Analysing the Fitness of GFP variants

In this assignment, you will parse another set of GFP mutagenesis data, analyse it and compare the results to the GFP variants we analysed in Exercises 1+2. The article describing generation and interpretation of this variant set is https://www.nature.com/articles/nature17995 (Sarkisyan_nature_GFP.pdf on Absalon).

There are “skeletons” of the code with all tasks in both R and Python on Absalon to get you started. The R code is also printed below.

Hand-in on Absalon:

- ❏ Your code in R or Python (or C++ if you prefer that)
- ❏ Answers to the questions in the task outline below, highlighted in bold also
- ❏ Plots as indicated in the different tasks

Please compile all but the source code into a single PDF and clearly indicate which task each answer belongs to. **Deadline: 2020 Oct 23, 23:59**

Questions?

Please ask on Absalon in the discussion forum.

```
# Task 1: quality control and translation
# ---------------------------------------

# quality control step 1: remove sequences that are too long, too
short, or have gaps, like in exercises 1+2
# useful function: nchar()

# translate to protein

# quality control step 2: remove sequences with premature STOP
codons

# Here, unlike in the GFP dataset we discussed in class, the
uniqueBarcodes column tells us how often a particular sequence has
been observed, while the medianBrightness tells us how bright
fluorescence in those cells was (with stdErr estimating the
measurement error, if available).
# How many unique barcodes (=DNA variant sequences) are found? How
many unique protein sequences after cleanup? What is the most common
protein sequence that is not wild-type? Include your answers in your
hand-in.
```

Task 2: protein-level variants

Next, determine differences to the native protein sequence. For simplicity we will consider each position independently, i.e., regardless of whether this mutation was only observed in context of other mutations. So, if a protein with A23P and S84Q is reported to have a medianBrightness of 3.5, then the brightness of A23P is 3.5 and the brightness of S84Q is 3.5 (in that protein). We also want to average the brightness across all contexts, so if i.e. there were 3 unique(!) proteins containing the A23P mutation, the averaged brightness of A23P is mean(brightness(protein1),brightness(protein2),brightness(protein3)).
#It may be convenient to generate a dataframe spanning all 20 possible amino acids at all positions, like we did in Exercise 2.

- see discussion in class on Oct 6 for useful functions to determine differences between wild-type and variant sequence

# As a control for the averaged data across different sequences, create a subset of the dataset where only single-mutation sequences are considered.
# useful functions: subset() and
substitution_distance <- function(s1,s2) { mapply(function(c1,c2) sum(c1!=c2), strsplit(s1,''), strsplit(s2,'')) }

Then compare the medianBrightness of those single-mutant sequences to the averaged data you created above. In our example above that would mean comparing the brightness of the A23P single mutant protein to the average brightness over all proteins that contain an A23P mutation. You should get scatter plots analogous to those we discussed in class on Oct 9th, see also https://www.biorxiv.org/content/biorxiv/early/2020/05/26/2020.05.26.116756/F7.large.jpg?width=800&height=600&carousel=1. Include the stdErr in the plot, using geom_errorbar() and geom_errorbarh(). **Are the deviations you observe beyond what you expect based on the experimental error? Submit plot and discussion as part of your hand-in.**

Next, pick 2 amino acids from your first and last name, respectively -> AA1, AA1
Visualise the distributions of mutations from AA1 and from AA2 to all other amino acids across all positions in the sequence. Then do the same for mutations from any amino acid into AA1 and AA2 - do they differ? **What would you expect based on biochemistry vs. what do you observe?**
 # - for synonymous mutations?

 # - for missense mutations?
 # This is analogous to Exercise 3.
 # Submit the plots for all 4 distributions as part of your handin. Of course you can plot all 4 distributions in the same figure, so long it is clear which line corresponds to which dataset.

```
# Task 3: summary matrix
# ----------------------
```

summarise the results across all variants in a 20x20 matrix showing the wild-type and target amino acids, as we did in the exercises in class. **Submit a plot of the matrix (see e.g. ex. 3) as part of your homework assignment.** You can use the colnames.by.AA to sort the amino acids:

```
mut.data$wt.ord <- factor( mut.data$aa.wt, levels =
colnames.by.AA)
mut.data$mut.ord <- factor( mut.data$aa.mut, levels =
colnames.by.AA)
# useful function: ddply(mut.data, ..., summarise, ...)
# https://colorbrewer2.org/ for colour schemes
```

```
# Task 4: compare to the other GFP mutagenesis dataset
# ----------------------------------------------------
```

Task 4.1
load in the native DNA from exercise 1
compare it to the nativeDNA included above (e.g. by pairwise sequence alignment), then translate both sequences to protein and compare those.
Write a short paragraph describing what you observe.

Task 4.2
Load in the GFP dataset we parsed in exercise 2, including the cleanup steps, translation to protein and identification of differences to the wt sequence.
 # How many variants (wt, position, mut.aa) are
 # - observed in both datasets?
 # - only observed in the Sarkisyan dataset?
 # - only observed in the dataset we worked with in class?

Task 4.3
For the variants found in both datasets, create a scatterplot to compare their averaged medianBrightness (see task 2) vs. log(bright/dim) ratio. Briefly describe what trends you observe, and whether those are what you would expect.

```
# Submit the scatter plot and discussion as part of your hand-in.
```