R Functions

Isabel Philip (A16855684)

Today we will get more exposure to functions in R. We call functions to do all our work and today we will learn how to write our own.

A fisrt silly function

Note that arguments 2 and 3 have default values (because we set y=0 and z=0) so we don't have to supply them when we call our function.

```
add <- function(x,y=0, z=0){
  x + y + z
}
```

```
add(1,1)
```

```
[1] 2
```

```
add(1, c(10,100))
```

```
[1]  11 101
```

```
add(1)
```

```
[1] 1
```

```
add(100,1,1)
```

```
[1] 102
```

A second more fun function

Let's write a function that generates random nucleotide sequences.

We can make use of a the in-built `sample()` function in R to help us here.

```
sample(x=1:10, size=9)
```

```
[1]  1  6  7 10  4  9  5  8  3
```

> Doesn't repeat numbers becuase the functions has replacements as False, therefore no repeats will occur

```
sample(x=1:10, size=11, replace= TRUE)
```

```
 [1]  5  9  9  3  3  2 10  4  2  4  9
```

> now can sample multiple times (above the sample size) since repeats can NOW occur
>
> Q. Can you use `sample()` to generate a random nucleotide sequence length of 5?

```
sample(x=c("A","T","G","C"), size=5, replace = TRUE)
```

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

> c() makes it a vector
>
> Q. Generate a function `generate_dna()`that makes a nucleotide sequence of a user specified length.

Every function in R has at least 3 things: - a **name** (in our case "generate_dna") - one or more **input arguments** (the length of the sequence we want) - a **body** (that does the work)

```
generate_dna <- function(length=5){
  bases <- c("A","T","G","C")
  sample(bases, size=length, replace = TRUE)
}
```

> length = 5 acts as a **default**, so when no number is inputted, the length will automatically be 5

```
generate_dna()
```

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

> Can you write a `generate_protein()` function that returns amino acids seqeunce of the user requested length?

```
aa <- bio3d ::aa.table$aa1[1:20]
```

```
generate_protein <- function(length=5){
  sample(aa, size=length, replace = TRUE)
}
```

```
generate_protein(10)
```

```
 [1] "Y" "P" "L" "E" "I" "M" "E" "E" "I" "S"
```

I want my output of this function not to be a vector with one amino acids per element, but rather a one element single string (like a word)

```
bases <- c("A","T","G","C")
paste(bases, collapse = "")
```

```
[1] "ATGC"
```

```
paste(generate_protein(), collapse = "")
```

```
[1] "IAIMK"
```

```
generate_protein <- function(length=5){
  aa <- bio3d ::aa.table$aa1[1:20]
  s <- sample(aa, size=length, replace = TRUE)
  paste(s, collapse = "")
}
```

```
generate_protein()
```

```
[1] "AEDMC"
```

> Q. Generate a protein sequence from length 6 to 12?

```
generate_protein(6)
```

```
[1] "WKYHLT"
```

```
generate_protein(7)
```

```
[1] "ADLHLPS"
```

```
generate_protein(8)
```

```
[1] "IHGQKTQI"
```

We can use the useful utility function `sapply()` to help us "apply" our function over all the values 6 to 12.

```
ans <- sapply(6:12,generate_protein)
ans
```

```
[1] "IIVYTH"       "GTFFDAH"      "EITQLKQQ"     "PLFTILKNK"    "EVMWNCSFQK"
[6] "QHRSYGAFEPG"  "EWHQHQFKDSLG"
```

X stands for a list, while FUn stands for the function you want it to be applied to

```
cat( paste(">ID", 6:12, sep="", "\n", ans, "\n", collapse = "") )
```

```
>ID6
IIVYTH
>ID7
GTFFDAH
>ID8
EITQLKQQ
>ID9
PLFTILKNK
>ID10
EVMWNCSFQK
>ID11
QHRSYGAFEPG
>ID12
EWHQHQFKDSLG
```

collapse was added since there was an additional space being added for ID7-ID12. with the collapse, it's fixed so everything is on order and in line. It doesn't change the rest of the code.

Are any of these sequences unique in nature - i.e never found in nature/ We can search "refseq-protein" and look for 100% Ide and 100% Coverage

Yes.

ID6 and ID7 are **not unique** as they can be found in nature with 100% identity and 100% coverage. ID6 is identical to multiple organisms, one being usherin [Echinops telfairi] (which holds the exact alignment sequence as ID6 = YFGFDH). ID7 is also identical to multiple organisms, one being uncharacterized protein BU24DRAFT_429207 [Aaosphaeria arxii CBS 175.79] (which holds the exact alignment sequence as ID7 = QMHGTMP).

ID8 through ID12 are all **unique** in nature as they are not found through BLAST as having 100% identity and coverage with other organisms. In other words, they have low coverage and/or low percent identity (along with high E-values), indicating that the searches that are showing up should not be completely trusted. In addition, the alignments do not exactly match, having gaps or starting off oddly to match the seqeunce.