

A two-step procedure for detecting change points in genomic sequences

Arfa Anjum¹, Seema Jaggi^{2,*}, Shwetank Lall³, Eldho Varghese⁴, Anil Rai¹, Arpan Bhowmik³ and Dwijesh Chandra Mishra¹

¹Centre for Agricultural Bioinformatics,

²Agricultural Education Division, and

³Division of Design of Experiments, ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012, India

⁴Fishery Resources Assessment Division, ICAR-Central Marine Fisheries Research Institute, Kochi 682 018, India

The field of whole genomic studies and investigations is currently focused on change-point detection. Over time, various segmentation techniques have been proposed to identify these change points. To effectively locate segments within a genome, it is helpful to pinpoint the intervals or boundaries between them, which are known as change points. By treating these change points as outliers, they can be identified. The anomalies or outliers in a dataset are the observations which are significantly different from the rest of the observations. They can be attributed to some measurement errors or properties of the data themselves. Studying the fluctuations over different segments also revealed the heterogeneity between consecutive segments. In this paper, anomaly identification approach or influential point detection has been discussed and studied in cow genome data of chromosome 25. Furthermore, the observed anomalies have been confirmed to determine whether or not they are true change points. The two-step technique resulted in the identification of change sites based on observed abnormalities and is efficient in terms of calculation time and cost. This study aims to detect any anomalies in genomic data and determine the exact points at which the data segment significantly differed from the rest of the segments. We have developed relevant R codes for data processing and applied methodologies.

Keywords: Anomalies, change points, genomic sequences, segmentation, two-step procedure.

CHANGE-point detection is a popular area of research in whole genomic studies and genomic investigations. The whole genome of an organism can be represented by a linear or circular DNA molecule, which consists of a strand of the letters A (adenine), T (thymine), G (guanine) and C (cytosine). Over the last few decades, research has shown that genomes often contain a wide range of distinct and diverse properties, which can be referred to as the ‘blueprint’ of an organism. Regulatory sequences, protein-coding regions, promoters, operons and other functionally essential features of genomes have been discovered in recent decades through

various studies. Others, such as horizontally transmitted areas, prophages, repetitive sequences, etc. have evolutionary relevance.

Due to the non-uniformity of the DNA sequence, statistical features are often unevenly distributed throughout. Strong signals can be found in certain areas, such as the protein-coding regions, while weaker signals are present in non-coding regions. In the genome, there exist dinucleotides (CpG islands) in low and higher numbers, the quality of which depict different inference. In computational biology, segmenting sequential genomic data based on location is a prevalent problem. Several segmentation techniques have been proposed over time. In order to identify segments in a genome, it is useful to identify intervals/boundaries between segments known as change points.

Many new segmentation algorithms have been recently developed to break a genomic sequence computationally into pieces based on predefined structural and functional factors. Change-point detection has long been used to study DNA sequences for a variety of purposes. Segmentation methods for categorical variables, for example, have been developed to find patterns of gene prediction¹. Amplification, mutation and deletion of genomic regions have been identified using the point-of-change technique^{2,3}. Finding transcription units like expressed versus unexpressed loci⁴ or operons⁵ is difficult. Data from multiple time intervals, tissues and cell types are segmented to compare changes in genomic organization.

There are numerous segmentation methods for genomic characteristics and time-series data^{6,7}. The majority of the work has been done with simulated or cancerous data. However, research on genuine molecular data is uncommon. The segmentation issue involves separating an ordered series of genomic data into uniform, roughly constant intervals and has swiftly gained popularity in computational biology. Genomic activity and gene regulation were earlier studied using segmentation methods on genomic, transcriptomic, epigenomic and proteomic datasets as input. Due to the widespread availability of genomics datasets, the number of segmentation methods employed has expanded recently.

Many genomic investigations, such as detecting copy-number variations or discovering transcribed areas, encounter

*For correspondence. (e-mail: seema.jaggi@icar.gov.in)

change-point issues. Change points can now be located at the nucleotide resolution because of the advancements in next-generation sequencing methods. Knowledge of the precise positions of many change sites in genomic sequences is useful for various biological functions. In the case of piecewise regression models, a filtering solution is available for the sequential multiple change-point identification problem⁸. The recent resurgence of regression analysis has coincided with the tremendous advancement of genetic engineering^{1,9-11}.

An outlier response, a sequence of trials over time or a change in location might cause an anomaly in genetic data. Although various statistical approaches for identifying outliers have been reported, identifying actual outliers remains difficult, especially in high-dimensional genomic data. This study offers an effective method for finding anomalies in whole genomic data.

Outlying observations, also known as influencing observations, can have a destructive or constitutive effect on the correct functioning of an organism. The detection of influential observations using a linear regression approach has been a popular study topic¹²⁻¹⁵. Cook's *D*, DFBETA, DFFITS, Grubbs' test, Dixon's test, Rosner's test, Atkinson's *Ci* and COVRATIO are some of the most commonly used measurements for this¹⁶. Cook's *D* is a popular measure for detecting outliers employing a linear regression technique¹⁷. However, distinguishing actual outliers from non-outliers remains difficult, especially when dealing with high-dimensional genomic data. The most difficult part is differentiating mild outliers from regular observations and disguising actual outliers¹⁸.

The main aim of this study was to find the point where actual changes take place in a genome, i.e. find the anomalies and their exact locations in the genome, which may be responsible for several functions in the genome such as disease causing or for performing vital genomic actions. We wanted to find the point of the segment which was significantly different from rest of the genomic sequence.

Materials and methods

Data descriptions

The *Bos taurus* (cow) genome data obtained from the National Centre for Biotechnology Information (NCBI), USA, genome database was considered for this study. The genome of *B. taurus* is 3000 MB in size. It is arranged into 29 pairs of autosomes and 2 sex chromosomes, according to current estimates. The chromosomes of cattle are acrocentric. The chromosome used in this study is the smallest, i.e. chromosome 25. Fasta file was downloaded for chromosome 25, with accession number NC_037352.1. The size of the genome under study was 42.35 Mb. Table 1 provides details of chromosome 25.

Variables were chosen for the study based on a few earlier studies to identify the actual change points present in geno-

mic sequences. Earlier, several studies were conducted based on simulated data and human genome data to know the relationship between Copy Number Variation (CNV) and gene expression data¹⁹. Apart from this, genetic association studies were also attempted using SNP, CNV and gene expression data²⁰. All the variables chosen signify specific properties, and we wanted to study their correlation and effect on each other over the entire genome.

The variable considered for this study was the GC content of the entire genome. This is also known as the G + C ratio or GC ratio. From Fasta file, GC content was extracted using R-script and the following formula: $(G + C / (A + G + C + T))$. CpG island, CNV and SNP were also initially considered for the study, but these variables have many null values in a given segment, which make no contribution when finding an outlier or an anomaly in a given segment. Therefore, only GC content is considered for this study. Besides, the GC variable also explains and provides information about the expression levels of genes, their thermostability active transcription, bendability of DNA and B-Z transition.

Data preparation was done using R software. Initially, the GC content of 1000 consecutive nucleotide sequences was considered as a region. The next region was from 1001 to 2000 and so on. Totally 42,350 regions with their respective GC content formed chromosome 25 of *B. taurus*. These data were plotted as a graph with respect to their quartile value. Summary statistics were obtained using R software. Table 2 summarizes the statistics of the 1st, 2nd (median) and 3rd quartile for GC data.

Methodology

Anomalies or outliers in statistics are data points that differ significantly from the others, as the name implies. That is, data values that appear to be out of phase with the other data values disrupt the broad distribution of a dataset. When the number of observations is minimal and one-dimensional, it is much easier to find these outliers. When

Table 1. Details of chromosome 25

Type	Name	Accession no.	Size (Mb)	GC%	Gene
Chr	25	NC_037352.1	42.35	47.1	1006

Table 2. Summary statistics of GC data used in this study

Statistics	GC
Minimum	0.2122
1st Quartile	0.4154
Median	0.4565
Mean	0.4705
3rd Quartile	0.5115
Maximum	0.8559

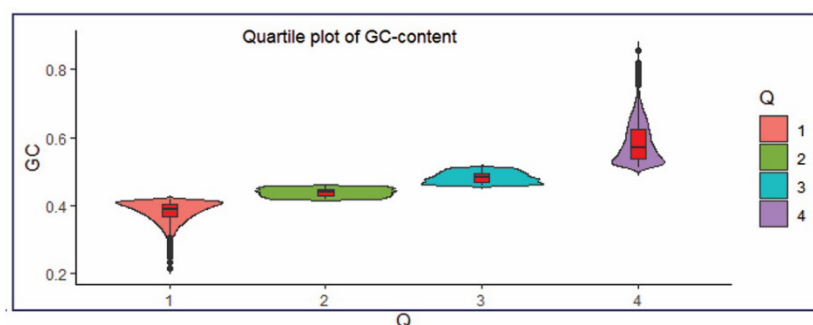


Figure 1. Distribution of GC over different quarters, viz. Q_1 , Q_2 , Q_3 and Q_4 .

there are thousands of observations or multi-dimensions, such as in genomic data, more precise and reliable methods of locating them are necessary.

Anomalies in a dataset can have a wide range of consequences for the data. The fact that the data are skewed is one of the most significant consequences, as it alters the general statistical distribution of data in terms of mean, variance and other properties. It could also imply a bias in the data used in the analysis. Anomalies might arise in genomic data because of inherent unpredictability also. In certain domains, anomalies are eliminated as they are caused by faulty techniques. However, in some areas, anomalies are preserved as they may contain valuable information, and their removal may result in the loss of sensitive data.

This work aimed to find anomalies in genomic data in order to pinpoint the specific time at which a data segment diverged considerably from the rest of the segments.

A number of techniques, ranging from simple descriptive statistics (such as minimum, maximum, histogram, boxplot and percentile) to more formal techniques like the Hampel filter, Grubbs, Dixon and Rosner tests and Cook's statistics are available to identify anomalies in a dataset. However, these methods have significant drawbacks. For example, Cook's statistics only apply to regression models, not univariate time-series data. Therefore, it was not used to detect anomalies. The Hampel filter method only returns a single outlier value for the entire dataset, which is insufficient for huge genomic datasets. Grubb's test detects only one outlier at a time, which is either a higher or lower value of data, which is inconsistent with the present study. The Dixon test is similar to Grubb's test for determining whether a single low or high result is an outlier. If more than one outlier is found, the Dixon test must be done on each one separately. This test is useful when the sample size is small ($n \leq 25$). The Rosner test is designed to limit the effects of swamping and masking, where an outlier with a similar value remains undetected. The Rosner test, unlike the Dixon test, works best when the sample size is large ($n > 20$).

Due to the closeness in data types, the time-series method of anomaly detection was used in this study to detect

anomalies because genomic data are location-specific and ordering is significant. Anomaly detection in time-series data is based on the decomposition of the data into the trend, seasonal and residual components. Since both genomic sequence and time-series data are position-specific, they are similar.

Let y_t represent the observation at the t th location. The additive decomposition can be represented as

$$y_t = T_t + R_t + S_t, \quad (1)$$

here T_t is the trend-cycle component, S_t the seasonality component and R_t is the remainder component.

In genomic data, the effect of change in position is less as compared to changing trends with the season in time series data, so additive decomposition is used here, and S_t is assumed to be zero. The idea is to first remove any trend (T_t) in the data and then find outliers in the remainder series (R_t). The final anomalies are estimated by the given equation. Anomalies are identified using \hat{R}_t values and finding interquartile ranges from them. If Q_1 denotes the 25th percentile and Q_3 the 75th percentile of the remainder values, then the interquartile range (IQR) is defined as

$$\text{IQR} = Q_3 - Q_1. \quad (2)$$

Observations are labelled as outliers/anomalies if they are $< Q_1 - 3(\text{IQR})$ or $> Q_3 + 3(\text{IQR})$ (ref. 21). If the remaining values are normally distributed, then the probability of an observation being identified as an anomaly is approximately 1 in 427,000. This method is given in the R-package 'forecast', which can be used to detect multiple outliers in a time-series data²².

An obvious question in the present study is whether these outliers/anomalies are actual change points or not. First, the distribution of data was tested using the Anderson–Darlings test and found to have a low P -value. To validate this further, the non-parametric Kolmogorov–Smirnov (KS) test was employed. Using a two-sample KS test, a comparison was made among consecutive segments concerning the identified anomalies.

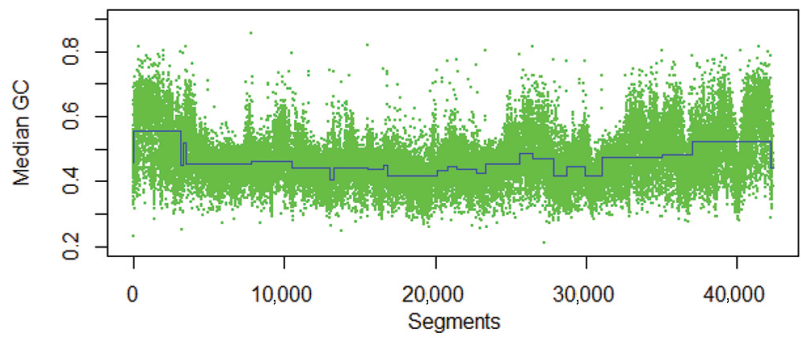


Figure 2. Segmentation based on mean GC content of *Bos taurus* chromosome 25.

Table 3. Anomalies identified along with their position on the genome

Anomaly no.	Position
1	38
2	3,174
3	3,350
4	3,484
5	7,775
6	10,473
7	12,983
8	13,248
9	15,542
10	16,544
11	16,800
12	16,867
13	20,100
14	20,819
15	21,403
16	21,506
17	22,677
18	23,300
19	25,592
20	26,438
21	27,846
22	28,685
23	29,889
24	30,999
25	35,053
26	37,018
27	42,246

Table 4. Change points obtained using Kolmogorov–Smirnov two-sample test

Change point	Anomaly position
38	38
3,174	3,174
3,350	3,350
3,484	3,484
7,775	7,775
10,473	10,473
12,983	12,983
13,248	13,248
15,542	15,542
16,544	16,544
16,867	16,800
20,100	16,867
20,819	20,100
21,403	20,819
22,677	21,403
23,300	21,506
25,592	22,677
26,438	23,300
27,846	25,592
28,685	26,438
29,889	27,846
30,999	28,685
35,053	29,889
37,018	30,999
42,246	35,053
	37,018
	42,246

Results and discussion

The total size of the *B. taurus* genome of chromosome 25 size was 42.35 Mb, whose GC content of 1000 consecutive nucleotides resulted in 42,350 regions using R-code. The distribution of GC content over different quartiles is presented in Figure 1 through a violin plot. It can be seen that quarter 4 has a much-elongated distribution compared to the other quarters. A violin plot shows the distribution pattern of data pertaining to GC over different quartiles. As can be seen in Figure 1, the data have a skewed distribution.

The total number of anomalies identified was 27 (Table 3).

The next step was to determine whether these anomalies were actually change points or not. For this, the data were first checked for normality using the Anderson–Darling test. For the given dataset, the P -value obtained was $<2.2e^{-16}$, which is very low. It also signified that the data were not normally distributed. Hence, the KS two-sample test was used to test consecutive segments with respect to the identified anomalies. Twenty-five out of 27 anomalies were identified as actual change points. Table 4 presents the identified change points.

Figure 2 depicts the 25 change points obtained using this two-step procedure. This method has resulted in the

identification of change points based on the anomalies detected and efficiently reduces computational time and cost.

As the data size was large, depicting all points in a single figure made it challenging to locate change points properly. To show these segments with more clarity in the plot, between two change points identified, the median value of all datasets was taken to represent a single value. This helps in obtaining a proper plot and a more insightful graph.

Conclusion

A change point in genomic data marks the boundary of two segments; hence, the identification of change points is the most common method of segmentation. The segments differ from each other with respect to statistical and biological properties. On the other hand, outliers or anomalies in data are simply observations having significantly different values or properties compared to the overall dataset. Influential positions or anomaly detection procedure has been described and the same has been investigated in cow genome data. Further, the anomalies detected have been validated to know whether they are actual change-points or not.

As discussed earlier, the common outlier detection procedures are not applicable for location-dependent genomic data like cattle chromosome 25 GC data. Hence, a time-series approach was employed in this study. Initially, 27 anomalies were found in the data. Further, the anomalies detected were validated to determine whether they were actual change points or not. Out of 27 anomalies, 25 were identified as change points using a two-sample KS test, resulting in 26 segments. KS test is the most powerful non-parametric test accounting for the non-normal nature of the data. The two-step procedure resulted in the identification of change points based on the anomalies detected. It is efficient in terms of reduction in computational time and cost. Relevant R codes for data processing and extraction have been developed.

Apart from location dependency, the large size of genomic data poses a major challenge in segmentation. The present study can also be extended to the multivariate scenario. Due to the computational infeasibility of the KS test in multivariate data, we propose using Cramer's test, a powerful multivariate two-sample test^{23,24}. The source code for implementing the proposed method is available with the authors and can be accessed upon request.

Conflict of interest: The authors declare that they have no conflict of interest.

1. Braun, J. V. and Muller, H. G., Statistical methods for DNA sequence segmentation. *Stat. Sci.*, 1998, **12**(2), 142–162.
2. Zhang, Z., Lange, K. and Sabatti, C., Reconstructing DNA copy number by joint segmentation of multiple sequences. *BMC Bioinform.*, 2012, **13**(1), 1–15.
3. Erdman, C. and Emerson, J. W., A fast Bayesian change point analysis for the segmentation of microarray data. *Bioinformatics*, 2008, **24**(19), 2143–2148.

4. Zeller, S. R., Henz, S., Laubinger, D., Weigel, and Rättsch, G., Transcript normalization and segmentation of tiling array data. *Pac. Symp. Biocomput.*, 2008, **13**, 527–538.
5. Bischler, T., Kopf, M. and Voß, B., Transcript mapping based on dRNA-seq data. *BMC Bioinform.*, 2014; doi:10.1186/1471-2105-15-122.
6. Elhaik, E., Graur, D. and Josić, K., Comparative testing of DNA segmentation algorithms using benchmark simulations. *Mol. Biol. Evol.*, 2010, **27**(5), 1015–1024.
7. Girimurugan, S. B., Liu, Y., Lung, P. Y., Vera, D. L., Dennis, J. H., Bass, H. W. and Zhang, J., iSeg: an efficient algorithm for segmentation of genomic and epigenomic data. *BMC Bioinform.*, 2018, **19**(1), 1–15.
8. Fearnhead, P. and Liu, Z., On-line inference for multiple change point problems. *J. R. Stat. Soc.: Ser. B*, 2007, **69**(4), 589–605.
9. Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M., Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 2004, **5**(4), 557–572.
10. Zhang, N. R. and Siegmund, D. O., A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 2007, **63**, 22–32.
11. Jeng, X. J., Cai, T. T. and Li, H., Optimal sparse segment identification with application in copy number variation analysis. *J. Am. Stat. Assoc.*, 2010, **105**(491), 1156–1166.
12. Snee, R., Regression diagnostics: identifying influential data and sources of collinearity. *J. Qual. Technol.*, 1983, **15**(3), 149–153.
13. Cook, R. D., Detection of influential observation in linear regression. *Technometrics*, 1977, **19**(1), 15–18.
14. Cook, R. D., Influential observations in linear regression. *J. Am. Stat. Assoc.*, 1979, **74**, 169–174.
15. Peña, D., A new statistic for influence in linear regression. *Technometrics*, 2005, **47**, 1–12.
16. Budhlakoti, N., Rai, A. and Mishra, D. C., Statistical approach for improving genomic prediction accuracy through efficient diagnostic measure of influential observation. *Sci. Rep.*, 2020, **10**(1), 1–11.
17. Hayes, B. and Goddard, M., Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 2001, **157**, 1819–1829.
18. Lourenço, V. M. and Pires, A. M., M-regression, false discovery rates and outlier detection with application to genetic association studies. *Comput. Stat. Data Anal.*, 2014, **78**, 33–42.
19. Ortiz-Estevéz, M., De Las Rivas, J., Fontanillo, C. and Rubio, A., Segmentation of genomic and transcriptomic microarrays data reveals major correlation between DNA copy number aberrations and gene-loci expression. *Genomics*, 2011, **97**(2), 86–93.
20. Momtaz, R., Ghanem, N. M., El-Makky, N. M. and Ismail, M. A., Integrated analysis of SNP, CNV and gene expression data in genetic association studies. *Clin. Genet.*, 2018, **93**(3), 557–566.
21. Tukey, J. W., Exploratory data analysis. *Addison-Wesley Ser. Behav. Sci.*, 1977, **2**, 131–160.
22. Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M. and Wang, E., Package 'forecast'; <https://cran.r-project.org/web/packages/forecast/forecast.2020>
23. Baringhaus, L. and Franz, C., On a new multivariate two-sample test. *J. Multivar. Anal.*, 2004, **88**, 190–206.
24. Justel, A., Peña, D. and Zamar, R., A multivariate Kolmogorov–Smirnov test of goodness of fit. *Stat. Probab. Lett.*, 1997, **35**(3), 251–259.

ACKNOWLEDGEMENTS. The first author (A.A.) thanks Professor Zhiwu Zhang (Washington State University, Pullman, USA) for guidance during an academic visit. Financial assistance received from the Indian Council of Agricultural Research (ICAR), New Delhi through the National Agricultural Higher Education Project and UGC-MANF fellowship is duly acknowledged. We thank Indian Agricultural Research Institute, ICAR-Indian Agricultural Statistics Research Institute and ICAR-Central Marine Fisheries Research Institute for providing the necessary facilities to carry out this study.

Received 15 September 2022; accepted 17 September 2023

doi: 10.18520/cs/v126/i1/54-58