2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE), Mexico, City. Mexico. September 11-13, 2019.

Comparison between CNA Estimators and WGS technology based on the Refinement of Breakpoints using the Confidence Masks.

1st Jorge Munoz-Minjares
Department of Eng. in Productive Systems
Universidad Tencnologica de Salamanca
Salamanca, Mexico
jmunoz@utsalamanca.edu.mx

2nd Yuriy S. Shmaliy
Department of Electronics Engineering
Universidad de Guanajuato
Salamanca, Mexico
shmaliy@ugto.mx

3rd Tatiana G. Popova
Department of UNITE 830 INSERM
Centre de Recherche, Institute Curie
Paris, France
Tatiana.Popova@curie.fr

4th Janette Perez-Chimal
Department of Mechatronics
Universidad Tecnologica de Salamanca
Salamanca, Mexico
rperez@utsalamanca.edu.mx

5th Jose Lopez-Robles
Department of Eng. in Productive Systems
Universidad Tecnologica de Salamanca
Salamanca, Mexico
jlopez@utsalamanca.edu.mx

6th Misael Lopez-Ramirez
Department of Electronics Engineering
Instituto Tecnologico de Aguascalientes
Aguascalientes, Mexico
m.lopezramirez87@gmail.com

***Abstract*—Alterations in some sections of genomes are commonly related with genetic disorders, which are well known as Structural Aberrations (SAs). The SAs are usually best known as Copy Number Alteration (CNA), and the identifications of this aberrations is essential to diagnose a diseases and to provide an specific treatment. Many algorithms have been developed to estimate the breakpoints and segmental levels of CNAs with the minimum error using the data obtained by modern technologies of hybridization. In spite of recent advances in high resolution data, the breakpoints estimated of CNAs are regularly inconsistent by two principal causes: an extensive variability of measurements and the different designs of the algorithms. Although the estimation of the CNAs is extremely important, still much less attention is given to the estimation accuracy and it is difficult to select the best estimator. In this work, we propose to use the confidence masks based on asymmetric exponential power distribution (AEP) to remove the breakpoints estimated from Ovarian Cancer samples using the CBS and PELT algorithms at a specific probability. Next, the refined breakpoints are compared with a reference obtained by the Whole Genome Sequencing (WGS) technology. Finally, the effectiveness of each estimator is registered by the F_1 test, Area Under Curve (AUC) and Youden's J statistic to determine the best algorithm.**

***Index Terms*—Copy Number Aberrations, confidence masks, breakpoints, estimators.**

I. Introduction

Some genetic disorders, such as cancer, are associated with somatic aberrations in DNA, which are commonly called copy number alterations (CNAs) [1]. Despite the great improvements of high resolutions technologies to obtain the genome chromosomal data, still the measurements contain undesirable artifacts. Among the best known technologies are Array Comparative Genomic Hybridization (aCGH) [2], High Resolution CGH (HRCGH) [3], and Whole Genome Sequencing (WGS) [4]. Based on the definition given by the NCI Dictionary of Genetic Terms, the WGS process helps to obtain almost completely the measurement of chromosomal DNA, so the CNAs estimated using this technology can be considered a *Gold Standard* reference.

The principal factors that affect the CNAs measurements obtained using the technologies above mentioned are:

- nature of biological material (tumor is contaminated by normal tissue, relative values and unknown baseline for copy number estimation),
- technological biases (quality of material and hybridization/sequencing), and
- intensive random noise [5]–[7].

Consequently, the measurements shows an intensive variability causing inconsistencies between the CNAs estimated with different algorithms [8]. Added to this, the mathematical design of the algorithms plays a crucial role in the estimation of breakpoints and levels of CNAs.

Nowadays, it is difficult to choose the best estimator of CNAs because there are few tools designed to this purpose. This task becomes even more challenging given that no one estimator can guarantee an existence of CNAs estimated.

For this reason, the confidence masks algorithm was developed to label the CNAs with a specific probability [9], which works with a fundamental concept called *jitter distribution*. The phenomenon of jitter distribution is immerse in the breakpoints of CNAs and its behavior have been widely studied and constantly improved [10].

In section II we developed the mathematical concepts of probabilistic confidence mask and its improvement respect to the jitter distribution. In section III, the Circular Binary Segmentation (CBS) and Prudence Exact Linear Time (PELT) algorithms are briefly described. The comparison of CNAs processed at several probabilities with the Whole Genome Sequencing (WGS) is showed in the section IV. Finally, the

978–1–7281–4840–319$31.00 ©2019 IEEE

results based on different metrics and the Conclusions are presented in sections V and VI, respectively.

II. Modified Confidence Masks

The confidence masks is an algorithm proposed in [9] that compute the upper and lower boundaries of CNAs estimated and guarantee its existence at a required probability. This algorithm is based in the concept of *jitter distribution*, which was initially computed using the skew Laplace (SkL) law.

The SkL distribution gives some inaccuracies when the segmental signal-to-noise ratio (SNR) ranges below unity [10]. For this reason, in [8] was proposed to modify the confidence masks replacing the skew Laplace distribution with the Asymmetric Exponential Power (AEP) distribution to approximate the jitter distribution in CNAs. The AEP function showed to be more accurate than SkL distribution for low and extra–low Signal-to-Noise-Ratio SNR.

The segmental SNRs are defined and calculated with respect to the lth breakpoint as

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

where $\Delta_l = a_{(l+1)} - a_l$, is the segmental difference and σ_l^2 and σ_{l+1}^2 are the segmental variances, which correspond to the measurements to the left l–segment and right $(l+1)$–segment.

Based on the conclusions made in [11], according to the approximation error using the SkL distribution respect to the value of SNR, it was decided to apply the probabilistic masks modified with the AEP distribution to evaluate the CNAs estimated by different algorithms.

A. Asymmetric Exponential Power Distribution.

According to the results obtained in [11] and the unusual curve of jitter pdf, it was concluded that the AEP distribution is ideal to approximate the jitter distribution [12], which is a generalization of the Gaussian and Laplace laws.

The AEP distribution is computed using the parameters of shape $\alpha_l > 0$, scale $\sigma_l > 0$, location $\theta_l = 0$, and skew $\kappa_l > 0$, and defined as

$$p(k|\bar{p}_l, \bar{q}_l) = \frac{\alpha_l}{\sigma_l \Gamma\left(\frac{1}{\alpha_l}\right)} \frac{\kappa_l}{1+\kappa_l^2} \begin{cases} \bar{p}_l^{k^{\alpha_l}}, & k \geqslant 0\,, \\ \bar{q}_l^{|k|^{\alpha_l}}, & k \leqslant 0\,, \end{cases} \tag{2}$$

where $\bar{p}_l = e^{-\frac{\kappa_l^{\alpha_l}}{\sigma_l^{\alpha_l}}}$, $\bar{q}_l = e^{-\frac{1}{\kappa_l^{\alpha_l}\sigma_l^{\alpha_l}}}$, and $\Gamma(x)$ is the Gamma function.

The constants α_l, κ_l and σ_l to different values of SNRs were found with a highest precision fitting the experimental normalized and computed histograms [11]. This procedure was deveveloped minimizing the Kolmogorov–Smirnov distance [13] defined as,

$$\mathrm{d}_{KS} = \max|F_0(x) - S_N(x)|\,, \tag{3}$$

where $F_0(x)$ is the population cumulative distribution of (2) and $S_N(x)$ is the observed cumulative step function.

The distance d_{KS} was calculated using (3) and selected the minimum one to set the most appropriate values of α_l, κ_l, and σ_l for various symmetric SNRs $\gamma_l^- = \gamma_l^+$. Also, the functions that describe the behavior of the constants α_l and σ_l were approximated in the mean square error (MSE) sense as

$$\alpha_l(\gamma_l) = 1 - \frac{a_1}{\gamma_l^{b_1}}\,, \tag{4}$$

$$\sigma_l(\gamma_l) = a_2 \gamma_l^{b_2}\,, \tag{5}$$

where $a_1 = 0.389$, $b_1 = 0.1394$, $a_2 = 1.142$ and $b_2 = -0.6289$. Note that, for $\gamma_l^- \neq \gamma_l^+$, the shape and scale factors are provided by $\alpha_l(\gamma_l^\pm) = \frac{\alpha_l(\gamma_l^+)+\alpha_l(\gamma_l^-)}{2}$ and $\sigma_l(\gamma_l^\pm) = \frac{\sigma_l(\gamma_l^+)+\sigma_l(\gamma_l^-)}{2}$ [8].

The skew factor κ_l is found modifying the equation proposed in [12]

$$\kappa_l = \left[\frac{\bar{X}_{\alpha_l}^-}{\bar{X}_{\alpha_l}^+}\right]^{\frac{1}{2(\alpha_l+1)}} \tag{6}$$

by substituting $\bar{X}_{\alpha_l}^- = \gamma_l^+$ and $\bar{X}_{\alpha_l}^+ = \gamma^-$ for the asymmetric case and $\kappa_l = 1$ otherwise. The jitter approximation using several simulated measurements for symmetric SNRs was showed with high accuracy in [8], [14].

B. Improved Masks based on AEP distribution.

After approaching the jitter distribution with the AEP-based approximation (2), the probabilistic masks can be modified using the equations provided in [9] for the SkL distribution replacing p_l and q_l respectively,

$$\bar{p}_l = e^{-\frac{\kappa_l^{\alpha_l}}{\sigma_l^{\alpha_l}}}\,, \tag{7}$$

$$\bar{q}_l = e^{-\frac{1}{\kappa_l^{\alpha_l}\sigma_l^{\alpha_l}}}\,. \tag{8}$$

So, the right-hand jitter $\bar{k}_l^{\mathrm{R}}$ and the left-hand jitter $\bar{k}_l^{\mathrm{L}}$ can be defined specifying $\bar{k}_l^{\mathrm{R}}(\vartheta)$ and $\bar{k}_l^{\mathrm{L}}(\vartheta)$ as

$$\bar{k}_l^{\mathrm{R}} = \left\lfloor \frac{\sigma_l}{\kappa_l} \ln \frac{(1-\bar{p}_l)(1-\bar{q})}{\xi(1-\bar{p}_l\bar{q}_l)} \right\rfloor\,, \tag{9}$$

$$\bar{k}_l^{\mathrm{L}} = \left\lfloor \sigma_l \kappa_l \ln \frac{(1-\bar{p}_l)(1-\bar{q}_l)}{\xi(1-\bar{p}_l\bar{q}_l)} \right\rfloor\,. \tag{10}$$

where $\lfloor x \rfloor$ means a maximum integer lower than or equal to x. Note that the background functions of (9) and (10) were obtained in [15] by equating (2) to $\xi(N_l) = \mathrm{erfc}(\vartheta/\sqrt{2})$ and solving for k_l.

Provided (9) and (10), the jitter left boundary $\bar{J}_l^{\mathrm{L}}$ and right boundary $\bar{J}_l^{\mathrm{R}}$ can finally be defined as, respectively,

$$\bar{J}_l^{\mathrm{L}} \cong \hat{n}_l - \bar{k}_l^{\mathrm{R}}\,, \tag{11}$$

$$\bar{J}_l^{\mathrm{R}} \cong \hat{n}_l + \bar{k}_l^{\mathrm{L}}\,, \tag{12}$$

and substituted in the algorithm earlier designed in [9] for the SkL-based confidence masks. So, the modified probabilistic masks can be formalized, which will be used to discard CNAs estimated using several algorithms and compared below with WGS estimates.

III. Estimators description.

In order to compare the CNAs estimated based on specialized methods with WGS technology, we use the algorithms implemented in R, which is a free software environment for statistical computing and graphics.

Circular Binary Segmentation (CBS) is one of the most useful algorithms to estimate CNAs and is implemented in R called BINSEG. In [16], it was proposed the CBS to translate noisy intensity measurements into regions of equal copy number. This algorithm is based on the partitions of a genome into constant segments, detecting copy numbers alterations and the change–point (breakpoint).

The CBS method is applied in R using the functions *cpt.mean* and *cpt.var* setting the next arguments: penalty=*"Manual"*, pen.value=*"log(n)"*, where *n* is the length of DNA measurement, method=*"BinSeg"*, Q=50 the maximum number of changepoints to search, test.stat=*"Normal"* the assumed test statistic, minseglen=10 the minimum segment length.

Another important estimator implemented in R is the Pruned Exact Linear Time (PELT). This method was introduced to find breakpoints and is computationally efficient in several applications, such as CNAs estimate.

The PELT estimator finds the minimum of the cost functions, such as the negative log likelihood, quadratic loss, cumulative sums or those based on both the segment log–likelihood and the length of the segment. Next, the Optimal Partition (OP) and location of breakpoints are obtained having a linear computational cost respect to the number of observations n, under mild conditions, so the computational efficiency of PELT is $\mathcal{O}(n)$ [17]. Also, this procedure requires a penalty for inserted changepoints. This algorithm is applied using the same parameters that for CBS method, replacing the argument method=*"PELT"*.

IV. Comparison of estimators

The comparison of CNAs estimated using the algorithms mentioned above with the WGS estimates is carry out following the procedure described in the pseudo-code in 1.

First, this methodology imply to process the microarray measurement y_n (ratio) estimating the breakpoints $\hat{\boldsymbol{b}}$ of CNAs with any method. The CNA data considered in this work were Ovarian cancer samples and sequenced using shallow Whole Genome Sequencing (WGS) technology; copy number estimations were obtained using ControlFreec tool [18] and randomly modified for anonymization. The figure 1 illustrate an example of breakpoints estimated using different techniques. Here it is possible to identify a notorious inconsistency in the location of breakpoints estimated.

Next, the modified probabilistic confidence masks $\mathcal{B}^{UB}_{l|\alpha\mathrm{E}}$ and $\mathcal{B}^{LB}_{l|\alpha\mathrm{E}}$ are applied to test the CNAs estimated with a specified probability tuning the parameter $\boldsymbol{\vartheta}$.

So, the CNAs and its breakpoints $\bar{\boldsymbol{b}}$ that possibly exist at a specified probability are compared with the ideal estimation of breakpoints $\boldsymbol{b}$. These breakpoints are provided by the method (WGS), which is considered a *gold standard*. However, the comparison is delimited to breakpoints $\boldsymbol{b}$ of segments with a length of 1, 2 and 3 Mbp seeking the best conditions for each algorithm. This restriction is based on the argument that many times medical experts made their analysis and diagnostics according to a specific length of estimated CNAs.

Algorithm 1 Algorithm to compare the breakpoints estimators respect to an ideal estimation obtained using the Whole Genome Sequencing technology. The modified probabilistic masks remove the breakpoints that unsatisfied a given probability, which is specified with the ϑ–sense.

Input: y_n, $\boldsymbol{\vartheta}$, $\mathbf{b}_i|threshold$
L =length($\boldsymbol{\vartheta}$)
$\hat{\boldsymbol{b}} \leftarrow$ BP_estimator(y_n) ◁ Estimated breakpoints.
for $i = 1 : L$ **do**
  $\bar{\boldsymbol{b}} \leftarrow \mathcal{B}^{UB}_{l|\alpha\mathrm{E}}, \mathcal{B}^{LB}_{l|\alpha\mathrm{E}}(y_n,\hat{\boldsymbol{b}},\vartheta_i)$ ◁ Breakpoints refined.
  $\boldsymbol{R} \leftarrow \hat{\boldsymbol{b}} \cap \bar{\boldsymbol{b}}$ ◁ Points removed.
  $\boldsymbol{X} \leftarrow \hat{\boldsymbol{b}} \cap \boldsymbol{b}$ ◁ Match breakpoints
  $\boldsymbol{Z}_1 \leftarrow \boldsymbol{R} \cap \boldsymbol{X}$ ◁ Points removed from match breakpoints
    $TP_i = length(\boldsymbol{X}) - length(\boldsymbol{Z}_1)$
    $TN_i = length(\boldsymbol{R}) - length(\boldsymbol{Z}_1)$
    $FP_i = length(\hat{\boldsymbol{b}}) - TP_i - length(\boldsymbol{R})$
    $FN_i = length(\boldsymbol{b}) - TP_i$
  $\boldsymbol{TPR} \leftarrow (TP_i/(TP_i + FN_i))$ ◁ True Negative Rate
  $\boldsymbol{FPR} \leftarrow (FP_i/(FP_i + TN_i))$ ◁ True Positive Rate
end for
Output: $\boldsymbol{TPR}, \boldsymbol{FPR}$

Finally, the matched breakpoints $\boldsymbol{X}$, where the location of $\hat{\boldsymbol{b}} = \boldsymbol{b}$, the removed points $\boldsymbol{R}$ from estimated breakpoints and from match points $\boldsymbol{Z}_1$, are considered to obtain the Receiver Operating Characteristic (ROC) curve.

So, it is obtained four possible results True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) points, which are represented in Table I. The True Negative points are obtained erasing a False Positive using the confidence masks.

V. Results

The results of the comparison are analyzed using the graph method based the ROC curve and computing different metrics to evaluate binary classifier systems.

TABLE I: Possible cases of comparison between a particular breakpoints detector –CBS or PELT– and Whole Genome Sequencing estimation, represented with the symbols ◯ and ▼, respectively. The case of True Negatives is given when a False Positive is removed using the confidence masks, it is illustrated with the symbol ⊗.

Method	TP	FP	TN	FN
Breakpoint estimator	◯	◯	⊗	
Reference NGS	▼			▼

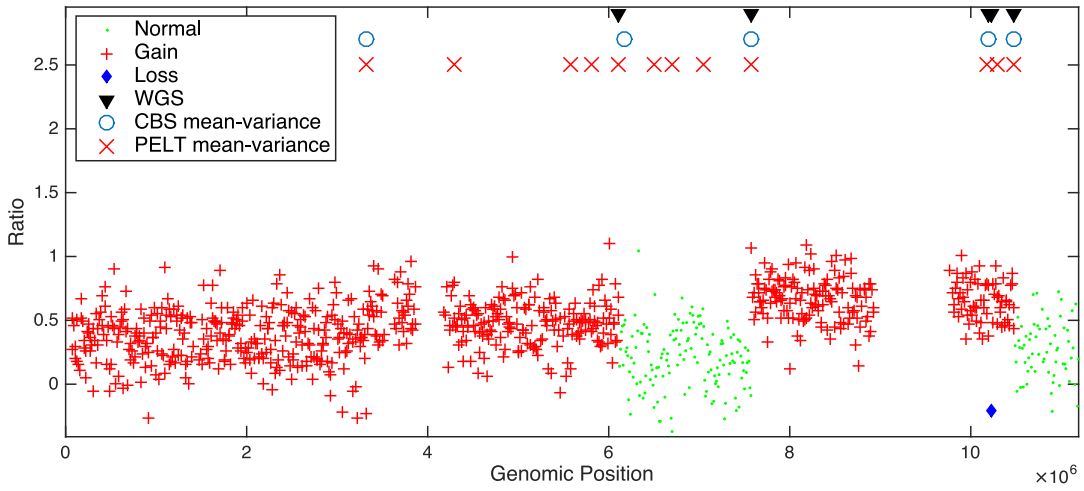


Fig. 1: Estimated breakpoints of Chromosomes 4 from sample 6 of Ovarian cancer using the methods CBS mean-variance (circle), PELT (cross) based on mean-variance and WGS technique (inverted triangle). The measurement shows Gain (plus sing), Loss (diamond) and normal (points) levels.

A. ROC Curve

Based on the computed points TP, FN, TN, and FP it is possible to show the results using a ROC curve. The ROC space is created plotting the TPR against the FPR [19]. The TPR and FPR are also known as sensitivity and as the fall–out or probability of false alarm, respectively. These parameters are calculated using the below equations

$$TPR = \frac{TP}{P} = \frac{TP}{TP+FN} \tag{13}$$

$$FPR = \frac{FP}{N} = \frac{FP}{FP+TN} = 1 - TNR \tag{14}$$

where TNR is defined as the specificity and computed as TNR=TN(TN+ FP), P the number of real positive cases and N the number of real cases in the data. The diagonal line between (0,0) and (1,1) divide the better (upper area) and worst (lower area) classification.

So, the Figures 2a and 2b show the results of comparison in the range of ϑ from 0.6745 to 20 (left to right) in the ROC space of each method with three delimiters for CNAs of WGS: 1 (cross), 2 (plus-sign), and 3 (circle) Mbp.

The curve generated by CBS based on mean is far from being a good estimator to the CNA measurements. Also, the PELT estimator based on mean shows some bad results, but this performance change to the database limited at 2 Mbp and 3Mbp when the confidence masks are applied with a discriminant probability $\geq 9.45\vartheta$. The best results are generated by the CBS and PELT estimators based on mean-variance parameters to all the established limits at a probability $> 0.6745\vartheta$.

B. Metrics

We compare two algorithms the CBS and PELT based on different parameters of the measurement *mean* and *mean* and *variance*. To determine the effectiveness of estimator, we used the analysis of the F-test score, the Area Under the Curve (AUC) and the Youden's J statistic.

1) F-test: The F_1 score is the harmonic average of the precision and recall [20]. The equation to compute this parameter is described as

$$F_1 = \frac{(\beta^2+1) \times P \times R}{\beta^2 \times P \times R} \tag{15}$$

where P is precision, R is recall, and β is the relative importance given to recall over precision. If recall and precision are of equal weight, $\beta = 1$. The F_1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

The precision P or Positive Predictive Value (PPV) is computed as,

$$PPV = \frac{TP}{TP+FP} \tag{16}$$

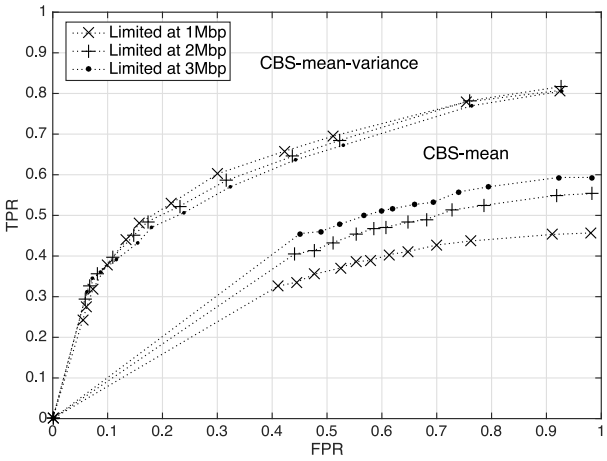
and the the recall R or True Positive Rate (TPR) is defined in equation (13).

The figure 3, show the maximum value of F_1 to each one algorithm with different delimitation 1 Mbp, 2 Mbp and 3 Mbp. The algorithm PELT based on mean obtain the maximum score in the three cases.

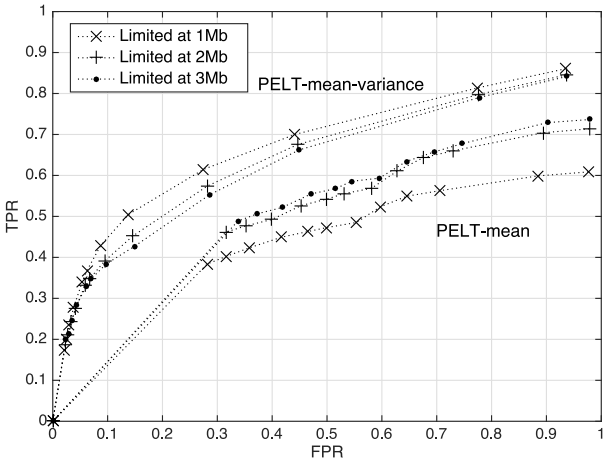
However, the F_1 test omit the true negative results, which represent the breakpoints removed by the confidence masks. For this reason, it is required to analyze the results given by the comparison with different metrics.

2) Area Under the ROC Curve and Youden's J statistic: Area Under the ROC Curve (AUC) is a fraction of the area of the unit square, which takes values between 0 and 1 [21]. Then, an acceptable estimator should have an AUC greater than 0.5.

The trapezoidal method can be used to approximate the integration under the curve ROC. Thus, it is possible define the formula to compute the area over an interval a to b as



(a) CBS.



(b) PELT.

Fig. 2: True positive rate against the false positive rate based on the results of comparison between a) CBS based on mean and mean-variance, b)PELT based on mean and mean-variance, and with WGS estimates at three thresholds: Mega base pairs for a range ϑ from 0.6745 to 20.

$$\int_a^b f(x)dx \approx \frac{1}{2}\sum_{n=1}^{N}(x_{n+1}-x_n)[f(x_n)+f(x_{n+1})] \quad (17)$$

where $f(x)$ is the sensitivity (TPR) and x is the False Positive Rate (FPR). The values of AUC to each one estimator delimited at different levels are showed in the table II, indicating the maximum values with bold numbers. The best algorithm to the first, second and third delimitation (1Mbp, 2Mbp and 3Mbp) was PELT based on mean and variance.

The Youden's J statistic is a well know parameter to evaluate the performance of a diagnostic, which is obtained using the parameters of TPR and specificity (TNR) as

$$J = TPR - TNR - 1 \quad (18)$$

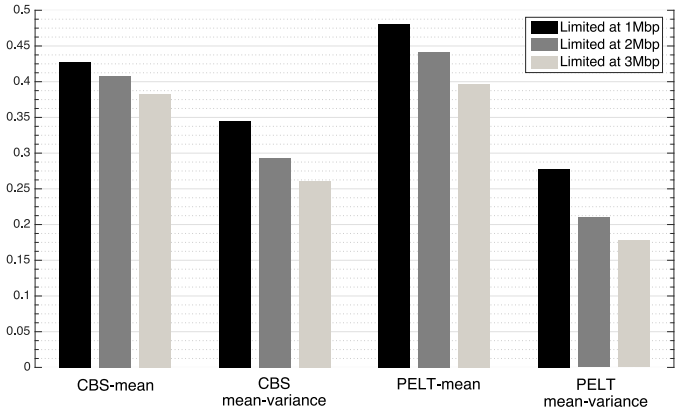


Fig. 3: F-measurement of accuracy.

TABLE II: Area Under the Curve and maximum Youdens J values of each estimator at several restrictions.

ESTIMATOR	Area Under the Curve		
	1 Mbp	2 Mbp	3 Mbp
CBS-mean	0.2347	0.2668	0.2872
CBS-mean-variance	0.5678	0.5611	0.5480
PELT-mean	0.3611	0.4021	0.4044
PELT-mean-variance	**0.6182**	**0.5911**	**0.5785**
	Maximum Youdens J statistic		
	1 Mbp	2 Mbp	3 Mbp
CBS-mean	-0.0847	-0.0360	0.041
CBS-mean-variance	0.3220	0.3118	**0.2905**
PELT-mean	0.1023	0.1464	0.1510
PELT-mean-variance	**0.3669**	**0.3062**	0.2865

This index has the value zero if the test reports the same proportion of positive tests for both control and disease groups. It has the value unity when, and only when, there are neither false positives nor false negatives resulting from the test [22]. The maximum values of J of each estimator with different conditions are showed in the table II.

Based in this metric, the PELT mean-variance shows a greater J level than CBS mean-variance to the first restriction (1Mbp), while an opposite behavior occurs with the delimitation of 2 and 3 Mbp. However, the higher levels of Youden's J of CBS mean-variance are obtained when the confidence masks are applied at a high probability $\geq 9.45\vartheta$

VI. Conclusions

The Circular Binary Segmentation (CBS) and Pruned Exact Linear Time (PELT) algorithms generate different CNAs estimations despite using the same parameters of reference *mean* and *mean-variance*. The comparison of CBS and PELT estimators was carried out by modified confidence masks based on AEPD (asymmetric exponential power distribution), which helped to improve the accuracy of each algorithm at a specified probability.

For comparison, we can declared that the PELT based on the mean-variance is the best estimator to this Ovarian cancer

database. The Area Under the Curve (AUC) supports this premise obtaining the higher value for all delimitations 1, 2 and 3 Mega base-pair (Mbp). In spite of the Youden's J parameter is greater to CBS mean-variance than PELT mean-variance with the limitation of 2 and 3Mbp, the probability required to obtain this values is higher to CBS. Furthermore, the F_1 parameter was insufficient to determine the worst and best algorithm because the comparison of this work is widely based on the True Negative (TN) results.

Finally, we can conclude that the PELT mean-variance obtained its best performance at a averaged probability of $1-1.11\times^{-16}$. So, the modified probabilistic confidence masks may play a crucial role in detecting actual chromosomal changes.

REFERENCES

[1] N.A. Graham, A. Minasyan, A. Lomova, A. Cass, N.G. Balanis, et al, "Recurrent patterns of DNA copy number alterations in tumors reflect metabolic selection pressures," Mol Syst Biol vol. 13, pp. 914, 2017.

[2] F. Forozan, R. Karhu, J. Kononen, A. Kallioniemi, O.P. Kallioniemi, et al. "Genome screening by comparative genomic hybridization," Trends Genet vol. 13, pp. 405–409, 1997.

[3] M.R. Speicher, N.P. Carter, "The new cytogenetics: Blurring the boundaries with molecular biology," Nat Rev Genet vol. 6, pp. 782–792, 2005.

[4] P.C. Ng, E.F. Kirkness, "Whole genome sequencing", Methods Mol. Biol. vol. 628, pp. 215–226, 2010.

[5] F. Zare, M. Dow, N, Monteleone, A. Hosny, S. Nabavi, et al., "An evaluation of copy number variation detection tools for cancer using whole exome sequencing data," BMC Bioinformatics vol. 18, pp. 286, 2017

[6] T. Popova, V. Boeva, E. Manie, Y. Rozenholc, E. Barillot, et al., "Analysis of somatic alterations in cancer genome: From SNP arrays to next generation sequencing," Sequence and Genome Analysis I Humans, Animals and Plants. Ltd IP (ed) iConcept Press Ltd, 2013.

[7] J.U. Munoz, Y.S. Shmaliy, "Estimates of the breakpoints in genome copy number alteration profiles with masks," Biomed. Signal Process Contr. vol. 10, pp. 238–248, 2017.

[8] J.U. Munoz-Minjares, Y.S. Shmaliy, T. Popova and R.J. Perez–Chimal, "Matching Confidence Masks with Experts Annotations for Estimates of Chromosomal Copy Number Alterations," International Conference on Bioinformatics and Biomedical Engineering, Springer, Cham. pp. 85-94, 2018.

[9] J.U. Munoz-Minjares and Y.S. Shmaliy, "Confidence masks for genome DNA copy number variations in applications to HR–CGH array measurements," Biomed. Signal Process. Contr. vol. 13, pp. 337–344, Sep. (2014).

[10] J.U. Munoz, Y.S. Shmaliy, "Improving estimates of genome CNVs with confidence masks using SNP array data," Biomed. Signal Process. Contr. vol. 31, pp. 238–248, 2017.

[11] J.U. Munoz-Minjares, J. Cabal, and Y.S. Shmaliy, "Effect of noise on estimate bounds for genome DNA structural changes," WSEAS Trans. on Biology and Biomedicine, vol. 11, pp. 52–61, Apr. 2014.

[12] A. Ayebo and T.J. Kozubowski, "An asymmetric generalization of Gaussian and Laplace laws," Journal of Probability and Statistical Science vol.1, no. 2, pp.187–210, 2003.

[13] F.J. Massey, "The Kolmogorov–Smirnov test for goodness of fit," Journal of the American Statistical Association, vol. 46, no. 253, pp.68–78, 1951.

[14] Jorge Munoz–Minjares, Yuriy S. Shmaliy, Ro. Olivera–Reyna, Re. Olivera–Reyna and R.J. Perez–Chimal,"Jitter representation in SCNA breakpoints using asymmetric exponential power distribution," 14th Int. Conf. on Electrical Engineering, Computing Science and Automatic Control, 2017.

[15] J.U. Muñoz, J. Cabal and Y.S. Shmaliy,"Probabilistic bounds for estimates of genome DNA copy number variations using HR-CGH microarray," Proc. 21st European Signal Process. Conf. (EUSIPCO), Marrakech, Marocco, Sep. 2013.

[16] A. B. Olshen, E.S. Venkatraman, R. Lucito and M. Wigler, "Circular binary segmentation for the analysis of array-based dna copy number data," Biostatistics, vol. 5, no. 4, pp. 557-572, 2004.

[17] R. Killick, P. Fearnhead, and I.A. Eckley, "Optimal detection of changepoints with a linear computational cost," Journal of the American Statistical Association, vol. 107. no. 500, pp. 1590–1598, 2012.

[18] V. Boeva, T. Popova, K. Bleakley, P. Chiche, J. Cappo, G. Schleiermacher and E. Barillot, "Control–FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data," Bioinformatics, vol. 28, no. 3, pp. 423-425, 2011.

[19] J.A. Hanley and B.J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, vol. 143, no. 1, pp. 29-36, 1982.

[20] W.B. Frakes and R. Baeza-Yates, "Information Retrieval: Data Structures & Algorithms," Englewood Cliffs: Prentice Hall ISBN:0-13-463837-9, 1992.

[21] T. Fawcett, "An introduction to ROC analysis", Pattern recognition letters, vol. 27, no. 8, pp. 861-874, 2006.

[22] W.J Youden, "Index for rating diagnostic tests," Cancer, vol. 3, pp. 32-35, 1950.