Biomedical Signal Processing and Control 31 (2017) 238–248

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Improving estimates of the breakpoints in genome copy number alteration profiles with confidence masks

Jorge Muñoz Minjares, Yuriy S. Shmaliy*

Department of Electronics Engineering, Universidad de Guanajuato, 36885 Salamanca, Mexico

ARTICLE INFO

Article history:
Received 12 November 2015
Received in revised form 8 March 2016
Accepted 18 August 2016

Keywords:
Copy number alterations
Jitter distribution
Breakpoints

ABSTRACT

Chromosomal structural changes in human body known as copy number alterations (CNAs) are often associated with disease such as cancer. Therefore, accurate estimation of the CNAs using high resolution technologies is on a front line of bioinformatics and engineering. Since the Laplace distribution recently justified to represent jitter in the CNA breakpoints is not sufficiently accurate to estimate small changes, we propose a more accurate approximation based on the modified Bessel function of the second kind and zeroth order. We develop the relevant confidence masks to bound the CNA estimates for the given probability. The masks are applied to test the estimates obtained using the profile copy number of single nucleotide polymorphism (SNP) array data. We also show how to improve the estimates for the required confidence probability by removing some unlikely existing breakpoints.

1. Introduction

It is well known that the deoxyribonucleic acid (DNA) of a genome essential for human life often demonstrates structural changes [1–3] called *genome copy number alterations* (CNAs) [4–6] which are associated with disease such as cancer [7]. Analysis of the breakpoint locations in the CNAs is still an important issue because it helps detecting structural alterations, load of alterations in the tumor genome, and absolute segment copy numbers. Thus, efficient estimators are required to extract information about the breakpoints with an accuracy acceptable for medical needs. To produce the CNAs profile, several technologies have been developed such as the comparative genomic hybridization (CGH) [8], High Resolution CGH (HR-CGH) [9], and Whole Genome Sequencing [10]. Most recently, the single nucleotide polymorphism (SNP) technology [11] showed a significant improvement in the detection of small genetic changes [12,13]. Even so and in spite of their high resolution, the modern methods still demonstrate the inability in getting good estimates of the breakpoint locations because of the following factors: (1) nature of biological material (tumor is contaminated by normal tissue, relative values and unknown baseline for copy number estimation), (2) technological biases (quality of material and hybridization/sequencing), and (3) intensive random noise. By virtue of this, probing is often provided at low signal-to-noise ratios (SNRs) that requires special approaches to refine the breakpoint location estimates. Although the exact source of the CNAs noise remains unclear, its certain correlation with the guanine-cytosine content (GC-content) along the genome and with other genomic features was shown in [15].

Unluckily, no one estimator – optimal or robust – is able to give a fully true picture in large noise [15]. The problem complicates by an inability of providing multiple probing in short time [16] and thus to improve the estimates statistically. Provided a single probing, an estimator cannot guarantee that all of the detected changes were estimated with a maximum probability. Accordingly, as shown in [17], some small changes may be diagnosed as unlikely existing if to test the estimates by the confidence masks. One may also suppose that some other small changes were not detected.

The SNP data are typically represented in SNP Index with the nth probe, $n_l \in [1, M]$, where M is the number of the probes [18]. In the CNAs structure, the ith edge (or *breakpoint*) corresponds to the n_lth discrete point. The CNA profile is represented by the Log R ratios (LRRs), which are the log-transformed ratios of the experimental and normal reference SNP intensities centered at zero for each sample [19].

The CNAs function demonstrates the following fundamental properties [20] which are of importance for the estimator design:

- It is piecewise constant (PWC) and sparse with a small number of alterations on a long base-pair length.
- Its constant values are integer, although this property is not survived in the log 2 ratio.

* Corresponding author.
E-mail address: shmaliy@ugto.mx (Y.S. Shmaliy).

http://dx.doi.org/10.1016/j.bspc.2016.08.009
1746-8094/

- The measurement noise in the log R ratio is highly intensive and can be modeled as additive white Gaussian.

An example of the CNAs probes with a single breakpoint and two segments is shown in Fig. 1a. Here, the lth segment a_l and $(l+1)$th segment a_{l+1} are represented with the noise standard deviations σ_l and σ_{l+1} and segmental difference $\Delta_l = a_{l+1} - a_l$ corresponding to the breakpoint at $n = 200$. The signal-to-noise ratios (SNRs) in the lth segment and $(l+1)$th segment can be specified as [21], respectively,

$$\gamma_l^- = \frac{\Delta_l^2}{\sigma_l^2}, \quad \gamma_l^+ = \frac{\Delta_l^2}{\sigma_{l+1}^2} \tag{1}$$

for supposedly constant segmental values.

What many issues face in practice of the CNA detection follows directly from Fig. 1a: *an intensive noise does not allow for an exact detection of the breakpoints and precise estimates of segmental levels*. In fact, accuracy in the estimates of the segmental levels is strongly affected by the segmental noise [22] which is known to be white Gaussian [20]. In turn, accurate detection of the breakpoint locations often becomes unavailable due to low segmental SNRs. As has been shown in [23] and reproduced in Fig. 1b, the problem complicates by jitter in the breakpoints which has the Laplace-like distribution.

In view of the aforementioned issues, many approaches have been developed during decades in order to provide denoising while preserving edges in such signals. Approaches with nonlinear wavelet-based processing with thresholding [24–26] have received a considerable interest because these are highly efficient in Gaussian noise. The nonlinear smoothers based on robust statistics were developed with this aim in [27–30]. Referring to the fact that time-variant linear structures are able to produce effects similarly to the nonlinear ones, adaptive and time-variant smoothers were suggested and investigated in [31–34]. Also, the forward–backward (FB) filters and smoothers were used [35,36].

In addition, several methods have been developed to refine the breakpoints. A solution to this problem has been proposed in [37] where the breakpoint detection was based on the Adaptive Weights Smoothing (AWS). Another method was developed in [38] to syntony blocks narrowing the breakpoint regions by aligning each breakpoint sequence against its orthologous sequences in other species. In [39] the authors have presented a computationally validated approach using local *de novo* assembly to gain a more comprehensive picture of the structural variants found within a genome.

Although the jitter distribution [23] in the breakpoints can be approximated with the skew discrete Laplace density shown as SkL in Fig. 1b [40,17], the Laplace-based distribution has appeared to be accurate enough only for the SNR values exceeding unity [22]. Otherwise, when the SNR is low, $\gamma_l^{\pm} < 1$, and when it is extremely low, $\gamma_l^{\pm} < 1$, the Laplace distribution becomes too rough. Accordingly, the confidence masks created based on the Laplace distribution narrow possible bounds of the estimated chromosomal changes and cannot efficiently be used to improve the estimates. A more correct probabilistic model of jitter in the breakpoints is thus required.

In this paper, we provide extensive investigations of jitter in the CNAs breakpoints, propose an approximation for the jitter distribution using the modified Bessel functions, and show that the approximation proposed fits the CNAs probes much better than the Laplace distribution for any reasonable SNR value of practical interest. We then introduce a statistical framework to compute the confidence lower boundary (LB) and upper boundary (UB) masks, test the CNA estimates by the masks for data obtained using the SNP technology, although it could be used any other, and show how to improve the CNAs estimates for the given confidence probability by removing some unlikely existing breakpoints. The rest of the paper is organized as follows. In Section 2, we discuss the jitter distribution and propose an efficient approximation using the modified Bessel functions. In Section 3, we form the probabilistic masks for the given confidence probability. In Section 4, we test by the masks the estimates obtained using the the SNP technology. In Section 5, we show how to improve the estimates by removing some unlikely existing breakpoints. Finally, concluding remarks are drawn in Section 6.

2. Jitter distribution in the breakpoints

Referring to an insufficient accuracy of the Laplace distribution in representing the jitter in the breakpoints with low SNRs, in this section we propose and analyse a new approximation to fit measurement data equally well for arbitrary segmental SNRs.

2.1. Approximation with discrete skew Laplace distribution

In order to derive the jitter distribution, the following supposition was made in [41,15]. Suppose that all of the probes to the left of the breakpoint belong to the segment a_l and all of the probes to the right from the breakpoint belong to the segment a_{l+1}. Otherwise, the probability that one or more probes belong to another segment represents the jitter probability. It has been shown in [41,15] that, under such a supposition, jitter in the breakpoints of the CNAs measured in white Gaussian noise can be approximated with the discrete skew Laplace probability density function (pdf) recently derived in [42],

$$p(k|d_l, q_l) = \frac{(1-d_l)(1-q_l)}{1-d_l q_l} \begin{cases} d_l^k, & k \geqslant 0, \\ q_l^{|k|}, & k \leqslant 0, \end{cases} \tag{2}$$

where $0 < d_l = e^{-(\kappa_l/\nu_l)} = P(B_l)^{-1} - 1 < 1$, $0 < q_l = e^{-(1/\kappa_l\nu_l)} = P(A_l)^{-1} - 1 < 1$, $\kappa_l = \sqrt{\frac{\ln x_l}{\ln(x_l/\mu_l)}}$, $\nu_l = -\frac{\kappa_l}{\ln x_l}$,

$$x_l = \frac{\phi_l(1+\mu_l)}{2(1+\phi_l)}\left(1 - \sqrt{1 + \frac{4\mu_l(1-\phi_l^2)}{\phi_l^2(1+\mu_l)^2}}\right), \tag{3}$$

$$\mu_l = \frac{P(A_l)[1-P(B_l)]}{P(B_l)[1-P(A_l)]}, \tag{4}$$

$$\phi_l = \frac{P(A_l) + P(B_l) - 1}{[1-2P(A_l)][1-2P(B_l)]}, \tag{5}$$

$P(A_l)$ is the probability that all of the probes to the left from the supposed breakpoint belong to a_l and $P(B_l)$ is the probability that all of the probes to the right belong to a_{l+1}.

Extensive investigation of pdf (2) in applications to the CNAs-like signals measured in white Gaussian noise have reveled the following [41]:

- Density (2) is reasonably accurate if the SNRs exceed unity, $\gamma_l^-, \gamma_l^+ > 1$, and highly accuracy for $\gamma_l^-, \gamma_l^+ > 1$.
- It is also reasonably accurate if at least one of the segmental SNRs exceed unity, $\gamma_l^- > 1$ or $\gamma_l^+ > 1$, and highly accuracy if $\gamma_l^- > 1$ or $\gamma_l^- > 1$.
- The approximation error is large when $\gamma_l^-, \gamma_l^+ < 1$ and can be unacceptable if $\gamma_l^-, \gamma_l^+ < 1$.

An overall conclusion which can be made following [21,22] is that the Laplace distribution (2) fits only easily seen breakpoints. If the chromosomal changes are not brightly pronounces, the Laplace distribution can be useless in making any decision about the CNAs structures via the estimates. A more correct jitter distribution is thus required.