IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)
e-ISSN: 2279-0853, p-ISSN: 2279-0861.Volume 16, Issue 7 Ver. I (July. 2017), PP 120-127
www.iosrjournals.org

Change Detection and Application to CGH Arrays

*Dr. M. Shahidul Islam
Professor Department of Statistics Shahjalal University of Science & Technology Sylhet, Bangladesh
shahed-sta@sust.edu
**Corresponding author: Dr. M. Shahidul Islam **

Abstract: *Copy number changes, also called chromosome gains or losses in the DNA content, have drawn recent attention in advancement of cancer research. Array CGH (Comparative Genomic Hybridization) is a molecular-cytogenetic method for genomewide screening for such loss and gain regions referring to genetic alterations. In this article, we present a simple but very effective method to uncover the locations of copy number changes through the use of modified maximal overlap discrete wavelet transform (MODWT) devised with some threshold approach. Implementations to simulated data having autocorrelation structure as well as to real CGH array confirm the excellent performance of this procedure.*

Key Words: *change detection; array CGH; wavelet; lattice plot; R package WaveCD.*

I. Introduction

A normal human cell contains two copies of each genomic segment, but this copy number changes from two in case of genetic alterations (Rancoita2009; Wang2005). 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. An advancement in cancer research can be availed through precise identification of the regions with DNA copy number alterations. *Comparative Genomic Hybridization* (CGH) microarray is a technique for measuring such changes (PinkelAlbertson2005). More recently, cDNA and oligonucleotide arrays have become popular for CGH. 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 Lai2005 and Willenbrock2005 for comparative reviews of the analysis methods.

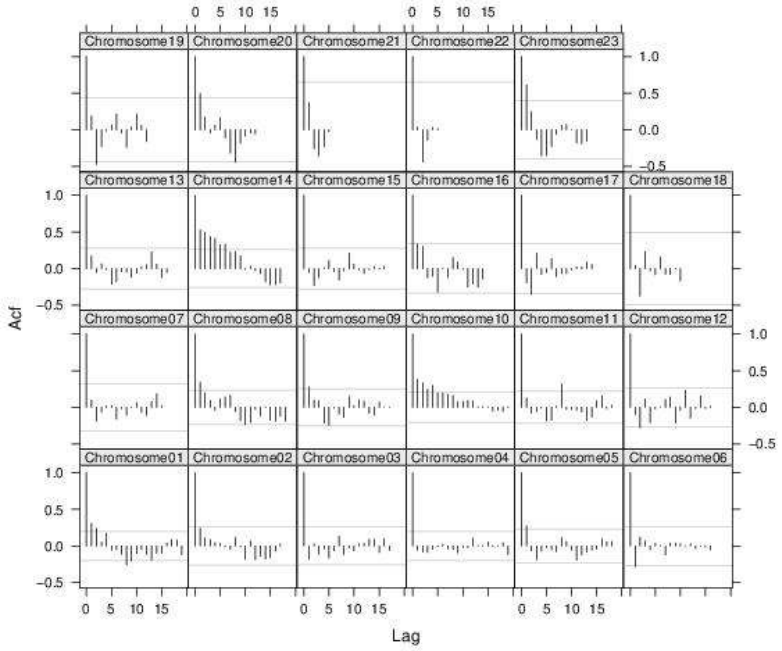

Figure 1: In this data set, 2400 *bacterial artificial chromosome* (BAC) clones were measured each with three replicates (Snijders2001). Measurements for log base 2 intensity ratio are provided. Average relative DNA copy number sequences of the three replicates in first 12 chromosomes are shown in this figure. The change points, detected by the proposed MOWDT, are marked as vertical red lines. The approach is applied independently on each chromosome.

In a CGH experiment, a test sample is labeled with red fluorescent dye (Cy5) and a reference sample is labeled with green fluorescent dye (Cy3). They are applied together on the array chip. The cDNA segments from both samples are then bound with complementary cDNA segments on the array, termed hybridization, and the corresponding dyes are thereafter left on the array to be caught by laser scanner. The ratios of the intensities of the red dye over the green dye, referring to test over reference sample, are therefore used to measure the relative DNA copy numbers. 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. 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 identifying the regions with DNA copy number alterations. Most notable methods are moving average to the process of ratios and used normal versus normal hybridization to compute the threshold (Pollack2002); maximum likelihood to fit mixture models corresponding to gain, loss and normal regions (Hodgson2001); 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 (Lingjaerde2001). The algorithm *Cluster Along Chromosomes* (CLAC) builds hierarchical clustering-style trees along each chromosome arm, and then selects the clusters by controlling the *False Discovery Rate* (FDR) at a certain level (Wang2005). Bayesian Piecewise Constant Regression (BPCR) for such change detection was also proposed Hutter2007. This method works through recognizing the data as deviation from piecewise constant function and thus aims at finding the lengths and end points of each constant function. In a recent article, it was claimed that mBPCR performs as an improved version of BPCR and so can detect small changes even in comparably much noisy data Rancoita2009. Among other recent methods, a mean and variance change-point (MVCM) model for change point detection can be used Chen2009.

In the present method, the log-ratio sequence is viewed as a time series sequence along the genome and possible correlation between clones at closer physical locations is taken into account. The challenge of change detection can be solved through edge detection technique (Yu2007). There are different wavelet methods available; however, for change detection Haar wavelets are often preferred (Brillinger1994). In the MATLAB *User's Guide* the problem of detecting discontinuities in a signal is discussed and it is recommended that Haar wavelets be used (MatLab2009). Similar to the method proposed by Wang1995, this is completely exploratory and requires examining graphs at different levels. Inevitable confusion may also arise for series with unknown change points. In the proposed modified approach, Haar wavelet at level 1 is considered for any series. We prefer MODWT over discrete wavelet transform (DWT) in order to overcome the limitations of dyadic length requirement and sensitivity of the starting point of the time series (Percival2000). Once a jump point is detected through some threshold guide, that value is deleted and rerun the procedure with updated threshold value to differentiate a signal from noise. Thus, for a series with substantial length, the method assures a probability of almost zero that a noise would be detected as change points. We present simulation results which mimic real data sets in the sense that autocorrelation along the series is considered. The justification of using autocorrelation model is provided in Appendix. Nevertheless, real array CGH is also analyzed and presented in lattice-style graph. The whole process is implemented in freely available software R and the package WaveCD is available at CRAN (Islam2010).

II. Theoretical Background

Let $Z_t, t=1,2,\ldots,n$ be the measure of the relative DNA copy numbers of n clones along each chromosome. 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.

2.1 Wavelets

Wavelets are well established in the mathematical sciences (Daubechies1992) 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 satisfies the following conditions:

$$\int_{-\infty}^{\infty} \psi(u)\mathrm{d}u = 0 \qquad (1)$$

$$\int_{-\infty}^{\infty} \psi^2(u)\mathrm{d}u = 1 \qquad (2)$$

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 tell 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.

Let dilation and translation through the integers s and λ get the mother function $\psi(x)$ to generate wavelets basis:

$$\psi_{s,\lambda}(x)=2^{s/2}\psi(2^{s}x-\lambda) \qquad (3)$$

The family $\{\psi_{s,\lambda}(x)\}$ for $s\in Z$ and $\lambda\in N$ with dyadic dilation and translation refers to MODWT.

2.2 Change Detection

Note that for any CGH data, there could be many change points along the genome and these points 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 adjacent clone. Copy number sequences along the genome can therefore be envisaged as a time series. As in time series analysis, Z_t is expected to be autocorrelated and this means that methods which assume independence may be expected to give incorrect or spurious results (Granger2001).

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. Methods are available for detecting discontinuities in a signal, but these require exploration at different levels of wavelet coefficients (MatLab2009).

The underlying model or null hypothesis may be written as $Z_t=f(t)+a_t, t=1,\ldots,n$ where $a_t \sim \text{NID}(0,\sigma^2)$ and $f(t)$ is a smooth function. We are interested in testing an abrupt change in the function $f(t)$. This model is more general and focuses on detecting the change points where a jump or sharp cusp occurs. A sharp cusp occurs at point t_0 if there exists a constant $K>0$ such that

$$|f(t_0+h)-f(t_0)|\geq K|h|^{\alpha} \qquad (4)$$

for all h as $h\to 0$ and $0\leq\alpha<1$. When $\alpha=0$, the function has a jump.

2.3 Threshold Guide

If $Z_1, Z_2, \ldots, Z_n$ are normally distributed with mean 0 and variance σ^2, then using the universal threshold (UT), described in Johnstone1997, we get

$$\text{Prob}(\max_{i=1,2,\ldots n}|Z_i|>\sigma\sqrt{()2\log n})\to 0 \qquad (5)$$

For detection of change points, Wang1995 recommended to use Daubechies wavelets 1 and up and look for jumps in the wavelet coefficients graphically using the threshold value $\sigma\sqrt{()2\log n}$. In this case, the unknown value σ may be estimated robustly from the lowest level of wavelet coefficients. In practice, it is not very satisfactory to examine the wavelet coefficient plots at different levels and then select the jump points, since this is a subjective and tedious procedure. We need to use some automated procedure for selecting appropriate levels for any given series.

III. Proposed Approach

By intuition, we can assume that any normal or abnormal regions comprise of more than one observation. That is, if any jump contains single observation then this refers to some noise in the series. We apply MODWT at level one and record the observation numbers where the wavelet coefficients are greater than the universal threshold. In order to verify that the wavelet coefficients correspond to right jump detection, we delete the observations where the jumps are detected and rerun the procedure. The signs and magnitudes of the new wavelet coefficients (NWC) corresponding adjacent to the deleted observations are noted. Large wavelet coefficients but with different signs than that of the original coefficients do not refer to any real jumps. However, a jump is differentiated from random noise when the signs are same but the coefficients are greater than the UT. Under the theorem stated in subsection

2.3, it follows that for a given series of size 20, the expected proportion of times the maximum value of the series exceeds the threshold is approximately 18%. With increasing n up to 2^{19}, the proportion will not exceed 7.5% (Johnstone1997). In modified approach we delete the value with large coefficients and rerun the MODWT. Thus we have another new set of coefficients with updated threshold level. The probability of exceeding threshold values in consecutive two runs get substantially decreased. Therefore, even for small value of n, say n=20, the probability that maximum values in initial and updated series exceed the thresholds becomes as low as 3%.

IV. Simulated Examples

Let Z_t, t=1,2,…,n be the observations along a specific chromosome arm. Our aim is to find jump points that are present in the series. In this section we present two simulated examples to demonstrate the performance of the proposed method.

4.1 Two Loss/Gain Regions

A set of 270 observations are simulated in two blocks representing two chromosomes. The series size for two chromosomes are 150 and 120 respectively.

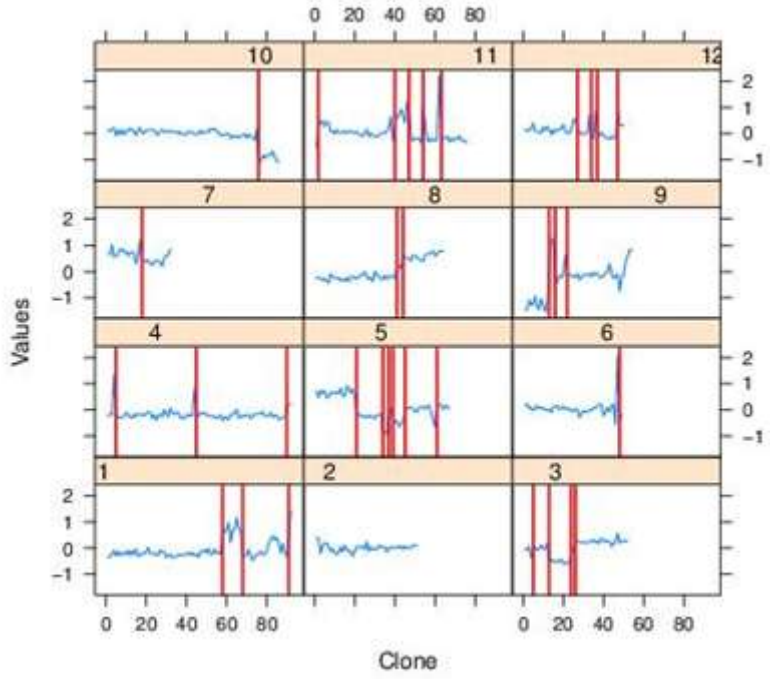

Figure 2: Application of wavelet method to the series with error term following AR(1). The error model is $e_t=\phi e_{t-1}+a_t$, where $a_t \sim \text{NID}(0,\sigma_a)$. Here we consider ϕ=0.8 and σ_a=0.2. The method was able to detect the jump points at exact places.

We consider same model for both the settings as $Z_t=\mu_t+e_t$, where μ_t takes on values 0, 0.7, and −0.7. That is,

$$\mu_{t1}=\begin{cases} 0; & 1\le t\le 80 \\ -0.7; & 81\le t\le 110 \\ 0; & 111\le t\le 150 \end{cases} \quad \textit{forchromosome}1$$

$$\mu_{t2}=\begin{cases} 0; & 1\le t\le 40 \\ -0.7; & 41\le t\le 70 \\ 0; & 71\le t\le 120 \end{cases} \quad \textit{forchromosome}2$$

For each chromosome, we consider that the innovation term follows simple autoregressive (AR) process of order 1. The expression for this AR(1) process can be expressed as,

$$e_t=\phi e_{t-1}+a_t, \quad a_t \sim \text{NID}(0,\sigma_a) \qquad (6)$$

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 simulate series of observations for different values of ϕ, say 0.4, 0.6 and 0.8. The proposed method is applied for each instances and the performance is evaluated. It is found that wavelet approach provides exact change detection for all the cases. To get a glimpse of the performance of the method, only case with ϕ=0.8 is presented in Figure 2.

4.2 Seven Jump Points

The data set consists of 200 observations having 7 jump points at 50, 60, 92, 106, 145, 169 and 180. We consider the model as $Z_t=\mu_t+e_t$, where $e_t \sim NID(0,0.1)$ and μ_t is expressed as:

$$\mu_t = \begin{cases} 0; & 1\le t<50 \\ 0.6; & 50\le t<60 \\ 0; & 60\le t<92 \\ -0.6; & 92\le t<106 \\ 0; & 106\le t<145 \\ 0.6; & 145\le t<169 \\ 1; & 169\le t<180 \\ 0; & 180\le t\le 200 \end{cases}$$

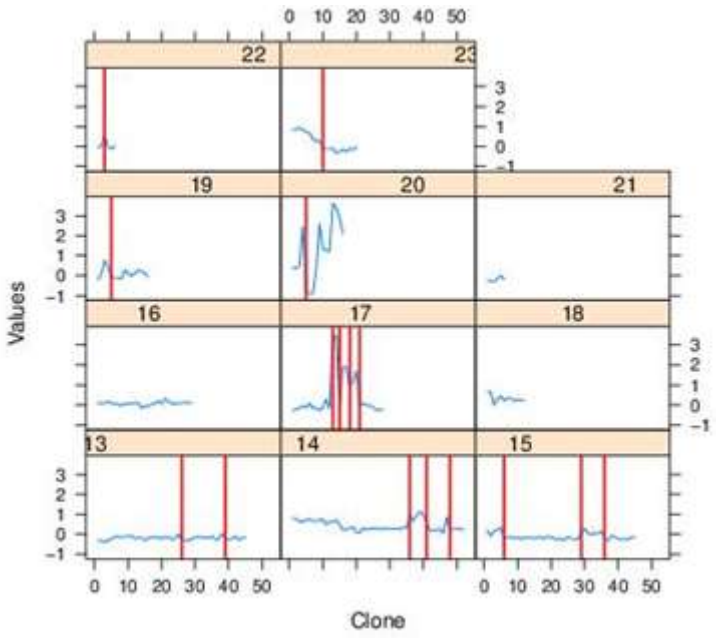

Figure 3: This represents a series, $Z_t=\mu_t+e_t$, with seven jump points located at 50, 60, 92, 106, 145, 169 and 180. Here $e_t \sim NID(0,0.1)$ and the mean value μ_t takes different values at different intervals. The proposed method detects the jump points at exact locations.

Unlike the previous example, here the error term does not have any autocorrelation structure. This is a typical example where there are two successive gain regions within second chromosome. Figure 3 reveals that the wavelet method is capable of detecting all break points at the right places.

V. Application to CGH Array

We apply our proposed method to a CGH array where 2400 bacterial artificial chromosome (BAC) clones were measured each with three replicates (Snijders2001). 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 and 4.

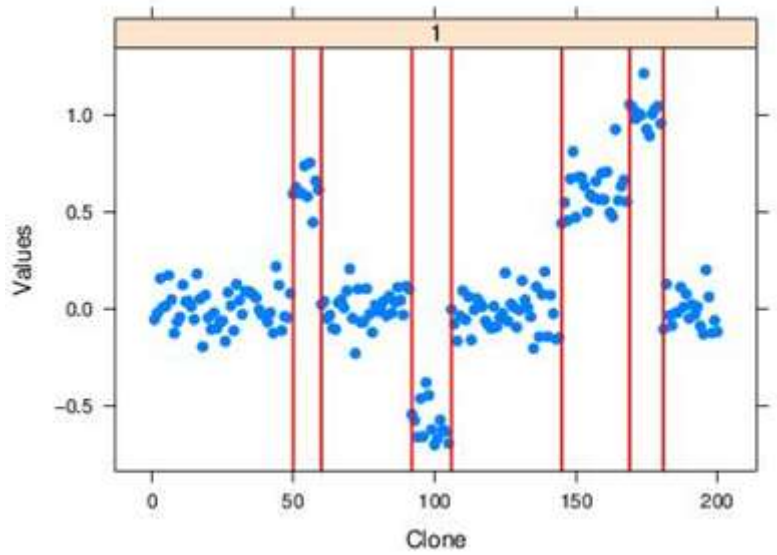

Figure 4: Representation of break points detected by wavelet method for copy number changes in last 11 chromosomes. In this data set, 2400 BAC clones were measured each with three replicates (Snijders2001). The figure demonstrates different possible break points along several chromosomes.

The figures also demonstrate the break points that are detected using the wavelet 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. Many jump points are apparent from the implementation of the wavelet method. We see from Figure 1 in Section

that there are jump points in all chromosomes but 2. Nevertheless, many of the chromosomes hold multiple breaks. Figure 4 presents the chromosome-wise abnormal regions for other chromosomes. This figure also depicts the presence of several jump points along many chromosomes.

VI. Conclusion

In this article, a wavelet method has been proposed to identify the positions of the change in DNA copy number throughout the genome. Discrete wavelet transform has two limitations; namely dyadic length requirement and sensitivity of the starting of the time series. To overcome such limitations, *maximum overlap discrete wavelet transformation* (MODWT) is used and break points are detected using universal threshold. A noise is differentiated from a real signal through rerunning the procedure with modified threshold. Thus the probability of calling a normal region to be a region of genetic alteration approximates to zero with the increasing size of clones. Through simulated examples, it was demonstrated that the method performs quite well in selecting the break points which mimic abnormal regions in a real CGH data. The autocorrelation structure among the clones might be very crucial to consider. The first simulation study clearly demonstrates that the method is invariant of strict assumption of independent and identically distributed observations. One more advantage of the method is that it is computationally very efficient and applicable to very large number of clones. Although many change points are flagged in the real CGH data, biological interpretation is needed for final comment. R package WaveCD is freely available for the implementation of the proposed approach. All the results and lattice graphs can be reproduced using this package.

VII. Appendix

Autocorrelation function (ACF) is useful in detecting the presence of correlation among the successive observations. We find the break points for sequence of observations corresponding to each chromosome and obtain the residuals by subtracting the mean of the selected region from the observations in that region. That is, $e_{tkj}=z_{tkj}-\mu_{tk}$ is jth residual for kth region in tth chromosome. ACF plots are presented for the residuals obtained from the application in CGH array described in Section **Error! Reference source not found.**.

Figure 5 is constructed to show the autocorrelation behavior of the error process for each chromosome. It seems that the residuals are not quite IID within each of the chromosome. The residuals in chromosome numbers 1, 8, 10, 14 and 23 demonstrate the presence of strong autocorrelation. This justifies and depicts the simulation study presented in first example of Section **Error! Reference source not found.**.

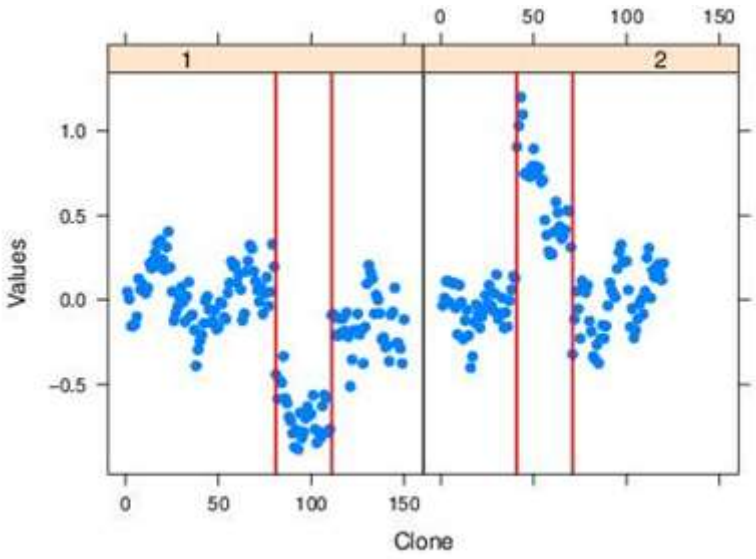

Figure 5: ACF plots for the residuals obtained for all chromosomes using the data set in section **Error! Reference source not found.**. The residuals are obtained by subtracting region mean from corresponding observations. The residuals in chromosomes 1, 8, 10, 14 and 23 indicate the presence of autocorrelation.

References

[1]. [Brillinger, 1994]Brillinger, D.R. (1994). Some river wavelets, *Environmetrics*, **5**, 211–220.

[2]. [Chen and Wang, 2009]Chen, J. and Wang, Y. P.(2009). A statistical change point model approach for the detection of dna copy number variations in array cgh data, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **6**, 529–541.

[3]. [Daubechies, 1992]Daubechies, I. (1992). Ten lectures on wavelets, *Society for Industrial and Applied Mathematics*, Philadelphia, PA, USA.

[4]. [Granger, 2001]Granger, C.W.J. (2001). *Essays in econometrics: collected papers of clive W. J. Granger*, vol. II, Cambridge University Press, Cambridge.

[5]. [Hodgson *et al*., 2001]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.

[6]. [Hutter, 2007]Hutter, M. (2007). Exact bayesian regression of piecewise constant functions, *Bayesian Analysis*, **2**, no. 4.

[7]. [Islam, 2010]Islam, M. S.(2010). WaveCD: Wavelet change point detection for array CGH data [url= http://CRAN.R-project.org/package=ascrda], [R package version 1.0].

[8]. [Johnstone and Silverman, 1997]Johnstone I.M. and Silverman, B.W. (1997). Wavelet threshold estimators for data with correlated noise, *Journal of the Royal Statistical Society B*, **59(2)**, 319–351.

[9]. [Lai *et al*., 2005]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, *Bioinformatics*, **21(19)**.

[10]. [Lingjaerde *et al*., 2001]Lingjaerde, O.C.,Baumbusch, L.O., Lisestol, K., Glad, I.K. and Borrsen-Dale, A. (2001). Cgh explorer: a program for analysis of array-cgh data, *Bioinformatics*, **21(6)**.

[11]. [Misiti *et al*., 2009]Misiti, M., Misiti, Y., Oppenheim, G., and Poggi, J.M. (2009). *Wavelet toolbox user's guide*, The MathWorks, Inc., MA, 4 Ed.

[12]. [Percival and Walden, 2000]Percival, D.B. and Walden, A. T. (2000). *Wavelet methods for time series analysis*, Cambridge University Press, Cambridge.

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

[14]. [Pollack *et al*., 2002]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*, **99**, 12963–-12968.

[15]. [Rancoita *et al*., 2009]Rancoita, P.M.V., Hutter, M., Bertoni, F. and Kwee, I. (2009). Bayesian dna copy number analysis, *BMC Bioinformatics*, **10(10)**.

[16]. [Snijders *et al*., 2001]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. and Albertson, D. (2001). Assembly of microarrays for genome-wide measurement of dna copy number, *Nature Genetics*, **29**, 263–264.

[17]. [Wang *et al*., 2005]Wang, P., Kim, Y., Pollack, J., Narasimhan, B. and Tibshirani, R. (2005). A method for calling gains and losses in array cgh data, *Biostatistics*, **6(1)**, 45–58.

[18]. [Wang, 1995]Wang, Y. (1995). Jump and sharp cusp detection by wavelets, *Biometrika*, **82**, 385–397.

[19]. [Willenbrock and Fridlyand, 2005]Willenbrock, H. and Fridlyand, J. (2005). A comparison study: applying segmentation to array cgh data for downstream analyses, *Bioinformatics*, **21(22)**, 4084–4091.
[20]. [Yu *et al*., 2007]Yu, T, Ye, H., and Sun, W., (2007). A forward-backward fragment assembling algorithm for the identification of genomic amplification and deletion breakpoints using high-density single nucleotide polymorphism (snp) array, *BMC Bioinformatics*, **8(145)**.

*Dr. M. Shahidul Islam. "**Change Detection and Application to CGH Arrays**." IOSR Journal of Dental and Medical Sciences (IOSR-JDMS) 16.7 (2017): 120-127.