*Indian Journal of Agricultural Sciences* **92** (7): 892–6, July 2022/Article
https://doi.org/10.56093/ijas.v92i7.118040

# Segmentation of genomic data through multivariate statistical approaches: comparative analysis

ARFA ANJUM[1], SEEMA JAGGI[1]*, SHWETANK LALL[1], ELDHO VARGHESE[2], ANIL RAI[1], ARPAN BHOWMIK[1] and DWIJESH CHANDRA MISHRA[1]

*ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012, India*

Received: 18 November 2021; Accepted: 25 April 2022

ABSTRACT

Segmenting a series of measurements along a genome into regions with distinct characteristics is widely used to identify functional components of a genome. The majority of the research on biological data segmentation focuses on the statistical problem of identifying break or change-points in a simulated scenario using a single variable. Despite the fact that various strategies for finding change-points in a multivariate setup through simulation are available, work on segmenting actual multivariate genomic data is limited. This is due to the fact that genomic data is huge in size and contains a lot of variation within it. Therefore, a study was carried out at the ICAR-Indian Agricultural Statistics Research Institute, New Delhi during 2021 to know the best multivariate statistical method to segment the sequences which may influence the properties or function of a sequence into homogeneous segments. This will reduce the volume of data and ease the analysis of these segments further to know the actual properties of these segments. The genomic data of Rice (*Oryza sativa* L.) was considered for the comparative analysis of several multivariate approaches and was found that agglomerative sequential clustering was the most acceptable due to its low computational cost and feasibility.

**Keywords**: Genome, Multivariate analysis, Segmentation, Sequential clustering

Sequencing has transformed research tremendously. Since the discovery of DNA structure, scientists have been evaluating genetic sequences. External or internal events can influence position-ordered genomic data, which may create a sudden structural shift in the data set. Change-points help to locate genetic variability, which is needed to research extensively. Segmentation is one way to discover transition points in genomic data and identifies homogenous zones considering the variability between sequences (Braun and Muller 1998). In recent years, technologies that accurately locate and size change-points have gained prominence.

Segmentation can be performed using statistical or algorithmic approaches. HMM (Bleakely and Vert 2011), Bayesian technique (Husmeier *et al*. 2002), nonparametric two-sample tests (Killick *et al*. 2012, Rigaill *et al.* 2012) and LASSO-based change-point detection (Omranian *et al.* 2015) have been studied for structural changes. iSeg (Girimurugan *et al*. 2018), Segmentr (Mello and Florencia 2019) are some examples of algorithmic approaches. Several univariate approaches were utilized to detect change-points, but multivariate techniques were rarely employed on genomic data. Multivariate signifies that multiple dependent variables are combined to produce a single result. A multicomponent system is the result of component interactions. High-throughput technology can track changes in genes and proteins. Alterations in the correlation structure may be caused by changes in the behaviour of the system's components. Multiple time series segments can be utilized on time-resolved biological data to discover key changes as system breakpoints.

Here, we tried to use multivariate approaches to molecular genomic data. Several studies were done on time series sequence data or on transcriptomics data (Microarray/ChIpseq or RNA seq data) or simulated data sets (Du Y *et al*. 2013, Omranian *et al*. 2013) due to genomic data properties: large size, presence of extreme or influential observations, too many zeroes, missing values are computational challenges when dealing with genomic data along with low variance-covariance determinant. Omranian *et al.* (2015) developed a regularised regression-based technique for detecting multivariate time-series breakpoints and segments. The sequential character of genetic data must be considered when applying statistical techniques, especially segmentation. In this paper, we have attempted to compare the different multivariate techniques for the segmentation of genome sequences.

## MATERIALS AND METHODS

*Data Description*: The data of Rice (*Oryza sativa*

[1]ICAR-Indian Agricultural Statistics Research Institute, New Delhi; [2]ICAR-Central Marine Fisheries Research Institute, Kochi.
*Corresponding author email: seema.jaggi@icar.gov.in

*japonica* group cv. Nipponbare) genome from National Centre for Biotechnology Information (NCBI) were considered for this study (2021). Fasta file was downloaded for Chromosome 1, with accession number NC_029260. The four variables considered for the study are GC content, CpG island, SNP and CNV. From Fasta file, GC content is extracted by using R-script and the formula followed for this is [G+C/(A+G+C+T)]. CpG island and CNV variables essentially represent a stretch of genome sequence and thus cannot be directly used in the segmentation process as corresponding recordings of these variables for a basic unit are not available. To handle this kind of data problem, we have taken proportion values for the respective region or unit. This method got inspired by Ortiz-Estevz *et al*. (2011), in which they used a segmented CNV approach. From Information Commons for Rice (IC4R), SNP (Single Nucleotide Polymorphism) data were taken. IC4R database contains 18 million single nucleotide polymorphisms (SNPs) discovered through resequencing of 5152 rice accessions and provides an ultra-high density rice variation map, and these SNPs are openly accessible. The CpG island data is downloaded from NCBI, Genome data viewer, and the CNV, fourth variable of data for rice genome was taken from NCBI Gene Expression Omnibus under the accession ID GSE42769 (Wang *et al*. 2013).

Variables were chosen for study based on a few studies in recent years to know the pattern in genomic sequences. Previously, studies were made on human or simulated data to understand the relations between CNV and Gene expression by Ortiz-Estevez *et al.* (2011), integrated study of SNP, CNV and gene expression in genetic association studies was made by Momtaz *et al*. (2018). All the variables chosen signify for specific properties and interest is to know about their correlation and effect on each other over the entire genome. Data preparation with the four variables was done using R software. Initially, the whole data was broken into small units of size 100 bp, resulting in 43,270 units with observations for all four variables. This data is plotted with the help of a graph concerning their quartile value. Summary statistics were obtained using R software, and the summary statistics of 1st, 2nd, 3rd quartile for GC, SNP, CpG and CNV are given in Table 1.

Correlation among genomic variables (SNP, GC, CpG and CNV) indicated a significant correlation, even though the strength of linear relation is small except for CpG with GC where the correlation is 0.771.

Table 1 Summary statistics of the data used for the segmentation

| Statistic | SNP | GC | CpG | CNV |
|---|---|---|---|---|
| Minimum | 0.0 | 0.17 | 0.0 | 0.0 |
| Ist Quartile | 6.0 | 0.37 | 0.0 | 0.0 |
| Median | 15.0 | 0.42 | 0.0 | 0.0 |
| Mean | 20.4 | 0.44 | 0.55 | 0.003 |
| 3rd Quartile | 31.0 | 0.49 | 1.0 | 0.0 |
| Maximum | 192.0 | 0.76 | 5.0 | 0.6 |
| SD | 18.27 | 0.09 | 0.83 | 0.03 |

As discussed in the data description, the variables considered in this study, namely GC, CNV and CpG, except SNP, cannot be computed for each base pair. These variables are a sequence property, not individual base-pair property. Hence, the genomic data is divided in basic units of size 1000 bp for the sake of applying any statistical technique.

Let $x_t$ be a unit of size n (=1000) where $1 < t < N$, N being the total number of units. Consecutive groups of such basic units form a window. We have taken the window size as 100 and used a sliding size of 100, which results in non-overlapping windows each of size 100 basic units. Using a sliding size of less than 100 will result in overlapping windows and can be used in segmentation methods supporting the sliding window concept. A detailed flow diagram of the sliding window approach has been given in figure 1. The problem is to identify one or more change-points in the sequence $x_t$. Let $x_{c1}, x_{c2}, \ldots, x_{cp}$ be p change-points in the sequence.

Let us define the problem taking the size of $S_i$, $S_i$ being a window at $i^{th}$ position

$$\text{Null hypothesis, } H_0: S_i = S_{i+1}$$

$$\text{The alternate hypothesis, } H_1: S_i \neq S_{i+1}$$

A statistical test, checks whether any two consecutive windows are the same, if a change point is not detected. The problem of finding the change-points can be seen as a variant of a simple two-sample test with one constraint of preserving the sequential nature of the original data. In other words, a change point divides the sequence into two different segments. To achieve this, two approaches can be employed, namely the divisive approach and the agglomerative approach.

*Divisive*: In this approach, all windows are considered to have belonged to a single segment. Now a change point is identified to divide this into two segments. Furthermore, the new segments are segmented by identifying change-points inside them.

*Agglomerative*: In the agglomerative approach, all the individual windows are considered as segments, and consecutive windows are tested for equality. Change-points are identified when the null hypothesis gets rejected by the test.

Sliding window methodology is adopted for this work. Firstly, observations on characteristics of interest are taken for each window, then size of the segment (=n) and sliding window (=k) is fixed. The parameters are estimated for each segment starting from 1 to n, k to n+k, 2k to n+2k, and so on. After that, the two consecutive segments are tested for significance and the significant segments are determined. This is continued for entire sequence and final segments are obtained.

*Multivariate statistical tools used for segmentation*

Hotelling $T^2$: It is a multidimensional extension of the student's t-statistic, which is now a commonly used tool for detecting differentially expressed genes in gene testing.

This has wide application potential in genome association studies, microarray process control and data control charts.

The null hypothesis states that all response variables' group means are equal.

The null hypothesis $H_0$: $\mu_{Si} = \mu_{Si+1}$ against the alternate hypothesis $H_1$: $\mu_{Si} \neq \mu_{Si+1}$

The two-sample Hotelling's $T^2$ test statistic is given by:

$$T^2 = \left(\bar{\S}_{s_i} - \bar{\S}_{s_{i+1}}\right)' \left[S\left(\frac{1}{n_{s_i}} + \frac{1}{n_{s_{i+1}}}\right)\right]^{-1} \left(\bar{\S}_{s_i} - \bar{\S}_{s_{i+1}}\right) \quad (1)$$

$$S = \frac{\left(n_{s_i} - 1\right)S_{X_{s_i}} + \left(n_{s_{i+1}} - 1\right)S_{X_{s_{i+1}}}}{\left(n_{s_i} - 1\right) + \left(n_{s_{i+1}} - 1\right)} \quad (2)$$

where S, the pooled covariance matrix of the sample for X and $X_{Si+1}$; $\bar{X}$, the mean of the sample, and the sample for each random variable $x_i$ in X has $n_{Si}$ elements.

Multivariate Cramer's test: It is a nonparametric multivariate technique given as:

$$C_{n_i n_{i+1}} = \frac{n_i n_{i+1}}{n_i + n_{i+1}} \begin{pmatrix} \frac{2}{n_i n_{i+1}} \sum_{p,q}^{n_i n_{i+1}} \phi\left(\left\|\vec{s}_{i_p} - \vec{s}_{i+1_q}\right\|^2\right) - \frac{1}{n_i^2} \\ \sum_{p,q=1}^{n_i} \phi\left(\left\|\vec{s}_{i_p} - \vec{s}_{i_q}\right\|^2\right) - \frac{1}{n_{i+1}^2} \\ \sum_{p,q=1}^{n_{i+1}} \phi\left(\left\|\vec{s}_{i+1_p} - \vec{s}_{i+1_q}\right\|^2\right) \end{pmatrix} \quad (3)$$

where $\vec{S}_{i_p}$ and $\vec{S}_{i_p}$, independent random vectors. The random vectors were assumed to be identically distributed. Here, $n_i$ is the number of observations in $S_i$ segment and $n_{i+1}$ denotes the number of observations in $S_{i+1}$ segment. The function $\phi$ is the kernal function [Franz C (2000), Baringhaus and Franz (2004)].

Sequential clustering approach: There are several clustering approaches available for the natural grouping of entities. The usual clustering approach cannot be applied in genomic segmentation since it does not preserve the sequential nature of the data. Hence, a sequential clustering approach is used here. Let $\{s_1, s_2, ..., s_{\left\lceil \frac{N}{n} \right\rceil + 1}\}$ be the $\left\lceil \frac{N}{n} \right\rceil + 1$ windows considered as initial segments. The goodness of fit statistic is given by the following expression:

$$\hat{T}_{\left\lceil \frac{N}{n} \right\rceil + 1}(S;\alpha) = \sum_{i=1}^{\left\lceil \frac{N}{n} \right\rceil} \hat{Q}\left(S_i, S_{i+1}; \alpha\right) \quad (4)$$

where α, the power of Euclidean distance whose value ranges between $0 < \alpha < 2$. Q is the term measuring between and within distance of two segments based on each observation of the two segments under consideration.Using the resampling technique, a maximum of S is determined to find segments.

A greedy approach is employed to obtain an estimated solution since computing the real maximum of the goodness-of-fit statistic for a given starting segmentation would be too computationally expensive. To tackle the problem of overfitting, the series of goodness-of-fit statistics can be penalized. This is achieved by using the penalty parameter, which calculates a penalty depending on the position of change-points. As a result, the positions of the change-points are calculated using maximization

$$\tilde{S}_k = \hat{S}_k + \text{penalty}\left(\vec{\tau}(k)\right) \quad (5)$$

where $\vec{\tau}(k) = \{\tau_1, \tau_2, ..., \tau_k\}$ is the set of change-points associated with the goodness of fit statistic $S_k$ (James and Matteson 2015).

Multivariate Kolmogorov Smirnov test: It is another multivariate nonparametric test proposed by Justel, Pena and Zamar (1997) for multivariate data. This test is an extension of the Kolmogorov–Smirnov two-sample test statistic to a multivariate setup. The test uses a statistic that is built using Rosenblatt's transformation, and an algorithm is available to compute it in the bivariate case. Due to the difficulty in computing the empirical distribution function under a multivariate setup, application potential is limited.

A detailed flow chart for the implementation of the developed algorithm has been given in Fig 1. The methodology was implemented by developing the script in R software (https://www.r-project.org/) and is available with the authors.

## RESULTS AND DISCUSSION

We have attempted four multivariate approaches for the segmentation of genomic data. Most of the multivariate segmentation techniques tend to fail in the presence of large missing data. We ignored the missing observations while applying these techniques. The assumption of normality could not be valid in genomic data, and hence any inference based on Hotelling $T^2$ is questionable, and therefore, we investigated the application of nonparametric techniques for segmentation. Kolmogorov-Smirnov of the two-sample test is the most obvious choice for such a case, but unfortunately, working out the empirical distribution function is very difficult because of the multivariate nature of the data. Cramer's multivariate two samples test looked promising and was pursued in both approaches, divisive and agglomerative as well. In the divisive approach, equally spaced five locations were chosen on the genome sequence, and five samples were found, each having the start as genome start and end at one of the five locations decided earlier. All the samples were tested against the whole genome. The test failed due to the complexity and computational cost as the test had to perform a subsampling procedure (Bootstrapping or Monte Carlo methods) to get the region of rejection. In the agglomerative approach, 100 successive windows were tested against the next 100 windows. This approach failed initially because of the presence of too many zeroes; in some cases, all the windows under testing had only zeroes in one or more variables except GC content. To avoid this, we added a condition in the program to include only the variables with a non-zero sum in the two samples being

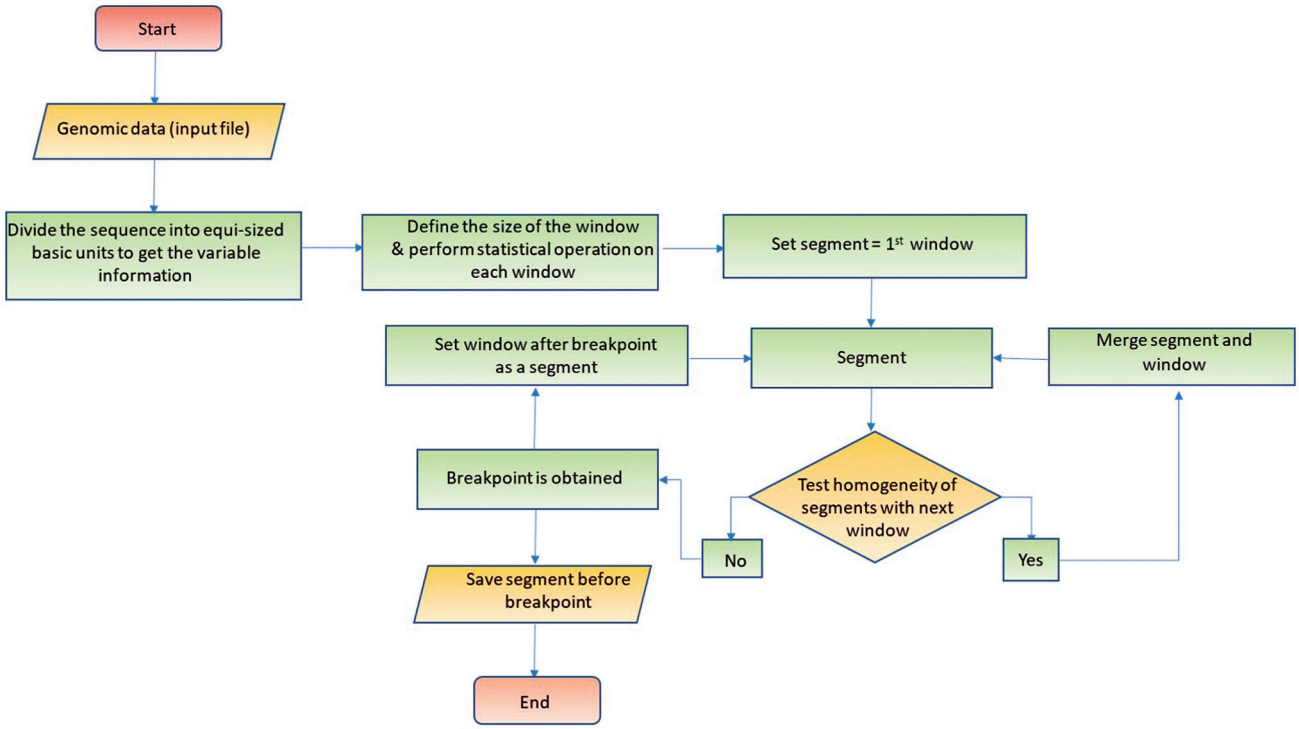


Fig 1 Flow chart of computational steps in multivariate segmentation.

tested. This resulted in 230–260 segments under different tuning parameters settings. On closer inspection, it was found that the change-points obtained might be due to the presence of extreme outlying or influential observations. These observations made it very hard to accept the null hypothesis, and hence most of the window samples were identified as separate segments. Removal of these outlying observations will miss the objective of finding a robust segmentation methodology on real genomic data. We have made a comparative analysis of four multivariate techniques and analysed the capability of handling various situations.

Three techniques, viz. Hotelling $T^2$, Multivariate Kolmogorov-Smirnov test and Multivariate Cramer's test ignore the autocorrelation or sequential nature of the data considered. Sequential clustering provides a suitable candidacy for the task. Simple clustering treats all observations as independent and can group any observation without considering the order, but sequential clustering takes this into account. We have used an energy-based agglomerative sequential clustering greedy approach to get the segments as the computation cost is heavily reduced by providing an initial guess for segments, and the obtained

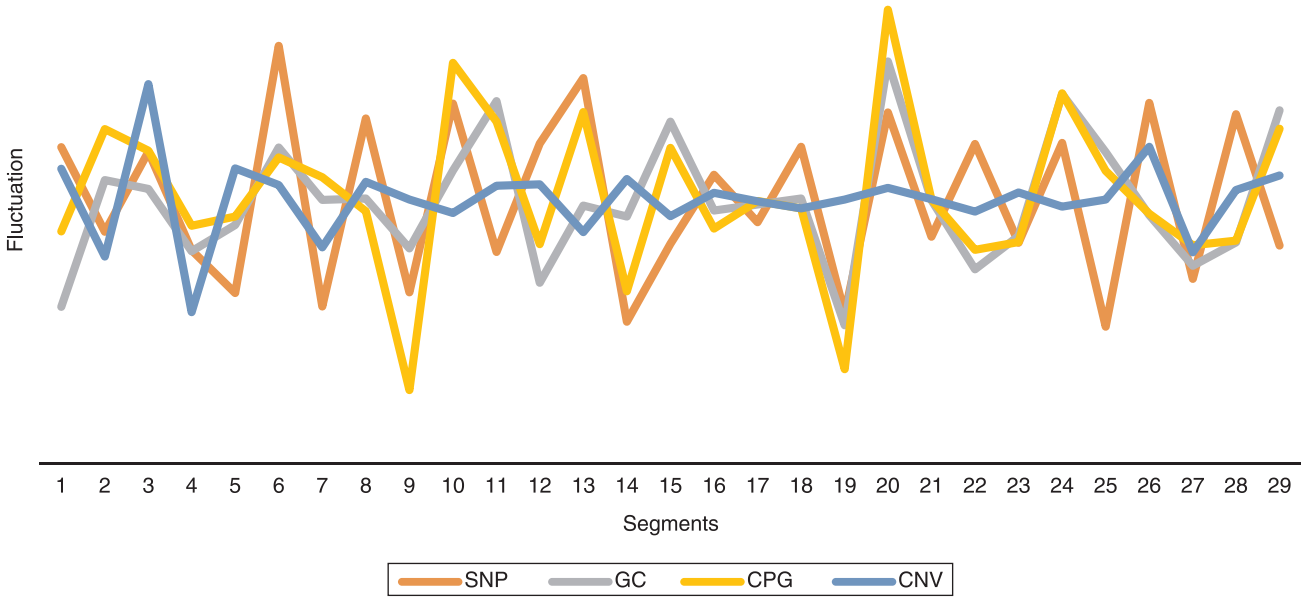


Fig 2 Relative fluctuations of different variables over consecutive segments.

Table 2 Distinct segments (start and end points) obtained through sequential clustering along with summary information pertaining to each variable

| Start | End | SNP | GC | CPG | CNV |
|---|---|---|---|---|---|
| 1 | 200000 | 2949 | 0.45626 | 82 | 1 |
| 200001 | 1400000 | 29569 | 0.4359 | 419 | 13 |
| 1400001 | 1900000 | 9255 | 0.4396 | 241.385 | 0 |
| 1900001 | 3500000 | 44031 | 0.44162 | 920.539 | 35 |
| 3500001 | 11400000 | 141503 | 0.43183 | 4150.62 | 4 |
| 11400001 | 12500000 | 218 | 0.42701 | 542.127 | 7 |
| 12500001 | 15700000 | 93652 | 0.43682 | 1831.4 | 29 |
| 15700001 | 15800000 | 896 | 0.43675 | 61.4077 | 0 |
| 15800001 | 16700000 | 21891 | 0.43692 | 532.808 | 3 |
| 16700001 | 17000000 | 2015 | 0.42764 | 69.243 | 1 |
| 17000001 | 19600000 | 64733 | 0.43309 | 1272.02 | 2 |
| 19600001 | 20200000 | 8961 | 0.45175 | 381.858 | 2 |
| 20200001 | 23600000 | 86680 | 0.43596 | 1876.91 | 21 |
| 23600001 | 23700000 | 4850 | 0.43478 | 71.7838 | 0 |
| 23700001 | 25500000 | 45570 | 0.43155 | 979.139 | 7 |
| 25500001 | 28300000 | 47372 | 0.44626 | 1796.77 | 2 |
| 28300001 | 30800000 | 54061 | 0.44419 | 1466.05 | 5 |
| 30800001 | 33200000 | 41627 | 0.44326 | 1400.55 | 4 |
| 33200001 | 33500000 | 8205 | 0.44342 | 169.919 | 0 |
| 33500001 | 33600000 | 575 | 0.4196 | 24.5171 | 0 |
| 33600001 | 34500000 | 20011 | 0.44573 | 543.898 | 2 |
| 34500001 | 38900000 | 66678 | 0.44556 | 2658.2 | 10 |
| 38900001 | 39600000 | 17987 | 0.43235 | 356.198 | 0 |
| 39600001 | 41100000 | 26281 | 0.42539 | 642.556 | 2 |
| 41100001 | 41400000 | 8463 | 0.44539 | 188.758 | 0 |
| 41400001 | 41500000 | 414 | 0.45435 | 68.3438 | 0 |
| 41500001 | 41600000 | 2244 | 0.45135 | 65.6305 | 1 |
| 41600001 | 41700000 | 737 | 0.43879 | 56.9431 | 0 |
| 41700001 | 42800000 | 25848 | 0.43064 | 540.746 | 2 |
| 42800001 | 43270000 | 6936 | 0.44754 | 293.925 | 3 |

SNP, CNV and CpG values are the total values that lie in a particular segment, whereas GC content values are mean values that lie in that segment, which can be obtained by Count (G + C)/Count (A + T + G + C).

segments are reported in Table 2 along with summary information about each variable within each segment.

But apart from GC content, all other variables are total or counted as their values. This makes it difficult to compare different segments due to their varying scales. Therefore, a numerical transformation has been made to each of the variables, except GC content, for making a relative comparison among the change-points concerning each variable (Fig 2). The figure shows sequential clustering is efficient in capturing the variability in the data set and identifying the change-points.

Most of the available techniques have their roots in multivariate time series analysis, while the different nature of real biological data is so obviously evident. In this study, it was found that segmenting real multivariate genomic data is challenging. Due to the huge size and presence of outlying observations, a nonparametric, robust, and computationally cheap technique is needed. Out of all the techniques considered in this study, only energy-based greedy agglomerative sequential clustering was found useful. This study highlights the potential of greedy heuristics for deeper exploration. Another extension could be to explore the possibility of dimension reduction techniques or a Bayesian approach for the segmentation.

## REFERENCES

Baringhaus L and Franz C. 2004. On a new multivariate two-sample test. *Journal of Multivariate Analysis* **88**: 190–206.

Bleakley K and Vert J P. 2011. The group fused lasso for multiple change-point detection. *Technical Report HAL-00602121*. Computational Biology Center, Paris.

Braun J V and Muller H G. 1998. Statistical methods for DNA sequence segmentation. *Statistical Science* **13**(2): 142–62.

Du Y, Murani E, Ponsuksili S and Wimmers K. 2014. biomvRhsmm: Genomic Segmentation with Hidden Semi-Markov Model. *BioMed Research International* **2014**: 1–12.

Franz C. 2000. 'A statistical test for the multidimensional two-sample problem'. Diploma Thesis, University of Hanover, Germany.

Girimurugan S B, Liu Y, Lung P Y, Vera D L, Dennis J H, Bass H W and Zhang J. 2018. iSeg: An efficient algorithm for segmentation of genomic and epigenomic data. *BMC Bioinformatics* **19**(1): 1–15.

Husmeier D and Wright F. 2002. A Bayesian approach to discriminate between alternative DNA sequence segmentations. *Bioinformatics* **18**(2): 226–34.

James N A and Matteson D S. 2015. ecp: An R package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software* **62**(7): 1–25.

Justel A, Pena D and Zamar R. 1997. A multivariate Kolmogorov–Smirnov test of goodness of fit. *Statistics & Probability Letters* **35**(3): 251–59.

Killick R, Fearnhead P and Eckley I A. 2012. Optimal detection of change-points with a linear computational cost. *Journal of the American Statistical Association* **107**(500): 1590–98.

Mello T and Florencia L. 2019. *Segmentr: Segment data minimizing a cost function*. Retrieved from https://CRAN.R-project.org/package=segmentr

Momtaz R, Ghanem N M, El-Makky N M and Ismail M A. 2018. Integrated analysis of SNP, CNV and gene expression data in genetic association studies. *Clinical Genetics* **93**(3): 557–66.

Omranian N, Mueller-Roeber B and Nikoloski Z. 2015. Segmentation of biological multivariate time-series data. *Scientific Reports* **5**(1): 1–6.

Ortiz-Estevez M, De Las Rivas J, Fontanillo C and Rubio A. 2011. Segmentation of genomic and transcriptomic microarrays data reveals major correlation between DNA copy number aberrations and gene-loci expression. *Genomics* **97**(2): 86–93.

Rigaill, G, Lebarbier E and Robin S. 2012. Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and Computing* **22**: 917–29.

Wang Y, Wu C, Ji Z, Wang B and Liang Y. 2011. Non-parametric change-point method for differential gene expression detection. *PLoS ONE* **6**(5): 1–16.