Identifying genes altered by a drug in temporal microarray data: A case study

Nicoleta Serban* Larry Wasserman*

September 1, 2003

Abstract. DNA microarrays are suited for parallel analysis of gene expression across different tissues or populations of cells. An important application of microarray techniques is identifying genes altered by a particular drug of interest. This process allow biologists to target drug therapies to particular diseases, and, eventually, to gain more knowledge about the biological processes responsible for diseases. Such an application is described in this paper and is focused on the effect of a drug, once prescribed for diabetes, which is a genetically heterogeneous diseases. Our case study moves beyond the simple 'rules' of previously published studies by tailoring our analysis to the data in hand. We identify significant systematic sources of variability which are potential issues for other microarray datasets. Subsequently, we apply two novel nonparametric multiple hypothesis test to identify differentially expressed genes and we find a set of genes which appear to change in expression level over time in response to the drug treatment. Finally, we address the problem of identification of co-expressed genes among the ones obtained from multiple hypothesis testing using a novel cluster analysis based on nonparametric regression procedure, REACT. We evaluate our methodologies on a complex synthetic dataset.

Keywords: microarray, poly(dT), expressed sequence tag (EST), false discovery rate, runs test, risk estimation and adaptation after coordinate transformation (REACT).

1 Introduction

This paper is a statistical study of a microarray dataset for treated adipose cells. Our analysis looks for a particular form of deviation in a very large data set. However, these kind of data contains significant systematic sources of variation and bias which could overpower the effect of interest. To anticipate undesirable deviations, we first focus on data validity analysis. We take into account several of these deviations in turn and develop four filters for removing their effects.

Following up the data validity, we analyze the response to the drug treatment. We consider two different applications of hypothesis testing to capture those genes which change in expression level under different hypotheses. They are the *2-time-difference test*, and the *multiresolution runs test*. The former test could be also applied to compare two or more experimental conditions when at least one has replicated measurements. The latter is a test based on the longest run test statistic. We control the rate of false positives by applying the False Discovery Rate to multiple testing. The cluster membership of the genes which respond similarly to the drug treatment can then be analyzed for further insights. We cluster expression profiles by clustering their cosine transforms. This method has its roots in the estimation procedure called REACT [2,3] which is for the first time connected to cluster analysis: we refer to it as *REACT clustering*.

The paper is organized as follows. In the following subsections, we present the experimental background and the main questions of interest raised by the experiment, and describe the microarray data format and data pre-processing. Additionally, we introduce the synthetic data used for validating our methodologies. Section 2 is reserved entirely for the methods we use in our analysis. In section 3 we analyze the microarray data following three steps: data validity (removing sources of variability), identification of differentially expressed genes and cluster analysis of those sequences identified in the previous step. It is followed by the analysis of the synthetic datasets. A summary of the methods applied to the data and of our conclusions (section 5) finalize the analysis.

1.1 Microarray Data

The analysis of the gene expression data set from the microarray experiment is addressing two different problems.

1. Identify and remove *systematic sources of variation and bias*.

2. Determine *the effect of the drug treatment on the gene expression level* under the experimental conditions in our study.

1.1.1 Background on the experiment

Why is our experiment important? We study a family of drugs which are used in humans to treat diabetes and obesity by increasing insulin sensitivity. Decreased insulin

*Carnegie Mellon University, Department of Statistics

sensitivity is a hallmark of both diabetes and obesity, and may be one of the major causes of their development. The ability of those drugs to increase insulin sensitivity can be investigated by identifying which genes change expression in response to the drug treatment in adipose tissue. Briefly, if we find that expression level of some genes is altered in our experiment, the biologist will ask: "Does increasing the expression of those genes increase insulin sensitivity in adipose tissue?". If the answer is "yes" then biologists can begin to develop other drugs that will increase the level of expression of those genes. It is also a good chance to understand more about how this family of drugs and insulin work.

What is the experiment? We examined data from spotted cDNA microarrays (Research Genetics, Carlsbad, California) experiment. The experiment was finished by February, 2002 at the University of Pittsburgh (Peters *et al.*). The spotted cDNA microarray experiment consists of a time-sequenced sampling of differential expression in mRNA from (3T3L1 cultured) adipose cells originally obtained from mice. These cells were treated with a drug, *troglitazone*, which is a member of a family of drugs known as thiazolidendiones (TZD's). The drug was mixed with a chemical DMSO (detergent) which makes the drug soluble. In our experiment, the drug treatment of the cells lasted for different periods of time ranging from 0 hours to 24 hours.

1.1.2 Data characteristics

Genes. There are 5355 DNA probes on the microarray. Among all probes, there only 4696 sequences of DNA with identifiable sequences of nucleotides and 139 probes consisting of tgDNA (total genetic DNA). We identify two types of sequences: known genes and expressed sequence tags (*EST's*) which are short sequences (100-500 nucleotides) typically from one end of a cloned cDNA [4]. In this presentation, *all the probes are called genes or sequences* and the distinction between EST's, genes and tgDNA will be made any time it is necessary.

Measurements. The data consist of 47 measurements of mouse adipose cells. These measurements were reported on 20 arrays (filters/chips), each being used at least twice. Thus we have data from two uses for all 20 arrays (40 measurements) and from the third use for only 7 filters. For each measurement, target cDNA was obtained by mRNA extraction and reverse transcription (into complementary DNA). Then the cDNA targets were hybridized to microarrays containing 5355 probes. Each of the 47 hybridizations produced images, which were processed using the software package Pathways 3. The main quantity of interest reported by the image analysis methods is the intensity for each probe on each array. After image processing, the gene expression data can be summarized by a matrix of intensities with 47 columns (corresponding to the number of arrays) and 5355 rows (corresponding to the number of probes).

Treatments. The adipose cells were treated in three different ways, with a solvent called DMSO (*control C+*) or without (*control C-*), and the *test drug*, a mixture of the drug troglitazone and DMSO (T). It is necessary to see the difference between treating with or without DMSO in order to distinguish between the genes which respond to the drug and genes which respond to the chemical, DMSO. A comparison between the two control groups helps in identification of any unexpected effect of DMSO.

Sampling scheme for treatment time. Each measurement lasted for *a period of time ranging from 0 to 24 hours*. The sampling scheme is in Table 1.1. In the first table the time ranges from 0 to 6 hours (360 minutes) and in the second table the time ranges from 7 to 24 hours. Note that at time 6 hours (360') there are 5 measurements for the first use reported on chips 7, 8, 9, 10 and 11, and a control measurement for the second use on the chip 2. At time 0 there are only measurements in the control group without DMSO for chip 1, for the first and second use. At 24 hours there are measurements for the test drug (chip 20, first use) and for the control with DMSO (chips 3 and 7, the third use).

Data transformation and normalization. Normalization can be performed in different ways depending on the availability of control DNA sequences, the number of genes that are expected to respond to the drug or other considerations. The method we consider for our data is *a global linear normalization* based on a constant adjustment. A global linear normalization forces the log intensities to have median equal to zero at each array, making the median of the experiment array the same as that of the baseline array. After logarithm transformation and normalization, an $array$ is transformed into $\log(array) - median(\log(array))$. This appears to perform well because we expect only a relative small proportion of the genes to vary significantly in expression between mRNA samples[20].

1.2 Synthetic data

We generate the synthetic data according to the regression model:

$$Y_j = f(t_j) + \sigma\epsilon_j$$

with $j = 1, \ldots, m$. We want to use these synthetic data to evaluate the nonparametric methods introduced in this paper. Subsequently, we need f to take different shapes. The regression functions for f are:

$$\mathbf{f_1(t)} = I_{\{t\in S\}}(t)\left(\left(\tfrac{2-5t}{2}\right) \bigwedge \left(\left(\tfrac{5t-2}{3}\right)^2 + \sin\tfrac{5\pi t}{2}\right)\right) + I_{\{t\in S^c\}}(t)\left(\left(\tfrac{2-5t}{2}\right) \bigwedge \left(\tfrac{5t-2}{3}\right)^2\right)$$

$$\mathbf{f_2(t)} = \left(\tfrac{2-5t}{2}\right) \bigwedge \left(\left(\tfrac{5t-2}{3}\right)^2 + \sin\tfrac{5\pi t}{2}\right)$$

$$\mathbf{f_3(t)} = I_{\{t\in T^c\}}(t)\left(\left(\tfrac{2-5t}{2}\right) \bigwedge \left(\left(\tfrac{5t-2}{3}\right)^2 + \cos(2\pi t)\right)\right) +$$

minutes	0	15	30	45	60	75	90	105	120	135	150	165	180	240	300	360
treatment		T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
use no.		2	2	2	2	2	2	2	1	2	1	2	1	1	2	1
chip no.		3	4	5	6	7	8	9	3	10	4	11	5	6	12	7-11
control	C-				C+/C-											C+
use no.	1/2				1/3											2
chip no.	1				2/1											2

hours	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
treatment	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
use no.	2	1/2/3	2	1	2	1	2	1	2	1	2	1	2	1	3	1	3	1
chip no.	13	12/20/4	14	13	15	14	16	15	17	16	18	17	19	18	5	19	6	20
control										C+								C+
use no.										3								3
chip no.										2								3/7

Table 1.1: *Within a cell the first letter indicates whether it is a treatment or a control with or without DMSO, the first number indicates the use number of the array specified underneath.*

$$I_{\{t\in T\}}(t)\cos(2\pi t)$$

$$\mathbf{f_4}(\mathbf{t}) = \cos(2\pi t)$$

where S and T are intervals: $S = \left(\frac{2}{5}, \frac{4}{5}\right)$, $T = \left(\frac{1}{5}, \frac{4}{5}\right)$.

The first two curves have similar pattern with some perturbation at the early and late time points (see Figure 1.1, left upper plot). Similarly, the last two curves, f_3 and f_4, have different patterns at the late hours(see Figure 1.1, bottom plot). We consider in the analysis the reversed (negated) curves (see Figure 1.1, right plots). In this way, we have a larger number of patterns in these data.

The synthetic data contains 75 curves for each of eight patterns on different scale. For simplicity, we take the noise ϵ_i being normal distributed with standard deviation lying in $[.1, .6]$.

Thus the synthetic data over unequally spaced time points $t_1, \ldots, t_m$ ($m = 30$ or 35) have:

$$\begin{aligned}
&Y_{ij} = N_m(0, \sigma_i), \text{ where } i = 1, \ldots 1400,\\
&Y_{ij} = \mu_i f_1(t_j) + N_m(0, \sigma_i I_m), \text{ where } i = 1401, \ldots 1550,\\
&Y_{ij} = \mu_i f_2(t_j) + N_m(0, \sigma_i I_m), \text{ where } i = 1551, \ldots 1700,\\
&Y_{ij} = \mu_i f_3(t_j) + N_m(0, \sigma_i I_m), \text{ where } i = 1701, \ldots 1850,\\
&Y_{ij} = \mu_i f_4(t_j) + N_m(0, \sigma_i I_m), \text{ where } i = 1851, \ldots 2000
\end{aligned}$$

where the scale μ_i takes $1 \le |\mu_i| \le 2$ to have a significant signal and σ_i can take any value in $[.1, .6]$.

We believe that the synthetic data have the structure and the complexity of a dataset provided by a microarray experiment. However, the generated data assume independent curves which is not necessarily the case of microarray data. In section 2.1 this issue is addressed.

2 Methods

2.1 Two multiple hypothesis tests

The gene expression data provided by the microarray experiments are large-scale data that capture the behavior of thousands of genes simultaneously. This suggests an approach of multiple inference. The question to be addressed is to detect differentially mRNA abundance as cellular responses to the environmental change (drug treatment). This suggests an approach of hypothesis testing.

To correct for the multiplicity problem we consider the error rate to be the False Discovery Rate (FDR), *the expected proportion of errors among the rejected hypotheses*. See [1,7,9,17]. We use Benjamini-Hochberg (1995) method; see figure below where the N sorted p-values are on the y-axis and the uniform quantiles, $(\frac{1}{N}, \frac{2}{N}, \ldots, \frac{N}{N})$, on the x-axis.

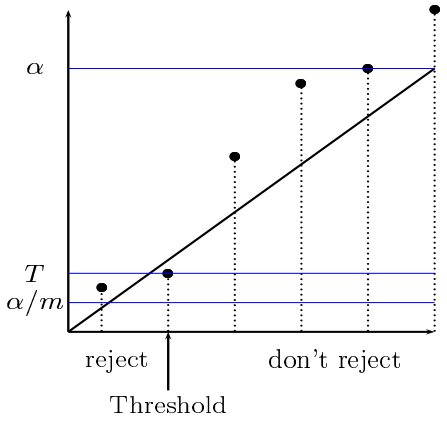

Figure 2.1: *Rejection Rule with FDR.*

However, the FDR procedure assumes independent test statistics and the gene expression levels tend to be correlated. This issue is bypassed by the argument given in Storey & Tibshirani [18]. This is, under *"loose dependence" and large number of tests*, FDR behaves as if the tests were independent. The "loose dependence" applies to the microarray data because genes tend to be dependent in small groups [18].

We propose two different applications of hypothesis test-

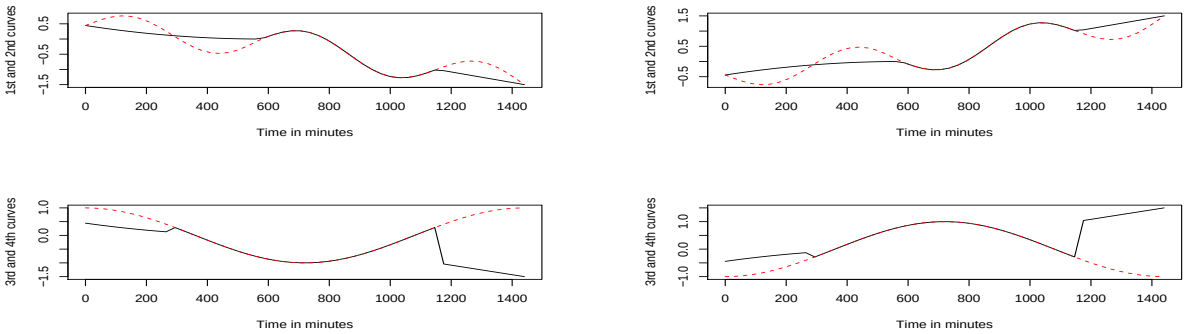

Figure 1.1: *The two plots on the left represent the four curves: f_1, f_2, f_3, and f_4; we plotted the first two patterns in the upper panel and the last two in the bottom panel. The plots on the right are the same curves but with negative means: $(-1)f_1$, $(-1)f_2$, $(-1)f_3$, and $(-1)f_4$. The time t applied to the four function is the time on x-axis rescaled to have values in $I = [0, 1]$.*

ing in order to capture gene expression levels which change at specific time points and gene expression profiles that change over all different treatment times.

2.1.1 The 2-time-difference test

This test provides candidate genes whose expression level changes between two treatment times.

Assume that the genes are divided into two classes: genes which change in intensity level between the two treatment times ("affected") and genes which don't change in intensity level ("unaffected"). Thus the test for gene i is:

$$H_{i0} : \text{gene is unaffected}$$

$$H_{i1} : \text{gene i is affected}$$

for $i = 1, ..., N$ (where N is the number of genes which are tested). The task is to *test simultaneously* for all N genes if the difference in the expression (intensity) levels at the same time point t_1 has the same distribution as the difference in the expression levels at different time points t_1 and t_2 (H_0).

***P*-value estimation in the 2-time-difference test.** Assume that at time t_1 there are r replications (i.e. there are r measurements under the treatment which lasted for the period of time t_1). Denote the replications at time t_1 for gene i by $X_i = (X_{i1}, X_{i2}, ..., X_{ir})$. Take also a measurement at time t_2 for gene i: Y_i.

Define the observed null difference for gene i and a fixed j (in $1, ..., r$) as follows:

$$\widehat{D_{ij}} = \left| \frac{1}{r-1} \sum_{k=1,..,j-1,j+1,..,r} X_{ik} - X_{ij} \right|.$$

Define the differences in expression for gene i between time t_1 and time t_2

$$\hat{d_{ij}} = \left| \frac{1}{r-1} \sum_{k=1,..,j-1,j+1,..,r} X_{ik} - Y_i \right|$$

The distribution of the null difference called F_0 can be estimated by the empirical distribution of $\widehat{D}_{ij}$ with $i = 1, ..., N$ which are identically distributed F_0:

$$\hat{F}_0(t) = \frac{1}{N} \sum_{i=1}^{N} I(\widehat{D_{ij}} < t).$$

It follows that the p-values can be estimated using the estimated $\hat{F}_0$ such as:

$$\hat{P}_i = 1 - \hat{F}_0(d_{ij}) = \frac{1}{N} \sum_{k=1}^{N} I(\widehat{D_{kj}} > d_{ij})$$

where $I()$ is the indicator function.

We have evidence against the "unaffected" hypothesis (H_0) for gene i if $\hat{P}_i < \hat{P}_{\hat{k}(.1)}$ where $\hat{k}$ is estimated based on the FDR procedure.

2.1.2 Multiresolution runs test.

The multiresolution runs test captures candidate genes whose expression changes over all experimental times. Even though this test is more computationally intensive than other runs tests, we prefer it over many other runs tests because on synthetic data, it displays higher power.

We assume for each genes i that $Y_{ij} = \mu_{ij} + \epsilon_{ij}$ where median$(\epsilon_{ij}) = 0$ and ϵ_{ij} are independent, identically distributed for $j = 1, \ldots, m$ and $i = 1, \ldots, N$ (N = # of genes and m = # of time points).

The null hypothesis for each gene i is:

$$H_0 : \mu_{i1} = \cdots = \mu_{im},$$

that is the intensity over time for gene i is randomly distributed with the same distribution as the error ϵ_{ij}.

***P*-value estimation in the multiresolution runs test.** Given a sequence of real numbers $x = (x_1, \ldots, x_m)$ let $r^+(x)$ be the longest run of positive numbers, let $r^-(x)$

be the longest run of negative numbers, and let $r(x) = \max\{r^+(x), r^-(x)\}$.

Let $R_1, \ldots, R_m$ be independent such that $\mathbb{P}(R_i = 1) = \mathbb{P}(R_i = -1) = 1/2$, and $\mu(k) = \mathbb{E}(r(R_1, \ldots, R_k))$ and $\sigma(k) = \sqrt{\mathrm{Var}(r(R_1, \ldots, R_k))}$ which can be computed by simulation, for $k = 1, \ldots, m$.

The test statistic is defined as

$$W(r,s) = \min_a \frac{|r(Y_r - a, \ldots, Y_s - a) - \mu(s - r + 1)|}{\sigma(s - r + 1)}$$

and

$$W = \max_{1 \le r < s \le k} W(r,s).$$

This test statistic has a similar form to the one presented by Dúmbgen and Johns[6]. Define G as the distribution of the test statistic W under H_0. We computed the cdf of W as $G(w) = \mathbb{P}(W \le w)$ under H_0 by simulation. The p-value for gene is computed as follows:

$$\hat{P}_i = \frac{1}{B} \sum_{b=1}^{B} (W(R_1^b, \ldots, R_m^b) > W(Y_{i1}, \ldots, Y_{im}))$$

where $R_1^b, \ldots, R_m^b$ is a sample of size m from the Rademacher distribution (i.e. $\mathbb{P}(R_i^b = 1) = \mathbb{P}(R_i^b = -1) = 1/2$).

Once the p-values are estimated, we reject the null hypothesis (randomness) for gene i if $\hat{P}_i < \hat{P}_{\hat{k}(.1)}$ where $\hat{k}(.1)$ is estimated based on the FDR procedure.

2.2 REACT clustering

The final objective is to identify groups of genes with similar expression profiles (i.e. co-expressed gene). The biological underpinning is that co-expressed genes are likely to be co-regulated and are hence co-expression can suggest functional pathways and interactions between genes.

In this context, a natural mathematical description of similarity in expression profiles is the correlation coefficient.

Clustering in Fourier space. An alternative to clustering using correlation is to cluster in Fourier space as we propose below. Similar to Beran's paper on REACT (risk estimation and adaptation after coordinate transformation) [2], we first transform the data using the cosine transform and then we estimate a curve (X) linearly: αX where f falls in a class of function. In our case, it is the nested selection class:

$$\alpha \in \mathcal{F}_{NNS} = \{(1, \ldots, 1), (1, \ldots, 1, 0), \ldots, (0, \ldots, 0)\}.$$

The smoothing algorithm is summarized in Figure 2.2.

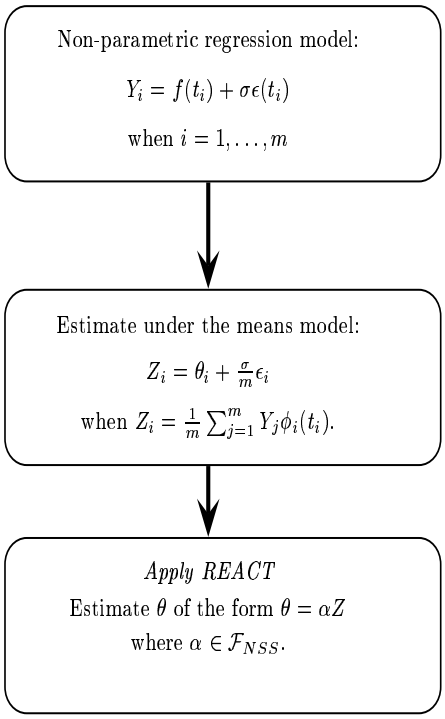

Figure 2.2: *Smoothing algorithm.*

The transforms of the observed expression profiles are obtained as follows.

Coordinate transformation. Let Y_{ij} be the expression level for gene i and time t_j. Here $1 \le i \le N$ and $1 \le j \le m$, where N is the number of genes and m is the number of time points. Assume that

$$Y_{ij} = f_i(t_j) + \sigma_i \epsilon_{ij} \tag{1}$$

where $E(\epsilon_{ij}) = 0$.

Step 0: Estimate σ_i^2, variance for gene i over time with

$$\hat{\sigma}_i^2 = \frac{1}{m-1} \sum_{j=2}^{m-2} (C_j^2 (A_j Y_{i(j-1)} + B_j Y_{i(j+1)} - Y_{ij})^2) \tag{2}$$

where $A_j = (t_{(j+1)} - t_j)/(t_{(j+1)} - t_{(j-1)})$, $B_j = (t_j - t_{(j-1)})/(t_{(j+1)} - t_{(j-1)})$, and $C_j^2 = (A_j^2 + B_j^2 + 1)^{-1}$ [**?**].

Step 1: Transform time to go from 0 to 1. Let $0 = t_1 < t_2 < \cdots < t_m = 1$ denote the ordered time points.

Step 3: Choose $k \le m$, the smoothing parameter.

Step 4: Let $\phi_0(t) \equiv 1$, $\phi_1(t) = \sqrt{2}\cos(2\pi t)$, $\phi_j(t) = \sqrt{2}\cos(2j\pi t)$, etc. the cosine basis.

Define the matrix

$$\Phi = \begin{pmatrix} \phi_1(t_1) & \phi_2(t_1) & \cdots & \phi_k(t_1) \\ \phi_1(t_2) & \phi_2(t_2) & \cdots & \phi_k(t_2) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_1(t_p) & \phi_2(t_p) & \cdots & \phi_k(t_p) \end{pmatrix}.$$

Step 5: Now perform a Gram-Schmidt orthogonalization on the columns of Φ to make the columns orthogonal. Denote the new matrix by Ψ.

Step 6: Given a profile $Y_i = (Y_{i1}, \ldots, Y_{im})$, define $\hat{\theta}_i = (\hat{\theta}_{i1}, \ldots, \hat{\theta}_{ik})$ by $\hat{\theta}_{ir} = \frac{1}{m}\sum_{j=1}^{m} \psi_{rj} Y_{ij}$. Thus $\hat{\theta}_i$ is *the cosine transform* of Y_i. Note that $\hat{f}_i(t_j) = \sum_{r=1}^{k} \hat{\theta}_{ir}\psi_{rj}$ is a smoothed version of the profile Y_i.

The one detail left is to choose k, the smoothing parameter.

Risk estimation. The method of estimating the smoothing parameter is based on minimizing the estimated risk over the nested selection class. If we had many observations on a single curve, we would use the method in Beran, Dümbgen (1998) and Beran (2000) for estimating k. Our problem is slightly different because we have many curves but fewer observations per curve. Our idea is to combine information across curves to estimate the risk. More specifically, we group curves that are similar in distance from being constant and average their risk estimators. Here are the steps.

We first divide the set of curves ordered by the estimated variance over time in groups of 20 to 50 curves (otherwise we combine curves with very different amounts of wiggliness) and for each group C_g of curves we apply the algorithm below. Finally, we obtain an estimated smoothing parameter $\hat{k}_g$ for each group of curves, C_g. Then the mode $\hat{k} = mode_g \hat{k}_g$ is an estimate of the smoothing parameter for all the curves in the data.

Let $R(k) = \mathbf{E}\int(f(t) - \hat{f}(t))^2 dt$ denote the risk where $k = 1, \ldots, m-1$. We minimize the average risk within each group C_g containing N_g genes: $R(k) = \frac{1}{N_g}\sum_{i=1}^{N_g} R_i(k)$.

First,

$$\mathbf{E}(\hat{\theta}_{ir}) = \frac{1}{m}\sum_{j=1}^{m} f_i(t_j)\psi_r(t_j) \approx \int f(t)\psi_r(t) = \theta_{ir},$$

$$\mathbf{V}(\hat{\theta}_{ir}) = \frac{1}{m^2}\sum_{j=1}^{m} \sigma_i^2 \psi_r^2(t_j) \equiv \nu_r^2.$$

We can estimate ν_r^2 by inserting $\hat{\sigma}_i$ defined in step 0 for σ_i and obtain:

$$\hat{\nu}_{ir}^2 = \frac{1}{m^2}\sum_{j=1}^{m} \hat{\sigma}_i^2 \psi_r^2(t_j)$$

Now $\mathbf{V}(\hat{f}_i(t)) = \sum_{r=1}^{k} \nu_{ir}^2 \psi_r^2(t)$ so the integrated variance is

$$\mathcal{V}_i(k) \equiv \int \mathbf{V}(\hat{f}_i(t))dt = \sum_{r=1}^{k} \nu_{ir}^2.$$

and its estimate is $\hat{\mathcal{V}}_i(k) = \sum_{r=1}^{k} \hat{\nu}_{ir}^2$.

Also, $\mathbf{E}(\hat{f}_i(t)) = \sum_{r=1}^{k} \theta_{ir}\psi_r(t)$. Hence the bias is

$$b(t) = \mathbf{E}(\hat{f}_i(t)) - f_i(t) = \sum_{r=k+1}^{\infty} \theta_{ir}\psi_r(t)$$

and the integrated squared bias is

$$\mathcal{B}_i^2(k) = \int b^2(t)dt = \sum_{r=k+1}^{\infty} \theta_{ir}^2 \approx \sum_{r=k+1}^{m} \theta_{ir}^2.$$

Since $\mathbf{E}(\hat{\theta}_{ir}^2 - \nu_r^2) = \theta_{ir}^2$, an estimate of the squared bias is $\hat{\mathcal{B}}_i^2(k) = \sum_{r=k+1}^{m} \left(\hat{\theta}_{ir}^2 - \nu_r^2\right)_+$ (here we have taken the positive part to avoid negative estimates).
Finally, our estimate of $R(k)$ is

$$\hat{R}(k) = \frac{1}{N_g}\sum_{i=1}^{N_g} \hat{\mathcal{V}}_i(k) + \frac{1}{N_g}\sum_{i=1}^{N_g} \hat{\mathcal{B}}_i^2(k).$$

Now we choose $\hat{k}_g$ by minimizing $\hat{R}(k)$ over $1 \le k \le (m-1)$.

Cluster finding. We apply *k-means* to the cosine transform data. *k-means* is a non-hierarchical algorithm which clusters according to a distance measure which in our case is the Euclidean distance. The problem with non-hierarchical clustering methods is that the number of clusters has to be known a priori. There are different ways to identify the number of clusters in the literature, many of them based an ad hoc procedures. We estimate the number of clusters using *the gap method* [19].

The gap method. Tibshirani et al. proposes testing under the null hypothesis whether the number of clusters is 1 versus the alternative hypothesis, the number of clusters is greater than 1. The null distribution, called the reference distribution, is the uniform distribution under which a clustering method would provide only one cluster. The test statistic is called the "gap" statistics. Tibshirani et al. propose taking the null distribution to be a uniform distribution on the hyper-rectangle over the range of the observed data which is a rough approximation to the convex hull comprising the observed data. We consider the null distribution to be a uniform over the hyper-ellipsoid enclosing the observed data, a more accurate approximation than the hyper-rectangle.

3 Microarray Data Analysis

3.1 Can the microarray data support the analysis?

For all its strengths, microarray technology has its drawbacks. These start with the unstable mRNA in the cells and storage after sampling and end with errors in image analysis. All the limitations in microarray process add error to the signal. The variability between and within arrays

that is not due to the differential expression in the abundance of mRNA should be identified and removed. A large portion of the statistical analysis of microarray data involves identifying and quantifying sources of variation, and this is aimed at distinguishing between experimental error (noise) and inherent variability (signal) resulting from the actual biological phenomena under study. These sources of systematic variability and bias could lead to the misinterpretation of the expression data. In our study, we account for chip-to-chip variability using normalization. Additionally, we remove four sources of systematic variability found in our microarray data by developing four corresponding data filters.

Reuse filter. A first filter applied to the data is over arrays. We account for chip-use to chip-use systematic variability because we would expect random error added to the signal due to chip reuse. We keep only *the measurements which are reported on chips used the first time* because the intensity data reported on reused filters appear to be unreliable. This step reduces the data from 47 arrays to 20.

Poly(dT) tracts filter. We define by consecutive 5' T residues or poly-dT tracts a string of T's which appears in front of the nucleotide sequence of a gene. The impact of the long poly(dT) tract sequences as discussed in section 3.1.2 is to be taken into account. The large variability and the large intensity of those sequences prefixed by long poly(dT) tracts suggests that they are a source of systematic bias that is not explained by the biological process. Because the variance increases sharply starting with a poly-dT tract of length 11 the filtered data will contain only those genes whose *sequences are prefixed by a T-string of length smaller than 11* [12]. The data after removing these sequences contain 3824 DNA sequences.

Replicates filter. Another source of systematic variability is due to genes which are not constantly expressed over replicates. The replicates in our data are 5 measurements at treatment time 6 hours for test drug. However, the 5 arrays at 6 hours are reported on 5 different chips. Even though the normalization is designed to remove the variation from an array to another, we may still find genes which are not approximately constantly expressed over the 5 replicates. Thus the third filter to the data is as follows.

For each gene i, consider the intensity values replicated for 6 hours, $x_{i1}, .., x_{i5}$, and form the ratios of any combination of two intensities $r_{i_{kj}} = \frac{x_{ik}}{x_{ij}}$. Then all those *genes which have* $|max_{k=1,..,5;j=1,..,5}(r_{i_{kj}})| < 1.5$ are "good" genes and included in the set of genes to be further analyzed. The number of genes after first and second filters is 1150.

Control filter. We would expect to see genes which respond to DMSO, the chemical added to the drug troglitazone in the test drug treatment. Thus the genes which are differentially expressed under the test drug could respond either to the drug, troglitazone, or to the DMSO. In order to remove the effect of this chemical, we compare $C+$ (with DMSO) to $C-$ (without DMSO) and eliminate *the genes which are not constantly expressed under both controls*. In this way the genes which are responding to the chemical are eliminated and we are certain that the ones which are differentially expressed under the test drug respond to the drug and not to the DMSO. These genes are filtered out. We consider only those genes which have the ratio of the intensities from the two controls in use-1 data less than ± 1.5.

The data are reduced from 5355 DNA sequences to 1055, and from 47 measurements to 20. These are the data further analyzed.

We will now describe the two most important of our filters in more depth.

3.1.1 Reuse filter

The current data contain 27 reuses of the arrays (second and third uses). Thus a problem may be raised by the random variability due to filter reuse.

Comparing use-1 data to use-2 data. In Figure 3.1(a), the curves of the 20 genes with the highest activity across experimental conditions are plotted for both use-1 and use-2 data over ordered time.

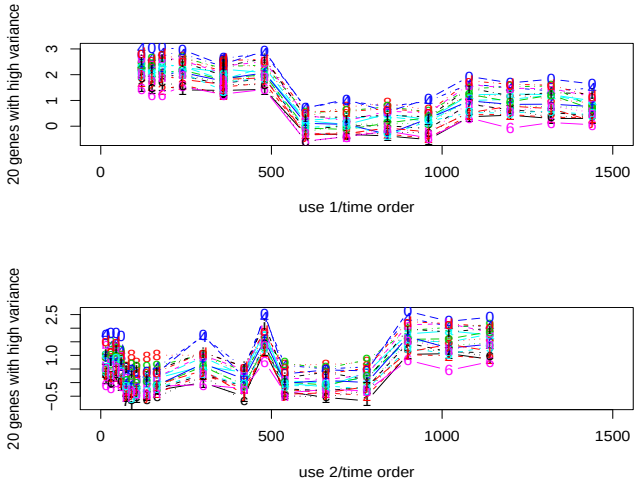

Figure 3.1 (a): *(a)Time series plots of the use-1 data and use-2 data of the 20 most variable genes.*

Another way to look at the time series plots is to smooth the expression profile. We smooth by applying a generalized additive model (smoothing splines)[13] with the response variable the use-1 data (top plot in Figure 3.1(b)) and then the use-2 data (bottom plot in Figure 3.1(b)) of the same set of 20 genes with the highest variability.

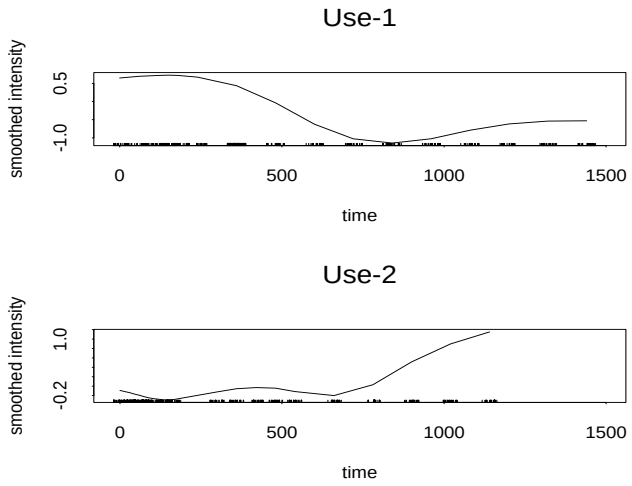

Figure 3.1 (b): *Smoothed time series plots from part (a).*

The plots show that there isn't a similar pattern over time under use-1 and use-2 for the 20 gene expression profiles. The difference in expression pattern over time for use-1 data and use-2 data implies that the biological process is not consistently replicated over the two uses. One would expect to see similar biological variation for the 20 genes under the same test drug treatment no matter the use.

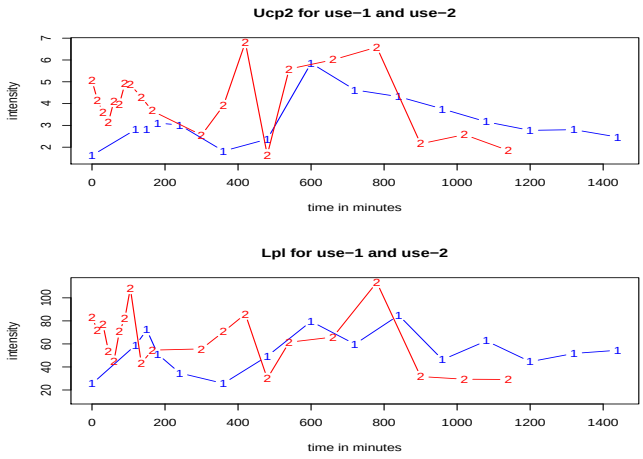

Figure 3.2: *Time series of Ucp2 and Lpl on the unlog scale; Use-1 is line 1 and use-2 is line 2.*

Northern blot for two genes. We conducted a different experimental method, Northern blot, for detecting and quantifying the RNA abundance of two sequences. The two are *uncoupling protein 2* (Ucp2) and *lipoprotein lipase*(Lpl).

The use-1 intensity data and use-2 intensity data curves over time for the two genes are are in Figure 3.2. The first line (blue) corresponds to use-1 intensities and the second line (red) represents the use-2 intensities over time. There are measurements at 8 hours for both use-1 and use-2 which allow a comparison between control and 8 hours over different uses. In use-1, there is a rise from control, $C-$ (at time 0 in the plot), to 8 hours (480 minutes). On the other hand, there is a depression from control, $C-$ to 8 hours for use-2.

The northern blot estimates for the two genes are consistent with the expression level over time obtained from the microarray experiment, use-1 arrays. There is a slightly amplified change from control to different hours in the Northern Blot method (for example, Northern blot reported 3 fold raise from control to 8 hours for Ucp2 compared to the time series plot in Figure 3.2 where control is around 0 and 8 hours is around 3). *Thus the Northern blot experiments agree with use-1 microarray data but conflict with the use-2 data for the two changing genes.*

We conclude that the use-2 and use-3 data appear to be unreliable and we didn't include them in our analysis.

3.1.2 Poly(dT) tracts filter

A poly(dT) tract sequence is prefixed by a T string: TTTTTTTTTTTTTTACAAATGTTATACATTTTT ATTTTTGTTCTTTTTGTAGGAAAAAAATACA...

The nucleotide sequences of the genes and EST's in our study were obtained from NCBI's Entrez nucleotide database. Browsing over the sequences of the 200 most active (variable) gene expression levels over experimental conditions, about 85% of them are prefixed by a sequence of consecutive 5' dT residues of length greater than 5. In contrast, 39% of the total number of sequences have poly(dT) tracts of length at least 5.

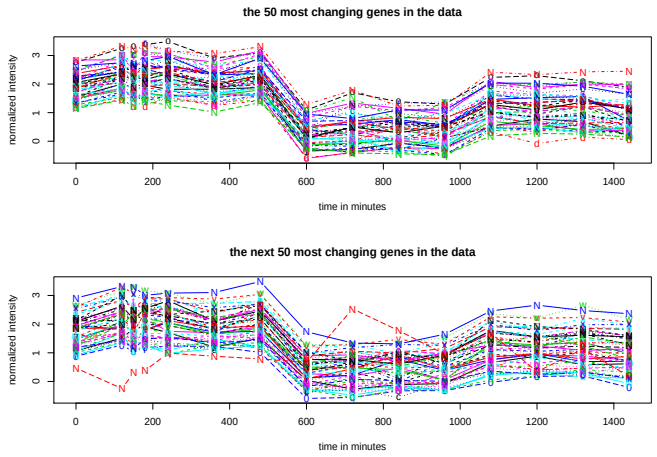

Figure 3.3: *The time series plots for the 100 sequences (50 in the upper plot and 50 in the bottom one) with the largest variance over all experimental conditions.*

Expression pattern similarity. Those sequences with large variance follow a similar pattern over time. Among the first 100 most variable sequences in these data, only one of them[1] is off the pattern as Figure 3.3 shows. On the other hand, only 10 of those 100 sequences don't contain long poly-dT tracts. One wouldn't expect to see such a similarity in expression pattern over time for the most

[1]It has the accession number AI428396 and starts with one T followed by a C.

changing DNA sequences. These results raise the question of whether there might be a systematic source of variation in the data *due to cross-hybridization* (see Figure 3.4).

We presented this artifact in the paper *Evidence of cross-hybridization artifact in Expressed Sequence Tags(ESTs) on cDNA microarrays*(2002), by Handley, D., Serban, N., Peters, D., O'Doherty, R., Field, M., Wasserman, L., Spirtes, P., Scheines, R., Glymour, C., submitted.

We concluded that this cross-hybridization is potentially an issue for any single-dye spotted cDNA microarrays. In the two-dye design, we would expect that any cross-hybridization would be equally (or nearly) represented by each dye, and therefore the resulting artifactual signal components would cancel. Our analysis of a two-dye data set supports this [12].

3.2 Which genes are responding to the drug treatment?

The two types of hypothesis testing described in Section 2.1 are applied to the microarray data with the following results.

2-time-difference test results. The test compares the normalized intensities measured at two different treatment times when one of them has to have replicates. Our microarray experiment has 5 replicates at 6 hours (arrays measured under treatment time of 6 hours and test drug). Thus it is possible to find candidate genes which change in the intensity level between 6 hours and 4 hours, 6 hours and 8 hours and so on.

Using the 2-time-difference test and FDR controlled at the level $\alpha = 0.1$ ($FDR \leq .1$), the 5 replicates at time 6 hours are compared to arrays of intensities at 3 hours, 4 hours, ..., 24 hours and a control array (Table 3.1). The 1st row consists of the experimental time points and the 2nd row consists of the number of significant gene expression levels according to the 2-time-difference test applied to the arrays at 6 hours and the array at the experimental time in the 1st row.

The intersection of the significant genes[2] from the tests *6 hours vs. 3 hours, 10 hours, 12 hours, 14 hours* contains 158 genes. There are 40 known genes among the 158 sequences.

The 40 genes but one, *glutathione transferase zeta 1 (maleylacetoacetate isomerase)* are not among the significant sequences when testing 6 hours vs. control $C-$, which has only two known sequences, the second one is the *adipsin*. Because they do not change significantly from control to 6 hours, but they change significantly from 6 hours to other different hours, these genes will be good candidates for the genes which change in intensity level from control $C-$ to the other different hours (except 6 hours). We take the intersection of sets of significant genes from testing 6 hours vs. the 4 treatment times because they are not very close to 6 hours but not very far away. One might expect no change in expression level in 2 hours (for example, 4 hours vs. 6 hours) or no significant change when measured between 6 hours and 20 hours. We also observe that the largest change happens around 10 hours. Thus our intersection is neither restrictive nor extremely wide.

[2]In this context, a significant gene is one whose intensity level changes between two time points. We hope that by capturing the change in intensity level we will identify those sequences which are differentially expressed.

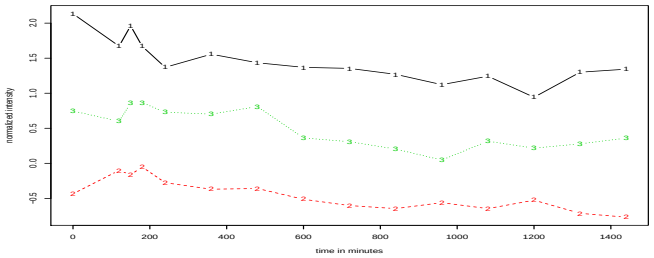

Figure 3.5: *Significant sequences from multiresolution runs test.*

Multiresolution runs test results. Even though the run tests are good tools for detecting patterns over time they are not powerful enough. With only 15 different time points, it's extremely difficult to differentiate between randomness and pattern.

We obtain 3 'significant' genes when the multiresolution runs test is applied at the level of significance $\alpha = .1$. All three are EST's. The plot of the expression profiles of the 3 sequences identified using multiresolution runs test is in Figure 3.5. Those three sequences have an approximately constant intensity over the last experimental time points. Thus the only nonrandom pattern detected with only 15 time points is the constant expression.

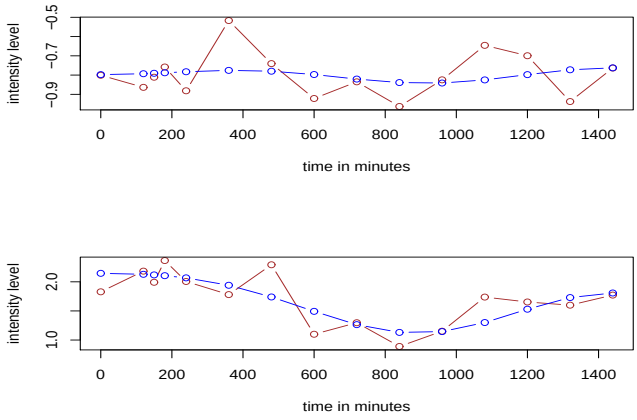

Figure 3.6: *Curves of two genes randomly chosen; for each gene, the observed (red line) and the smooth data with smoothing parameter $\hat{k} = 4$ (blue curve) is plotted.*

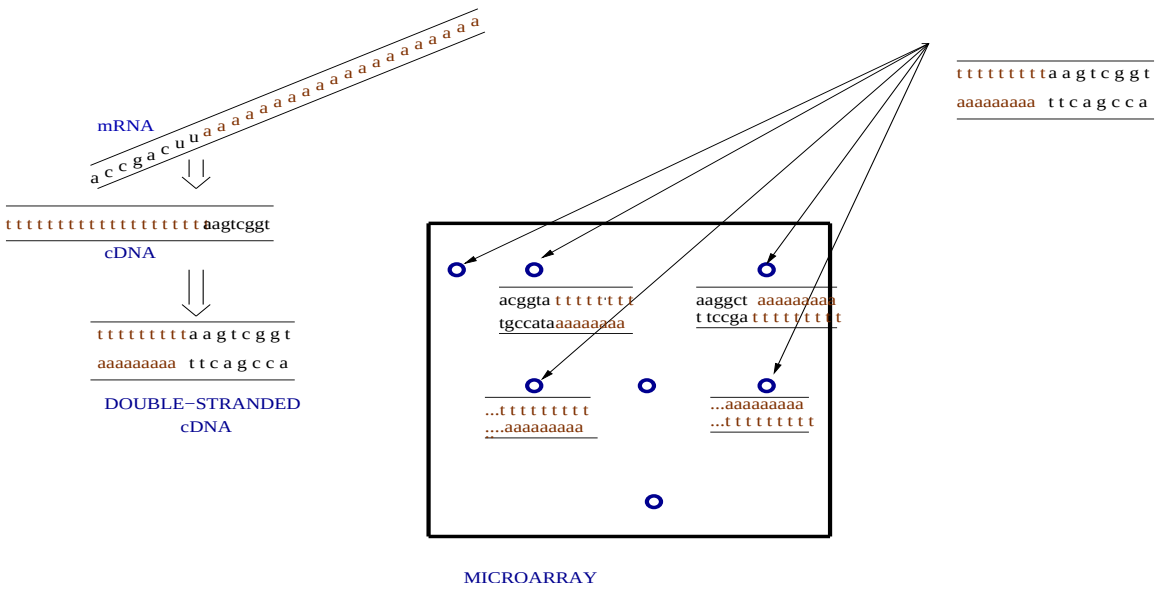

Figure 3.4: Cross-hybridization in our microarray experiment.

C-	2 hrs	2.5 hrs	3 hrs	4 hrs	8 hrs	10 hrs	12 hrs	14 hrs	16 hrs	18 hrs	20 hrs	22 hrs	24 hrs
21	322	408	354	282	249	456	427	353	383	218	242	298	250

Table 3.1: Testing with 2-time-difference test applied to the 1055 genes

3.3 Which Genes Vary Together?

The next step is to examine the extremes, e.g. sequences with significant differential expression. We apply the cluster analysis described in Section 2.2 to the 158 sequences (extremes).

Cosine - transform data. The smoothing parameter is estimated to be $\hat{k} = 4$. For two randomly chosen genes, the data (brown line) and the smoothed data (blue line) with the smoothing parameter $\hat{k} = 4$ are in Figure 3.6.

Identifying the number of clusters. The cosine transforms of the observed expression profiles are evaluated with the gap method (the clustering algorithm is k-means) in order to estimate the number of clusters. The number of clusters is estimated to be $\hat{K} = 2$ for both the set of significant sequences from 2-time-difference test (158 sequences) and the set of all sequences (1055 sequences). Thus we identified two main patterns. Among the 158 significant sequences, 134 fall in the first cluster and 24 in the second.

The average curves and smooth average curves over time of the significant sequences in each cluster are shown in Figure 3.7. According to these plots, the sequences in the first cluster have a rise around 8 - 12 hours, and the sequences in the second cluster have a depression around 8 - 10 hours. The separation between the 2 clusters obtained with the REACT clustering method is evident in the figure below. In this figure we plot the second estimated mean, $\hat{\theta}_2$, vs. the third estimated mean, $\hat{\theta}_3$, for each significant curve.

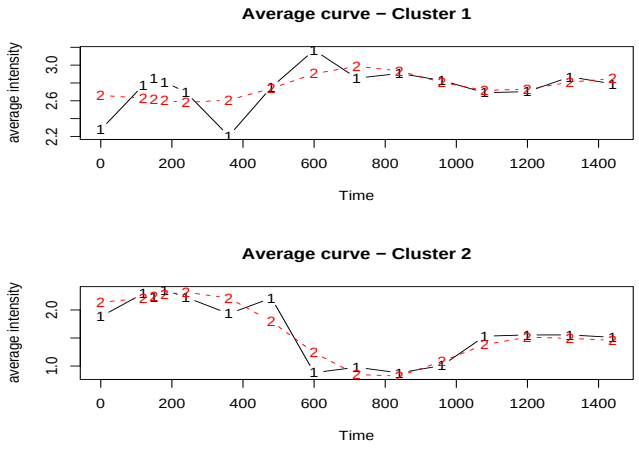

Figure 3.7: *Average intensity (black line) and smooth average intensity (red line) for the significant sequences in each cluster.*

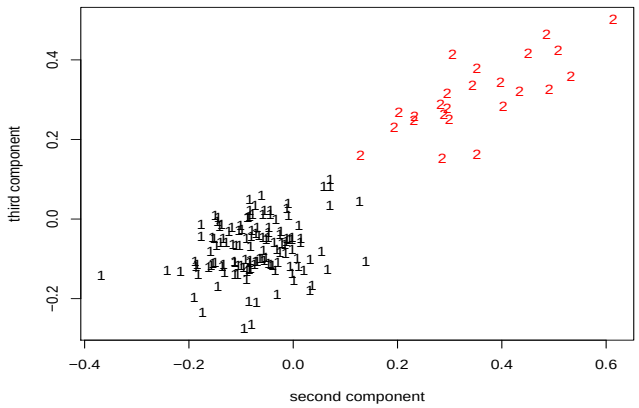

Figure 3.8: *The second vs. the third mean of the significant genes*

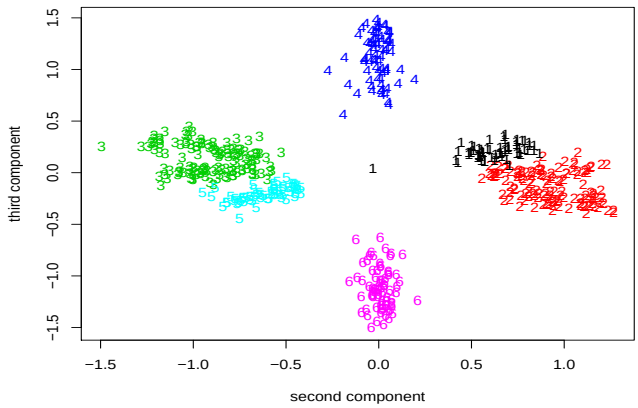

Figure 4.1: *The second vs. the third component for the significant curves*

4 Synthetic Data Analysis

Multiresolution runs test. Only the multiresolution runs test is applied to the synthetic data constructed in Section 1.2. The number of time points is large ($m = 30$) compared to the number of different experimental times in the microarray data ($m = 15$). There are 1400 curves with random signal and 600 curves with signal from one of the functions described in 1.2.

When the multiresolution runs test is applied to the synthetic data 468 curves are found to have a nonrandom pattern after FDR correction for multiplicity. None of them has a random signal (false positive). However, a high rate of false negatives (in this case, 132 false negatives) is the greatest disadvantage. An alternative is to conduct a larger number of experiments. For example, when $m = 35$, that is we have 35 different treatment times, the number of true positives increases to 537 and only 1 false positive. Thus with the cost of 5 other experiments, we gain 11.5% of the true positives. In real life applications, this gain may be extremely valuable.

Cluster Analysis. We cluster the significant curves in synthetic data with $m = 35$ (538 significant profiles over time where one is false positive). A first step is to estimate the smoothing parameter for those curves. With the method described in Section 2.2, we estimate $\hat{k} = 9$. The smoothing parameter is large because the variability of random error is small.

The gap method is applied to the cosine transform data of the synthetic data as described in 2.2. The number of cluster is estimated to be $\hat{K} = 6$. Then k-means on the cosine transform data with Euclidean distance clusters the curves coming from f_1 and f_2 together in one cluster and the negative curves coming from $(-1)f_1$ and $(-2)f_2$ in another cluster (two clusters so far). However, the curves coming from f_3 and f_4 (as well as for their negatives $(-1)f_3$ and $(-1)f_4$) are not clustered together. Thus the small perturbations in the first two curves are disregarded by our algorithm but the larger change at the late hours for f_3 and f_4 are not missed by the clustering method. It's worth mentioning that for a smaller number of time points ($m = 25$) the cluster algorithm doesn't identify the different pattern in the last two curves providing only 4 clusters.

The 6 clusters of the significant curves obtained with the REACT clustering method are very well defined across the second and the third estimated means, $\hat{\theta}_2$ and $\hat{\theta}_3$ (see Figure 4.1).

5 Conclusions

The methodologies introduced in this paper provide a means of identifying gene expression levels which change significantly under different experimental conditions and a means of estimating the cluster memberships of gene expression profiles. We present two novel methods of hypothesis testing and a novel algorithm for cluster analysis. We also present four systematic sources of variation and bias in microarray data from this level of experiment. We apply our statistical methods to a microarray data for treated adipose mouse cells and a complex synthetic dataset.

For the gene expression data, we identify two clusters of sequences which showed a change in expression around 8 to 14 hours. The known genes among those ones identified are enumerated in the Appendix. We consider only the middle treatment times because we believe that the cell

response to the drug we are testing for happens during this period of time. However, for differentiating expression levels which change over all treatment times we propose the multiresolution runs test. Unfortunately, we don't have a large number of treatment times to account for changes with this test. However, we anticipate that this approach will demonstrate its utility as experiments of this kind continue to produce even larger datasets.

The two genes for which we had quantified the mRNA abundance using Northern blot are among the significant genes obtained from the application of 2-time-difference test. According to their expression level measured by Northern blot, the two genes are differentially expressed. This supports our result.

Importantly, we screened out four sources of variability (sources of false positives) which are not due to differential expression. Because of the limitations of microarray experiments, this step is essential in the analysis. In this way, we eliminate false positives which could lead to misinterpretation in the reported results.

To prove the validity of our methods, we generate a complex synthetic dataset. The multiresolution runs test applied to these data shows to be a good tool to identify curve changes when a reasonable number of treatment times are considered. Additionally, the cluster analysis indicates to be a very good estimate of the clusters we consider in these synthetic data.

We clustered only significant genes because most of the procedures on estimating the number of clusters are highly biased by the presence of random variables [8] (in the temporal framework, a random variables is a random curve over time).

The C functions used in this paper as well as other new applications of multiple hypothesis and clustering we are currently developing are available upon request.

ACKNOWLEDGMENTS

The authors would like to thank to the members of the Gene Group, David Peters, Peter Spirtes, Dan Handley, Robert O'Doherty, Richard Scheines and Clark Glymour, for their valuable work and input on this project. This work was supported by NASA grant NCC2-1227.

6 Appendix

For biological interest we list the sets of genes identified by our analysis (genes whose expression level changes significantly according to 2-time-difference test).

Cluster 1: "tissue inhibitor of metalloproteinase 3", "zinc finger protein 46","ADP-ribosyltransferase 5", "capping protein alpha 1", "Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed (fox derived)", "branched chain keto acid dehydrogenase E1, beta polypeptide", "tumor necrosis factor receptor superfamily, member 7", "inhibin beta E", "nuclear receptor subfamily 4, group A, member 2", "zinc finger protein 147", "aldolase 1, A isoform", "isocitrate dehydrogenase 1 (NADP+), soluble", "Rous sarcoma oncogene", "ribosomal protein, mitochondrial, S7", "neuroblastoma ras oncogene", "ribosomal protein L27a", "lipoprotein lipase", "Ngfi-A binding protein 2", "ATPase-like vacuolar proton channel", "uncoupling protein 2, mitochondrial", "integral membrane protein 2 B", "Unc-51 like kinase 1 (C. elegans)", "solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 5", "calmodulin", "collapsin response mediator protein 1", "glutathione transferase zeta 1 (maleylacetoacetate isomerase)", "proteolipid protein (myelin)", "phosphoenolpyruvate carboxykinase 1, cytosolic", "ADP-ribosylarginine hydrolase", "transformed mouse 3T3 cell double minute 2", "chaperonin subunit 2 (beta)", "selenocysteine lyase", "ubiquitin-conjugating enzyme E2I", "NADH dehydrogenase flavoprotein 1", "interferon activated gene 203", "ubiquitin fusion degradation 1 like", "T-box 2"

Cluster 2: "gamma-glutamyl transpeptidase","inter-alpha trypsin inhibitor, heavy chain 2", "RAB11a, member RAS oncogene family"

References

[1] Benjamini, Y., Hochberg, Y. (1995), *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*, Journal of Royal Statistical Society, B, 57, 1.

[2] Beran, R. (2000), *REACT Scatterplot Smoothers: Superefficiency through basis economy*, JASA, 95, # 449, pp 155-171.

[3] Beran, R., Dúmbgen, L. (1998), *Modulation of estimators and confidence sets*, Annals of Statistics, 26, 5, pp 1826-1856.

[4] Bolnado, M. F., Lennon, G., Soares, M.B. (1996), *Normalization and subtraction: two approaches to facilitate gene discovery.* Genome Res. 6(9): 791-806.

[5] Brown, T.A.(1999), *Genomes*, John Wiley & Sons, NY.

[6] Dúmbgen, L., Johns, R. B. (2000). Confidence bands for isotonic median curves via sign tests. Preprint.

[7] Efron, B., Storey, J. D., Tibshirani, R.(July 2001), *Microarrays, Empirical Bayes Methods, and False Discovery Rates*, Journal of the American Statistical Association, 96.

[8] Fridlyand, J., Dudoit, S. (Sept 2001), *Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method*, technical report # 600.

[9] Genovese, C., Wasserman, L.(Dec 2001), *False Discovery Rates*, Technical report.

[10] Gasser, T., Sroka, L. , Jennen-Steinmetz, C. (1986), *Residual variance and residual pattern in nonlinear regres-*

sion, Biometrika, 73, 3, pp. 625-633.
[11] Gibbsons, J. D., Chakraborti, S. (1991), *Nonparametric Statistical Inference*, Marcel Dekker, Inc., 3rd Edition.
[12] Handley, D., Serban, N., Peters, D., O'Doherty, R., Field, M., Wasserman, L., Spirtes, P., Scheines, R., Glymour, C. (Nov. 2002), *Evidence of cross-hybridization artifact in expressed sequence tags (ESTs) on cDNA microarrays*, technical report.
[13] Hastie, T., Tibshirani, R.(1990), *Generalized Additive Models*, Chapman &Hall/CRC, Boca Raton.
[14] Hastie, T., Tibshirani, R.,Friedman, J.H.(2001), *The elements of Statistical Lerning: Data Mining and Prediction*, Springer Series in Statistics.
[15] Mosteller, F., Rourke, R. E. (1941), *Note on an application to quality control charts*, Annals of Mathematical Statistics, 12, 228-232.
[16] Pollard, K. S., Van del Lann, M. J. (2002), *A method to identify significant clusters in gene expression data.*, technical report.
[17] Storey, J. D.(2001, June), *A Direct Approach to False Discovery Rates*, Journal of Royal Statistical Society, B.
[18] Storey, J. D. and Tibshirani, R.(2002), *Estimating FDR under Dependence with Applications to DNA microarrays*, technical report
[19] Tibshirani, R., Walther, G., Hastie, T. (Dec 2000), *Estimating the number of clusters in a dataset via the Gap statistic.* Technical report, published in JRSSB,2000.
[20] Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P., *Normalization for cDNA Microarray Data*, technical report.
[21] Wichern, D. W., Johnson, R. A. (1982), *Applied Multivariate Statistical Analysis*, Prentice-Hall, NJ.

Web site references

[1] National Center for Biotechnology Information, Entrez search and retrieval system, *http://www.ncbi.nlm.nih.gov/Entrez/*