class06-inclass

Tin Nguyen

Quarto

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see https://quarto.org.

> Q. Write your first R function, add() to add some numbers

```
#Adding two number
add <- function(x,y){
  sum <- x + y;
  return(sum)
}

add(1,2)
```

```
[1] 3
```

> Q. Write a second function, generate_dna() to return nucleotide sequences of a user specifed length.

```
#Genearte DNA
generate_dna <- function(length){
  dna<-sample(c("A","G","C","T"),length, replace = TRUE)
  return(dna)
}

generate_dna(10)
```

```
 [1] "T" "T" "T" "A" "T" "A" "C" "A" "A" "A"
```

> Q. Write a third function, generate_protein() to return protein sequences of different lengths and test whether these sequences are unique in nature.

```r
#Genearte Protein
generate_protein <- function(length, seq_id = "id.1", description = NULL){

  amino_acid <- c("A", "R", "N", "D", "C", "Q", "E", "G", "H", "I",
                  "L", "K", "M", "F", "P", "S", "T", "W", "Y", "V")
  protein_sequence <-paste(sample(amino_acid,length, replace = TRUE), collapse ="")

  header <- if (!is.null(description)) paste(">", seq_id, description) else paste(">", seq_id

  fasta <- paste0(header, "\n" , protein_sequence)
  cat(fasta, "\n")
  return(fasta)
}


generate_protein(20)
```

```
> id.1
CDVNPWGPKVAPNQRIRAGR

[1] "> id.1\nCDVNPWGPKVAPNQRIRAGR"
```

```r
generate_protein(20, "id.2")
```

```
> id.2
PSGMVMDVKFLDKRSWASSA

[1] "> id.2\nPSGMVMDVKFLDKRSWASSA"
```

```r
#
protein_lengths <- 6:12
seq_ids <- paste0("id.", seq_along(protein_lengths))

sapply(seq_along(protein_lengths), function(i) {
  generate_protein(protein_lengths[i], seq_id = seq_ids[i])
})
```

```
> id.1
RIMEEL
```

```
> id.2
YAPCKDC
> id.3
RFQYCTMT
> id.4
WFEHVNINF
> id.5
MCCGCYMDTV
> id.6
KPVWMSSYVPM
> id.7
FSHKCSTKAVPQ
```

```
[1] "> id.1\nRIMEEL"       "> id.2\nYAPCKDC"      "> id.3\nRFQYCTMT"
[4] "> id.4\nWFEHVNINF"    "> id.5\nMCCGCYMDTV"   "> id.6\nKPVWMSSYVPM"
[7] "> id.7\nFSHKCSTKAVPQ"
```

> Q. Determine if these sequences can be found in nature or or are they unique? Why or why not?

I Blastpseached my FASTA format sequences against NR and found that length 6, 8 , … are not unique and can be found in the data bases with 100% coverage and 100% identity.

Random seqeucnes of length 9 and above are unique and can't be found in the databases.

Word size/window size for protein is 9