MSc in Biostatistics
Bayesian Inference: Project 1

The genetic information of an organism is encoded in DNA molecules, arranged on structures called *chromosomes*, in its cells. DNA molecules are polymers consisting of long sequences of monomers called nucleotides, each of which contains a base. There are four types of base: adenine (A), guanine (G), cytocine (C), and thymine (T). Interest lies in the *isochore* structure of a given chromosome. Isochores are regions of the chromosome where the proportion of bases being of type C or G is roughly constant.

The file dataproject1.R contains 5 data sets of CG content along regions of Human chromosome 1. Each data set consists of the number of bases that are C or G within 100 consecutive windows of 5,000 bases. You are asked to decide whether each data set comes from one or two Isochores.

Write a report presenting your solutions and answering the above question. Submit your report together with a printout of the R code you used to obtain the numerical results. The deadline for this project is the day of the final exam.

Approach

1. Choose a model for the number of CG bases in a window of 5000 in one isochore. This will give you the likelihood of the observed number of CG bases in a window of 5000 bases in that isochore.

2. Consider suitable models to represent your data under hypotheses for (i) a single isochore; (ii) two isochores.

3. Perform a Bayesian analysis to choose between your two models; you will need to choose a (and justify your choice of) prior.

4. You should present suitable quantitative results and a discussion which must answer the question that is set.