Peridocity, Change Detection and Prediction in Microarrays

Chapter 1

A Method for Analysis of CGH Microarray Data

1.1 Introduction

Genomic DNA copy number at any locus in a genome represents the number of copies of DNA. Copy number changes might trigger the occurrence of many diseases; for example, genetic alternations frequently cause tumorigenesis. Thus, studying these copy number changes draws huge interest in cancer research. Deletions of copy numbers contribute to the alterations in the expression of tumor-suppressor genes, whereas amplifications contribute to the alterations in oncogenes. The changes in gene expression modify the normal growth control and survival pathways. Thus, for understanding disease phenotype and for localizing important genes, it is important to characterize the DNA copy number changes. *Comparative Genomic Hybridization* (CGH) microarray is a technique for measuring such changes (Pinkel and Albertson, 2005). As a high throughput technique, it offers many advantages over other cytogenetic techniques such as *Fluorescence In Situ Hybridization* (FISH). While early experimental techniques were only able to detect chromosomal changes at the whole chromosomal or

whole arm level, the CGH arrays using *Bacterial Artificial Chromosome* (BAC) clones have been widely used. More recently, cDNA and oligonucleotide arrays have become popular for CGH. The shorter probes on these arrays provide design flexibility and greater coverage, and the resultant high-throughput CGH data have prompted the development of various methods for data analysis. See Lai *et al.* (2005) and Willenbrock and Fridlyand (2005) for comparative reviews of the analysis methods.

In a CGH experiment, a test sample labelled red (Cy5) is hybridized to a reference normal sample labelled green (Cy3), and the resulting data consists of the ratio of the fluorescence intensities from test versus reference sample, indexed by the physical location of the clones on the genome. The arrays in CGH experiment are constructed with the assumption that the ratio of binding of test and control DNA is proportional to the ratio of the copy numbers of the corresponding DNA sequences. Alterations in DNA copy number typically occur through the gain or loss of chromosomal segments. In a homogenous cell population the actual DNA copy number profile of the genome consists of a series of plateaus of constant copy number, bounded by sharp transitions. Thus the alterations correspond to the regions of concentrated high or low log-ratios on the genome.

Various methods have already been proposed to study and solve the challenge of efficiently identifying the regions with DNA copy number alterations. For example, Pollack *et al.* (2002) applied a moving average to the ratios and used normal versus normal hybridization to compute the threshold; Hodgson *et al.* (2001) used a maximum likelihood to fix mixture models corresponding to gain, loss and normal regions; Lingjaerde *et al.* (2001) employed a simple smoothing to signs of neighbours and significance is described by comparing both the height and weight of the observed segments with their joint null distribution. Wang *et al.* (2005) proposed an algorithm *Cluster Along Chromosomes* (CLAC), which builds hierarchical clustering-style trees along each chromosome arm (or chromosome), and then selects the clusters by controlling the *False Discovery Rate* (FDR) at a certain level.

The log-ratio sequence is viewed as a time series sequence along the genome by considering the possible correlation between clones at closer physical locations on the genome. The problem of change point detection in such series is closely related to the problem of detecting discontinuities in signal processing and edge-detection in image analysis. Wavelet methods are widely used for these problems. For example, for detecting discontinuity, one method recommends using the Haar Wavelet and looking at the lowest two levels of detail (Matlab, 2007). The MatLab approach is purely exploratory. Wang (1995) proposed a method for identifying the jumps in a time series by checking if wavelet transformation of the data has large absolute values across fine scale levels.

We propose a new method for determining the change point of log-ratio. Maximum overlapping discrete wavelet transform (MODWT) is employed for this purpose. The method can automatically and efficiently detect the change points and hence the gain and loss regions along the whole genome. This method utilizes Wang's threshold value to define significant jumps from the previous region. Double application of MODWT at level one is used to confirm the presence of true abnormal regions in the sequence.

The organization of the chapter is as follows. In Section 1.2 we introduce the models and applications of wavelet methods to the CGH data. Some simulated examples are demonstrated in Section 1.3 to show the performance of the proposed method. Section 1.4 is devoted to the application of the method to real CGH data. A brief discussion is presented in the Section 1.5.

1.2 Notation and Models

Microarray based CGH provides the relative copy number of the spotted DNA sequences by monitoring the differential hybridization of two samples to the sequences on the array. Let $z_t, t = 1, 2, ..., n$ be the measure of the relative DNA copy numbers of n clones along the genome. Usually z_t is the logarithm with base 2 of the intensity ratio of test sample versus the reference sample. There are systematic variations in microarray experiments and so normalization

procedures are applied to remove those noises. We assume here that all the data are normalized. To identify or screen the genes that have DNA copy number gain or loss is equivalent to describe the genes locations on the genome where the DNA copy numbers increase or decrease. Assuming the DNA copy number follows a distribution F_0 in a region on the genome, and after the location k, the distribution is changed to F_1; so we can write,

$$\begin{aligned} z_1, z_2, \ldots, z_k &\sim F_0 \\ z_{k+1}, z_{k+2}, \ldots, z_n &\sim F_1 \end{aligned}$$

That is equivalent to finding the change point k, where the distribution of the relative copy numbers are different on both sides of k. Note that for the CGH data, there may be many change points along the genome, which define the regions of gains or losses of the copy numbers. If the clones on the genome are close enough, they might affect each other on copy numbers. Thus we can assume that the copy number of a clone on the genome is associated with that of the previous clone. The copy numbers sequence along the genome can therefore be envisaged as a time series. Determination of change points is equivalent to the determination of abrupt change along the sequence. Wavelets are ideally suited for this purpose.

1.2.1 Wavelet Methods

Wavelets are well established in the mathematical sciences (Daubechies, 1992) and have been successfully applied in fields such as signal and image processing, numerical analysis and statistics. Wavelets literally means small waves. A function $\psi(.)$ defined over the entire real axis is called a wavelet if $\psi(.) \to 0$ as $t \to \pm\infty$ and satisfying the following conditions:

$$\int_{-\infty}^{\infty} \psi(u)\mathrm{du} = 0 \tag{1.1}$$

$$\int_{-\infty}^{\infty} \psi^2(u)\mathrm{du} = 1 \tag{1.2}$$

Wavelets are functions that can be used to describe a signal efficiently by

breaking it down into its components at different scales and following their evolution in the time domain. Wavelets tells us the changes in averages in a time series. These changes in averages are computed in terms of weighted average differences of the series over different time scales, denoted by λ. The variation of λ can provide information about how averages of $x(.)$ over many different scales can change from one period of length λ to the next. The collection of variables $\{W(\lambda, t) : \lambda > 0, -\infty < t < \infty\}$, defined in Equation 1.3, is called continuous wavelet transform (CWT).

$$W(\lambda, t) = \int_{-\infty}^{\infty} \psi_{\lambda,t}(u)x(u)\mathrm{du} \tag{1.3}$$

In Equation 1.3, $W(\lambda, t)$ is proportional to the difference between two adjacent averages of scale λ. Here the transformed series $x(.)$ ia a function of translation parameter t and scale parameter λ. The transforming function $\psi_{\lambda,t}(u)$ is called the mother wavelet.

Discrete wavelet transformations map data from the time domain to the wavelet domain (Percival and Walden, 2000); however, the difference from CWT is that the scale λ and translation parameter t are no longer continuous. These transformations result in a vector of the same size. If we have a series of size N, wavelet transformations can be defined by the matrices of dimension $N \times N$.

The partial DWT is a special orthonormal transformation:

$$(z_0, \ldots, z_{N-1}) \longleftrightarrow (W_1, ..., W_J, V_j),$$

where W_j is a vector of length $N_j = N/2^j$; V_J has the same length as that of W_J and N_J. For simplicity we have assumed that N is a multiple of 2^J. The vector W_j is called the vector of wavelet coefficients at level j and is associated with changes or differences on scale 2^{j-1}. Vector V_J, which is the scaling coefficients at level J, is associated with averages on scale 2^{J-1}.

We can write,

$$W = \Gamma X,\ \Gamma = \begin{pmatrix} \Gamma_1 \\ \Gamma_2 \\ \vdots \\ V_J \end{pmatrix}$$

In practice the DWT is computed using the pyramid algorithm which requires only $O(N)$ flops. There are two practical limitations of DWT; these are:

- Series should be of dyadic length and
- Selecting different starting point for the series changes the result of the analysis.

First problem can be dealt through polynomial extensions of the scaling coefficients, then the DWT can be practically implemented for any size of the series. However, it is not a trivial task to select an appropriate number of end points to fit or the order of fit (Constantine and Percival, 2003). The second problem refers that the DWT is not a shift-invariant transform and so shifting the time series circularly can totally change the DWT. Maximum overlap discrete wavelet transformation (MODWT) is used to overcome such limitations. Thus MODWT provides us the advantage of making the series length and shift-invariant.

Our goal here is to identify the change point in DNA sequence. We focus on change-point approaches to data dependent thresholding (Ogden and Parzen, 1976). The primary idea is to divide the wavelet coefficients into groups of small coefficients containing primary noise and one of large coefficients containing significant signal. Hypothesis testing techniques are employed to obtain an appropriate threshold and a test to determine if the set of coefficients at that scale contains significant signal if coefficients exceed the threshold.

1.2.2 Wang's Threshold

The underlying model or null hypothesis may be written as $z_t = f(t) + a_t, t = 1, \ldots, n$ where $a_t \sim NID(0, \sigma^2)$ and $f(t)$ is a smooth function, ie. continuous and differentiable. This is a classic model in time series. Examples include the

polynomial trend analysis (Fisher, 1921) and the lowess polynomial seasonal adjustment of Clevelend *et al.* (1990). For forecasting purposes, the ARMA family and its extensions are more useful models. Wang (1995) and Ogden and Parzen (1976) consider the problem of testing an abrupt change in the function $f(t)$. Ogden and Parzen (1976) approach the problem as one of estimating $f(t)$ when step functions are present and $f(t)$ is otherwise constant and their approach is again, like in MatLab, exploratory. Wang (1995) model is more general and focuses on detecting the change points at which a jump or sharp cusp occurs. A sharp cusp occurs at point t_0 if there exists a constant $K > 0$ such that

$$|f(t_0 + h) - f(t_0)| \geq K|h|^{\alpha} \tag{1.4}$$

for all h as $h \to 0$ and $0 \leq \alpha < 1$. When $\alpha = 0$, the function has a jump. Wang (1995) shows that, asymptotically with probability 1, all wavelet coefficients will have absolute value less than the universal threshold value $\sigma\sqrt{2\log(n)}$ provided there are no jumps or cusps. The unknown value of σ may be estimated robustly by the median absolute value of the wavelet coefficients at level 1 divided by 0.6745.

1.2.3 Practical Implementation Details

We follow the notation in the book by Percival and Walden (2000) which is also used in the S-Plus, R and MatLab software. The principal differences are that level 1 refers highest time domain resolution and filter width rather than half-width is used in the naming of the Daubechies wavelets. So for example, Wang's $D(1)$ corresponds to $D(2)$ which is also equivalent to the Haar wavelet. Another difference is that Wang pads the real data so that n is a power of 2 because he uses Mallat's algorithm. For detection of changepoints, Wang (1995) recommends examining plots of the absolute empirical wavelets at various levels j and finding those values which exceed the threshold line and are larger than others. Daubechies wavelets are denoted as: $D(k)$, $k = 2, 4, \ldots, 20$. The level j should be chosen as small as possible in order to obtain the highest time domain

resolution.

In practice, it is not possible to examine the wavelet coefficient plots at different levels and then select the jump points. Therefore, we need to use some automated procedure for selecting appropriate levels for different series. MODWT at level one serves as a good strategy for this purpose. By intuition, we can think that the gain or loss region cannot contain a single observation. We apply MODWT at level one and record the observation numbers where the wavelet coefficients are greater than the Wang's threshold value. In order to verify that the wavelet coefficients are directing to right jump detection, we delete the observations where the jumps were detected and rerun the procedure. If the new wavelet coefficient adjacent to the deleted observation is again greater than the Wang's threshold and has the same sign as that of the previous coefficients, then the deleted observation in previous step is considered as true signal of jump detection.

To reach a conclusion of the analysis, we have to define the loss or gain region. We can define a threshold beyond which a region is called to be loss or gain region according to the sign of the wavelet coefficients. The selection of the threshold using some multiple test procedure is discussed in the following subsection. The region with multiple testing value, say q-value, greater than the threshold is colored as red and the region having multiple testing value less than the threshold is colored as green. Thus red corresponds to the gain region and green corresponds to the loss region. We put a line in each detected region to represent the mean.

1.2.4 Testing Region Means Using Bootstrap

We have $z_t, t = 1, 2, \ldots, n$ as the observations along a specific chromosome arm. The observations in ith region and tth position can be expressed as

$$z_{ti} = \mu_i + e_t,\ i = 1, 2, \ldots k \text{ and } t = 1, 2, \ldots, n$$

The error term e_t follows AR(p) process, the order of which can be estimated. That is,

$$e_t = \phi_1 e_{t-1} + \phi_2 e_{t-2} + \ldots + \phi_p e_{t-p} + a_t \tag{1.5}$$

where ϕ_1, ϕ_2, ..., ϕ_p are autoregressive parameters and $e_t \sim N(0, \sigma_a^2)$.

Suppose we have only one region and we would like to test whether the region mean is significantly different from zero. A t-test procedure that considers corrected variance of $\bar{z}$ in an AR(p) error process would seem to work for such case. A short simulation study with an AR(1) process was done to see the power of this test procedure. Table 1.1-1.2 reveal that the method does not perform very well even for small ϕ values. Hence with the increase of magnitude of ϕ, the method becomes incapable of handling such situation regardless of the series length. Moreover, series length refers to the length in a particular gain/loss region, which in real CGH data will not be very large.

Power comparison; $n = 50$, $\sigma_a = 0.2$

ϕ	$\mu = 0$	$\mu = 0.5$
0.0	0.056	0.942
0.1	0.084	0.9074
0.3	0.118	0.7472
0.5	0.108	0.5186
0.7	0.137	0.3100
0.9	0.236	0.2862

Table 1.1: Power of the test $\mu = 0$ in an AR(p) setting with series length 50. Here we consider standard deviation for error term to be 0.2. The test is done at 0.05 level of significance. The column of $\mu = 0$ indicates that the type I error increases with the increase of ϕ.

Power comparison; $n = 100$, $\sigma_a = 0.2$

ϕ	$\mu = 0$	$\mu = 0.5$
0.0	0.0548	0.9988
0.1	0.0718	0.9950
0.3	0.0780	0.9406
0.5	0.0788	0.7216
0.7	0.0970	0.3962
0.9	0.1656	0.2010

Table 1.2: Power of the test $\mu = 0$ in an AR(p) setting. Here we consider sample size to be 100 and the standard deviation for error term to be 0.2. The test is done at 0.05 level of significance. The column of $\mu = 0$ indicates that the type I error increases with the increase of ϕ.

To overcome lack of power of the test in such phenomenon, we can resort to parametric bootstrapping procedure. This simple method can be outlined in the following few steps:

Step 1 Find the region means using MODWT procedure and then find $e_{ti} = y_{ti} - \hat{y}_i$.

Step 2 Consider the autoregressive order of the process to be 1 and so this is our selected model.

Step 3 Estimate the parameters and innovation variance from the model selected in step 2.

Step 4 Simulate a mean-zero stationary Gaussian AR(p) time series, say e^*, with parameters $\hat{\phi}$ and innovation variance $\hat{\sigma}$ found in step 3. For null model $\mu = 0$, and so $y = e$. Do the simulation procedure large number of times, say $B = 10^4$ times.

Step 5 Find the means for each simulated series in all regions, $\bar{y}^*_{\gamma_1}, \bar{y}^*_{\gamma_2}, \ldots, \bar{y}^*_{\gamma_k}$, where the superscript *denotes the bootstrap sample. The p-value for region i is defined as, $p_i = \#\{\bar{y}^*_{\gamma_i} \geq \bar{y}_{\gamma_1}\}/B$

In step 2, we used AR(1) process instead of general AR(P). As our main goal is not the model selection but testing the region means, considering such restricted model will not affect the final result. However, in the presence of large series, we can find the order of the AR(p) process from the series using BIC criterion.

Bootstrapping power comparison; $n = 50$, $\sigma_a = 0.2$

ϕ	$\mu = 0$	$\mu = 0.5$
0.0	0.064	1.00
0.1	0.064	1.00
0.3	0.066	1.00
0.5	0.08	1.00
0.7	0.118	0.998
0.9	0.244	0.752

Table 1.3: The table shows the power of the bootstrapping method for testing $\mu = 0$ in AR(1) process when $n = 50$ and $\sigma_a = 0.2$. For any value of σ, FPR is very high in this case.

Bootstrapping power comparison; $n = 100$, $\sigma_a = 0.2$

ϕ	$\mu = 0$	$\mu = 0.5$
0.0	0.054	1.00
0.1	0.056	1.00
0.3	0.064	1.00
0.5	0.062	1.00
0.7	0.072	1.00
0.9	0.134	0.83

Table 1.4: The table shows that power of the test $\mu = 0$ using bootstrap method. The AR(1) series has length 100 and $\sigma_a = 0.2$. There is a little improvement of FPR than that for $n = 50$, but still this is higher than 0.05 for all value of σ.

The simulation study, presented in Tables 1.3-1.5, suggests that the bootstrapping method works well for testing mean in large series. The *False Positive*

Bootstrapping power comparison; $n = 200$, $\sigma_a = 0.2$

ϕ	$\mu = 0$	$\mu = 0.5$
0.0	0.048	1.00
0.1	0.048	1.00
0.3	0.052	1.00
0.5	0.054	1.00
0.7	0.060	1.00
0.9	0.124	0.996

Table 1.5: The table shows that power of the bootstrapping method for testing $\mu = 0$ when $n = 200$ and $\sigma_a = 0.2$. There is substantial improvement in power as well as in FPR.

Rate (FPR) of the test is still high for large ϕ and short series. Nonetheless, this test procedure works better than the previously mentioned one.

If there is only one region present in the study, the decision about the test can be done using this obtained p-value. However, in a GCH data analysis there will be several gain and loss regions and so the overall decision depends on multiple test method. Having obtained the p-values for all regions using the aforementioned bootstrap procedure, we need to calculate the multiple test values using some standard method. Benjamini and Hochberg (1995) proposed a method for multiple testing using *False Discovery Rate* (FDR). Another more recent approach, called q-value, was proposed by Storey (2002). To deal with multiple testing, Pounds *et al.* (2004) introduced spacings LOESS histogram, or SPLOSH. This aims at estimating conditional FDR which is the expected proportion of false positives given we have r significant features. In the genome wide study of testing periodicity, SPLOSH revealed to be most conservative while q-value approach seems to be liberal in detecting the correct number of periodic genes. However, unlike the number of genes, the number of jump points or the number of regions will not be even hundreds. So it would be expected that all these methods would produce similar results in this simulation.

1.2.5 Determination of Gains and Losses

Assume that the relative copy number is a smooth function $f(k)$, where k denotes the position of the clone on the gene. To find the change points of $f(k)$, we can determine abrupt change of the function $f(k)$ through wavelet coefficients. The test threshold is calculated using the universal threshold $\sigma\sqrt{2\log(n)}$. Any wavelet coefficients that exceed the point are specified as the position of abnormal change in DNA copy numbers. Once we specify distinct regions using the threshold, we need to define them as loss, gain or normal region through another preselected threshold T_2.

$$\mathcal{R}_i = \begin{cases} \text{Call gain,} & \text{if } \mathcal{M}_i > T_2 \\ \text{Call loss,} & \text{if } \mathcal{M}_i < -T_2 \\ \text{Call normal,} & \text{if } -T_2 \leq \mathcal{M}_i \leq T_2 \end{cases}$$

where $\mathcal{M}_i$ is the multiple test value of the i-th region. If we would like to call a region to be gain or loss region at a q-value of 0.05, then this is our selected T_2.

1.3 Simulated Examples

Let z_t, $t = \{1, 2, \ldots, n\}$ be the observations along a specific chromosome arm. In this section we present few simulated examples to demonstrate the performance of the proposed method. A comparison of the method with CLAC is provided. A preselected threshold of $q = 0.05$ is used to call a gain or loss region in all the simulated examples and real data.

1.3.1 Example-1: White noise series

Data of length 1040 are generated such that $z_t \sim N(0, 0.15^2)$. This means that no loss or gain region is present in the data shown in Figure 1.1. The proposed method, applied to raw data, worked well in providing the true feature of the series.

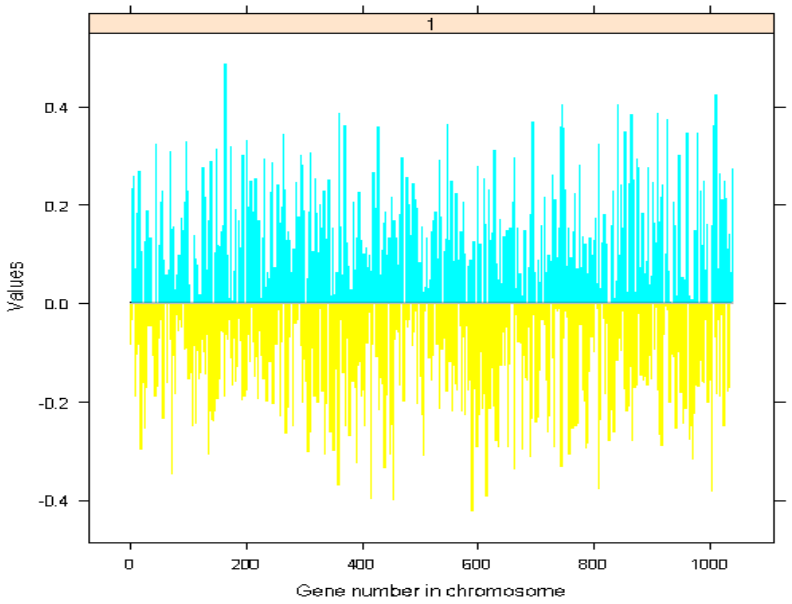

Figure 1.1: White noise series, where there is no jump. The data are simulated from $N(0, 0.15^2)$. The mean value of the region is almost in the zero line.

1.3.2 Example-2: Null Case: Smooth signal plus white noise

Some data, $n = 200$, was generated by adding random noise to a smooth curve presented in Figure 1.2. That is, the observations follow the relationship $z_t = g(x_t) + \epsilon_t$, where $g(x_t)$ is the smooth part and $\epsilon_t \sim (N(0, \sigma^2)$.

Wavelet method is applied to this data for plausible jump detection. We see from Figure 1.3 that the method is able to correctly detect the absence of any break points.

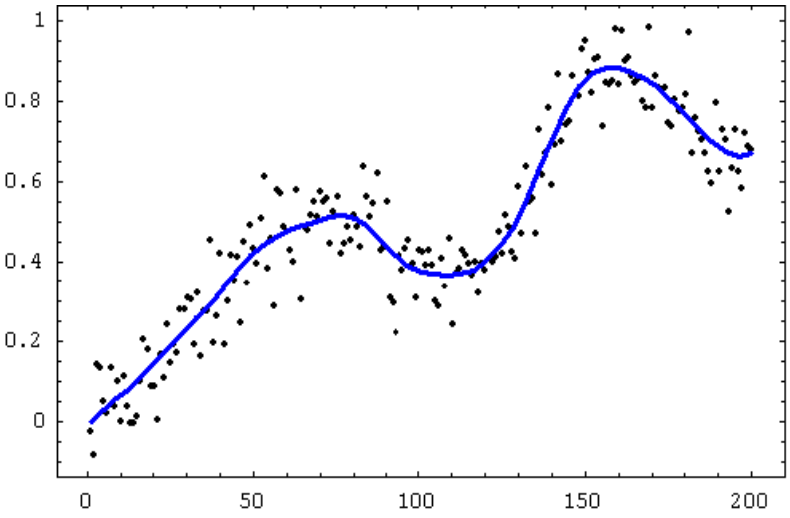

Figure 1.2: Scatter plot of simulated observations obtained by adding random noise to a smooth curve, which is also shown. Apparently there is no sharp jump point in the series.

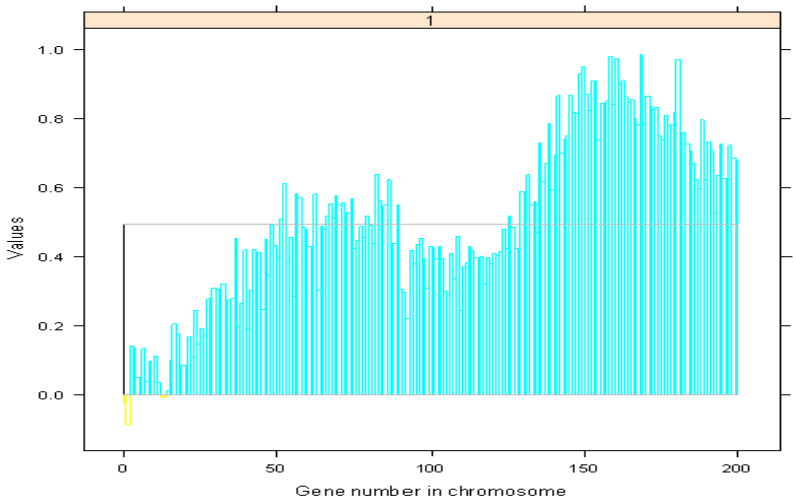

Figure 1.3: Application of wavelet method to the series shown in Figure 1.2. There is only one region, that is no jump was detected in this series. The mean value is given by a line.

1.3.3 Example-3: Two loss/gain regions

Data set with $n = 270$ observations are simulated in two blocks representing two chromosomes. The model in both chromosomes is $z_t = \mu_t + e_t$, where μ_t takes on values 0, 0.7, and -0.7. That is,

$$\mu_{t1} = \begin{cases} 0, & 1 \leq t \leq 80 \\ -0.7, & 81 \leq t \leq 110 \\ 0, & 111 \leq t \leq 150 \end{cases} \quad \text{for chromosome 1}$$

$$\mu_{t2} = \begin{cases} 0, & 1 \leq t \leq 40 \\ -0.7, & 41 \leq t \leq 70 \\ 0, & 71 \leq t \leq 120 \end{cases} \quad \text{for chromosome 2}$$

For each chromosome,

$$e_t = \phi e_{t-1} + a_t, \quad a_t \sim \text{NID}\,(0, \sigma_a^2) \tag{1.6}$$

Since $\text{Var}\,(e_t) = \sigma_a^2/(1-\phi^2)$, we can write the innovation variance, $\sigma_a^2 = (1-\phi^2)\text{Var}\,(e_t)$. We consider three cases with ϕ values 0.4, 0.6 and 0.8. Here we do not provide the graphs for case $\phi = 0.6$ as it gives the similar result as that for $\phi = 0.4$. Figures 1.4 and 1.6 show that the proposed method detects the jump points at right places in all three cases. CLAC method is applied in all data sets. For the implementation of CLAC method, normal array is generated from AR(1) process with corresponding value of ϕ used in the original data. This method seems to work well with low values of ϕ, as can be seen in Figure 1.5. However, Figure 1.7 indicates that the detection of loss and gain region is not perfect in the presence of high autocorrelation. It should be noted that the performance of the method relies on the selection of normal array.

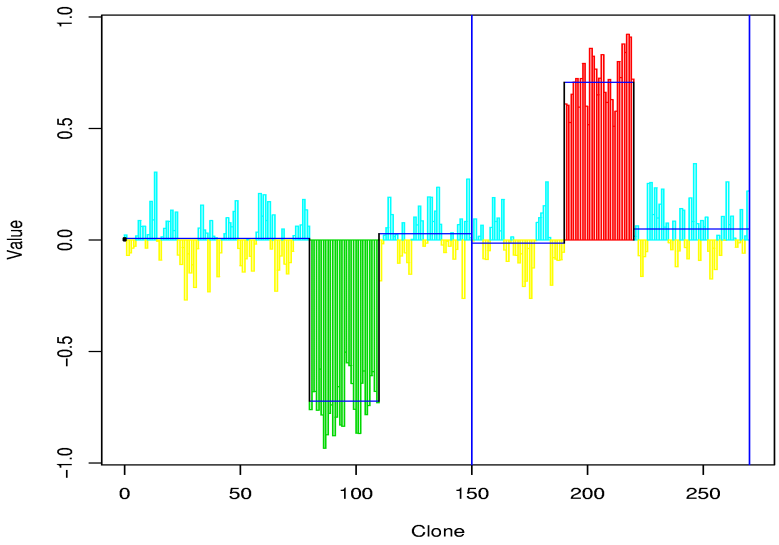

Figure 1.4: Application of wavelet method to the series with error term following AR(1) with $\phi = 0.4$. The method can detect the gain and loss region.

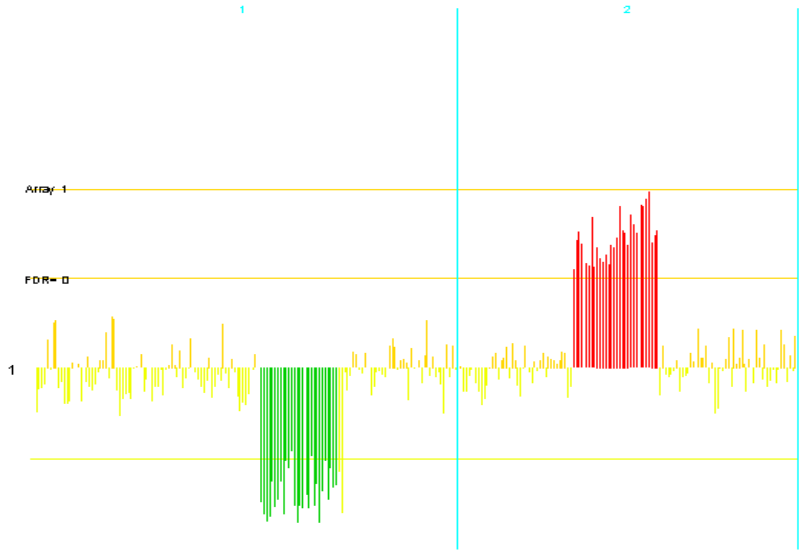

Figure 1.5: Application of CLAC method to the series with error term following AR(1) with $\phi = 0.4$. Gain and loss region is detected at the right places for this value of ϕ.

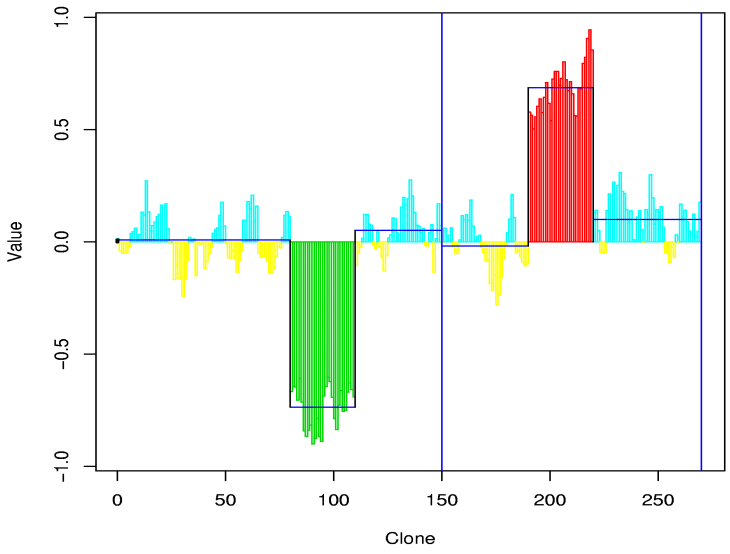

Figure 1.6: Application of wavelet method to the series with error term following AR(1) with $\phi = 0.8$. The method can detect correct gain and loss region.

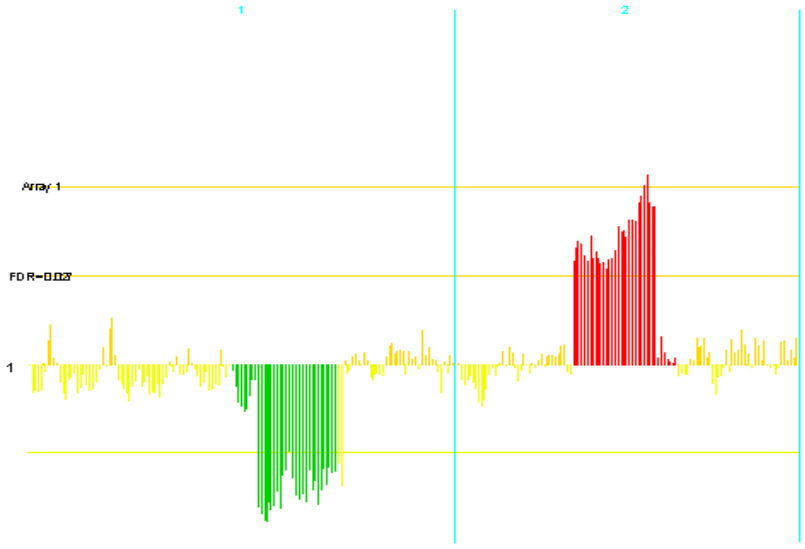

Figure 1.7: Application of CLAC method to the series with error term following $AR(1)$ with $\phi = 0.8$. We do not get exact detection of gain and loss region.

1.3.4 Example-4: Seven jump points

The data set consists of 200 observations having 7 jump points at 50, 60, 92, 106, 144, 169 and 181. Error terms are i.i.d normal with mean 0 and standard deviation 0.5. We split the series into two chromosomes where 141 genes are assigned to first one and 59 genes assigned to second one. This is a typical example where there are two successive gain regions within second chromosome. We see from Figure 1.8 that the proposed method can detect the break points exactly and define the loss and gain regions according to the preselected threshold value.

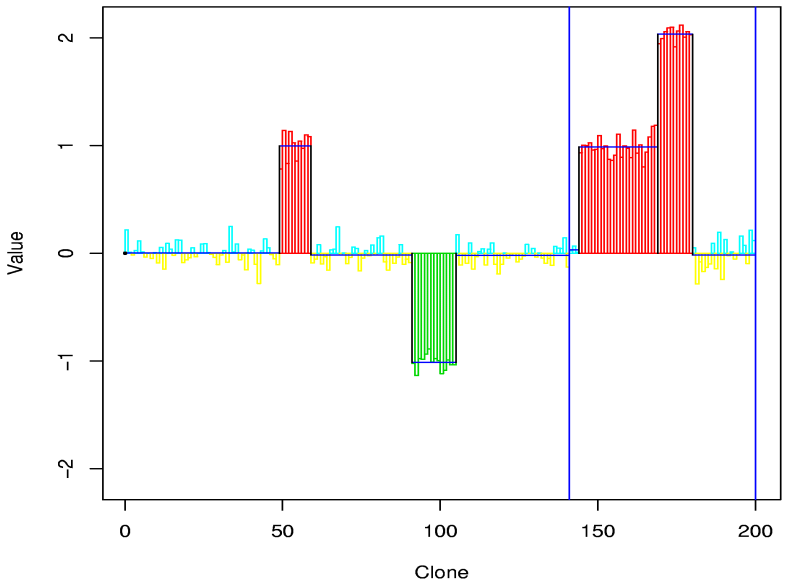

Figure 1.8: The series has seven jump points. Observations are split between two chromosomes such that first 141 observations are in chromosome 1 and rest 59 observations are assigned to chromosome 2. The proposed method correctly detects the jump points.

1.3.5 Smoothing the data

Wang (1995) suggested using simple moving average smoothing (MAS) with specific window size before applying the approach. If $\hat{z}$ be the running mean

with neighbourhood size k, then the smoothed series would be:

$$\hat{z}_i = \frac{1}{2k+1}(z_{i-k} + z_{i-k+1} + \ldots + z_{i+k}) \tag{1.7}$$

for $i = k+1, k+2, \ldots, n-k$. For the other observations, say for $i = 1, 2, \ldots k$ and $i = n-k+1, n-k+2, \ldots, n$, define $u = \max(1, i-k)$ and $v = \min(n, i+k)$

$$\hat{z}_i = \frac{1}{\nu - u + 1}, (z_u + z_{u+1} + \ldots + z_{nu}) \tag{1.8}$$

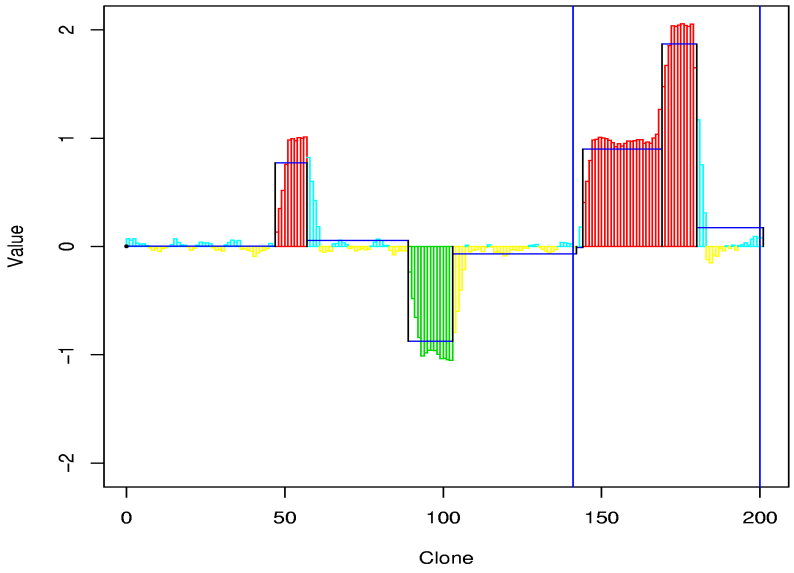

Figure 1.9: Wavelet method is applied to the same data that was demonstrated in Figure 1.8, but smoothing is done before the analysis. There are shifts in the jump point detection.

Investigations revealed that this smoothing result some shift in the break point detection when wavelet method is applied. For example, the series in Example 4 was smoothed and the wavelet method was applied thereafter. Figure 1.9 shows that the number of regions detected is correct; nevertheless, the detection points are not at the appropriate places.

1.4 Application to real data

We apply the proposed method in two real CGH data sets. The method detects several loss and gain regions. A comparison of the method with CLAC is illustrated through the second example.

1.4.1 Application-1

In CGH array, 2400 BAC clones were measured each with three replicates (Snijders *et al.*, 2001). Measurements for log base 2 intensity ratio are provided. Average relative DNA copy number sequences of the three replicates along the genome is shown in Figure 1.10. The figure also demonstrates the gain or loss regions that are detected using this method. As we can see, the measures are mostly along the zero line, which indicates that the test sample has the same DNA copy numbers as that of reference sample.

The log ratios along the genome are considered as a time series sequence. The proposed method is then applied to calculate the wavelet coefficients and to determine the abnormal positions. There are number of loss and gain regions detected by this method. To have better view of the loss and gain regions we use `trellis` plot in Figure 1.11 and 1.12 which provide plotting of the log-ratio intensities for individual chromosomes.

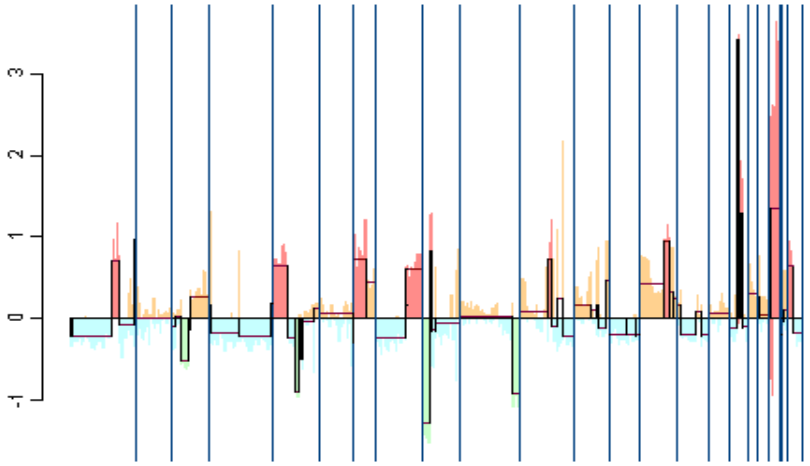

Figure 1.10: Application of wavelet method to CGH data set from Snijders *et al.* (2001). There are many gain/losss regions in the whole genome.

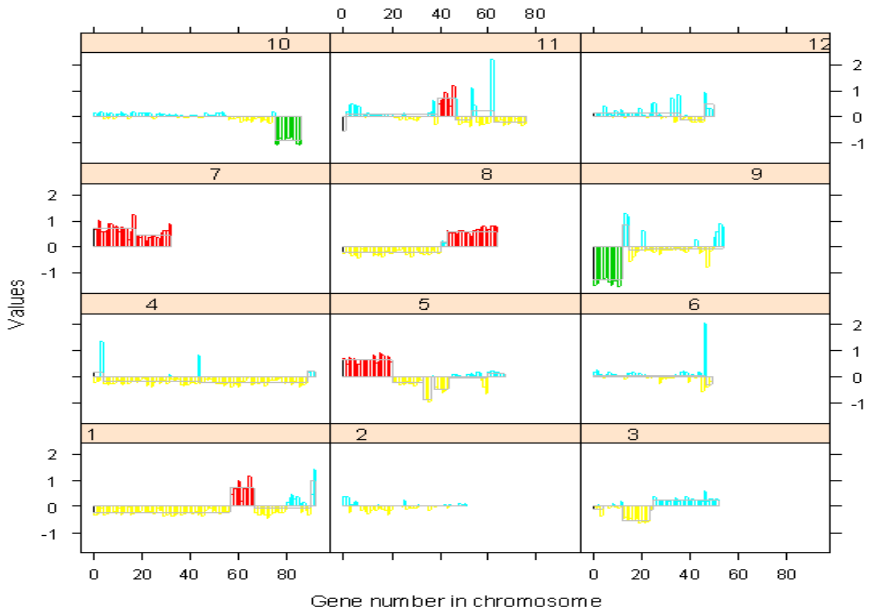

Figure 1.11: For better view of the loss/gain regions, here we plot the first 12 chromosomes in trellis plot. There are presence of abnormal regions in chromosome number 1, 5, 7, 8, 9 and 11.

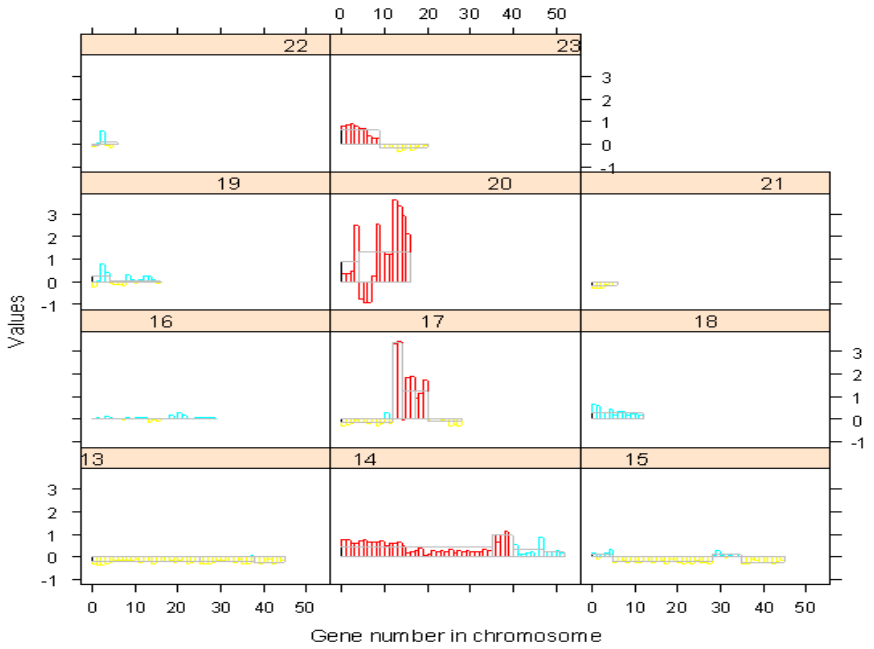

Figure 1.12: For better view of the loss/gain regions, here we plot the last 11 chromosomes in trellis plot. There are presence of abnormal regions in chromosome number 14, 17, 20 and 23.

1.4.2 Application-2

We apply the proposed method to one of the examples found in `R` library `clac`. The package has data set `BACArray` and the column `DiseaseArray` has 9980 observations containing 4 arrays, one of which is analyzed for comparison. Wavelet method detected two gain regions colored as red in Figure 1.13. Figures 1.14 and 1.14 of individual chromosome explicitly show that the chromosome 18 and 23 are the regions with copy number amplification. One normal array from the `clac` package is picked and then CLAC method is applied to the array. The outcome, presented in Figure 1.16, also indicates that chromosome 18 and 23 refer to the amplified regions for DNA copy number.

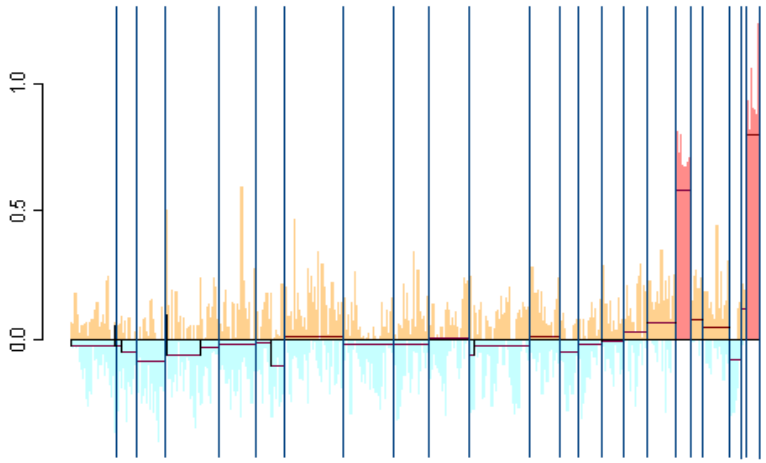

Figure 1.13: This CGH Array is obtained from R package `clac`. The wavelet method detects only two gain regions in this data set.

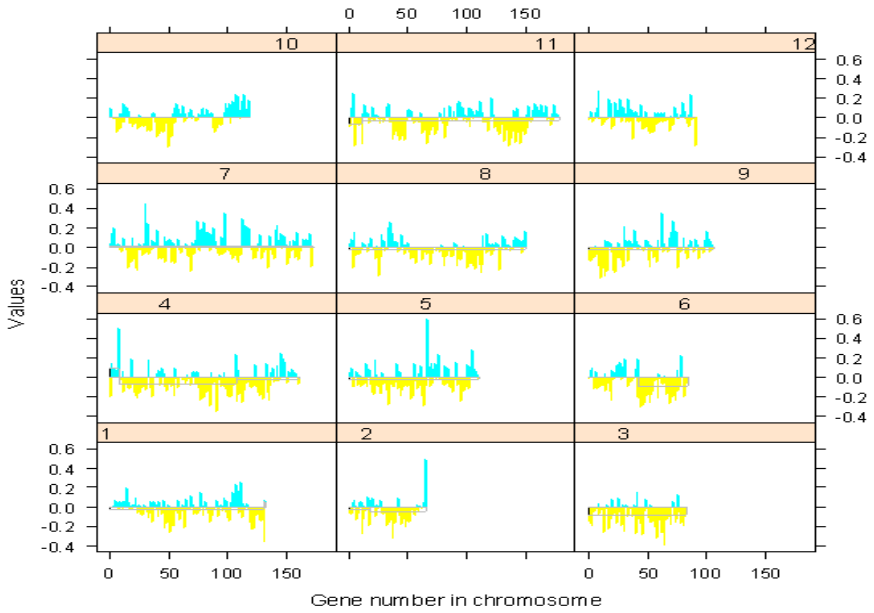

Figure 1.14: For better view of the loss/gain regions, here we plot the first 12 chromosomes. No abnormal regions were detected in these chromosomes.

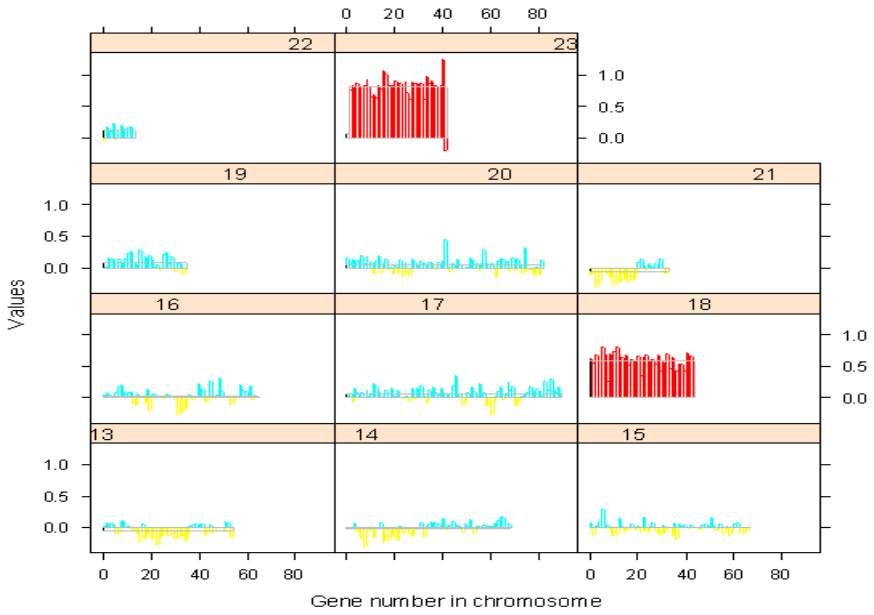

Figure 1.15: Results of individual chromosomes 13 to 23 are presented. Chromosomes 18 and 23 refer to regions of abnormal gain in DNA copy numbers.

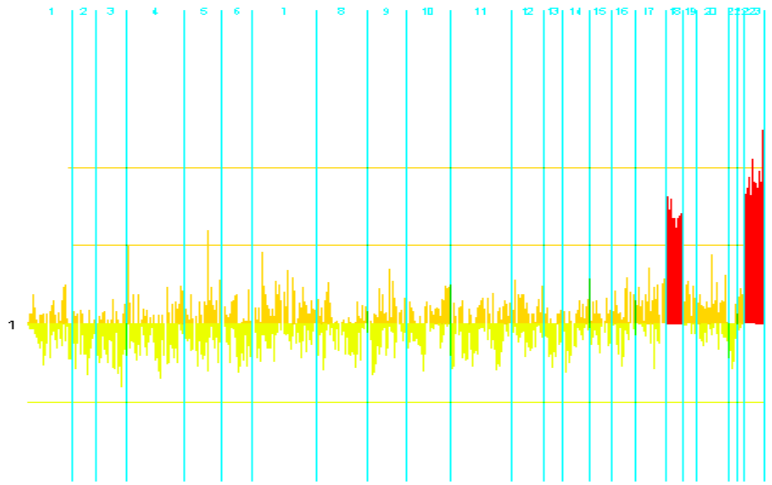

Figure 1.16: The CGH Array described in Application-2 was analyzed using CLAC method. It shows there are two gain regions in chromosome number 18 and 23.

1.5 Discussion

In this chapter we have proposed wavelet method to identify the abnormal DNA copy number positions on genome. Discrete wavelet transform has two limitations; namely dyadic length requirement and sensitivity of the starting of the time series. To overcome such limitations, we use maximum overlap discrete wavelet transformation (MODWT) in this analysis. The positions of the break points were detected using Wang's threshold. Calling a region to be gain, loss or normal depends on the selection of another threshold T_2. Through the simulated examples we demonstrate that the method performs quite well in selecting the break points and hence the abnormal regions in a time series sequence. Moreover, the procedure reports several abnormal regions in two real CGH arrays.

CLAC algorithm, proposed by Wang *et al.* (2005), uses some normal array for detecting deletion and amplification regions. Independence and normality of the clones are two strong assumptions; but the procedure of Jong *et al.* (2003) depends on these assumptions. ACF plots of the estimated errors from the fitted model are presented in Appendix. It is evident from the plots that consideration of i.i.d. observations in the sequence would not be realistic. Our propose method does not assume that the observations be i.i.d. In a short simulation example we show how the detection of the change points shifts when a moving average smoothing is used before applying the wavelet method.

Appendix

Autocorrelation function (ACF) is useful in detecting the presence of correlation among the successive observations. In this study, we observe the residuals by subtracting the mean of any selected region from the observations in that region. That is, $e_t = z_t - \mu_t$ is the residual for t-th clone. ACF plots are presented for the residuals obtained from the application in the real data described in Section 1.4.

ACF plot for individual chromosomes for Application-1

Figure 1.17 is constructed to show the autocorrelation behavior of the error process for each chromosome. It seems that the residuals are not quite i.i.d. within each of the chromosome. The residuals in chromosome numbers 1, 8, 10, 14 and 23 demonstrate the presence of strong autocorrelation. This can be a justification to use a simulation study in Example-3 of Section 3.

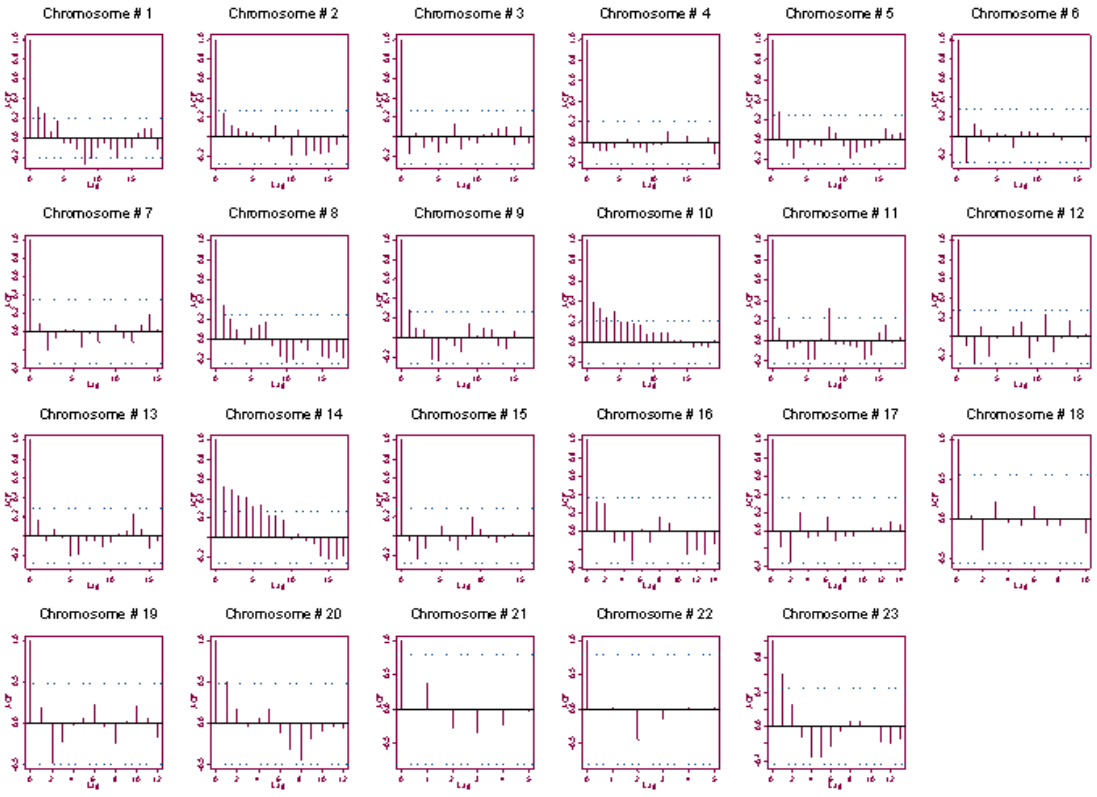

Figure 1.17: The figures show the ACF plot for the residuals obtained for chromosome 1 to 23 using the data set in subsection 1.4.1. The residuals in few of the chromosomes indicate the presence of high autocorrelation.

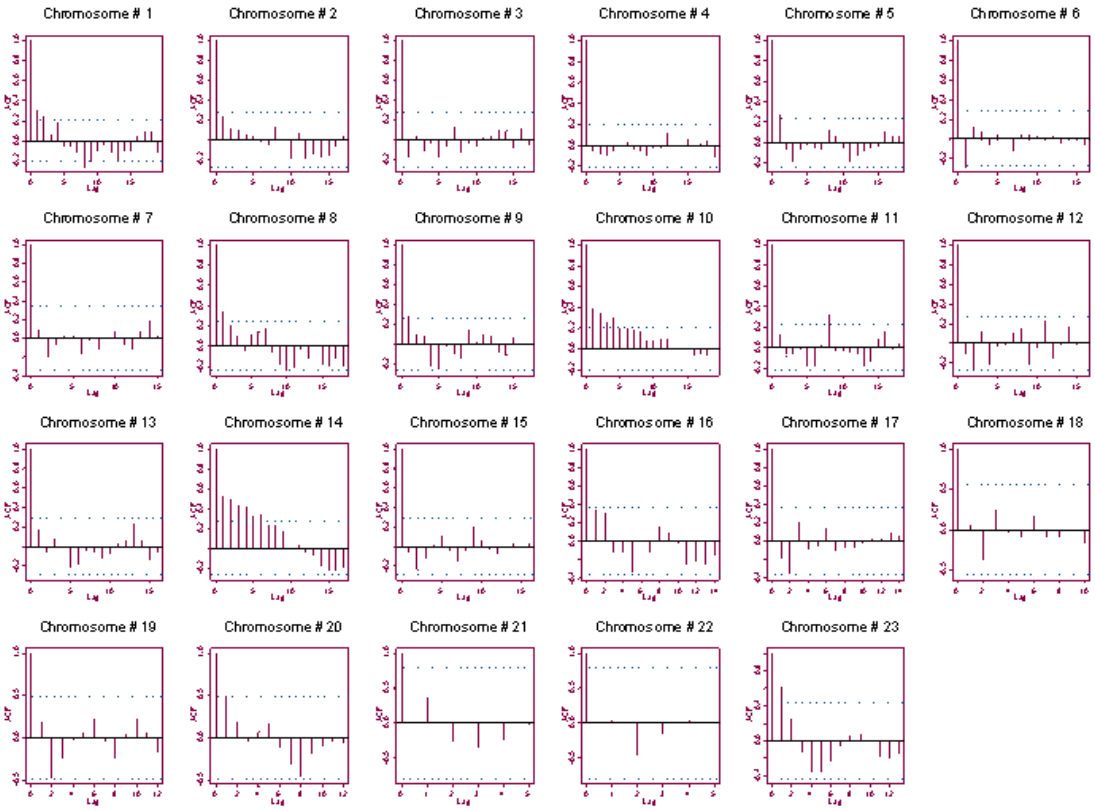

Figure 1.18: The figures show the ACF plot for the residuals obtained for chromosome 1 to 23 using data set in subsection 1.4.2. The residuals in few of the chromosomes indicate the presence of high autocorrelation.

ACF plot for individual chromosomes for Application-2

Here the residuals are obtained from real data mentioned in subsection 4.2. The ACF plots in Figure 1.18 indicate the presence of high autocorrelation among the residuals in chromosome numbers 1, 4, 7, 8, 9, 10, 11, 13, 14 and 21. Therefore, considering the residuals to be i.i.d. would not be realistic in detecting the abnormal regions in this CGH Array.

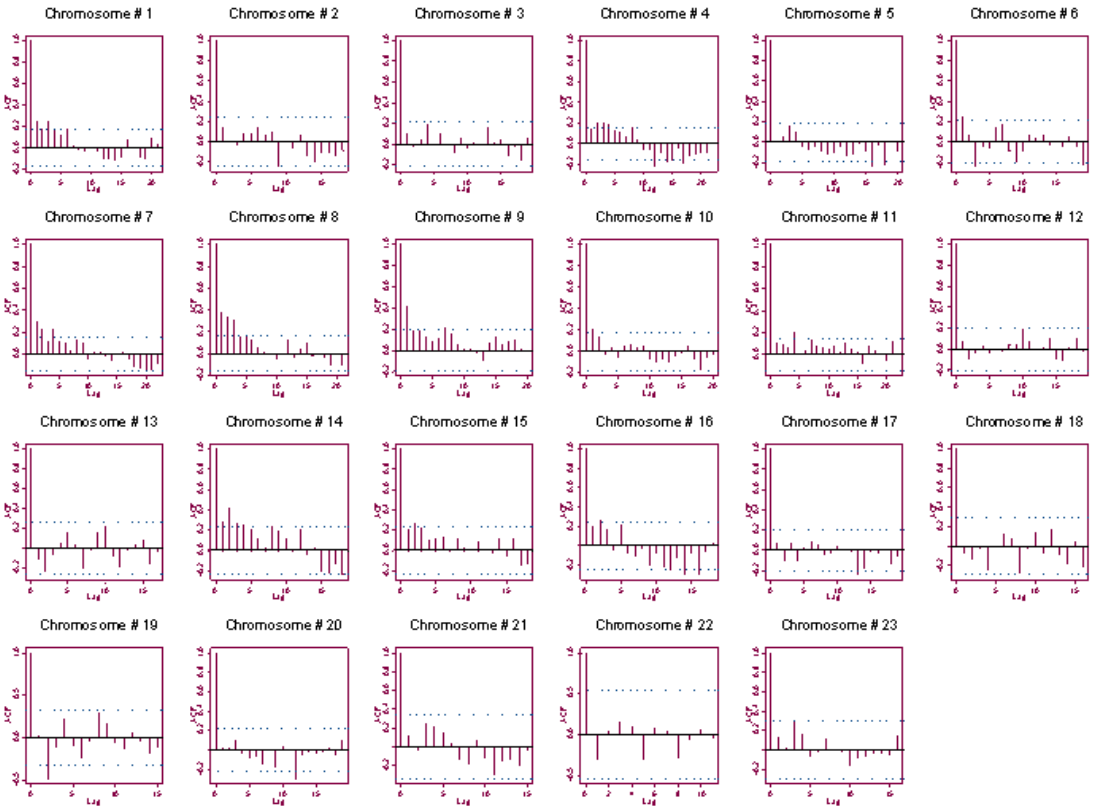

Figure 1.19: The figures show the ACF plot for the residuals obtained from the normal array described in subsection 1.4.2. There exists high autocorrelation among the residuals within some of the chromosomes.

ACF Plot for Normal Array from Application-2

The normal array described in subsection 4.2, are analyzed for the presence of autocorrelation in the error term. Figure 1.19 reveals that there is presence of dependence characteristic in residuals within many of the chromosomes; for example, we can note the presence of high autocorrelation in chromosome numbers 1, 4, 7, 8, 9 and 14.

Bibliography

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289-300.

Clevelend, R.B., Clevelend, W.S., McRae, J.E. and Terpenning, I. (1990) STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *J of Official Statistics*, **6**, 3-73.

Constantine, W.L.B., Percival, D.B. (2003) S+Wavelets 2.0. *Insightful Corporation*, Seattle, WA.

. Daubechies, I. (1992) Ten Lectures on Wavelets. *Philadelphia: Society for Industrial and Applied Mathematics*.

Fisher, R.A. (1921). Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk. *J. Agric. Sci.*, **11**: 107-135.

Hodgson, G., Hager, J., Volik, S., Hariono, S., Wernick, M., Moore, D., Nowak, N., Albertson, D., Pinkel, D., Collins, C., Hanahan, D. and Gray, J.W. (2001). Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature Genetics* **29**, 491.

Jong, K., Marchiori, E., Vaart, A., Ylstra, B., Weiss, M., Meijer, G. (2003). Chromosomal breakpoint detection in human cancer. *In LNCS*, (2611), Springer.

Lai, W.R., Johnson, M.D., Kucharlapari, R. and Park, P.J. (2005) Comparative analysis of algorithm for identifying amplications and deletions of rray CGH data. *Bioinformation*, **21**, 3763-3770.

Lingjaerde, O.C., Baumbusch, L.O., Lisestol, K., Glad, I.K. and Borrsen-Dale, A. (2005) CGH Explorer: a program for analysis of array-CGH data. *Bioinformation*, **21**, 821-822.

Matlab (2007). Detecting Discontinuities and Breakdown Points. *In Wavelet Toolbox: Wavelet Applications*.

Ogden, T. and Parzen, E. (1976). Data dependent wavelet thresholding in non-parametric regression with change-point applications. *Computational Statistics & Data Analysis*, **22**, 53-70.

Percival, D.B. and Walden, A.T. (2000). *Wavelet Methods for Time Series Analysis*. Cambridge University Press.

Pinkel, D. and Albertson, D.G. (2005) Array comparative hybridization and its applications in cancer. *Nature Genetics*, **37**, S11-S17.

Pollack J.R., Sorlie, T., Perou, C.M., Rees, C.A., Jeffry, S.S., Lonning, P.E., Tibshirani, R., Botstein D., Borrsen-Dale, A. and Brown, P.O. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences*, USA, **99**, 1296312968

Pounds,S. and Cheng C. (2004) Improving false discovery rate estimation, *Bioinformatics*, **20**, 1737-1745.

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown N., Conroy, J., Hamilton, G., Hindle, A.K, Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A. N., Pinkel, D., Albertson, D. (2001)Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics* **29**, 263 - 264.

Storey,J.D.(2002) A direct approach to false discovery rates. *J. R. Stat. Soc. B*, **64**, 479-498.

Wang, Y. (1995). Jump and Sharp Cusp Detection by Wavelets. *Biometrika*, **82**, 385-397.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B., Tibshirani, R. (2005). Studies in crop variation. I. A method for calling gains and losses in array CGH data. *Biostatistics*, **6**, 1, 4558

Willenbrock H, Fridlyand J. (2005). A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, **21**, 40844091.