Genomes in 3D: some issues in multidimensional scaling

Wally Gilks

Department of Statistics, University of Leeds

There is growing evidence that the three-dimensional (3D) layout of genomes (the set of chromosomes) within in the nucleus of a cell is conserved over millions of years. Over recent years, laboratory techniques have been developed that provide increasingly high-resolution information on the 3D structure of the genome of any species. The latest of these techniques, called 5C and Hi-C, use high-throughput DNA sequencing to produce millions of pairs of short DNA reads, each pair identifying two small pieces of the genome which are juxtaposed within the nuclear space. This allows a matrix of DNA-DNA contacts within the genome to be compiled.

In principle, these contact matrices hold information on the 3D configuration of the genome. Extracting this information may be done using multidimensional scaling. However, this is not entirely straightforward. The data contain substantial amounts of noise and the experimental procedures require aggregation of contacts over millions of cells, introducing an additional source of variability.

After briefly introducing the biological, experimental and bioinformatic background, I will present data analyses, simulation results and statistical theory which highlight difficulties in analysing the resulting contact matrices, and describe some preliminary statistical methods to address some of the issues arising.