Open **Day4-stringr-lect-mito.rmd**

Load libraries and run the join function to creat the object `mito_proteins`

mito_proteins

```
glimpse(mito_protein)
```

```
$ yeast_mito_homolog_score    <chr> "NoMitoHomolog", "NoMitoHomolog", "NoMitoHomolog", "NoMitoH...
$ rickettsia_homolog_score    <chr> "NoHomolog", "NoHomolog", "NoHomolog", "NoHomolog", "NoHomo...
$ msms_score                  <chr> "50-75ambig", NA, "50-75ambig", NA, NA, NA, NA, NA, NA, "25...
$ mcarta2_score               <dbl> -1.4533, -3.3441, -4.0790, -13.4554, -10.6961, -8.2616, -10...
$ mcarta2_fdr                 <dbl> 0.459, 0.614, 0.679, 0.924, 0.887, 0.866, 0.887, 0.767, 0.4...
$ mcarta2_list                <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,...
$ mcarta2_evidence            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "target...
$ hg19_chromosome             <fct> 19, 10, 12, 12, 1, 22, 3, 12, 12, 3, 3, 1, 1, 4, 15, 2, 11,...
$ hg19_start                  <dbl> 58858171, 52559168, 9220303, 8975149, 33772366, 43088126, 1...
$ hg19_stop                   <dbl> 58864865, 52645435, 9268558, 9029381, 33786699, 43116876, 1...
$ msms_num_tissues            <dbl> 1, NA, 1, NA, NA, NA, NA, NA, NA, 1, NA, NA, NA, 2, NA, 4, ...
$ msms_num_peptides_unique    <dbl> 24, NA, 67, NA, NA, NA, NA, NA, NA, 9, NA, NA, NA, 24, NA, ...
$ msms_num_spectra            <dbl> 153, NA, 326, NA, NA, NA, NA, NA, NA, 11, NA, NA, NA, 168, ...
$ msms_total_intensity        <dbl> 8.11e+09, NA, 6.93e+09, NA, NA, NA, NA, NA, NA, 4.51e+08, N...
$ msms_percent_coverage       <dbl> 60, NA, 55, NA, NA, NA, NA, NA, NA, 33, NA, NA, NA, 53, NA,...
$ tissues                     <chr> "placenta", NA, "placenta", NA, NA, NA, NA, NA, NA, "liver"...
```

These columns contain strings!


stringr
www.rstudio.com

mito_proteins

```
glimpse(mito_protein)
```

```
$ yeast_mito_homolog_score    <chr> "NoMitoHomolog", "NoMitoHomolog", "NoMitoHomolog", "NoMitoH...
$ rickettsia_homolog_score    <chr> "NoHomolog", "NoHomolog", "NoHomolog", "NoHomolog", "NoHomo...
$ msms_score                  <chr> "50-75ambig", NA, "50-75ambig", NA, NA, NA, NA, NA, NA, "25...
$ mcarta2_score               <dbl> -1.4533, -3.3441, -4.0790, -13.4554, -10.6961, -8.2616, -10...
$ mcarta2_fdr                 <dbl> 0.459, 0.614, 0.679, 0.924, 0.887, 0.866, 0.887, 0.767, 0.4...
$ mcarta2_list                <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,...
$ mcarta2_evidence            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "target...
$ hg19_chromosome             <fct> 19, 10, 12, 12, 1, 22, 3, 12, 12, 3, 3, 1, 1, 4, 15, 2, 11,...
$ hg19_start                  <dbl> 58858171, 52559168, 9220303, 8975149, 33772366, 43088126, 1...
$ hg19_stop                   <dbl> 58864865, 52645435, 9268558, 9029381, 33786699, 43116876, 1...
$ msms_num_tissues            <dbl> 1, NA, 1, NA, NA, NA, NA, NA, NA, 1, NA, NA, NA, 2, NA, 4, ...
$ msms_num_peptides_unique    <dbl> 24, NA, 67, NA, NA, NA, NA, NA, NA, 9, NA, NA, NA, 24, NA, ...
$ msms_num_spectra            <dbl> 153, NA, 326, NA, NA, NA, NA, NA, NA, 11, NA, NA, NA, 168, ...
$ msms_total_intensity        <dbl> 8.11e+09, NA, 6.93e+09, NA, NA, NA, NA, NA, NA, 4.51e+08, N...
$ msms_percent_coverage       <dbl> 60, NA, 55, NA, NA, NA, NA, NA, NA, 33, NA, NA, NA, 53, NA,...
$ tissues                     <chr> "placenta", NA, "placenta", NA, NA, NA, NA, NA, NA, "liver"...
```

These columns contain strings!

Counting

How many AA's are in each protein?

String to be evaluated

```
str_count(mito_protein$sequence)
```

Counting

How many lysines are in each protein?

String to be evaluated

```
str_count(mito_protein$sequence, “K”)
```

“pattern” (what to *count*)

Exercise 1

Using the str_count() function, determine how many instances of lysine next to a arginine there are.

```
str_count(mito_protein$sequence, “KR”)
```

Logical evaluation

TRUE or FALSE

Presence or Absence

```
str_detect()
```

Logical evaluation

(R/K/N)CP(K/hydrophobic)(L/M)– heme regulatory motif

String to be evaluated

```
str_detect(protein$sequence, “RCP.L”)
```

“pattern” (what to *detect*)

Logical evaluation

(R/K/N)CPX(L/M)– heme regulatory motif

String to be evaluated

```
str_detect(protein$sequence, “RCP.L”)
```

“pattern” (what to *detect*)

Questions

Run the following code:

```
str_detect(protein$sequence, “RCP.L” )
```

1) How is this the same or different from str_count()?

2) What does the '.' stand for in this code?

3) How does this compare to the sequence motif we want to find?

Logical evaluation

(R/K/N)CPX(L/M)– heme regulatory motif

Regular Expressions (regex) are language used to describe patterns in strings

```
str_detect(mito_protein$sequence, "RCP.L")
```

'.' = will accept ANY character in this position

Logical evaluation

(R/K/N)CPX(L/M)– heme regulatory motif

(R/K/N)CPX(L/M)
=/=
RCPXL

```
str_detect(mito_protein$sequence, “RCP.L”)
```

How can we add more information to our pattern?

Logical evaluation

(R/K/N)CPX(L/M)– heme regulatory motif

```
str_detect(mito_protein$sequence, "RCP.(L|M)")
```

Will accept L *OR* M

Brackets indicate a space with multiple options

Exercise 2

How would you alter this code to accept R, K or N in the first position?

```
str_detect(mito_protein$sequence, “RCP.(L|M)”)
```

```
str_detect(mito_protein$sequence, “(R|K|N)CP.(L|M)”)
```

Exercise 2

How did you know it worked?? Try replacing ‘str_detect’ with ‘str_view’ (add “match” argument to see what is TRUE

```
str_detect(mito_protein$sequence, “(R|K|N)CP.(L|M)”)
```

```
str_view(mito_protein$sequence, “(R|K|N)CP.(L|M)”, match=TRUE)
```

Logical evaluation

```
str_view(mito_protein$sequence, "(R|K|N)CP.(L|M)", match=TRUE)
```

MAGTYSSTLKTLEDLTLDSGYGAGDSCRSLSLSSSKSNSQALNSSAQQHRGAAWWCYSGSMNSRHNSWDTVNTVLPEDPEVADLFSRCPRLPELEEFPWTEGDVARVLRKGAGGRRLPQFSAEAVRRLAGLLRRALIRV
MHQRHPRARCPPLCVAGILACGFLLGCWGPSHFQQSCLQALEPQAVSSYLSPGAPLKGRPPSPGFQRQRQRRAAGGILHLELLVAVGPDVFQAHQEDTERYVLTNLNIGAELLRDPSLGAQFRVHLVKMVILTEPEG
MESVVRRCPFLSRVPQAFLQKAGKSLLFYAQNCPKMMEVGAKPAPRALSTAAVHYQQIKETPPASEKDKTAKAKVQQTPDGSQQSPDGTQLPSGHPLPATSQGTASKCPFLAAQMNQRGSSVFCKASLELQEDVQEMNA
MVTAAMLLQCCPVLARGPTSLLGKVVKTHQFLFGIGRCPILATQGPNCSQIHLKATKAGGDSPSWAKGHCPFMLSELQDGKSKIVQKAAPEVQEDVKAFKTDLPSSLVSVSLRKPFSGPQEQEQISGKVTHLIQNNMPGI
MLGSLGLWALLPTAVEAPPNRRTCVFFEAPGVRGSTKTLGELLDTGTELPRAIRCLYSRCCFGIWNLTQDRAQVEMQGCRDSDEPGCESLHCDPSPRAHPSPGSTLFTCSCGTDFCNANYSHLPPPGSPGTPGSQGPQAA
MNGVAFCLVGIPPRPEPRPPQLPLGPRDGCSPRRPFPWQGPRTLLLYKSPQDGFGFTLRHFIVYPPESAVHCSLKEEENGGRGGGPSPRYRLEPMDTIFVKNVKEDGPAHRAGLRTGDRLVKVNGESVIGKTYSQVIALI
MAWRCPRMGRVPLAWCLALCGWACMAPRGTQAEESPFVGNPGNITGARGLTGTLRCQLQVQGEPPEVHWLRDGQILELADSTQTQVPLGEDEQDDWIVVSQLRITSLQLSDTGQYQCLVFLGHQTFVSQPGYVGLEGLPT
MNTKDTTEVAENSHHLKIFLPKKLLECLPRCPLLPPERLRWNTNEEIASYLITFEKHDEWLSCAPKTRPQNGSIILYNRKKVKYRKDGYLWKKRKDGKTTREDHMKLKVQGMECLYGCYVHSSIVPTFHRRCYWLLQNPI
MEQPWPPPGPWSLPRAEGEAEEESDFDVFPSSPRCPQLPGGGAQMYSHGIELACQKQKEFVKSSVACKWNLAEAQQKLGSLALHNSESLDQEHAKAQTAVSELRQREEEWRQKEEALVQREKMCLWSTDAISKDVFNKSI
MASLLPLLCLCVVAAHLAGARDATPTEEPMATALGLERRSVYTGQPSPALEDWEEASEWTSWFNVDHPGGDGDFESLAAIRFYYGPARVCPRPLALEARTTDWALPSAVGERVHLNPTRGFWCLNREQPRGRRCSNYHVE
MPKVMKDVVHPLGGEEPSMARAVVRSVGGFTLGLSLATAYGLLELLVEGHSPWGCLVGTLTLAAFLSLGMGFSRQVRATVLLLLPQAFSRQGRTLLLVAAFGLVLQGPCANTLRNFTRASEAVACGAELALNQTAEVLQE
MERNVLTTFSQEMSQLILNEMPKAEYSSLFNDFVESEFFLIDGDSLLITCICEISFKPGQNLHFFYLVERYLVDLISKGGQFTIVFFKDAEYAYFNFPELLSLRTALILHLQKNTTIDVRTTFSRCLSKEWGSFLEESYE
MGSKDHAVFFREMTQLILNEMPKAGYSSILNDFVESNFFVIDGDSLLVTCLGVKSFKWGQNLHFFYLVECYLVDLLSNGGQFTIVFFKDAEYAYFDFPELLSLRTALILHLQHNTNIDVQTEFSGCLSQDWKLFLEQHYE
MPARTAPARVPTLAVPAISLPDDVRRRLKDLERDSLTEKECVKEKLNLLHEFLQTEIKNQLCDLETKLRKEELSEEGYLAKVKSLLNKDLSLENGAHAYNREVNGRLENGNQARSEARRVGMADANSPPKPLSKPRTPRE
MTPELMIKACSFYTGHLVKTHFCTWRDIARTNENVVLAEKMNRAVTCYNFRLQKSVFHHWHSYMEDQKEKLKNILLRIQQIIYCHKLTIILTKWRNTARHKSKKKEDELILKHELQLKKWKNRLILKRAAAEESNFPERS

Using stringr with dplyr

We have a code chunk that determines presence of a motif in our sequence, but how can we add this information to our data frame?

dplyr function

```
mito_protein <- mutate(mito_protein, hrm_motif = str_detect(sequence,
      "(R|K|N)CP.(L|M)"))
```

Column to create

stringr function