seq2R: Analyzing compositional asymmetries in DNA

Nora M. Villanueva[1*], **Marta Sestelo**[2], **Javier Roca-Pardiñas**[1]

1. Department of Statistics and Operation Research, University of Vigo, Spain
2. Department of Mathematics, University Autonomous of Barcelona, Spain
*Contact author: nmvillanueva@uvigo.es

Keywords: DNA sequence, change points, nonparametric models, testing procedure, first derivatives.

Understanding the mutational processes that shape DNA sequences is fundamental to better comprehend how genomes evolve. These mutations do not equally affect both complementary strands of DNA when these mutations are associated with molecular processes that are also asymmetric affecting differently both strands (e.g. transcription, DNA repair or replication; Touchon et al., 2005). Over the years, different compositional analyzes were carried out to detect the location of compositional changes points in mitochondrial genomes (Grigoriev, 1998; Reyes et al., 1998, Faith and Pollock, 2003). Identifying these change points in a statistical framework can be a challenging task. Numerous methodological approaches have been developed to analyze change points models, i.e. Bayesian estimation, maximum-likelihood estimation, least squares regression or nonparametric regression. 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 nonparametric regression models. Our procedure is based on two steps. Firstly, we propose an initial approach of the regions with possible change points in which the first derivative is different to zero. The regression curve and its first derivative are estimated by local linear kernel smoothers (Fan and Gijbels, 1996; Wand and Jones, 1995) and the bandwidths are automatically selected using cross-validation techniques (Golub et al., 1979). Secondly, we asses if there are true change points in those regions, specifically, where and how many they are, with a testing procedure.

References

Faith, J. J., Pollock, D. D., (2003). Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* **165**, pp.735–745.

Fan, J and Gijbels, I (1996). *Local polynomial modelling and its applications. Monographs on statistics and applied probability series 66*. Chapman & Hall.

Golub, G. H., Heath, M. and Wahba, G. (1979). Generealized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, pp. 215–56.

Grigoriev, A., (1998). Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Research* **26** (10), pp. 2286–2290.

Reyes, A., Gissi, C., Pesole, G., Saccone, C., (1998). Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Molecular Biology and Evolution* **15** (8), pp. 957–66.

Touchon, M., Rocha, E. P., (2008). From GC skews to wavelets: A gentle guide to the analysis of compositional asymmetries in genomic data. *Biochimie* **90** (4), pp. 648–659.

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing.*. Chapman & Hall.