

Short overview of the *sequences* package

Laurent Gatto

lg390@cam.ac.uk

Cambridge Center for Proteomics

University of Cambridge

July 3, 2013

1 Introduction

The dummy *sequences* package is used to illustrate the *Advanced R programming and package development*. It describes classes and methods to manipulate generic and biological sequences. If you are interested in real sequence manipulation in R, have a look at *Biostrings*¹, *seqinr*² or *ape*³ and possibly others.

2 Using *sequences*

Let's start by loading the package and read a fasta sequence that is provided with the package.

```
> library(sequences)
> fastafilename <- dir(system.file(package="sequences",dir="extdata"),
+                       full.name=TRUE,
+                       pattern="fasta$")
> fastafilename

[1] "/tmp/RtmpaWpzkG/Rinst40b02d163b23/sequences/extdata/aDnaSeq.fasta"

> myseq <- readFasta(fastafilename)
> myseq
```

Object of class DnaSeq

Id: example dna sequence

Length: 132

Alphabet: A C G T

Sequence: AGCATACGACGACTACGACACTACGACATCAGACACTACAGACTACTACGACTACAGACATCAGACACTACATATTTACAT

¹<http://www.bioconductor.org/help/bioc-views/release/bioc/html/Biostrings.html>

²<http://seqinr.r-forge.r-project.org/>

³<http://cran.r-project.org/web/packages/ape/index.html>

Printing the sequence displays it's sequence numbering the lines.

```
> print(myseq)

> example dna sequence
 1  AGCATACGA
10  CGACTACGAC
20  ACTACGACAT
30  CAGACACTAC
40  AGACTACTAC
50  GACTACAGAC
60  ATCAGACACT
70  ACATATTTAC
80  ATCATCAGAG
90  ATTATATTAA
100 CATCAGACAT
110 CGACACATCA
120 TCATCAGCAT
130 CAT
```

This creates an instance of class DnaSeq that can be transcribed with the transcribe method.

```
> transcribe(myseq)
```

Object of class RnaSeq

Id: example dna sequence -- transcribed

Length: 132

Alphabet: A C G U

Sequence: AGCAUACGACGACUACGACACUACGACAUCAGACACUACAGACUACUACGACUACAGACAUCAGACACUACAUAUUUACAU

```
> barplot(gccount(seq(myseq)))
```

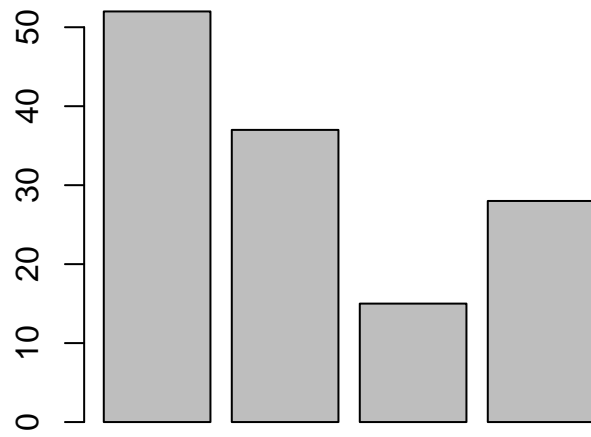


Figure 1: Number of A, C, G and T bases in the `myseq` object.

3 Background

This package is developed as part of the *Advanced R programming and package development* (ARPD) course, taught by Laurent Gatto and Robert Stojnic. The course has originally been set up and run as an intense 1 day course in the Graduate School of Life Sciences of the University of Cambridge. Since March 2011, the course has been run on a regular basis in the Bioinformatics Teaching Facility in the Department of Genetics, Cambridge.

On the 28 and 29 November, a 2 day version was taught at the EMBL in Heidelberg, at Wolfgang Huber's invitation (see figure 2).



Figure 2: Delegates and organisers, EMBL, Heidelberg, 28 - 29 November 2011

Acknowledgements Several people have been contributed to make this course possible. David P. Judge, initially helped us to set up the course in the Bioinformatics Teaching Facility at the Cambridge University. Wolfgang Huber, invited us at the EMBL in Heidelberg, in November 2011,

4 Session information

- R version 3.0.1 Patched (2013-06-16 r62969),
x86_64-unknown-linux-gnu
- Locale: LC_CTYPE=en_GB.UTF-8, LC_NUMERIC=C, LC_TIME=en_GB.UTF-8,
LC_COLLATE=C, LC_MONETARY=en_GB.UTF-8, LC_MESSAGES=en_GB.UTF-8,
LC_PAPER=C, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C,
LC_MEASUREMENT=en_GB.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: Rcpp~0.10.4, sequences~0.5.7
- Loaded via a namespace (and not attached): tools~3.0.1