Short overview of the *sequences* package

Laurent Gatto*

December 3, 2014

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")

## Loading required package:  Rcpp
## This is package 'sequences'

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

## [1] "/tmp/Rtmp6egddS/Rinst42111894325b/sequences/extdata/aDnaSeq.fasta"
## [2] "/tmp/Rtmp6egddS/Rinst42111894325b/sequences/extdata/moreDnaSeqs.fasta"

myseq <- readFasta(fastafilename[1])
myseq
```

*lg390@cam.ac.uk

[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

```
## Object of class DnaSeq
##  Id: example dna sequence
##  Length: 132
##  Alphabet: A C G T
##  Sequence: AGCATACGACGACTACGACACTACGACATCAGACACTACAGACTACTACGACTACAGACATCAGACACTACATATTTA
```

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: AGCAUACGACGACUACGACACUACGACAUCAGACACUACAGACUACUACGACUACAGACAUCAGACACUACAUAUUUA
```

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

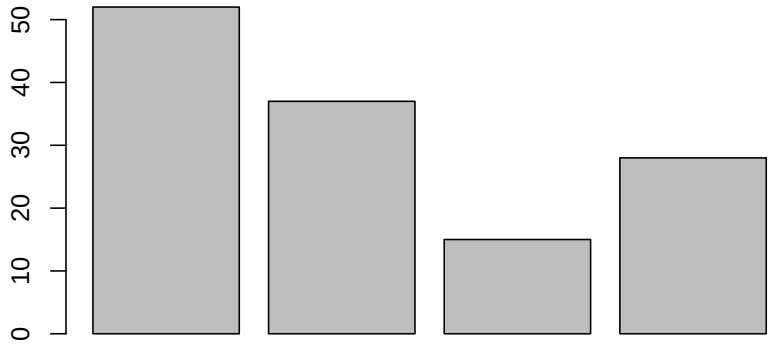

Figure 1: Number of A, C, G and T bases in the `myseq` object.

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 Facilty in the Department of Genetics, Cambridge.

In November 2011 and December 2012, 2 day courses were 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 Facilty at the Cambridge University. Wolfgang Huber, invited us at the EMBL in Heidelberg.

4 Session information

- R Under development (unstable) (2014-11-01 r66923), `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=en_GB.UTF-8`, `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 0.11.3, sequences 0.5.9
- Loaded via a namespace (and not attached): evaluate 0.5.5, formatR 1.0, highr 0.4, knitr 1.8, stringr 0.6.2, tools 3.2.0