A NOVEL ALGORITHM FOR THE ANALYSIS OF ARRAY CGH DATA

Mehrnoush Khojasteh[1, 2], Bradley Coe[1], Sohrab Shah[3], R. K. Ward[2], W. L. Lam[1], Calum MacAulay[1]

[1] British Columbia Cancer Research Center, Vancouver, BC, Canada
[2] Electrical & Computer Eng. Dept., [3] Computer Science Dept., Univ. of British Columbia, BC, Canada

ABSTARCT

DNA copy number aberrations are common in cancer and other diseases. Newly developed array CGH technologies enable simultaneous measurement of DNA copy numbers for tens of thousands of sites within a genome. In array CGH experiments, DNA copy number of a test DNA sample relative to the DNA copy number of a reference DNA sample is measured. These relative measurements are mapped to their corresponding chromosomal locations. DNA copy gains and losses are then detected as deviations from the normal reference at specific chromosomal locations. In this paper, we introduce a novel algorithm to automatically identify the regions of DNA copy number gain and loss from array CGH data through a multi-scale edge detection algorithm. We demonstrate the method on two array CGH datasets. Our results show that this method can be successfully applied for the analysis of array CGH biological data.

1. INTRODUCTION

The DNA copy number of a region within a genome is the number of copies of the region's DNA sequence present in the genome being measured. In a normal diploid cell, the normal copy number is two for all the autosomes (non-sex chromosomes). Variations in genomic copy number are common in cancer and other diseases. These variations are a result of genetic events causing discrete gains and losses in contiguous segments of the genome. For this reason, efforts have been made over the last ten years to generate whole genome copy number maps from a single experiment. Microarray-based comparative genomic hybridization (array CGH) provides a means to quantitatively measure DNA copy number aberrations and map them directly onto known genome sequences [2].

In an array CGH experiment, a DNA sample of interest, called the test sample, and a diploid reference sample are first labeled with different dyes and then mixed. This combined sample is then hybridized to the microarray and imaged. This results in test and reference intensities for each DNA clone on the array. Since it is generally assumed that the reference sample does not have any copy number aberrations, clones with test intensities significantly greater than the reference intensities are indicative of copy number gains in the test sample at those positions. Similarly, significantly lower intensities in the test sample are signs of copy number losses [2].

The analysis model that is typically used for these data is that the base-two logarithm of the ratio of the intensities for each clone in the experiment is considered as a function of that clone's location in its chromosome. An example of how a genomic profile may look is illustrated in Figure 2. The statistical methods for analyzing array CGH data are thus aimed at identifying regions of gains or losses of copy number within chromosomes. Downstream analyses involve classifying the samples and finding copy number alterations that are associated with known biological markers.

In the analysis of array CGH data, broadly, there are two estimation problems. One is to identify the locations in which copy number transitions or "breakpoints" occur and thus identifying the regions of gain or loss in DNA copy number; the other one is to infer the statistical significance of the identified regions. Several methods have been proposed for portioning clones into sets with the same copy number. These methods include a hidden Markov model approach [9], a non-parametric change-point method [1], and a wavelet smoothing approach [10]. Two comparison studies of different methods are found in [8] and [13]. While various algorithms have been developed for analysis of high throughput array CGH data, there is yet no proven best method for data analysis. Different approaches typically suffer from one or all of the following drawbacks: imposing strong distribution assumptions on the array CGH data which generally are not true, need of tuning a set of parameters, minimal consideration about the spatial structure of the data including the distance between clones, length of the clones, and clone overlaps.

In this paper, we propose using edge detection as a novel approach to locate breakpoints in the copy number data. The edge detection framework is a natural approach to this problem as it estimates the locations along the chromosome where abrupt changes (edges) in the copy number occur. These breakpoints split the chromosomes into regions of equal copy number while accounting for the noise in the data. We then apply a nonparametric statistical test to determine the statistical significance of regions with altered copy numbers.

This paper is organized as follows: In Section 2 we describe the edge detection algorithm used. Section 3 presents the detailed structure of our proposed algorithm. In Section 4 we present the result of analysis on two array CGH data sets. Conclusions are discussed in Section 5.

2. EDGE DETECTION

Typically in signal and image processing, an edge detector smoothes the signal at some scale and then detects sharp variation points from the local maxima of the first derivative. However, one smoothing scale may not be enough to remove noise while preserving good localization and instead edge detection must be performed at different scales of resolution [6]. The wavelet transform can focus on localized signal structures with a zooming procedure that progressively reduces the scale parameter and is thus suited to our particular edge detection problem [12].

2.1. Wavelet Analysis

The wavelet transform of $f(x)$ with respect to wavelet $\psi(x)$ is

$$W_s f(x) = f * \psi_s(x), \text{ where } \psi_s(x) = \frac{1}{s}\psi\left(\frac{x}{s}\right)$$

The term modulus maximum is used to describe any point at which $|W_s f(x)|$ is locally maximum. Singularities of $f(x)$ are detected by finding the locations where the wavelet modulus maxima converge at fine scales. If the wavelet function $\psi(x)$ is the first derivative of a smoothing function $\theta(x)$, the wavelet transform $W_s f(x)$ is a multi-scale differential operator. In this case, the wavelet modulus maxima are the maxima of the first order derivative of $f(x)$ smoothed by $\theta(x)$ [12]. If $\theta(x)$ is Gaussian, it is guaranteed that a modulus maximum at a scale always propagates towards finer scales. In other words, as scale decreases, additional modulus maxima may appear but existing ones can not disappear [3].

Linear smoothing with a Gaussian has two effects: 1) removal of fine-scale features and 2) dislocating the features that survive. The latter undesirable effect may be overcome by tracking coarse extrema to their fine scale locations. Thus a coarse scale may be used to identify extrema and a fine scale to localize them [3]. While coarse to fine tracking solves the problem of localizing the large scale events, it doesn't solve the multi-scale integration problem. In other words, the most appropriate upper bound scale to use (the coarsest scale) is still unknown. We introduce a method that based on a scale-space representation of the edges, adaptively finds the appropriate smoothing scale for each segment of the signal. The details of this method are described in Section 3.

3. PROPOSED ALGORITHM

3.1. Breakpoint Detection

Below we describe each step of our proposed procedure.

1. The clones, for which the copy numbers are measured, are unequally spaced along the chromosomes. Also some of the clones are overlapping. So the first step of our analysis is then to define a set of equally spaced data points from the original array CGH data. To achieve this, a set of 2^J points is defined, where J is the smallest integer so that $2^J \geq N$ and N is the number of clones in each chromosome arm. Each chromosome arm is analyzed separately because of the gap that exists between the clones from two arms of the chromosomes.

A piecewise constant interpolation is then applied to the data. The resulting signal is then sampled over the new set of points according to the following criteria:

$$f'(x_i) = \begin{cases} f(C_j) & \text{, if } x_i \text{ belongs to only one clone } (C_j) \\ \dfrac{\sum_j f(C_j)}{K} & \text{, if } x_i \text{ belongs to more than one clone } (C_j, j=1,2,\ldots,K) \\ f(C_j) & \text{, if } x_i \text{ doesn't belong to a clone but the closest clone to } x_i \text{ is } C_j \end{cases}$$

where C_j, $j=1,2,\ldots,N$ are the clones and $f(C_j)$ is the ratio measurement for clone C_j.

2. A discrete wavelet transform is computed at scales $s = a^j$, with $a = 2^{1/v}$, where v is the number of intermediate scales in each interval $[\,2^j, 2^{j+1})$. In our analysis, we used $j=1, 2, \ldots, J$ and $v=5$ (2^J is the length of the discrete signal).

The wavelet transform modulus maxima are computed across all the scales. Next, these maxima are chained together into maxima lines that start from larger scales and propagate all the way towards finest scale. WAVELAB software [5] is used to perform the wavelet analysis.

At this point, we have found edges at all the scales of the wavelet transform. These edges may be viewed as chains in the scale-space plane that start at some larger scale and propagate to the finest scale. An example is shown in Figure 1.

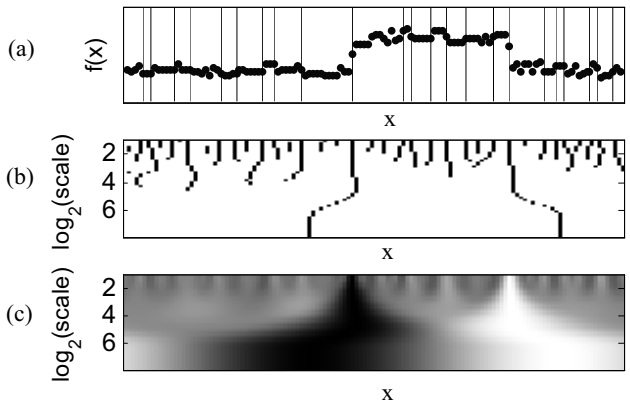

Figure 1. (a) signal and multi-scale detected edges, (b) modulus maxima of the wavelet transform chained together across the scales, (c) modulus of the wavelet transform across the scales

3. The next step is to select the appropriate scale or degree of smoothing which is equivalent to selecting the appropriate starting scale for the chains. So the following recursive algorithm is used:

I. The algorithm is initialized with the scale equal to the largest scale. The set of "true edges" is initialized with the chromosomal positions of first and last data points.
II. At each scale, the chains that start at that scale are located.
III. These chains are tracked to the finest scale to obtain the accurate position of their corresponding edges. These positions are called the "potential edges".
IV. The "potential edges" between an adjacent pair of "true edges" are located. For each pair of adjacent true edges:
   a. Based on the positions of the potential edges in between these true edges, "regions" of chromosome that are bound by two adjacent edges are defined.
   b. Each pair of adjacent regions is tested to find out if they are statistically different in the central tendency of their underlying data points.

c. If two regions are not significantly different, the common edge separating them is considered as a false edge and is removed.

V. The edges that were preserved at the previous step are added to the set of true edges. Next, the scale is reduced by one and algorithm goes back to step 2.

To compare each pair of adjacent regions, Wilcoxon rank sum test was used which tests the hypothesis that two independent samples come from distributions with equal medians. A significance level, α, is required by the test.

3.2. Determining the Statistical Significance of Altered Regions

The breakpoints split the chromosomes into regions of equal copy number while accounting for the noise in the data. Next, regions with altered copy numbers need to be identified. To do this, a set of clones are selected (as explained below) from the experiment that acts as a reference sample of non-altered clones. Next, each region is compared against that reference sample to test the hypothesis that the region being compared has a median equal to that of reference population.

The reference method is generated as follows: The median of the log ratios from each region is calculated. The central 5% range of the empirical distribution of these medians is then determined. Regions whose medians are in this range are selected. Data points belonging to these regions are considered as the reference sample. A permutation test was used to compare regions against the reference sample. A random sample with the same length as the region in question is selected from the reference sample many times (about 10,000 times). The median of the random sample is compared to the median of the sample in question each time. The p-value of the test is calculated as the number of random samples with medians equal or higher than the sample of interest. Regions whose corresponding p-values are less than a significance level determined by the user are considered as altered.

4. BIOLOGICAL VALIDATION

4.1. Coriell cell lines data

We demonstrate our approach on publicly available dataset by Snidjers et al [2]. The dataset consists of single array CGH experiments on 15 fibroblast cell lines. Each array contains 2276 mapped BACs (Bacterial Artificial Chromosomes). The true copy number changes are known for these cell lines and confirmed by spectral kariyotyping. So, this dataset is used as a proof of principle. This dataset has been previously used in [9], [1], and [10] to show performance of different methodologies.

There were a total of 8 alterations that covered the whole chromosomes and 15 alterations that covered parts of the chromosomes. Our method identified all of the known changes except for two cases where the change involves only one or two clones at one end of the chromosome (cell line GM03563 chromosome 9 and cell line GM01536 chromosome 4). Our methodology also found a number of changes not detected by spectral karyotyping. The number of false positives covering parts of the chromosomes in each cell line, at the significance level of 0.001, ranged from 2 to 14. As also mentioned in [1], these were because of local trends in the data. There were also false positives that covered the entire chromosome. These false positives were caused by improper normalization of data that allowed the average level of the $\log_2$ ratios for a chromosome to be significantly different from zero while the chromosome did not have any confirmed changes. Typical examples of one confirmed change and several false positives detected by our method can be found in Figure 2.

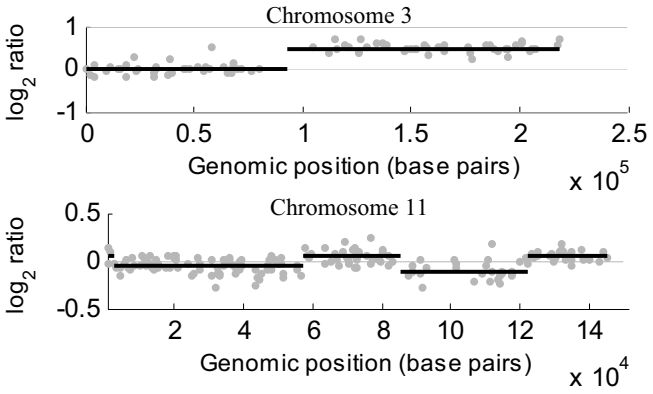

Figure 2. Analysis results for Coriell cell line GM03563, top: a confirmed change, bottom: "false positives".

4.2. H526 Lung Cancer Cell Lines data

We also applied our method to 7 replicate array CGH experiments comparing the lung cancer cell line H526 DNA to normal male genomic DNA performed on the Submegabase Tiling Resolution (SMRT) BAC microarray [4]. Each SMRT microarray consists of 26318 overlapping BAC clones spotted in duplicate. The variable used for the analysis was the normalized base 2 logarithm of the over reference ratio averaged for duplicate clones [11].

Overall visual inspection of the results showed quite similar results to the manual segmentation of data performed by experts.

As we did not have external confirmation of all the alterations observed in these datasets, we based our evaluation on the similarity of the analysis results from 7 independent replicate experiments. Since these datasets are obtained from replicate experiments, the true copy number changes are theoretically the same in all of them. At a fixed statistical significance level, each clone on each array is either identified as not-altered or is identified as Gained or Lost. The true state for each clone is assumed to be the state of the corresponding clone in the majority of the arrays. Having the true state for each clone, the number of true positives (clones identified as altered while actually altered) and false positives (the clones identified as altered while actually non-altered) can be determined for each array. The results from this analysis are reported in Table 1.

We also analyzed these data using DNAcopy [7] version 1.1.2 that is an implementation of the methodology introduced in [1]. According to the comparison studies [8] and [13], this method has the best performance among other methods being compared. This method, takes in a parameter alpha between 0 and 1 that controls the sensitivity of the segmentation. Increasing alpha leads to more breakpoints, and, possibly, more false positive breakpoints. We analyzed these data at varying levels of alpha from 0.01 to 0.9.

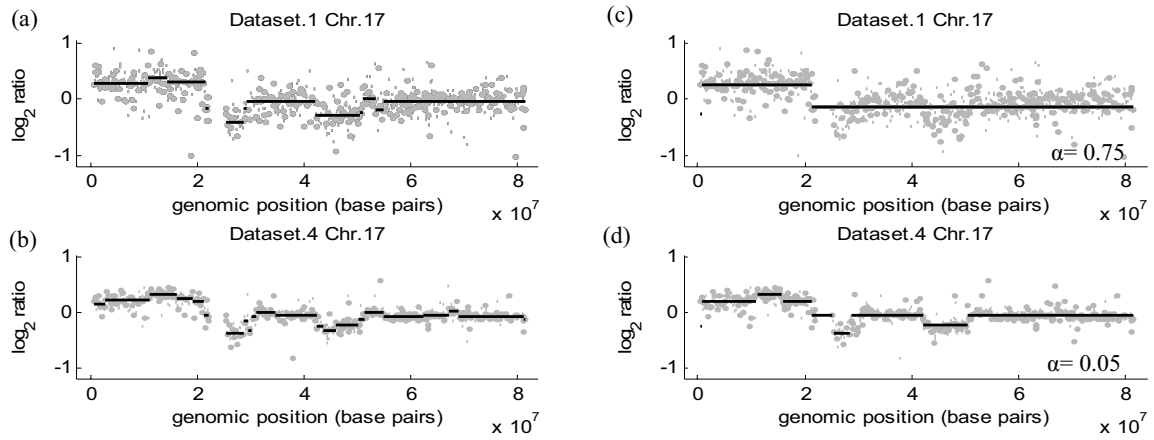

Figure 3. (a) and (b) results of our method, (c) and (d) DNAcopy results, dark lines represent segmented regions, (a) and (c) are the results for the noisy dataset and (b) and (d) are the results for the data set with lower noise.

Visual comparison results show that our method outperforms DNAcopy in detecting low level aberrations especially in noisy samples. As an example, Figure 3 shows the same chromosome in two datasets with different levels of noise analyzed by DNAcopy and by our method. DNAcopy performs well in the dataset with lower noise, however, in the other dataset, at alpha levels less than or equal 0.75 it fails to detect the lower level aberrations. Further increasing alpha in this case results in lots of false positive breakpoints. Our method, on the other hand, performs well in both datasets.

Dataset	TPR	FPR
1	85.8%	1.7%
2	93.7%	2.5%
3	93.3%	0.7%
4	98.9%	19.6%
5	97.2%	6.2%
6	96.4%	4.4%
7	99.2%	13.4%

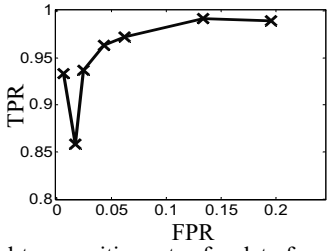

Table 1. (a) False positive and true positive rates for data from 7 H526 cell line datasets, (b) plot of true positive vs. false positive.

5. CONCLUSIONS AND DISCUSSION

A new methodology for the automatic identification of regions of copy number gain and loss from array CGH data is introduced. Our approach is based on multi-scale edge detection to locate breakpoints in the copy number data which split the chromosomes into regions of equal copy number while accounting for the noise in the data. The multi-resolution approach to finding the breakpoints is a natural framework that allows identification of copy number changes at different levels.

Some of the advantages of our proposed methodology are: 1) it does not impose a statistical distribution assumption on the data, 2) Unlike most methods, it doesn't need an arbitrary threshold or parameter, 3) it takes into account the spatial structure of the data including the distance between the clones and clone overlaps, 4) it reports the statistical significances of the identified altered regions.

We applied our procedure to public data from array CGH experiment with confirmed copy number changes. The procedure identified all but two expected alterations.

We also applied our methodology to data from a set of replicate array CGH experiments on lung cancer cell line H526 performed on SMRT microarrays. Our comparisons demonstrate that our method outperforms the state of the art methodology in analyzing these data. Unfortunately, there is no biological verification for the changes detected in the data from these experiments; However in spite of different degrees of noise in these data, alterations found in the data sets were mostly consistent with each other.

REFERENCES

[1] A.B. Olshen et al, "Circular binary segmentation for the analysis of array-based DNA copy number data", *Biostatistics*, Vol. 5, No. 4, pp. 557-572, 2004.

[2] A.M. Snijders et al, "Assembly of Microarrays for Genome-wide Measurement of DNA Copy Number", *Nat. Genet.* Vol. 29, No. 3, pp. 263-4. 2000.

[3] A.P. Witkin, "Scale Space Filtering", *Proc. Int. Joint conf. in Artificial intelligence*, 1983.

[4] A.S. Ishkanian et al, "A Tiling Resolution DNA Microarray with Complete Coverage of the Human Genome", *Nat. Genet.* Vol. 36, No. 3, pp. 299-303, 2004.

[5] D. Donoho et al, WAVELAB 802, www-stat.stanford.edu /~wavelab/, Oct. 2005.

[6] D. Marr and E. Hildreth, "Theory of Edge Detection", *Proc. Royal. Soc. London*, Vol. 207, pp. 187-217, 1980.

[7] E. S. Venkatraman and Adam B. Olshen, "A package for analyzing DNA copy data", http://bioconductor.org/repository/ devel/vignette/DNAcopy.pdf.

[8] H. Willenbrock et al, "A comparison Study: Applying Segmentation to Array CGH data for Downstream Analyses", *Bioinformatics*, Epub September 14, 2005.

[9] J. Fridlyand et al, "Hidden Markov models approach to the analysis of array CGH data", *Journal of multivariate analysis*, Vol. 90, No. 1, pp. 132-153, 2004.

[10] L. Hsu et al, "Denoising array-based comparative genomic hybridization data using wavelets", *Biostatistics*, Vol. 6, No.2, pp. 211-226, 2005.

[11] M. Khojasteh et al, "A Stepwise Framework for the Normalization of array CGH Data", *BMC Bioinformatics*, In press.

[12] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, Janurary 1998.

[13] W.R. Lai et al, "Comparative Analysis of Algorithms for Identifying Amplifications and Deletions in Array CGH Data", *Bioinformatics*, Vol. 21, No. 19, pp. 3763-70, 2005.