

Class 6: R functions

Serena (PID: A18556865)

All functions in R have at least 3 things:

- A **name**, we pick this and use it to call the function.
- Input **arguments**, there can be multiple comma separated inputs to the function.
- The **body**, lines of R code that do the work of the function.

Our first wee function:

```
add <- function(x, y=1){  
  x + y  
}
```

Let's test our first function

```
add(c(1,2,3), y=10)
```

```
[1] 11 12 13
```

```
add(10)
```

```
[1] 11
```

```
add(10,100)
```

```
[1] 110
```

A second function

Let's try something more interesting. Make a sequence generation tool.

The `sample()` function could be useful here

```
sample(1:10, size = 3)
```

```
[1] 2 4 6
```

Change this to work with nucleotides A C G and T and return 3 of them

```
n <- c("A", "C", "G", "T")
sample(n, size = 15, replace = TRUE)
```

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

Turn this snippet into a function that returns a user specified length dna sequence. Let's call it `generate_dna()`...

```
generate_dna <- function(len=10, fasta = FALSE) {
  n <- c("A", "C", "G", "T")
  v <- sample(n, size = len, replace = TRUE)

  # Make a single element vector
  s <- paste(v, collapse = "")

  cat("Well done you!\n")

  if(fasta) {
    return(s)
  } else{
    return(v)
  }
}
```

```
generate_dna(5)
```

Well done you!

```
[1] "G" "A" "A" "A" "C"
```

```
s <- generate_dna(15)
```

Well done you!

```
s
```

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

I want the option to return a single element character vector with my sequence all together like this: "GGAGTAC"

```
generate_dna(10, fasta = FALSE)
```

Well done you!

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

A more advanced example

Make a third function that generates protein sequence of a user specific length and format

```
generate_protein <- function(size= 15, fasta=TRUE) {  
  aa <- c(  
    "A","R","N","D","C","E","Q","G","H","I",  
    "L","K","M","F","P","S","T","W","Y","V")  
  
  seq <- sample(aa, size= size, replace = TRUE)  
  
  if (fasta) {  
    return(paste(seq,collapse = ""))  
  } else {  
    return(seq)  
  }  
}
```

Try this out...

```
generate_protein(10)
```

```
[1] "SGFWPRQLDM"
```

Q.Generate random protein sequences between lengths 5 and 12 amino-acids.

```
generate_protein(5)
```

```
[1] "MWMFP"
```

```
generate_protein(6)
```

```
[1] "AVQEFD"
```

One approach is to do this by brute force calling our function for each length 5 to 12.

Another approach is to write a `for()` loop to iterate over the input valued 5 to 12

A very useful third R specific approach is to use the `sapply()` function.

```
seq_lengths <- 6:12
for (i in seq_lengths) {
  cat(">", i, "\n")
  cat(generate_protein(i))
  cat("\n")
}
```

```
> 6
QMKDQH
> 7
RNTYVCI
> 8
LTGYPKEG
> 9
EPCGFNVSF
> 10
PNAQRWGGVD
> 11
LHTWHCEAFMM
> 12
TNPPLFEWYELP
```

```
sapply(5:12, generate_protein)
```

```
[1] "RMCTQ"      "THTHPN"      "DTFGWSN"      "VIRTMNGL"      "LHDYKCLFY"  
[6] "HSSMWGQVPF" "IGQGEHPFRLY" "MSWLWLGLVADC"
```

Key-Point: Writing functions in R is doable but not the easiest thing in the world. Starting with a working snippet of code and then using LLM tools to improve and generalize your function code is a productive approach.