Podatki v FASTA formatu

A. Blejec

November 3, 2009

Contents

1 Branje FASTA format

www.ncbi.nlm.nih.gov/enterz

```
> lfn <- "C:/_Y/R/Bioinformatika/Perl/BRCA1.fasta"
> lfn
[1] "C:/_Y/R/Bioinformatika/Perl/BRCA1.fasta"
```

Preberemo samo header

```
> header <- readLines(lfn, n = 1)
> header
[1] ">gi|9739178|gb|AF284812.1| Homo sapiens BRCAI (BRCA1) gene, exon 20 and p
```

Preberemo celo datoteko. Da dobimo samo DNA zaporedje, odrežemo prvo vrstico

```
> data <- readLines(lfn, n = -1)
> data
[1] ">gi|9739178|gb|AF284812.1| Homo sapiens BRCAI (BRCA1) gene, exon 20 and p
[2] "ATATGACGTGTCTGCTCCACTTCCATTGAAGGAAGCTTCTCTTTCTCTTATCCTGATGGGTTGTGTTTGG"
[3] "TTTCTTTCAGCATGATTTTGAAGTCAGAGGAGATGTGGTCAATGGAAGAAACCACCAAGGTCCAAAGCGA"
[4] "GCAAGAGAATCCCAGGACAGAAAGGTAAAGCTCCCTCCCTCAAGTTGACAAAAATCTCACCCCACCACTC"
[5] "TGTATTCCACTCTGTATTCCACTCCCCTTTGCAGAGATGGGCCGCTTCATTTTGTAAGACT"
> dna <- data[-1]
> dna
[1] "ATATGACGTGTCTGCTCCACTTCCATTGAAGGAAGCTTCTCTTTCTCTTATCCTGATGGGTTGTGTTTGG"
[2] "TTTCTTTCAGCATGATTTTGAAGTCAGAGGAGATGTGGTCAATGGAAGAAACCACCAAGGTCCAAAGCGA"
[3] "GCAAGAGAATCCCAGGACAGAAAGGTAAAGCTCCCTCCCTCAAGTTGACAAAAATCTCACCCCACCACTC"
[4] "TGTATTCCACTCTGTATTCCACTCCCCTTTGCAGAGATGGGCCGCTTCATTTTGTAAGACT"
```

Zaporedje je v vektorju `dna`. Vektor zlepimo v eno samo dolgo zaporedje.

```
> paste(dna, collapse = "")
[1] "ATATGACGTGTCTGCTCCACTTCCATTGAAGGAAGCTTCTCTTTCTCTTATCCTGATGGGTTGTGTTTGGTTT
```

SessionInfo

Windows XP (build 2600) Service Pack 3

- R version 2.10.0 (2009-10-26), `i386-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, methods, splines, stats, utils
- Other packages: Hmisc 3.7-0, survival 2.35-7
- Loaded via a namespace (and not attached): cluster 1.12.1, grid 2.10.0, lattice 0.17-26