Advanced R Programming for Bioinformatics.

Exercises for Session 10: Large data

1. The file `sisg.nc` is a netCDF file with a subset of SNP data from the HapMap project.
(a) Read in the first ten SNPs for each person
(b) For each person, compute the proportion of SNPs at which they are heterozygous (i.e. have genotype==1)

2. The file `SEAflights.db` is a SQLite database with the same data as `SEAflights.csv`. Read in the arrival and departure delays for all flights from SFO.

3. Create a SQLite database with the data from `sisg.nc` and compare the speed of reading and writing in the two formats.

4. Design an R class as a front-end to netCDF files as follows;
(a) the object will store the connection to the netCDF file
(b) a method for '[' stores which rows/columns are selected, but does not read or modify the file
(c) a method for `as.matrix()` (1 or 2 dimensional) returns the object converted to a matrix, i.e. reads in the data
(d) for enthusiastic people: the object optionally stores a function as a transformation for each variable, which is applied when the data are read in