seq2R: Detecting DNA compositional change points

Nora M. Villanueva1, Marta Sestelo1, Javier Roca-Pardiñas1

1. Department of Statistics and Operation Research, University of Vigo
*Contact author: nmvillanueva@uvigo.es

Keywords: kernel, bootstrap, DNA, change points, derivatives

Identifying compositional change points in a statistical framework can be a challenging task. Numerous methodological approaches have been developed to analyse change points models, i.e. Bayesian estimation, maximum-likelihood estimation, least squares regression or nonparametric regression. Part of our philosophy is to make easier for others to use a new statistical methodology. With that in mind, we implement in a user-friendly and simply R package, seq2R, a methodology that identifies and locates compositional change points in DNA sequences by fitting regression models and their first derivatives. Since the estimation procedure of this methodology with large datasets implies a high computational cost, Fortran is used as the programming language. Our approach to assess the change points, is based on detecting the maximum or minimum of the first derivative of nonparametric regression models with binary response. To estimate the regression curve and its first derivative, we apply local linear kernel smoothers. Additionally, bandwidths are automatically selected using cross-validation techniques. Inference implies the construction of confidence intervals which can be obtained by bootstrap methods. The choice of the bandwidth and the usage of bootstrap resampling techniques may entail high computational cost. To considerably reduce this cost and render the operational procedures, we apply binning techniques.