TP- PAM

Madalina Ciortan

December 17, 2016

TP: Substitution matrices

Read data from the file (note that columns names and row names should be discarded for subsequent computation and that the values given should be divided by 10000).

```
setwd("C:\\workspace\\bio-info-db\\pam")
P1<-read.table('pam1.mat', sep = "\t", header=T, colClasses=c("character",rep("integer",20)))
P1 <- P1[1:20,2:21]
P1 <-as.matrix(P1)

P1 <- P1/10000
#Print P1 diag
diag(P1)
```

```
##  [1] 0.9867 0.9913 0.9822 0.9859 0.9973 0.9876 0.9865 0.9935 0.9912 0.9872
## [11] 0.9947 0.9926 0.9874 0.9946 0.9926 0.9840 0.9871 0.9976 0.9945 0.9901
```

Compute the probability matrix corresponding to any PAM distance n (corresponding to an evolutionary distance of n PAM, i.e. the PAMn probability matrix).

```
mypower <- function(x,n){
  p<-x
  prod<-diag(nrow(x))
  while ( n > 0 ){
    if(n %% 2 == 1) {
      prod <- prod %*% p
    }
    n <- floor(n/2)
    p <- p %*% p
  }
  prod
}
P10 <- mypower(P1, 10)
diag(P10)
```

```
##  [1] 0.8759146 0.9167487 0.8368855 0.8695232 0.9733598 0.8834091 0.8746325
##  [8] 0.9372861 0.9158135 0.8800826 0.9486147 0.9290079 0.8811620 0.9476147
## [15] 0.9286849 0.8525370 0.8792085 0.9762689 0.9466169 0.9063296
```

Compute the scoring matrix corresponding to your PAMn matrix.

score(i,j) = log(Pn / fj)

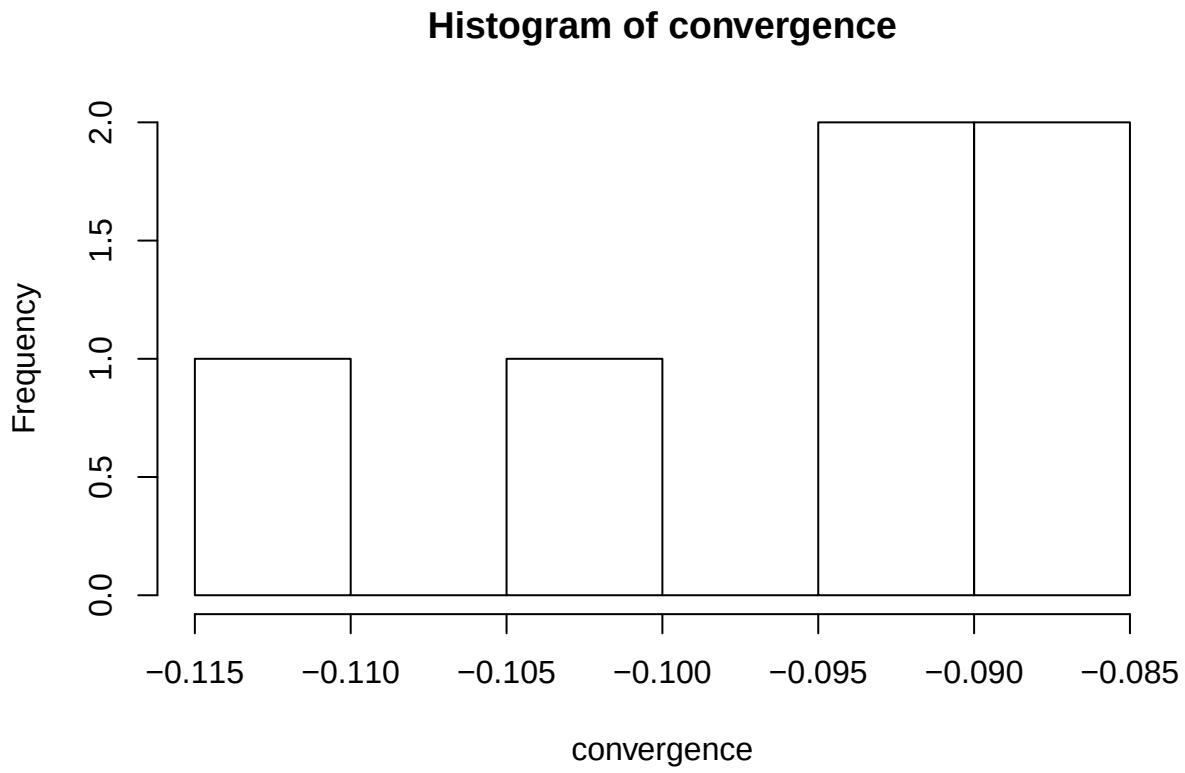
```
freq <-read.table('aafreq.mat', sep = "\t", header=F)
f <- freq$V2

score <- function(x, i, j) {
  log10(x/f[i])
}
```

```
scoringMatrix <- function(mat) {
  matrix(mapply(score, mat, row(mat), col(mat)), nrow = nrow(mat))
}

diag(scoringMatrix(P10))
```

```
##  [1] 1.002943 1.349466 1.320606 1.267183 1.469759 1.366378 1.242856
##  [8] 1.022482 1.430328 1.376322 1.047671 1.059534 1.768965 1.374572
## [15] 1.260298 1.167285 1.098994 1.989569 1.499053 1.144373
```

Check the convergence of the values in the columns of the PAM probability matrix when n becomes large: These values should approach the amino acid frequencies.

```
convergence <- c (sum (f - diag(mypower(P1, 1500))),
                  sum (f - diag(mypower(P1, 1650))),
                  sum (f - diag(mypower(P1, 1800))),
                  sum (f - diag(mypower(P1, 2000))),
                  sum (f - diag(mypower(P1, 2150))),
                  sum (f - diag(mypower(P1, 2300))))

hist(convergence)
```



Plot the relation between the evolutionary distance (in PAM unit) and the percentage of identity between the sequences

```
nIndexes <- seq(1,1500, 20) # take n values at 20 Pam distance from 1 to 1500
distance <- function (value) {
```

```
  sum(diag(mypower(P1, value)))
}

hist(sapply(nIndexes, distance), nclass = 50)
```