assignment 2 part 2

Anders Albrechtsten

September 11, 2020

1 Estimate allele frequency on the mitochondrion

In this part of the assignment you will estimate allele frequencies of the mitochondrion.

1.1 The data

We will use one of the 1000 Genomes individuals that has been sequencing using illumina short read sequencing. The reads have been mapped to the human reference genome. To get the pileup format we can use the command

```
samtools mpileup MT/MTNA12003.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.bam \
 -r MT: > mt.pileup
```

This however contains information regarding the read start points and also indels. To make things a bit easier I have made a clean version without this information for you to use in this assignment. You can find it here:

```
/data/albrecht/advBinf/MT/MTnice.pileup
```

2 Model

The assignment is to estimate the allele frequencies of the four alleles (A,C,G,T) at each position of the mitochondrion of the chosen individual. Note that the mitochondrion is haploid but there might still be several versions present in each individual which is known as Heteroplasmy[1] and in this case the allele frequencies of the four allele in a given locus may not all be 0 or 1. You should make a model that estimates the four allele frequencies at each position an the four allele frequencies should sum to one. The model should take in to account that mitochondrion is haploid and it should take into account possible errors in the sequencing data which is reflected in the base quality score.

[1] https://en.wikipedia.org/wiki/Heteroplasmy

- Write the likelihood model that uses both the observed bases and the quality scores for a single site. The frequencies of the four bases are the parameters. Remember that the true bases are not observed. NB! you have to explain all of the notation i.e. variables/parameters you use in your model
- Write the Q (estimation) and M step of the EM algorithm that you will need for the optimization. Use the same notation (variables/parameters) as you use to describe the likelihood.

3 Implement

Implement the EM algorithm for the model and estimate the fractions. For this data the base quality offset is 33 (most common offset). Therefore, you will have to convert from ASCII to Integer, offset with 33 and then convert to a probability using the inverse of the phredScaling. Example in R of reading in data and converting the quality scores into values

```
r<- read.delim("MTnice.pileup",as.is=T,head=F)
bases<-strsplit(r[,5],"")
quality<-strsplit(r[,6],"")

fun<-function(x){
   y <- R.oo::charToInt(x)-33 #offset
   10^(-y/10)
}
quality<-lapply(quality,fun)
quality[[1]]

dat <- read.delim("MTnice.pileup",as.is=T,comment.char="",head=F)
names(dat) <- c("CHR","POS","REF",c("depth","bases","Qscore"))
head(dat,1)

library(R.oo) #package with charToInt function to convert ascii to a value

## bases for individual 1 as a list
bases <- strsplit(dat$bases,"")
## bases for site 2 (note the [[ notation for lists)
bases[[2]]
## base on read 3 for for site 2
bases[[2]][3]

## ascii qualities for individual 1
asciiQ <- strsplit(dat$Qscore,"")
##  quality values values
Q <- lapply(asciiQ,function(x) R.oo::charToInt(x) - 33 )
```

```
## quality  for site 2
Q[[2]]
## qualty values for base on read 3 for site 2
Q[[2]][3]
```

NB. If your implementation is very slow then note that some calculations do not depend on the parameters and therefore does not need to be calculated in each EM iteration.

- Estimate the allele frequencies for all sites
- How many sites are there where the allele frequency of the most common allele is less than 0.9?

4 Hand in

You should hand in a pdf with your answers alongside the code (R,python,C/C++) you wrote for your analysis