ESTIMATING THE LOCATION OF CHANGE-POINTS IN DNA SEQUENCES VIA THE CROSS-ENTROPY METHOD

Sofronov G.Yu., Evans G.E., [1]Keith J.M., Kroese D.P.

Department of Mathematics, The University of Queensland, Brisbane, Qld 4072, Australia
[1]School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434 Brisbane, Qld 4001, Australia

The genomes of complex organisms, including the human genome, are known to vary in GC content along their length. That is, they vary in the local proportion of the nucleotides G and C, as opposed to the nucleotides A and T. Changes in GC content are often abrupt, producing well-defined regions. We model the problem as a multiple change-point problem, that is, a problem in which sequential data is separated into segments by an unknown number of change-points, with each segment supposed to have been generated by a different process. Multiple change points are important in many biological applications and, particularly, in analysis of DNA sequences. Besides the aforementioned problem, multiple change points can be applied in segmenting protein sequences (20 character alphabet) according to hydrophobicity. We are also interested in identifying segments that are conserved between two species. We use the Cross-Entropy method to find estimates of change-points as well as parameters of the process on each segment. Numerical experiments have illustrated the effectiveness of the approach. We obtain estimates for the locations of change-points in artificially generated sequences and compare the accuracy of these estimates to those obtained via other methods such as IsoFinder and MCMC. Lastly, we provide examples with real data sets to illustrate the usefulness of our method.