

A COMPARISON STUDY OF SIGNAL EXTENSIONS METHODS FOR WAVELET DENOISING OF ARRAY CGH DATA

Yuhang Wang

*Department of Computer Science and Engineering, Southern Methodist University,
Dallas, TX 75205, USA
Email: yuhangw@engr.smu.edu*

Array-based comparative genome hybridization (array CGH) is a recently developed high-throughput technique to detect DNA copy number aberrations. Typically, array CGH data is noisy. Wavelet denoising was previously shown to have superior performance for denoising array CGH data. However, the effect of different signal extensions methods on the performance of wavelet denoising in this particular application has not been previously studied. In this paper, we performed a comparison study of three signal extensions methods (zero-padding, periodic extension, and symmetrization) for wavelet denoising of array CGH data using realistically generated synthetic data. Empirical results suggest that the zero-padding method outperforms the other two methods by 0.9–1.2% in terms of the overall root mean squared error. The difference is statistically significant ($P < 0.01$) in at least 80% of all test cases.

1. INTRODUCTION

Genomic instabilities are often associated with the development and progression of cancer. Amplification or deletion of chromosomal segments can lead to cancer development⁹. Both deletion and amplification change the copy numbers of tumor DNA.

Array-based comparative genome hybridization (array CGH) is a recently developed high-throughput technique to detect DNA copy number aberrations. These high-throughput approaches yield data consisting of \log_2 transformed fluorescence intensity ratios of tumor and reference normal DNA samples. The intensity ratios provide information about DNA copy number aberrations. Typically, array CGH data is very noisy.

1.1. Related Work

To address the noise issue, a few denoising and smoothing schemes have been used for array CGH data analysis. In Ref. 1, the locally weighted regression and scatterplot smoothing (LOWESS) was used. Eilers *et al.*⁵ introduced a quantile smoothing method based on the minimization of sum of absolute errors. Hsu *et al.*⁶ proposed a wavelet denoising scheme using the maximal overlap discrete wavelet transform (MODWT⁸).

Recently, Lai *et al.*⁷ reported an extensive comparative analysis of 11 algorithms for identifying amplifications and deletions in array CGH data, includ-

ing the aforementioned three denoising and smoothing approaches. Their results show that the three denoising and smoothing schemes (LOWESS, quantile smoothing and wavelets) outperform other methods when the noise level is high in the data. Their Receiver Operating Characteristic (ROC) analysis also confirmed the superior performance of wavelet denoising, which is consistent with other studies in the wavelet literature.

Classically, the discrete wavelet transform (DWT) is defined for signals with length of some power of 2. To apply DWT to signals of other sizes, some signal extension methods are needed. Popular methods include zero-padding, periodic extension, and symmetrization¹¹. Border distortions arise when the DWT is applied to the extended signals. Note that although MODWT is defined for signals of any length, it still implicitly uses periodic extension under the scene⁸.

In general, the length of array CGH data on a chromosome is not a power of 2. To the best of our knowledge, the effect of border distortions on the performance of wavelet denoising of array CGH data has not been studied. In this paper, we empirically compare different signal extensions methods in array CGH data denoising.

To facilitate such comparison, we need to generate realistic synthetic array CGH data so that the underlying “ground truth” is known. Lai *et al.*⁷ proposed an simulation model for array CGH data.

However, their model has the following drawbacks: 1) The aberrations were added only to the center of the artificial chromosome, and therefore, the boundary effects were not examined; 2) Their model assumes the probes are equally spaced along the chromosome. However, this is not the case in real array CGH data. In general, the physical distances between adjacent probes along the chromosome are not uniform. This is true for all of currently available CGH arrays. For example, for the UCSF HumArray2 BAC arrays used in Ref. 10, the minimum, median, maximum distances are 1 kb, 829 kb, 29349 kb, respectively.

Willenbrock & Fridlyand¹² also described a simulation model for generating synthetic array CGH data. Their model was designed realistically using a primary breast tumor data set of 145 samples. In particular, the aberrations can occur anywhere in the chromosome in their model. However, their model still assumes the probes to be equispaced along the chromosome. In this paper, we extend their model by placing nonequispaced probes along the chromosomes.

1.2. Our Contributions

In summary, the contributions of this paper are as follows:

- (1) The effect of different signal extensions methods on the performance of wavelet denoising of array CGH data is examined empirically.
- (2) The model for generating synthetic array CGH data due to Willenbrock & Fridlyand¹² is extended to take into account the unequal spacing of probes along the chromosome.

1.3. Outline of The Paper

The remainder of the paper is organized as follows. Section 2 gives a brief review of the wavelet denoising and signal extensions methods. Section 3 describes the model used to generate synthetic array CGH data. Section 4 presents the experiment results on performance comparison of wavelet denoising of synthetic array CGH data using different signal extension methods. Section 5 concludes the paper.

2. WAVELET DENOISING AND SIGNAL EXTENSION METHODS

Pioneered by Donoho and Johnstone⁴, wavelet denoising techniques have been shown to have asymptotic near-optimality properties over a range of function spaces of inhomogeneous smoothness. In general, wavelet denoising works as follows:

- (1) Extend the noisy signal if needed.
- (2) Apply DWT to the extended noisy data to obtain DWT coefficients.
- (3) Apply some thresholding rule to the resulting coefficients, zeroing out small wavelet coefficients whose absolute values are below a certain threshold.
- (4) Apply the inverse DWT to the thresholded coefficients to obtain the denoised data.

When the length of the signal is not divisible by 2^{J_0} , where J_0 is the maximum level of wavelet decomposition, the signal must be extended. Here we consider three signal extension methods: 1) Zero-padding, which assumes that the signal is zero outside the original support; 2) Periodic extension, which assumes the original signal is periodic; 3) Symmetrization, which extends the original signal by symmetric boundary value replication. For more details about these methods, see Ref. 11.

3. A REALISTIC MODEL FOR GENERATING SYNTHETIC ARRAY CGH DATA

The synthetic array CGH data on a chromosome was generated using the following steps:

- (1) As suggested by Willenbrock & Fridlyand¹², chromosomal segments with DNA copy number 0, 1, 2, 3, 4, and 5 were generated with probability 0.01, 0.08, 0.81, 0.07, 0.02, and 0.01, respectively. The lengths for the segments were determined by random sampling from the corresponding empirical length distribution given in Ref. 12.
- (2) Following the same model in Ref. 12, each sample was assumed to be a mixture of tumor cells and normal cells, and the proportion of tumor cells P_t was drawn from a uniform distribution

between 0.3 and 0.7. The expected \log_2 ratio of intensity, computed as $\log_2((cP_t + 2(1 - P_t))/2)$ is then the latent true signal.

- (3) Gaussian noise of mean 0 and variance σ^2 was added to the latent true signal.
- (4) Nonequispaced probes were placed on the chromosome. The distances between adjacent probes were randomly drawn from the empirical distribution of distances obtained from the UCSF HumArray2 BAC array.

Using the model described in the previous section, for each noise level σ of 0.1, 0.125, 0.15, 0.175, and 0.2, we generated synthetic data for 1000 artificial chromosomes of length 200 Mb.

4. EXPERIMENT RESULTS

4.1. Experiment Setup

To evaluate the effect of different signal extension methods on the performance of wavelet denoising of array CGH data, we compared the root mean squared errors (RMSE) obtained after wavelet denoising using the three signal extension methods. In the experiments, we used the Haar wavelet because it fits well with the piece-wise constant nature of true DNA copy number data. In all experiments, the maximum level of wavelet decomposition $J_0 = 3$ was used. Three wavelet thresholding rules were employed: SURE³, hybrid SURE³ and soft thresholding with the universal threshold². Experiments were carried out using the MATLAB Wavelet Toolbox.

4.2. Empirical Results

For each of the 1000 artificial chromosomes, the RMSE between the denoised signal and the latent true signal was computed after wavelet denoising using different signal extension methods. Table 1 shows the average RMSEs obtained using zero-padding, symmetrization, and periodic extension. To evaluate the statistical significance of the differences, we also computed the P values of paired t-test of the RMSEs, as shown in Table 2.

From the results, we can observe that, on average, zero-padding outperforms periodic extension by 1.2%, and symmetrization by 0.9% in terms of the overall root mean squared error. The performance

difference between zero-padding and periodic extension is statistically significant ($P < 0.01$) in all test cases, whereas the difference between zero-padding and symmetrization is statistically significant in 80% of all test cases.

Table 1. Comparison of average RMSE obtained from the 1000 artificial chromosomes using zero-padding (zpd), symmetrization (sym), and periodic extension (per). SU, U and ST denote SURE, hybrid SURE and soft-thresholding rules respectively.

σ	Thres- hold	Signal Extension Methods		
		zpd	sym	per
0.1	SU	0.04779	0.04841	0.04834
	H	0.04414	0.04426	0.04500
	ST	0.04626	0.04613	0.04734
0.125	SU	0.05923	0.05999	0.05972
	H	0.05380	0.05406	0.05468
	ST	0.05540	0.05550	0.05643
0.15	SU	0.07058	0.07168	0.07107
	H	0.06284	0.06328	0.06362
	ST	0.06404	0.06429	0.06501
0.175	SU	0.08204	0.08344	0.08273
	H	0.07284	0.07331	0.07367
	ST	0.07397	0.07432	0.07504
0.2	SU	0.09177	0.09322	0.09229
	H	0.07922	0.08017	0.08026
	ST	0.07963	0.08044	0.08073

Table 2. P values of paired t-test of the RMSEs obtained from the 1000 artificial chromosomes after wavelet denoising.

σ	Thres- hold	zpd vs	zpd vs	sym vs
		sym	per	per
0.1	SU	1.17×10^{-12}	2.23×10^{-10}	0.49
	H	0.15	0	4.89×10^{-11}
	ST	0.16	0	0
0.125	SU	3.70×10^{-14}	1.90×10^{-7}	0.0153
	H	6.42×10^{-3}	0	8.34×10^{-7}
	ST	0.30	0	2.68×10^{-12}
0.15	SU	0	4.85×10^{-5}	1.14×10^{-5}
	H	9.61×10^{-7}	3.00×10^{-15}	4.55×10^{-3}
	ST	0.0127	0	1.20×10^{-8}
0.175	SU	0	9.16×10^{-8}	2.11×10^{-5}
	H	2.42×10^{-5}	3.90×10^{-14}	5.61×10^{-3}
	ST	1.74×10^{-3}	0	3.14×10^{-7}
0.2	SU	0	4.97×10^{-3}	1.31×10^{-6}
	H	1.00×10^{-15}	0	0.56
	ST	4.99×10^{-12}	0	0.0471

5. CONCLUSIONS

In this paper, we empirically compared the effect of different signal extension methods on the performance of wavelet denoising of array CGH data. Experimental results suggest that the zero-padding method is the best for this application.

ACKNOWLEDGMENTS

The author would also like to acknowledge Jane Fridlyand for her assistance with providing parameters about UCSF CGH arrays used in the generation of synthetic data. This work was partially supported by the J. Lindsay Embrey Trustee Assistant Professorship from SMU.

References

1. BEHESHTI, B., BRAUDE, I., MARRANO, P., THORNER, P., ZIELENSKA, M., AND SQUIRE, J. A. Chromosomal localization of DNA amplifications in neuroblastoma tumors using cDNA microarray comparative genomic hybridization. *Neoplasia* 5, 1 (2003), 53–62.
2. DONOHO, D. L., AND JOHNSTONE, I. M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81, 3 (1994), 425–455.
3. DONOHO, D. L., AND JOHNSTONE, I. M. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* 90, 432 (1995), 1200–1224.
4. DONOHO, D. L., JOHNSTONE, I. M., KERKY-ACHARIAN, G., AND PICARD, D. Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 2 (1995), 301–369.
5. EILERS, P. H., AND DE MENEZES, R. X. Quantile smoothing of array CGH data. *Bioinformatics* 21, 7 (2005), 1146–53.
6. HSU, L., SELF, S. G., GROVE, D., RANDOLPH, T., WANG, K., DELROW, J. J., LOO, L., AND PORTER, P. Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics* 6, 2 (2005), 211–26.
7. LAI, W. R., JOHNSON, M. D., KUCHERLAPATI, R., AND PARK, P. J. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* 21, 19 (2005), 3763–70.
8. PERCIVAL, D. B., AND WALDEN, A. T. *Wavelet methods for time series analysis*. Cambridge University Press, Cambridge, UK, 2000.
9. PINKEL, D., AND ALBERTSON, D. G. Array comparative genomic hybridization and its applications in cancer. *Nat Genet* 37 Suppl (2005), S11–S17.
10. SNIJDERS, A. M., SCHMIDT, B. L., FRIDLYAND, J., DEKKER, N., PINKEL, D., JORDAN, R. C., AND ALBERTSON, D. G. Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma. *Oncogene* 24, 26 (2005), 4232–42.
11. STRANG, G., AND NGUYEN, T. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, Wellesley, MA, 1996.
12. WILLENBROCK, H., AND FRIDLYAND, J. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* 21, 22 (2005), 4084–91.