Interfacing With Common Resources: NetCDF

Nishant Gopalakrishnan
Fred Hutchinson Cancer Research Center

17-18 February, 2011
Introduction

ncdf R package

Resources
Handling Large Data

- text files
- binary files
- read and process in chunks
- NetCDF
Network Common Data Form (NetCDF)

Set of data formats, programming interfaces software libraries to read/write scientific data

- NetCDF software library
  - NetCDF version 3
  - NetCDF4

- R packages
  - ncdf
    - Warning: character array implementation is inefficient.
  - ncdf4
    - Multiple unlimited dimensions
    - Data compression
    - Not available on Windows
NetCDF Model

- **Variables**: N-dimensional arrays of data: char, byte, int, double
- **Dimensions**
  - Axes of data arrays
  - Name, length
  - Unlimited dimension
- **Attributes**: Annotate Variables with meta data
Using ncdf R package

► Define dimensions
  > sampDim <- dim.def.ncdf("Sample", "id", seq_len(NROWS))
  > snpDim <- dim.def.ncdf("SNP", "id", seq_len(NCOLS))

► Define variable
  > snpDat <- var.def.ncdf("Genotype", units ="", dim = list(sampDim, snpDim),
  + missval = 0L, prec = "byte")
Using ncdf R package

- Create file, write variable
  ```
  nc <- create.ncdf("myFile.nc", snpDat)
  put.var.ncdf(nc, "Genotype", mat)
  close(nc)
  ```
Using ncdf R package

- Read variable from file
  ```r
  nc <- open.ncdf("myFile.nc")
  myVar <- get.var.ncdf(nc, "Genotype")
  close(nc)
  ```

- Read slices of data
  ```r
  nc <- open.ncdf("myFile.nc")
  sliceRow <- get.var.ncdf(nc, "Genotype", start = c(30, 100), count = c(10, 20)) ## samples 30:40, snps 100:120
  sliceRow <- get.var.ncdf(nc, "Genotype", start = c(1, 1000), count = c(-1, 100)) ## all samples, snps 1000:1100
  close(nc)
  ```
Resources

- http://cran.r-project.org/web/packages/ncdf/ncdf.pdf
- http://www.unidata.ucar.edu/software/netcdf/docs/netcdf-tutorial.html