Case study 1: Detecting CG-rich genome regions

Real DNA sequences are heterogeneous throughout genomes in terms of the nucleotide usage, particularly CG content within and between genomes has been of interest (eg, it can be used as one of features for gene prediction and has been linked to optimal growth temperatures in prokaryotes). See example figure below (from Melodelima & Gautier (2008) BMC Genomics 9:632; doi.org/10.1186/1471-2164-9-632).

Formally define an HMM with two hidden states for annotating a given DNA sequence with CG-rich (*R*) and CG-poor (*P*) regions. CG-rich sequences have high content of nucleotides C and G. Hidden states *R* and *P* have equal initial probabilities, and transition probabilities are a_{RR}=0.5, a_{RP}=0.5, a_{PR}=0.6, a_{PP}=0.4. Nucleotides T, C, A, and G are emitted from states *R* and *P* with probabilities 0.2, 0.3, 0.2, 0.3 and 0.3, 0.2, 0.3, 0.2 respectively. Draw a state diagram for this model. For tasks 1 and 2, write down all the steps and calculations manually first, then implement the same procedures in R or python.

(1) Use a Viterbi algorithm to define the most likely sequence of hidden states for sequence GGACTGAA.

(2) Find P(*x*) the probability of sequence *x* = GGCA by both forward and backward algorithms. Repeat the computation for the forward algorithm using the scaling variables (see annex).

(3) Use this model to annotate a real DNA sequence (found in the file "Sequence_case1.txt"). Do you think this model is good at describing this data? Explain and propose a simple way to improve this model.

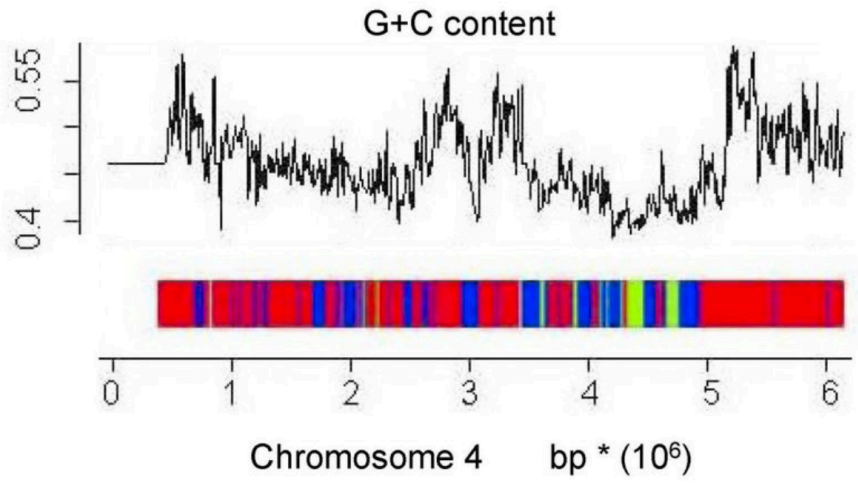

Figure 4
Distribution of isochores along *T nigroviridis* chromosome 4. The detected H, L and M isochores are shown in red, green and blue, respectively. To check the consistency of isochore prediction, the graph is shown alongside a plot of the GC content along the chromosome.