OXWASP PROBABILITY AND APPROXIMATION MODULE MICRO-PROJECT: CHANGE POINTS FOR GENOMIC DATA

One natural model for a long sequence of categorical observations $x_1, x_2, \ldots, x_n$ is a discrete Markov chain: Suppose there are k categories, numbered $1, \ldots, k$. Given that $x_t = i$, there are probabilities P_{ij} that $x_{t+1} = j$, for $j = 1, \ldots, k$, independent of any of the states $x_1, \ldots, x_{t-1}$ preceding x_t. If these probabilities remain the same the Markov chain is called *homogeneous*; otherwise, it is *inhomogeneous*.

In many cases it is useful to model the sequence as *piecewise Markov*, with different Markov transition probabilities in different parts of the sequence. Deciding where the changepoints are, the points in the sequence where one transition probability matrix gives way to another, is then an important statistical problem.

Consider how you would design an algorithm to detect changepoints in a long sequence. Things to consider:

(1) Do you want to group the sequence into fixed blocks, or try to be more flexible? If fixed blocks, what might be a reasonable criterion for choosing the block size?
(2) How do you estimate the transition probabilities, and represent the stochastic uncertainty?
(3) How should you set a threshold for a change having occurred?

Try applying your ideas either to simulated data or to genomic sequence data, such as one of the chromosomes of the nematode C. elegans. (The downloads are in the `FASTA` format, which may be read into `R` using the `read.fasta` command in the `inseqr` package.) As a reminder, the genome of nearly all organisms consist of long sequences (chromosomes) of *DNA nucleotides*, of one of four types, denoted by the letters A, C, G, T. Changes in distribution and transition probabilities may represent important structural regions of the chromosome.