Biological Data Analysis (CSE 182) : Assignment 1

Name: Amit Elia

Date: April 2nd 2022

Libraries used:

```
library(Biostrings)
library(dplyr)
library(stringr)
```

References:

https://stackoverflow.com/questions/21263636/read-fasta-into-a-dataframe-and-extract-subsequences-of-fasta-file
https://www.biostars.org/p/274312
https://www.edureka.co/community/2091/how-to-import-text-file-as-a-single-character-string

1. I have read and agree to the AI, grading, and syllabus policies.
2. System Description:
    - **Platform:** Windows 10 PC
    - **Scripting language:** R
    - **Editor:** Rstudio
    - **Simple program:**

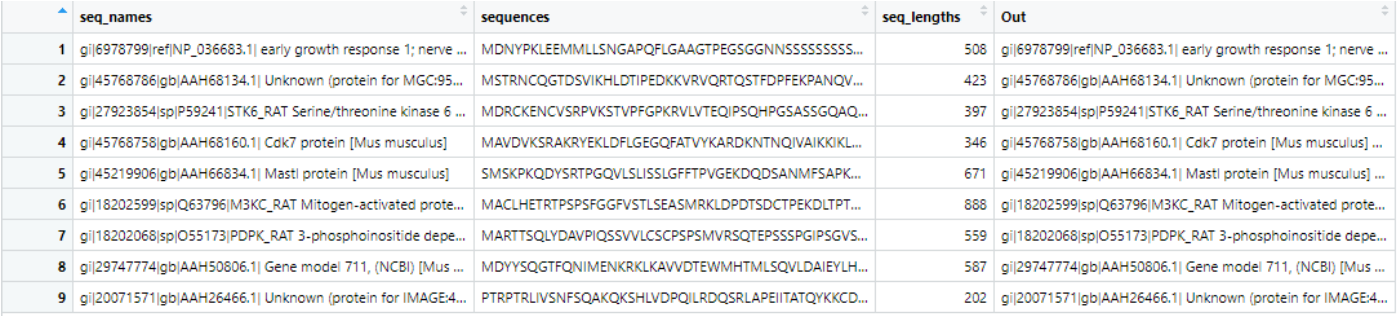
```
> rm(list = ls()) #clear environment
>
> print("Hello Bioinformatics")
[1] "Hello Bioinformatics"
```

3. Sent.
4. Commands:
   - readAAStringSet - a Biostrings method that reads an amino-acid fasta file into XStringSet format that is used to easily subset sequences.
   - data.frame(...) - creates a new df object with given vectors.
   - df["Out"] <- paste(dfseq_names, dfseq_lengths) - used to get output vector for each sequence by concatenating the sequence description to the length.

   Output:

```
> #print output
> df$Out
 [1] "gi|6978799|ref|NP_036683.1| early growth response 1; nerve growth factor-induced gene [Rattus norvegicus] 508"
 [2] "gi|45768856|gb|AAH67618.1| Serum/glucocorticoid regulated kinase [Danio rerio] 433"
 [3] "gi|45768786|gb|AAH68134.1| Unknown (protein for MGC:95907) [Mus musculus] 423"
 [4] "gi|27923854|sp|P59241|STK6_RAT Serine/threonine kinase 6 (Aurora-A) (ratAurA) 397"
 [5] "gi|45768720|gb|AAH67812.1| Cyclin L1 [Homo sapiens] 526"
 [6] "gi|45768758|gb|AAH68160.1| Cdk7 protein [Mus musculus] 346"
 [7] "gi|45219906|gb|AAH66834.1| Mastl protein [Mus musculus] 671"
 [8] "gi|18202599|sp|Q63796|M3KC_RAT Mitogen-activated protein kinase kinase kinase 12 (MAPK-upstream kinase) (MUK) 888"
 [9] "gi|4835224|emb|CAB42902.1| protein kinase ATN1 like protein [Arabidopsis thaliana] 370"
[10] "gi|40787731|gb|AAH64804.1| SLK protein [Homo sapiens] 617"
[11] "gi|18202068|sp|O55173|PDPK_RAT 3-phosphoinositide dependent protein kinase-1 (Protein kinase B kinase) (PkB kinase) 559"
[12] "gi|34191428|gb|AAH36504.2| C9orf96 protein [Homo sapiens] 700"
[13] "gi|29747774|gb|AAH50806.1| Gene model 711, (NCBI) [Mus musculus] 587"
[14] "gi|28856169|gb|AAH48033.1| Serine/threonine kinase 3 (STE20 homolog, yeast) [Danio rerio] 492"
[15] "gi|20071571|gb|AAH26466.1| Unknown (protein for IMAGE:4485517) [Mus musculus] 202"
[16] "gi|45709347|gb|AAH67695.1| Unknown (protein for MGC:85918) [Danio rerio] 320"
```

5. Commands:
    - filter(df, grepl('Mus musculus|Rattus norvegicus|_RAT', seq_names)) - dplyr command that outputs a filtered data frame based on the condition that seq_names column contains one of the patterns between '|'.



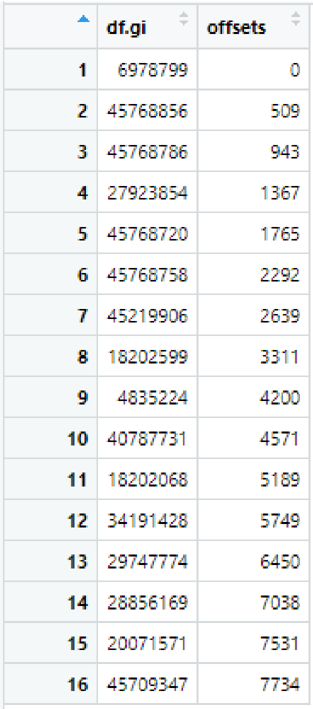
	seq_names	sequences	seq_lengths	Out
1	gi|6978799|ref|NP_036683.1| early growth response 1; nerve ...	MDNYPKLEEMMLLSNGAPQFLGAAGTPEGSGGNNSSSSSSSSS...	508	gi|6978799|ref|NP_036683.1| early growth response 1; nerve ...
2	gi|45768786|gb|AAH68134.1| Unknown (protein for MGC:95...	MSTRNCQGTDSVIKHLDTIPEDKKVRVQRTQSTFDPFEKPANQV...	423	gi|45768786|gb|AAH68134.1| Unknown (protein for MGC:95...
3	gi|27923854|sp|P59241|STK6_RAT Serine/threonine kinase 6 ...	MDRCKENCVSRPVKSTVPFGPKRVLVTEQIPSQHPGSASSGQAQ...	397	gi|27923854|sp|P59241|STK6_RAT Serine/threonine kinase 6 ...
4	gi|45768758|gb|AAH68160.1| Cdk7 protein [Mus musculus]	MAVDVKSRAKRYEKLDFLGEGQFATVYKARDKNTNQIVAIKKIKL...	346	gi|45768758|gb|AAH68160.1| Cdk7 protein [Mus musculus] ...
5	gi|45219906|gb|AAH66834.1| Mastl protein [Mus musculus]	SMSKPKQDYSRTPGQVLSLISSLGFFTPVGEKDQDSANMFSAPK...	671	gi|45219906|gb|AAH66834.1| Mastl protein [Mus musculus] ...
6	gi|18202599|sp|Q63796|M3KC_RAT Mitogen-activated prote...	MACLHETRTPSPSFGGFVSTLSEASMRKLDPDTSDCTPEKDLTPT...	888	gi|18202599|sp|Q63796|M3KC_RAT Mitogen-activated prote...
7	gi|18202068|sp|O55173|PDPK_RAT 3-phosphoinositide depe...	MARTTSQLYDAVPIQSSVVLCSCPSPSMVRSQTEPSSSPGIPSGVS...	559	gi|18202068|sp|O55173|PDPK_RAT 3-phosphoinositide depe...
8	gi|29747774|gb|AAH50806.1| Gene model 711, (NCBI) [Mus ...	MDYYSQGTFQNIMENKRKLKAVVDTEWMHTMLSQVLDAIEYLH...	587	gi|29747774|gb|AAH50806.1| Gene model 711, (NCBI) [Mus ...
9	gi|20071571|gb|AAH26466.1| Unknown (protein for IMAGE:4...	PTRPTRLIVSNFSQAKQKSHLVDPQILRDQSRLAPEIITATQYKKCD...	202	gi|20071571|gb|AAH26466.1| Unknown (protein for IMAGE:4...

    - writeXStringSet - writes a new .fasta file from a XStringSet object.

Output File : fasta_mouse_rat.fasta

6. Commands:
    - ○ writeLines() - method to write a string vector into a file with sep = “@”
    - ○ df$gi <- str_extract(df$seq_names, "\\|.*?\\|") - stringr command that extracts substrings from a string vector with a given syntax. In this case it looks for the first substring between two ‘|’ which is the gi number.
    - ○ df$gi <- str_remove_all(df$gi, "[\\|\\|]") - removes the ‘|’ characters from each string.
    - ○ Offsets for each sequence in data.seq were obtained using a for-loop.

   Output Files: data.seq, data.in.
    - ○ Data.in:



	df.gi	offsets
1	6978799	0
2	45768856	509
3	45768786	943
4	27923854	1367
5	45768720	1765
6	45768758	2292
7	45219906	2639
8	18202599	3311
9	4835224	4200
10	40787731	4571
11	18202068	5189
12	34191428	5749
13	29747774	6450
14	28856169	7038
15	20071571	7531
16	45709347	7734

7. Commands:
    - ○ readChar() - reads a txt file into a single string.
    - ○ locations <- gregexpr(pattern = query, seq_data) - returns a list of all indices of pattern in seq_data.

   Output:

```
> output <- getSeqFunc("MHIQITDFGTAKVLSPDS")
> output
[1] 18202068
```

8. About 3.5 hours. I did not ask for help. I used my past knowledge of R and RStudio during BENG185 and my time at Chen's lab at the radiology department here at UCSD. I enjoyed the assignment because it refreshed my use of R data frames without being too stressful. I would like to get feedback on the formatting of my script.