COMPREHENSIVE STUDY OF DNA COPY NUMBER ANALYSIS USING SIGMA FILTER

Abdullah K. Alqallaf and Ahmed H. Tewfik

Department of Electrical and Computer Engineering, University of Minnesota
200 Union Street. SE, Minneapolis, MN 55455, USA
alqal001@umn.edu, tewfik@umn.edu

ABSTRACT

DNA copy number aberrations are characteristic of many genomic diseases including cancer. Microarray-based Comparative Genomic Hybridization (aCGH) is a recently developed high-throughput technique used to detect DNA copy number (DCN) aberrations. Unfortunately, the observed copy number changes are corrupted by noise, making aberration boundaries hard to detect. In the first part of this paper, we propose a novel technique to analyze DCN aberrations based on the Sigma filter algorithm. We establish its superior performance for denoising DCN data and low computational complexity as compared to previous techniques. We present a comparison study between our approach and other smoothing and statistical approaches, the wavelet-based, LookAhead, CGH segmentation and HMM. We provide examples using real data to illustrate the performance of the algorithms. In the second part of this paper, we extend our algorithm by considering the effect of nonuniform physical distance between the probes in the aCGH data. Finally, we provide simulated and real data examples to study this effect.

1. INTRODUCTION

DNA copy number aberrations are associated with the development and progression of many genomic diseases including cancer, where amplifications and deletions of DNA copy number can contribute to variations in the expression of oncogenes and tumor suppressor, respectively. Microrray-based Comparative Genomic Hybridization (aCGH) is an approach for genome-wide scanning of differences in DNA copy numbers (DCN). It provides a high-resolution method to map and measure relative changes in DNA copy number simultaneously at thousands of genomic loci. By mapping the genomic locations of the genes responsible of genetic defects, it will be easier to characterize the genomics diseases as well as identify the targets for therapeutic involvement.

Generally, Microarray experiments contain many sources of errors due to human factors, array printer performance, labeling, and hybridization efficiency [7]. One should therefore consider denoising the data as a pre-processing step to uncover the true DCN changes before drawing inferences on the patterns of aberrations in the data samples. Smoothing techniques are particularly suitable for data denoising as they do not require a parametric model to find structures in the data. We review several of these techniques in the next section. In this paper, we use the Sigma filter algorithm to denoise the raw data because it has low computational complexity and useful properties for breakdown point detection. It is particularly well suited for handling the variations in aCGH data. Simulation studies show that denoising data prior to testing can achieve greater power in detecting the aberration regions than using the raw data without denoising. We present a comparison study between efficient techniques including the Sigma filter [10,11], Wavelet-based [3,4], LookAhead [1], CGH segmentation [9], and HMM [2]. We illustrate these methods on a typical Bacterial Artificial Chromosome (BAC) arrays.

In the second part of this paper, we extend our algorithm by considering the nonuniform spacing distance between the probes. In addition, we use a multidimensional Sigma filter to process the DNA copy number profiles. Finally, we demonstrate the extended algorithm using simulated and real data examples. The rest of this paper is structured as follows: Prior work is presented in section 2. In section 3, we introduce one-dimensional Sigma filter algorithm and the aberration boundaries detection scheme. A comparison study of proposed algorithms is given in Section 4. Section 5 is devoted for presenting the Multi-dimensional Sigma filter with the consideration of physical distances between the probes of the DCN data. Finally, section 6 is a conclusion of our observed results.

2. PRIOR WORK

Generally, DNA Copy Number (DCN) variation detection techniques fall into two categories: statistical model based approaches and smoothing techniques. In the statistical model based algorithms, the noise free signal and noise models are required. Unfortunately, these models are usually unknown or impossible to describe adequately with simple random processes. As a result, the important details (boundaries) of the DCN aberration regions will be included in the smoothing process. In addition, the techniques are computationally costly. The lookAhead algorithm [1] presents the DCN data as simple form of optimization problem over real-valued vectors of signal. It seeks to maximize the scoring function over all subintervals in the input vector. The CGH segmentation [9] is a likelihood method that avoids underestimating the number of segments in the data by using different penalty functions. It is pointed out that a homogeneous variance assumption among different regions can have important consequences in the model. A different kind of modeling approach involves the use of Hidden Markov Models (HMMs) [2], in which the underlying copy numbers are the hidden states with certain transition probabilities. Smoothing techniques provide an alternative method for processing the DCN data, that are characterized by small and long intervals with of sharp transitions and singularities at edges (starting and ending points). In these methods, local operators are applied to the noisy data. Only those points in a small local neighborhood are involved in the computation. The main advantage of these techniques is their computational efficiency. They can process the data in parallel without waiting for their neighboring points to be processed. In a one-dimensional wavelet-based denosing algorithm, the wden function of [3,4] is used, with the choice of Haar wavelet, since it has the shortest support among all orthogonal wavelets and soft Stein's Unbiased Risk Estimate (SURE) thresh-

olding rule with maximum wavelet coefficient level of 3. More recent study [14] used wavelet footprints to obtain a basis for representing the DCN data that is maximally sparse then Sparse Bayesian Learning is applied to infer the copy number changes from noisy array probe intensities.

3. SIGMA FILTER APPROACH

In this paper, we present the Sigma filter algorithm as a local smoothing technique. It is motivated by the sigma probability of the Gaussian distribution. It is conceptually simple but effective noise smoothing technique. The Sigma filter algorithm had been used in image processing as two-dimensional image smoothing tool to preserve the high intensity pixels (edges) and to smooth the low intensity pixels [10,11]. The basic idea is to replace the center point to be processed by the average of only those neighboring points having their intensities within a fixed sigma range of that point excluding the points out of that range. The advantage of this discriminative feature comparing to the other techniques has great impact on preserving the boundary points (edges) of the aberration regions and smoothing only the points in the neighboring of the boundary edges with low variance. Here we present the Sigma filter as a one-dimensional local smoothing technique to denoise the DCN data.

3.1. ONE-DIMENSIONAL SIGMA FILTER

A good model for describing aCGH data is:

$$y[i] = f[i] + \varepsilon_i . \qquad i=1,2,\ldots,N \qquad (1)$$

Let $\boldsymbol{y} = \{y[i]$ be a vector of size N which represents the observed intensities of aCGH data in our case and $\boldsymbol{f} = \{f[i]\}$ is the true relative intensity vector of aCGH data with the same size. We assume that ε_i is an additive white Gaussian noise with zero mean and standard deviation σ. The one-dimensional *Sigma filter* output $\hat{\boldsymbol{y}}$ vector is the smoothed intensities used to estimate the true relative intensity vector $\boldsymbol{f}$ from the observation vector $\boldsymbol{y}$. The one-dimensional *Sigma filter* procedure is described as follows:

1. Start with the choice of the siding window W, where W is the window size of odd length.
2. Select the intensity range $(y_i-\Delta, y_i+\Delta)$, where $\Delta=2\sigma$ is used as a threshold. Set

$$\delta_k = \begin{cases} w, & \text{if } (y_i - \Delta) \le y_k \le (y_i + \Delta) \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

   where w is a weighting function centered at y_i. The obvious choice is w=1.
3. Sum all copy number intensity values that lie within the intensity range in the window, i.e, calculate

$$L = \sum_{k=i-n}^{n+i} \delta_k . \qquad (3)$$

4. Select a constant value K such that K is less than the half of the sliding window, i.e., a sliding window with 7 points, K should be less than 4.

   If $L > K$, then compute the average by dividing the sum of the number of intensities that lie in that window.

$$\hat{y}_i = \frac{\sum_{k=i-n}^{n+i} \delta_k y_k}{L}. \qquad (4)$$

   If $L \le K$, then $\hat{y}_i =$ the average of the four immediate neighbors.
5. Continue until the entire DCN data is processed.
6. Relax the threshold Δ by a small amount (e.g. 10%) and go back to step 2 for the next iteration.

Note that the value of K should be carefully selected to remove the isolated spot noise without destroying thin features and subtle details. The threshold range of two-sigma is generally large enough to include 95.5% of the DCN intensities from the same distribution in the window, yet in most cases it is small enough to exclude the DCN intensities representing high-contrast edges of the aberration regions.

3.2. ABERRATION BOUNDARIES DETECTION

After denoising the DCN data, we apply the following scoring scheme to detect the boundaries of the aberration intervals. Assign a score for each interval $I \subseteq N$ with respect to the baseline (the x-axis) which is the average of this interval I if it pass a given threshold value τ as follows:

1. Start at i=0.
2. Scan for a value $\hat{y}_i$ such that $|\hat{y}_i| > \tau$. Let the location of this value be $i=i_o$.
3. Go back to $i=i_o$-4.
4. Evaluate the scoring average

$$S(I) = \sum_{n=i_o-4}^{I} \hat{y}_i \Big/ (I+4). \qquad (5)$$

   if $|S(I)|>\tau$ then increment I by 1 and go back to 4.
   if $|S(I)|<\tau$ stop and go to step 5.

5. We will stop adding elements to that interval if the score is below the threshold value. If this score S of interval I exceeds the threshold value τ, then $\hat{y}(I)$ is marked as an aberrant interval as follows:

$$\left\{ \begin{array}{ll} \text{H1: } \textit{Amplification interval} & \textit{if } S(I) \ge \tau \\ \text{H2: } \textit{Normal interval} & \textit{if } |S(I)| < \tau \\ \text{H3: } \textit{Deletion interval} & \textit{if } S(I) < -\tau \end{array} \right\}$$

6. Go back to step 2 starting at I+1.
7. Continue until the entire DCN data is processed.

4. COMPARISON OF ALGORITHMS USING REAL DATA (*CORIEL CELL LINES*)

In this section, we compare various algorithms for processing aCGH data. Obviously, it is hard to evaluate various algorithms with different parameters, but we can evaluate their performance based on the output results for detecting the DCN variation regions using the Receiver Operating Characteristic (ROC) curves.

To generate the ROC curves for each noisy signal, we calculated the true positive (TPR) and false positive rates (FPR) for various threshold levels. These levels varied from the minimum log-ratio value to the maximum. We defined the TPR and the FPR as follows:

$$TPR = \frac{Number\ of\ probes\ \ \geq \tau\ inside\ the\ alternation\ regions}{total\ number\ of\ probes\ inside\ the\ alternation\ regions}. \quad (6)$$

$$FPR = \frac{Number\ of\ probes \geq \tau\ outside\ the\ alternation\ regions}{total\ number\ of\ probes\ outside\ the\ alternation\ regions}. \quad (7)$$

Each threshold value results in a TPR and a FPR, represented by a point on the ROC curve.

4.1. Material

In this section, we used the MDA-MB-453 sample data of the Bacterial Artificial Chromosomes (BACs) array cell lines [8] to demonstrate the ROC curves of the algorithms under study. We used this data set because the genomic alterations were previously characterized by cytogenetics that is the true copy number changes are known for these cell lines which are easily detectable by manual inspection of the profiles, so that we can use it as a proof of principles. The other reason to use this data set is that it has been analyzed by other methods, namely, the HMM method [2] which we used as a reference for the real data example. Missing data (empty cells) of the *Coriel* cell line data sets were filled with the nearest neighboring average.

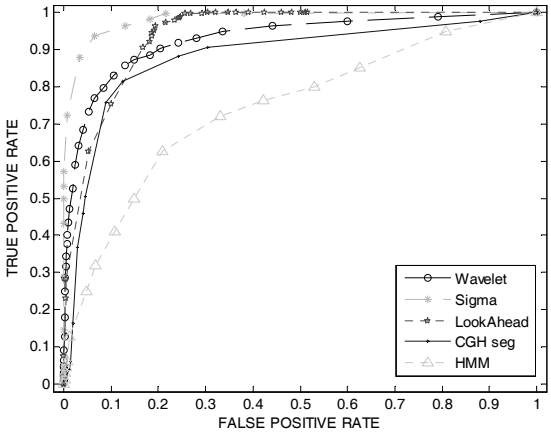

Figure 1. The receiver operating characteristic (ROC) curves for MDA-MB-453 data sample of Coriel cell line of *Sigma filter*, *Wavelet*, *LookAhead*, *CGH segmentation*, and *HMM* algorithms.

4.2. Results

4.2.1. Performance

From the ROC curves for MDA-MB-453 data sample sets of *Coriel cell lines* as shown in Figure 1, all algorithms did well for detecting the aberration regions of this data since it has a high-resolution except for the *HMM* technique. As the threshold level increases, we observed that the smoothing techniques and especially the *Sigma filter* gave better detection results than the statistical models (high TPR values at low FPR).

4.2.2. Computational Complexity

The other measure of performance was the algorithm run-time. The statistical models, *HMM*, *CGH segmentation* and *LookAhead* algorithms have higher order complexities comparing with the smoothing techniques, *Sigma filter (Sfilt)*, and *Discrete Wavelet Transform (DWT)* algorithms. If we have $\boldsymbol{N}$ data points corresponding to the number of clones (probes) in the input data set. The *CGH segmentation* and the *HMM* algorithms use $\boldsymbol{O}(\boldsymbol{N}^2)$ which is the complexity of dynamic programming, that will segment the data into a fixed number of segments. The *LookAhead* algorithm uses $\boldsymbol{O}(\boldsymbol{N}^{1.5})$. The *Sigma filter* and *Discrete Wavelet Transform* algorithms use $\boldsymbol{O}(\boldsymbol{N})$.

5. EXTENDED SIGMA FILTER

In this section, we extend our algorithm from one-dimensional to multi-dimensional Sigma filter to process the DNA copy number profiles to take advantage of the parallel computation technique to further increase speed. In addition, we consider the nonuniform physical distance between the probes. Finally, we demonstrate the extended algorithm using simulated and real data examples.

5.1. Method

5.1.1. Multi-dimensional Sigma filter

The extension of one dimensional *Sigma filter* to multi-dimensional cases is quite straightforward. All aspects of the one-dimensional *Sigma filter* remain the same except that the algorithm is applied to a two-dimensional neighborhood instead of a one-dimensional interval. The only difference in this case is the standard deviation σ used in the threshold function ($\Delta=2\sigma$) which will be the standard deviation of the whole genome. The main reason for using the two-dimensional *Sigma filter* is that it allows the parallel processing the chromosomes DCN data without waiting for their neighboring points to be processed which is computationally efficient. This is possible because the noise level is the same for all chromosomes since the data profiles are collected from the same source of errors and same organism. This reduces the chance of false positive or/and false negative points. Let,

$$\boldsymbol{Y} = \boldsymbol{F} + \boldsymbol{\epsilon}. \quad (8)$$

Figure 2. Multi-dimensional *Sigma filter*.

In the above equation, $\boldsymbol{Y}$ is the observed DCN data matrix of size (MxN) and $\boldsymbol{\epsilon}$ is the matrix of additive white Gaussian noise and corresponds to each chromosome vector. In Figure 2, $\hat{\boldsymbol{Y}}$ is the output matrix of the multi-dimensional *Sigma filter* used to estimate the true relative intensity matrix $\boldsymbol{F}$ of the same size, where M is the number of chromosomes in process and N is the number of probes of each chromosome. In other word, (M•N) is the length of the whole genome in process.

5.1.2. The effect of physical distance

Most prior works considered the DNA copy number profiles as discrete signals under the assumption that the probes are uniformly distributed along the chromosomes. This assumption may lead to wrong decisions with false positive or/and false negative points.

More recent studies [12,13] show that considering the nonuniform spacing distance between the probes of the DCN data profiles could be beneficial for detecting and measuring the DNC variations. In this section, we use the smoothing techniques, *Sigma filter* (*Sfilt*) and *Discrete Wavelet Transform* (*DWT*), to address the effect of considering the nonuniform physical distance between the probes

because they are computationally efficient and the Receiver Operating Characteristic (ROC) analysis of the first part of this paper confirms the superior performance of smoothing techniques over statistical models. We present the comparison study using simulated and real data examples.

As we have discussed in the DCN data model in the first part of this paper, here we present the DCN profiles as nonuniformly distributed discrete signals which can be modeled as follows:

$$y[x_i] = f[x_i] + \varepsilon_i. \qquad i=1,2,\ldots,N \qquad (9)$$

,where x_i in this case is the nonuniform distributed probe at i^{th} location along the x-axis. The x_i's are not uniformly distributed and the distance between two adjacent probes x_i and x_{i+1} may vary randomly. The $y[x_i]$ and $f[x_i]$ are the observed and true intensities of the DCN data probe location x_i. The ε_i represent independent identically distributed random variable from the Gaussian distribution of zero mean and σ^2 variance. Our goal here is to estimate the true relative intensities $f[x_i]$'s such that the smoothed intensities $\hat{y}[x_i]$ have small root mean square errors (RMSE's). We calculated the Root mean Square Errors (RMSE) values to evaluate the performance of different algorithms output for the simulated data profiles as follows:

$$err = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}[x_i] - f[x_i])^2}{N}}. \qquad (10)$$

The suggested procedure to solve this problem can be summarized as follows:

1. Insert uniform markers between the original probes using the average distance between the adjacent probes as a guideline for the spacing of the artificial markers. Let p be the average of the spacing distance between the adjacent probes. Let,

$$p = \frac{\sum_{i=1}^{N-1}(x_{i+1} - x_i)}{N-1}, \qquad (11)$$

and $\tilde{U}=\{\tilde{u}_j | \tilde{u}_j = kp,\ k=1,2,\ldots,\lfloor x_N / p \rfloor\}$ be the set of locations at which we may insert markers along the chromosome. To avoid overlapping the new markers with the original markers, we only insert markers at location in a subset U of $\tilde{U}$, such that:

$$U=\{u_j: u_j \in \tilde{U},\ |u_j - x_i| \geq p/2 \text{ for all } i=1,2,\ldots,N\}.$$

2. Apply the nearest neighbor interpolation to obtain the interpolated data $Y(u_j)$. Here we use the nearest neighbor interpolation instead of other interpolation methods such as linear, or spline interpolations because the aCGH data follow a piece-wise constant function.
3. Merge the original data and the interpolated data to obtain the merged data $\tilde{Y}$ such that:

$$\tilde{Y}=\{Y[x_i] : i=1,2,\ldots,N\} \bigcup \{Y[u_j] : u_j \in U\}.$$

4. Process the interpolated data using the Multi-dimensional *Sigma filter*.
5. Apply the scoring scheme of section 3 to detect the boundaries of the aberration intervals.

Note that for a fair comparison the locations of the original probes are changed. New artificial markers are simply added along the chromosome in between the original nonuniform probes. When comparing the denoising results, only the values in the original probes are used.

5.2. Material

5.2.1. The generation of realistic simulated data

We will use simulated data initially to compare the true signals with the denoised signals. For this, we use the same procedure used in [12] to generate simulated DCN profiles which is an extended design of [13]. In [13] the authors assume the probes are uniformly distributed along the chromosomes. In [12] they extended the model by placing nonuniform probes by randomly distribution. The process of generating the simulated data can be summarized as follows:

1. Construct the true DCN profile based on the distribution of a real data example.
2. Add white Gaussian noise with zero mean and specified noise level σ.
3. Pick the noisy DCN profile randomly at different probe distance locations based on the real nonuniform distance distributions between the probes of the real data.

We call it realistic data generation model because it uses the distributions of real DCN profiles with different noise levels and random probe locations. Following the same data model suggested by [13] which assumes the chromosomal segmentations of DCN 0, 1, 2, 3, 4, and 5 were generated with probabilities of 0.01, 0.08, 0.07, 0.02, and 0.01 respectively. The length of each chromosome sample is determined by random sampling from the corresponding length distribution of real DCN profiles. Each sample was assumed to be a mixture of tumor and normal cells. The proportion of tumor cells P_t was drawn from a uniform distribution between 0.3 and 0.7. The expected relative true $\log_2$ ratio intensity level of the DCN profiles computed as $\log_2((c\ P_t+2(1-P_t))/2)$, where c is a constant. Next a white Gaussian noise with zero mean and specified noise level σ^2 added to the relative true data to generate the noisy DCN profiles. Then, from the uniformly distributed noisy data we randomly pick the probes according to the distribution obtained by [13]. Thus, the physical distances between the probes are randomly distributed.

We use this model to generate 100 simulated DCN profiles with different noise levels σ^2 of 0.1, 0.15, and 0.2 and 200 Mb lengths.

5.3. Results of simulated and real data profiles

In this section, we demonstrate the effect of considering the nonuniform distances between the probes on the denoising techniques *Sigma filter* and *Discrete Wavelet Transform* using simulated and real data examples.

5.3.1. Simulated data results

Table 1 shows that the average of the root mean square errors (RMSEs) values measured by (10) of 100 simulated data sets generated randomly as mentioned in the previous section at different noise levels using the *Sigma filter* and *Discrete Wavelet Transform* algorithms with and without considering the effect of nonuniform spacing distance between the probes of simulated DCN data profiles.

σ	Sigma filter	DWT	Sigma filter	DWT
	Without the physical distance effect		With the physical distance effect	
0.1	0.0254	0.0325	0.0248	0.0297
0.15	0.0507	0.0521	0.0345	0.0515
0.2	0.0672	0.0753	0.0621	0.0722

Table 1. The average of root mean square errors (RMSE's) values for 100 simulated DCN data with 3 different noise levels using *Sigma filter* and *DWT* algorithms.

From the results of the RMSE's values in table 1, we can observe that on average the *Sigma filter* algorithm outperformes the *Discrete Wavelet Transform* approach. In addition, The *Sigma filter* algorithm that consider the nonuniform spacing effect between the probes achieved better performance than the variant that does not consider it.

5.3.2. Real data examples

In this section, we present real data examples used in the first part of this paper with the consideration of the effect of the nonuniform spacing distribution between the probes in the aCGH data to confirm our observation results of the simulated examples. We processed the MDA-MB-453 data sample of Coriel cell line using the Multi-dimensional *Sigma filter* algorithm.

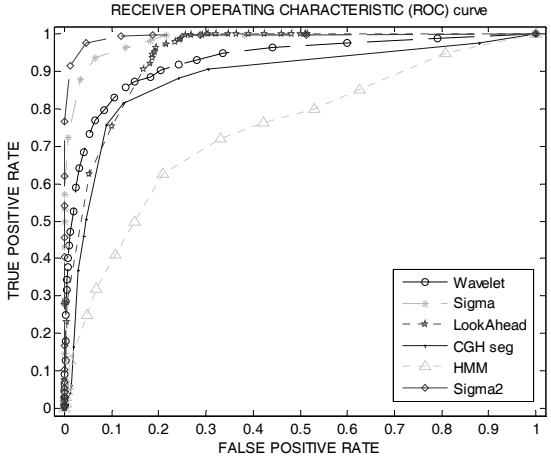

Figure 3. The receiver operating characteristic (ROC) curves for MDA-MB-453 data sample of Coriel cell line of *Sigma filter*, *Wavelet*, *LookAhead*, *CGH segmentation*, *HMM*, and *Sigma2 filter* algorithms.

Figure 3 shows the ROC curves for MDA-MB-453 data sample of Coriel cell line for the proposed algorithms. One can obviously see the advantage of considering the Multi-dimensional *Sigma filter* and the nonuniform spacing distance between the probes of the chromosomes on the output result of the ROC curves of *Sigma2* algorithm.

6. CONCLUSION

In this paper, we investigated the performance characteristics and complexities of various algorithms. The comparison study shows that our proposed algorithm, *Sigma filter*, works very efficient due to the discriminative characteristic of preserving edges. The comparison also shows that the *Sigma filter* algorithm is computationally efficient compared to the other evaluated algorithms. In the second part of the paper we show that the *Sigma filter* can achieve even better smoothing capabilities by considering the effect of the nonuniform physical distance between the probes of the aCGH data.

REFERENCES

[1] Doron L., Yonatan A., Amir B., Nathan L., and Zohar Y. (2006). Efficient Calculation of Interval Scores for DNA Copy Number Data Analysis. Journal of Computational Biology, Pp. 215-228.

[2] Fridlyand J., Snijders A., Pinkel, D., Albertson D. G. and Jain, A. N. Application of Hidden Markov Models to the analysis of the array CGH data. (Special Genomic Issue of Journal of Multivariate Analysis, June 2004, V. 90, pp. 132-153)

[3] David Donho et al, WAVELAB 802, "http://www.stat.stanford.edu/~wavelab/".

[4] Stephane Mallat, A Wavelet Tour of signal processing, Academic Press, San Diego, January 1998.

[5] Wang, P., Kim, Y., Pollack, J., Narasimhan, B. and Tibshirani, R. (2004). A method for calling gains and losses in array CGH data. Biostatistics (in press).

[6] Olshen, A. and Venkatraman, E. (2002). Change-point analysis of array-based comparative genomic hybridization data. ASA Proceedings of the Joint Statistical Meetings, 2530–2535.

[7] Churchill GA. Fundamentals of experimental design for cDNA microarray. Nat Genet Sup. 2002;32:490. doi:10.1038/ng1031.

[8] Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J. P., Gray, J. W., Jain, A. N., Pinkel, D. and Albertson, D. G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number by CGH. Nature Genetics 29, 263–264.

[9] Franck P., Stephane R. Marc L., Christian V., and Jean-Jacques D. A statistical approach for array CGH data analysis. BMC Bioinformatics 2005, 6:27 doi:10.1186/147-2105-6-27.

[10] J.S. Lee, "Digital Image Smoothing and the Sigma Filter," CVGIP, Vol 24, pp. 255-269, 1983.

[11] C. H. Kuo and A. H. Tewfik, "Multiscale Sigma Filter and Active Contour for Image Segmentation" , ICIP,1999.

[12] Wang, Y. and Wang S. (2007) `A novel stationary wavelet denoising algorithm for array-based DNA copy number data', Int. J. Bioinformatics Research and Applications, available at "http://engr.smu.edu/~yuhangw/papers/waveletdenoising_dcn.pdf".

[13] Willenbrock, H. and Fridlyand, J. (2005) `A comparison study: applying segmentation to array CGH data for downstream analyses', Bioinformatics, Vol. 21, No. 22, pp.4084-4091.

[14] Pique-Regi R, Tsau ES,Ortega A, Seeger RC, Asgharzadeh S: "Wavelet footprints and sparse Bayesian learning for DNA copy number change analysis ", in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Hawaii 2007, available at "http://biron.usc.edu/~piquereg/papers/SBL_WF_ICASSP07.pdf"