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

Arfa Anjum[1], Seema Jaggi[2,*], Shwetank Lall[3], Eldho Varghese[4], Anil Rai[1], Arpan Bhowmik[3] and Dwijesh Chandra Mishra[1]

[1]Centre for Agricultural Bioinformatics,
[2]Agricultural Education Division, and
[3]Division of Design of Experiments, ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012, India
[4]Fishery 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.

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 prediction[1]. 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 loci or operons[5] 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