TakeHome exercises Vertex

Claudio Zanettini

Jan 11, 2010

Contents

crispR

I wrote the functions in a `R package`. Source files of the package are in `crispR_0.0.0.9000.tar.gz`. Installing the source will automatically check and install any missing dependencies.

Installation

```
# install.packages("devtools")
devtools::install_local("path_to_local_package")
```

Part1 Answers

a) the code for the function

The function is called `find_proto` and is the first function in the`protospacers.R` file.

b) the code to call the function with the example variables (and others, if desired)

```
library(crispR)
find_proto(d_seq = "TGATCTACTAGAGACTACTAACGGGGATACATAG",
           l = 2,
           PAM = "NGG")
```

start_p	end_p	protospacer	PAM	strand
20	21	AA	CGG	+
21	22	AC	GGG	+
22	23	CG	GGG	+

..or using DNA of the Dopamine Transporter (DAT internal data):

```r
library(crispR)
print(DAT)
```

```
## [1] "TTTGCAAACGCTCGCATGTCACCGAAGGCGCAACAGCTCCGATTTTGAAATTTCCAACACGGCCCTCAAGTTGAAAGTTTTCCAAAAAAATTTAA
```

```r
library(crispR)
prot <- find_proto(d_seq = DAT,
                   l = 20,
                   PAM = "NGG")
head(prot, 10)
```

start_p	end_p	protospacer	PAM	strand
6	25	AAACGCTCGCATGTCACCGA	AGG	+
40	59	CGATTTTGAAATTTCCAACA	CGG	+
109	128	TGTGAATGTGAAGTGAAATT	CGG	+
110	129	GTGAATGTGAAGTGAAATTC	GGG	+
115	134	TGTGAAGTGAAATTCGGGTT	TGG	+
122	141	TGAAATTCGGGTTTGGCATT	CGG	+
128	147	TCGGGTTTGGCATTCGGCAT	TGG	+
136	155	GGCATTCGGCATTGGTTGTG	TGG	+
158	177	GAGCTTTTTTCTAAGTTTTC	TGG	+
180	199	GATATTTTTCAAAAGTCTCA	AGG	+

c) the time complexity for the function (in big-O notation)

I am not explicitly using any loop, but my function is in any case iterating and looking at each nucleotide of the sequence by using `grep` (`stringr` and regular expressions).

time Complexity: *O(n)*

Part2 Answers.

a) The code for the function

The function is called `find_FASTA` and is the second function in the`protospacers.R` file.

b) The source of the FASTA file used for the reference genome in the example problem

I downloaded the Reference Genome Sequence GRCh38 from here.

c) How many candidate guide (protospacer) sequences were identified in the example problem

A total of 54 protospacers were identified on strand (+). Please note the arguments “start”, “end” and “l” are 1-indexed and intervals are fully closed.

d) The list of candidate guide (protospacer) sequences in a tab-delimited file...

A tab-delimited file called solution.txt is in the current archive.

Dependencies

All the dependencies are listed in the file DESCRIPTION.

Time needed to right the code

A quick and dirty version can be written probably in 1 hour or less. I polished the code, wrote the documentation too, and in total it took me few hours... but I also spent quite some time thinking about the reverse complementary strand!