CGH Data Modeling and Smoothing in Stationary Wavelet Packet Transform Domain.

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

[1]Department of Computer Science and Engineering, University of Texas at Arlington, Texas, USA
[2]Department of Electrical Engineering, University of Texas at Arlington, Texas, USA

Email: Heng Huang*- heng@uta.edu; Nha Nguyen - nhn3175@exchange.uta.edu; Soontorn Oraintara - oraintar@uta.edu; An Vo - vpnan@gauss.uta.edu;

*Corresponding author

Abstract

Background: Array-based comparative genomic hybridization (array CGH) is a highly efficient technique, allowing the simultaneous measurement of genomic DNA copy number at hundreds or thousands of loci and the reliable detection of local one-copy-level variations. Characterization of these DNA copy number changes is important for both the basic understanding of cancer and its diagnosis. In order to develop effective methods to identify aberration regions from array CGH data, many recent research work focus on both smoothing-based and segmentation-based data processing. In this paper, we propose stationary packet wavelet transform based approach to smooth array CGH data. Our purpose is to remove CGH noise in whole frequency while keeping true signal by using bivariate model.

Results: In both synthetic and real CGH data, Stationary Wavelet Packet Transform (SWPT) is the best wavelet transform to analyze CGH signal in whole frequency. We also introduce a new bivariate shrinkage model which shows the relationship of CGH noisy coefficients of two scales in SWPT. Before smoothing, the symmetric extension is considered as a preprocessing step to save information at the border.

Conclusions: We have designed the SWTP and the SWPT-Bi which are using the stationary wavelet packet transform with the hard thresholding and the new bivariate shrinkage estimator respectively to smooth the array CGH data. We demonstrate the effectiveness of our approach through theoretical and experimental exploration of a set of array CGH data, including both synthetic data and real data. The comparison results show that our method outperforms the previous approaches.

Background

Gene amplifications or deletions frequently contribute to tumorigenesis. When part or all of a chromosome is amplified or deleted, a change in DNA copy number results. Characterization of these DNA copy number changes is important for both the basic understanding of cancer and its diagnosis. Cancer researchers currently use array comparative genomic hybridization (array CGH) to identify sets of copy number changes associated with the particular cancer or its congenital and developmental disorders. In array CGH, because the clones contain

sequences information directly connecting with the genome database, array CGH offers rapid genome-wide analysis at high resolution and the information it provides is directly linked to the physical and genetic maps of the human genome. Bacterial Artificial Chromosomes (BAC) based CGH arrays were amongst the first genomic arrays to be introduced [1] and are routinely used to detect single copy changes in the genome, owing to their high resolution in the order of 1 Mb [1, 2]. More recently Oligonucleotide aCGH [3, 4] was developed to allow flexibility in probe design, greater coverage, and much higher resolution in the order of 35-100 Kb [5].

In order to develop effective methods to identify aberration regions from array CGH data, the previous research works focus on both smoothing-based [5–9] and segmentation-based data processing [10–14]. The array CGH is very noisy. For example, in cDNA array CGH data, the signal to noise ratio is often approximately 1 (0 dB) [15]. Research in this area has been active in the last few years. Beheshti *et al.* proposed to use the robust locally weighted regression and smoothing scatterplots (lowess) method in [6]. Eilers and Menezes [7] perform a quantile smoothing method based on the minimization of the sum of absolute errors to create sharper boundaries between segments. Hsu *et al.* [8] investigated the usage of maximal overlap discrete wavelet transform (MODWT) in the analysis of array CGH data. They have shown translation invariant wavelets are promising methods for array CGH data smoothing and also observed that the denoising techniques may miss singleton clones that have small changes but somehow are consistent across tumors. In 2005, Lai [16] compared 11 different algorithms for analyzing array CGH data. Many smoothing and estimation methods were included in [16] such as CGHseg (2005) [17], Quantreg (2005) [7], CLAC (2005) [18], GLAD (2004) [11], CBS (2004) [14], HMM (2004) [19], MODWT (2005) [8], Lowess [6], ChARM (2004) [13], GA (2004) [12], ACE (2005) [20]. Lai concluded that Wavelet, Quantreg and Lowess method gave better detection results (higher true position rate and lower false position rate) than other methods. So, the wavelet based smooth was considered as the promising approach. More recently Y. Wang and S. Wang [5] extended the stationary wavelet (SWT) denoising and regression for nonequispaced data, because the physical distance between adjacent probes along a chromosome are not uniform, even vary drastically. However, if a signal is decomposed by using SWT or MODWT, we get unequal sub-bands and a long high frequency sub-bands. Because true CGH signals include many step functions, they contain important information in high frequency. If long high frequency is used to remove noise, maybe, some high frequency true information of CGH will be loosen.

In this paper, we propose to use the Stationary Wavelet Packet Transform (SWPT) to denoise the array CGH data. Because, in SWPT, all sub-bands are also shift invariant, each sub-band provides a shiftable description of signal in a specific scale as the same SWT or MODWT. SWPT analyzes signal to many equally frequency sub-bands. So, information in both of low and high frequency sub-band are saved. Moreover, new bivariate shrinkage function is used in SWPT instead of universal thresholding at the first time, soft thresholding [21–23] and BayesShrink [24]. We demonstrate the effectiveness of our approach through theoretical and experimental exploration of a set of array CGH data, including both synthetic data and real array CGH data. The comparison results show that our method outperforms the previous approaches about $6.4\% - 57.9\%$. Let see detail results in next section.

Results and Discussion

In this section, results of our proposed methods such as the SWPT and the SWPT-Bi will be compared to the other efficient smooth methods such as the Lowess [16], the Quantreg [7, 25], the SWTi [5], the DTCWTi-bi [26]. In our experiments, the artificial chromosomes are generated using the methods proposed in [27] and [5]. Finally, real data examples are showed to make sure that our methods are still better the others.

Synthetic data

First, we describe how to create synthesis data as follow.

Artificial Chromosome Generation

Willenbrock and Fridlyand [27] proposed a simulation model to create the synthetic array CGH data with equally spaced along the chromosome. More recently Y. Wang and S. Wang [5] extended this model by placing unequally spaced probes along

chromosome. As suggested in [27] and [5], 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 [27]. 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 [27], the $log2ratio$ is calculated by

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

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

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. Because the distances between two probes 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 $\sigma_n = 0.1, 0.125, 0.15, , 0.175, 0.2, 0.25$ and 0.275.

Comparison by RMSE

In this section, we will present the results when applying six methods such as the Lowess [16], the Quantreg [7,25], the SWTi [5], the DTCWTi-bi [26] and our methods the SWPT and the SWPT-Bi. One thousand artificial chromosomes with seven different noise levels $\sigma_n = 0.1, 0.125, 0.15, , 0.175, 0.2, 0.25$ and 0.275 are denoised.

The denoising results of all methods are shown in the Figure 1. We can see that the proposed SWPT and SWPT-Bi methods yield the better performance than the others. The SWPT and SWPT-Bi outperform the Lowess by $43.4\% - 55\%$ and $48.4\% - 54.2\%$ respectably, the Quantreg by $50.3\% - 53.7\%$ and $49.5\% - 57.9\%$ respectably and the SWTi by $27.5\% - 31.5\%$ and $26.8\% - 35.3\%$ in terms of the root mean squared errors (RMSEs). If compared the DTCWTi-bi, the SWPT-Bi gets better by $6.4\% - 17.9\%$ for seven noise level and the SWPT performs better by $1\% - 19.2\%$ for six noise levels $(0.1 - 0.225)$. For all noise levels, the SWPT-Bi consistently achieves much better results than the others.

Some examples of wavelet denoising results by using the Lowess, the Quantreg, the SWTi, the DTCWTi-bi, the SWPT and the SWPT-Bi methods are shown in Figure 2 at the noise level of $\sigma = 0.2$. In those Figures, the black solid lines represent the latent true signals, the blue points stand for the noisy DNA copy data log_2ratio 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.

At the copy three $c = 3$ (from 1 $kbase$ to 1.4x10^4 $kbase$) as shown in Figure 2, the log_2ratio value of the latent true signal is 0.3598, but these values of the Quantreg, the SWTi and the DTCWTi-bi based denoised signal in Figure 2 are from 0.2262 to 0.4966, from 0.1774 to 0.3828 and from 0.09233 to 0.6182 respectably. These values can cause a mistake when we segment the DNA copy number data. However, the denoised data using the Lowess, the SWPT and the SWPT-Bi will be segmented correctly as the copy three (from 0.2 to 0.4) because the log_2ratio values are from 0.2129 to 0.3619, 0.2794 to 0.3649 and from 0.2565 to 0.3964. At the copy two $c = 2$ (from 1.4x10^4 $kbase$ to 1.2x10^5 $kbase$), the denoised data in the second sub-figure (denoised by Quantreg) of Figure 2 has an amplitude of 0.2262 which will make an error in segmentation process, while the denoised data in other sub-figures of Figure 2 will give a correct segmentation. In this copy, the denoised signals using DTCWTi-bi, the SWPT and SWPT-Bi are approximately the latent true signals, while the denoised data using the Lowess, the Quantreg and the SWTi have many ripples. At the copy zero $c = 0$ (from 1.2x10^5 $kbase$ to 1.79x10^5 $kbase$), if we use TPR (true position rate = number of denoised probes belove -0.4 / total of true probes), the Lowess, the Quantreg and our methods gave a ratio of 22 over 34 instead of 17/34 and 14/34 of the SWTi and DTCWTi-bi. However, the denoised signals of the Lowess, the SWPT and the SWPT-Bi look better than of the Quantreg. At the copy two $c = 2$ (form 1.79x10^5 $kbase$ to the end of the chromosome), the fourth, fifth and sixth sub-figures's signal (denoised by DTCWTi-bi, SWPT the and SWPT-Bi) of Figure 2 look smoother than the others. Furthermore, the denoised signals at the first sub-figure (the Lowess's) and the second sub-figure (the Quantreg's) may cause error when segmentation because denoised signals change from -0.3133 to 0.101 (Lowess) and from -0.2119 to 0.2084 (Quantreg).

From above results, we can see that our proposed SWPT and SWPT-Bi methods with the stationary wavelet packet transform are better than the others. Now, real data will be used to test five smoothing methods as follow.

Real Data Examples

In this paper, the BAC array data on 15 fibroblast cell lines [8, 28] has been used to show that denoising by the SWPT and the SWPT-Bi are better than by the others such as the Lowess, the Quantreg, the SWTi and the DTCWTi-bi. This data set is from Stanford University, which can be freely downloaded at [29]. Because the true copy number changes are known for these cell lines, we choose these data as a proof of principles. We pick up the chromosome 1 of GM13330 from these data and apply six algorithms for denoising. In Figure 3, the number copies are two and four. At the copy two (from 1 *kbase* to 1.56x10^5 *kbase*), the SWPT and SWPT-Bi based smoothed signals are smoother than the others. With the copy four, from 1.56x10^5 *kbase* to the end of this chromosome, the performance of the Lowess, the SWPT and the SWPT-Bi based denoising methods are the better than of the Quantreg, the SWTi and the DTCWTi-bi. From the above figures, we can believe that our methods perform better than the oethers in denoising of real CGH data.

Conclusions

In this paper, we explored the stationary wavelet packet transform method with the new bivariate shrinkage estimator in array CGH data denoising study. In the simulation situations, the denoising results from the SWPT and the SWPT-Bi are much better (improve $6.4\% - 57.9\%$) than the previous methods in terms of the root mean squared error measurement at different noise levels. Furthermore, we also demonstrate our method by using the real array CGH data. In our future work, we will develop a smoothing and segmentation combinatorial algorithm to improve the aberration regions identification from DNA copy number data.

Methods

Our methods named the SWPT (SWPT and universal shrinkage function) and the SWPT-Bi (SWPT and bivariate shrinkage function) will be introduced. First, let review wavelet transform and see how SWPT operates.

Wavelet Methods

We will provide a brief review of wavelet transforms which were used for array CGH data smoothing and is used by this paper. We should note that the simple wavelet transform will be introduced firstly and the SWPT will be mentioned finally.

Discrete Wavelet Transform

The discrete wavelet transform (DWT), showed in Figure 4, based on the octave band tree structure, can be viewed as the multiresolution decomposition of a signal. It takes a length N sequence, and generates an output sequence of length N using a set of lowpass and highpass fiters followed by a decimator. It has $N/2$ values at the highest resolution, $N/4$ values at the next resolution, and $N/2^L$ at the level L. Because of decimation, the DWT is a critically sampled decomposition. However, the drawback of DWT is the shift variant property. In signal denoising, the DWT creates artifacts around the discontinuities of the input signal [30]. These artifacts degrade the performance of the threshold-based denoising algorithm.

Stationary Wavelet Transform

The stationary wavelet transform (SWT) [30], showed in Figure 4, is similar to the DWT except that it does not employ a decimator after filtering, and each level's filters are upsampled versions of the previous ones. The SWT is also known as the shift invariant DWT. The absence of a decimator leads to a full rate decomposition. Each subband contains the same number of samples as the input. So for a decomposition of L levels, there is a redundant ratio of $(L + 1) : 1$. However, the shift invariant property of the SWT makes it 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 [30].

Dual-tree Complex Wavelet Transform

A dual-tree structure that produces a dyadic complex DWT, showed in Figure 4, is proposed by Kingsbury [31, 32]. In the case of 1-D signals, the structure consists of two binary trees of multiresolution decomposition of the same signal. It is therefore an overcomplete representation with a redundant ratio of 2 : 1. In the two trees, the filters are designed in such a way that the aliasing in one branch in the first tree is approximately canceled by the corresponding branch in the second tree. The relation between the wavelet filters of the two trees yields shift-invariant property [31].

The analysis FB for the DTCWT is an iterative multi-scale FB. Each resolution level consists of a pair of two-channel FBs. The purpose of the dual-tree CWT is to provide a shiftable and scalable multiresolution decomposition. The input signal is passed through the first level of a multiresolution FB. The low frequency component, after decimation by 2, is fed into the second level decomposition for the second resolution. The outputs of the two trees are the real and imaginary parts of complex-valued subbands. To reconstruct the signal, the real part and imaginary part are inverted to obtain two real signals, respectively. These two real signals are then averaged to obtained the final output. For more details of the construction of the dual-tree, the reader is referred to [33].

Discrete Wavelet Packet Transform

We continue with an another basic othornormal wavelet transform. Discrete wavelet packet transform (DWPT), which can be readily computed by using a very simple adjustment of the pyramid algorithm for DWT, will be mentioned. All of DWPT scales are performed at the same level j. The jth level DWPT decomposes the frequency interval $[0, 1/2]$ into 2^j equal and individual intervals, each of which has $N/2^j$ values if taking a length N sequence. DWPT still keeps a shift variant property.

Stationary Wavelet Packet Transform

Stationary Wavelet Packet Transform (SWPT), showed in Figure 4, still keeps two important properties of SWT such as shift invariance and redundancy. In the SWPT, both scaling and wavelet coefficients are subjected to the high-pass and low-pass filter when computing the next level coefficients. At the given level L, there are 2^L scales with the same length as the input signal's. The redundant ratio is $(2^L) : 1$ for a decomposition of L levels. SWPT is really combination of SWT and DWPT. So, it is very useful in denoising of DCN data. After wavelet transform, reader should be introduced a new shrinkage function to remove noise of CGH data in SWPT domain as follow.

New Bivariate Shrinkage Function for SWPT-Based Denoising.

In this sub-section, the bivariate shrinkage function which describes the relationship of child and parent (Figure 4) coefficients will be reminded. Because SWPT, which decomposes a signal into many subbands at the same scale, just has child and cousin coefficients (Figure 4) at the same level, new bivariate shrinkage function will be developed to exploit the relationship between child and cousin coefficients.

A simple denoising algorithm via wavelet transform consists of three steps: decompose the noisy signal by wavelet transform, denoise the noisy wavelet coefficients according to some rules and take the inverse wavelet transform from the denoised coefficients. To estimate wavelet coefficients, some of the most well-known rules are universal thresholding, soft thresholding [21–23] and BayesShrink [24]. In these algorithms, the authors assumed that wavelet coefficients are independent. Sendur and Selesnick [34] has recently exploited the dependency between coefficients and proposed a non-Gaussian bivariate pdf for the child coefficient w_c and its parent w_p. Nguyen *et el* [26,35] applied that function to recover CGH data successfully and got some promising results.

Now basing on the idea in [34], we try to discover the connection of child and cousin coefficients in SWPT with CGH data. We assume that we get the DNA copy number data Y which 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.

$$Y = D + n. \tag{2}$$

After decomposing the data Y by the SWPT, we get the coefficients $\mathbf{y}_k$. In the wavelet domain, those coefficients can be formulated as

$$\begin{aligned} y_1 &= w_1 + n_1, \\ y_2 &= w_2 + n_2, \end{aligned} \tag{3}$$

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 continue being 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:

$$\begin{array}{rcl} y_3 & = & w_3 + n_3, \\ y_3[n] & = & h[n] * y_2[n] \\ & = & \sum_{k=1}^{N}(h[n-k].y_2[k]), \end{array} \quad (4)$$

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

$$\mathbf{y} = \mathbf{w} + \mathbf{n}, \quad (5)$$

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

$$p_{\mathbf{n}}(\mathbf{n}) = \frac{1}{2\pi {\sigma_n}^2} \exp(-\frac{n_1^2 + n_3^2}{2\sigma_n^2}). \quad (6)$$

The standard MAP estimator [34] of $\mathbf{w}$ from $\mathbf{y}$ is followed as

$$\widehat{\mathbf{w}}(\mathbf{y}) = \arg\max_{\mathbf{w}}[\log(p_{\mathbf{n}}(\mathbf{y}\text{-}\mathbf{w})) + \log(p_{\mathbf{w}}(\mathbf{w}))]. \quad (7)$$

As [34], we propose a non-gaussian bivariate pdf for w_1 and w_3 as

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

With this pdf, two variables w_1 and w_3 are really dependent. Let us define:

$$\begin{array}{rcl} f(w) & = & \log(P_w(w)) \\ & = & \log(\frac{k}{2\pi\sigma^2}) - \frac{\sqrt{2k}}{\sigma}\sqrt{|w_1|^2 + |w_3|^2}. \end{array} \quad (9)$$

By using (6), (7) becomes:

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

Solving above equation is the same solving of two following equations:

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

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

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)

$$f_{w_1}(\widehat{w}) = \frac{\sqrt{2k}w_1}{\sigma\sqrt{|w_1|^2 + |w_3|^2}}. \quad (13)$$

$$f_{w_3}(\widehat{w}) = \frac{\sqrt{2k}w_3}{\sigma\sqrt{|w_1|^2 + |w_3|^2}}. \quad (14)$$

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

$$\widehat{w_1}.(1 + \frac{\sqrt{2k}\sigma_n^2}{\sigma r}) = y_1, \quad \widehat{w_3}.(1 + \frac{\sqrt{2k}\sigma_n^2}{\sigma r}) = y_3, \quad (15)$$

where $r = \sqrt{|\widehat{w_1}|^2 + |\widehat{w_3}|^2}$. Drawing r from (15):

$$r = (\sqrt{|y_1|^2 + |y_3|^2} - \frac{\sqrt{2k}{\sigma_n}^2}{\sigma})_+. \quad (16)$$

If replacing r by (16) into (15), the MAP estimator can be written as:

$$\hat{w_1} = \frac{(\sqrt{|y_1|^2 + |y_3|^2} - \frac{\sqrt{2k}{\sigma_n}^2}{\sigma})_+}{\sqrt{|y_1|^2 + |y_3|^2}}.y_1, \quad (17)$$

where $(u)_+$ is defined by

$$(u)_+ = \begin{cases} 0 & \text{if } u < 0, \\ u & \text{otherwise.} \end{cases} \quad (18)$$

Replacing y_3 from (4) to (17), we can rewrite the MAP estimator as

$$\hat{w_1} = \frac{y_1(\sqrt{|y_1|^2 + |\sum_{k=1}^{N}(h[n-k]y_2[k])|^2} - \frac{\sqrt{2k}{\sigma_n}^2}{\sigma})_+}{\sqrt{|y_1|^2 + |\sum_{k=1}^{N}(g[n-k].y_2[k])|^2}} \quad (19)$$

In (19), σ can be estimated by

$$\hat{\sigma} = \sqrt{(\hat{\sigma}_y^2 - \hat{\sigma}_n^2)_+}, \quad (20)$$

where $\hat{\sigma}_n$ is the noise deviation which is estimated from the finest scale wavelet coefficients by using a robust median estimator [22] as follows

$$\hat{\sigma}_n^2 = \frac{median(|y_i|)}{0.6745}. \quad (21)$$

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

$$\hat{\sigma}_y^2 = \frac{1}{M}\sum_{y_i \in N(k)} |y_i|^2, \quad (22)$$

where M is the size of the neighborhood $N(k)$. In the packet wavelet transform, the cousin scales have not any parent scale. In this case, we can use hard thresholding estimator [21] to recover cousin coefficients $\hat{w_{cs}}$:

$$\hat{w_{cs}} = (y_{cs} - \sigma_n\sqrt{2\log N})_+, \tag{23}$$

Now, after getting new bivariate shrinkage functions, we should compare this new function to the bivariate function of Sendur [34] as the table 1. From this table, our function has four different parts with Sendur's. Now, we have one more pre-processing step to save data at the border of CGH data. That is signal extension which will be discussed more as follow.

Signal Extension

CGH data is finite signal. If we apply wavelet smooth method directly, we may get error at the border of denoised signal. So, extension step is a very important preprocessing step before denoising. There are three main extension methods. According to the book [36] (chapter 8), symmetric extension is the best if applied to a filtered image because we can save information at the border better. With CGH data, we also need save the information at the border. So, we recommend that symmetric extension method should be used as a preprocessing step before denoising. Let's assume that the length of the CGH 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 [37]. If N is not a power of two, we can extend our signal to make sure $N = 2^j$ by using symmetric extension method. Finally, the SWPT-Bi will be detailed in next part.

Proposed Method

The DWT with the redundant ratio of $1:1$ is efficient for the denoising applications. However, the DWT creates artifacts around the discontinuities of the input signal [30] because it is shift-variant. To overcome this problem, SWT [5] or MODWT [8] and DTCWT [26,35] with translation invariant property was proposed for signal denoising. It has been shown that many of the artifacts could be suppressed by a redundant representation of the signal [30]. One important thing is that CGH data contains many step functions which their information is in both low frequency and high frequency. The above wavelet methods have one disadvantage which some high frequency components of CGH data were removed. In this paper, the SWPT will be used to overcome some above problems because it keeps shift invariant property and looks for signal both in low frequency and in high frequency band for denoising operation. Several methods were proposed for selecting thresholding values such as hard universal [21, 22] and ununiversal thresholding [23]. However, the dependency between wavelet coefficients are not exploited in these methods. Thus, we propose the usage of shift invariant SWPT and new bivariate shrinkage estimator which takes advantage of the dependency between wavelet coefficient and its cousin for array-based DNA copy number data denoising.

Our purpose is to find $\hat{D}$ from Y so that the root mean squared error (RMSE) (24) is the smallest.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}^{N}(\hat{D}_i - D_i)^2}, \tag{24}$$

and N is the number of input samples, $D = \{D_i\}$ and $\hat{D} = \{\hat{D}_i\}$.

We propose a stationary wavelet packet transform and new bivariate shrinkage function based smooth method (SWPT-Bi). The SWPT-Bi can be summarized as follows:

Step 1 : *Extend Y by using symmetric extension method and decompose new data Y' by the SWPT to L levels as (25). The numbers of decomposition levels [38] (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 [38] which yields the best denoising result. In this paper, we use $J = 4$ as the same in [8] and [5].

Step 2 : *Calculate the noise variance $\hat{\sigma}_n^2$ and the marginal variance $\hat{\sigma}^2$ for wavelet coefficient y_k by using (21), (22) and (20).*

Step 3 : *Estimate the child coefficients $\hat{w}_c = \hat{w}_1$ as in (19) and estimate the cousin coefficients $\hat{w}_{cs}$ as in (23). In this case, $k = 1.45$ should be chosen.*

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

We also propose one simple method SWPT. In the SWPT method, hard thresholding [22] method is used. The SWPT method can be summarized as follows:

Step 1 : *Extend Y by using symmetric extension and decompose new data using the SWPT.*

Step 2 : *Estimate the noise variance σ_n^2 by (21).*

Step 3 : *Find the denoised coefficients from noisy coefficients as follow*

$$\hat{w}_i = (y_i - \sigma_n \sqrt{2 \log N})_+, \qquad (26)$$

where N is length of y.

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

References

1. Pinkel D, Segraves R, Sudar D, Clark S, *et al*: **High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays**. *Nat Genet* 1998, **20**:207–211.
2. Snijders AM, Nowak N, Segraves R, Blackwood S, *et al*: **Assembly of microarrays for genome-wide measurement of DNA copy number**. *Nat Genet* 2001, **29**(3):263–264.
3. Brennan C, Zhang Y, Leo C, Fenga B, *et al*: **High-resolution global profiling of genomic alterations with long oligonucleotide microarray**. *Cancer Res* 2004, **64**:4744–4748.
4. Pollack J, Perou C, Alizadeh A, Eisen M, *et al*: **Genome-wide analysis of DNA copy-number changes using cDNA microarrays**. *Nat Genet* 1999, **23**:41–46.
5. Wang Y, Wang S: **A Novel Stationary Wavelet Denoising Algorithm for Array-Based DNA Copy Number Data**. *International Journal of Bioinformatics Research and Applications* 2007, **3**(2):206 – 222.
6. Beheshti B, Braude I, Marrano P, Thorner P, Zielenska M, Squire J: **Chromosomal localization of DNA amplifications in neuroblastoma tumors using cDNA microarray comparative genomic hybridization**. *Neoplasia* 2003, **5**:53–62.
7. Eilers P, de Menezes R: **Quantile smoothing of array CGH data**. *Bioinformatics* 2005, **21**:1146–1153.
8. LHsu, SGSelf, DGrove, TRandolph, KWang, JJDelrow, LLoo, PPorter: **Denoising array-based comparative genomic hybridization data using wavelets**. *Biostatistics(Oxford,England)* 2005, **6**(2):211–226.
9. Pollack J, Sorlie T, Perou C, Rees C, *et al*: **Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors**. *Proc. Natl Acad. Sci.* 2002, **99**:12963–12968.
10. Daruwala R, Rudra A, Ostrer H, Lucito R, Wigler M, Mishra B: **A versatile statistical analysis algorithm to detect genome copy number variation**. *Proc. Natl Acad. Sci.* 2004, **101**:16292–16297.
11. Hupe P, Stransky N, Thiery JP, Radvanyi F, Barillot E: **Analysis of array CGH data: from signal ratio to gain and loss of DNA regions**. *Biostatistics* 2004, **20**:3413–3422.
12. Jong K, Marchiori E, Meijer G, Vaart A, Ylstra B: **Breakpoint identification and smoothing of array comparative genomic hybridization data**. *Bioinformatics* 2004, **20**:3636–3637.
13. Myers C, Dunham M, Kung S, Troyanskaya O: **Accurate detection of aneuploidies in array CGH and gene expression microarray data**. *Bioinformatics* 2004, **20**:3533–3543.
14. Olshen A, Venkatraman E, Lucito R, Wigler M: **Circular binary segmentation for the analysis of array-based DNA copy number data**. *Biostatistics* 2004, **5**:557–572.
15. Bilke S, Chen QR, Whiteford CC, Khan J: **Detection of low level genomic alterations by comparative genomic hybridization based on cDNA microarrays**. *Bioinformatics* 2005, **21**(7):1138–1145.
16. Lai W, Johnson M, Kucherlapati R, Park P: **Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data**. *Bioinformatics* 2005, **21**:3763–3770.
17. Picard F, *et al*: **A statistical approach for array CGH data analysis**. *BMC Bioinformatics* 2005, **6,27**.
18. Wang P, *et al*: **A method for callling gains and losses in arrray CGH data**. *Bioinformatics* 2005, **6**.
19. Fridkyand J, *et al*: **Hidden markov models approach to the analysis of arrray CGH data**. *J.Multivariate Anal.* 2004, **90**:132–153.
20. Lingjaerde O, *et al*: **CGH-Exploxer: a program for analysis of array-CGH data**. *Bioinformatics* 2005, **21**:821–822.
21. Donoho D, Johnstone I: **Ideal spatial adaptation by wavelet shrinkage**. *Biometrika* 1994, **81**:425–455.
22. Donoho D: **De-Noising by soft-thresholding**. *IEEE Trans. on Inf. Theory* 1995, **41**(3):613–627.
23. Johnstone I, Silverman B: **Wavelet Threshold Estimators for Data with Correlated Noise**. *Journal of the Royal Statistical Society* 1997, (59):319–351.
24. Chang S, Yu B, Vetterli M: **Adaptive wavelet thresholding for image denoising and compression**. *IEEE Trans.Image processing* Sept2000, **9**:1532–1546.
25. Li Y, Zhu J: **Analysis of array CGH data for cancer studies using fused quantile regression**. *Bioinformatics* 2007, **23**:2470–2476.
26. Nguyen N, Huang H, Oraintara S, Vo A: **A New Smoothing Model for Analyzing Array CGH Data**. *IEEE BIBE* 2007.

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

28. AMSnijders, NNowak, RSegraves, SBlackwood: **Assembly of microarrays for genome wide measurement of DNA copy number by CGH**. *Nature Genetics* 2001, **29**:263–264.

29. University S: **Assembly of microarrays for genome-wide measurement of DNA copy number**[http://www.nature.com/ng/journal/v29/n3/suppinfo/ng754_S1.html].

30. Coifman R, Donoho D: **Translation-invariant denoising**. *Wavelets and Statistics* 1995, **103 of Lecture Notes in Statistics**:125–150.

31. Kingsbury NG: **Image Processing with Complex Wavelets**. *Phil. Trans. Royal Society London A* 1999, **357**(1760):2543–2560.

32. Kingsbury NG: **Complex wavelets for shift invariant analysis and filtering of signals**. *Journal of Applied and Computational Harmonic Analysis* 2001, **10**(3):234–253.

33. Selesnick IW, Baraniuk RG, Kingsbury NC: **The dual-tree complex wavelet transform**. *IEEE Signal Processing Magazine* 2005, **22**(6):123–151.

34. Sendur L, Selesnick I: **Bivariate Shinkage Function for Wavelet-Based Denoising exploiting Interscale Dependency**. *IEEE Transaction on Signal Processing* November 2002, **50**(11).

35. Nguyen N, Huang H, Oraintara S, Wang Y: **Denoising of Array-Based DNA Copy Number Data Using The Dual-tree Complex Wavelet Transform**. *IEEE BIBE* 2007.

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

37. Coifman R, Wickerhauser M: **Entropy-based Algorithms for best basis selection**. *IEEE Trans. on Inf. Theory* 1992, **38**:713–718.

38. Bruce A, Gao H: **Understanding waveshrink: Variance and bias estimation**. *Biometrika* 1996, **83**:727–745.

Figures

Figure 1 - Comparison by RMSE

Comparison of average RMSEs obtained from the 1,000 artificial chromosomes with each of the 7 noise levels using the Lowess, the Quantreg, the SWTi, the DTCWTi-bi and our methods such as the SWPT and the SWPT-Bi.

Figure 2 - Example of wavelet denoising results

Example of wavelet denoising results at the noise level of $\sigma = 0.2$ using the Lowess, the Quantreg, the SWTi, the DTCWTi-bi and our methods such as the SWPT and the SWPT-Bi..

Figure 3 - Real Data Examples

The wavelet denoising results of array CGH data on chromosome 1 in the real signal GM13330 using some methods such as the Lowess, the Quantreg, the SWTi, the DTCWTi-bi and our methods such as the SWPT and the SWPT-Bi.

Figure 4 - Wavelet Transform

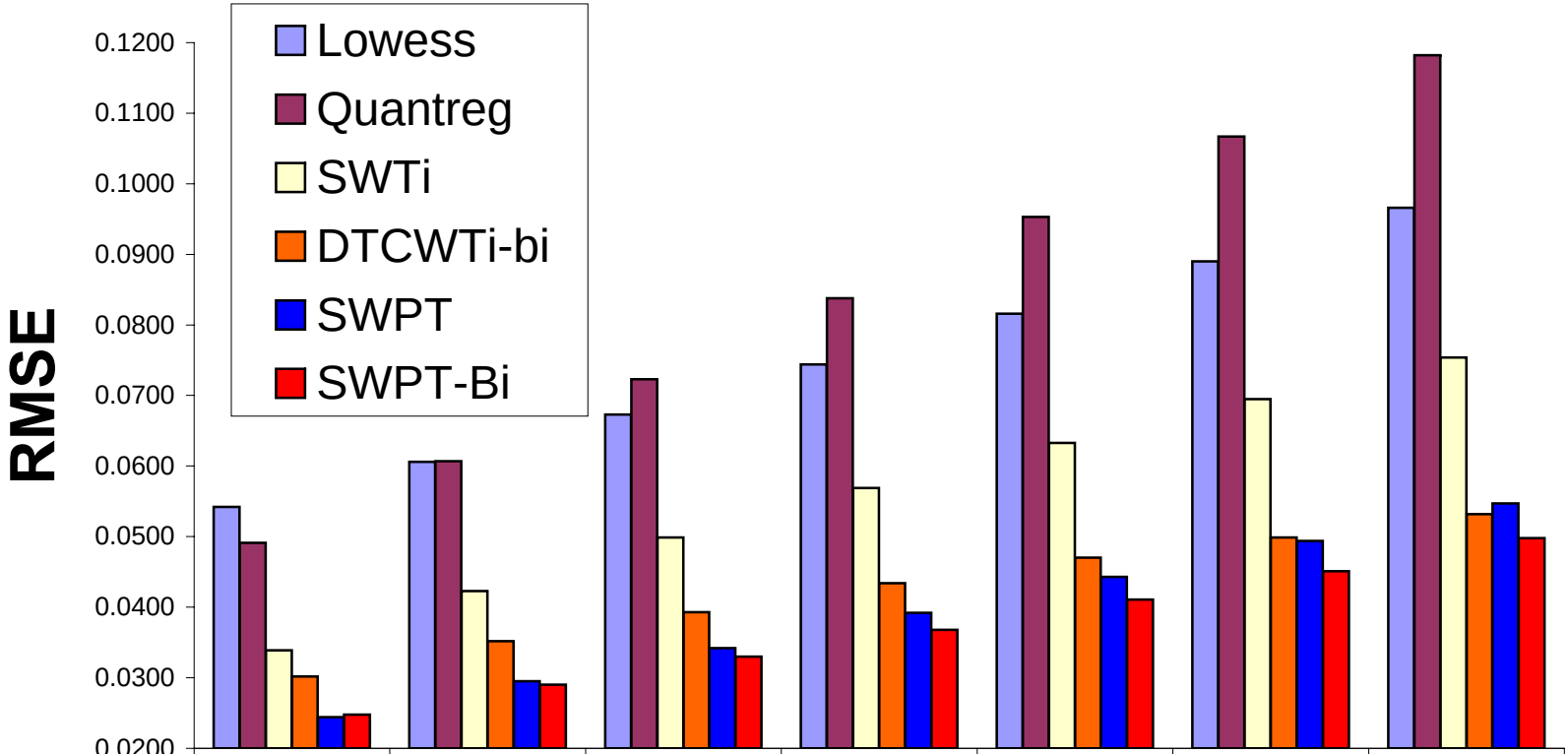
Analysis filter bank and the position of child, parent and cousin coefficients of discrete wavelet transform (DWT), stationary wavelet transform (SWT), dual tree wavelet complex transform (DTCWT), discrete wavelet packet transform (DWPT) and stationary wavelet packet transform (SWPT).

Tables

Table 1 - Comparison table of our new bivariate shrinkage function and function in [34]

Comparison Table		
Method	New bivariate shrinkage function.	Bivariate shrinkage function in [34].
Applying to	CGH data.	image.
Relationship	child and cousin coefficient.	child and parent coefficient.
y_3	$y_3 = h * y_2$, where h is a low pass filter.	$y_p = g * y_2$, where g is a high pass filter.
Transform	SWPT and DWPT.	DWT, SWT and DTCWT.

Figure 1-Comparison by RMSE



	0.100	0.125	0.150	0.175	0.200	0.225	0.250
Lowess	0.0542	0.0606	0.0673	0.0744	0.0816	0.0890	0.0966
Quantreg	0.0491	0.0607	0.0723	0.0838	0.0953	0.1067	0.1182
SWTi	0.0339	0.0423	0.0499	0.0569	0.0633	0.0695	0.0754
DTCWTi-bi	0.0302	0.0352	0.0393	0.0434	0.0470	0.0499	0.0532
SWPT	0.0244	0.0295	0.0342	0.0392	0.0443	0.0494	0.0547
SWPT-Bi	0.0248	0.0290	0.0330	0.0368	0.0411	0.0451	0.0498

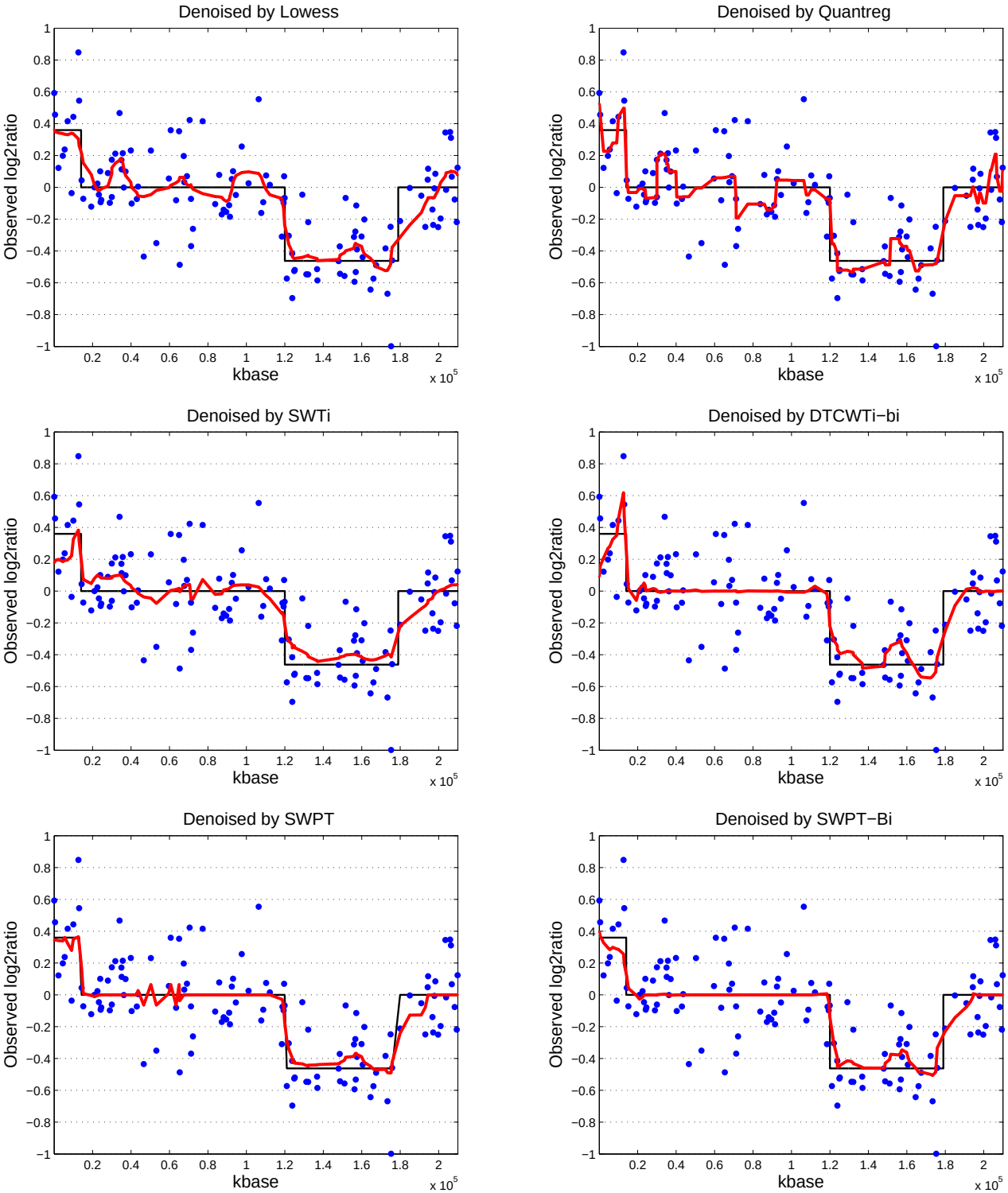


Figure 2 - Example of wavelet denoising results.

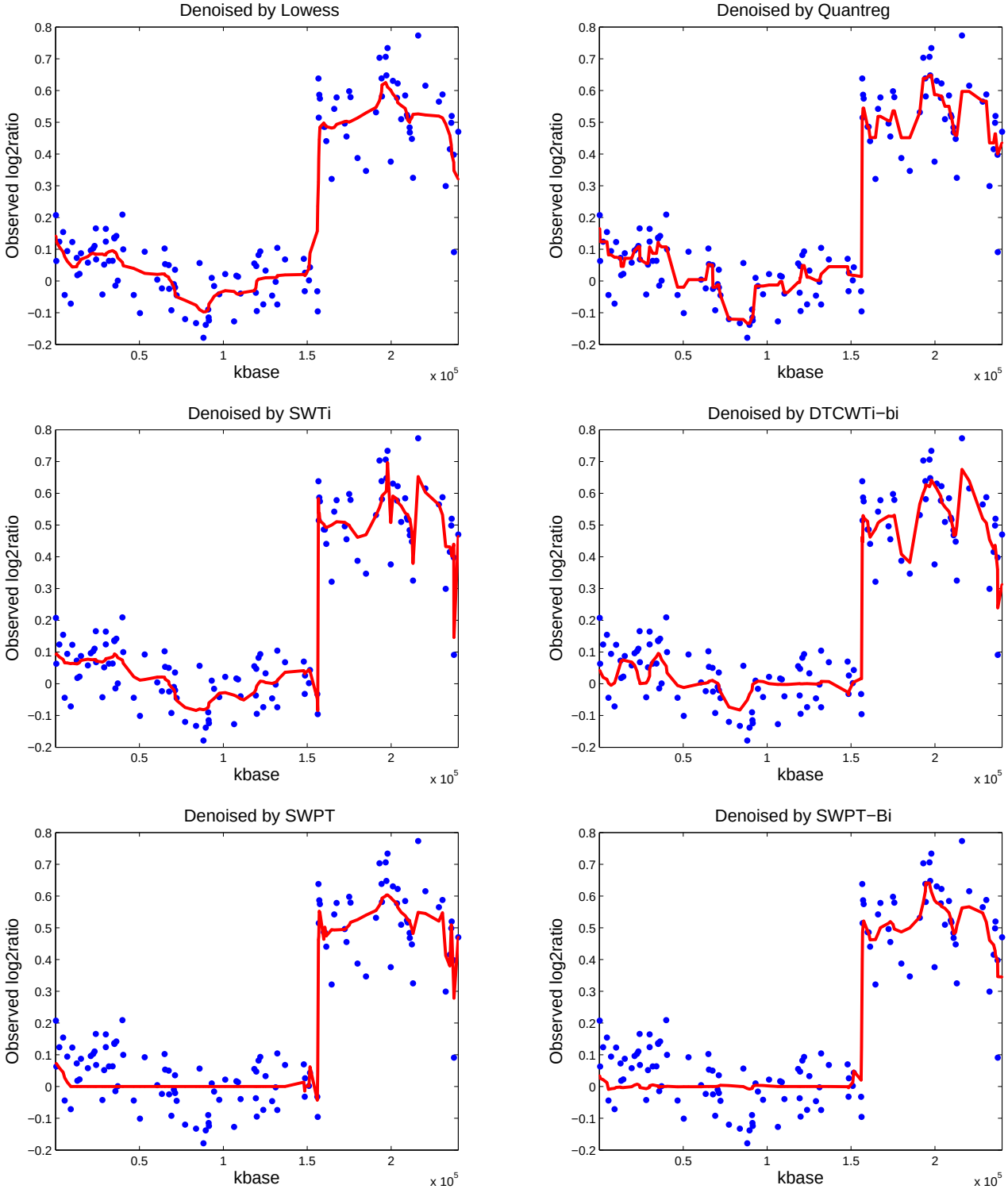


Figure 3 - Real data examples.

Figure 4 – Wavelet Transform