Enzyme recognition sequence in DNA

Joaquín Rodrigo Ponce de León Conconi

5/14/2021

We first need to import the library stingr

```
library(stringr)
```

Next we define our example DNA

```
dna <- "GAATTCGAGCTCGGTACCCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTT"
```

Now we define a function that takes the DNA sequence and the enzyme recognition sequence (both 5' to 3'), and displays "The recognition sequence was not found" if the enzyme recognition sequence was not found in the DNA strand or displays the DNA sequence showcasing in lowercase letters the enzyme recognition sequence along with an answer TRUE or FALSE to confirm that the final answer is the same as the DNA sequence taken by the function, which should always be TRUE. Also, if the DNA sequence is shorter than the enzyme recognition sequence, the function returns an error.

```
enz_seq <- function(seq_dn, seq_en){ #DNA sequence and enzyme recognition sequence asked
  seq_dn <- toupper(seq_dn)#change seq_dn uppercase
  seq_en <- toupper(seq_en)#change seq_en to uppercase
  seq_dns <- strsplit(seq_dn, "") #seq_dn is split
  len_seq <- length(seq_dns[[1]]) #The number of base pairs is counted from the first element of the li
  if (length(strsplit(seq_en, "")[[1]])>len_seq){
    return("The DNA sequence is shorter than the enzyme recognition sequence")
  }else{
    A <- "" #Blank final answer variable
    if (str_detect(seq_dn, seq_en)==F){ #If the sequence is not detected the it returns...
      return("The recognition sequence was not found")
    }else{ #Else, the code to display the DNA seq. along with the enzyme recognition sequence is saved
      locs <- str_locate(seq_dn, seq_en) #The locations within seq_dn where seq_en begins and ends
      if (locs[1]==1){#If the beginning of seq_en is in index 1 of seq_dn, then
        A <- c(tolower(seq_dns[[1]][locs[1]:locs[2]]), seq_dns[[1]][locs[2]+1: len_seq])
      }else if (locs[2]==len_seq){#Else if the end of seq_en is in the last index of seq_dn
        A <- c(seq_dns[[1]][1:locs[1]-1], tolower(seq_dns[[1]][locs[1]:locs[2]]))
      }else{
        A <- c(seq_dns[[1]][1:locs[1]-1],tolower(seq_dns[[1]][locs[1]:locs[2]]),seq_dns[[1]][locs[2]+1:
      }
    }
    A <- paste(na.omit(A), collapse = "")
    return(c(A, identical(toupper(A), seq_dn)))
  }
}
```

Let's do some examples with different enzymes that have distinct recognition sequences

```
enz_seq(dna, "GAATTC")#EcoRI
```

```
## [1] "gaattcGAGCTCGGTACCCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTT"
## [2] "TRUE"
```

```
enz_seq(dna, "GAGCTC")#SacI
```

```
## [1] "GAATTCgagctcGGTACCCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTT"
## [2] "TRUE"
```

```
enz_seq(dna, "GGTACC")#KpnI
```

```
## [1] "GAATTCGAGCTCggtaccCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTT"
## [2] "TRUE"
```

```
enz_seq(dna, "CCCGGG")#SmaI
```

```
## [1] "GAATTCGAGCTCGGTAcccgggGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTT"
## [2] "TRUE"
```

```
enz_seq(dna, "GGATCC")#BamHI
```

```
## [1] "GAATTCGAGCTCGGTACCCGGggatccTCTAGAGTCGACCTGCAGGCATGCAAGCTT"
## [2] "TRUE"
```

```
enz_seq(dna, "TCTAGA")#XbaI
```

```
## [1] "GAATTCGAGCTCGGTACCCGGGGATCCtctagaGTCGACCTGCAGGCATGCAAGCTT"
## [2] "TRUE"
```

```
enz_seq(dna, "GTCGAC")#SalI
```

```
## [1] "GAATTCGAGCTCGGTACCCGGGGATCCTCTAGAgtcgacCTGCAGGCATGCAAGCTT"
## [2] "TRUE"
```

```
enz_seq(dna, "CTGCAG")#PstI
```

```
## [1] "GAATTCGAGCTCGGTACCCGGGGATCCTCTAGAGTCGACctgcagGCATGCAAGCTT"
## [2] "TRUE"
```

```
enz_seq(dna, "GCATGC")#SphI
```

```
## [1] "GAATTCGAGCTCGGTACCCGGGGATCCTCTAGAGTCGACCTGCAGgcatgcAAGCTT"
## [2] "TRUE"
```

```
enz_seq(dna, "AAGCTT")#HindIII
```

```
## [1] "GAATTCGAGCTCGGTACCCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCaagctt"
## [2] "TRUE"
```

```
enz_seq(dna, "aagctt")#You can even enter the senquence in lowercase
```

```
## [1] "GAATTCGAGCTCGGTACCCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCaagctt"
## [2] "TRUE"
```

```
enz_seq(dna, "TTTACG")#Negative control with sequence that's not in dna
```

```
## [1] "The recognition sequence was not found"
```

```
enz_seq("AAATGCCGTGATGCCGTTTTAGGCTGCAG", "GGATCC")#Negative control with DNA sequence that doesn't cont
```

```
## [1] "The recognition sequence was not found"
```

```
enz_seq("GGATCC", dna)#Negative control placing the BamHI recognition sequence in the dna position
```

```
## [1] "The DNA sequence is shorter than the enzyme recognition sequence"
```