters of co-regulated genes. The co-regulation is defined in terms of the nominal gene expression measurements across different experiment conditions. However, under certain conditions such as time, temperature and doses, it is as well interesting not to only cluster based on similarity of gene expression measurement but also on the trends. In this section, we focus on the clustering of dose-response relationship under the monotone constrains.

4.1 The δ -biclustering Method

δ -biclustering is a node deletion based algorithm introduced by Cheng and Church (2000) to find a subset of genes and conditions with high similarity score. The similarity between members of a bicluster is defined in terms of the mean squared residue score of the cluster. The lower the mean squared residue scores, the more homogeneous is the cluster. The δ -biclustering method is based on a belief that every entry in a gene expression matrix can be expressed in terms of its row mean, column mean, the overall mean of the expression matrix and some random error. In other words, for a gene expression matrix \mathbf{Y} with entries y_{ij} , the residue of expression value of gene i under condition j can be expressed as :

$$r_{ij} = y_{ij} - y_{Ij} - y_{iJ} + y_{IJ}, \quad (7)$$

and the mean squared residue score of matrix Y is defined as:

$$H_{IJ} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} r^2. \quad (8)$$

where y_{IJ} is the overall mean of the expression matrix \mathbf{Y} ; y_{Ij} is the mean expression of gene i and y_{iJ} is the mean expression of condition j . The number of genes and conditions

are denoted with $|I|$ and $|J|$ respectively. Note that residual in (7) can be expressed in the form of a two-way ANOVA model without interaction:

$$y_{ij} = \mu + \alpha_i + \beta_j + r_{ij}, \quad (9)$$

with $\mu = y_{IJ}$; $\alpha_i = y_{Ij} - y_{IJ}$ and $\beta_j = y_{iJ} - y_{IJ}$.

For a perfect cluster, the mean squared residue score is zero. However, such clusters of genes may be less informative as it may contain only one gene per cluster. Hence, it may be more sufficient to find subsets of genes and conditions with mean squared residue score below a pre-specified threshold $-\delta$. Cheng and Church (2000) proposed a suit of node deletion algorithms that evolve in cycles starting from a gene expression matrix \mathbf{Y} until a bicluster that satisfies δ -criterion is found. Several cycles of the algorithm are then applied on the data by replacing the initially found biclusters with random data. The δ values used by Cheng and Church (2000) were elucidated from the existing clusters based on other clustering methods, which may not always be readily available.

4.2 Order Restricted Clustering for Dose-response microarray data

Within the dose-response setting, β_j in 9 is the effect of the j th dose level. for genes with monotonic dose-response curve $\beta_0 \leq \beta_1 \leq \beta_2 \leq \dots \leq \beta_k$. Our aim is to find subset of genes which form a cluster, i.e share the same dose-response curve shape. Further more, in this paper we focus on monotone clustering along the complete dose range. This means that, in contrast with Cheng and Church (2000) , we do not aim to find subset of both genes and conditions, but only subset of genes. Here, all dose levels are included in all clusters.

4.2.1 δ -clustering

Applying the δ -biclustering algorithm in only one dimension offers a clustering method where number of clusters is not required to be specified but implicitly controlled by the degree of

homogeneity assumed for a cluster. However, the choice of a δ value to achieve a desired degree of homogeneity is not readily available (Prelic *et al.*, 2006). We propose a relative δ criterion, where a cluster is a subsets of genes with mean squared residue score smaller than certain proportion ($0 \leq \lambda \leq 1$) of the heterogeneity in the observed data. Additionally, our aim is to find non-overlapping clusters of genes. To achieve this, we propose that members of initially found clusters should be deleted from the observed data before another cycle of the node deletion algorithms is applied. To ease the problem of local minima (Prelic, *et al.* 2006), we introduced an additional parameter ϕ that indicates the minimum number of genes in a cluster. Note that for $\lambda = 0$, the algorithm searches for clusters of genes with mean squared residue score of 0, which may result in as many clusters as the number of genes in the data set. On the other hand, specifying λ to be 1 considered all the genes as one cluster. Any values of λ between 0 and 1 reflect the degree of homogeneity expected of a cluster. We defined the algorithm to carry out this task as "algorithm 1". This algorithm can also be applied to cluster time-course microarray experiments as well.

Algorithm 1: δ -clustering

Input: Y , a matrix of real numbers; ϕ , minimum number of genes in a cluster; and λ :
 $0 \leq \lambda \leq 1$

Output: Y_{IJ} , a cluster that is a submatrix Y with a row set I with all the columns, with a score no larger than δ or $I \leq \phi$

Initialization: $\delta = \lambda * H_P$, where H_P is the mean squared residue score of the observed data.

Iteration:

1. apply node deletion algorithms proposed by Cheng and Church (2000) only in gene direction while the dose levels are kept fixed
2. if mean squared residue of the reduced matrix satisfies δ criterion or the number of genes in the reduced matrix is at most ϕ , then output the reduced matrix as a cluster.

3. delete members of cluster found in step 2.
4. Repeat steps 1 to 3 on the non-clustered genes until every gene belongs to a cluster.

4.2.2 δ -Clustering within the dose-response microarray setting

We turn now to discuss the case in which the primary interest is to identify gene clusters among the genes for which monotone trends was detected. In this stage, we focus the discussion to the case that the isotonic means $\hat{\boldsymbol{\mu}}^*$ are used for clustering. A typical dose-response microarray data Y has entries y_{ijk} corresponding to expression level of gene i under dose j from subject/sample k . Usually, different subjects/samples are used for different doses. In order to find clusters of genes with a similar dose-response relationship, it is required that gene-expression measurements under increasing doses are constrained to be monotone using isotonic regression. Thus, a new matrix Y^* of the isotonic means is obtained. The gene effects (α_i), isotonic dose effects (β_j^*) and overall mean (μ) can be defined as shown below:

$$\begin{aligned}
 \mu &= \sum_{i \in I, j \in J} \frac{y_{ij}^*}{|I||J|} && \text{overall mean} \\
 \alpha_i &= \sum_{j \in J} \frac{y_{ij}^*}{|J|} - \mu && \text{effect of the } i\text{th row - gene_}i \\
 \beta_j^* &= \sum_{i \in I} \frac{y_{ij}^*}{|I|} - \mu && \text{effect of the } j\text{th column - dose_}j
 \end{aligned} \tag{10}$$

Under the constrain of monotone trends, a directional inference for dose-response relationship can either be of the two directions; monotone increasing profile or a monotone decreasing decreasing profiles. To account for the direction of the dose-response relationship in the clustering process, a likelihood ratio statistic (Barlow, *et al.* 1972; Robertson *et al.* 1988) is used to assign each gene to a direction, where is most likely. The clustering algorithm is applied specific to each direction to find cluster of genes with monotone increasing trends and those with monotone decreasing trends. The linear model for δ - clustering algorithm using a reduced gene expression matrix based only on the isotonic means is is

given by the following model

$$y_{ij}^* = \mu + \alpha_i + \beta_j^* + r_{ij} \quad (11)$$

Algorithm 2 : Order restricted clustering based only on the isotonic means

Input: Y^* , a matrix of isotonic means, ϕ , minimum number of genes in a cluster; and λ :
 $0 \leq \lambda \leq 1$

Output: Y_{IJ} , a cluster that is a submatrix Y with row set I with all the columns, and a score no larger than δ or $I \leq \phi$

Initialization: $\delta = \lambda * H_P$, where H_P is the mean squared residue score of Y^* using 11.

Iteration:

1. Using global likelihood ratio statistic, assign each gene to a direction
2. Apply Algorithm 1 using the linear model in equation 11 specifically to each direction.

Algorithm 2 discuss above uses the isotonic matrix \mathbf{Y}^* as an input. Hence, one can argue that that algorithm 2 ignores the variability among the sample at each dose level and use only the variability between dose levels for the clustering. Interestingly, the linear model formulation for the clustering method can be explored further by taking into account the variability due to the replicates at each dose level. The mean squared residue score under this setting is similar to the variance expression in Barlow *et al.* (1972) under an ordered alternative hypothesis. Let y_{ijk} be the k th replicate of gene $_i$ at the j th dose level. We consider the following gene-specific model

$$y_{ijk} = \mu + \alpha_i + \beta_j^* + r_{ijk}, \quad (12)$$

where $\beta_j^* \leq \beta_{j+1}^*$ for $j < j + 1$, (for upward trend) and $\beta_j^* \geq \beta_{j+1}^*$ for downward trend. Note that the gene-specific model (12) is identical to the two way model in Barlow *et al.* (1972). The residuals sum of square in Barlow *et al.* (1972), calculated under ordered constrain, is given by

$$SS = \sum_{ijk} (y_{ijk} - \mu - \alpha_i - \beta_j^*)^2. \quad (13)$$

Algorithm 3 : Order restricted clustering based on observed data and isotonic means

Input: Y , a matrix of data with replicates; Y^* , a matrix of isotonic means, ϕ , minimum number of genes in a cluster; and λ : $0 \leq \lambda \leq 1$

Output: Y_{IJ}^* , a cluster that is a submatrix Y^* with row set I and columns set J , with a score no larger than δ or $I \leq \phi$

Initialization: $\delta = \lambda * H_P$, where H_P is the mean squared residue score of Y using 12.

Iteration:

1. Using global likelihood ratio statistic, assign each gene to a direction
2. Apply Algorithm 1 using the linear model in 12 specifically to each direction.

4.2.3 Choice of clustering parameters space(λ and ϕ)

Algorithms 1-3 discussed in the previous sections require to choose the values of the clustering parameters λ and ϕ . Note that for a given number of clusters there are two sources of variability which needed to be taken into account when selecting λ and ϕ : (1) the total variability (or the total sum of squared residual, TSS) and (2) the within cluster sum of squared residual (WSS). Let $R(\lambda, \phi)$ be the ratio WSS/TSS given by

$$R(\lambda, \phi) = \frac{\sum_q \sum_{ij} (y_{ij} - \mu_q - \alpha_i - \beta_{jq}^*)^2}{\sum_{ij} (y_{ij} - \mu - \alpha_i - \beta_j^*)^2}. \quad (14)$$

Let n_c be the number of clusters and N the number of genes, $1 \leq n_c \leq N$. Note that $n_c = 1$ for $\lambda = 1$ and $n_c = N$ for $\lambda = 0$ and $\phi = 1$ (provided that there are no identical genes in dataset). Hence, $R(\lambda, \phi) = 0$ for $n_c = N$ since WSS=0. On the other hand, $R(\lambda, \phi) = 1$ for $n_c = 1$ since in that case WSS=TSS. Between the extreme cases $R(\lambda, \phi)$ will decrease from 1 to 0 as the number of clusters increases from 1 to N . Since the total sum of squared of residual (TSS) is fixed, our wish is to find cluster parameters than minimizes within cluster sum of squared residuals(WSS). Let $S(\lambda, \phi)$ be the with cluster sum of squared for clustering based on a given choice of λ and ϕ , The most plauabile clustering parameters with the one that minimizes

$$S(\lambda, \phi) = \sum_q \sum_{ij} (y_{ij} - \mu_q - \alpha_i - \beta_{jq}^*)^2 \quad (15)$$

This will serve as goodness-of-fit measure of the a choice of clustering parameters given the data. However, looking for clustering parameters that minimizes $S(\lambda, \phi)$ will always favour clustering parameters that results in large number of cluster ,i.e, as $n_c \rightarrow N$, $S(\lambda, \phi)$ approaches 0. This suggest that there is a need to penalized for the complexity of the clusters based on a gene set of parameters. let $S_p((\lambda, \phi))$ be such a measure of the trade off between the goodness-of-fit and the complexity of the resulting clusters. Then $S_p((\lambda, \phi))$ can be defined as

$$S_p(\lambda, \phi) = S(\lambda, \phi) + G(n_c) \quad (16)$$

where $G(n_c)$ is the penalty term. The performance of the proposed method depend largely on the data as well as the specified clustering parameters. For most of real life data, these parameters are often unknown. We suggest a re-sampling based method to obtain a possible set of parameters. For a given dataset, G re-sampled datasets of size N are obtained by sampling N genes with replacement from the dataset. A clustering algorithm is applied on each re-sampled dataset using all the possible choice of parameters from a pre-defined sets. The most likely choice for the clustering parameters for each of the G re-sampled

dataset is selected using a penalized ratio of within cluster sum of squared residual (WSS) and total sum of squared residual (TSS). While the most frequently chosen parameters are considered as a plausible choice for clustering the original dataset. Intuitively, the number of resulting clusters (n_c) from any dataset is bounded between 1 and N , where N is the number of genes in the dataset. The number of cluster will be 1 if $\lambda = 1$. It will be N if $\lambda = 0$ and $\phi = 1$ provided there are no two genes in the dataset that are similar. Consequently, the ratio of WSS and TSS will be zero if number of resulting cluster equals the number of genes in the cluster because WSS is zero. The ratio will be 1 if number of the resulting cluster is 1 because WSS is equal to the TSS . We therefore choose a penalty such that when the number of cluster is 1, the penalty is 0. When the number of resulting clusters is N , the penalty is 1. Hence, the choice of the possible clustering parameters can be said to be a trade of between the amount of homogeneity of the resulting clusters and the number of clusters. In other words, the ratio of WSS and TSS favours large number of clusters, while the penalty favours small number of clusters. The penalized ratio of WSS and TSS is described in equation 17.

$$p \frac{WSS}{TSS} = \frac{\sum_k \sum_i \sum_j (y_{ij} - \mu_k - \alpha_i - \beta_{jk}^*)^2}{\sum_i \sum_j (y_{ij} - \mu - \alpha_i - \beta_j^*)^2} + \frac{\log(n_c)}{\log(N)} \quad (17)$$

5 Application to Data

The data set used in this paper is the same data used by Lin *et al.* (2007) and Lin *et al.* (2008). The global likelihood ratio test statistics \bar{E}_{01}^2 using isotonic means was used to test the significance of the dose-response trends in the data. The distribution of the test statistic \bar{E}_{01}^2 was approximated by permutations and correction for multiple testing was done using the Benjamini-Hochberg (BH) procedure for controlling False Discovery Rate (FDR, Benjamini and Hochberg, 1995). There were 3499 genes found to be significant out of 16,998 genes. Out of the 3499 significant genes, 1600 genes have monotone increasing profiles, while 1899 genes have monotone decreasing profiles. The clustering algorithms are illustrated using 1600 genes with monotonic increasing profiles.

5.1 Clustering of dose-response data using only the isotonic means

In order to choose a plausible clustering values for λ and ϕ , we generated 1000 re-sampled datasets by randomly sampling with replacement from the data of genes with monotonic increasing profiles. Each re-sampled dataset contains 100 genes. We pre-defined λ as $\{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90\}$. We set the plausible values for ϕ as $\{5, 10, 15, 20\}$. For each re-sampled dataset, we consider a parameter set $\{\lambda, \phi\}$. For example, $\{\lambda = 0.05, \phi = 5\}$ and obtain clustering of the re-sampled dataset using "algorithm 2". The most likely choice for the clustering parameters was chosen as the one with a minimum a penalized ratio of within cluster sum of squared residual and total clusters sum of squared residual. We present in Figure 3 the proportion of times a parameter set is chosen based on the 1000 re-sampled data. These proportion does not necessarily sum to one because it is possible for more than one parameter sets to be chosen as possible clustering parameters for a re-sampled dataset. The plot shows that the parameter sets are not equally likely. For the dataset at hand, some values of λ and ϕ are more likely than others. The most frequently chosen is parameter set $\{\lambda = 0.1, \phi = 10\}$.

Figure 3 about here.

The clustering algorithm (Algorithm 2) for clustering dose-response microarray data using only the isotonic means was applied on the genes with monotone increasing profiles. The parameters set $\lambda = 0.1$ and $\phi = 10$ was used. There are 24 clusters of genes obtained with the specified parameters, We present in Figure 4 some of the resulting clusters . The plots show that most of the clusters contain genes with a similar profiles.

Figure 4 about here.

5.2 Clustering of dose-response data with replicates at each dose

Clustering of dose response microarray data with replicated can be done using "algorithm 3". This may be necessary in order to investigate the effects of the extra-variability due to the replicates at each dose may have on the resulting clusters. if the variability between the replicated at each dose is small, we at least expect clusters that are as good as the one obtained using only the isotonic means. if the within gene variability due to the replicates is much more than between gene variability, the resulting cluster may be worse than clustering using only the isotonic means. WE applied "algorithm 3" using the same choice of parameters used for "algorithm 2".The resulting clusters are presented in Figure 6. It can be observed from the plots that resulting clusters are as good as those obtained using only the isotonic means. Additionally, there are genes with a similar dose-response patterns but belongs to different clusters. This is due to the noise - intensity relationship in gene expression experiments. Specifically, Chudin et al, (2001), Tu et al, 2002 and Hochreiter, et al , 2006 pointed out that the variance of the noise in gene expression measurements depends on the signal strength. Which may implies that the stronger the intensity, the larger the variance. This may explain the local minima associated with the δ -biclustering algorithm (Prelic, *et al.* 2006). For the proposed algorithm, the noise-intensity relationship implies an intrinsic ordering of the resulting clusters. Clusters of genes with strongers signal come later than those with weak signal. Hence, it offers a direction for the exploration of the resulting clusters.

Figure 6 about here.

6 Discussion

One of the interests in dose-response microarray experiments is to find cluster of genes with similar dose-response relationship under an increasing doses of a therapeutic compound. In the present paper, we propose a clustering method for dose-response microarray data. Our proposal is motivated by the δ - biclustering proposed by Cheng and Church, (2000),

where they defined a bicluster to be a subset of genes and a subset of conditions with a 'high similarity score', using mean squared residue score. For the proposed clustering method, the δ values is modified to be data dependent. It is expressed as a pre-specified proportion of heterogeneity in the observed data. The method shares some features of the standard clustering method, it partitions genes in a dose-response microarray data into non-overlapping groups. The benefits of the clustering methods in the context of dose-response microarray experiments is in three folds; (1) it can be used for clustering and visualization of finitely many dose-response patterns in microarray data; (2)it can be used as an additional pre-processing method to exclude genes with non-monotone profiles prior to testing for significant dose effects; (3) It can as well be used as a clustering and visualization tool for significant dose-response patterns after testing procedures as in this paper.

The clustering algorithms were applied to a real real life data. In the first instance, each gene in the dose response microarray data was assigned to either a monotonic increasing profile or a monotonic increasing profiles using the global likelihood ratio test. Specifically to each direction, the clustering algorithms were applied. In this paper, we illustrate our prosed method using genes with monotone increasing profiles. One major challenge for the clustering algorithm is in choosing the plausible values of the clustering parameters. A such, we proposed a re-sampling based method. A parameter set that most frequently chosen based on 1000 re-sampled data set was considered. For genes with monotone increased profiles, a parameter set $\lambda = 0.1, \phi = 10$ was chosen. The method was illustrated under two settings, namely; clustering using only the isotonic means; and clustering taking into account the replicates at each dose level. Under both settings, most of the resulting clusters contains genes with a similar dose-response profile,

References

- Affymetrix (2004) GenChip Expression Analysis Technical manual, Rev. 4. *Affymetrix Santa Clara, CA*
- Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society, Biotstatistics*, **57**, 289-300.
- Barlow, R.E., Bartholomew, D.J., Bremner, M.J. and Brunk, H.D. (1972) *Statistical Inference under order restriction*, New York: Wiley.
- Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2002) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, **19**, 185-193.
- Chudin, E., Walker, R., Kosaka, Alan., Wu, X.S., Rabert, D., Chang, T. K. and Kreder D.E. (2002) Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biol.* , **3(1)**, research0005.1research0005.10.
- Cheng, Y. and Church, G.M. (2000) Biclustering of Expression Data *Proc. Conf. Intelligent Systems for Molecular Biology (ISMB)*, **55**, 93-104.
- Ge, Y., Dudoit, S. and Speed, P.T. (2003) Resampling based multiple testing for microarray data analysis. *technical report*, **633**, University of Berkeley.
- Hubell, E., Liu, W.M. and Mei, R. (2002) Robust estimators for expression analysis. *Bioinformatics*, **18(2)**, 1585-1592.
- Hochreiter, S., Clevert, D. and Obermayer, K. (2006) A new summarization method for affymetrix probe level data. *Bioinformatics*, **22(8)**, 943-949.
- Lin, D., Shkedy, Z., Yekutieli, D., Burzykowki, T., Göhlmann, H.W.H., De Bondt, A., Perera, T., Geerts, T., Bijmens, L. (2007) Testing for trend in dose-response microarray experiments: comparison of several testing procedures, multiplicity, and resampling-based inference. *Statistical Application in Genetics and Molecular Biology*, **6(1)**, article 26.

- Lin, D., Shkedy, Burzykowki, T., Göhlmann, H.W.H., De Bondt, A., Perera, T., Geerts, T., Bijnens, L. (2008) Classification of trends in dose response microarray experiments using information theory selection methods. *Proceedings of Joint Statistical Meeting*, Special issue.
- Madeira, S.C. and Oliveira, A.L. (2004) Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **1(1)**, 24-45.
- Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Buhlmann, P., Gruissem, W., Hennig, L., Thiele, L. and Zitzler, E. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22(9)**, 1122-1129.
- Tu, Y., Stolovitzky, G. and Klein, U. (2002) Quantitative noise analysis for gene expression microarray experiments *Proc. Natl Acad. Sci. USA*, **99**, 14031-14036.

Table 1: The set of seven possible monotonic dose-response models for an experiment with four dose levels. μ_i is the mean response of dose level

Model	Up: Mean Structure	Down: Mean Structure
g_1	$\mu_0 = \mu_1 = \mu_2 < \mu_3$	$\mu_0 = \mu_1 = \mu_2 > \mu_3$
g_2	$\mu_0 = \mu_1 < \mu_2 = \mu_3$	$\mu_0 = \mu_1 > \mu_2 = \mu_3$
g_3	$\mu_0 < \mu_1 = \mu_2 = \mu_3$	$\mu_0 > \mu_1 = \mu_2 = \mu_3$
g_4	$\mu_0 < \mu_1 = \mu_2 < \mu_3$	$\mu_0 > \mu_1 = \mu_2 > \mu_3$
g_5	$\mu_0 = \mu_1 < \mu_2 < \mu_3$	$\mu_0 = \mu_1 > \mu_2 > \mu_3$
g_6	$\mu_0 < \mu_1 < \mu_2 = \mu_3$	$\mu_0 > \mu_1 > \mu_2 = \mu_3$
g_7	$\mu_0 < \mu_1 < \mu_2 < \mu_3$	$\mu_0 > \mu_1 > \mu_2 > \mu_3$

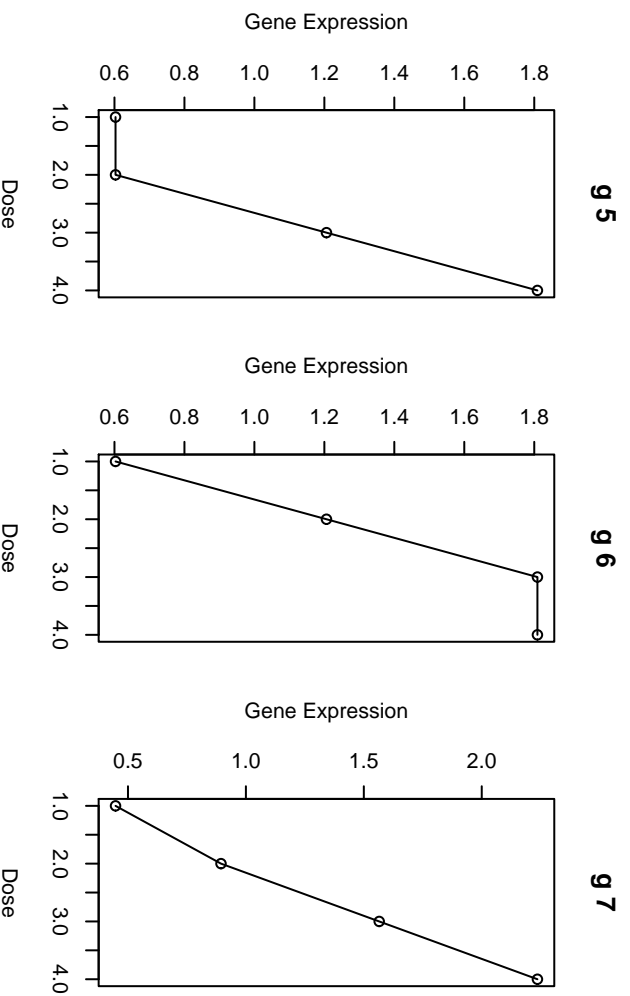
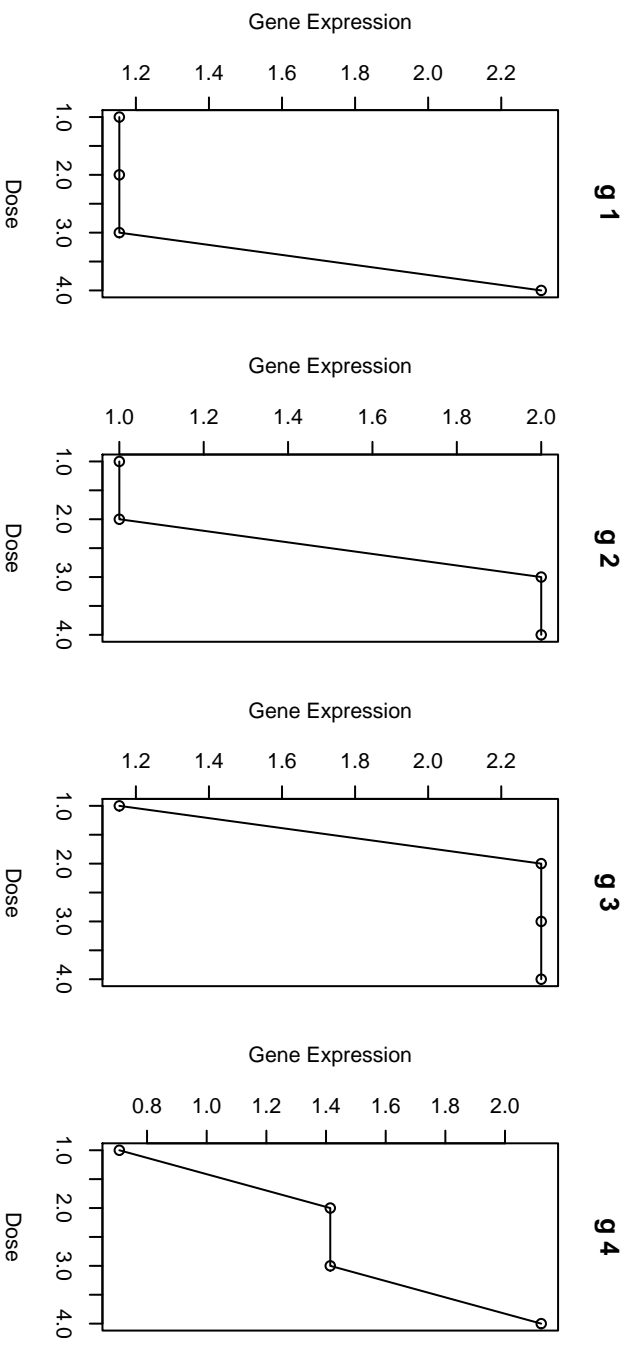


Figure 1: The set of the seven possible monotonic dose response curves for an experiment with four dose levels.

Dose 4

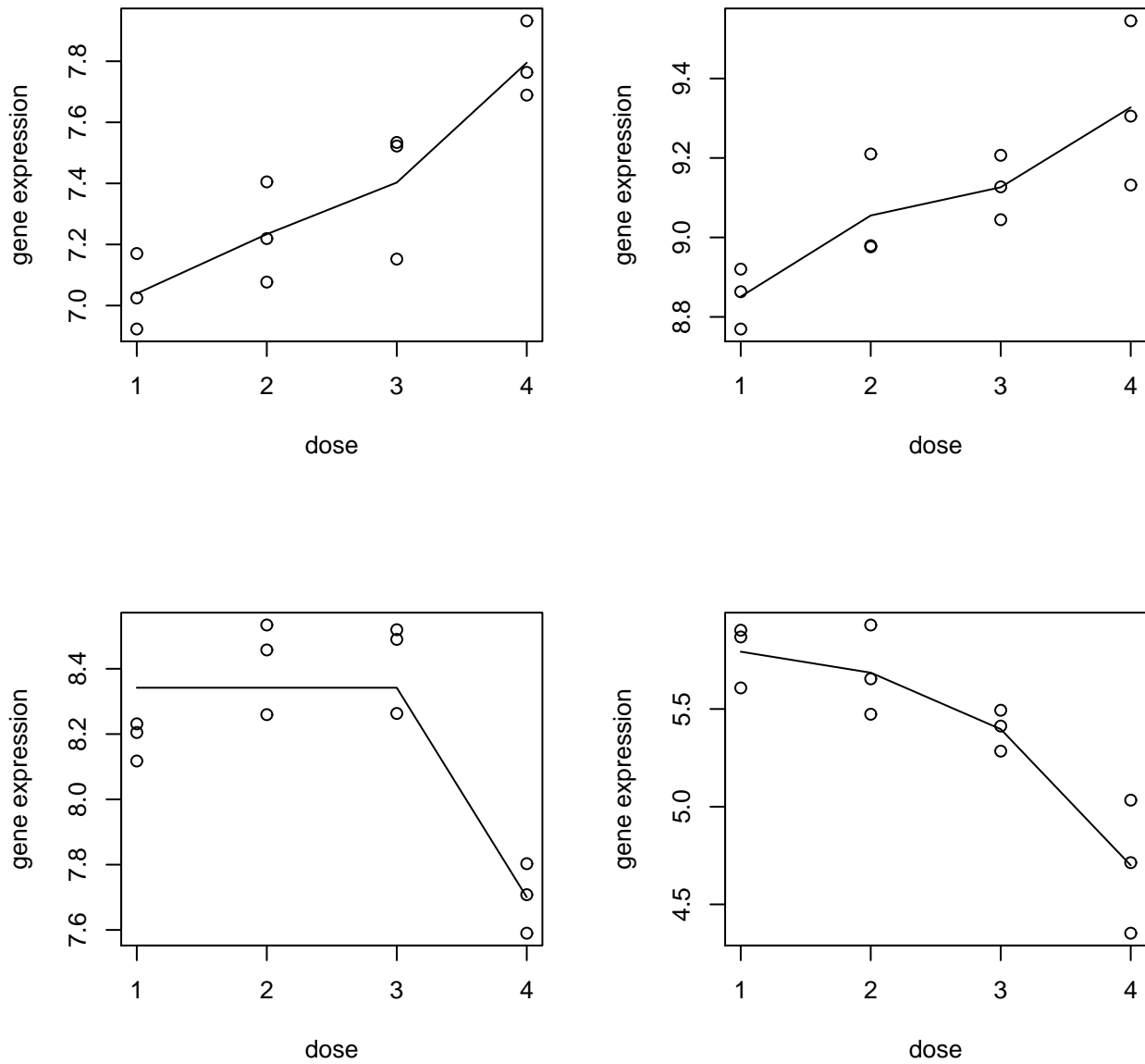


Figure 2: Example of genes with significant dose-response relationship from dose-response microarray experiments 1: the scatter points denote the observed data , while the solid line represent the isotonic means

Dose 6

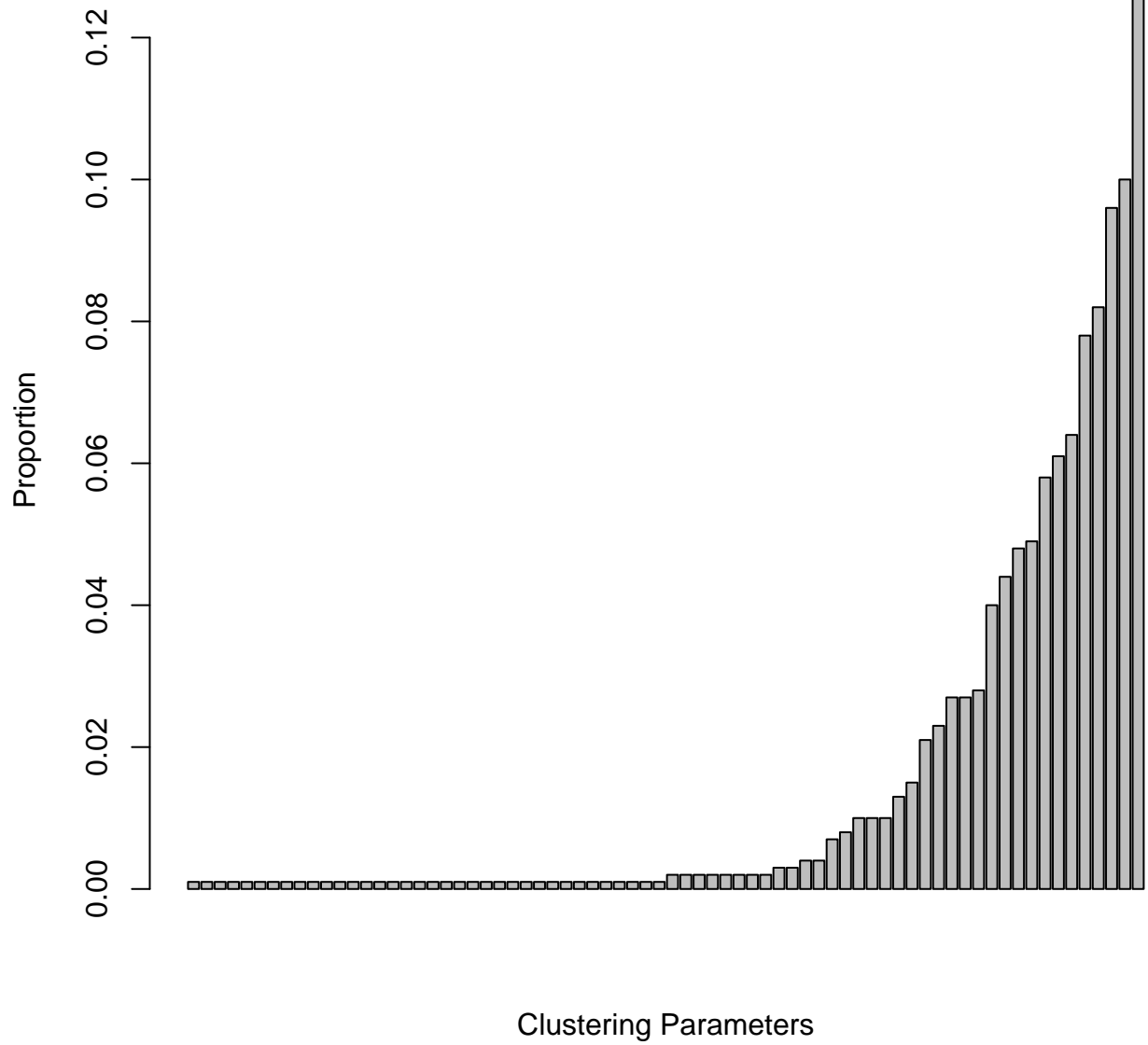


Figure 3: The choice of clustering parameters based on a 1000 re-sampled data

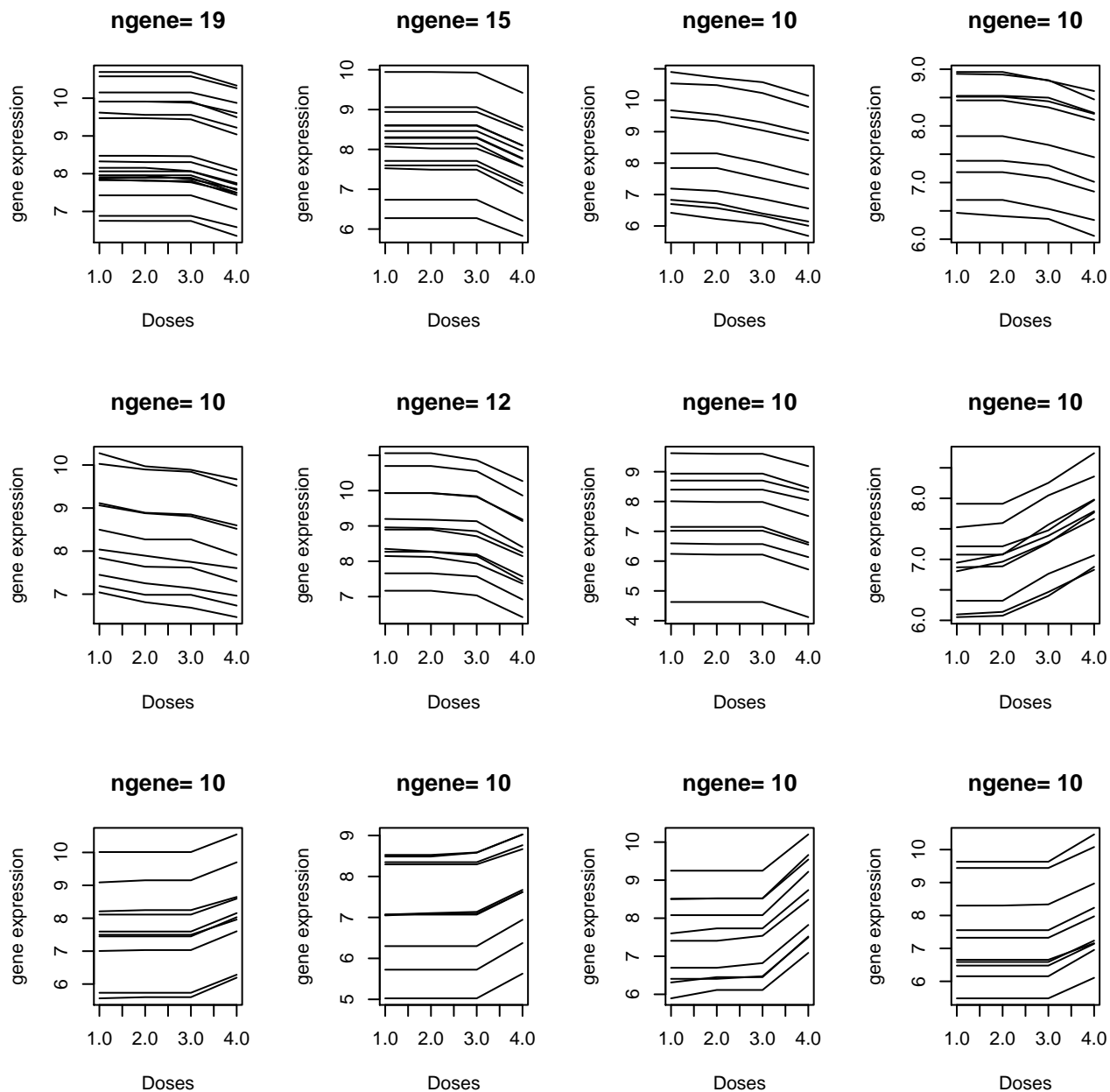


Figure 4: Some of resulting clusters from using applying order restricted clustering on dose-response microarray experiment 1, using observed data and isotonic means: the raw gene expression measurements

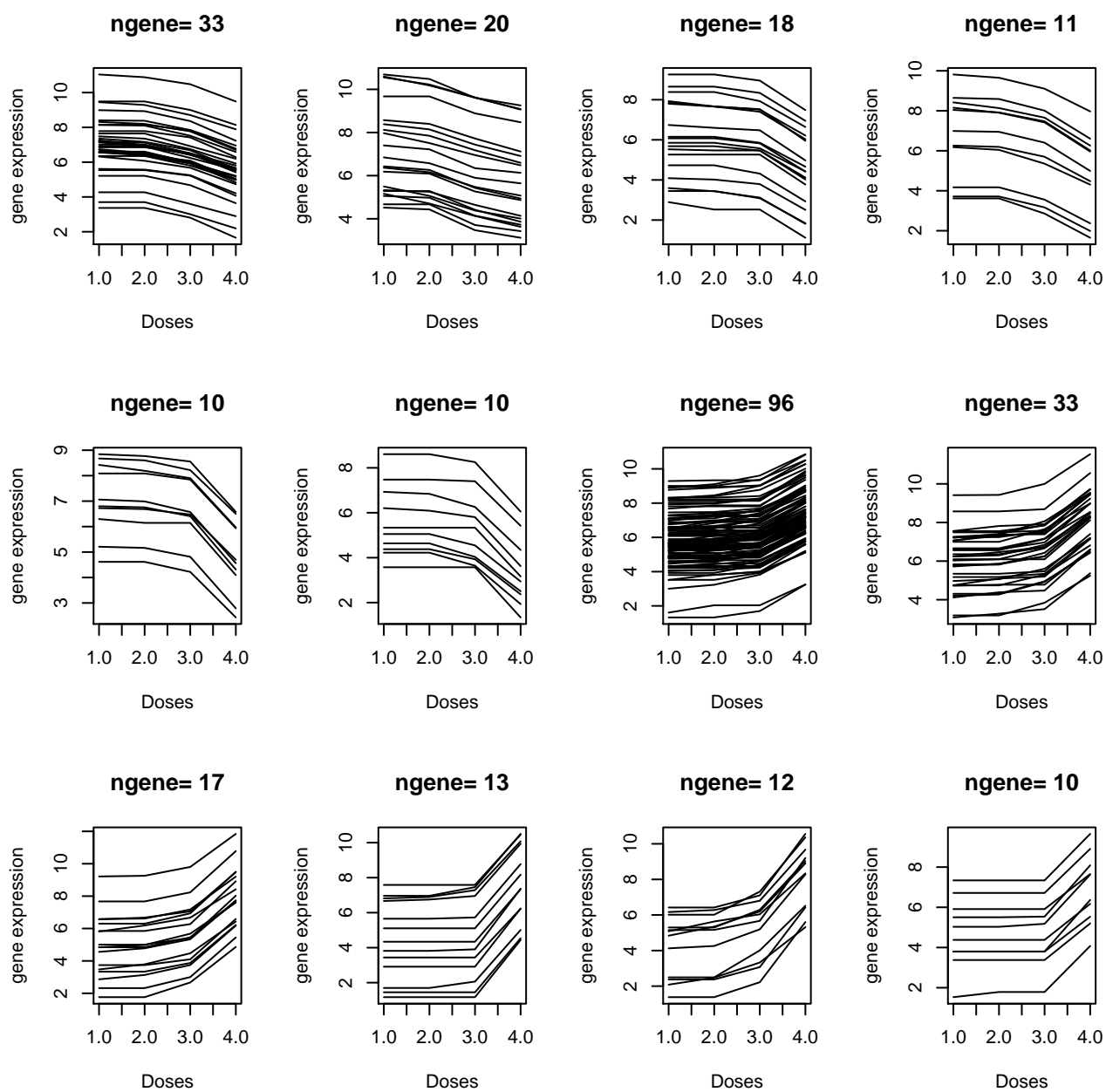


Figure 5: Some of resulting clusters from using applying order restricted clustering on dose-response microarray experiment 1, using only the isotonic means: the raw gene expression measurements

