Stationary Wavelet Packet Transform and Dependent Laplacian Bivariate Shrinkage Estimator For Array-CGH Data Smoothing

Nha Nguyen[1,2], Heng Huang[1]*, Soontorn Oraintara[2] and An Vo[3]

September 6, 2009

Abstract

Array based comparative genomic hybridization (aCGH) has merged as a highly efficient technique for the detection of chromosomal imbalances. Characteristics of these DNA copy number aberrations provide the insights into cancer, and they are useful for the diagnostic and therapy strategies. In this paper, we propose a statistical bivariate model for aCGH data in the stationary wavelet packet transform (SWPT) and apply this bivariate shrinkage estimator into the aCGH smoothing study. Because our new dependent Laplacian bivariate shrinkage estimator covers the dependency between wavelet coefficients and the shift invariant SWPT results include both low and high frequency information, our dependent Laplacian bivariate shrinkage estimator based SWPT method

[1]Department of Computer Science and Engineering, University of Texas at Arlington, TX, USA.
[2]Department of Electrical Engineering, University of Texas at Arlington, TX, USA.
[3]The Feinstein Institute for Medical Research, North Shore LIJ Health System, New York, USA.
*Corresponding Author.

(named as SWPT-LaBi) has fundamental advantages to solve aCGH data smoothing problem compared to other methods. In our experiments, two standard evaluation methods, the Root Mean Squared Error (RMSE) and the Receiver Operating Characteristic (ROC) curve, are calculated to demonstrate the performance of our method. In all experimental results, our SWPT-LaBi method outperforms the previous most common used aCGH smoothing algorithms on both synthetic data and real data. Meantime, we also propose a new synthetic data generation method for aCGH smoothing algorithms evaluation. In our new data model, the noise from real aCGH data is extracted and used to improve synthetic data generation. Implementation and data will be available under software tab at: http://ranger.uta.edu/∼heng/aCGH and http://naaan.org/nhanguyen/

Keywords: DNA Copy Number, Array Comparative Genomic Hybridization, Smoothing, Stationary Wavelet Packet Transform

1 INTRODUCTION

Gene amplifications or deletions frequently contribute to tumorigenesis. When part or all of a chromosome is amplified or deleted, there are changes in DNA copy number results. Characterization of these DNA copy number changes is important for both fundamental understanding of cancers and their diagnosis. For cancers study, researchers currently use array Comparative Genomic Hybridization (aCGH) to identify sets of copy number changes associated with the particular cancer or its congenital and developmental disorders. In aCGH data, because the clones contain sequences information directly connecting with the genome database, aCGH offers rapid genome-wide analysis in high resolution and those information is directly linked to the physical and genetic maps of the human genome. Bacterial Artifi-

cial Chromosomes (BAC) based aCGH arrays were amongst the first genomic arrays to be introduced (Pinkel *et al.*, 1998) and are routinely used to detect single copy changes in the genome, owing to their high resolution in the order of 1 Mb (Pinkel *et al.*, 1998; Snijders *et al.*, 2001). More recently Oligonucleotide aCGH (Brennan *et al.*, 2004; Pollack *et al.*, 1999) was also developed to allow flexibility in probe design, greater coverage, and much higher resolution in the order of 35-100 Kb (Wang *et al.*, 2007).

Because aCGH is very noisy, many diseases related chromosomal aberrations are buried by noise. For example, in cDNA array CGH data, the signal to noise ratio is often approximately 1 (0 dB) (Bilke *et al.*, 2005). In order to develop effective methods to identify aberration regions from array CGH data, many research works focus on both smoothing/denosing-based and segmentation-based data processing. Segmentation-based methods target to model data as a series of discrete segments with unknown boundaries and unknown heights. Since the boundary points are highly possible to be identified as aberration region, the false positives are introduced. Smoothing-based methods reduce noise by comparing each data point to its adjacent ones and reduce the number of identified false aberration regions.

Beheshti *et al.* proposed to use the robust locally weighted regression and smoothing scatterplots (lowess) method in paper (Beheshti *et al.*, 2003). Eilers and Menezes (Eilers *et al.*, 2005) performed a quantile smoothing method based on the minimization of the sum of absolute errors to create sharper boundaries between segments. Hsu *et al.* (L.Hsu *et al.*, 2005) investigated the usage of maximal overlap discrete wavelet transform (MODWT) in the analysis of array CGH data. In 2005, Lai (Lai *et al.*, 2005) compared 11 different algorithms for analyzing array CGH data. Many smoothing and estimation methods were included in(Lai *et al.*, 2005) such as CGHseg (2005) (Picard & *et al.*, 2005), Quantreg

(2005) (Eilers *et al.*, 2005), CLAC (2005) (Wang *et al.*, 2005), GLAD (2004) (Hupe *et al.*, 2004), CBS (2004) (Olshen *et al.*, 2004), HMM (2004) (Fridkyand *et al.*, 2004), MODWT (2005) (L.Hsu *et al.*, 2005), Lowess (Beheshti *et al.*, 2003), ChARM (2004) (Myers *et al.*, 2004), GA (2004) (Jong *et al.*, 2004), and ACE (2005) (Lingjaerde *et al.*, 2005). Based on empirical experiments, Lai (Lai *et al.*, 2005) concluded that MODWT, Quantreg and Lowess methods gave better detection results (higher true position rate and lower false position rate) than other methods. Meantime, the wavelet (MODWT) based smoothing method was considered as the most promising approach. More recently Y. Wang and S. Wang (Wang *et al.*, 2007) extended the stationary wavelet (SWT or MODWT) denoising for nonequal spaced data, because the physical distances between adjacent probes along a chromosome are not uniform, even vary drastically. In paper (Nguyen *et al.*, 2007), Nguyen *et al.* developed another wavelet based method using DTCWTi-bi (Dual tree complex wavelet transform - interpolation - bivariate shrinkage function) technique with better performance. However, if a signal is decomposed by using SWT (MODWT) or DTCWT, we get nonuniform sub-bands and a wide sub-band in high frequency. Because true aCGH signals include many step functions, they contain important information in high frequency. The above wavelet based methods do not offer enough sub-bands in high frequency for smoothing operation.

In this paper, we propose to use shift invariant SWPT with dependent Laplacian bivariate shrinkage estimator (named as SWPT-LaBi) for aCGH data smoothing. In the SWPT, all sub-bands are shift invariant and each sub-band provides a shiftable description of signal in a specific scale as the same as the SWT or the MODWT. Such shift invariant property is crucial to apply wavelet based method into aCGH data smoothing. Although the Discrete Wavelet Transform (DWT) with the redundant ratio of $1:1$ is efficient for computation, it

is not suitable for aCGH smoothing application. Because DWT creates artifacts around the discontinuities of the input signal (Coifman *et al.*, 1995) and is shift-variant. Because the SWPT also decomposes signal to many uniform frequency sub-bands, information in both of low and high frequency sub-bands are captured. However the previous wavelet based methods lose the information in high frequency (L.Hsu *et al.*, 2005; Wang *et al.*, 2007; Nguyen *et al.*, 2007).

Moreover, we propose dependent Laplacian bivariate shrinkage function to exploit the dependency between wavelet coefficients and its cousin in SWPT to improve the performance. We demonstrate the effectiveness of our approach through theoretical and experimental explorations of a set of aCGH data, including real data and synthetic data with both gaussian and real noise. It is the first time to create synthetic aCGH data using real noise. Our new synthetic data generation model provides a more accurate validation way to evaluate aCGH smoothing algorithms. We compare the performance between our method and previous methods by root mean squared error (RMSE) and receiver operating characteristic (ROC) curve that are the standard performance comparison criterions. The experimental results show that our method outperforms the previous approaches about $5\% - 59.3\%$ under gaussian noise and $7.9\% - 51.8\%$ under real noise.

2 METHODOLOGY

2.1 Stationary Wavelet Packet Transform

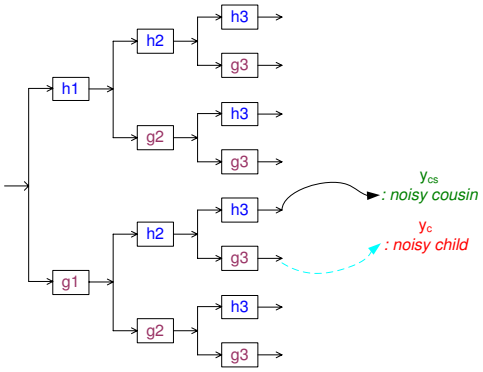


Figure 1: The 3 level SWPT filter bank structure.

Stationary Wavelet Packet Transform (SWPT), shown in Fig. 1, is a generalization of Stationary Wavelet Transform (SWT). First, a signal is decomposed into a low frequency sub-band and a high frequency sub-band by using two channels filter bank. Similar to the SWT, the SWPT does not employ a decimator after filtering. Then the low frequency sub-band as well as the high frequency sub-band can be decomposed into a second-level low and high frequency sub-band, and the process is repeated as in Fig. 1. Each level's filter are upsampled versions of the previous ones. The absence of a decimator leads to a full rate decomposition. Each sub-band contains the same number of samples as the input. So for a decomposition of L levels, there is a redundant ratio of $2^L : 1$. However, the absence of a decimator makes the SWPT shift invariant. This shift invariant property makes the SWPT preferable for the usage in various signal processing applications such as denoising and classification because it relies heavily on spatial information. It has been shown that many of the artifacts could be suppressed by a redundant representation of the signal (Coifman *et al.*, 1995). In the SWT,

the low frequency subband is itself decomposed into two second-level sub-bands. Therefore, the the SWT has nonuniform frequency supports, while the SWPT has uniform frequency supports. As a result, the SWPT offers a richer range of possibilities for signal analysis. With the uniform shift-invariant sub-bands, the SWPT can capture more information from the aCGH data. Thus, we propose to use the SWPT to smooth the aCGH data.

2.2 Dependent Laplacian Bivariate Shrinkage

A general wavelet based denoising algorithm consists of three steps: decompose the noisy signal by wavelet transform, denoise the noisy wavelet coefficients according to specific rules, and take the inverse wavelet transform from the denoised coefficients. The second step is crucial to the whole algorithm. To estimate wavelet coefficients, the most well-known rules are universal thresholding, soft thresholding (Donoho *et al.*, 1994; Donoho, 1995; Johnstone *et al.*, 1997), and BayesShrink (Chang *et al.*, 2000). In these algorithms, the authors assumed that wavelet coefficients are independent. Sendur and Selesnick (Sendur *et al.*, 2002) have recently exploited the dependency between coefficients and proposed a non-Gaussian bivariate function for the child coefficient w_c and its parent w_p in the complex wavelet transform domain. Nguyen *et al.* (Nguyen *et al.*, 2007) successfully applied that function into the complex wavelet transform domain to recover aCGH data and got promising results.

Because the SWPT offers a richer range of shift-invariant sub-bands than the complex wavelet transform and SWT, it is natural to use SWPT to denoise aCGH data. However, the SWPT, which decomposes a signal into many uniform sub-bands at the same scale, only has child and cousin coefficients as in Fig. 1 and a new method is required to capture the relationship between them. In order to solve this problem, we develop a bivariate shrinkage function

which models the relationship of child and cousin coefficients in the SWPT operation of aCGH data.

For any DNA copy number data Y, we can assume it includes the deterministic signal D and the independent and identically distributed (IID) Gaussian noise n. This Gaussian noise has zero mean and variance σ_n^2.

$$Y = D + n. \tag{1}$$

After decomposing the data Y by the SWPT, we get the wavelet coefficients and those coefficients can be formulated as

$$\begin{aligned} y_1 &= w_1 + n_1, \\ y_2 &= w_2 + n_2, \end{aligned} \tag{2}$$

where y_1 and y_2 are noisy wavelet coefficients, w_1 and w_2 are true coefficients, w_2 represents the cousin of w_1 (child), n_1 and n_2 are independent Gaussian noise coefficients. If the cousin scale y_2 is decomposed, we will get detail and approximation coefficients. Let's call y_3 as approximation coefficients of y_2. We can calculate y_3 from y_2 by the follow equation:

$$\begin{aligned} y_3 &= w_3 + n_3, \\ y_3[n] &= h[n] * y_2[n] = \textstyle\sum_{k=1}^{N} (h[n-k].y_2[k]), \end{aligned} \tag{3}$$

where $h[n]$ is the scale filter and N is the length of signal y_2. In general, we can write

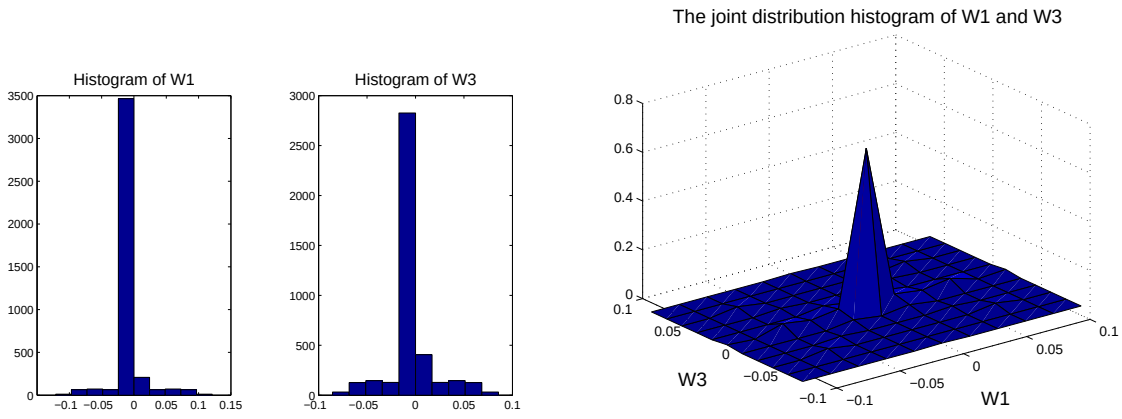
$$\mathbf{y} = \mathbf{w} + \mathbf{n}, \tag{4}$$

where $\mathbf{y} = (y_1, y_3)$, $\mathbf{w} = (w_1, w_3)$ and $\mathbf{n} = (n_1, n_3)$. The noise pdf should be followed as

$$p_{\mathbf{n}}(\mathbf{n}) = \frac{1}{2\pi {\sigma_n}^2} \exp(-\frac{n_1^2 + n_3^2}{2\sigma_n^2}). \tag{5}$$

The standard MAP estimator of $\mathbf{w}$ from $\mathbf{y}$ (Sendur *et al.*, 2002) is followed as

$$\widehat{\mathbf{w}}(\mathbf{y}) = \arg\max_{\mathbf{w}} [\log(p_{\mathbf{n}}(\mathbf{y} - \mathbf{w})) + \log(p_{\mathbf{w}}(\mathbf{w}))]. \tag{6}$$



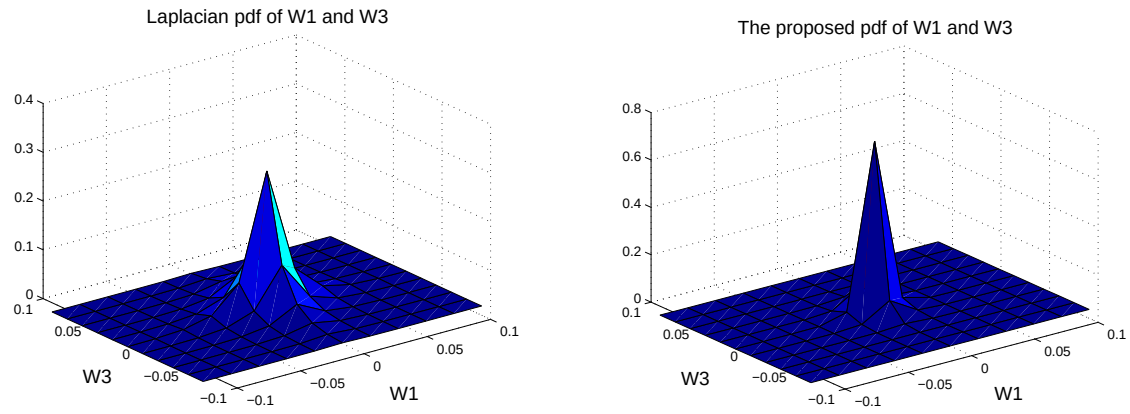
(a) (b) (c)

Figure 2: The histograms computed from true CGH signal. (a) Histogram of w_1 , (b) Histogram of w_3, (c) Joint distribution of w_1 and w_3 created from decomposition of true CGH signal.

The Fig. 2 (a) and (b) illustrate the histograms of the wavelet coefficient w_1 (child) and the approximation coefficient w_3 of w_2(cousin). The w_1 and w_2 are computed from CGH data by using the SWPT. The Fig. 2 (c) shows the joint distribution of w_1 and w_3. We try to find a model for the empirical histogram in Fig. 2 (c). First, we assume that this joint distribution is an independent Laplacian as follows

$$p_{\mathbf{w}}(\mathbf{w}) = \frac{1}{2\sigma^2}\exp(-\frac{\sqrt{2}}{\sigma}|w_1| + |w_3|). \tag{7}$$

It is clear that the independent Laplacian distribution in Fig. 3 (a) does not fit well the empirical histogram in Fig. 2 (c). So, it is not possible to model the empirical histogram with the independent Laplacian distribution. In (Sendur *et al.*, 2002), a general joint pdf which is combined by the independent Laplacian pdf and the dependent component are proposed for image in complex wavelet transform. However, the parameters of the model is tunable.



(a) (b)

Figure 3: (a) The Laplacian pdf with two variables: w_1 and w_3, (b) The proposed pdf with two variables: w_1 and w_3.

So, in the case of the SWPT coefficients of the aCGH data, we propose using this bivariate model with two specific parameters as follows

$$p_{\mathbf{w}}(\mathbf{w}) = \frac{1}{2\pi\sigma^2}\exp(-\frac{\sqrt{3}}{\sigma}\sqrt{|w_1|^2+|w_3|^2}-\frac{\sqrt{2}}{\sigma}(|w_1|+|w_3|)). \tag{8}$$

We can see that the proposed bivariate pdf in Fig. 3 (b) fits well the empirical histogram in Fig. 2 (c). With this pdf, two variables w_1 and w_3 are really dependent and the Eq.(8) is named as dependent Laplacian bivariate model. Let us define

$$f(w) = \log(P_w(w)) = \log(\frac{1}{2\pi\sigma^2}) - \frac{\sqrt{3}}{\sigma}\sqrt{|w_1|^2+|w_3|^2}-\frac{\sqrt{2}}{\sigma}(|w_1|+|w_3|). \tag{9}$$

By using Eq.(5), Eq.(6) becomes:

$$\widehat{\mathbf{w}}(\mathbf{y}) = \arg\max_{\mathbf{w}}[\log(\frac{1}{2\pi{\sigma_n}^2}) - \frac{(y_1-w_1)^2+(y_3-w_3)^2}{2\sigma_n^2} + f(w)]. \tag{10}$$

Solving (10) is the same as solving the two following equations:

$$\frac{(y_1-w_1)}{\sigma_n^2} + f_{w_1}(\widehat{w}) = 0, \tag{11}$$

$$\frac{(y_3 - w_3)}{\sigma_n^2} + f_{w_3}(\widehat{w}) = 0, \tag{12}$$

where f_{w_1} and f_{w_3} represent the derivative of $f(w)$ with respect to w_1 and w_3, respectively. We can get f_{w_1} and f_{w_3} from (9) as

$$f_{w_1}(\widehat{w}) = -(\frac{\sqrt{3}w_1}{\sigma\sqrt{|w_1|^2 + |w_3|^2}} + \frac{\sqrt{2}}{\sigma} sign(w_1)), \tag{13}$$

$$f_{w_3}(\widehat{w}) = -(\frac{\sqrt{3}w_3}{\sigma\sqrt{|w_1|^2 + |w_3|^2}} + \frac{\sqrt{2}}{\sigma} sign(w_3)), \tag{14}$$

where $sign(w)$ is defined as follow:

$$sign(w) = \begin{cases} 0 & \text{if } w = 0, \\ \frac{w}{|w|} & \text{otherwise.} \end{cases} \tag{15}$$

Substituting (13) and (14) into (11) and (12) gives

$$\begin{aligned} \widehat{w_1} \cdot (1 + \tfrac{\sqrt{3}\sigma_n^2}{\sigma r}) &= (|y_1| - \tfrac{\sqrt{2}\sigma_n^2}{\sigma})_+ \cdot sign(y_1) = soft(y_1, \tfrac{\sqrt{2}\sigma_n^2}{\sigma}), \\ \widehat{w_3} \cdot (1 + \tfrac{\sqrt{3}\sigma_n^2}{\sigma r}) &= (|y_2| - \tfrac{\sqrt{2}\sigma_n^2}{\sigma})_+ \cdot sign(y_2) = soft(y_2, \tfrac{\sqrt{2}\sigma_n^2}{\sigma}), \end{aligned} \tag{16}$$

where $r = \sqrt{|\widehat{w_1}|^2 + |\widehat{w_3}|^2}$ and $(u)_+$ is defined by

$$(u)_+ = \begin{cases} 0 & \text{if } u < 0, \\ u & \text{otherwise.} \end{cases} \tag{17}$$

Drawing r from (16)

$$\begin{aligned} r^2 &= \frac{soft(y_1, \frac{\sqrt{2}\sigma_n^2}{\sigma})}{(1 + \frac{\sqrt{3}\sigma_n^2}{\sigma r})} + \frac{soft(y_2, \frac{\sqrt{2}\sigma_n^2}{\sigma})}{(1 + \frac{\sqrt{3}\sigma_n^2}{\sigma r})}, \\ (r + \frac{\sqrt{3}\sigma_n^2}{\sigma})^2 &= soft(y_1, \frac{\sqrt{2}\sigma_n^2}{\sigma}) + soft(y_2, \frac{\sqrt{2}\sigma_n^2}{\sigma}), \\ r &= (\sqrt{soft(y_1, \tfrac{\sqrt{2}\sigma_n^2}{\sigma}) + soft(y_2, \tfrac{\sqrt{2}\sigma_n^2}{\sigma})} - \tfrac{\sqrt{3}\sigma_n^2}{\sigma})_+ \\ &= (R - \tfrac{\sqrt{3}{\sigma_n}^2}{\sigma})_+ . \end{aligned} \tag{18}$$

If replacing r by (18) into (16), the MAP estimator can be written as

$$\hat{w_1} = \frac{(R - \frac{\sqrt{3}\sigma_n{}^2}{\sigma})_+}{R} \cdot soft(y_1, \frac{\sqrt{2}\sigma_n^2}{\sigma}), \tag{19}$$

where R is as follows

$$R = \sqrt{soft(y_1, \frac{\sqrt{2}\sigma_n^2}{\sigma})^2 + soft(y_3, \frac{\sqrt{2}\sigma_n^2}{\sigma})^2}. \tag{20}$$

The (19) is called as dependent Laplacian bivariate shrinkage function. In (19) and (20), σ can be estimated by

$$\hat{\sigma} = \sqrt{(\hat{\sigma}_y^2 - \hat{\sigma}_n^2)_+}, \tag{21}$$

where $\hat{\sigma}_n$ is the noise deviation which is estimated from the finest scale wavelet coefficients by using a robust median estimator (Donoho, 1995) as follows

$$\hat{\sigma}_n^2 = \frac{\text{median}(|y_i|)}{0.6745}. \tag{22}$$

$\hat{\sigma}_y$ is the deviation of observation signal estimated by

$$\hat{\sigma}_y^2 = \frac{1}{M} \sum_{y_i \in N(k)} |y_i|^2, \tag{23}$$

where M is the size of the neighborhood $N(k)$. In the packet wavelet transform, the cousin scales have not any parent scales. In this case, we can use hard thresholding estimator (Donoho *et al.*, 1994) to recover cousin coefficients $\hat{w_{cs}}$:

$$\hat{w_{cs}} = (y_{cs} - \sigma_n \sqrt{2 \log N})_+. \tag{24}$$

2.3 SWPT-LaBi Algorithm

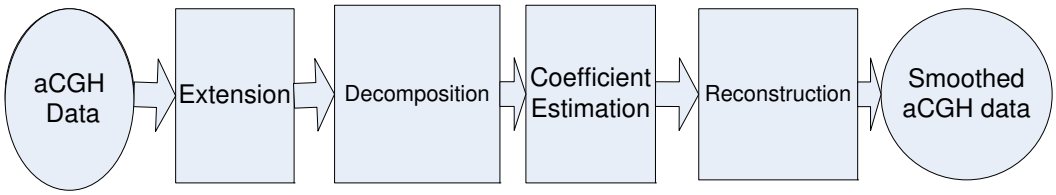


Figure 4: The flowchart of SWPT-LaBi method.

The aCGH data is a finite signal. If we apply wavelet smoothing method directly, the errors will exist at the border of denoised signal. Thus, extension step is a very important preprocessing step before denoising. There are three main extension methods. According to the book (Strang *et al.*, 1996) (chapter 8), symmetric extension is the best one to be applied to a filtered image because it can save information at the border better. With aCGH data, we also need save the information at the border. Therefore, we adopt the symmetric extension method as the preprocessing step before denoising. Let's assume that the length of the aCGH signal is N. In order to get the best performance in the wavelet denoising algorithm, the length of the input signal is required to be a power of two (Coifman *et al.*, 1992). If N is not a power of two, we can extend our signal to make sure $N = 2^j$ using symmetric extension method. Fig. 4 is the flowchart of our SWPT-LaBi algorithm which can be summarized as follows:

Step 1 : *Extend aCGH data Y using symmetric extension method and decompose new data $Y^{'}$ by the SWPT to L levels as Eq.(25). The numbers of decomposition levels (Bruce* et al.*, 1996) (at the remark* 11*) can be computed by*

$$L = \log_2(N) - J, \tag{25}$$

where $J = 3,4,5,6$. This is a perfect number of levels (Bruce et al.*, 1996) which yields the best denoising result. In this paper, we use $J = 4$ as the same as (L.Hsu* et al.*, 2005) and (Wang* et al.*, 2007).*

Step 2 : *Calculate the noise variance $\hat{\sigma}_n^2$ and the marginal variance $\hat{\sigma}^2$ for wavelet coefficient y_k by using Eq.(22), Eq.(23) and Eq.(21).*

Step 3 : *Estimate the child coefficients $\hat{w}_c = \hat{w}_1$ as in Eq.(19) and estimate the counsin coefficients $\hat{w}_{cs}$ as in Eq.(24).*

Step 4 : *Reconstruct data $\hat{D}$ from the denoised coefficients $\hat{w}_c$ and $\hat{w}_{cs}$ by taking the inverse SWPT.*

The error of smoothing result could be measured by the root mean squared error (RMSE) that is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}^{N}(\hat{D}_i - D_i)^2}, \tag{26}$$

where N is the number of input samples, $D = \{D_i\}$ and $\hat{D} = \{\hat{D}_i\}$ are the values of data points before and after smoothing.

3 EXPERIMENTAL RESULTS AND DISCUSSION

We conducted both standard simulation study (all previous aCGH smoothing studies used the same experimental setup) and real data analysis to evaluate the performance of our method in identifying regions of genomic alterations. The experimental results of our method are compared with several other most common used aCGH smoothing methods in literature: Lowess (Beheshti *et al.*, 2003), Quantreg (Eilers *et al.*, 2005; Li *et al.*, 2007), Smoothseg (Huang *et al.*, 2007), SWTi (Wang *et al.*, 2007) (the same as MODWT (L.Hsu *et al.*,

2005)), and DTCWTi-bi (Nguyen *et al.*, 2007). The standard Root Mean Squared Error (RMSE) defined in Eq.(26) and Receiver Operating Characteristic (ROC) curve are used to evaluate the performance of the above six methods.

Willenbrock and Fridlyand (Willenbrock *et al.*, 2005) proposed a standard simulation model to create the synthetic array aCGH data. This model has been widely used to evaluate aCGH data smoothing algorithms. Y. Wang and S. Wang (Wang *et al.*, 2007) improve this model with unequally spaced probes. In our experiments, we first create the synthetic data using the combination of two above methods. After that, we propose a new synthetic data model using real aCGH noise. Although most papers related to aCGH data study assumed the Gaussian noise existing in dataset, some researchers doubted on this noise estimation (Huang *et al.*, 2007). Thus, we improve the synthetic data generation by adding real noise into ground truth copy numbers. Both synthetics aCGH datasets are used in our validation.

3.1 Standard Synthetic Data Generation

In Willenbrock and Fridlyand (Willenbrock *et al.*, 2005) model, a primary tumor dataset of 145 samples is segmented and the probes are equally spaced along the chromosome. But the real aCGH data has randomly space between two probes. More recently Y. Wang and S. Wang (Wang *et al.*, 2007) extended this model by placing unequally spaced probes along chromosome.

The primary tumor data set is segmented using DNA copy number levels from the em-

pirical distribution of segment mean values *smv* as

$$c = \begin{cases} 0 \quad \text{(0 copies)} & : smv < -0.4, \\ 1 \quad \text{(one copy)} & : -0.4 < smv < -0.2, \\ 2 \quad \text{(two copies)} & : -0.2 < smv < 0.2, \\ 3 \quad \text{(three copies)} & : 0.2 < smv < 0.4, \\ 4 \quad \text{(four copies)} & : 0.4 < smv < 0.6, \\ 5 \quad \text{(five copies)} & : smv > 0.6. \end{cases}$$

The synthetic DNA copy number data on a chromosome is generated with Gaussian noise as follows:

1. Determine copy number probability and the distribution of segment length. As suggested in (Willenbrock *et al.*, 2005) and (Wang *et al.*, 2007), the chromosomal segments with DNA copy number $c = 0, 1, 2, 3, 4$ and 5 are generated with probability $0.01, 0.08, 0.81, 0.07, 0.02$ and 0.01. The lengths for segments are picked up randomly from the corresponding empirical length distribution given in (Willenbrock *et al.*, 2005).

2. Compute $log2ratio$. Each sample is a mixture of tumor cells and normal cells. A proportion of tumor cells is P_t, whose value is from a uniform distribution between 0.3 and 0.7. As in paper (Willenbrock *et al.*, 2005), the $log2ratio$ is calculated by

$$log2ratio = \log_2 \left(\frac{cP_t + 2(1 - P_t)}{2} \right), \tag{27}$$

where c is the assigned copy number. The expected $log2ratio$ value is then the latent true signal.

3. Add Gaussian noises. Gaussian noises with zero mean and variance σ_n^2 are added to the latent true signal. Till now, we get the equally spaced CGH signal.

4. Create unequally spaced probes. Because the distances between probe k and probe $k+1$ are randomly, the best way to get these distances is from the UCSF HumArray2 BAC array. Thus, we create a real CGH signal from the equally spaced CGH signal when the unequally spaced probes are placed on the chromosome. Now, we have many artificial chromosomes of length 200 Mbase which are created by many noise levels $\sigma_n = 0.1, 0.125, 0.15, 0.175$, 0.2, 0.225 and 0.25.

3.2 New Synthetic Data Model

In our new synthetic data model, we still follow the four above steps but in the third step, the real noise should be added instead of Gaussian noise. There are many aCGH data source such as (Stanford, 2001), (GBM, 2005), (NCBI, 2008), but only data from (NCBI, 2008) can be used to get real noise. Because the number of probes in (Stanford, 2001) and (GBM, 2005) are not enough. Data from (Stanford, 2001) has hundreds of probes and data from (GBM, 2005) has about several thousand probes. Both of them have not enough probes to estimate the correct distribution of noise. However, the length of data from (NCBI, 2008) is long enough (more than ten thousands of probes). For example, from (NCBI, 2008), chromosome 13 of GSM232967 has 18323 probes. If we use 64 bins, the distributions of noise from the above chromosomes are shown in Fig. 5. Now, it is easy to create arrays with random values under the above distributions. These arrays are added into true signal to create simulated data with real noise. During this step, we have to randomly choose chromosomes that only have the copy two (zero means). There are many chromosomes which can be used to extract

real noise model, *e.g.* chromosome $1, 3, 4, 6, 8, 9, 10, 12, 13, 14, 17, 18, 19, 20$ of GSM232967 and chromosome 18 of GSM232968.

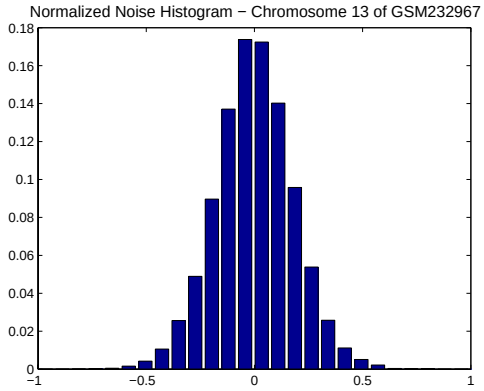


Figure 5: Normalized distribution of real noise from chromosome 13 of GSM232967.

3.3 Performance Evaluation by RMSE

In this section, we will compare the experimental results of Lowess (Lai *et al.*, 2005), Quantreg (Eilers *et al.*, 2005; Li *et al.*, 2007), Smoothseg (Huang *et al.*, 2007), SWTi (Wang *et al.*, 2007), DTCWTi-bi (Nguyen *et al.*, 2007), and our SWPT-LaBi methods. One thousand artificial chromosomes with Gaussian noise in seven different levels $\sigma_n = 0.1,\ 0.125,\ 0.15,\ 0.175,\ 0.2,\ 0.25$ and 0.275 are denoised. Meantime, simulated chromosomes with real noise is also used to test above six methods. Five methods compared to our method are summarized as follows:

- Lowess: This is the locally weighted scatter plot smooth using least squares linear polynomial fitting. It uses a first-degree polynomial instead of second-degree polynomial in Loess. This method is used to compare in (Lai *et al.*, 2005).

- Smoothseg: A smooth segmentation method (Huang *et al.*, 2007) for array CGH data

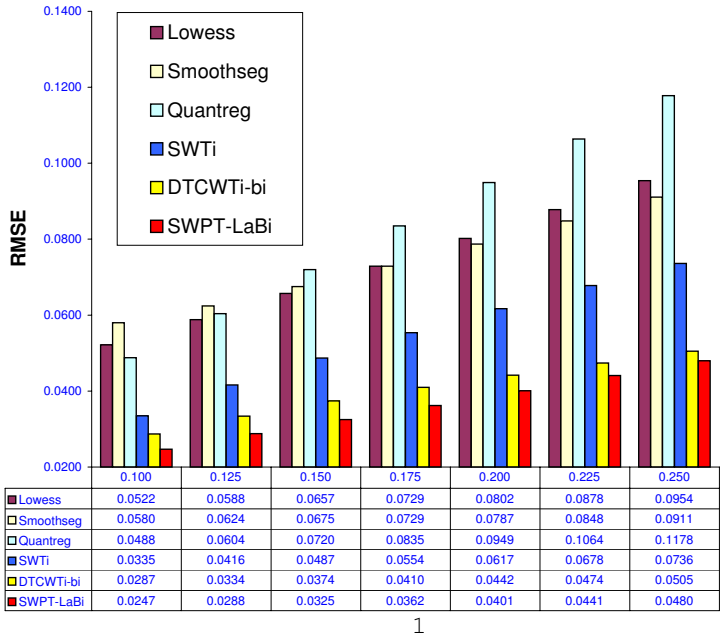
analysis is based on a doubly heavy-tail-random-effect model. This heavy-tailed model on error term deals with outliers in observations. To deal with possible jumps in the copy-number pattern, the i.i.d Cauchy distribution is proposed for modeling the second-order differences of original data. The denoised data is estimated by the iterative weighted least-squares algorithm.

- Quantreg: This is a quantile regression method which has been used by Eilers in (Eilers *et al.*, 2005). The total variation was used as the roughness penalty. In 2007, Li (Li *et al.*, 2007) modified this method by incorporating the physical distance between adjacent clones.

- SWTi: SWTi method comes from paper (Wang *et al.*, 2007). Compared with SWPT-bi, SWTi method has two different steps: 1) the aCGH data which has the unequal distances between two samples is interpolated to reduce the difference of those distances; 2) the array CGH signal is decomposed by the SWT; 3) the term by term thresholding is applied to estimate the SWT coefficients (Wang *et al.*, 2007).

- DTCWTi-bi: This method comes from (Nguyen *et al.*, 2007). It follows five steps: 1) Interpolate the DNA copy number data; 2) Use zero-padding and decompose new data by DTCWT; 3) Calculate the noise variance and the marginal variance; 4) Estimate the coefficients by using a bivariate estimator which shows a relationship of child and parent coefficients; 5) Reconstruct data from the denoised coefficients by taking inverse DTCWT.

The denoising results of all methods are shown in the Fig. 6. The proposed SWPT-LaBi method has a better performance than the others. The SWPT-LaBi outperforms the Lowess



	0.100	0.125	0.150	0.175	0.200	0.225	0.250
Lowess	0.0522	0.0588	0.0657	0.0729	0.0802	0.0878	0.0954
Smoothseg	0.0580	0.0624	0.0675	0.0729	0.0787	0.0848	0.0911
Quantreg	0.0488	0.0604	0.0720	0.0835	0.0949	0.1064	0.1178
SWTi	0.0335	0.0416	0.0487	0.0554	0.0617	0.0678	0.0736
DTCWTi-bi	0.0287	0.0334	0.0374	0.0410	0.0442	0.0474	0.0505
SWPT-LaBi	0.0247	0.0288	0.0325	0.0362	0.0401	0.0441	0.0480

Figure 6: Comparison of average RMSEs obtained from the 1,000 artificial chromosomes with each of the 7 noise levels (Gaussian noise)

by $49.7\% - 52.7\%$, the Quantreg by $47.3\% - 57.4\%$, the Smoothseg by $49.4\% - 59.3\%$, the SWTi by $26.3\% - 35\%$ and the DTCWTi-bi by $5\% - 13.9\%$ in terms of the root mean squared errors. For all noise levels, the SWPT-LaBi consistently achieves much better results than others.

In the experiments using real noise synthetic data, 15 chromosomes are used to create real noise for synthetic data as description in section 3.1. We also run 1000 chromosomes with six above methods. The Fig. 7 shows that the RMSE of the SWPT-LaBi is 0.0453 and is the smallest one if compared to the Lowess (0.0771), the Smoothseg (0.0813), the Quantreg (0.0940), the SWTi (0.0700), and the DTCWTi-bi (0.0492). Our proposed method outperforms all previous methods between 7.9% and 51.8% on the synthetic data with real noise.

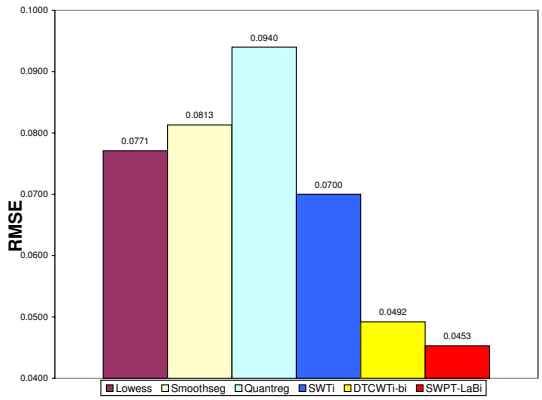


Figure 7: Comparison of average RMSEs obtained from the 1,000 artificial chromosomes with real noise

3.4 Performance Evaluation by ROC Curve

Paper (Lai *et al.*, 2005) introduced another method to evaluate aCGH smoothing algorithms by ROC curve. Several hundred artificial chromosomes, consisting of 100 probes and with the square-wave signal in the center of the chromosome, are created from four templates. In 2007, Huang *et al.* (Huang *et al.*, 2007) modified this setting to make the problem harder. The modification decreased the width of the center square-wave and increased the noise level. However people usually want to test the performance of methods not only at the middle of signal but also the border of signal. Therefore, we keep four templates with the aberration widths of 5, 20, 30 and 40. We add three more templates (one or two aberrations) with the aberration widths of 20, 10 and 5 on the border.

From seven genomic templates with one or two aberrations, 270 samples are generated with unequal space probes. The ROC profiles of SWTi, DTCWTi-bi, Lowess, Quantreg, Smoothseg, and our SWPT-LaBi methods are calculated. Fig. 8 illustrates the ROC curves

with different noise levels: $\sigma = 0.125, 0.175, 0.2$. The TPR is defined as the number of probes inside the aberration whose absolute values are above the threshold level divides by the number of probes in the aberration. The FPR is defined as the number of probes outside the aberration or the number of probes inside the copy two region whose absolute value are above the threshold level divides by the total number of probes outside the aberration. The threshold level is changed from 0 to 1.

In Fig 8 (a,b,c), our SWPT-LaBi method clearly performs better than other methods. If we just compare some methods in time domain such as Lowess and Quantreg , the Lowess looks better. This result also agrees with the experimental results in paper (Lai *et al.*, 2005). In low noise level ($\sigma = 0.125$), in Fig 8 (a), most methods operate well except Smoothseg. If we increase noise, the Quantreg gets worse in Fig 8 (b,c). Of course, with Gaussian noise, the Smoothseg always gets worse than the others because it is designed to operate with the student't noise. From three above figures, our method always give out the best results.

We also get real noise from the chromosome 13 of GSM232967 and create 270 simulated aCGHs with real noise from seven genomic templates with one or two aberrations. Fig 8 (d) shows the ROC curve results of 270 above simulated aCGHs with real noise. The performance of Smoothseg becomes better but still gets worse than our method. In this case, the SWPT-LaBi is the best in the low FDR area that is more meaningful in practical aCGH applications. In summary, our proposed method is the best one in ROC curve comparison.

3.5 Demonstration Using Real Data

We discuss two more examples of smoothing results with simulated data (Gaussian noise) and real aCGH data in Fig. 9 and Fig. 10. In these figures, Lowess, Quantreg, Smoothseg,

SWTi, DTCWTi-bi, and SWPT-LaBi are still used to smooth aCGH signal. The black solid lines represent the latent true signals, the blue points stand for the noisy DNA copy data ($\log_2 ratio$) at the probe loci and the red lines correspond to the denoised data. We should note that the line connecting the denoised data points is only for visualization purpose.

Fig. 9 shows the smoothing result of synthetic data with Gaussian noise. The signal of that figure is separated to four segments. To easily recognize these segments, we name them as A, B, C, and D shown in Fig. 9 (f). The $\log_2 ratio$ values of the latent true signal are 0.3598 at the segment A (the copy three $c = 3$), zero at the segments B and D (the copy two $c = 2$), -0.4624 at the segment C (the copy zero $c = 0$). If the RMSE values are calculated for six denoised signals, the SWPT-LaBi has the smallest RMSE value (0.0322). The RMSE values of Lowess, Quantreg, Smoothseg, SWTi, and DTCWTi-bi are 0.0503, 0.0532, 0.0462, 0.0444, and 0.0377, respectively. In aCGH data, the segment A and C which may be causal to cancer are important. Therefore, we consider the segment A and C as the abnormal segments, the segment B and D as the normal segments. We calculate the RMSE values of the abnormal segments and the normal segments separately. For the abnormal segments, Lowess, Quantreg, Smoothseg, SWTi, DTCWTi-bi, and SWPT-LaBi give out 0.0522, 0.0626, 0.0679, 0.0658, 0.0758, and 0.0520 in term of the RMSE values respectively. For the normal segments, the RMSE values are 0.0717 (Lowess), 0.0741 (Quantreg), 0.0608 (Smoothseg), 0.0583 (SWTi), 0.0416 (DTCWTi-bi), and 0.0408 (our method). In both cases, our method has the smallest RMSE value. Therefore, the SWPT-LaBi method gives out the best result for both the abnormal segments and normal segments. In visible, the denoised signals using the SWPT-LaBi also tracks the true signal better than the others, and the denoised results of Lowess, Quantreg, Smoothseg, SWTi, and DTCWTi-bi methods

have some ripples.

We also perform six methods on the real aCGH data that is the BAC array data on 15 fibroblast cell lines (Snijders *et al.*, 2001; L.Hsu *et al.*, 2005). This dataset comes from Stanford University, which can be freely downloaded at (Stanford, 2001). Because the true copy number changes are already known for these cell lines by manual inspection of profiles, we choose these data as a proof of principles. We pick up the chromosome 14 of GM01750 from these data and perform our method and other algorithms for denoising.

Fig. 10 illustrates six sub-figures for six methods. From Fig. 10, we know there are only the copy four and the copy two in this aCGH data. At the copy two, the denoised signals using the DTCWTi-bi and SWPT-LaBi look smoother than results of other methods. At the copy four, the SWPT-LaBi in Fig. 10 (f) still gives more exact segmentation than Fig. 10 (c) and (e). In summary, the performance of the SWPT-LaBi based denoised signal is the best one in Fig. 10.

4 CONCLUSION

In this paper, we propose the dependent Laplacian bivariate shrinkage estimator to improve the SWPT method in aCGH data denoising study. In our experiments, the denoising results of our SWPT-LaBi method are much better than the previous methods in terms of the root mean squared error measurement (improve $5\% - 59.3\%$) and the ROC curve at different Gaussian noise levels. Furthermore, we also propose to use real noise to improve the traditional aCGH synthetic data generation. Since our new synthetic data generation model is a better approximation of real aCGH data, it can more accurately evaluate the aCGH smoothing algorithms. This new synthetic aCGH with real noise is also exploited in evaluation, and

our method still outperforms others. Meantime, we also use the real aCGH data to demonstrate our SWPT-LaBi approach is better than other most common used smoothing methods.

5 ACKNOWLEDGMENTS

The research described in the paper has been supported by Dr. Heng Huang's start up funding from University of Texas at Arlington.

6 DISCLOSURE STATEMENT

No competing financial interests exist.

References

Beheshti, B. and Braude, I., Marrano, P. and Thorner, P. and Zielenska, M. and Squire, J. (2003) Chromosomal localization of dna amplifications in neuroblastoma tumors using cdna microarray comparative genomic hybridization. *Neoplasia,* **5**, 53–62.

Bilke, S. and Chen, Q. R. and Whiteford, C. C. and Khan, J. (2005) Detection of low level genomic alterations by comparative genomic hybridization based on cdna microarrays. *Bioinformatics,* **21** (7), 1138–1145.

Brennan, C. and Zhang, Y. and Leo, C. and Fenga, B. and *et al.* (2004) High-resolution

global profiling of genomic alterations with long oligonucleotide microarray. *Cancer Res,* **64**, 4744–4748.

Bruce, A. and Gao, H. (1996) Understanding waveshrink: variance and bias estimation. *Biometrika,* **83**, 727–745.

Chang, S. and Yu, B. and Vetterli, M. (Sept.2000) Adaptive wavelet thresholding for image denoising and compression. *IEEE Trans.Image processing,* **9**, 1532–1546.

Coifman, R. and Donoho, D. (1995) Translation-invariant de-noising. *Wavelets and Statistics,* 103 of Lecture Notes in Statistics, 125–150.

Coifman, R. & Wickerhauser, M. (1992) Entropy-based algorithms for best basis selection. *IEEE Trans. on Inf. Theory,* **38**, 713–718.

Donoho, D. (1995) De-noising by soft-thresholding. *IEEE Trans. on Inf. Theory,* **41** (3), 613–627.

Donoho, D. and Johnstone, I. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika,* **81**, 425–455.

Eilers, P. and de Menezes, R. (2005) Quantile smoothing of array cgh data. *Bioinformatics,* **21**, 1146–1153.

Fridkyand, J. and *et al.* (2004) Hidden markov models approach to the analysis of arrray cgh data. *J.Multivariate Anal.,* **90**, 132–153.

Computational Genomics (2005). Supplement material of lai w, 2005. http://compbio.med.harvard.edu/Supplements/Bioinformatics05b.html

Huang, J. and *et al.* (2007) Robust smooth segmentation approach for array cgh data analysis. *Bioinformatics,* **23**, 2463–2469.

Hupe, P. and Stransky, N. and Thiery, J. P. and Radvanyi, F. and Barillot, E. (2004) Analysis of array cgh data: from signal ratio to gain and loss of dna regions. *Biostatistics,* **20**, 3413–3422.

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

Jong, K. and Marchiori, E. and Meijer, G. and Vaart, A. and Ylstra, B. (2004) Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics,* **20**, 3636–3637.

Lai, W. and Johnson, M. and Kucherlapati, R. and Park, P. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data. *Bioinformatics,* **21**, 3763–3770.

Hsu, L. and *et al.* (2005) Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics(Oxford,England),* **6** (2), 211–226.

Li, Y. & Zhu, J. (2007) Analysis of array cgh data for cancer studies using fused quantile regression. *Bioinformatics,* **23**, 2470–2476.

Lingjaerde, O. & *et al.* (2005) Cgh-exploxer: a program for analysis of array-cgh data. *Bioinformatics,* **21**, 821–822.

Myers, C. and Dunham, M. and Kung, S. & Troyanskaya, O. (2004) Accurate detection of

aneuploidies in array cgh and gene expression microarray data. *Bioinformatics,* **20**, 3533–3543.

Nguyen, N. and Huang, H. and Oraintara, S. & Vo, A. (2007) A new smoothing model for analyzing array cgh data. *IEEE BIBE,* .

National Center for Biotechnology Information (2008). Gene expression omnibus - GEO. http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE9220

Olshen, A. and Venkatraman, E. and Lucito, R. and Wigler, M. (2004) Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics,* **5**, 557–572.

Picard, F. and *et al.* (2005) A statistical approach for array cgh data analysis. *BMC Bioinformatics,* **27**.

Pinkel, D. and Segraves, R. and Sudar, D. and Clark, S. and *et al.* (1998) High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nat Genet,* **20**, 207–211.

Pollack, J. and Perou, C. and Alizadeh, A. and Eisen, M. and *et al.* (1999) Genome-wide analysis of dna copy-number changes using cdna microarrays. *Nat Genet,* **23**, 41–46.

Sendur, L. and Selesnick, I. (November 2002) Bivariate shinkage function for wavelet-based denoising exploiting interscale dependency. *IEEE Transaction on Signal Processing,* **50** (11).

Snijders, A. M. and Nowak, N. and Segraves, R. and Blackwood, S. and *et al.* (2001) As-

sembly of microarrays for genome-wide measurement of dna copy number. *Nat Genet,* **29** (3), 263–264.

Strang, G. and Nguyen, T. (1996) *Wavelets and filter banks*. Wellesley-Cambridge Press.

Stanford University (2001). Assembly of microarrays for genome-wide measurement of dna copy number. http://www.nature.com/ng/journal/v29/n3/suppinfo/ng754_S1.html

Wang, P. and *et al.* (2005) A method for callling gains and losses in arrray cgh data. *Bioinformatics,* **6**, 45–58.

Wang, Y. and Wang, S. (2007) A novel stationary wavelet denoising algorithm for array-based dna copy number data. *International Journal of Bioinformatics Research and Applications,* **3** (2), 206 – 222.

Willenbrock, H. and Fridlyand, J. (2005) A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics,* **21** (22), 4084–4091.

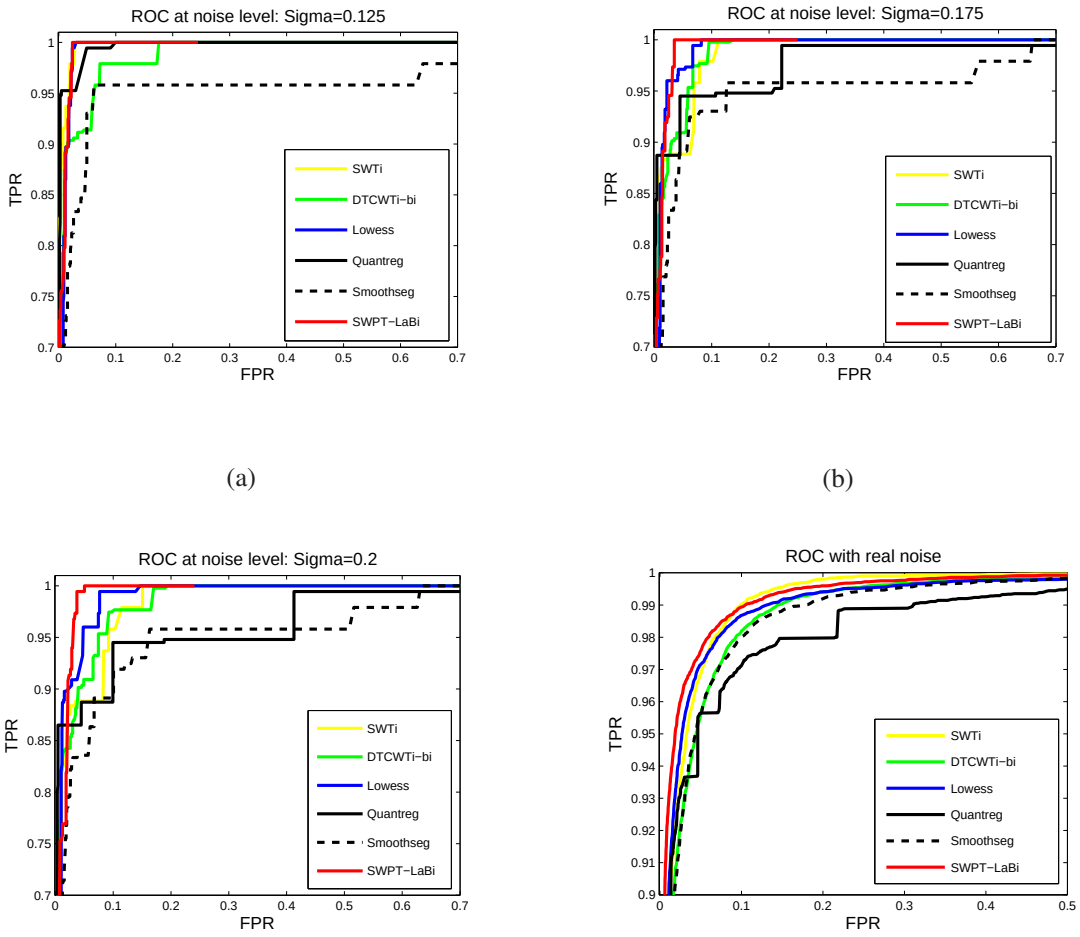


Figure 8: Receiver operating characteristic (ROC) curves obtained from the 270 artificial chromosomes (generated from 7 genomic templates) with each of the different noise levels using the SWPT-LaBi and other most common used CGH algorithms such as SWTi, DTCWTi-bi, Lowess, Quantreg, and Smoothseg.

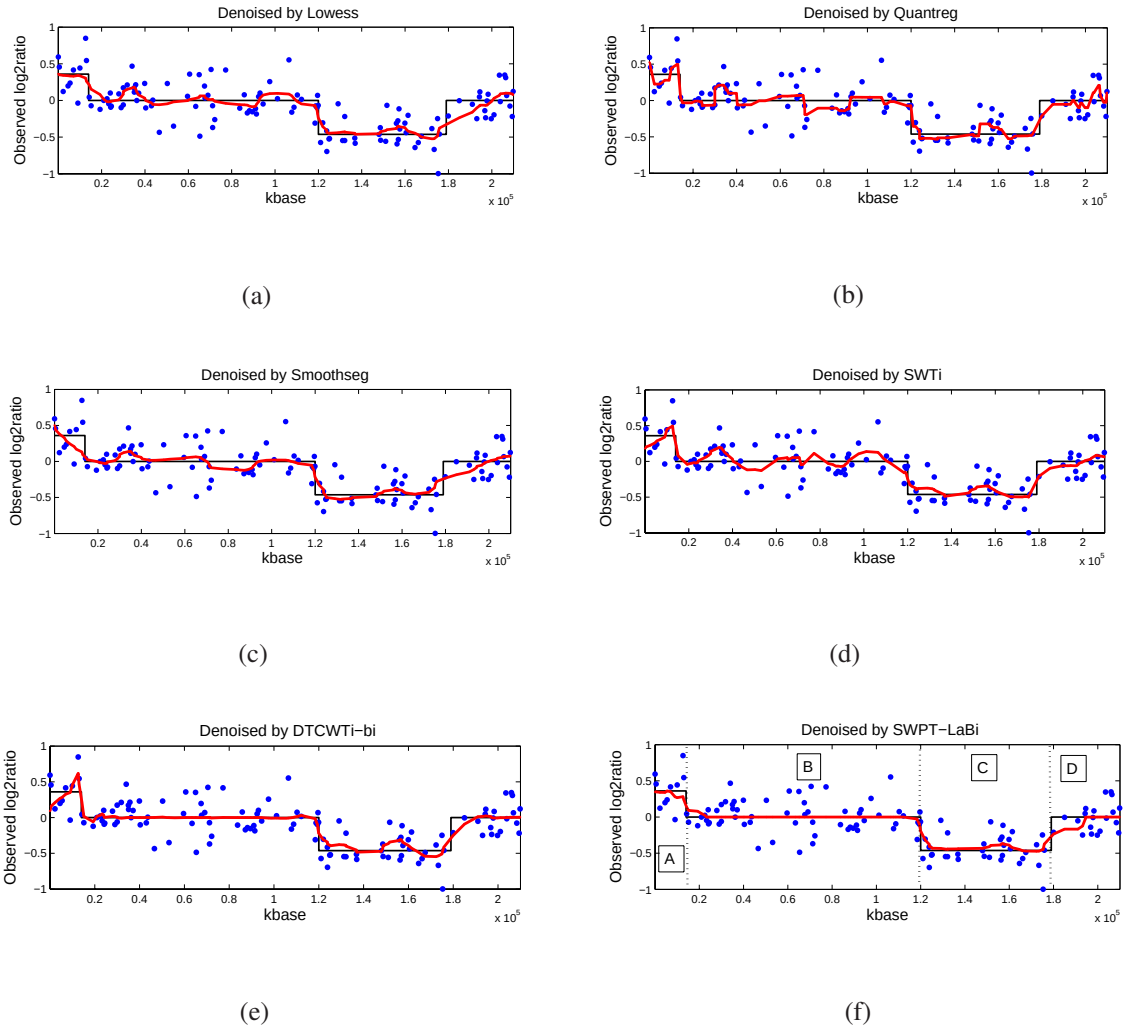
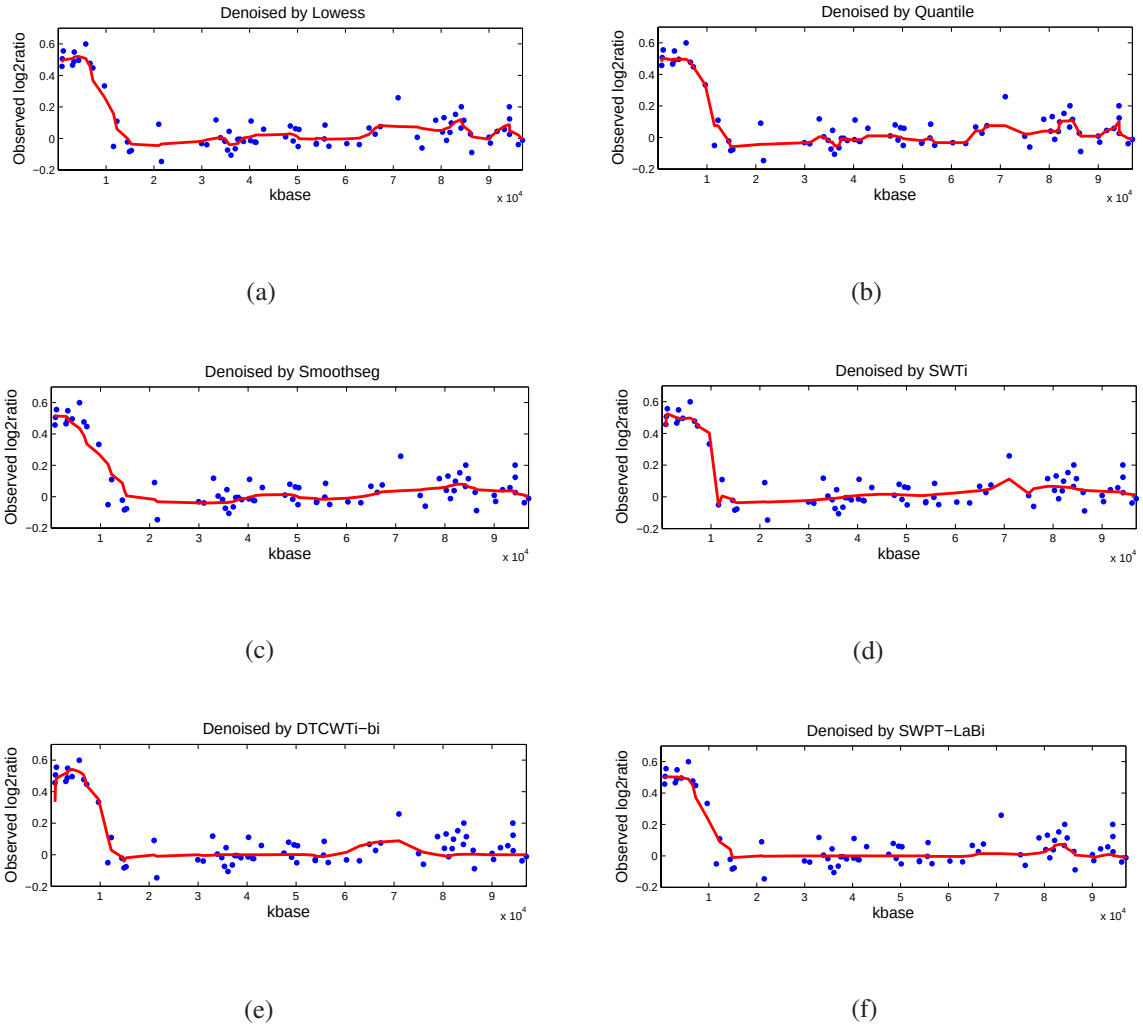


Figure 9: Example of smoothing results at the noise level of $\sigma = 0.2$ of six methods



(a) (b)

(c) (d)

(e) (f)

Figure 10: The wavelet denoising results of six methods on chromosome 14 in the real signal GM01750