

Denoising of Array-Based DNA Copy Number Data Using The Dual-tree Complex Wavelet Transform

Nha Nguyen*, Heng Huang[†], Soontorn Oraintara[‡] and Yuhang Wang[§]

*Department of Electrical Engineering, University of Texas at Arlington

Email: nhn3175@exchange.uta.edu

[†]Corresponding Author

Department of Computer Science and Engineering, University of Texas at Arlington

Email: heng@uta.edu

[‡]Department of Electrical Engineering, University of Texas at Arlington

Email: oraintar@uta.edu

[§]Department of Computer Science and Engineering, Southern Methodist University

Email: yuhangw@engr.smu.edu

Abstract—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 works focus on both smoothing-based and segmentation-based data processing. In this paper, we propose to use the dual-tree complex wavelet transform to smooth the array CGH data. We demonstrate the effectiveness of our approach through theoretic 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.

I. INTRODUCTION

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 [6], [7], [8], [9], [5] and segmentation-based data processing [10], [11], [12], [13], [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. 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 into N levels using SWT or MODWT, the redundant ratio is $(N + 1) : 1$. It makes denoising algorithm more computationally intensive with worse denoising results.

In this paper, we propose to use the dual-tree complex wavelet transform (DTCWT) to denoise the array CGH data. Because in DTCWT all complex subbands are shift invariant, each complex subband provides a shiftable description of signal in a specific scale. Its representation has a redundant ratio of $2 : 1$. We demonstrate the effectiveness of our approach through theoretic 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 (SWT or MODWT) about 19% – 34.4%.

II. WAVELET METHODS

In this section, we provide a brief review of wavelet transforms which were used for array CGH data smoothing and is used by this paper.

A. Discrete Wavelet Transform

The discrete wavelet transform (DWT), based on the octave band tree structure, decomposes an input signal using a set of lowpass and highpass filters followed by a decimator. Since the number of samples contained by each subband at level N is the number of input samples divided by 2^N , the DWT is not a redundant representation. The total number of wavelet coefficients over all subbands is equal to the length of the input signal. The DWT has been widely and successfully used in several signal processing applications such as denoising, estimation, and compression. However, one of the disadvantages of the DWT is that it is shift-variant. The DWT is highly dependent on its relative position with respect to the downsampling lattice. In signal denoising, the DWT is known to create artifacts around the discontinuities of the input signal [16]. These artifacts degrade the performance of the threshold-based denoising algorithm.

B. Stationary Wavelet Transform

The SWT [16] is similar to the DWT except that it does not employ a decimator after filtering, and the filters are upsampled at each level of decomposition. This is also known as the redundant or translation invariant DWT. The absence of a decimator leads to a redundant signal representation. Each subband contains the same number of samples as the input. So for a decomposition of N levels there is a redundant ratio of $(N + 1) : 1$. However, the translation 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 [16]. Each level's filters are upsampled versions of the previous ones.

C. Dual-tree Complex Wavelet Transform

A dual-tree structure that produces a dyadic complex DWT is proposed by Kingsbury [17], [18]. Since array CGH data are one dimensional signals, in this paper we only talk about the 1-D case of dual-tree CWT. 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 over-complete 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 [17].

The analysis FB for the DTCWT, shown in Fig. 1(a), is an iterative multiscale 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

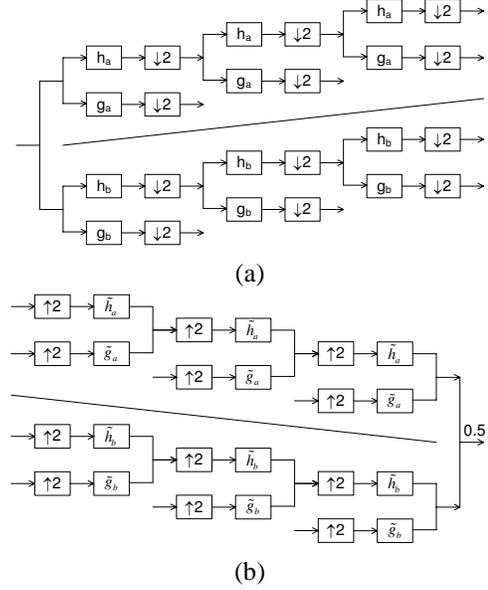


Fig. 1. A 3 level dual-tree CWT. (a) Analysis FB, (b) Synthesis FB .

decomposition. According to Fig. 1(a), 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. For more details of the construction of the dual-tree, the reader is referred to [19].

The most important property of the DTCWT is that all complex subbands are shift invariant in the sense that there is no significant aliasing in the decimated complex subbands. Therefore, each complex subband provides a shiftable description of signal in a specific scale. By construction of the dual-tree CWT, each pair of corresponding filters has the Hilbert transform relation [19]. Therefore, the equivalent complex filter for each subband has one-sided frequency support. The real part of the complex filter is symmetric while the imaginary part is anti-symmetric.

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 obtain the final output. Fig. 1(b) shows the synthesis FB for the inverse DTCWT.

III. DNA COPY NUMBER DATA DENOISING USING DUAL-TREE CWT ALGORITHM

Basic wavelet-based signal denoising algorithms use wavelet transforms such as DWT and SWT, and hard or soft thresholding. The DWT with the redundant ratio of $1 : 1$ is efficient for the denoising applications. However, because the DWT is shift-variant, it creates artifacts around the discontinuities of the input signal [16]. These artifacts degrade the performance of the threshold-based denoising algorithm [16]. To overcome this problem, SWT 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 [16]. However, if a signal is decomposed into N levels, the redundant ratio is $(N + 1) : 1$. It makes denoising algorithm more computationally intensive. We propose the use of DTCWT for array-based DNA copy number data denoising. Several methods were proposed for selecting thresholding values such as SURE, Fixed, Mixture and Minimax [20], [21]. In our experiments, the above thresholding methods are performed on array CGH signals. The experimental results show that the universal threshold works well as them for DTCWT in array CGH signal denoising.

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 .

$$Y = D + n. \quad (1)$$

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

$$RMSE = \sqrt{\frac{1}{N} \sum_i^N (\hat{D}_i - D_i)^2}, \quad (2)$$

and N is the number of input samples. After decomposing the data Y by the DTCWT, we get the complex coefficients W_i . All complex coefficients whose magnitudes are less than a particular threshold are set to zero as follows:

$$W_i = \begin{cases} 0 & \text{if } |W_i| \leq \delta_U, \\ W_i & \text{if } |W_i| > \delta_U. \end{cases} \quad (3)$$

After that, the subband coefficients denoised are used to reconstruct the data \hat{D} . Next, we discuss how to choose the threshold values.

The universal threshold is defined in [20], [21] by:

$$\delta_U \equiv \sigma_n \sqrt{2 \log(N)}, \quad (4)$$

where N denotes the number of samples in data Y and σ_n is the standard deviation of Gaussian noise which we want to remove. In real situations, the variance of noise to be removed is unknown. So Donoho [20] proposed a special method to estimate this value by using the following equation:

$$\sigma_n \equiv \frac{\text{median} \left(|W_{1,0}^{(D)}|, |W_{1,1}^{(D)}|, \dots, |W_{1,N_1-1}^{(D)}| \right)}{0.6745}, \quad (5)$$

where N_1 is the length of DWT subband at level 1.

We assume that noise in the DNA data is IID. The universal threshold denoising method can be summarized as follows:

- Step 1: Insert zeros into Y as in section IV-C and decompose new data using the DTCWT.
- Step 2: Estimate the noise variance σ_n^2 by (5).
- Step 3: Calculate the threshold by (4)
- Step 4: Find the denoised coefficients from noisy coefficients by (3).
- Step 5: Reconstruct data \hat{D} from the denoised coefficients by taking inverse DTCWT.

For the SWT, only the scaling coefficients are denoised. However, for the DTCWT, all subband coefficients are denoised. In DTCWT method, we get complex subbands, thus the absolute values of the real SWT coefficients are replaced by the magnitudes of complex coefficients (DTCWT). This gives out a better result than the method using the absolute values of real and imaginary subbands separately.

Meantime, we also propose the DTCWT with interpolating (DTCWTi) method to denoise array CGH data which will be discussed in section IV-D.

IV. EXPERIMENTS AND DISCUSSIONS

In our experiments, the artificial chromosomes are generated as in [22] and [5]. Since they are unequally spaced data, we apply the interpolating method with the Pseudo-markers [5] for them. In order to guarantee the number of data points to be a power of two, the zero-padding is implemented. During our experiments, the denoising results of different methods are compared.

A. Artificial Chromosome Generation

Willenbrock and Fridlyand [22] proposed a simulation model to create the synthetic array CGH data. In their model, a primary tumor dataset of 145 samples is segmented and the probes are equally spaced along the chromosome. More recently Y. Wang and S. Wang [5] extended this model by placing unequally spaced probes along chromosome. The primary tumor data set is segmented using DNAcopy number levels from the empirical distribution of segment mean values smv as

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

The synthetic DNA copy number data on a chromosome was generated as follows:

- Step 1 Determine copy number probability and the distribution of segment length. As suggested in [22] and [5], the chromosomal segments with DNA copy number $c = 1, 2, 3, 4$ 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 [22].
- Step 2 Compute $\log_2 ratio$. 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 [22], the $\log_2 ratio$ is calculated by

$$\log_2 ratio = \log_2 \left(\frac{cP_t + 2(1 - P_t)}{2} \right), \quad (6)$$

where c is the assigned copy number. The expected $\log_2 ratio$ value is then the latent true signal.

Step 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.

Step 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 $\sigma_n = 0.1, 0.125, 0.15, 0.175$ and 0.2 .

B. DNA Copy Number Data Interpolation

DNA copy number data contains a lot of noises and the distances between two samples vary greatly. Kovac [23] proposed a new method to change this kind of data to equally spaced data and got good performance in denoising application. However, in his method, the numbers of new samples were created densely. In [5] DNA copy number data interpolation with pseudo-markers was proposed. Their method has some advantages: the number of new samples is not dense and the content of data does not change much. Suppose that the observed DNA copy number data Y at M probe locations on the chromosome of length L is

$$Y(x_i) = D(x_i) + n_i,$$

where $i = 1, 2, \dots, M$, x_i are the probe locations with $0 \leq x_1 \leq x_2 \leq \dots \leq x_M \leq L$, and $D(x_i)$ is the latent true signal at location x_i . We know that x_i are not equally spaced. Therefore, the equally spaced pseudo-markers between sparse probes are inserted and this interpolating method can be summarized as the following steps:

Step 1 Create a set P of locations along the chromosome at the uniform spacing of q by

$$P = \{p_j | p_j = kq, k = 0, 1, 2, \dots, \lfloor L/q \rfloor\},$$

where q is the average distance between adjacent probes.

Step 2 Form a set P' of locations which will be inserted to DNA copy number data $Y(x_i)$. A pseudo-marker can be inserted at each location p_j . However, to avoid the worst cases when the original marker and a pseudo-marker overlap or their distance is arbitrarily small, a subset P' of P is formed by

$$P' = \{p'_j | p'_j \in P, |p'_j - x_i| \geq q/2 \text{ for all } i = 1, 2, \dots, M\}.$$

P' includes the points which are in the set P and satisfies a condition that the distances from that point to the original locations must be larger than $q/2$ but less than $3q/2$.

Step 3 Apply the nearest neighbor interpolation to obtain the interpolated values for $Y(p'_j)$.

Step 4 Merge the original signal $Y(x_i)$ and the interpolated signal $Y(p'_j)$ by

$$Y' = \{Y(x_i) : i = 1, 2, \dots, M\} \cup \{Y(p'_j) : p'_j \in P'\},$$

and make sure that x_i and p'_j are in ascending order.

The new CGH data $Y'(x_i)$ instead of the original $Y(x_i)$ will be denoised.

C. Zero-padding

After interpolating, the length of the CGH signal is N . In order to get best performance in the wavelet denoising algorithm with thresholding, the length of the input signal is required to be a power of two [24]. If N is not a power of two, we can apply the zero-padding method to make sure that N' is the value we need with $N' = 2^j$. For example, we get $N = 120$ after interpolating, we must insert more eight zeros and then we get $N' = 128$. The numbers of decomposition levels can be computed by

$$L = \log_2(N) - 4. \quad (7)$$

In the above example, we get $L = 3$. This is a perfect number of levels which yields the best denoising results.

D. Experiments Design

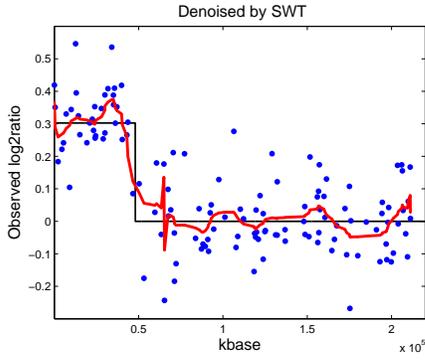
In our experiments, the signals $Y(x_i)$ (after zero-padding) and the new interpolated signal $Y'(x_i)$ are decomposed by the DTCWT. We call them as DTCWT and DTCWTi, respectively. The universal thresholding method is applied to denoise the noisy wavelet coefficients. We compare our method's results with the other method's such as the SWT and the SWTi [5].

DTCWT The array CGH signal is generated as in section IV-A and applied the zero-padding method as in section IV-C. Then, this signal is decomposed by the DTCWT to L levels as equation (7). The universal thresholding value is applied to denoise the noisy DTCWT coefficients as steps in section III. DWT coefficients of the array CGH signal at level 1 is used to estimate noise by equation (5). Finally, we reconstruct the signal by taking the inverse DTCWT from the denoised coefficients.

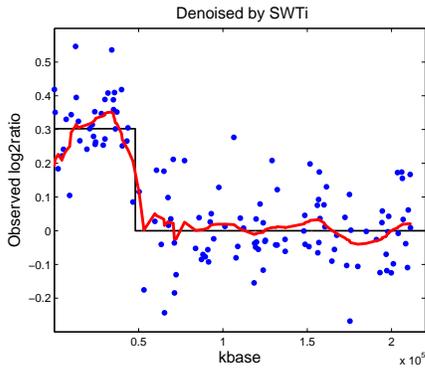
DTCWTi Using the same steps in DTCWT method. However, the array CGH signal is interpolated by section IV-B before inserting zeros. In step 1, instead of denoising array CGH signal, the signal with interpolation is processed. The rest of steps are as the same step 2 to step 5 as in section III.

SWT This method comes from paper [5]. Compared with the DTCWT, the SWT method has two different steps: 1) the array CGH signal is decomposed by the SWT; 2) the real coefficients of SWT subbands are denoised instead of the complex coefficients. We also use the universal thresholding in this method.

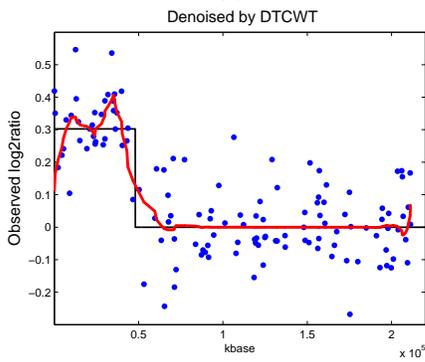
SWTi It extends the SWT method by interpolating array CGH signal before decomposition. They apply universal thresholding to denoise the SWT coefficients [5].



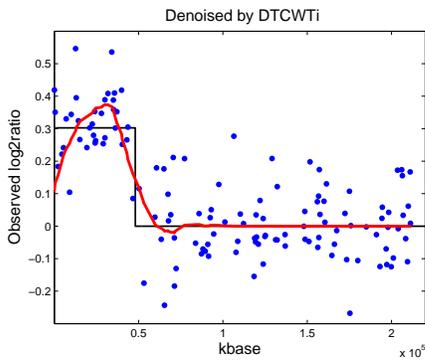
(a)



(b)

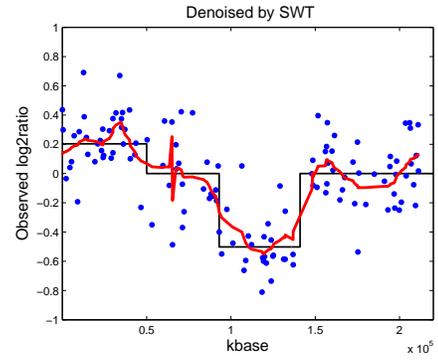


(c)

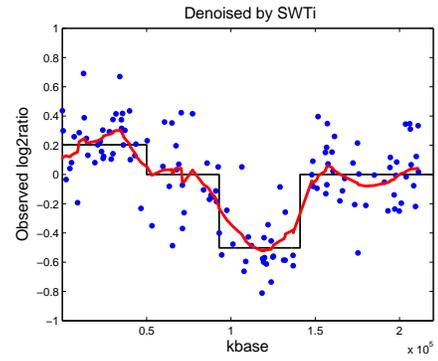


(d)

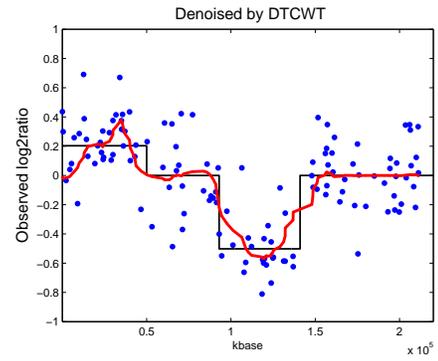
Fig. 2. Example of wavelet denoising results at the noise level of $\sigma = 0.1$ using (a) SWT, (b) SWTi, (c) DTCWT and (d) DTCWTi.



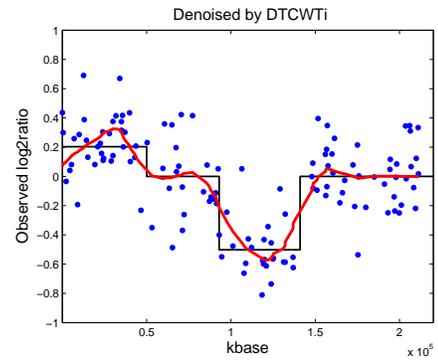
(a)



(b)



(c)



(d)

Fig. 3. Example of wavelet denoising results at the noise level of $\sigma = 0.2$ using (a) SWT, (b) SWTi, (c) DTCWT and (d) DTCWTi.

TABLE I
COMPARISON OF AVERAGE RMSEs OBTAINED FROM THE 1,000
ARTIFICIAL CHROMOSOMES WITH EACH OF THE 5 NOISE LEVELS USING
SWT, SWTi, DTCWT AND DTCWTi

σ	SWT	SWTi	DTCWT	DTCWTi
0.1	0.0368	0.0338	0.0305	0.0298
0.125	0.0460	0.0422	0.0354	0.0351
0.15	0.0548	0.0497	0.0399	0.0393
0.175	0.0633	0.0567	0.0442	0.0435
0.2	0.0715	0.0631	0.0481	0.0469

E. Empirical Results

The denoising results of the SWT, SWTi, DTCWT and DTCWTi methods are shown in the table I corresponding to the five noise levels $\sigma_n = 0.1, 0.125, 0.15, 0.175$ and 0.2 . As shown in the table I, the DTCWT yields much better performance than the SWT and the SWTi. The average RMSEs of the DTCWTi is less than that of the DTCWT. The DTCWTi provides the much better denoising results than the SWTi. Moreover, the DTCWT is more efficient and has less computation than the SWT because the redundancy ratio of the DTCWT 2 : 1 is much less than that of the SWT 4 : 1 (if number of level decomposition $L = 3$).

We run the denoising algorithm for 1,000 artificial chromosomes with five different noise levels. The denoising results of the SWT, SWTi, DTCWT and DTCWTi methods are shown in the table I corresponding to the five noise levels $\sigma_n = 0.1, 0.125, 0.15, 0.175$ and 0.2 . As shown in the table I, the DTCWT yields much better performance than the SWT and the SWTi. The average RMSEs of the DTCWTi is less than that of the DTCWT. At the noise level of $\sigma_n = 0.2$, the average RMSEs of the SWTi method is 0.0631, while that of the DTCWTi is 0.0469. The DTCWTi provides the much better denoising results than the SWTi. Moreover, the DTCWT is more efficient and has less computation than the SWT because the redundancy ratio of the DTCWT 2 : 1 is much less than that of the SWT 4 : 1 (if the number of level decomposition $L = 3$).

Some examples of wavelet denoising results by using four denoising methods (SWT, SWTi, DTCWT and DTCWTi) are shown in Fig. 2 at the noise level of $\sigma = 0.1$ and in Fig. 3 at noise level $\sigma = 0.2$. From these figures, we can see that DTCWTi denoising method provides the better performance than the others.

The RMSEs between the denoised signals and the latent true signals are computed to evaluate these four methods and the average RMSEs results (from 1000 artificial chromosomes) are shown in table I. From table I, we can see that, on average, the DTCWTi outperformed the SWT by 19% – 34.4%, and the SWTi by 11.8% – 22.7% in terms of the RMSE. In all of the noise levels, the DTCWTi achieved much better results than the others.

F. Real Data Examples

We use the BAC array data on 15 fibroblast cell lines [25], [8] to show that denoising by the DTCWTi is better than

by the SWTi. This data set can be freely downloaded at <http://www.nature.com/ng/journal/v29/n3/supinfo/ng754S1.html>. Because the true copy number changes are known for these cell lines, we choose these data as a proof of principles. We pick up two chromosomes from this data and use the SWT, the SWTi, the DTCWT and the DTCWTi for denoising. These are the chromosome 9 of MPE600 and the chromosome 14 of GM01750.

In Fig. 4, number copy is from zero to two. With the copy two, Fig. 4 (c) and (d) show that the DTCWT and DTCWTi give the smoother denoised signal than the SWT and SWTi. With the zero copy and the one copy, the performance of the DTCWTi denoising method is the best in Fig. 4 (d).

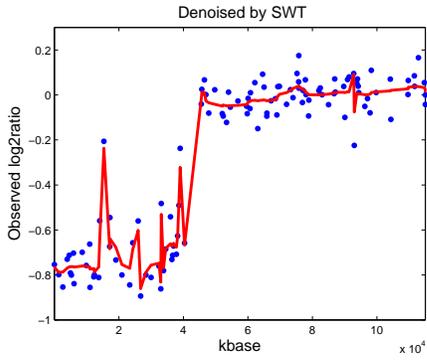
In figure 5, number copy is from two to four. The figure 4 (d) shows that denoising by DTCWTi still provides the better results than the others. From two above figures, it is easy to see that DTCWT and, specially, DTCWTi perform better than SWT and SWTi in denoising of real CGH data.

V. CONCLUSIONS

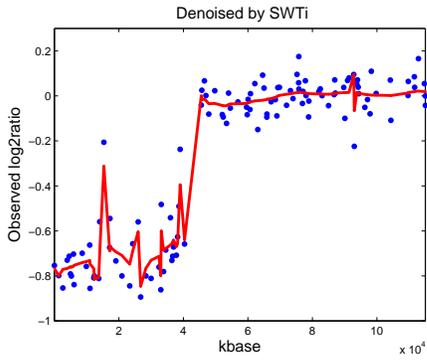
In this paper, we explored the dual-tree complex wavelet transform method in array CGH data denoising study. The unequal spacing of probes on the chromosome is taken into account. In the simulation situations, the denoising results from DTCWT and DTCWTi are much better (improve 19%–34.4% and 11.8% – 22.7%) than 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.

REFERENCES

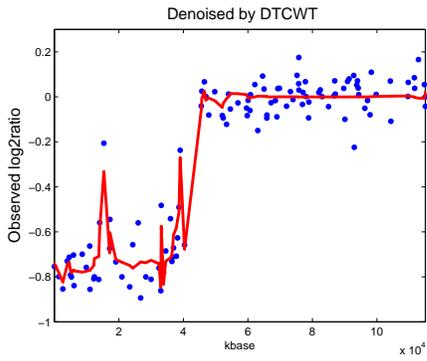
- [1] D. Pinkel, R. Seagraves, D. Sudar, S. Clark, and *et al.*, “High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays,” *Nat Genet*, vol. 20, pp. 207–211, 1998.
- [2] A. M. Snijders, N. Nowak, R. Seagraves, S. Blackwood, and *et al.*, “Assembly of microarrays for genome-wide measurement of dna copy number,” *Nat Genet*, vol. 29, no. 3, pp. 263–264, 2001.
- [3] C. Brennan, Y. Zhang, C. Leo, B. Fenga, and *et al.*, “High-resolution global profiling of genomic alterations with long oligonucleotide microarray,” *Cancer Res*, vol. 64, pp. 4744–4748, 2004.
- [4] J. Pollack, C. Perou, A. Alizadeh, M. Eisen, and *et al.*, “Genome-wide analysis of DNA copy-number changes using cDNA microarrays,” *Nat Genet*, vol. 23, pp. 41–46, 1999.
- [5] Y. Wang and S. Wang, “A novel stationary wavelet denoising algorithm for array-based DNA copy number data,” *International Journal of Bioinformatics Research and Applications*, vol. 3, no. 2, pp. 206 – 222, 2007.
- [6] B. Beheshti, I. Braude, P. Marrano, P. Thorner, M. Zielenska, and J. Squire, “Chromosomal localization of dna amplifications in neuroblastoma tumors using cna microarray comparative genomic hybridization,” *Neoplasia*, vol. 5, p. 53C62, 2003.
- [7] P. Eilers and R. de Menezes, “Quantile smoothing of array cgh data,” *Bioinformatics*, vol. 21, p. 1146C1153, 2005.
- [8] L. Hsu, S. Self, D. Grove, T. Randolph, K. Wang, J. Delrow, L. Loo, and P. Porter, “Denoising array-based comparative genomic hybridization data using wavelets,” *Biostatistics(Oxford,England)*, vol. 6, no. 2, pp. 211–226, 2005.
- [9] J. Pollack, T. Sorlie, C. Perou, C. Rees, and *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.*, vol. 99, p. 12963C12968, 2002.



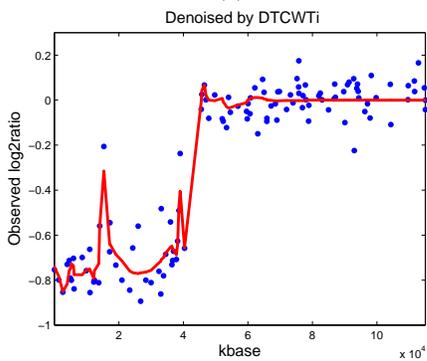
(a)



(b)

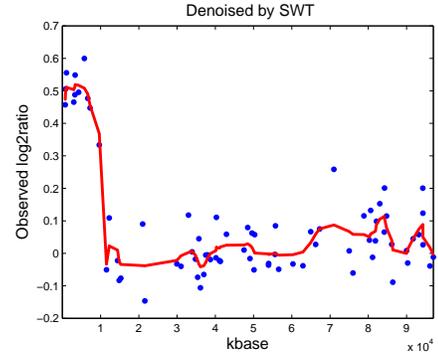


(c)

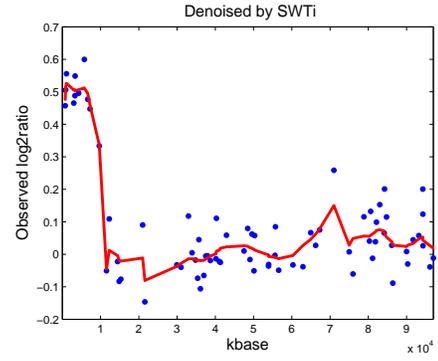


(d)

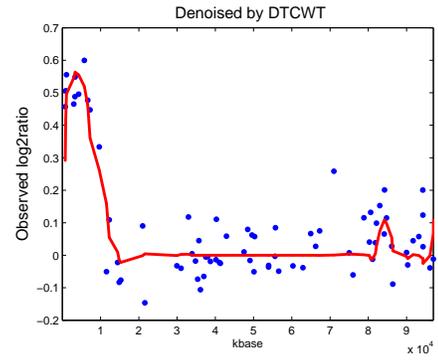
Fig. 4. The wavelet denoising results of array CGH data on chromosome 9 in the real signal MPE600 using (a) SWTi, (b) SWTi, (c) DTCWT and (d) DTCWTi.



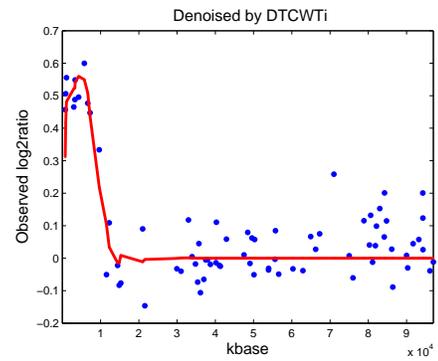
(a)



(b)



(c)



(d)

Fig. 5. The wavelet denoising results of array CGH data on chromosome 14 in the real signal GM01750 using (a) SWTi, (b) SWTi, (c) DTCWT and (d) DTCWTi.

- [10] R. Daruwala, A. Rudra, H. Ostrer, R. Lucito, M. Wigler, and B. Mishra, "A versatile statistical analysis algorithm to detect genome copy number variation," *Proc. Natl Acad. Sci.*, vol. 101, pp. 16 292–16 297, 2004.
- [11] P. Hupe, N. Stransky, J. P. Thiery, F. Radvanyi, and E. Barillot, "Analysis of array cgh data: from signal ratio to gain and loss of dna regions," *Biostatistics*, vol. 20, pp. 3413–3422, 2004.
- [12] K. Jong, E. Marchiori, G. Meijer, A. Vaart, and B. Ylstra, "Breakpoint identification and smoothing of array comparative genomic hybridization data," *Bioinformatics*, vol. 20, pp. 3636–3637, 2004.
- [13] C. Myers, M. Dunham, S. Kung, and O. Troyanskaya, "Accurate detection of aneuploidies in array cgh and gene expression microarray data," *Bioinformatics*, vol. 20, pp. 3533–3543, 2004.
- [14] A. Olshen, E. Venkatraman, R. Lucito, and M. Wigler, "Circular binary segmentation for the analysis of array-based dna copy number data," *Biostatistics*, vol. 5, pp. 557–572, 2004.
- [15] S. Bilke, Q. R. Chen, C. C. Whiteford, and J. Khan, "Detection of low level genomic alterations by comparative genomic hybridization based on cdna microarrays," *Bioinformatics*, vol. 21, no. 7, pp. 1138–1145, 2005.
- [16] R. Coifman and D. Donoho, "Translation-invariant de-noising," *Wavelets and Statistics*, vol. 103 of Lecture Notes in Statistics, pp. 125–150, 1995.
- [17] N. Kingsbury, "Image processing with complex wavelets," *Phil. Trans. Royal Society London A*, vol. 357, no. 1760, pp. 2543–2560, Sept 1999.
- [18] —, "Complex wavelets for shift invariant analysis and filtering of signals," *Journal of Applied and Computational Harmonic Analysis*, vol. 10, no. 3, pp. 234–253, May 2001.
- [19] I. Selesnick, R. Baraniuk, and N. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 123–151, Nov 2005.
- [20] D. Donoho, "De-noising by soft-thresholding," *IEEE Trans. on Inf. Theory*, vol. 41, no. 3, pp. 613–627, 1995.
- [21] D. Donoho and I. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.
- [22] H. Willenbrock and J. Fridlyand, "A comparison study: applying segmentation to array CGH data for downstream analyses," *Bioinformatics*, vol. 21, no. 22, pp. 4084–4091, 2005.
- [23] A. Kovac and B. Silverman, "Extending the scope of wavelet regression methods by coefficient-dependent thresholding," *Journal of the American Statistical Association*, vol. 95, no. 449, pp. 172–183, 2000.
- [24] D. Percival and A. Walden, *Wavelet Methods for Time Series Analysis*. Cambridge University Press, 2006.
- [25] A. Snijders, N. Nowak, R. Segraves, S. Blackwood, and *et al.*, "Assembly of microarrays for genome-wide measurement of DNA copy number by CGH," *Nature Genetics*, vol. 29, pp. 263–64, 2001.