Bioinformatika

A. Blejec

12. december 2012

Povzetek

⋈

Kazalo

1 Introduction to R

http://ablejec.nib.si/R

2 FASTA

Metoda opisna v **??** na strani 74:

```
> set.seed(1234)
> codes <- c("A","C","T","G")
> n <- 100
> x <- sample(codes,n, replace=TRUE)
> x <- paste(x,sep="",collapse="")
> x
[1] "ATTTGTAATTTTCGCGCCAACCAAAGTGGACCCTAGACGGTTCTCTTCAGACTTATCGAGGACAATCTATAGA
```

Za preglednejši način izpisa naredim funkcijo

```
> wrap <- function(x,n=50){
+ first <- seq(1,nchar(x),n)
+ return(substring(x,first,first+n-1))
+ }
> wrap(x)
```

```
[1] "ATTTGTAATTTTCGCGCCAACCAAAGTGGACCCTAGACGGTTCTCTTCAG"
[2] "ACTTATCGAGGACAATCTATAGAGATCACTGCATAGCCAGAGAAATCACT"
```

```
> k <- 2
> n <- nchar(x)
> words <- sapply(1:(n-2),FUN=function(x,sequence,k=1) substring(sequence,x,x+
> tbl <- table(words)
> tbl
```

```
words
AA AC AG AT CA CC CG CT GA GC GG GT TA TC TG TT
 7  7  9  8  8  5  4  6  9  4  3  3  6  8  3  8
```

Funkcija za iskanje besed v zapisih

```
> myGrep <- function(x,words) grep(x,words)
```

'Hash table'

```
> y <- lapply(sort(unique(words)),myGrep,words=words)
> names(y) <- sort(unique(words))
> str(y)
```

```
List of 16
 $ AA: int [1:7] 7 19 23 24 64 93 94
 $ AC: int [1:7] 20 30 37 51 62 78 98
 $ AG: int [1:9] 25 35 49 59 71 73 85 89 91
 $ AT: int [1:8] 1 8 55 65 69 75 83 95
 $ CA: int [1:8] 18 22 48 63 77 82 88 97
 $ CC: int [1:5] 17 21 31 32 87
 $ CG: int [1:4] 13 15 38 57
 $ CT: int [1:6] 33 43 45 52 67 79
 $ GA: int [1:9] 29 36 50 58 61 72 74 90 92
 $ GC: int [1:4] 14 16 81 86
 $ GG: int [1:3] 28 39 60
 $ GT: int [1:3] 5 26 40
 $ TA: int [1:6] 6 34 54 68 70 84
 $ TC: int [1:8] 12 42 44 47 56 66 76 96
 $ TG: int [1:3] 4 27 80
 $ TT: int [1:8] 2 3 9 10 11 41 46 53
```

3 Regularni vzorci

Imena in priimki iz naslovov

```
> email <- c("Miha.Novak@nib.si","Micka.Podlogar@gmail.com")
> email
[1] "Miha.Novak@nib.si"         "Micka.Podlogar@gmail.com"
> exp <- "(.*)\\.(.*)(\\@.*)"
> imePriimek <- gsub(exp,"\\1 \\2",email)
> imePriimek
[1] "Miha Novak"      "Micka Podlogar"
```

Zamenjava vrstnega reda besed

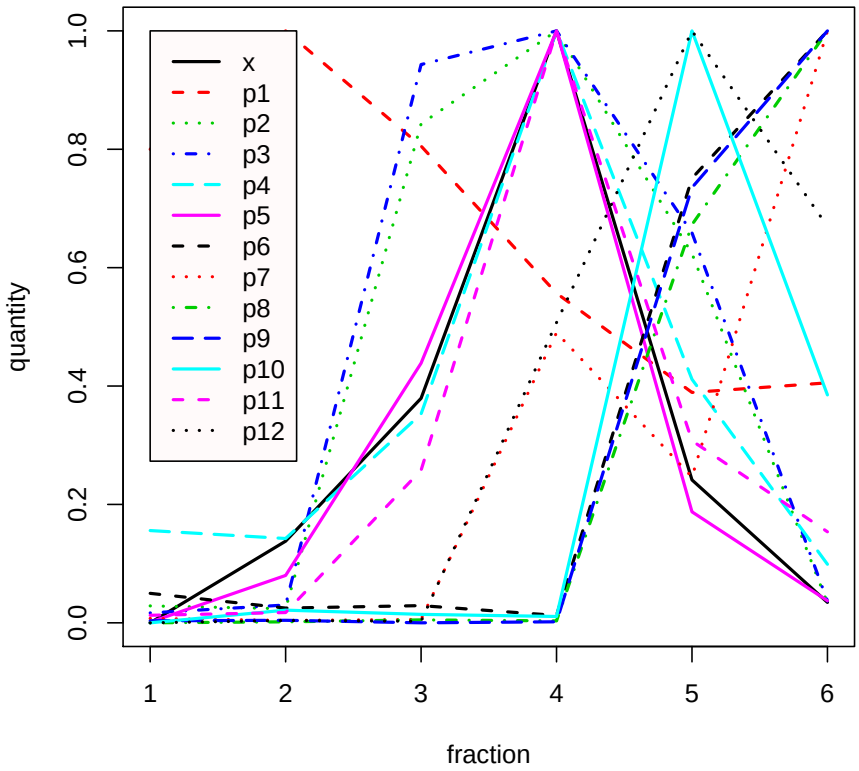
```
> imePriimek <- c("Miha Novak","Micka Podlogar")
> exp <- "(.*) (.*)"
> priimekIme <- gsub(exp,"\\2 \\1",imePriimek)
> priimekIme
[1] "Novak Miha"      "Podlogar Micka"
```

4 Primeri R

Primeri iz knjige Building Bioinformatics Solutions http://bixsolutions.net/the-book/

4.1 Simple R program to load a data matrix, scale it and plot the result

```
> # PROFILES.R
> #
> # Simple R program to load a data matrix, scale it and plot the result
>
> ## clear out the workspace first
> #rm(list = ls())
>
> # load data frame from web site
> X <- read.table("http://www.bixsolutions.net/profiles.csv", sep=",", header=
> # rangescale data by dividing by the maximum value in each column
> Xmax <- apply(X,2,max)
> Xscaled <- scale(X, scale=Xmax, center=FALSE)
> # plot columns in matrix as lines on a single graph
> matplot(Xscaled,type="l",xlab="fraction",ylab="quantity",col=1:6,lty=1:5,lwd
> # add legend to graph
> legend(x=1,legend=names(X),col=1:6,lty=1:5,lwd=2,bg="snow")
```



4.2 Simple R script to display a sequence with structural annotation

```
> # SSSEQ.R
> #
> # Simple R script to display a sequence with structural annotation
>
> # define sequence and secondary structure
> seq <- "GARVHMDGARLMNAAVALRIPPARLVEHCDSVSFCFSKG"
> struct <- c(0,0,2,2,2,2,2,1,1,1,1,0,0,0,0,0,0,0,0,0,0,
+                     1,1,1,1,0,0,0,0,0,2,2,2,2,2,0,0,0,0)
> residuecount <- 39;
> # set up the window for plotting
> #x11()  # may need to change to quartz() for Mac or windows() for PC
> plot.new()
> plot.window(c(0,40),c(-20,20))
> # plot a line representing the length of the sequence
> segments(0.5,0,39.5,0)
> # plot the sequence and features
> for (i in 1:residuecount) {
+     text(i,-2,substr(seq,i,i))  # write residue letter
+     if (struct[i] != 0) {
+         if (struct[i] == 1) boxcolour <- "dodgerblue4"  # alpha helix
+         if (struct[i] == 2) boxcolour <- "firebrick"    # beta sheet
+         rect(i-0.5,-1,i+0.5,1,col = boxcolour, border = NA)
+     }
+ }
> # plot a legend
> legend(x=0,y=8,legend=(c("alpha helix","beta sheet")),
+     pch=15,col=c("dodgerblue4","firebrick"),bg="snow")
> # krajša varianta
> #rect((1:39)-.5,-2,(1:39)+.5,2,col=c(NA,1,2)[struct+1],border=FALSE)
```


alpha helix
beta sheet
GARVHMDGARLMNAAVALR I PPARLVEHCDSVSFCFSKG

4.3 Funkcije

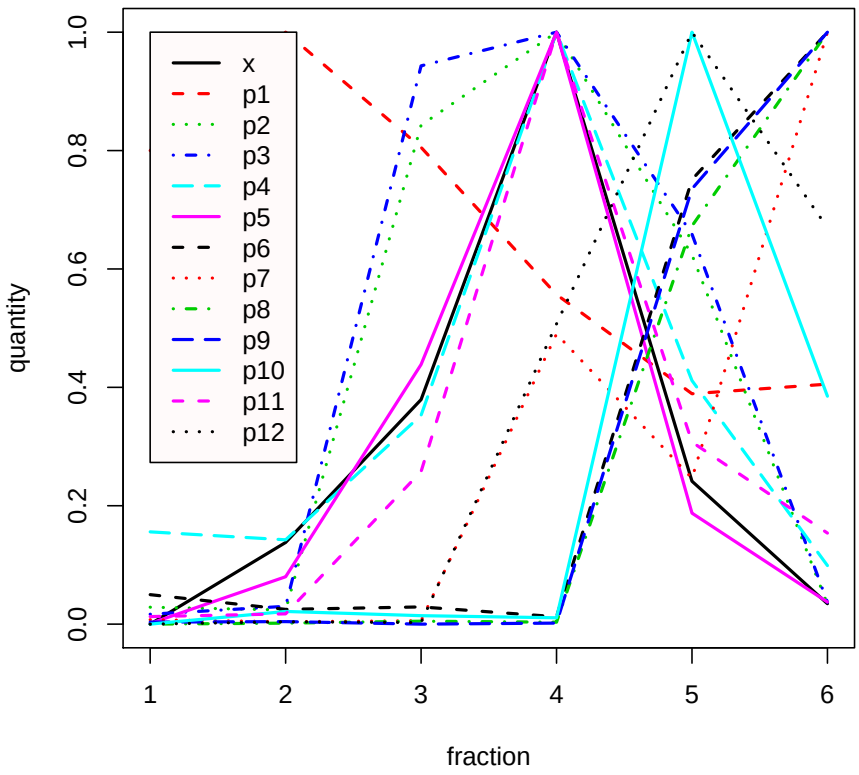
V R lahko definiramo nove funkcije, ki olajšajo kasnejše delo.

```
> # RANGESCALE.R
> #
> # R program to define a function to rangescale columns of a matrix
>
> rangescale <- function(X) {
+
+     Xmax <- apply(X, 2, max)
+     Xscaled = scale(X, scale=Xmax, center=FALSE)
+
+     return(Xscaled)
+ }
```

4.4 Ponovimo primer

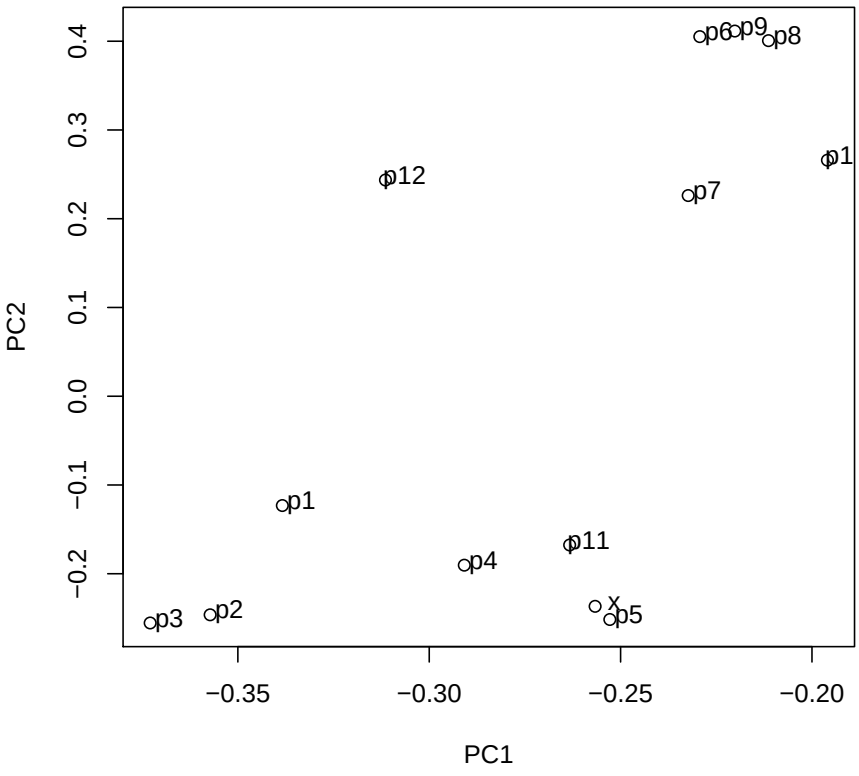
4.1

```
> # PROFILES.R
> #
> # Simple R program to load a data matrix, scale it and plot the result
>
> # clear out the workspace first
>
> # load data frame from web site
> X <- read.table("http://www.bixsolutions.net/profiles.csv", sep=",", header=
> # rangescale data by dividing by the maximum value in each column
> Xscaled <- rangescale(X)
> # plot columns in matrix as lines on a single graph
> matplot(Xscaled,type="l",xlab="fraction",ylab="quantity",col=1:6,lty=1:5,lwd
> # add legend to graph
> legend(x=1,legend=names(X),col=1:6,lty=1:5,lwd=2,bg="snow")
```



4.5 Statistična analiza podatkov - metoda glavnih komponent (PCA)

```
> # PCA_EXAMPLE.R
> #
> # Program to load in data matrix, calculate principal components
> # and plot resulting scores.
>
> #rm(list=ls())                    # clear workspace
>
> #source("rangescale.r")               # define our rangescale function
>
> # load data matrix from file
> X <- read.table("http://www.bixsolutions.net/profiles.csv", sep=",", header=
> Xscaled = rangescale(X)              #scale the profiles
> result = prcomp(Xscaled, center=FALSE)  # perform PCA
> # extract the scores matrix from the result
> scores=result$rotation
> # plot PC1 against PC2
> plot(scores[,1], scores[,2], xlab="PC1",ylab="PC2")
> # add labels to point (note 0.005,0.003 offset to avoid obscuring points)
> text(scores[,1]+0.005, scores[,2]+0.003, names(X))
>
```



SessionInfo

Windows 7 x64 (build 7601) Service Pack 1

- R version 2.15.1 (2012-06-22), `x86_64-pc-mingw32`
- Locale: `LC_COLLATE=Slovenian_Slovenia.1250`, `LC_CTYPE=Slovenian_Slovenia.1250`, `LC_MONETARY=Slovenian_Slovenia.1250`, `LC_NUMERIC=C`, `LC_TIME=Slovenian_Slovenia.1250`
- Base packages: base, datasets, graphics, grDevices, stats, utils
- Other packages: patchDVI 1.8.1584
- Loaded via a namespace (and not attached): tools 2.15.1