Statistical process control for large scale microarray experiments

Fabian Model, Thomas König, Christian Piepenbrock and Péter Adorján

Epigenomics AG, Kastanienallee 24, Berlin, D-10435, Germany

Received on January 24, 2002; revised and accepted on March 28, 2002

ABSTRACT

Motivation: Maintaining and controlling data quality is a key problem in large scale microarray studies. In particular systematic changes in experimental conditions across multiple chips can seriously affect quality and even lead to false biological conclusions. Traditionally the influence of these effects can be minimized only by expensive repeated measurements, because a detailed understanding of all process relevant parameters seems impossible.

Results: We introduce a novel method for microarray process control that estimates quality based solely on the distribution of the actual measurements without requiring repeated experiments. A robust version of principle component analysis detects single outlier microarrays and thereby enables the use of techniques from multivariate statistical process control. In particular, the $T^2$ control chart reliably tracks undesired changes in process relevant parameters. This can be used to improve the microarray process itself, limits necessary repetitions to only affected samples and therefore maintains quality in a cost effective way. We prove the power of the approach on 3 large sets of DNA methylation microarray data.

Contact: Fabian.Model@epigenomics.com

Keywords: DNA methylation; robust PCA; microarray; statistical process control; quality control.

INTRODUCTION

DNA microarrays are one of the most popular technologies in molecular biology today (Lockhart and Winzeler, 2000). They allow looking at the mRNA expression of thousands of genes at once and make applications such as marker identification, tissue classification, and discovery of new tissue subtypes possible (Golub et al., 1999). Recently it has been shown that microarrays can also be used to detect DNA methylation and that those results are comparable to mRNA expression analysis (Gitan et al., 2002; Adorján et al., 2002; Model et al., 2001).

Despite the popularity of microarray technology, there remain serious problems regarding measurement accuracy and reproducibility. Considerable effort has been put into the understanding and correction of effects such as background noise, signal noise on a slide and different dye efficiencies (Brown et al., 2001; Tseng et al., 2001; Dudoit et al., 2000). However, with the exception of overall intensity normalization (Zien et al., 2001), it has not been clear until now how to handle variations between single slides and systematic changes between slide batches. The detection of these two effects will be the focus of this paper.

Between-slide variations are very problematic because it is difficult to explicitly model the numerous different process factors which may distort the measurements. Some examples are concentration and amount of spotted probe during array fabrication, the amount of labeled target added to the slide and the general conditions during hybridization (Tseng et al., 2001). Other common, but often neglected, problems are handling errors such as accidental exchange of different probes during array fabrication (Knight, 2001). These effects can randomly affect single slides or whole slide batches. The latter is especially dangerous because it introduces a systematic error and can lead to false biological conclusions.

There are several ways to reduce between-slide variance and systematic errors. Removing obvious outlier chips based on visual inspection is an easy and effective way to increase experimental robustness. A more costly alternative is to do repeated chip experiments for every single biological sample and obtain a robust estimate for the average signal. With or without chip repetitions, randomized block design can further increase confidence in biological findings. Unfortunately, there are several problems with this approach. Outliers cannot always be detected visually and it is not feasible to make enough chip repetitions to obtain a fully randomized block design for all potentially important process parameters. However, when experiments are standardized enough, process dependent alterations are relatively rare events. Therefore, instead of reducing these effects by repetitions one should rather detect problematic slides or slide batches and repeat only those. This can only be achieved by controlling process stability.
Process stability control is well known in many areas of industrial production where multivariate statistical process control (MVSPC) is used routinely to detect significant deviations from normal working conditions. The major tool of MVSPC is the $T^2$ control chart, which is a multivariate generalization of the popular univariate Shewhart control procedure (Mason and Young, 2000).

Microarray chip production is rapidly evolving towards a high throughput industry. Therefore it seems natural to apply MVSPC for statistical quality control of microarray experiments. However, most of the relevant process parameters of a microarray experiment cannot be measured routinely in a high throughput environment. As an alternative, we propose using the measurement values of the microarrays themselves to control the stability of the production process. However, these measurements are extremely high dimensional and contain outliers, prohibiting the application of standard MVSPC methods. We show that it is nevertheless possible to apply MVSPC techniques, when using robust PCA (Hubert et al., 2002) to remove outliers and reduce data dimensionality.

Furthermore, we introduce novel methods that provide additional information about the nature of a process error (e.g., probe permutation versus change in probe concentration). We demonstrate, on three large DNA methylation microarray datasets, that this technique is a powerful tool for detecting process errors in microarray experiments.

The rest of the paper is structured as follows. In the second section we give a short introduction to the process that generated our microarray data. Then we describe the data sets used in this paper and point out typical sources of artefacts. In the third section we demonstrate how robust PCA can be used to detect abnormal hybridizations. This is an essential prerequisite for the application of statistical process control to microarray data. MVSPC is introduced in the fourth section and we develop a method to check whether all essential conditions stay constant over the course of an experimental series. Finally we conclude, in the fifth section, with a discussion of the importance of systematic quality control in large scale microarray experiments.

MICROARRAY DATA AND TYPICAL SOURCES OF ERROR

Microarrays to measure DNA methylation

Methylation is a modification of cytosine, which occurs either with or without a methyl group attached. Cytosine methylation can only appear together with guanine as CpG dinucleotide. Methylation is a particular relevant layer of genomic information because it plays an important role in expression regulation (Robertson and Wolffe, 2000). Methylation analysis has therefore the same potential applications as mRNA expression analysis or proteomics.

In order to measure the methylation state of different CpG dinucleotides by hybridization, sample DNA is bisulphite-treated to convert all unmethylated cytosines to uracil whereas methylated cytosines are conserved. For analysis, genes are then amplified by PCR using fluorescently labelled primers converting originally unmethylated CpG dinucleotides to TG and conserving originally methylated CpG sites. Pairs of PCR primers are multiplexed and designed complementary to DNA segments containing no CpG dinucleotides. This allows unbiased amplification of many alleles in one reaction. All PCR products from an individual sample are then mixed and hybridized to glass slides carrying a pair of immobilized oligonucleotides for each CpG position. Each of these detection oligonucleotides is designed to hybridize to the bisulphite converted sequence around a specific CpG site which is either originally unmethylated (TG) or methylated (CG). Hybridization conditions are selected to allow the detection of the single nucleotide differences between the TG and CG variants.

In the following, $N_D$ is the number of measured CpG positions per slide, $N_S$ is the number of biological samples in the study and $N_C$ is the number of hybridized chips in the study. For a specific CpG position $k \in \{1, \ldots, N_D\}$, the frequency of methylated alleles in sample $j \in \{1, \ldots, N_S\}$, hybridized onto chip $i \in \{1, \ldots, N_C\}$ can then be quantified as

$$m_{ik} = \log \frac{CG_{ik}}{TG_{ik}},$$

where $CG_{ik}$ and $TG_{ik}$ are the corresponding hybridization intensities (Adorján et al., 2002). This ratio is invariant to the overall intensity of the particular hybridization experiment and therefore gives a natural normalization of our data.

Here we will refer to a single hybridization experiment $i$ as experiment or chip. The resulting set of measurement values is the methylation profile $m_i = (m_{i1}, \ldots, m_{iN_D})'$.

Data sets

In our analysis we use data from three microarray studies. In each study the methylation status of about 200 different CpG dinucleotide positions from promoters, intronic and coding sequences of 64 genes was measured.

Temperature control Our first set of 207 chips comes from a control experiment where PCR amplificates of DNA from peripheral blood of 15 patients diagnosed with ALL or AML was hybridized at 4 different temperatures (38, 42, 44 and 46°C). We will use this data set to prove that our method can reliably detect shifts in experimental conditions.
Fig. 1. Typical artefacts in microarray based methylation analysis. The plots show the correlation between single or averaged methylation profiles. Every point corresponds to a single CpG position, the axis-values are log ratios. (a) A normal chip, showing good correlation to the sample average. (b) A chip classified as ‘unacceptable’ by visual inspection. Many spots showed no signal, resulting in a log ratio of 0 after thresholding the signals to $\epsilon > 0$. (c) A chip classified as ‘good’. Hybridization conditions were not stringent enough, resulting in saturation. In many cases pairs of CG and TG oligos showed nearly identical high signals, giving a log ratio around 0. (d) A chip classified as ‘acceptable’. Hybridization signals were weak compared to the background intensity, resulting in a high amount of noise. (e) Comparison of group averages over all 64 ALL/AML chips hybridized at 42°C and all 48 ALL/AML chips hybridized at 44°C. (f) Comparison of group averages over 447 regular chips from the lymphoma data set and the 200 chips with a simulated accidental probe exchange during slide production, affecting 12 CpG positions.

Lymphoma  The second data set, having 647 chips, comes from a study where the methylation status of different subtypes of non-Hodgkin lymphomas from 68 patients was analyzed. All chips underwent a visual quality control, resulting in quality classification as ‘good’ (proper spots and low background), ‘acceptable’ (no obvious defects but uneven spots, high background or weak hybridization signals) and ‘unacceptable’ (obvious defects). We will use this data set to identify different types of outliers and show how our methods detect them.

In addition we have simulated an accidental exchange of oligo probes during slide fabrication in order to demonstrate that such an effect can be detected by our method. The exchange was simulated in silico by permuting 12 randomly selected CpG positions on 200 of the chips (corresponding to an accidental rotation of a 24 well oligo supply plate during preparation for spotting).

ALL/AML  Finally we show data from a second study on ALL and AML, containing 433 chips from 74 different patients. During the course of this study 46 oligomers ran out of stock and had to be re-synthesized. As it turned out, some of them showed a significant change in hybridization behaviour, due to synthesis quality problems. We will demonstrate how our algorithm successfully detected this systematic change in experimental conditions.

Typical artefacts  In order to illustrate typical error sources we use the Lymphoma data set with its more than 9 repeated hybridization experiments $i$ for every single biological sample $j$. With this high number of replications for each biological sample the corresponding average methylation profile $\bar{m}_j$ can be estimated reliably. Here we use the $L_1$-median

$$\bar{m}_j = \text{argmin}_x \sum_{i \in R_j} \|m_i - x\|_2$$

(2)

to compute a robust estimate for the methylation profile of biological sample $j$ from its set of repetitions $R_j$. Outlier chips can then be detected relatively easily by their strong deviation from the sample methylation profile $\bar{m}_j$.

Figure 1a shows a typical chip classified as ‘good’ by visual inspection. The small random deviations from the sample median are due to the approximately normally distributed experimental noise. A typical chip classified as ‘unacceptable’ by visual inspection is shown in Figure 1b and can be easily identified by the fact that many of the
oligo pairs gave no signal which results in a log ratio of zero. The opposite case is shown in Figure 1c. This chip has very strong hybridization signals and was classified as ‘good’ by visual inspection. However, obviously the hybridization conditions have been too unspecific and most of the oligos were saturated. Figure 1d shows a chip classified as ‘acceptable’. Many of these chips give good measurements; however some of them have such weak correlation with the true methylation profile that they should be regarded as outliers.

Other potential error sources such as changing concentrations or handling errors during slide production will influence whole chip batches. Variations in hybridization buffer or salt concentration will systematically affect the melting temperature of the spotted oligos. Figure 1e shows this systematic effect by comparing hybridizations at two different temperatures. Finally, Figure 1f shows the simulation of an accidental probe exchange during slide production, affecting 12 CpG positions.

In the following we will introduce a method to detect systematic errors that does not rely on repeated hybridization experiments and makes no explicit assumptions about error sources. This will be achieved in three major steps. First outliers are removed by robust PCA. Then classical PCA is used for dimension reduction. Finally methods from MVSPC are applied to detect changes in experimental conditions.

DETECTING OUTLIER CHIPS WITH ROBUST PCA

Methods

As a first step we aim to detect single outlier chips. In contrast to statistical approaches based on image features of single slides (Brown et al., 2001) we will use the overall distribution of the whole experimental series. This is motivated by the fact that although image analysis algorithms will successfully detect bad hybridization signals, they will usually fail in cases of unspecific hybridization. The idea is to identify the region in measurement space where most of the chips \( m_i, i = 1 \ldots N_c \), are located. The region will be defined by its center and an upper limit for the distance between a single chip and the region center. Chips with deviations higher than the upper limit will be regarded as outliers.

A simple approach would be to separately define for every dimension (in our case CpG position) \( k \) the deviation of a chip \( i \) from the center \( \mu_k \) as

\[
t_k(i) = \frac{|m_{ik} - \mu_k|}{s_k},
\]

where \( \mu_k = \frac{1}{N} \sum_i m_{ik} \) is the mean and \( s_k^2 = \frac{1}{(N-1)} \sum_i (m_{ik} - \mu_k)^2 \) is the sample variance overall chips. Assuming that the \( m_{ik} \) are normally distributed, \( t_k \) multiplied by a constant follows a \( t \)-distribution with \( N-1 \) degrees of freedom. This can be used to define the upper limit of the admissible region for a given significance level \( \alpha \) (Mardia et al., 1979).

However, a separate treatment of the different dimensions is only optimal when they are statistically independent. As Figure 2 demonstrates it is important to take into account the correlation between different dimensions. It is possible that a point which is not detected as an outlier by a component-wise test is in reality an outlier (e.g., \( P_1 \) in Figure 2). On the other hand, there are points that will be erroneously detected as outliers by a component-wise test (e.g., \( P_2 \) in Figure 2). Because microarray data have usually a very high correlation, it is better to use a multivariate distance concept instead of the simple univariate \( t_k \)-distance. A natural generalization of the \( t_k \)-distance is given by Hotelling’s \( T^2 \) statistic, defined as

\[
T^2(i) = (m_i - \mu)' S^{-1} (m_i - \mu),
\]

where

\[
\mu = \frac{1}{N} \sum_i m_i, \quad S = \frac{1}{N-1} \sum_i (m_i - \mu)' (m_i - \mu)
\]
with mean \( \mu = (1/N_C) \sum_{i=1}^{N_C} m_i \) and sample covariance matrix \( S = 1/(N_C-1) \sum_{i=1}^{N_C} (m_i - \mu)(m_i - \mu)'. \) Assuming that the \( m_i \) are multivariate normally distributed, \( T^2 \) multiplied by a constant follows a \( F \)-distribution with \( N_C - N_D \) degrees of freedom and the non-centrality parameter \( N_D \). This can be used to define the upper limit of the admissible region for a given significance level \( \alpha \) (Mardia et al., 1979).

Two problems arise when we want to use the \( T^2 \)-distance for microarray data:

1. For fewer chips \( N_C \) than dimensions \( N_D \), the sample covariance matrix \( S \) is singular and not invertible.
2. The estimates for \( \mu \) and \( S \) are not robust against outliers (Lopuhaä and Rousseeuw, 1991).

The first problem can be addressed by using principle component analysis (PCA) to reduce the dimensionality of the measurement space (Mardia et al., 1979). This is done by projecting all measurement profiles \( m_i \) onto the first \( d \) eigenvectors having the highest variance. As a result we get the \( d \)-dimensional centered vectors \( \tilde{m}_i = P_{PCA}(m_i - \mu) \) in eigenvector space. After the projection, the covariance matrix \( \tilde{S} = \text{diag} (\tilde{s}_1, \ldots, \tilde{s}_d) \) of the reduced space is a diagonal matrix and the \( T^2 \)-distance of equation 4 is approximated by the \( T^2 \)-distance in the reduced space

\[
\tilde{T}^2(i) = \sum_{r=1}^{d} \frac{\tilde{m}_{ir}^2}{\tilde{s}_r}. \tag{5}
\]

Under the assumption that the true variances are equal to the observed variances \( \tilde{s}_j \), \( \tilde{T}^2 \) follows a \( \chi^2 \) distribution with \( d \) degrees of freedom. This can be used to define the upper limit of the admissible region for a given significance level \( \alpha \). However, the problem remains that the estimated eigenvectors and variances \( \tilde{s}_j \) are not robust against outliers.

We propose solving the problem of outlier sensitivity by using robust principle component analysis (rPCA) (Hubert et al., 2002). rPCA finds the first \( d \) directions with the largest scale in data space, robustly approximating the first \( d \) eigenvectors. The algorithm starts with centering the data with a robust location estimator. Here we use the \( L_1 \) median

\[
\mu_{L1} = \text{argmin}_x \sum_{i=1}^{N_C} \| m_i - x \|_2. \tag{6}
\]

In contrast to the simple component-wise median, this gives a robust estimate of the distribution center that is invariant to orthogonal linear transformations such as PCA (Lopuhaä and Rousseeuw, 1991).

Then all centered observations are projected onto a finite subset of all possible directions in measurement space. The direction with maximum robust scale is chosen as an approximation of the largest eigenvector (e.g., by using the \( Q_n \) estimator (Croux and Rousseeuw, 1992)). After projecting the data into the orthogonal subspace of the selected ‘eigenvector’ the procedure searches for an approximation of the next eigenvector. Following Hubert et al., we have chosen simply the finite set of possible directions as the set of centered observations themselves. Note that in our experience the concrete choice of robust estimators for location and scale has no crucial impact on the results.

After obtaining a robust projection of the data into a \( d \)-dimensional subspace we can compute the outlier insensitive \( \tilde{T}^2 \)-distance and its respective upper limit of the admissible region \( \tilde{T}^2_{UCL} \), also referred to as the upper control limit (UCL). For a given significance level \( \alpha \) it is computed as

\[
\tilde{T}^2_{UCL} = \chi^2_{d,1-\alpha}. \tag{7}
\]

Every observation \( m_i \) with \( \tilde{T}^2(i) > \tilde{T}^2_{UCL} \) is regarded as an outlier.

### Results

We tested the rPCA algorithm by comparing its performance to classical PCA on the Lymphoma dataset. The results are shown in Figure 3. The rPCA algorithm detected 97% of the chips with ‘unacceptable’ quality, whereas classical PCA detected only 29%. 10% of the ‘acceptable’ chips were detected as outliers by rPCA, whereas PCA detected 3%. rPCA detected 21 chips as outliers which were classified as ‘good’. These ‘good’ chips have all been confirmed to show saturated hybridization signals, not identified by visual inspection. This means that rPCA is able to detect nearly all cases of outlier chips identified by visual inspection. Additionally rPCA detects microarrays which have inconspicuous image quality but show an unusual hybridization pattern.

An obvious concern with this use of rPCA for outlier detection is that it relies on the assumption of normal distribution of the data. If the distribution of the biological data is highly multi-modal, biological subclasses may be wrongly classified as outliers. To quantify this effect we simulated a very strong cluster structure in the Lymphoma data by shifting one of the smaller subclasses by a multiple of the standard deviation. Only when the measurements of all 174 CpG of the subclass where shifted by more than 2 standard deviations a considerable part of the biological samples were wrongly classified as outliers.

This situation can only be reliably detected when there are repeated hybridization experiments for every sample.
Lymphoma dataset. The check, at any time point, data set fully describe the statistical behaviour of the process for some time under perfect working conditions. The idea is to observe the variables of a process for some time under perfect working conditions. The idea is to observe the variables of a process for some time under perfect working conditions. The idea is to observe the variables of a process for some time under perfect working conditions. Therefore we define the current state of the process to control is a microarray experiment and the only process variables we have are the log ratios of the actual hybridization intensities. A single observation is then a chip \( m_i \) and the HDS of size \( N_{HDS} \) is defined as \( \{ m_1, \ldots, m_{N_{HDS}} \} \). We have to be aware of a few important issues in this interpretation of statistical process control. Firstly, our data set has a multi-modal distribution which results from a mixture of different biological samples and classes. Therefore the assumption of normality is only a rough approximation and \( T^2 \) from equation 8 should be regarded with caution. Secondly, as we have seen in the last sections, microarray experiments produce outliers, resulting in transgression of the UCL. This means that sporadic violations of the UCL are normal and do not indicate that the process is out of control. The third issue is that we have to use the assumption that a microarray study will not change its data-generating distribution systematically over time. Therefore the experimental design has to be randomized or block randomized, otherwise a systematic change in the true biological data would be interpreted as an out of control situation (e.g., when all patients with the same disease subtype are measured in one block). Finally, one has to order the chips by their date of hybridization, which shows a very high correlation to most process parameters of interest.

In this case the fraction of outlier chips per sample can be computed. A high fraction would indicate a biological cause. We used a threshold of 50% outlier chips per sample to detect outliers resulting from biological effects. However, we never encountered such a situation in our data sets.

STATISTICAL PROCESS CONTROL

Methods

In the last section we have seen how outliers can be detected solely on the basis of the overall data distribution. Statistical process control expands this approach by introducing the concept of time. The idea is to observe the variables of a process for some time under perfect working conditions. The data collected during this period form the so-called historical data set (HDS). Under the assumption that all variables are normally distributed, the mean \( \mu_{HDS} \) and the sample covariance matrix \( S_{HDS} \) of the historical data set fully describe the statistical behaviour of the process.

Given the historical data set it becomes possible to check, at any time point, \( i \), how far the current state of the process has deviated from the perfect state by computing the \( T^2 \)-distance between the ideal process mean \( \mu_{HDS} \) and the current observation \( m_i \). This corresponds to equation 4 with the overall sample estimates \( \mu \) and \( S \) replaced by their reference counterparts \( \mu_{HDS} \) and \( S_{HDS} \). Any change in the process will cause observations with greater \( T^2 \)-distances. To decide whether an observation shows a significant deviation from the HDS we compute the upper control limit as

\[
T^2_{UCL} = \frac{p(n + 1)(n - 1)}{n(n - p)} F_{p,n-p,1-\alpha},
\]

where \( p \) is the number of observed variables, \( n \) is the number of observations in the HDS, \( \alpha \) is the significance level and \( F \) is the \( F \)-distribution with \( n - p \) degrees of freedom and the non-centrality parameter \( \nu \). Whenever \( T^2 > T^2_{UCL} \) is observed the process has to be regarded as significantly out of control (Mason and Young, 2000).

In our case the process to control is a microarray experiment and the only process variables we have are the log ratios of the actual hybridization intensities. A single observation is then a chip \( m_i \) and the HDS of size \( N_{HDS} \) is defined as \( \{ m_1, \ldots, m_{N_{HDS}} \} \). We have to be aware of a few important issues in this interpretation of statistical process control. Firstly, our data set has a multi-modal distribution which results from a mixture of different biological samples and classes. Therefore the assumption of normality is only a rough approximation and \( T^2 \) from equation 8 should be regarded with caution. Secondly, as we have seen in the last sections, microarray experiments produce outliers, resulting in transgression of the UCL. This means that sporadic violations of the UCL are normal and do not indicate that the process is out of control. The third issue is that we have to use the assumption that a microarray study will not change its data-generating distribution systematically over time. Therefore the experimental design has to be randomized or block randomized, otherwise a systematic change in the true biological data would be interpreted as an out of control situation (e.g., when all patients with the same disease subtype are measured in one block). Finally, one has to order the chips by their date of hybridization, which shows a very high correlation to most process parameters of interest.

Although it is certainly interesting to look how single hybridization experiments \( m_i \) compare to the HDS, we are more interested in how the general behaviour of the chip process changes over time. Therefore we define the current state...
data set (CDS) as \([m_{i-NCDS/2}, \ldots, m_i, \ldots, m_i+NCDS/2]\), where \(i\) is the time of interest. This allows us to look at the data distribution in a time interval of size \(NCDS\) around \(i\). In analogy to the classical setting in statistical process control we can define the \(T^2\)-distance between the HDS and the CDS as

\[
T^2_w(i) = (\mu_{HDS} - \mu_{CDS})^T \hat{S}^{-1}(\mu_{HDS} - \mu_{CDS}),
\]

where \(\hat{S}\) is calculated from the sample covariance matrices \(S_{HDS}\) and \(S_{CDS}\) as

\[
\hat{S} = \frac{(N_{HDS} - 1)S_{HDS} + (N_{CDS} - 1)S_{CDS}}{N_{HDS} + N_{CDS} - 2}.
\]

Although it is possible to use the \(T^2_w\)-distance between the historical and current data set to test for \(\mu_{HDS} = \mu_{CDS}\), this information is relatively meaningless. The hypothesis that the means of HDS and CDS are equal would almost always be rejected, due to the high power of the test. Of more interest is \(T_w\) itself, which is the amount by which the two sample means differ in relation to the standard deviation of the data.

In order to see whether an observed change of the \(T^2_w\)-distance comes from a simple translation, it is also interesting to compare the two sample covariances, \(S_{HDS}\) and \(S_{CDS}\). A translation in \(\log(CG/TG)\) space means that the hybridization intensities of HDS and CDS differ by only a constant factor (e.g. a change in probe concentration).

This situation can be detected by looking at

\[
L(i) = 2\left[\ln |\hat{S}| - \frac{N_{HDS} - 1}{N_{HDS} + N_{CDS} - 2} \ln |S_{HDS}| \right. \\
- \left. \frac{N_{CDS} - 1}{N_{HDS} + N_{CDS} - 2} \ln |S_{CDS}|\right],
\]

which is the test statistics of the likelihood ratio test for different covariance matrices (Hartung and Epelt, 1995). It gives a distance measure between the two covariance matrices (i.e., \(L = 0\) means equal covariances).

Before we can apply the methods described to a real microarray data set, we again have to solve the problem of finding a non-singular and outlier-resistant estimate of \(S_{HDS}\) and \(S_{CDS}\). In contrast to the last section, the simple approximation of \(S_{HDS}\) by its first principle components will not work here. The reason is that changes in the experimental conditions outside the HDS will not necessarily be represented in the first principle components of \(S_{HDS}\).

The solution is to first embed all the experimental data into a lower dimensional space by PCA. This works because any significant change in the experimental conditions will be captured by one of the first principle components. \(S_{HDS}\) and \(S_{CDS}\) can then be reliably computed in the lower dimensional embedding. The problem of robustness is simply solved by first using robust PCA to remove outliers before performing the actual embedding and before computing the sample covariances. A summary of our algorithm is given in Figure 4.

With the computed values for \(T^2\), \(T^2_w\) and \(L\) we can now generate a plot that visualizes the quality development of the chip process over time, a so called \(T^2\) control chart.

**Results**

The first example is shown in Figure 5, which demonstrates how our algorithm detects a change in hybridization temperature. As can be expected, the \(T^2\)-value grows with an increase in hybridization temperature. The systematic increase of the \(L\)-distance indicates that this is not only caused by a simple translation in methylation space. The process has to be regarded as clearly out of control, because almost all chips are above the UCL after the temperature change and the process center has drifted more than \(T_w = 4\) standard deviations away from its original location.

Figure 6 shows how our method detects the simulated handling error in the Lymphoma data set. The affected chips can be clearly identified by the significant increase in the \(T^2\)-distances as well as by their change in the covariance structure.

Finally, Figure 7 shows the \(T^2\) control chart for the ALL/AML study. It clearly indicates that the experimental conditions significantly changed twice over the course...
of the study. A look at the L-distance reveals that the covariance within the two detected artefact blocks is identical to the HDS. A change in covariance can be detected only when the CDS window passes the two borders. This clearly indicates that the observed effect is a simple translation of the process mean.

The major practical problem is now to identify the reasons for the changes. In this regard the most valuable information from the $T^2$ control chart is the time point of process change. It can be cross-checked with the laboratory protocol and the process parameters that have changed at the same time can be identified. In our case the two process shifts corresponded to the time of replacement of re-synthesized probe oligos for slide production, which were obviously delivered at a wrong concentration. After exclusion of the affected CpG positions from the analysis the $T^2$ chart showed normal behaviour and the overall noise level of the data set was significantly reduced.

**DISCUSSION**

Taken together, we have shown that robust principle component analysis and techniques of statistical process control can be used to detect flaws in microarray experiments. Robust PCA has proven to be able to automatically detect nearly all cases of outlier chips identified by visual inspection, as well as microarrays with unconspicuous image quality but saturated hybridization signals. The $T^2$ control chart we have introduced is a tool that facilitates the detection and assessment of even minor systematic changes in large scale microarray studies.

A major advantage of both methods is that they do not rely on explicit modeling of the microarray process as they are based solely on the distribution of the actual measurements. Having successfully applied our methods to the example of DNA methylation data, we assume that the same results can be achieved with other types of microarray platforms. The sensitivity of the methods improve with increasing study sizes, due to their multivariate nature. This makes them particularly...
suitable for medium to large scale experiments in a high throughput environment.

The retrospective analysis of a study with our methods can greatly improve results and avoid misleading biological interpretations. When the $T^2$ control chart is monitored in real time, a given quality level can be maintained in a very cost effective way. On the other hand, this allows for an immediate correction of process parameters. On the other hand, this makes it possible to specifically repeat only those slides affected by a process artefact. This guarantees high quality while minimizing the number of repetitions.

A general shortcoming of $T^2$ control charts is that they indicate that something merely went wrong, but not what was exactly the source. Therefore we have used the time point at which a significant change happened in order to identify the responsible process parameter. We have shown that changes in covariance structure provide additional information and permit discrimination between different problems such as changes in probe concentration and accidental handling errors. However, further work will be necessary to facilitate an efficient detection of error sources.

ACKNOWLEDGEMENTS

We would like to thank the members of the STAT Theory Club at TU Berlin, Jürgen Schweiger, Malte Kuss, Jörg Betzin and Ulrich Kockelkorn, for many enlightening discussions. We would also like to thank Evelyne Lipscher at Epigenomics for her many hours of dedicated artefact hunting.

REFERENCES


