Binary autoregressive geometric modelling in a DNA context

Sónia Gouveia[1], Manuel G. Scotto[2], Christian H. Weiβ[3], Paulo Jorge S. G. Ferreira[4]

1 — IEETA & CIDMA, University of Aveiro
2 — Department of Mathematics & CEMAT, IST, University of Lisbon
3 — Helmut-Schmidt University, Hamburg, Germany
4 — Department of Electronics, Telecommunications and Informatics & IEETA, University of Aveiro

Symbolic sequences occur in many contexts and can be characterized by integer-valued intersymbol distances or binary-valued indicator sequences. The analysis of the numerical sequences often sheds light on the properties of the original symbolic sequences. This work introduces new statistical tools to explore autocorrelation structure in indicator sequences, with application to deoxyribonucleic acid (DNA) sequences.

DNA is a long A, C, G and T sequence from which 4 sequences of inter-nucleotide distances (IND) can be derived as the consecutive distances between equal nucleotides. It is known that IND probability distributions deviate significantly from those assuming independent random placement (i.e. geometric distributions) and the deviations can be used to discriminate between species and to build phylogenetic trees. To investigate the extent to which autocorrelation explains the deviations, each 0–1 indicator sequence is endowed with a binary autoregressive (AR) model of optimum order. The corresponding binary AR geometric distribution is derived analytically and compared with the observed IND distribution by goodness-of-fit X2-testing.

The figure shows observed/expected frequencies for the human mitochondrial DNA: the expected optimum BinAR frequencies (red) are better adjusted to the observed ones (grey) than those expected for the geometric distribution (black). Overall results from several species (GenBank, http://ncbi.nlm.nih.gov/genbank) indicate that the statistical hypothesis of equal observed/expected frequencies is seldom rejected with a binary AR model instead of independence (76/136 vs 125/136 rejections at 1% level). Furthermore, binary AR modelling leads to a relevant median observed/expected deviation reduction (30% for A, 80% C, 90% G, 60% T). Therefore, these models are suitable to describe the dependences within a given nucleotide and encourage the development of a model-based framework to compact IND information.

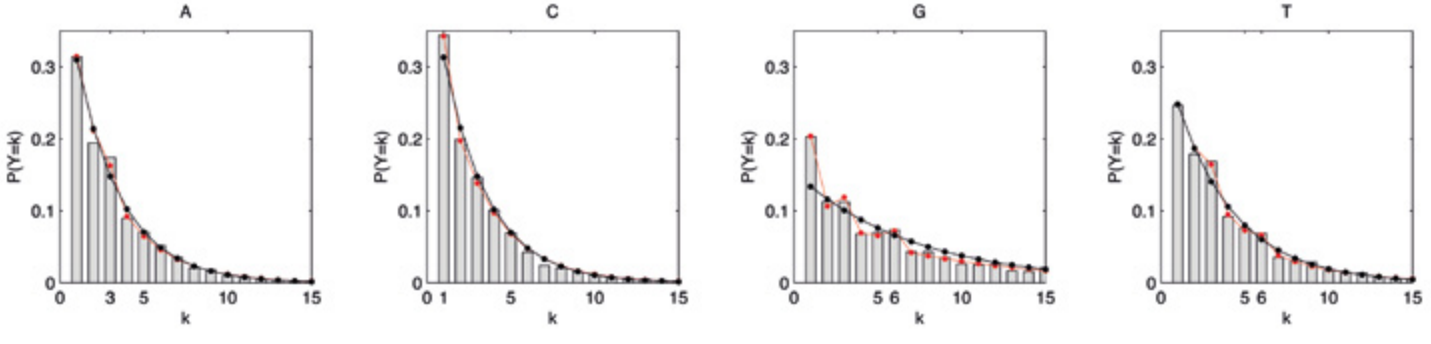

FIGURE 1
Probability mass functions of IND values (Y) for the Human species: observed frequencies (grey), expected geometric frequencies (black) and optimum order expected BinAR frequencies (red) with parameters estimated from the data. Chosen optimum order is (3,1,6,6) respectively for (A, C, G, T) nucleotides.