Analysis of microarray data for treated fat cells

Nicoleta Serban[*] Larry Wasserman[*], David Peters[†]
Peter Spirtes[§,¶] Robert O'Doherty[‡] Dan Handley[†,§]
Richard Scheines[§] Clark Glymour[¶,¶]

January 10, 2003

Abstract

DNA microarrays are perfectly suited for comparing gene expression in different populations of cells. An important application of microarray techniques is identifying genes which are activated by a particular drug of interest. This process will allow biologists to identify therapies targeted to particular diseases, and, eventually, to gain more knowledge about the biological processes in organisms. Such an application is described in this paper. It is focused on diabetes and obesity, which is a genetically heterogeneous disease, meaning that multiple defective genes are responsible for the diseases. The paper is divided in three parts, each dealing with a different problem addressed to our study. First we validate the data from our microarray experiment. We identified significant systematic sources of variability which are potentially issues for other microarray datasets. Second, we applied multiple hypothesis testing to identify differentially expressed genes. We found a set of genes which appear to change in expression level over time in response to a drug treatment. Third, we tried to address the problem of identification of co-expressed genes using cluster analysis. This last problem is still under discussion.

1 Overview

This paper is a statistical study which puts together *data validity* and investigation of genes which respond to the drug treatment. The analysis of the response to the drug treatment is based on two approaches: *multiple hypothesis testing* with rejection decision according to False Discovery Rate and *cluster analysis* of the cosine transformed expression profiles.

We first pre-process the data by normalization/transformation. An analysis of data validity is necessary because these data are being studied for the first time. It is important to

[*]Carnegie Mellon University, Department of Statistics
[†]University of Pittsburgh, School of Public Health, Human Genetics
[‡]University of Pittsburgh, School of Medicine, Department of Medicine and Department of Molecular Genetics and Biochemistry
[§]Carnegie Mellon University, Department of Philosophy
[¶]Institute for Human and Machine Cognition, University of West Florida

filter out the genes which don't provide reliable information, and eliminate incorrect directions in the further analysis. Related to data validity, we present an evidence of a significant cross-hybridization artifact on spotted single-dye cDNA microarrays. With the information gained in the data validity study, the two statistical methods, multiple hypothesis testing and clustering, are presented and applied to our microarray data. We consider three different applications of hypothesis testing to capture those genes which change in expression level under different hypothesis. They are *the 2-time-difference test*, *Wilcoxon rank sum test* and *runs test*. The identification of the differentially expressed genes is followed by their validation using a different experimental technology, *Northern blot*. 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 the cosine transformed data. In addition to global clustering, we test for *local clusters* and introduce a *cluster stability assessment procedure*.

This paper is organized as follows. In section 2, we present the biological and experimental background and the main questions of interest raised by the experiment, and describe the data format and data pre-processing. Section 3 is reserved entirely for data validation with a description of all the filters we applied to our data and a discussion of two of them. The problem of identification of differentially expressed genes is addressed in section 4. It is followed by a cluster analysis of the expression profiles in section 5. Section 4 and 5 have three parts: background on the method, the results of its application and a short discussion of the related results. A summary of the methods applied to the data and of our reconsiderations (section 6) finalize the analysis.

2 Introduction

2.1 Biological background and microarray technology

Gene expression. The biological information in a gene is read by proteins that initiate a series of biochemical reactions referred to as gene expression which is a two stages process: 1. transcription of the DNA into mRNA (with a pre-transcription into RNA and the mRNA is isolated from RNA) and 2. translation of the mRNA into the protein. The gene expression is measured as the abundance of mRNA. Not all the genes are expressed in all cells all the time. Thus we need to quantify the mRNA abundance of genes in order to identify those which are expressed under specific experimental conditions. This is what microarray technology tries to capture.

Microarray Technology. There are 4 main steps in quantifying mRNA abundance simultaneously for a set of genes with microarray technology: 1. mRNA extraction from the total RNA and reverse transcription into complementary DNA, 2. hybridization[1] to a microarray of complementary DNA, 3. geiger counter detection (scanning) of the microarray in order to capture the presence of bound DNA, and 4. interpretation of the scanned microarrays.The output is an image as the one in figure 2.1. Each spot corresponds to a DNA probe and the intensity of each spot measures the mRNA abundance of the corresponding DNA

[1]A cDNA probe binds to the spot where it finds the matching complementary sequence.

probe. Thus a measurement will provide *an array of intensities*, each intensity quantifying the mRNA abundance of a DNA probe. Each array comprises intensities for the same set of DNA probes.

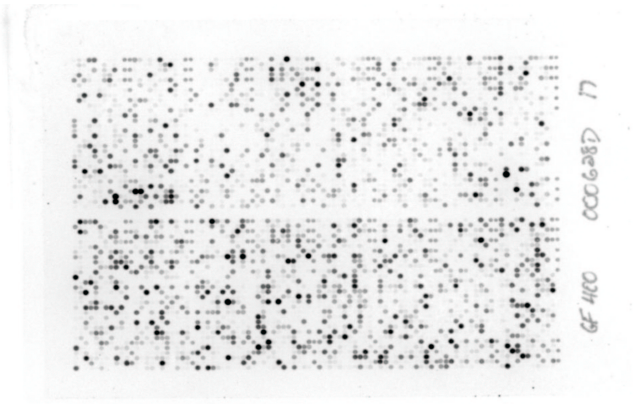


Figure 2.1: Microarray of one of the measurements in the experiment. Each spot corresponds to a DNA probe.

2.2 Background on the experiment and motivation

Why is our experiment significant? We study a family of drugs which are used in humans to treat **diabetes and obesity** by increasing insulin sensitivity. Decreased insulin 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 is understood by identifying which*genes change expression* in response to the drug treatment in adipose tissue. In short, if we find that expression level of gene A is increased in our experiment, the biologist will ask: Does increasing the expression of this gene 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 this gene. 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 a spotted cDNA microarray(Research Genetics, Carlsbad, California) experiment. The experiment was finalized on February, 2002 at the University of Pittsburgh (Peters et al. published data). The spotted cDNA microarray experiment consists of a time sequenced sampling of differential expression in mRNA from (3T3L1 cultured) fat cells originally obtained from mice. These cells were treated with a drug, troglitazone[2], which is a member of a family of drugs known as thiazolidendiones (TZD's). In our experiment, the drug treatment of the cells lasted for different periods of time ranging from 0 hours to 24 hours.

2.3 Scientific questions

For all its strengths, microarray technology has its drawbacks. They start with the fact that mRNA is unstable and they end with errors in scanning the images. All the limitations

[2]The drug was mixed with a chemical DMSO (detergent) which makes the drug soluble.

in microarray process add error to the signal. The variability between and within arrays that is not due to the difference in the abundance of mRNA should be identified and removed. It is difficult to identify all undesirable sources of variability, but at least it is possible to focus on those induced by the experiment.

In summary, 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.

2.4 Data

Genes. There are 5355 DNA probes which are hybridized to the microarray. Those probes are either fragments from DNA or total genomic DNA (tgDNA). Among all probes, there only 4696 sequences of DNA with identifiable sequences of nucleotides[3] and 139 probes consisting of tgDNA. Those fragments of DNA which are not genes are called expressed sequence tags (*EST's*)[4]. There are only 962 known genes[5]. For the ease of formulation, *all the probes are called genes* and the difference between EST's, genes and tgDNA will be pointed out any time it is necessary.

Measurements. The data consist of 47 measurements of mouse endothelial cells. These measurement were reported on 20 chips (filters), each chip being used at least twice (seven of them were used three times). 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 (number of arrays) and 5355 rows (number of probes).

Treatments. The fat cells were treated in three different ways, with a detergent called DMSO (*control C+*) or without (*control C-*), and *test drug*, a mixture of drug 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 will help in identification of any unexpected effect of DMSO.

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

[3]Nucleotides are basic molecular units which form the genetic code.

[4]ESTs are DNA sequences read from both ends of expressed gene fragments. ESTs are widely used in the discovery of new genes, and identification of coding regions in genomic sequences.

[5]A gene is a fragment of DNA which codes for a particular protein.

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 2.1: 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.

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 2.1. 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 and for the control with DMSO for chips 3 and 7, for the third use.

2.5 Preprocessing the data

Why do we transform and normalize? A rule of thumb in dealing with microarray data is to transform the data. There are a few reasons. One is that it makes *the data normalization additive*. Moreover, it makes *variation of intensities more independent of magnitude*, and gives a more realistic sense of variation as it is shown by the two histograms in figure 2.2 (a), first on the unlogged scale[6] and the second on the log scale of the same measurement. Another different reason is to attain constant variance across arrays. However, the constant variance cannot be realized without a good normalization to the data.

Next, we normalize the data in order to account for chip-to-chip variability. Individual microarray experiments may yield sampled intensities which are generally brighter or darker than similar intensities for a reference image. In order to compare different experiments, it is therefore necessary *to adjust for global shifts in intensity levels*.

What is our 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 react 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.

That is, for each measurement of the same sample the corresponding intensity value of each gene is divided by the median of the all intensities in the considered microarray. After logarithm transformation and normalization, an array is transformed into $\log(array) - median(\log(array))$ (figure 2.2 (b)). This appears to perform well because we expect only

[6]The scale cannot be finer because there are few genes with an extremely large intensity. The maximum is 21420 and the minimum is 12.54.

a relative small proportion of the genes will vary significantly in expression between mRNA samples[23].

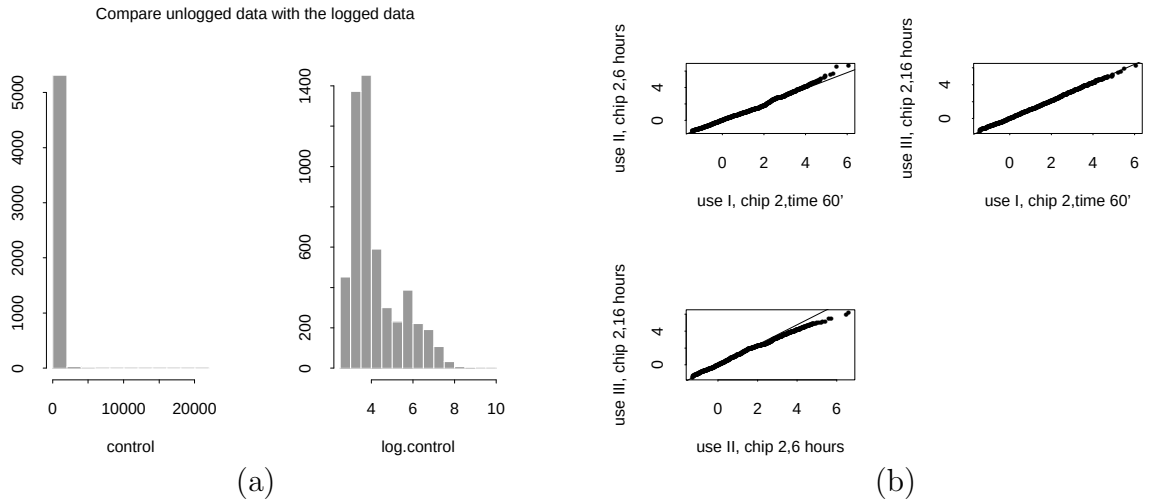


Figure 2.2: (a) Histogram of the intensity data from one array on unlogged (left) and logged scale (right);(b) The QQ-plots of the normalized measurements (log scale) for the chip 2 (different (different uses, control with DMSO).

3 Can the data support the analysis?

In the context of our microarray dataset, the main question to be addressed is to detect differentially mRNA abundance as cellular responses to the environmental change.

Systematic sources of variation. Microarray data typically contain many sources of variation, from unavoidable random error introduced during the performance of the experiment to measurement error associated with analysis of raw intensity data. 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.

We accounted for chip-to-chip variability using normalization in the previous section. Additionally, we identified four other sources of systematic variability or bias in our microarray data. Two of those sources of systematic variability are discussed. They are the systematic error in the data introduced by the filter reuse and by the sequences with long initial poly-dT tracts which is a systematic bias potentially common in single-dye spotted cDNA microarrays. The last two can be reduced by improving the experimental technology.

3.1 Filtering the data

The four source of systematic variability are removed as follows.

Reuse-Filter. A first filter applied to the data is over arrays. We accounted 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 will be discussed in section 3.2. This step reduced the data from 47 arrays to 20.

Poly(dT) tracts-Filter. The impact of the long poly-dT tracts sequences as discussed in section 3.3 is to be considered. The large variability of those genes which are prefixed by long poly-dT tracts suggests that they are a source of systematic variability that is not explained by the biological process. For this reason, these genes are filtered out and the normalization procedure is applied to the filtered data. 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*[11]. The data after removing these sequences contain 3824 DNA sequences.

Replicate-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 second filter to the data (3824 genes) 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 compared $C+$ (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 the genes which have the ratio of the intensities from the two controls in use-1 data less than ± 1.5.

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

3.2 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. There are several ways to compare use-1 data (the 20 arrays which are measured on filters used first time) to use-2 data. One way is plotting the expression levels for the same set of genes over time for use-1 data and use-2 data separately. This might be the most informative way to look at the two data sets expecting that *the time series of the expression levels of each gene will follow a similar pattern in use-1*

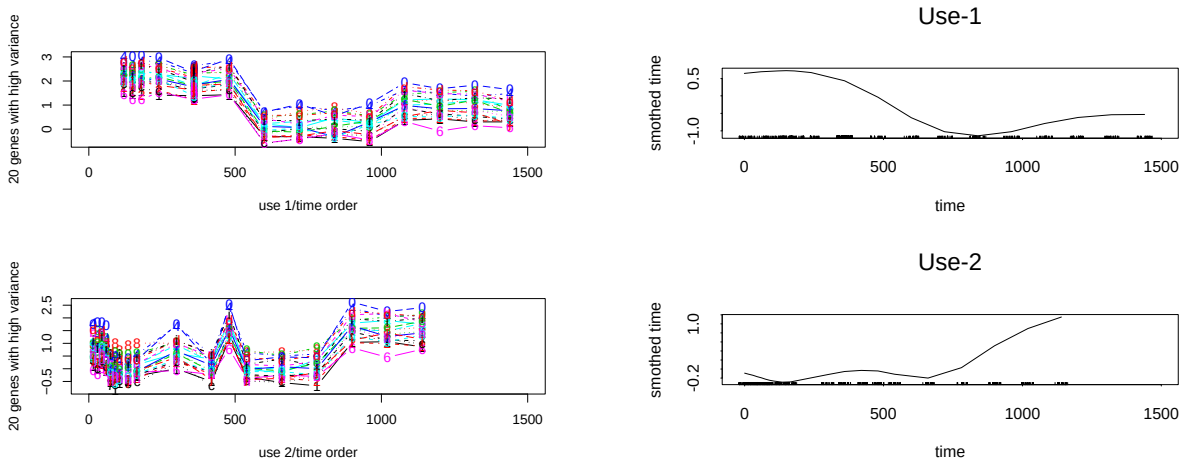
and use-2 data.

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 (in minutes).

It's interesting to observe that the expression profiles in use-2 plot (the second in the figure 3.1 (a)) follow a similar pattern to the ones in use-1 data but *with a shift in time.* On the other hand, there is a relationship between the chip number and time. For use-1 data the chips are ordered in time, i.e. chip 1 corresponds to the earlier time (150 minutes) and chip 20 corresponds to the latest time (24 hours). For the use-2 data there is a slight mismatch between chip number and time. Except for two chips, 2 and 20, all of the others are ordered by time. This difference might influence the shift shown in the curves.

Smoothed time curves for the 20 genes in use-1 and use-2 data. Another way to look at the time series plots is to smooth the pattern using generalized additive models[12]. Time is smoothed by applying generalized additive model (smoothing splines) 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 20 genes with the highest variability.

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



(a) (b)

Figure 3.1: (a)Time series plots of the use-1 data and use-2 data of the 20 most variable genes; (b) Smoothed time series plots from part (a);

Use-2 vs. use-1 vs. chip index. In this case, a more careful analysis of use-2 vs. use-1 data is necessary. Plotting the same 20 genes vs. chip indexes, the pattern in the use-2 plot follows the sha- pe in the use-1 plot (on a different scale only). The arrays in the control groups are included. *The concern is whether use-2 is a replication of use-1 with additional noise.*

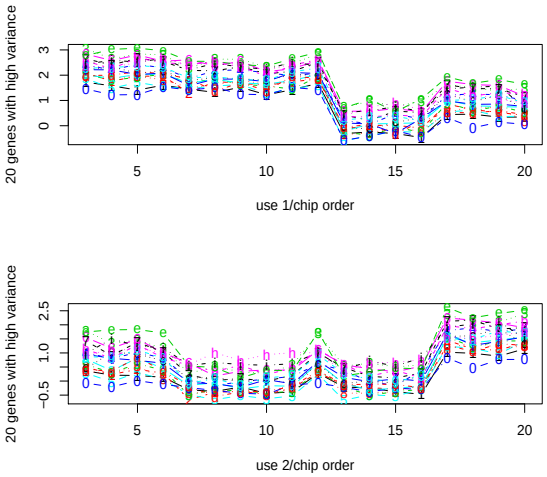


Figure 3.2: The time series plots of the 20 genes with the highest variance vs. the chip order for use-1 data and use-2 data

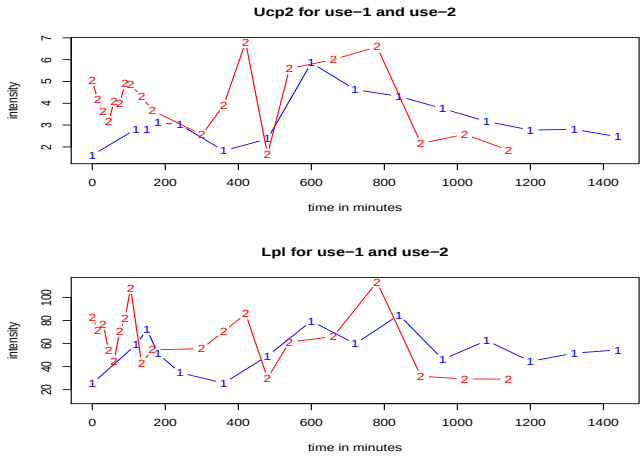


Figure 3.3: Time series of Ucp2 and Lpl on the unlog scale; Use-1 is line 1 and use-2 is line 2; intensities are normalized using filtered data only.

Changing genes. There are two known genes that *were observed being differentially expressed during the experiment.* They are *uncoupling protein 2* (Ucp2) and *lipoprotein lipase*(Lpl). In the use-1 data, the first gene, Ucp2, has the 488th largest variance and the second gene, Lpl, has the 166th largest variance. One would expect to see high activity for those genes which change significantly in expression level over time.

Changing genes in use-2 data for filtered data. The use-1 intensity data and use-2 intensity data curves over time for the two genes are are in figure 3.3. The first line (blue) corresponds to use-1 intensities and the second line (red) is the use-2 intensities. There are measurements at 8 hours for both use-1 and use-2 which allow a comparison between control and 8 hours treatments 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.

Northern blot on the two changing genes. We checked the difference in expression level for the two genes with the Northern Blot. (A description of this technique is in the Appendix.) The two genes showed a similar expression over time for the microarray experiment, use-1 arrays, and the Northern Blot experimental method. 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 plots in figure 3.3 where control is around 0 and 8 hours is around 3). *Thus the Northern blot experiments agree with use-1 microarray data but conflict the use-2 data for the two changing genes.* In this context, it might be useful to perform Northern blots on other genes in the data and validate the data from the three uses.

Is the filter reuse a significant source of variability in the data? The starting point of these data explorations is the source of variability due to chip reuse. The time series plots and the Northern blots for the two changing genes showed that the use-2 data needs a more detailed exploration before these data are analyzed for the identification of differentially expressed genes. For the moment, only the use-1 data is considered. The reliability of the intensity data from reused filters is *still an open question.*

3.3 Poly-dT tracts - Filter

Differentially expressed genes. The idea of ordering genes in the data according to their variance across different measurements (over time) comes from the fact that one would expect to see a quite large variance over time for the genes which change significantly in expression level. Of course, this variability could come either from different undetectable or detectable sources of variability or from the relevant signal.

How did we find the poly-dT[7] tracts to be an important source of variation? First, the sets of DNA sequences, whose expression levels were analyzed, had their sequences obtained from NCBI's Entrez nucleotide database. Browsing over the sequences of *the 200 most changing (variable) gene expression levels over experimental conditions, about 85% of them were prefixed by a sequence of consecutive 5' dT residues of length greater than 5.* In comparison, 39% of the total number of sequences had poly-dT tracts of length at least 5. Even going further to the 500 most variable genes in the expression data, we similarly find a large percentage of sequences with long poly-dT tracts.

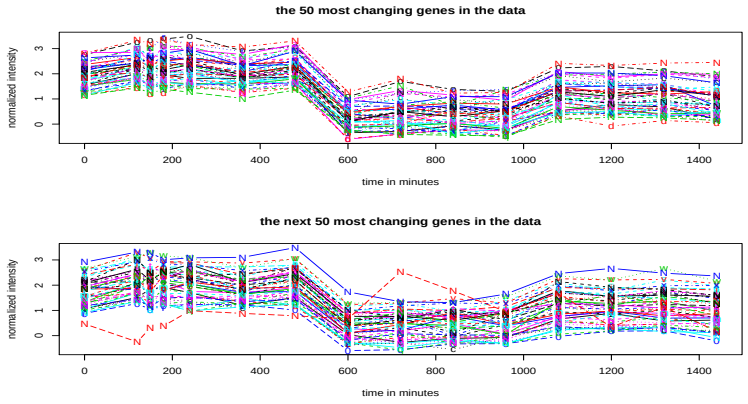


Figure 3.4: The time series plots for the 100 sequences (50 in the upper plot and 50 in the bottom one) with the largest variance under experimental conditions.

Similar pattern. 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**[8] *is off the pattern* as figure 3.4 shows. On the other hand, only 8 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 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**.

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

We conclude 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

[7]We defined 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.

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

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.

4 Which genes are responding to the drug treatment?

Next, we use multiple hypothesis testing to find a set of genes that respond to the drug treatment. In this context, we say that a gene responds to the test drug if it changes significantly in expression level over time.

4.1 Multiple hypothesis testing – Methods

Why multiple hypothesis testing? The gene expression data provided by the microarray experiments are large-scale data that capture the behavior of thousands of genes simultaneously. This suggests approaching multiple inferences. To control the multiplicity problem different methods were proposed. Among the most popular are Bonferroni correction, and more recently False Discovery Rate(FDR).

Comparison of Bonferroni correction and FDR. The Bonferroni method controls the probability of making one or more Type I errors among all hypotheses. The limitation of this procedure is the strong control. In many multiple inference problems the number of erroneous rejections should be considered and not only the question of whether any error was made. In this context, an error rate may be *the expected proportion of errors among the rejected hypotheses* ("discoveries") which defines the False Discovery Rate (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 [21]. That 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 depend in clusters[9].

FDR application. To use FDR, the computation of the p-values is necessary. For a detailed description of the procedure for the decision rule according to False Discovery Rate I refer to the Appendix (section 7.3). In the next subsections, three applications of hypotheses testing are applied with final goal of finding the p-value corresponding to each gene under specific assumptions. Finally, the rejection rule is a function of p-values according to FDR controlled at the level of significance $\alpha = 0.05$. In the last section, a discussion about modified FDR under a different estimate of a, Bonferroni correction and uncorrected case will be compared with respect to an application of the hypothesis testing considered in the analysis.

Only one application of hypothesis testing is described here. It is *the 2-time-difference test*. Other two applications are described in the Appendix. One is *the Wilcoxon rank sum test* which was previously applied to microarray data[5] (see section 7.4). The second is *the runs test* which often is used with time series data (see section 7.5).

The 2-time-difference test. This test provides candidates for genes whose expression level changes between two time points.

[9] "Genes tend to work in pathways, that is, small groups of genes interact to produce some overall process."

Assume that the genes are divided into two classes: genes which change between the two treatment times ("affected") and genes which don't change ("unaffected"). Thus the test for gene i is: H_{i0} : gene is unaffected vs. H_{i1} : gene i is affected for $i = 1, ..., g$ (where g is the number of genes which are tested). The task is *to test simultaneously* for all g genes if the difference in the expression levels at the same time point t_1 has the same distribution as the difference in the expression levels at the different time points t_1 and t_2.

P-value estimation in the 2-time-difference test. Assume that at time t_1 there are n replications (i.e. there are n 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 which is a vector of the form $(X_{i1}, X_{i2}, ..., X_{in})$. Consider 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 $1, ..., n$) as follows:

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

Define the differences in expression between time t_1 and time t_2

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

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

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

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

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

where $I()$ is the indicator function.

We have evidence against the "unaffected" hypothesis (H_0) for gene i if $P_i < P_{\hat{k}(.05)}$ according to FDR.

4.2 Multiple Hypothesis Testing–Results

In this section, I present the results according to each application of the hypothesis testing.

2-time-difference test. The 2-time-difference test was described in section 4.1. The test compares the normalized intensities for two different time points, one of them has to have replicates. Fortunately, 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 single differentially genes between 6 hours and 4 hours, 6 hours and 8 hours and so on.

Using the 2-time-difference test (FDR controlled at the significance level $\alpha = 0.05$), the 5 replicates at time 6 hours were compared to arrays of intensities at 3 hours, 4 hours, ..., 24 hours and 2 controls in the use-1 data (table 4.1). The 2nd column consists of the number

of genes which are significant for 2-time-difference test applied to 6 hours and each of the other time point. The 3rd and 4th columns indicate whether the two changing genes are significant (1) or not (0) based on the corresponding test.

Compare 6 hours vs.	# of signif genes	Ucp2	Lpl
2 hours	218	1	1
2 1/2 hours	313	1	1
3 hours	257	1	1
4 hours	140	1	1
8 hours	182	0	1
10 hours	360	1	1
12 hours	358	1	1
14 hours	270	1	1
16 hours	282	1	1
18 hours	153	1	1
20 hours	158	1	1
22 hours	227	1	1
24 hours	195	0	1
control C- 0 hours	4	0	0
control C+ 1 hour	158	0	1

Table 4.1: Testing with the 2-time-difference test applied to the screened 1055 genes

Agreement between the 2-time-difference test and time series plots. There is a close relationship between testing hypothesis results and the time series plots of the two *changing genes* (see figure 3.3). The time series plot of the Ucp2 gene shows that there is a mild rise from 6 hours to 8 hours, and roughly no change between 6 hours and control $C-$, and between 6 hours and 24 hours. This is also replicated by the test of 6 hours vs. control $C-$, 8 hours and 24 hours for Ucp2 (see column 3 in table 4.1). In any other case, there is a considerable change between 6 hours and any other time was tested against 6 hours which is also displayed in the time series plots.

Intersect the set of significant[10] genes according to 2-time-difference test. The intersection of the significant genes from the *tests 6 hours vs. 3 hours, 10 hours, 12 hours, 14 hours, 16 hours* contains 99 genes. There are *21 known genes* among the 99 from the intersection. The 21 genes found by intersecting the significant genes of the 5 tests are not among the genes which are significant from testing 6 hours to control $C-$, which has only one known gene, *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 expression level from control $C-$ to the other different hours (but 6 hours). I considered the intersection of sets of significant genes from testing 6 hours vs. the 5 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.

The list of 21 known genes and the list of the genes which change from $C-$ to 6 hours are in appendix.

Wilcoxon rank sum test. There are no time points other than 6 hours in the data with replicates on the same treatment. For this reason, the Wilcoxon test is applied only to compare the arrays before and after a single treatment time or to compare the replications at a time point (6 hours) to the data after or before that treatment time.

Testing for changes at 6 hours with Wilcoxon rank sum test. A first test compares the distribution of the 5 replicates at time 6 hours with the distribution of the arrays at 10 hours, 12 hours, 14 hours, 16 hours, 18 hours, 20 hours, 22 hours and 24 hours (two sided alternative) at the level $\alpha = 0.05$ (the level of significance at which we control

[10]In this context, a significant gene is one whose expression level changes between two time points.

FDR). There is a quite large number of genes which encounter significant change according to this test. About half of the genes (471 out of 1055) encounter significant change from 6 hours to the hours after. The same test is applied for one-sided alternative encountering for a depression (138 significant gene expression levels) or a rise (327 significant gene expression levels among them being the two changing genes, Ucp2 and Lpl) between 6 hours and the times after. Among the 471 genes, there are 87 known genes including the two changing genes, Ucp2 and Lpl.

Testing for changes at 8 hours with Wilcoxon rank sum test. A second approach to the Wilcoxon rank sum test is to consider a time point, let say 8 hours, and compare the arrays before and after this time point at the level $\alpha = 0.05$. Between 8 hours and 10 hours is when the genes with long initial poly-dT tracts showed a sharp depression. According to this test, 169 gene expression levels from which 38 known genes change significantly between the early time points and the late ones. The two changing genes don't significantly change according to this test.

Runs test There are different versions of this test. First, the sign vector corresponding to each gene is determined from comparison to the median level or from the differences of the intensities of consecutive treatment times. Then, there are two different ways to test the randomness, by using the number of runs or the length of the longest run. The level of significance for FDR control is $\alpha = 0.05$.

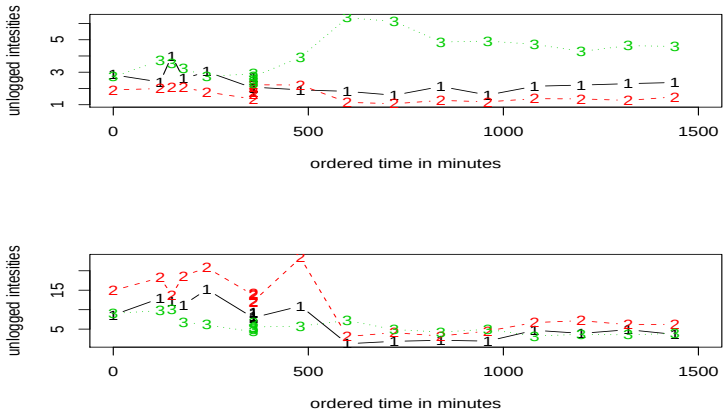


Figure 4.1: 6 genes which change in expression level (unlogged scale) over time according to runs test

The difference between the two rules is the following. When the magnitude of each gene expression level is compared to the median then the runs test is testing for fluctuations around the median level or a very long sequence of intensities above or below the median level (up-regulation or down-regulation). When the differences of the consecutive gene expression levels across time give the sign vector, then a large number of runs corresponds to a large number of rises and depressions, and a long run corresponds to a trend over time. The runs test based on the number of runs and on the longest run under the Rule 2 (signs of the differences of the consecutive time points) provide no significant genes. That is, there is not a significant trend over time for the gene expression levels and the pattern over time is random. On the other hand, there is no significant gene for the runs test based on the

number of runs using Rule 1, but there are 118 genes significant under the same Rule 1 but with the test based on the longest run. That is, all the genes are tested have random pattern in distribution above and below the median level, but there are significant genes which are up-regulated (above median level) or down-regulated (below median level) for a long time. The two changing genes are not significant according to this test.

There are 21 known genes whose expression levels change significantly according to runs test based on the longest run (see Appendix for a complete list).

4.3 Discussion

Common significant genes. Generally, the three tests don't agree on the set of significant genes. For example, the 2-time difference test (99 significant gene expression levels considered) and runs test (118 significant gene expression levels) have only 9 gene expression levels in common (only 1 known gene, *gamma-glutamyl transpeptidase*). However, a noticeable agreement exists between the 2-time-difference test and Wilcoxon rank sum test (6 hours compared to the time points after) set of changing gene expression levels including the ones for the two changing genes, Ucp2 and Lpl. This last agreement on the set of significant gene expression levels is explained by the fact that both tests compare roughly the same sets of arrays: replicates at 6 hours and the measurements for the time points after 6 hours. Thus the set of significant genes depends on the treatment times we test, on the hypothesis we make and the correction of multiplicity we consider.

Runs test vs. 2-time-difference test. The runs test (based on the longest run) provides genes whose expression levels are low for a while (over some of the time points) and then high or vice versa. This is different from what the 2-time-difference test tries to capture. The latter counts for the changes between time points over a period of time (considering the intersection), but these changes are not necessarily in the same direction. For example, the former test might give genes which are very low before 6 hours and very high after 6 hours. On the other hand, the latter test might provide genes which are depressed from 6 hours to 10 hours but rise from 6 hours to 12 hours.

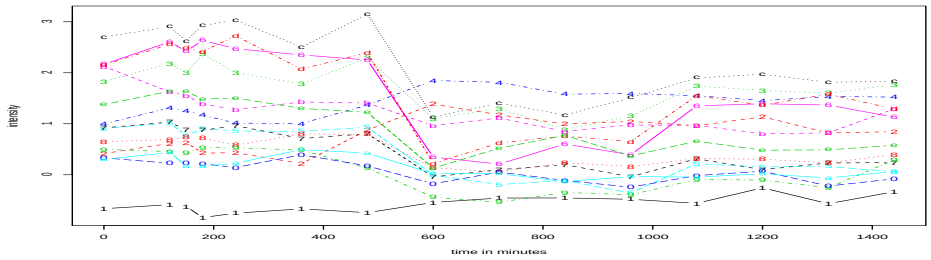


Figure 4.2: Common significant known genes from testing with runs test and Wilcoxon rank test applied to the arrays 3-8 hours vs. arrays 10-24 hours

Wilcoxon rank sum test vs. 2-time-difference test and runs test. There are only 12 genes in intersection of significant genes according to the 2-time-difference test and the significant genes according to Wilcoxon rank sum test when the arrays from the early treatment times (3-8 hours) are compared to the arrays of late times (10-24 hours). Two of them

are known genes: *"gamma-glutamyl transpeptidase"* and *"zinc finger protein 46"*. On the other hand, there are 99 common significant genes for the runs tests and the same Wilcoxon test. Among them 15 are known genes. Their expression curves are in figure 4.2. The large number of genes in the later intersection suggests that most of the genes identified with runs test are low before 8 hours with a rise after or are high before hours with a depression after.

Correction for multiple inference. We corrected the p-values provided by different applications of hypothesis testing with False Discovery Rate at the level of significance $\alpha = 0.05$ and in the conservative case, $a = 0$ (corresponding to the case when all the hypotheses are assumed to be null). We choose FDR with $a = 0$ because we don't want to be very restrictive (as in the case of Bonferroni correction) but we don't want to allow many genes in the set of significant genes (as in the case when we would estimate a). In table 4.2, different test corrections for multiplicity are compared. We control the set of p-values with False Discovery Rate under different estimates of a (see 7.2 for details), with Bonferroni correction as well as with uncorrected testing. The Bonferroni correction provides no significant genes according to our criteria. However, restricting the intersection only to the tests 6 hours vs 10 hours, 6 hours vs 12 hours, and 6 hours vs 14 hours (they provide the largest sets of significant genes), the intersection of the sets of significant genes consists of 36 genes (7 known genes including *lipoprotein lipase*, and *uncoupling protein 2*, mentioned in the appendix). Additionally, three genes, *adipsin and 2 EST's*, change in expression from 6 hours to control $C-$. At the other extreme, the uncorrected testing for multiplicity gives 164 gene expression levels which change from 6 hours to 3 hours, 10 hours etc. We also notice that the case of $\hat{a} = 0$ is the most conservative for the FDR correction.

control with	Bonferroni correction	$\hat{a} = 0$ FDR	$\hat{a} = max_{k=1}^{[n/2]} \frac{\hat{G}(k) - P_{(k)}}{1 - P_{(k)}}$ FDR	$\hat{a} = 2(\hat{G}(\frac{1}{2}) - \frac{1}{2})$ FDR	Uncorrected tests
number of significant genes	0	99	136	150	164
number of significant known genes	0	21	31	37	39

Table 4.2: The number of significant genes represent those genes which are change from 6 hours to 3 hours, 10 hours, 12 hours, 14 hours and 16 hours according to the 2-time-difference test and with the multiplicity correction specified in the first row.

The set of genes whose expression levels change significantly over time provided by the 2-time-difference test (which are a subset of those given by Wilcoxon rank sum test) and runs test need to be analyzed further. One experimental methodology is the Northern blot procedure mentioned earlier in the text and described briefly in the Appendix.

A statistical tool applied next is cluster analysis. The goal is to evaluate those significant genes according to their pattern over time and to identify other genes which vary similarly to the ones we found to be single differentially expressed or to identify common clusters for those genes.

5 Which Genes Vary Together?

The second approach in the analysis, clustering, is used to group genes with similar

patterns (co-expression). This similarity in pattern over time doesn't necessarily imply co-regulation of the genes in the same cluster. Grouping genes by co-regulation is a more complex and harder problem.

5.1 Cluster analysis – Methods

The main objective is to find at least *well-defined clusters of genes which are expressed similarly over time*. A better distance for quantifying the closeness between the two curves is the correlation coefficient rather than the Euclidean distance.

Clustering in the Fourier space. An alternative to clustering using correlation is to cluster in Fourier space. The profiles/curves of the gene expression levels over time are quite smooth. (The smoothness assumption implies that the curve functions are in a Sobolev space; estimating a function under a smoothness constraint can be reduced to the problem of estimating Normal means in an ellipsoid.)

We choose the Fourier transformation rather than "decorrelation" by Principal Component Analysis (PCA), which is a common approach to this problem, because each gene has a small number of observations over time (only 15 different time points). In this case, we would prefer to deal with a known orthogonal, orthonormal basis rather than estimating it from the data (as in PCA case).

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

Algorithm.

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 $Y_{ij} = f_i(t_j) + \sigma_j \epsilon_{ij}$ where $E(\epsilon_{ij}) = 0$.

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

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

where $A_j = \frac{t_{j+1} - t_j}{t_{j+1} - t_{j-1}}$, $B_j = \frac{t_j - t_{j-1}}{t_{j+1} - t_{j-1}}$, and $C_j^2 = (A_j^2 + B_j^2 + 1)^{-1}$[8].

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 smooth parameter (see below).

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

Step 5: Let

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

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 (we estimate $\hat{f}_i(t) = \sum_{r=1}^{k} \hat{\theta}_i r\psi_r(t)$) .
Step 7: Now cluster the $\hat{\theta}$ vectors using Euclidean distance. Omit the first component $\hat{\theta}_1$ (which corresponds to the average) from each $\hat{\theta}$ vector. This forces the clustering to use shape information only (a cluster will contain genes whose curves are similar but not necessary on the same scale).

One left detail is to choose k, the smooth parameter. We found this parameter using decision theory. The method is described in section 7.6 of the Appendix.

Clustering Routine. The clustering algorithm applied in this study is *k-means*. It is a non-hierarchical clustering method.

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, but in the current analysis only one is used. This is estimating the number of clusters using *the gap method* [22].

The gap method. Tibshirani et al. proposed 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 considered to be the uniform distribution under which a clustering method would provide only one cluster. The test statistics is called the "gap" statistics.

Tibshirani et al. proposed simulation from the hyper-rectangle over the range of the observed data which is a rougher approximation to the observed data convex hull than the hyper-ellipsoid. Along the same lines, the generated data could be also from a uniform distribution over the hyper-ellipsoid or hyper-rectangle shaped with the principal components of the data[22].

Simulation study on the the gap method. A simulation study with 8 different models according to Fridlyand & Dudoit[6] was performed. The gap method fails to find the number of clusters when there is overlap between the clusters and when there are noise variables in the data. Changing from hyper-rectangle to hyper-ellipsoid, not much improvement was gained (the 8 models are on low-dimensional data only).

Unfortunately, we would expect to see both overlap and noise variables in our data.

Local clustering. The gap procedure is a global test for determining the number of clusters in the data. We propose a test for local clustering in addition to a global clustering. With this procedure, we aim to find local clusters which are not identified by the gap method.

Our local clustering is based on hypothesis testing. The null hypothesis is that a set of data points is uniformly distributed over the hyper-ellipsoid which bounds the data points. The alternative hypothesis is that the data points come from a sum of two uniform distributions. This alternative hypothesis can be extended over a sum of more than two uniforms. However, we only want to test whether a set of data points forms a single cluster thus the alternative over two uniforms may suffice.

Simulation study on local clustering. I mentioned previously that the gap statistics fails when the clusters are overlapped or when noise variables are in the data. The same 8 models from Fridlyand & Dudoit[6] were applied to the local cluster test. As one would expect, the p-values are significantly large when the data is formed by three very well-separated clusters. On the other hand, irrelevant variables (noise), overlapping clusters and elongated clusters are handled by the local cluster test.

Cluster stability. The larger is the gap between those local clusters, the better is the clustering. We proposed to quantify the separation between local clusters by a stability assessment.

There exist several attempts to cluster stability assessment. One method developed for the microarray setting proposes to bootstrap the residuals from an ANOVA which would model the gene expression levels taking into account the gene effect, the variety effect (that could be the time) and other variables that may represent a source of inherent variation in the data (for example, array effect)[14]. Then the bootstrapped residuals would be added to the fitted values generating datasets with the same distribution as the observed one. Our stability algorithm (section 7.9 in Appendix) is along the same lines. We bootstrap from a multivariate normal distribution with the parameters given by the data to be clustered. Then we cluster the simulated data and count for those data points which stay in their cluster.

Our algorithm provides an estimate of the cluster stability (defined by the number of genes that stay in the cluster) and the corresponding standard error. In the meantime, the gene frequency in each cluster can be computed. This will give the proportion of genes which stay in the initial cluster 100% or 95% of the time, and, eventually, it will reveal overlapping (say, genes that are about 50% of the time in a cluster and 50% of the time in a different one). Thus we can see this stability assessment as a gap-overlap measure. In the case of a considerable overlap between clusters we may not separate the overlapped clusters.

5.2 Clustering analysis–Results

The cluster analysis is applied to two different forms of the expression data: the untransformed data and the cosine transformed data of the 1055 gene expression profiles. k-means routine with Euclidean distance is applied to these datasets.

Cosine transform data. The 1055 gene expression profiles are smoothed according to the procedure outlined in 5.1.

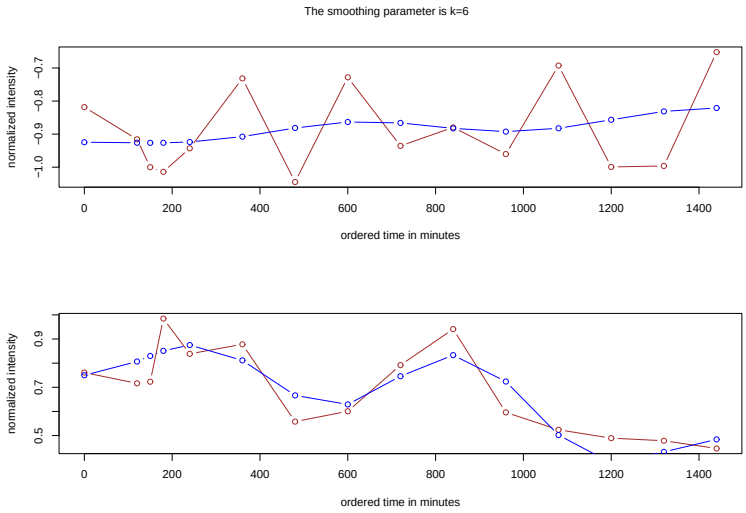


Figure 5.1: Curves of two genes randomly chosen; for each gene, the observed (red line) and the smooth data with smooth parameter $\hat{k} = 6$ (blue curve) is plotted.

First, the 1055 genes are divided in 20 groups (to account for different wiggliness in the curves), each containing about 52 genes. The mode is attained at $k = 6$ which is the smooth parameter for most of the curves in the data. For two randomly chosen genes, the data (brown line) and the smoothed data (blue line) with the smoothing parameter $k = 6$ are in figure 5.1.

Identifying the number of clusters. The untransformed expression profiles, and cosine transforms are evaluated with the gap method (the clustering algorithm is k-means) in order to estimate the number of clusters.

The first data set discussed is the *untransformed expression profiles*. The algorithm gap method is applied to intensity dataset for the 1055 profiles. The reference distribution is either uniform from the hyper-rectangle or from hyper-ellipsoid which covers the observed data. In the two cases, the number of clusters is estimated to be $\hat{K} = 16$ and, respectively, $\hat{K} = 3$. Then when reference distribution takes into account the shape of the data[22], the number of clusters is estimated with $\hat{K} = 6$ (hyper-rectangle) and $\hat{K} = 3$ (hyper-ellipsoid).

Second, the dataset is formed by the cosine transforms. For these *transformed data with all 6 components*, the number of clusters is estimated to be $\hat{K} = 6$ when the reference datasets are not corrected for the data shape (both hyper-rectangle and hyper-ellipsoid) and $\hat{K} = 3$ when the reference distribution takes into account the shape of the data. However, we omit the first component in each cosine transforms (the space dimension is 5). In this case, the gap method estimates the number of clusters as follows. When the reference datasets are uniform on the hyper-rectangle $\hat{K} = 3$. Then when the reference datasets are uniform on the ellipsoid, the number of cluster is estimated to be $\hat{K} = 2$.

Local clustering. The number of clusters is estimated to be $\hat{K} = 3$ for the untransformed dataset and to be $K = 2$ for the cosine transforms when the first component is omitted.

Next, the k-means algorithm and the local cluster test are applied in order to check for possible local clusters.

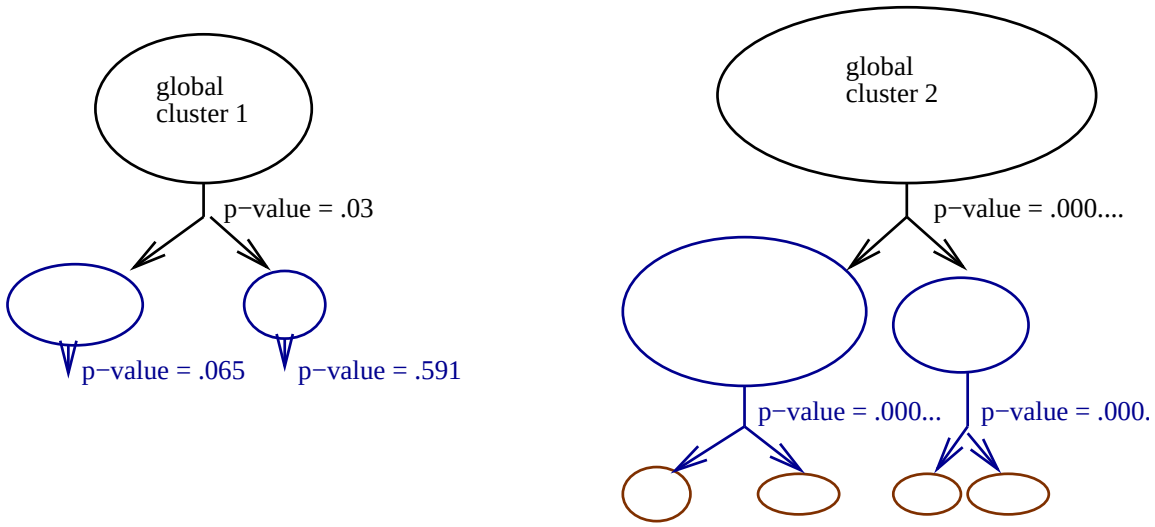


Figure 5.2: Clustering for transformed data according to gap statistics procedure (start with 2 clusters) and local testing, with a final number of 5 clusters at the level of significance 0.01[11]

The local cluster test applied to the untransformed intensity data (with 5000 reference datasets) gives p-values approximately 0 for the two largest clusters and about 0.0044 for the smallest cluster. Those small p-values indicates possible local clusters which are not identified by the gap statistics. We conclude that for each of the 3 global clusters identified by the gap statistics the test provides evidence against the null hypothesis. The p-values from testing the split of the six local clusters are large. *The final clustering with the untransformed data is into 6 clusters.*

Similarly, testing the null hypothesis of one cluster over the two global clusters obtained with the cosine transforms (omitting the first component) the p-values are as in figure 5.2. The final clustering with these data is: the first global cluster is not split but the second global cluster could be split into two and then into four other local clusters. However, the testing over the local clusters obtained from the first split is unnecessary. We may not want a significant overlap between clusters. The data are not so complex in structure that they need to be divided in very fine subclusters. We would like to capture a few patterns such as a depression followed by a rise or vice versa. This might be another reason to find clusters over the smoothed data. We keep the first global cluster and the two local clusters from the first split of the second global cluster, thus we obtained only *three clusters for the cosine transform data.*

Average curves in each cluster for the untransformed data. The average curves, over time, of the gene expression levels in each cluster are shown in figure 5.3 for each of the 6 clusters. There are three different plots in this figure. Each corresponds to a global cluster from which two local clusters are formed (only one split level). The curves were rescaled to the median 0 in order to be plotted on the same scale.

Stability for the untransformed data. The summary of the stability procedure is in tables 5.1 and 5.2). All 6 clusters have a high stability (higher than 90% for all of them). This fact shows an extremely small overlap within the hyper-balls estimated from the data points

[11]Tests are not corrected for multiplicity.

in each cluster. Similarly, all clusters have a quite large number of genes which fall in their cluster 100% of the time (among the 5000 simulations). This shows well-separated clusters. The two changing genes, Ucp2 and Lpl, have a similar pattern over time thus we would like to find them in the same cluster. However, they are in clusters 5 and 3, respectively.

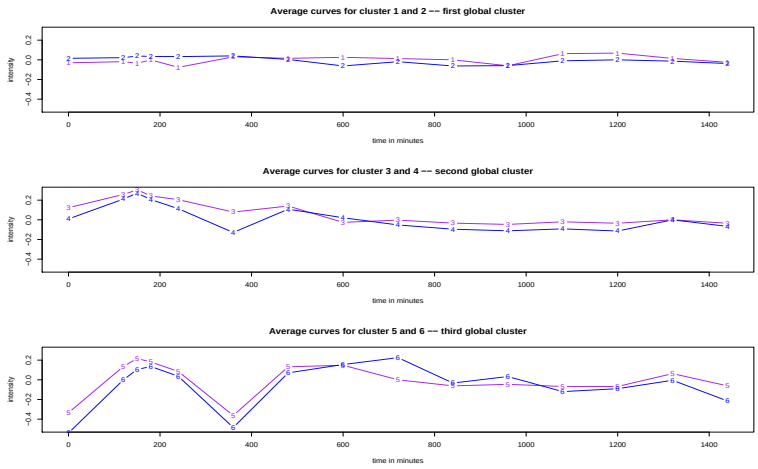


Figure 5.3: Rescaled average curves of the genes in each cluster over time for untransformed data; first plot corresponds to the clusters which are formed from the first global cluster etc.

Local vs. global clustering for untransformed data. The local clusters coming from the same global cluster have similar average expression pattern over time. However, the stability assessment shows that there is not much overlap between all 6 clusters, thus between clusters with the same global cluster. *Even though the local clusters have similar curves over time they are well-separated.*

cluster no.	size of the cluster	estimated cluster stability	estimated standard error of stability
1	340	99%	2.36
2	85	96%	2.1
3	263	90%	7.19
4	162	94%	5.28
5	141	93%	3.52
6	64	97%	2.5

Table 5.1: Estimated cluster stability for the untransformed data

cluster index	cluster size	no. genes that stays in the cluster 100% of the time	no. of genes that stays in the cluster 95% of the time	no. of genes that stays in the cluster 90% of the time	no. of genes that stays in the cluster 85% of the time
1	340	275	320	328	331
2	85	30	67	72	76
3	263	108	182	206	216
4	162	69	121	130	135
5	141	66	103	109	117
6	64	40	53	57	59

Table 5.2: The number of stable genes for each cluster according to the stability percentage

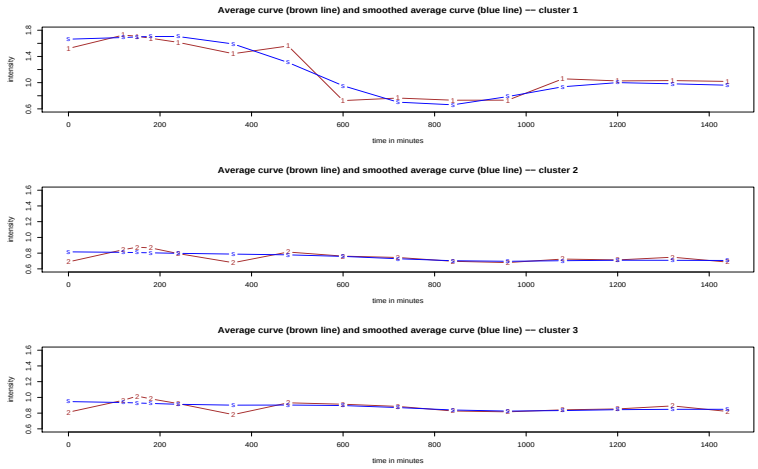


Figure 5.4: Average curves of the genes in local clusters over time for transformed data

Average curves in each cluster in the transformed data. The average curves over time of the genes in each cluster are shown in figure 5.4 for each of the 3 clusters. The two different curves represent the average curve over the gene expression level in each clusters (brown line) and the smoothed average curve (blue line).

Stability for the transformed data. Next, the stability of each cluster in the transformed data is estimated. First estimates of the cluster stability and their standard errors are computed over 5000 multivariate samples (table 5.3).

cluster no.	size of the cluster	estimated cluster stability	estimated standard error of stability
1	140	73%	34.3
2	649	32%	29.4
3	266	55%	9.4

Table 5.3: Estimated cluster stability for transformed data

Next, the gene stability is quantified.

cluster index	cluster size	no. of genes that stays in the cluster 90% of the time	no. of genes that stays in the cluster 80% of the time	no. of genes that stays in the cluster 60% of the time	no. of genes that stays in the cluster 50% of the time	overlap with clusters
1	140	19	57	103	124	–
2	649	0	0	8	97	clusters 3
3	266	19	52	136	168	cluster 2

Table 5.4: The number of stable genes for each cluster according to the stability percentage

There is no gene which falls in its cluster for at least 95% of the 5000 data simulations. Actually, there are extremely few genes which are stable in their cluster 80% (only 107 out of 1055).

Local vs. global clustering. The average curves for the local clusters coming from the same global cluster are similar. In this case, when we are specifically interested in changes in variation in the first hours or late hours, it would make sense to split the global clusters

to have a finer grouping of the genes. Otherwise, the global clustering we obtained with gap method is good enough. This is suggested by the low stability we obtained for the 3 clusters. We don't need to introduce unnecessary centers.

The curves of the two global clusters for the transformed data are in figure 5.5.

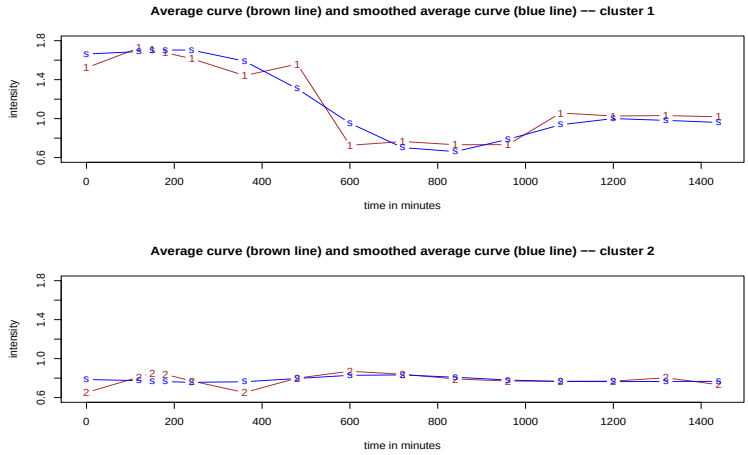


Figure 5.5: Average curves over the genes in each of the two global clusters; the brown line is the observed average curve and the blue line is the smoothed average curve

5.3 Discussion

Methodology. We are interested in finding well-defined clusters of genes whose expression levels have a similar pattern over time. We applied a classic clustering method, k-means. It is a non-hierarchical procedure for which the number of clusters is unknown. In our approach, we use the gap method as a tool of estimating the number of clusters in the data as it is described in Tibshirani et al., eventually with the change in the reference uniform distribution.

Euclidean distance vs. correlation coefficient. We propose a method which takes into account the correlation between curves. We believe that clustering the expression profiles has to be based on correlation and not on the Euclidean distance. The 99 significant genes according to 2-time-difference test are spread out over all six clusters obtained from untransformed data and Euclidean distance. Cluster 5 and 6 have most of the genes, 61 and 18, respectively. The plots of the significant genes according to their cluster are in figure 5.6. The four plots show that Euclidean distance is not a good measure of curve similarity.

The uniform over the hyper-ellipsoid vs. hyper-rectangle. In high dimensions, the difference between the hyper-rectangle (used by Tibshirani et al.) and hyper-ellipsoid which contains a given set of data points is large. (It increases with the number of dimensions.) Thus the convex hull which covers a set of data points is more closely estimated by a hyper-ellipsoid than by a hyper-rectangle in a high-dimensional space. A simulation study showed that the gap method fails to determine the number of clusters when there is overlap, and when some noise is added to the data. By replacing the reference uniform distribution on hyper-rectangle with the uniform on hyper-ellipsoid, there is not a significant gain in the

cluster estimation according to our simulation study in which we were actually considered data on a low-dimensional space only. However, a significant improvement is on the elongated clusters on a low-dimensional space where Fridlyand & Dudoit found their algorithm, *Clest*, very effective (in fact the only improvement over the gap method)[6].

In our case, the number of clusters appears to be better estimated when the uniform is from an hyper-ellipsoid. The number of clusters is $\hat{K} = 3$ for the hyper-rectangle reference datasets and $\hat{K} = 2$ for the hyper-ellipsoid reference datasets when the cosine transforms are clustered. The average curves for the three clusters are in figure 5.7. Cluster 1 and cluster 3 have similar average curves. From this point of view, we may want to merge the two clusters as the gap method with the reference datasets from hyper-ellipsoid estimates.

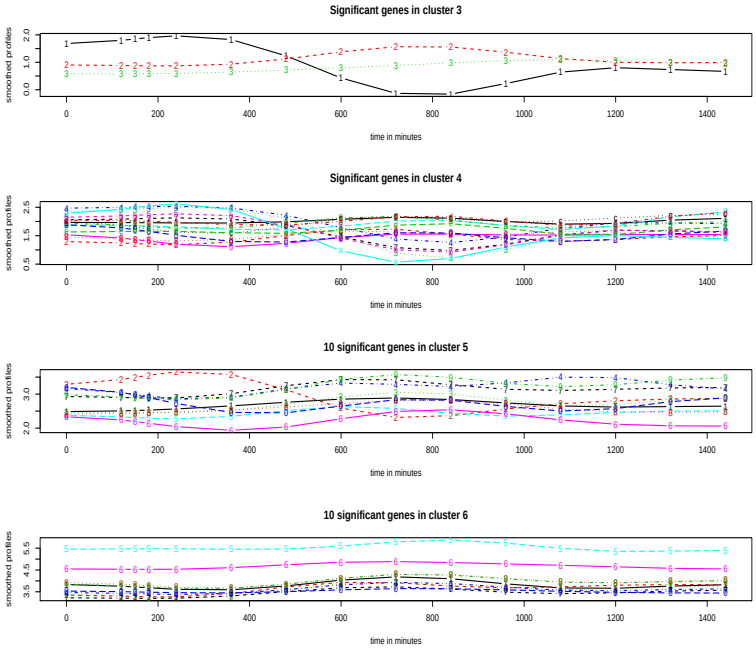


Figure 5.6: Smoothed expression profiles of the 99 significant genes clustered with k-means and Euclidean distance over untransformed data

Omitting the first component in the cosine transforms. Our cluster analysis was applied to the cosine transforms omitting the first component of the cosine transforms. There is a considerable difference between clustering with all the components in the cosine transforms and clustering with only 5 of them (omitting the first one). For example, when all six components are considered the gap method estimates 3 clusters instead of 2. The average curves for the 3 clusters in figure 5.8 are quite different from the ones in figure 5.7. Additionally, the smoothed expression profiles for the 99 significant genes from the multiple hypothesis testing reveal the importance of omitting the first component (see figure 5.9 and figure 5.10).

Global vs. local clustering. On the other hand, the gap procedure is a global estimate of the number of clusters which provides well-separated clusters but not local clusters. In figure 5.11, the three global clusters are the ones which are recognized by the gap procedure. The small local clusters can be identified by a lcoal testing as discussed in previous section.

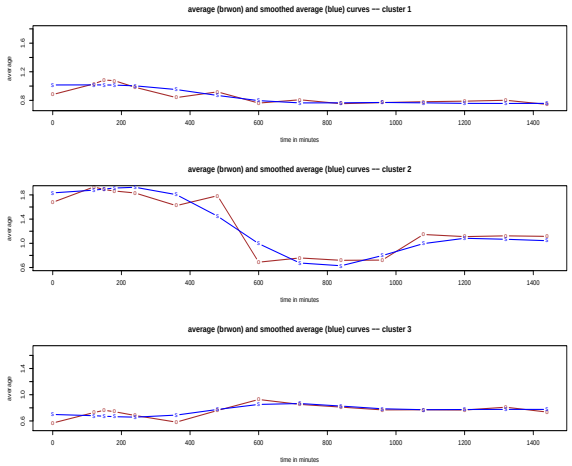


Figure 5.7: Average curves over the genes in each of the three clusters identified with the gap method when the reference datasets are from a uniform on a hyper-rectangle; the brown line is the observed average curve and the blue line is the smoothed average curve

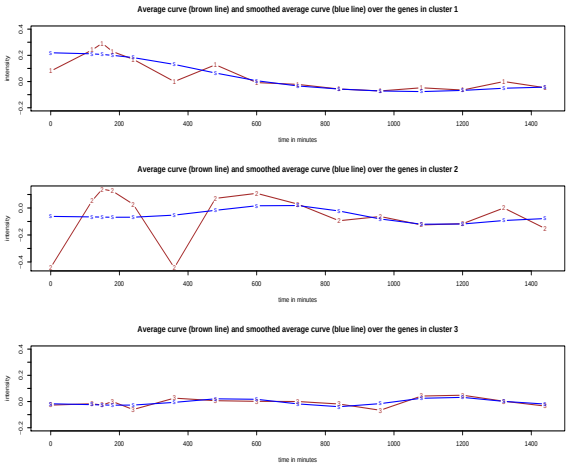


Figure 5.8: Average curves over the genes in each of the three global clusters; the brown line is the observed average curve and the blue line is the smoothed average curve

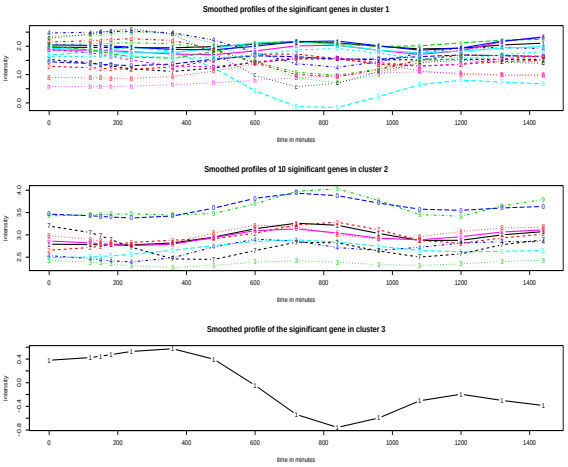


Figure 5.9: Smoothed expression profiles for the 99 significant genes according to 2-time-difference test and their clusters from clustering the cosine transforms **without omitting the first component**

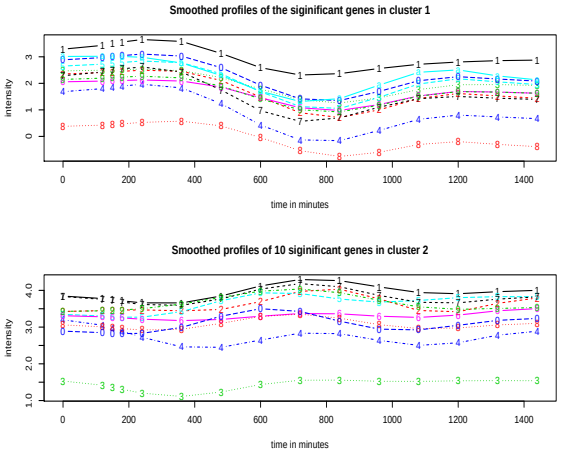


Figure 5.10: Smoothed expression profiles for the 99 significant genes according to the 2-time-difference test and their clusters from clustering the cosine transforms **omitting the first component**

We identified 3 global clusters for the untransformed and 2 global clusters for the cosine transforms (omitting the first component) with the gap method, and applied the local test to the those global clusters recursively. The clusters formed by splitting global clusters should have their average curves similar because they come from the same global cluster. It

is the case of clustering both the untransformed data and the cosine transforms (see figures 5.3 and 5.4).

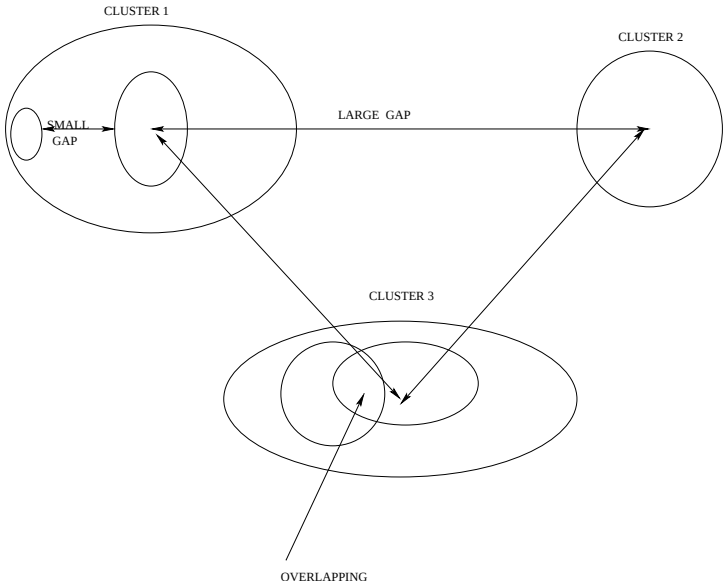


Figure 5.11: Three global clusters vs. small local clusters contained in the global clusters

Role of the cluster stability. The stability assessment showed a considerable overlap between the local clusters of the same global cluster for the transformed profiles. On the other hand, the 6 clusters coming from the untransformed data are quite stable. It followed that it might be a better clustering when we don't take into account the local clustering for the cosine transform data but not for the untransformed data. However, for a high-dimensional space, like the space where the untransformed data lie, one would expect to see a better separation between those hyper-balls (between the clusters). The transformed data in the Fourier space has only 5 dimensions. This is one third of the initial dimension (15 different time points). The difference in dimensions explains the overlap between the hyper-balls which cover the 3 clusters (unstable clusters under multivariate data samples).

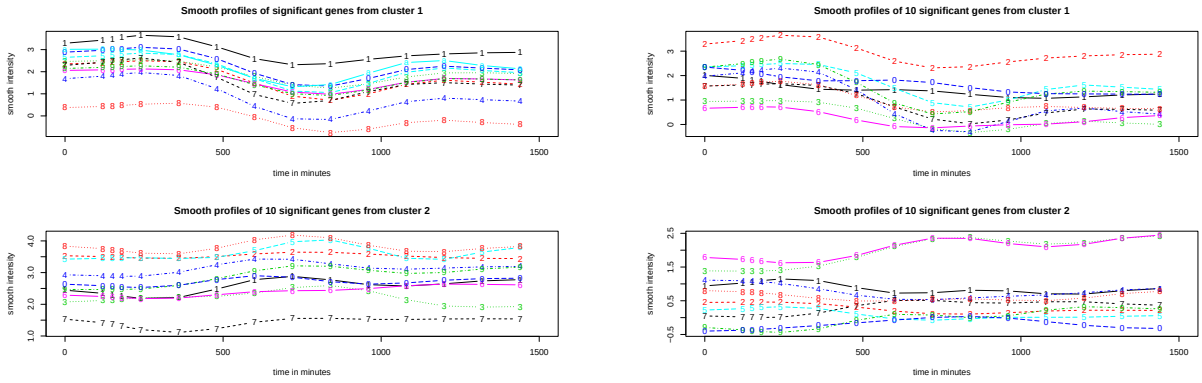
6 Summary

Statistical methodology. So far two different statistical methodologies were applied to the microarray data: multiple hypothesis testing and clustering.

Two tests, the 2-time-difference test and the runs test based on the longest run provide quite different sets of genes. From the first one, we've found a set of 21 known genes (99 genes and EST's) and from the second one we've found 21 known genes (118 genes and EST's) whose expression level changes over time. The two lists of known genes are provided in the appendix for those who have the biological background to interpret those results.

Moreover, a clustering algorithm was implemented with the final goal of finding well-defined clusters of genes whose expression levels over time have similar profiles. The clustering algorithm is applied to the transformed (normalized) data in the Fourier space to capture the similarity in expression profile. Two clusters are formed using the cosine transforms and omitting the first component.

The two statistical methods complement each other. That is, we might want to know if the genes whose expression levels significantly change over time share the same cluster and to find other genes which fall in the clusters where we find the genes of interest. For example, the two changing genes, uncoupling protein 2 (Ucp2) and lipoprotein lipase(Lpl) are in the second cluster which contain most of the significant genes according to 2-time-difference test. The significant genes grouped by their co-expression according to our cluster analysis are in figure 6.1 (a) and (b) (the known genes with the clusters they fall in are in appendix). The significant genes according to 2-time-difference test are quite well grouped by our clustering. For the first cluster the genes have a sharp depression before 14 hours and then a rise in expression level. The second cluster is formed by genes with an expression rise around 10-12 hours. On the other hand, most of the significant genes according to runs test are in the first cluster. Similarly, the first group contains those genes with a depression and genes with low variation in expression level. The second group consists of those genes with a rise in expression level or low variation in expression.



(a) (b)

Figure 6.1: (a) Smoothed expression profiles for the significant genes according to 2-time-difference test; the second plots contains the smoothed expression profiles for a sample of 10 genes out of 88; (b) Smoothed expression profiles for the significant genes (a sample of 10 genes in each cluster) according to runs test.

Thus a way to look at the clustering connected to the multiple hypothesis testing is to consider each of the single differentially expressed gene, known or unknown genes, reported by the hypothesis testing and check the cluster it is in. In this way we may find new genes which are co-expressed with the ones we have information about and give it a meaning in the biological process. For example, a biologist could test whether the 11 genes and EST's in cluster 1 which are differentially expressed according to the 2-time-difference test are truly co-expressed, and even better, whether they are truly co-regulated.

Filtering the data. We applied different filters to the data in order to remove the systematic variation and bias. Considering the two statistical methodologies applied to the filtered data, one would may wonder about the difference in results over the filtered and unfiltered data.

Use-2 data. We removed 27 arrays quantified on reused filters. Of course, we may have a better understanding of the biological variation over time with 47 measurements than with 20. Thus it is essential to prove (un)reliability on the 27 filters.

We saw that two genes which were observed as being differentially expressed during the performance of the experiment have quite different expression profiles from use-1 and use-2 data. Both use-1 and use-2 have measurements for the treatment time 8 hours and test drug. There are only 104 genes among 1055 (filtered data) and 649 genes among 3824 (data excluding the DNA sequences with long poly-dT tracts) which have the ratio of the normalized intensity in the two arrays larger than the threshold 1.5. We may conclude that there is not a significant difference between the two arrays (as one would expect when the two arrays were from the same use). On the other hand, 8-hour array is measured on the 20th chip under use-2 which corresponds to 24 hours under use-1. Additionally, 884 genes out of 1055 and 2773 genes out of 3824 have *both* the ratio of the intensity at 8 hours from *use-2* and intensity at 24 hours under use-1 and the ratio of the intensity at 8 hours from *use-1* and intensity at 24-hour array under use-1 smaller than 1.5. This shows that it might be possible that the signal from use 1 to be carried through use-2. Two significant genes according to the 2-time-difference test have their curves over time from use-1 and use-2 separately in figure 7.1. The first gene has equal intensities at 8 hours but different intensities at control $C-$ which corresponds to 0 hours from the two uses. On the other hand, the early hours are quite different for the first gene. The smoothed profiles of the same gene shows a quite different expression for the use-1 data and use-2 data. For the second gene, there is not approximately equal expression levels for 8 hours but the curves over time are similar (on a different scale).

Compare 6 hours vs.	Gene Set	Known genes	Ucp2	Lpl
1/4 hours	341	81	1	1
1/2 hours	334	79	1	1
3/4 hours	0	0	0	0
1 hours	0	0	0	0
1 1/4 hours	0	0	0	0
1 1/2 hours	0	0	0	0
1 3/4 hours	0	0	0	0
2 1/4 hours	348	99	1	1
2 3/4 hours	352	80	1	1
5 hours	215	45	0	1
7 hours	505	106	1	1
8 hours	160	34	0	0
9 hours	461	92	1	1
11 hours	484	104	1	1
13 hours	434	90	1	1
15 hours	174	35	0	1
17 hours	281	60	1	1
19 hours	157	26	0	0
control C- 0 hours	396	99	1	1
control C+ 6 hour	368	85	1	1

Table 7.1: Testing with the 2-time-difference test applied to the screened 1055 genes

Impact of use-2 data. Table 7.1 provides the number of genes which change significantly according to the 2-time-difference test from 6 hours (measured under use-1) and all the other 20 measurements from use-2. In this case, a large number of genes, 368, change in expression level from 6 hours to the control $C-$ as well as for very early treatment times (1/4 and 1/2 hours). Then a dramatic change happens at 3/4 hours until 2 1/4 hours when no gene changes in expression level as compared to 6 hours. Additionally, about half of the genes are changing in expression level from the 6 hours at use-1 and 7 hours at use-2.

Reliability of intensity data from reused chips. Based on the analysis of the intensity data from reused chips in section 3.2 and these reconsiderations, the intensity data from reused filters appear to be unreliable. It is possible that some of chips are not affected by the reuse. It is also possible that only some of the DNA probes on the microarrays to be altered due to reuse. In any case, it is not safe to draw conclusions from the data which is

partially unreliable.

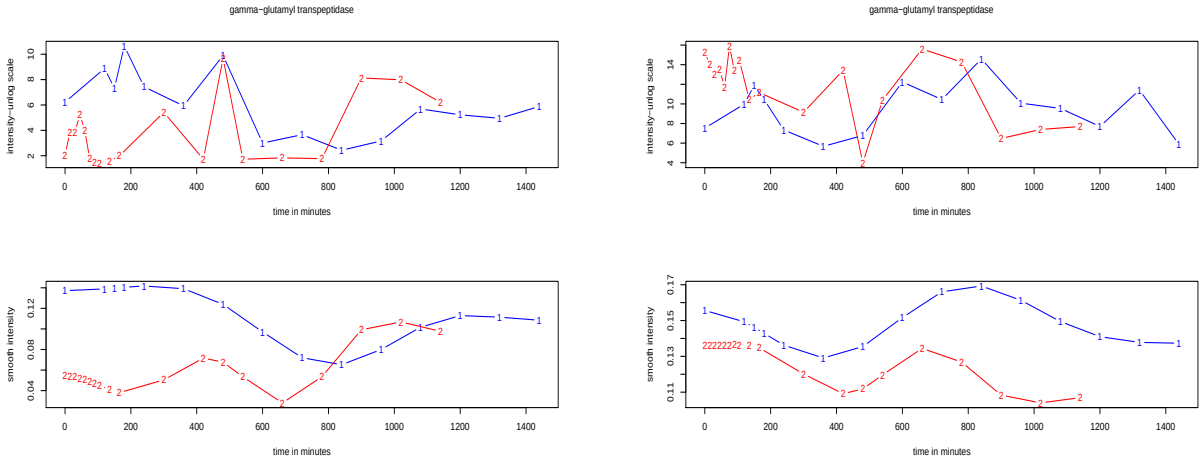


Figure 6.1: Time series on the unlogged scale of two significant genes; the blue line is the curve over the first use and the brown curve is over the second use in upper plots for the two genes; the bottom plots are the smoothed profiles of the same genes for use-1 and use-2 curves.

Impact of poly-dT tracts. Another important issue we found to have an impact on our analysis is the variability of DNA sequences with long initial poly-dT tracts. When those sequences *are not removed* the intersection of significant genes from testing 6 hours vs. 3 hours, 10 hours, 12 hours, 14 hours, and 16 hours contains 165 genes (19 known genes). In this set of significant genes there are 71 which are among the 99 identified with the filtered data and the same 2-time-difference test. There are three known genes which are among the 19 but not among the 21 significant genes found with the 1055 DNA sequences. They are *"macrophage expressed gene 1", "upstream transcription factor 2", "hydroxysteroid 17-beta dehydrogenase 1"*.

Applying the Wilcoxon test with the alternative that genes are down-regulated between 8 hours and 10 hours, 554 genes in the set of 5355 DNA sequences are significant according to this test (with the rejection rule according to FDR controlled at the level of significance 0.05). Among those 339 have initial poly-dT tracts longer than 10 and 435 have poly-dT tracts of any length. In contrast, there are only 123 genes which are significant when the same test is applied to the set of genes excluding those ones with long initial poly-dT tracts.

Thus the DNA sequences with long initial poly-dT tracts affect the multiple hypothesis testing. Additionally, we would expect to see a different clustering of the unfiltered gene set.

Validity of the known genes we found. There are reasons we may want to validate our DNA sequences we found. The most important one is the experimental error which so greatly affects the inherent variability. We identified several sources of systematic errors but we don't know whether some other sources of systematic error are still present. On the other hand, the rejection rule in the context of multiple hypothesis testing is based on False Discovery Rate controlled at the level of significance $\alpha = .05$. Thus, *in average*, we would make 1 out of 20 false positive. Thus we are not 100% certain about our results. The experimental methodology we use to validate our results is Northern Blot which was applied to the two 'changing genes'.

Cluster Analysis. The cluster analysis we proposed can be improved upon. Instead of applying it to the cosine transform data with the Euclidean distance, we can also cluster the smoothed profiles using the correlation coefficient as the distance between two data points. Further, a reliability assessment and a simulation study are important to rely on the conclusions we make about co-expression we identified with our cluster analysis.

These are the next steps in our analysis of the microarray data for treated fat cells and maybe different other experiments.

7 Appendix

7.1 Correlation analysis with respect to reuse issue

Correlation between arrays. The correlation coefficients for each pair of arrays which are measured on the same filter, different use (use-1 vs. use-2) are summarized as following (the first coefficient corresponds to the correlation between the array reported on chip-1, use-1 vs. array experimented on chip-1, use-2 and so on):

```
0.93 0.96 0.97 0.97 0.97 0.97 0.91 0.91 0.89 0.91 0.90 0.98 0.98 0.97
0.98 0.97 0.98 0.98 0.99 0.97
```

The lowest correlation happened to be for chip 9 which has use-1 at 6 hours and second use at 105 minutes, quite far away from each other. The largest correlation is for chip 19 which has use-1 at 22 hours and use-2 at 19 hours, reasonably close to each other. Thus *the high correlation between arrays coming from the same filter* could be induced by the time difference in use-1 and use-2. However, the correlation coefficients are quite large supporting the previous statement.

Use-1 vs. use-2 correlation by gene. Another way to look at the correlation is to calculate the correlation coefficient for each gene given its data from use-1 vs. the data from use-2 *ordered by the chip index*. These coefficients reflect the correlation between use-1 vs. use-2 according to the chip variable for each gene. There are in total 188 DNA fragments with correlation coefficients larger than 0.8. The summary of the correlation coefficients for all genes is as follows.

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
-0.665	0.159	0.370	0.348	0.555	0.936

Even the 100 genes with the highest variance have those coefficients in the interval (-0.356,0.410). A large correlation would imply that the measurements reported on first-use chips replicate a similar expression profile *over chip number* to the one reported on second-use filters. However, *the correlation coefficients are not extremely large* to make this assertion. From this point view, it is uncertain if the use-2 filters report similar signal to use-1 filters (because the filters weren't cleaned correctly after first use, for instance).

7.2 Northern blot

The Northern blot is a method of detecting and quantifying specific RNA sequences. RNA is separated by molecular weight on an electrophoretic gel, then transferred to a membrane where it is allowed to hybridize with an oligonucleotide probe specific to the RNA of interest. Excess probe is washed off, and the probe that remains is visualized through either an immunohistochemical stain or a radioactive tracer. The presence of a band on a Northern blot indicates RNA of the correct (or approximately so) expected nucleotide sequence, and the intensity of the band gives an indication of the amount of RNA present. The position of the band on the membrane is compared against known molecular weight markers, and is used to help verify the RNA detected is of the expected size.

7.3 Multiple hypothesis testing – False Discovery Rate

Suppose we test n identical hypothesis tests with independent statistics (or "loose dependence when n large) T_1,..,T_n. The n test define the same rejection region. Each test will provide a *p-value* (P_i of the i^{th} test).

Define the variable $H_i = 0$ if the i^{th} null hypothesis is true, and 1 otherwise. The $Pr(H_i = 0) = a$ and $Pr(H_i = 1) = 1-a$ are given a priori, following that $H_i \approx Bernoulli(a)$.

The distribution of the p-values under the null hypothesis is uniform: $P_i|H_i = 0 \approx U(0,1)$. The distribution of the p-values under the alternative hypothesis is usually unknown and is estimated by the data: $P_i|H_i = 1 \approx F$. Then the marginal distribution of P_i is: $G = (1-a)U + aF$.

The false discovery rate is defined as

$$FDR = E\left[\frac{V}{R}\middle| R > 0\right] P(R > 0)$$

where V is the number of rejections among the true null hypothesis and R is the number of total rejections. "The Benjamini & Hochberg method guarantees that $E[FDR] = (1-a)\alpha$ in the continuous case, which is conservative because E[FDR] is controlled below the level α".

The threshold that define the rejection rule for the multiple testing is estimated by the *sequential p-values method to control FDR* [1]:
Step 1: Set α (the level of significance at which FDR is controlled).
Step 2: Compute the p-values of n tests based on independent test statistics.
Step 3: Order the p-values: $p_{(1)}, ..., p_{(n)}$.
Step 4: Choose $\hat{k}_\alpha$ such that $\hat{k}_\alpha = argmax\{p_{k)} \leq \frac{i}{n}\frac{\alpha}{p_0}\}$ where p_0 is the proportion of true null hypotheses.
Step 5: Reject the null H_{0k} if $p_k \leq p_{\hat{k}_\alpha}$.

Modified FDR. Under the mixture model, $G = (1-a)U + aF$ and the B&H threshold $\hat{k}$ is equivalent to: $T_{BH} = sup\{t : G(t) = \frac{t}{\alpha}\}$.
In fact, a more powerful test is obtained when the threshold is given by:
$sup\{t : G(t) = \frac{(1-a)t}{\alpha}\} = sup\{t : \frac{(1-a)t}{G(t)} = \alpha\}$ because $E[FDR] \leq (1-a)\alpha$ [20,7].
In the modified FDR, all the steps in the algorithm described above are the same but adding

a step (2') in which a is estimated by $\hat{a} = \frac{\hat{G}(p)-F_0(p)}{1-F_0(p)} = \frac{\hat{G}(p)-p}{1-p}$ where p can be chosen as $p = \frac{1}{2}$[20]. This is a good estimate because for any p in (0,1), the following inequality is true: $a \geq \frac{G(p)-F_0(p)}{1-F_0(p)}$. Another estimate of a could be taken as $\hat{a} = max_{k=1}^{[n/2]} \frac{\hat{G}(k)-P_{(k)}}{1-P_{(k)}}$ which is based on the same inequality[7]. Also, replace in step 4, α by $\frac{\alpha}{1-\hat{a}}$.

7.4 Application of hypothesis testing – Wilcoxon rank sum test

What does it test? This test can differentiate genes which are potentially up-regulated or down-regulated at one time point t, comparing the measurements before t to the ones after t. It can also test if genes change from a time point to another comparing the replications provided for each of the time points (it will require replicates at both treatment times).

Setting. Suppose for each gene i, two independent samples of size n and m, $(X_{i1}, .., X_{in})$ and $(Y_{i1}, ..., Y_{im})$ for each gene $i = 1, .., g$ which are observations from two continuous distributions F_i, and G_i, respectively.
The null hypothesis is $H_0 : F_i \approx G_i$ (identical distributions) vs. a two-sided ($H_1 : G_i(x) = F_i(x - \theta), \theta \neq 0$) or one-sided alternative such as $H_1 : H_1 : G_i(x) = F_i(x - \theta), \theta > 0$ (gene i is up-regulated at the given time point t).

P-values estimation. Merge the two samples into one sample of length $N = (n + m)$, sort it, assign ranks to the sorted values (giving the average rank to any 'tied' observation) and define $R_{i1}, R_{i2}, ..., R_{in}$, the positions (ranks) taken by X_i's in the combined sequence of length N. The Wilcoxon rank-sum test statistic is $W_N = \sum_{k=1}^{n} R_{ik}$.

The distribution of the test of lengths (n, m) is given by the Wilcoxon distribution, $W(n, m)$, when n and m are small. The exact distribution can be expressed by a recursive scheme[9]. For large m and n, the distribution of W_N under H_0 can be approximated by the normal distribution with mean $E(W_N) = \frac{m(N+1)}{2}$ and variance $V(W_N) = \frac{m \times n(N+1)}{12}$.

The p-values are computed with $Pr(W_N < W_{i,obs})$ for one-sided alternative $H_1 : G_i(x) = F_i(x - \theta), \theta > 0$, for example.

Then the p-values are used in defining the rejection rule according to FDR.

7.5 Application of hypothesis testing – Runs test

What does it test? Tests based on runs test for randomness in the data against the tendency of genes to cluster or tendency of genes to mix. They can be applied to both qualitative and quantitative observations, and they are good tools to test randomness in time series. With runs test, we will capture genes whose expression changes over all treatment times. Here, the change is defined according to the test setting.

Settings. First the intensity vector over ordered time for each gene is transformed in a sign vector that will contain (+1)'s and (-1)'s according to a specific rule. Two rules are considered:

Rule 1: For each gene, find the median of the intensities over time. Then replace the intensity value with (+1) if it is above the median level and (-1) if it is below the median level.

Rule 2: For each gene, take the difference of consecutive time observations $X_i - X_{i-1}$ and assign (-1) if the difference is negative and (+1) if the difference is positive.

The definition of a "run" is a consecutive sequences of (+1)'s or (-1)'s only.

The null hypothesis, H_0, is 'gene i is random' vs. the two-sided alternative, H_a, 'gene i is not random' or one-sided alternative depending on the specific runs test used. The runs tests are based either on the number of runs or on the longest run. There is a third type runs test called runs up and down test which won't be discussed here [9].

P-values estimation. The exact distribution under the null hypothesis of the runs test based on the number of runs is easier to define than for the runs test based on the longest run.

The distribution of the number of runs R (Feller's distribution) can be derived from the joint distribution of the number of runs of (-1)'s, R_1, and the number of runs of (+1)'s, R_2. That is, given n_1 the number of (-1)'s and n_2 the number (+1)'s, then the probability of having an even R number of runs is: $P(R \text{ runs}) = 2\frac{\binom{n_1-1}{r/2-1}\binom{n_2-1}{R/2-1}}{\binom{n_1+n_2}{n_1}}$ and of having an odd R number of runs is: $P(R \text{ runs}) = 2\frac{\binom{n_1-1}{(R-1)/2}\binom{n_2-1}{(R-3)/2}+\binom{n_1-1}{(R-3)/2}\binom{n_2-1}{(R-1)/2}}{\binom{n_1+n_2}{n_1}}$.

The distribution of the length of the longest run K of (-1) or (+1) can be derived from the joint distribution of the R_{ij}, the number of runs of length j corresponding to the two types $i = 1, 2$. This distribution is not described here. It's a complicated sum over many indexes [9,16].

The alternative hypothesis can be one-sided or two-sided. For the test based on the number of runs, if the one-sided alternative is that the number of runs is large - a tendency of like genes to cluster - then $p-value = Pr(R \leq r.observed)$ and if the one-sided alternative is the the number of runs is small - tendency of genes to mix - then $p - value = Pr(R \geq r.observed)$. Similarly, the p-values are computed for the test based on the longest run according to the alternative.

7.6 Cluster Analysis – Clustering routines

The clustering routines applied in our cluster analysis is k-means. It is a non-hierarchical clustering method for which the number of clusters needs to be specified.

K-means clustering. The algorithm can be summarized as follwos:
Step 1: Partition the items into k initial clusters (given for example by the hierarchical clustering).
Step 2: Assign an item to the cluster whose centroid (mean) is the nearest and then recalculate the centroids.
Step 1 and Step 2 are reiterated until the assignments do not change.
In the analysis of the microarray data, the k-means procedure was applied with the initial division of the set of genes from the hierarchical clustering, average linkage.

7.7 Cluster Analysis – Choosing the smooth parameter

The cosine transform data is formed by the first k coefficients from the data transforma-

tion according to cosine basis. It remains to choose the smooth parameter, k.

We would expect to choose a small k because the gene profiles are quite smooth. Instead of choosing a random k, we use decision theory to decide k. The reasoning is as follows.

Let $R(k) = \mathbf{E}\int(f(t) - \hat{f}(t))^2 dt$ denote the risk. It is hard to estimate the risk based on a small number of observations. Instead, we choose the same k for curves grouped by their gene expression level variance over time (otherwise we combine curves with very different amounts of wiggliness) and minimize the average risk within each group containing N genes: $R(k) = \frac{1}{N}\sum_{i=1}^{N} R_i(k)$.

First, $\mathbf{E}(\hat{\theta}_{ir}) = \frac{1}{m}\sum_{j=1}^{m} f_i(t_j)\psi_r(t_j) =\approx \int f(t)\psi_r(t) = \theta_{ir}$ and $\mathbf{V}(\hat{\theta}_{ir}) = \frac{1}{m^2}\sum_{j=1}^{m}\sigma_i^2\psi_r^2(t_j) \equiv \nu_r^2$.

We can estimate ν_r^2 by inserting $\hat{\sigma}_i$ for σ_i: $\hat{\nu}_r^2 = \frac{1}{m^2}\sum_{j=1}^{m}\hat{\sigma}_i^2\psi_r^2(t_j)$.

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

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

The estimate is $\hat{\mathcal{V}}(k) = \sum_{r=1}^{k}\hat{\nu}_r^2$.

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

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

Now we choose k by minimizing $\hat{R}(k)$ over $1 \le k \le (m-1)$. For each group of genes, C_g, say $g = 1, ..., 20$, the corresponding k_g is estimated. Then the mode $mode_{g=1}^{20}(k_g)$ is considered the smoothing parameter.

7.8 Cluster Analysis – Gap method

Notation. The data consist of N genes with different intensities at m time points: $(x_{ij})_{i=1,..,N,j=1,..,m}$. Denote with D_c the sum of the distances of any two genes in cluster c.

In Tibshirani et al., the within-dispersion measure is defined as $W_k = \sum_{c=1}^{K} \frac{1}{2n_c} D_c$ with K being the number of clusters, which in fact is the pooled within-cluster variance under the Euclidean distance.

Algorithm for estimating the number of clusters.
Step 1: Cluster[12] the data with the number of clusters varying from 1 to a large value K (usually, in the analysis $K = 20$) and find W_k for each number of clusters (where W_k was defined above).
Step 2: Generate B ($B \geq 1000$) reference datasets from the uniform distribution. Then cluster each of the new data and find the within-dispersion measure W_{kb}^* with $b = 1, .., B$ and $k = 1, .., K$. The estimated gap statistic will be: $Gap(k) = \frac{1}{B} \sum_{b=1}^{B} log(W_{kb}^*) - log(W_k)$.
Step 3: Compute the standard deviation s_k as: $\sqrt{(1 + \frac{1}{B}) \frac{1}{B} \sum_{b=1}^{B} (log(W_{kb}^*) - \frac{1}{B} \sum_{b=1}^{B} log(W_{kb}^*))^2}$ and estimate the number of clusters with $\hat{k} = argmin_{k=1,..,K}(Gap(k) - Gap(k+1) + s_{k+1} \geq 0)$.

The algorithm is based on estimating the number of clusters k for which the observed $log(W_k)$ is the furthest away from the estimated expectation of $log(W_k)$ found from the reference (uniform) distribution.

7.9 Cluster Analysis – Local clustering

The gap procedure is a global test for determining the number of clusters in the data. We propose a test for local clustering in addition to a global clustering. With this procedure, we aim to find local clusters which are not identified by the gap method.

Algorithm. After breaking up the data in $C_1, \ldots, C_k$ clusters (using the gap method), we test them to see whether each cluster contains local clusters. The test for deciding whether C_j should be split into other clusters is the following:
Step 1: Run k-means (or any other clustering algorithm) with $k = 2$ on the data in cluster C_j only. Denote the two new clusters by C_{j1} and C_{j2}.
Step 2: Let $\tilde{C}_j$ be the hyper-ellipsoid[13] containing the points in C_j and similarly define $\tilde{C}_{j1}$ and $\tilde{C}_{j2}$.
Step 3: We will test H_0: the data are uniform over $\tilde{C}_j$ versus H_1: the data are a sum of uniforms over $\tilde{C}_{j1}$ and $\tilde{C}_{j2}$. Let $\mu = \mathsf{volume}(\tilde{C}_j)$, $\mu_1 = \mathsf{volume}(\tilde{C}_{j1})$, $\mu_2 = \mathsf{volume}(\tilde{C}_{j2})$. Under H_0 the density is $f(x) = 1/\mu$ for $x \in \tilde{C}_j$. Under H_1 the density is

$$g(x) = \frac{1}{2\mu_1} I_1(x) + \frac{1}{2\mu_2} I_2(x)$$

where I_1 is the indicator function for $\tilde{C}_{j1}$ and I_2 is the indicator function for $\tilde{C}_{j2}$.

Step 4: The Neyman-Pearson test statistic is

$$T = \frac{\prod_i g(X_i)}{\prod_i f(X_i)} = \left(\frac{1}{2r_1}\right)^{n_1} \left(\frac{1}{2r_2}\right)^{n_2}$$

where n_1 is the number of points in C_{j1}, and $r_1 = \mu_1/(\mu_1 + \mu_2)$ and similarly for n_2, r_2.
Step 5: Now we need to get the null distribution of T. Let n be the number of points in

[12]The clustering algorithm is k-means or Partition around menoids.

[13]We take the hyper-ellipsoid instead of the smallest convex set which covers C_j because, estimating the convex hull of a set of points is $O(N^{[d/2]+1})$ or exponential in the dimensions.

C_j (over the hyper-ellipsoid which covers C_j). Simulate n observations from a uniform over C_j, compute T and repeat. Repeat B times ($B \geq 1000$) to get $T_1, \ldots, T_B$. The $p - value$ is $B^{-1} \sum_{b=1}^{B} I(T < T_b)$.
Step 6: If the p-value is less than .05, split C_j otherwise stop.

7.10 Cluster Analysis – Cluster Stability

An important limitation of clustering methods is their interpretation. The methods themselves don't quantify *cluster stability*.

We propose the following algorithm to quantify the cluster stability.

Step1. Estimate the number, K, of clusters using the gap algorithm and local clustering.

Step2. Cluster the data into k clusters $C_1, ..., C_K$ and get the covariance matrices $\hat{S_1}, ..., \hat{S_K}$ of the clusters. Let $\mu_1, ..., \mu_K$ denote the cluster centers.

Estimate the stability of clusters as follows.

Step3. For each cluster k, generate a sample of length 1 from the multivariate normal distribution with mean given by the intensities of the genes in the cluster and the covariance matrix $\hat{S_k}$. Denote it by X_k^*.

Step4. Count the number of genes in the initial cluster C_k which minimizes the distance $||X_k^*(i) - \mu_j||$ when j is in $1, .., K$. This will give us the number of genes that stay in the cluster C_k $(k = 1, ..., K)$.

Repeat 3-4 5000 times.

7.11 Lists of genes provided by multiple hypothesis testing and clustering

In the followings, the list of genes which change in expression level according to hypothesis testing (the 2-time-difference test and the runs test) are included. It is also specified the clusters they fall in according to our cluster analysis over.

Genes whose expression level changes significantly according to 2-time-difference test with FDR correction.

In cluster 1 there is 1 known gene

[1] "gamma-glutamyl transpeptidase"

In cluster 2 there are 20 known genes

[1] "tissue inhibitor of metalloproteinase 3" [2] "zinc finger protein 46" [3] "tumor necrosis factor receptor superfamily, member 7" [4] "inhibin beta E" [5] "nuclear receptor subfamily 4, group A, member 2" [6] "zinc finger protein 147" [7] "aldolase 1, A isoform" [8] "neuroblastoma ras oncogene" [9] "ribosomal protein L27a" [10] "lipoprotein lipase" [11] "Ngfi-A binding protein 2" [12] "ATPase-like vacuolar proton channel" [13] "uncoupling protein 2, mitochondrial" [14] "integral membrane protein 2 B" [15] "Unc-51 like kinase 1 (C. elegans)" [16] "glutathione transferase zeta 1 (maleylacetoacetate isomerase)" [17] "phosphoenolpyruvate carboxykinase 1, cytosolic" [18] "ADP-ribosylarginine hydrolase" [19] "transformed mouse 3T3 cell double minute 2" [20] "T-box 2"

Genes whose expression level changes significantly according to 2-time-difference test with Bonferroni correction.

[1] "tumor necrosis factor receptor superfamily, member 7" [2] "ribosomal protein, mitochondrial, S7" [3] "inter-alpha trypsin inhibitor, heavy chain 2" [4] "lipoprotein lipase" [5] "uncoupling protein 2, mitochondrial" [6] "RAB11a, member RAS oncogene family" [7] "ATPase, Na+/K+ beta 3 polypeptide"

The list of known genes among the genes whose expression levels change significantly according to Wilcoxon rank sum test (compare 3-8 hours vs. 10-24 hours arrays).

In cluster 1 there are 16 known genes

[1] "gamma-glutamyl transpeptidase" [2] "homer, neuronal immediate early gene, 3" [3] "inter-alpha trypsin inhibitor, heavy chain 2" [4] "tenascin C" [5] "homeo box B5" [6] "DNA segment, Chr 13, Wayne State University 50, expressed" [7] "neuronal d4 domain family member" [8] "dopa decarboxylase" [9] "DNA segment, Chr 2, Wayne State University 88, expressed" [10] "cysteine knot superfamily 1, BMP antagonist 1" [11] "wee 1 homolog (S. pombe)" [12] "bisphosphate 3'-nucleotidase 1" [13] "RAB11a, member RAS oncogene family" [14] "DNA segment, human DXS9928E" [15] "ATPase, Na+/K+ beta 3 polypeptide" [16] "DNA polymerase epsilon, subunit 2"

In cluster 2 there are 22 known genes

[1] "solute carrier family 20, member 1" [2] "fos-like antigen 2" [3] "high mobility group protein I" [4] "cytochrome c oxidase, subunit VI a, polypeptide 1" [5] "collagen binding protein 1" [6] "hormonally upregulated Neu-associated kinase" [7] "poly(rC) binding protein 2" [8] "selenoprotein P, plasma, 1" [9] "CASP2 and RIPK1 domain containing adaptor with death domain" [10] "zinc finger protein 46" [11] "suppressor of K+ transport defect 3" [12] "Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed (fox derived)" [13] "branched chain keto acid dehydrogenase E1, beta polypeptide" [14] "branched chain keto acid dehydrogenase kinase" [15] "ribosomal protein, mitochondrial, S7" [16] "hydroxysteroid sulfotransferase" [17] "cytochrome c oxidase, subunit VIIIa" [18] "hemoglobin beta chain complex" [19] "beta-2 microglobulin" [20] "calmodulin" [21] "ribosomal protein S6 kinase, 90kD, polypeptide 2" [22] "small inducible cytokine subfamily A, member 22"

The list of known genes among the genes whose expression levels change significantly according to runs test (based on the longest run).

In cluster 1 there are 15 known genes

[1] "gamma-glutamyl transpeptidase" [2] "homer, neuronal immediate early gene, 3" [3] "carbonyl reductase" [4] "complement component 3" [5] "inter-alpha trypsin inhibitor, heavy chain 2" [6] "homeo box B5" [7] "adipsin" [8] "DNA segment, Chr 13, Wayne State University 50, expressed" [9] "proteasome (prosome, macropain) subunit, alpha type 1" [10] "solute carrier family 27 (fatty acid transporter), member 2" [11] "neuronal d4 domain family member" [12] "bisphosphate 3'-nucleotidase 1" [13] "RAB11a, member RAS oncogene family" [14] "ATPase, Na+/K+ beta 3 polypeptide" [15] "DNA polymerase epsilon, subunit 2"

In cluster 2 there are 6 known genes

[1] "solute carrier family 20, member 1" [2] "collagen binding protein 1" [3] "selenoprotein P, plasma, 1" [4] "hemoglobin beta chain complex" [5] "ribosomal protein S6 kinase, 90kD, polypeptide 2" [6] "complement component 1, q subcomponent binding protein"

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] **Boldrick J.C., et. al.(2002)** *Stereotyped and specific gene expression programs in human innate immune responses to bacteria*, Proc Natl Acad Sci. 99(2), 972-7.

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

[4] **Brown, L. D, Low, M. (1996)**, *A constrained risk inequality with applications to nonparametric functional estimation*, The annals of Statistics, 24, 4, pp. 2524-2535.

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

[6] **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.

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

[8] **Gasser, T., Sroka, L. , Jennen-Steinmetz, C. (1986)**, *Residual variance and residual pattern in nonlinear regression*, Biometrika, 73, 3, pp. 625-633.

[9] **Gibbsons, J. D., Chakraborti, S. (1991)**, *Nonparametric Statistical Inference*, Marcel Dekker, Inc., 3rd Edition.

[11] **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.

[12] **Hastie, T., Tibshirani, R.(1990)**, *Generalized Additive Models*, Chapman &Hall/CRC, Boca Raton.

[13] **Hastie, T., Tibshirani, R.,Friedman, J.H.(2001)**, *The elements of Statistical Lerning: Data Mining and Prediction*, Springer Series in Statistics.

[14] **Kerr, K. M., Churchill, G. A.(2001)**, *Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments*, Technical report.

[15] **Knudsen, S.(2002)**, *A Biologist's Guide to analysis of c DNA Microarray Data*, John Wiley & Sons, NY.

[16] **Mosteller, F., Rourke, R. E. (1941)**, *Note on an application to quality control charts*, Annals of Mathematical Statistics, 12, 228-232.

[17] **Pollard, K. S., Van del Lann,. M. J. (2002)**, *A method to identify significant clusters in gene expression data.*, technical report.

[18] **Ramsey, J. O., and Silverman, B. W.(1997)**, *Functional Data Analysis*, Springer-Verlag, NY.

[19] **Speed, P. T., Yang, Y. H. (2002, April)**, *direct versus indirect designs for cDNA microarray experiments*, technical report.

[20] **Storey, J. D.(2001, June)**, *A Direct Approach to False Discovery Rates*, Journal of

Royal Statistical Society, B.
[21] **Storey, J. D. and Tibshirani, R.(2002)**, *Estimating FDR under Dependence with Applications to DNA microarrays*, technical report
[22] **Tibshirani, R., Walther, G., Hastie, T. (2000,Dec)**, *Estimating the number of clusters in a dataset via the Gap statistic*. Technical report, published in JRSSB,2000.
[23] **Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P.**, *Normalization for cDNA Microarray Data*, technical report.
[24] **Wichern, D. W., Johnson, R. A. (1982)**, *Applied Multivariate Statistical Analysis*, Prentice-Hall, NJ.

Web site references

[1] **Stanford Microarray Database**, *http://genome-www5.Stanford.EDU/MicroArray/SMD/*
[2] **National Center for Biotechnology Information, Entrez search and retrieval system**, *http://www.ncbi.nlm.nih.gov/Entrez/*