Bioinformatika

A. Blejec

14. oktober 2009

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

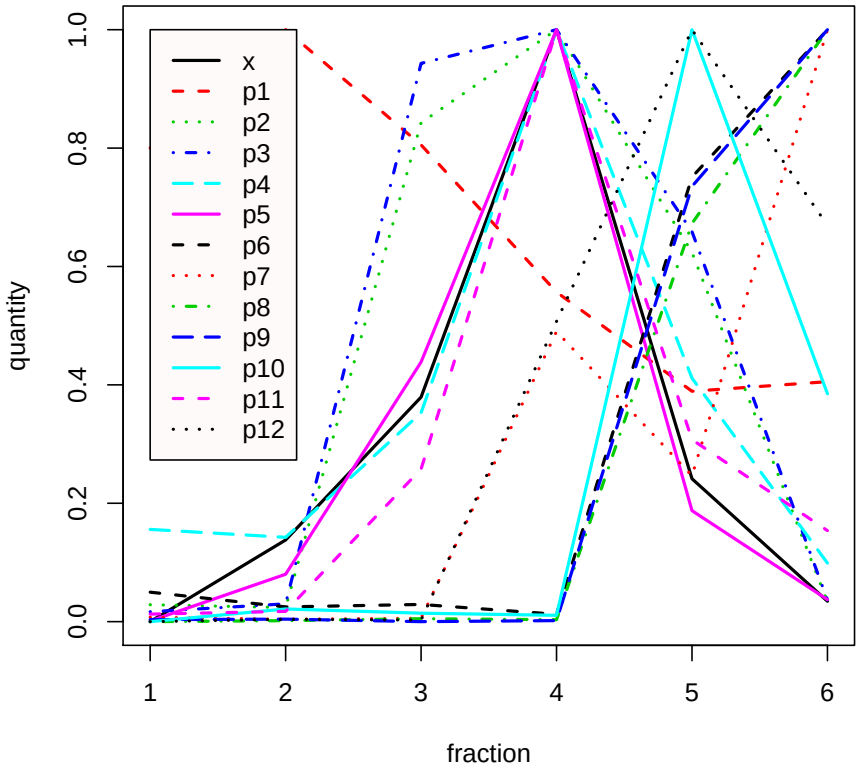
```
> imePriimek <- c("Miha Novak", "Micka Podlogar")
> exp <- "(.*) (.*)"
> priimekIme <- gsub(exp, "\\2 \\1", imePriimek)
```

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

```
> X <- read.table("http://www.bixsolutions.net/profiles.csv",
+     sep = ",", header = TRUE)
> Xmax <- apply(X, 2, max)
> Xscaled <- scale(X, scale = Xmax, center = FALSE)
> matplot(Xscaled, type = "l", xlab = "fraction", ylab = "quantity",
+     col = 1:6, lty = 1:5, lwd = 2)
> 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

```
> 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
> plot.new()
> plot.window(c(0, 40), c(-20, 20))
> segments(0.5, 0, 39.5, 0)
> for (i in 1:residuecount) {
+     text(i, -2, substr(seq, i, i))
+     if (struct[i] != 0) {
+         if (struct[i] == 1)
+             boxcolour <- "dodgerblue4"
+         if (struct[i] == 2)
+             boxcolour <- "firebrick"
+         rect(i - 0.5, -1, i + 0.5, 1, col = boxcolour,
+             border = NA)
+     }
+ }
> legend(x = 0, y = 8, legend = (c("alpha helix", "beta sheet")),
+     pch = 15, col = c("dodgerblue4", "firebrick"), bg = "snow")
```

4.3 Funkcije

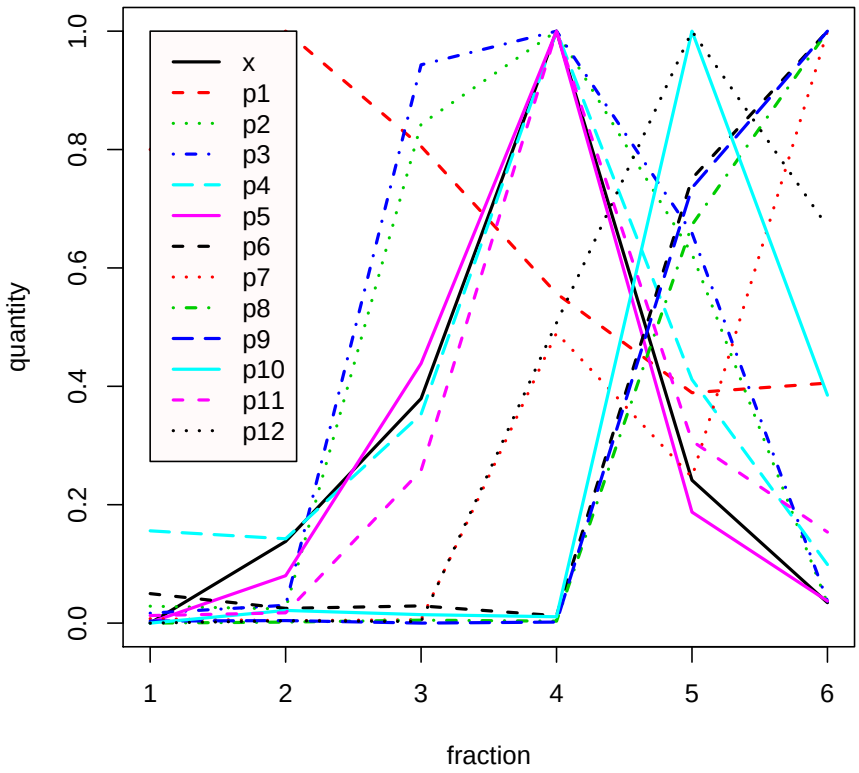
V R lahko definiramo nove funkcije, ki olajšajo kasnejše delo.

```
> rangescale <- function(X) {
+     Xmax <- apply(X, 2, max)
+     Xscaled = scale(X, scale = Xmax, center = FALSE)
+     return(Xscaled)
+ }
```

4.4 Ponovimo primer

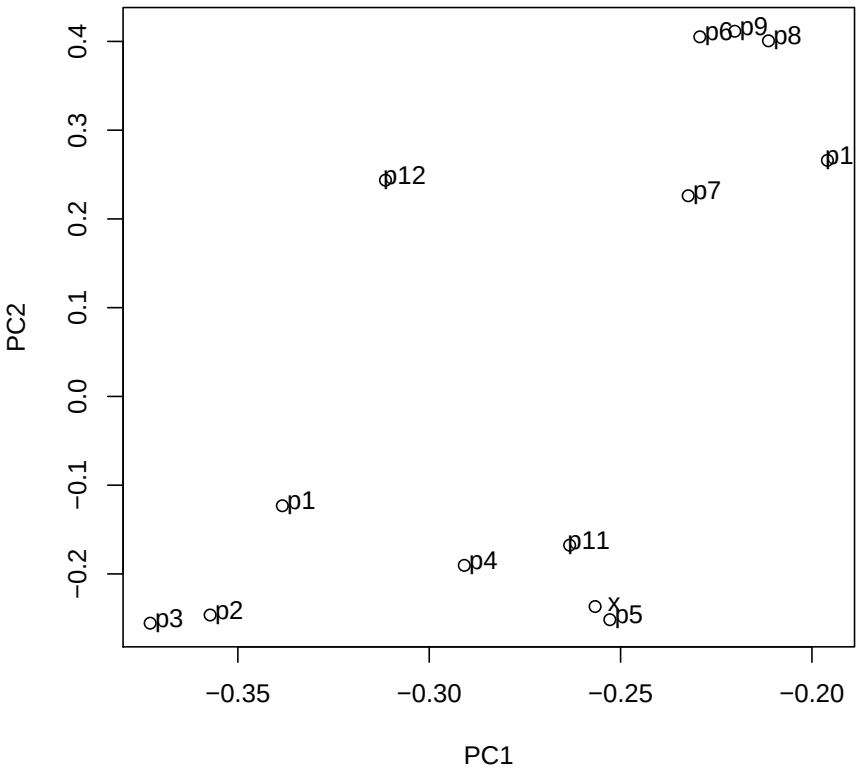
4.1

```
> X <- read.table("http://www.bixsolutions.net/profiles.csv",
+     sep = ",", header = TRUE)
> Xscaled <- rangescale(X)
> matplot(Xscaled, type = "l", xlab = "fraction", ylab = "quantity",
+     col = 1:6, lty = 1:5, lwd = 2)
> 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)

```
> X <- read.table("http://www.bixsolutions.net/profiles.csv",
+     sep = ",", header = TRUE)
> Xscaled = rangescale(X)
> result = prcomp(Xscaled, center = FALSE)
> scores = result$rotation
> plot(scores[, 1], scores[, 2], xlab = "PC1", ylab = "PC2")
> text(scores[, 1] + 0.005, scores[, 2] + 0.003, names(X))
```



SessionInfo

Windows XP (build 2600) Service Pack 3

- R version 2.8.0 (2008-10-20), `i386-pc-mingw32`
- Locale: `LC_COLLATE=Slovenian_Slovenia.1250;LC_CTYPE=Slovenian_Slovenia.1250;LC_MON`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils