22nd International Congress on Modelling and Simulation, Hobart, Tasmania, Australia, 3 to 8 December 2017
mssanz.org.au/modsim2017

An Improved Hybrid Algorithm for Multiple Change-point Detection in Array CGH Data

G. Y. Sofronov [a], **T. V. Polushina** [b], **M. W. Jayawardana** [c]

[a] *Department of Statistics, Faculty of Science and Engineering, Macquarie University, Sydney NSW 2109 Australia*

[b] *Department of Clinical Science, Faculty of Medicine and Dentistry, University of Bergen, 7804, NO-5020 Bergen, Norway*

[c] *Department of Statistics, Data Science and Epidemiology, Faculty of Health, Arts and Design, Swinburne University of Technology, Melbourne VIC 3122 Australia*
Email: mjayawardana@swin.edu.au

Abstract: A human genome is highly structured. Usually, the structure forms regions having patterns of a specific property. It is well-known that analysis of biological sequences is often confronted with measurements for the gene expression levels. When these observations are ordered by their location on the genome, the values form clouds with different observed means, supposedly reflecting different mean levels. The statistical analysis of these sequences aims at finding chromosomal regions with "abnormal" (increased or decreased) mean levels. Therefore, identifying genomic regions associated with systematic aberrations provides insights into the initiation and progression of a disease, and improves the diagnosis, prognosis and therapy strategies.

In this paper, we present a further extension of our work, where we propose a two-staged hybrid algorithm to identify structural patterns in genomic sequences. At the first stage of the algorithm, an efficient sequential change-point detection procedure (for example, the Shiryaev-Roberts procedure or the cumulative sum control chart (CUSUM) procedure) is applied. Then the obtained locations of the change-points are used to initialize the Cross-Entropy (CE) algorithm, which is an evolutionary stochastic optimization method that estimates both the number of change-points and their corresponding locations. The first-stage of the algorithm is very sensitive for the thresholds selection, and the identification of optimal thresholds will increase the accuracy of the results and further improve the efficiency of the algorithm. In this study, we propose an improved hybrid algorithm for change-point detection, which uses optimal thresholds for the sequential change-point detection procedure and the CE method to obtain more precised estimates. In order to illustrate the usefulness of the algorithm, we have performed a comparison of the proposed hybrid algorithms for both artificially generated data and real aCGH experimental data. Our results show that the proposed methodologies are effective in detecting multiple change-points in biological sequences.

Keywords: *Change-point detection, aCGH microarray data, CNVs, DNA copy number, combinatorial optimization, Cross-Entropy method*

1 INTRODUCTION

Recent biological studies show the close relationship between chromosomal copy number alterations and diseases like cancer and diabetes (for example, see Almal and Padh (2012)). The technique of microarray comparative genomic hybridization (array CGH) enables one to perform genome-wide screening for all possible regions with DNA copy number variations (CNV). In abnormal cells, mutations can cause a gene to be either deleted from the chromosome or amplified, that is, there are extra DNA copies of the gene. The analysis of aCGH data relates to one of the most important applications involving change point detection. Identifying abrupt changes is also very important in other biological applications such as recombination of viruses (Halpern, 2000), characterization of complete transcriptomes via high-density DNA tiling microarrays (Huber et al., 2006) or investigation of DNA sequences in general (see, for example, Karlin and Brendel (1993)).

In recent years, a number of change-point detection methods have been developed. Many of the segmentation approaches are discussed in Braun and Muller (1998), Algama and Keith (2014) and Priyadarshana and Sofronov (2015). These also include stochastic optimization methods (for example, genetic algorithm and the Cross-Entropy method) Priyadarshana and Sofronov (2015); Evans et al. (2011); Polushina and Sofronov (2011, 2013, 2014); Priyadarshana et al. (2013, 2015); Polushina and Sofronov (2016) and Markov chain Monte Carlo (MCMC) algorithms Algama and Keith (2014); Keith (2006); Keith et al. (2008); Sofronov (2011). Apart from applications in bioinformatics, segmentation methods can also be used in different fields including economics (Priyadarshana and Sofronov, 2012), ecology (López et al., 2010), and quality control (Sofronov et al., 2012).

All change-point problems can be divided in two large groups: retrospective (off-line) and sequential (online). In the first case, all data have been observed already and the problem is to estimate the number and the locations of the change-points. In the second (online) case, variables appear sequentially (one by one) and one does not know the future observations. Since in the case of DNA segmentation, all observations are known, sequential change-point detection methods can be used, for example, as an initial approximation for off-line methods (Priyadarshana et al., 2013).

The paper is structured as follows. Section 2 provides a statement of the multiple change-point problem in mathematical terms. Section 3 describes the proposed hybrid algorithm and a general Cross-Entropy method for multiple change-point problem. In Section 4, we perform the numerical analysis to assess the significance of the proposed procedure. Section 5 concludes the paper with a discussion.

2 THE MULTIPLE CHANGE-POINT PROBLEM

We model genome sequences as a multiple change-point process, that is, a process in which sequential data are separated into segments by an unknown number of change-points, with each segment supposed to have been generated by a different process. More formally, let us consider a sequence of observations $\mathbf{X} = (x_1, x_2, \ldots, x_L)$ of length L, in which the x_i's are independently distributed random variables. A segmentation of the sequence is specified by the number of change-points N and the corresponding locations of the change-points $\mathbf{C} = (c_1, c_2, \ldots, c_N)$, where $1 = c_0 < c_1 < \cdots < c_N < c_{N+1} = L + 1$. In this context, a change-point is defined as a boundary between two adjacent segments. The value of c_n is the sequence position of the rightmost character of the segment to the left of the n-th change-point. The segments are numbered from 0 to N as there will be one more segments than the number of change-points. The model assumes that within each segment the observations are distributed as normal with the mean μ_n and the variance σ_n^2, $n = 0, 1, \ldots, N$. Both the mean and the variance are not known in advance and the maximum likelihood method is used to obtain their estimates. The log-likelihood of the model is

$$ll(\mathbf{X} \mid N, \mathbf{C}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \sum_{n=0}^{N} \left[-\frac{\lambda_n}{2} \ln\left(2\pi\sigma_n^2\right) - \frac{1}{2} \sum_{i=c_n}^{c_{n+1}-1} \left(\frac{x_i - \mu_n}{\sigma_n} \right)^2 \right], \tag{1}$$

where the length of the n-th segment is defined as $\lambda_n = c_{n+1} - c_n$, $\mathbf{C} = (c_1, c_2, \ldots, c_N)$, $\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_N)$, and $\boldsymbol{\sigma}^2 = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_N^2)$.

If we assume that there is no change in the variance, that is, for all segments σ^2 stays the same, then the

log-likelihood function is

$$ll(\mathbf{X} \mid N, \mathbf{C}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \sum_{n=0}^{N} \left[-\frac{\lambda_n}{2} \ln\left(2\pi\sigma^2\right) - \frac{1}{2} \sum_{i=c_n}^{c_{n+1}-1} \left(\frac{x_i - \mu_n}{\sigma} \right)^2 \right]. \tag{2}$$

3 THE HYBRID CROSS-ENTROPY METHOD FOR THE MULTIPLE CHANGE-POINT PROBLEM

We can consider the multiple change-point detection problem as a combinatorial maximization problem of the log-likelihood function defined in (1) or (2). In general, let S be a real-valued performance (or objective) function on $\mathcal{X}$, where $\mathcal{X}$ is a finite set and the aim is to find the positions of change-points that correspond to the maximum value of S over $\mathcal{X}$. In order to solve this problem, we use the Cross-Entropy (CE) method, a stochastic optimization method that is used for the estimation of rare event probabilities (Rubinstein and Kroese, 2004, 2007). This estimation problem can be reformulated as an optimization problem.

We propose to combine a sequential change-point detection algorithm with the CE method to detect multiple change-points. So our hybrid algorithm consists of two stages.

Stage 1: We use the function `processStream` from the R package `cpm` Ross (2015). If we aim to detect changes in the mean (see model (2)), we use `cpmType=Student` Hawkins et al. (2003). To detect changes in mean and variance (see model (1)), we use `cpmType=GLR` Hawkins and Zamba (2005). Traditionally, in sequential analysis we select parameters to trade off between false alarm rate and average run length. However, in our case we will verify and update true change-points by the CE algorithm. Therefore here by selecting parameters, we allow to obtain relatively high false alarm rate with additional false positive change-points.

Stage 2: The obtained locations of the change-points in stage 1 are used as initial parameters for the Cross-Entropy algorithm. We created new functions `CE.Normal.Init.Mean` and `CE.Normal.Init.MeanVar` in the R package `breakpoint` (Priyadarshana and Sofronov, 2016) to obtain change-point estimates with initial values.

The proposed hybrid algorithm and its steps are explained in algorithm 1.

Algorithm 1 Proposed hybrid algorithm.

1: Run a sequential change-point detection algorithm to obtain initial estimates for the number (N) as well as the locations ($\mathbf{C}$) of change-points.
2: Based on the estimates of N and $\mathbf{C}$, initiate the CE algorithm to obtain more accurate locations of change-points.
3: For all pairs of adjacent segments, perform a two sample Student's *t*-test (for equal variances) or Welch's *t*-test (for unequal variances), in order to identify the most insignificant change-point with the largest *p*-value. The *p*-values are adjusted using a multiple comparison correction (for example, the Bonferroni correction Simes (1986)) to control the family wise error rate in multiple hypothesis testing.
4: Eliminate the most insignificant change-point from the solution and update the solution vector with the other estimates.
5: Initiate the CE algorithm with the new set of change-point locations.
6: Repeat steps 3, 4 and 5 until all change-points found are significant. Return $\mathbf{C}$: the vector of change-point locations. The length of this vector is the resultant number of change-points.

Let N be the maximum number of change-points and $\boldsymbol{c}$ be a set of the change-points, which is a non-decreasing N-dimensional vector. In the R package `breakpoint`, we can choose either normal or 4-parameter beta distributions to simulate the change-point positions. The CE method updates the parameters in each step and updating is continued until a stopping criterion is met. In each iteration an *elite* sample is defined as the best performing combinations of change-points with respect to the performance function score. The process is carried out until a specific stopping criterion is achieved. In each step, the simulation parameters are updated accordingly. The main steps of the CE algorithm are described in algorithm 2 using normal distribution to simulate change-point locations. Here we choose the simulation parameters under conditions that guarantee convergence of the CE algorithm Costa et al. (2007).

Algorithm 2 CE algorithm.

1: Choose initial sets for $\mu^{(0)} = \left(\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_N^{(0)}\right)$ and $(\sigma^2)^{(0)} = \left((\sigma_1^2)^{(0)}, (\sigma_2^2)^{(0)}, \ldots, (\sigma_N^2)^{(0)}\right)$. The length of both vectors is N. Set $t = 1$.

2: Generate a random sample $c^{(1)}, c^{(2)}, \ldots, c^{(N_1)}$ from the normal distributions with parameters $\left(\mu^{(t-1)}, (\sigma^2)^{(t-1)}\right)$, where $c^{(i)} = \left(c_1^{(i)}, c_2^{(i)}, \ldots, c_N^{(i)}\right)$, $i = 1, 2, \ldots, N_1$, is a change-point vector.

3: For each $i = 1, 2, \ldots, N_1$, order $\left(c_1^{(i)}, c_2^{(i)}, \ldots, c_N^{(i)}\right)$ from smallest to biggest.

4: Evaluate the performance of each $c^{(1)}, c^{(2)}, \ldots, c^{(N_1)}$. Define the elite sample, which is the best performing combinations of the change-points. Let $N_{elite} = \rho N_1$ be the size of the elite sample.

5: For all $j = 1, 2, \ldots, N$, estimate the parameters $\mu_j^{(t)}$ and $(\sigma_j^2)^{(t)}$ using the elite sample and update the current parameter sets as follows:

$$\mu_j^{(t)} = \frac{\sum\limits_{i \in I} c_j^{(i)}}{N_{elite}}, \qquad \left(\sigma_j^2\right)^{(t)} = \frac{\sum\limits_{i \in I} \left(c_j^{(i)} - \mu_j^{(t)}\right)^2}{N_{elite}},$$

where I is the set of indices of the best performing samples.

6: Stopping criterion is $\max_j (\sigma_j^2)^{(t)} < \varepsilon$.

7: If the stopping criterion is met, then stop the process and identify the combination of the positions of change points $c^{(i)}$ that minimizes the objective function. Otherwise set $t = t + 1$ and iterate from step 2.

4 Numerical Results

In this section, we include results of numerical experiments that illustrate the performance of the hybrid CE method. In the first example, we consider a synthetic sequence with a known distribution, which allows us to provide direct comparison with existing techniques. The second example uses aCGH experimental data.

4.1 Example 1: Artificial data with multiple change-points

We generated 100 Gaussian random sequences of length 3500 with 10 abrupt change-points (or 11 segments) and the standard deviation, $\sigma = 1$, and different signal-to-noise ratios (SNRs), $\mathrm{SNR} = \mathrm{Mean}/\sigma$. Table 1 displays the values of the parameters used for this simulation study.

Table 1. The parameter values for the artificial data

Segment	1	2	3	4	5	6	7	8	9	10	11
Length	200	550	150	250	500	250	400	600	200	150	250
SNR	0	2	4	2.5	0	2	3	4	2.5	3.5	1

Table 2 compares the number of change-points estimated by the CE method (function `CE.Normal.Mean` in the R package `breakpoint`), the hybrid CE method, cumSeg Muggeo (2012); Muggeo and Adelfio (2011) and DNAcopy (Seshan and Olshen, 2016; Olshen et al., 2004; Venkatraman and Olshen, 2007) for 100 generated sequences. The results show that the hybrid CE algorithm (CE hybrid) more accurately estimates the number of change-points compared to the CE algorithm. The DNAcopy method appears to be very effective while cumSeg tends to overestimate the number of change-points. We observed a significant improvement in the processing time in the proposed hybrid CE method over its standard implementation. Figure 1 shows the profile plots for the first sequence out of 100 generated sequences; all of the methods are in very good agreement.

4.2 Example 2: Real data (neuroblastoma)

In this example, we used a manually annotated data set (Hocking et al., 2013), where neuroblastoma array-CGH profiles were analysed. The data are available in the R package `neuroblastoma` (Hocking, 2013). In

Table 2. The number of change-points estimated by different segmentation methods for 100 generated sequences

Estimated number of change-points	7	8	9	10	11	12	13	14	Average processing time (in seconds)
CE	0	0	6	14	42	28	9	1	11.28
CE hybrid	1	3	39	51	6	0	0	0	5.39
cumSeg	0	0	1	23	69	3	4	0	0.26
DNAcopy	0	0	0	79	12	6	3	0	0.08

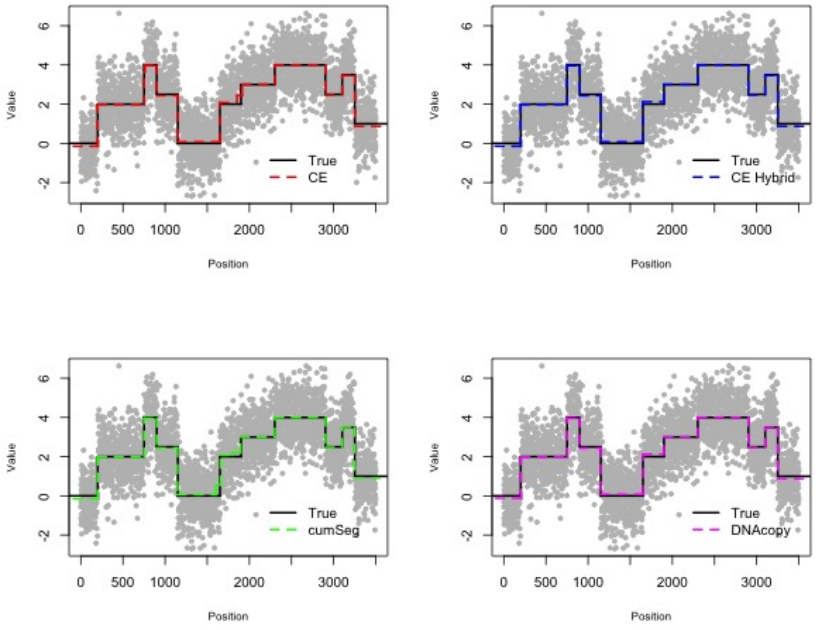

Figure 1. Profile plots of the CE algorithm (red dashed line), the hybrid CE algorithm (blue dashed line), cumSeg (green dashed line) and DNAcopy (magenta dashed line) for the artificial data; the true profile is the black solid line

this data set, we searched for the longest profile with at least two change-points. We choose the profile with `profile.id`=508, for which chromosome 11 is annotated as having at least one change-point.

Figure 2 shows the comparison of the algorithms (the CE algorithm, the hybrid CE algorithm, cumSeg, and DNAcopy) for chromosome 11. All methods identify the most significant change-point very accurately. They may however differ in estimating some smaller segments and the less significant change-points.

5 Conclusion

In this paper, we have proposed an improved hybrid Cross-Entropy algorithm, which uses the R packages `cpm` and `breakpoint`. We have compared the performance of the proposed hybrid algorithm with the standard CE algorithm (which is also implemented in the R package `breakpoint`), which does not use results from the sequential techniques as its initial parameters. We have also compared the algorithm with other well-known segmentation methods: cumSeg and DNAcopy.

Numerical results show that the hybrid CE algorithm tends to underestimate the correct number of change-points. This may be explained by the use of the Bonferroni correction, which can be conservative. Overall processing time was significantly improved using the hybrid implementation compared to the standard algo-

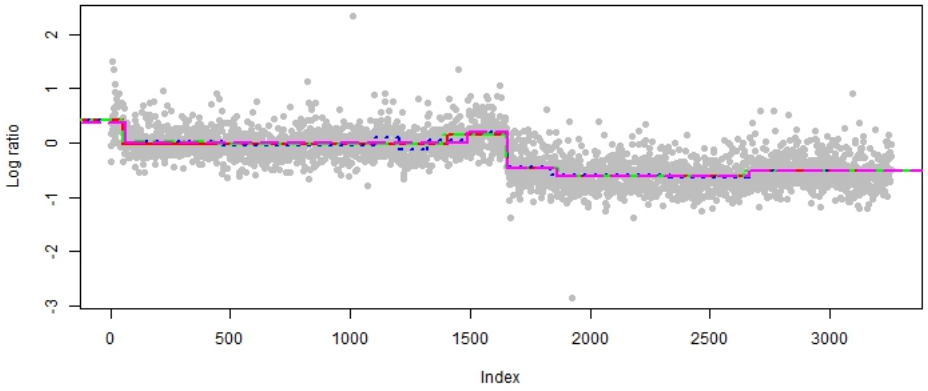

Figure 2. Array CGH profiles plots for chromosome 11: the CE algorithm (red solid line), the hybrid CE algorithm (blue dotted line), cumSeg (green dotdashed line) and DNAcopy (magenta longdashed line)

rithm. While the compared methods may differ in how they estimate the number of change-points, all of them appear to be very effective in identifying the locations of the change-points (see Figures 1 and 2).

REFERENCES

Algama, M. and J. Keith (2014). Investigating genomic structure using changept: A Bayesian segmentation model. *Computational and Structural Biotechnology Journal 10*(17), 107–115.

Almal, S. H. and H. Padh (2012). Implications of gene copy-number variation in health and diseases. *Journal of Human Genetics 57*(1), 6–13.

Braun, J. and H.-G. Muller (1998). Statistical methods for DNA sequence segmentation. *Statistical Science 13*(2), 142–162.

Costa, A., O. Jones, and D. Kroese (2007). Convergence properties of the cross- entropy method for discrete optimization. *Operations Research Letters 35*(5), 573–580.

Evans, G. E., G. Y. Sofronov, J. M. Keith, and D. P. Kroese (2011). Estimating change-points in biological sequences via the cross-entropy method. *Ann. Oper. Res. 189*(1), 155–165.

Halpern, A. L. (2000). Multiple-changepoint testing for an alternating segments model of a binary sequence. *Biometrics 56*(3), 903–908.

Hawkins, D., P. Qiu, and C. Kang (2003). The changepoint model for statistical process control. *Journal of Quality Technology 35*(4), 355–366.

Hawkins, D. M. and K. Zamba (2005). Statistical process control for shifts in mean or variance using a changepoint formulation. *Technometrics 47*(2), 164–173.

Hocking, T. D. (2013). *neuroblastoma: Neuroblastoma copy number profiles*. R package version 1.0.

Hocking, T. D., G. Schleiermacher, I. Janoueix-Lerosey, V. Boeva, J. Cappo, O. Delattre, F. Bach, and J.-P. Vert (2013). Learning smoothing models of copy number profiles using breakpoint annotations. *BMC Bioinformatics 14*(1), 164.

Huber, W., J. Toedling, and L. M. Steinmetz (2006). Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics 22*(16), 1963–1970.

Karlin, S. and V. Brendel (1993). Patchiness and correlations in DNA sequences. *Science 259*, 677–677.

Keith, J., D. Kroese, and G. Sofronov (2008). Adaptive independence samplers. *Statistics and Computing 18*(4), 409–420.

Keith, J. M. (2006). Segmenting eukaryotic genomes with the generalized Gibbs sampler. *J. Comp. Biol. 13*(7), 1369–1383.

López, I., M. Gámez, J. Garay, T. Standovár, and Z. Varga (2010). Application of change-point problem to the detection of plant patches. *Acta Biotheoretica 58*, 51–63.

Muggeo, V. M. (2012). *cumSeg: Change point detection in genomic sequences*. R package version 1.1.

Muggeo, V. M. R. and G. Adelfio (2011). Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics 27*(2), 161–166.

Olshen, A. B., E. Venkatraman, R. Lucito, and M. Wigler (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics 5*(4), 557–572.

Polushina, T. and G. Sofronov (2011). Change-point detection in biological sequences via genetic algorithm. In *Proceedings of the IEEE Congress on Evolutionary Computation (CEC'2011)*, pp. 1966–1971.

Polushina, T. and G. Sofronov (2014). Change-point detection in binary Markov DNA sequences by the cross-entropy method. In *2014 Federated Conference on Computer Science and Information Systems, FedCSIS 2014*, pp. 471–478.

Polushina, T. and G. Sofronov (2016, Oct). A cross-entropy method for change-point detection in four-letter dna sequences. In *2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 1–6.

Polushina, T. V. and G. Y. Sofronov (2013). A hybrid genetic algorithm for change-point detection in binary biomolecular sequences. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2013)*, pp. 1–8.

Priyadarshana, M., T. Polushina, and G. Sofronov (2013). A hybrid algorithm for multiple change-point detection in continuous measurements. In *International Symposium on Computational Models for Life Sciences, AIP Conference Proceedings*, Volume 1559, pp. 108–117.

Priyadarshana, M., T. Polushina, and G. Sofronov (2015). Hybrid algorithms for multiple change-point detection in biological sequences. *Advances in Experimental Medicine and Biology 823*, 41–61.

Priyadarshana, M. and G. Sofronov (2012). A modified cross-entropy method for detecting change-points in the Sri-Lankan stock market. In *Proceedings of the IASTED International Conference on Engineering and Applied Science, EAS 2012*, pp. 319–326.

Priyadarshana, W. and G. Sofronov (2015). Multiple break-points detection in array CGH data via the cross-entropy method. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 12*(2), 487–498.

Priyadarshana, W. and G. Sofronov (2016). *breakpoint: An R Package for Multiple Break-Point Detection via the Cross-Entropy Method*. R package version 1.2.

Ross, G. J. (2015). Parametric and nonparametric sequential change detection in R: The cpm package. *Journal of Statistical Software 66*(3), 1–20.

Rubinstein, R. and D. Kroese (2004). *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. New York: Springer-Verlag.

Rubinstein, R. and D. Kroese (2007). *Simulation and the Monte Carlo Method*. John Wiley & Sons.

Seshan, V. E. and A. Olshen (2016). *DNAcopy: DNA copy number data analysis*. R package version 1.48.0.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika 73*(3), 751–754.

Sofronov, G. (2011). Change-point modelling in biological sequences via the Bayesian adaptive independent sampler. *International Proceedings of Computer Science and Information Technology 5*, 122–126.

Sofronov, G., T. Polushina, and M. Priyadarshana (2012). Sequential change-point detection via the cross-entropy method. In B. Reljin and S. Stankovic (Eds.), *The 11th Symposium on Neural Network Applications in Electrical Engineering (NEUREL2012)*, pp. 185–188.

Venkatraman, E. and A. B. Olshen (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics 23*(6), 657–663.