Assignment3

Stefano Pellegrini

10/21/2020

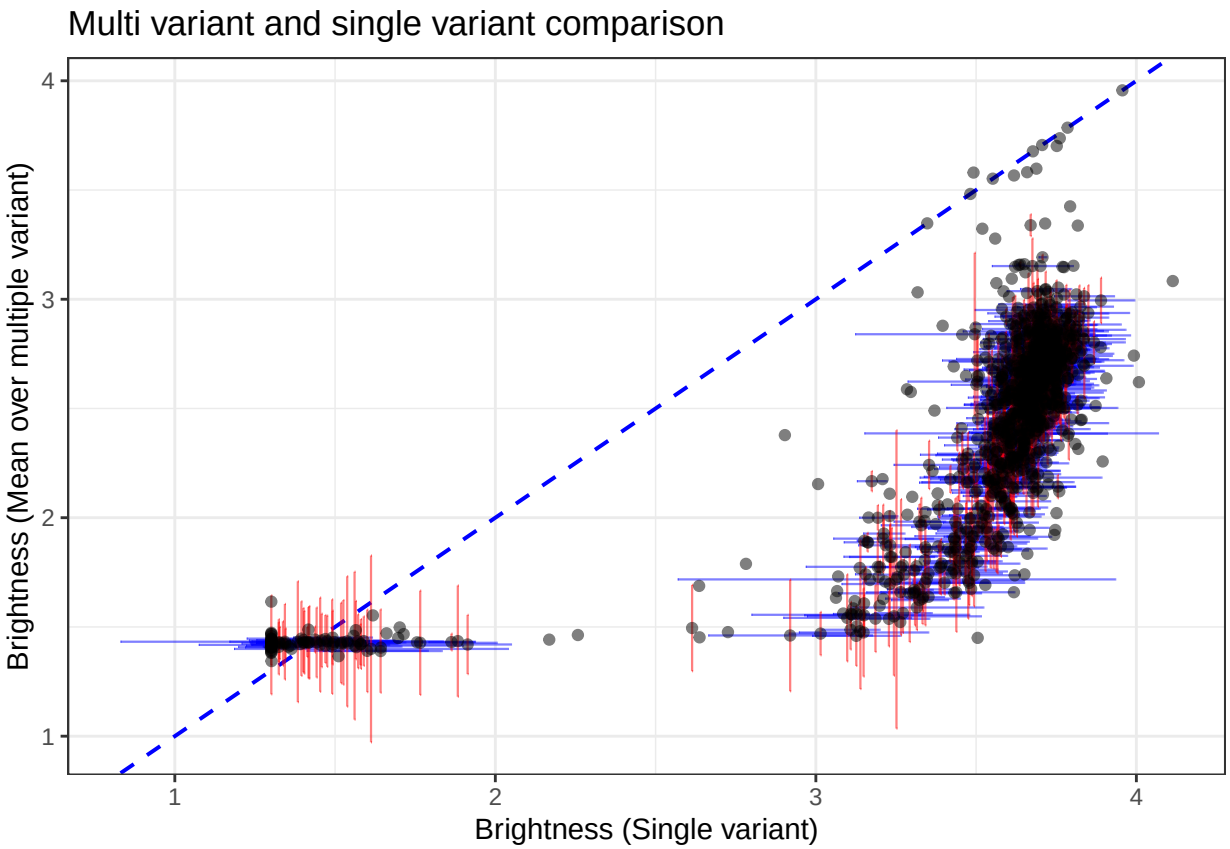
Task 1: quality control and translation

Quality control: remove sequences that are too long, too short, or have gaps. Then, remove sequences with premature STOP codons. 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.

After initial clean-up there are 56087 (DNA variant sequences) unique barcodes and 51716 unique proteins, so several DNA sequences that resulted in the same protein are present in the data. After merging the proteins duplicates (obtained from different DNA sequences), there are 51716 unique barcodes and 51716 unique proteins.
The most common protein (different from the wild-type) occurred 59 times, it has a median brightness of 3.698481, and it has the following sequence:
SKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTT
LSYGVQCFSRYPDHMKQHDFLKSAMPE GYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIE
LKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNT
PIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDELYK*.

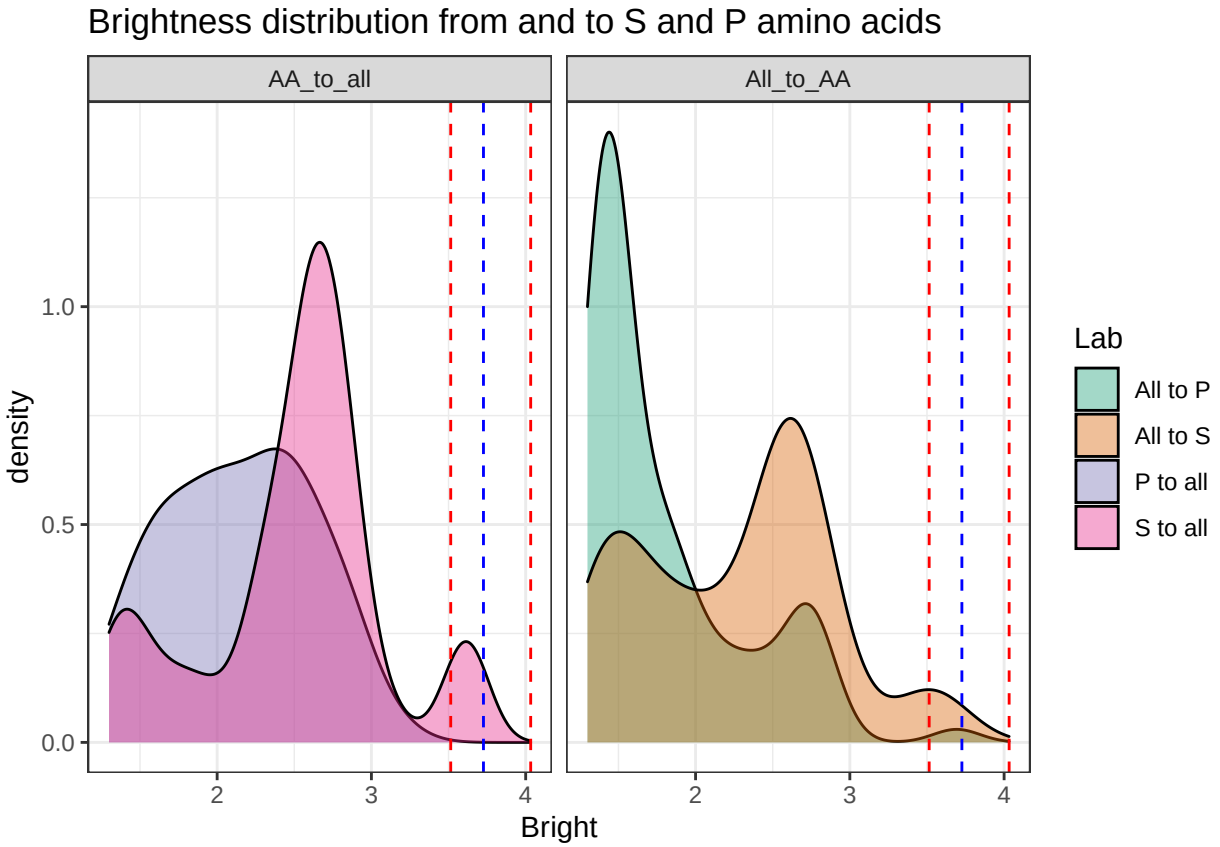
Task 2: protein-level variants

Next, determine differences to the native protein sequence. For simplicity we will consider each position independently. We also want to average the brightness across all contexts. As a control for the averaged data across different sequences, create a subset of the dataset where only single-mutation sequences are considered. Compare the medianBrightness of those single-mutant sequences to the averaged data you created above. 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.



The plot shows the comparison between the effect of single mutation and multiple mutations to the local fitness of avGFP, which was estimated by measuring the fluorescence levels of genotypes obtained by random mutagenesis. The blue dashed diagonal line rapresents the theoretical line where the fitness of the two groups is the same. We can see that the data points are shifted to the bottom right of the plot, indicating that the genotypes with multiple missense mutations are more likely to show a reduced fitness compared to genotypes with single mutation. One explaination is that these genotypes exhibit negative epistasis, where combination of neutral mutations had a deleterious effect on the protein, or that the cumulative effect of slightly deleterious mutations decreased the protein stability, therefore reducing the fluorescence. Also, we can observe that few data points are above the diagonal line, indicating the presence of positive epistasis. In these few cases the negative effect of a mutation was partially compensated by other mutations, which partially restored the fluorescence signal.

Next, pick 2 amino acids from your first and last name, respectively -> AA1, AA2. 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?



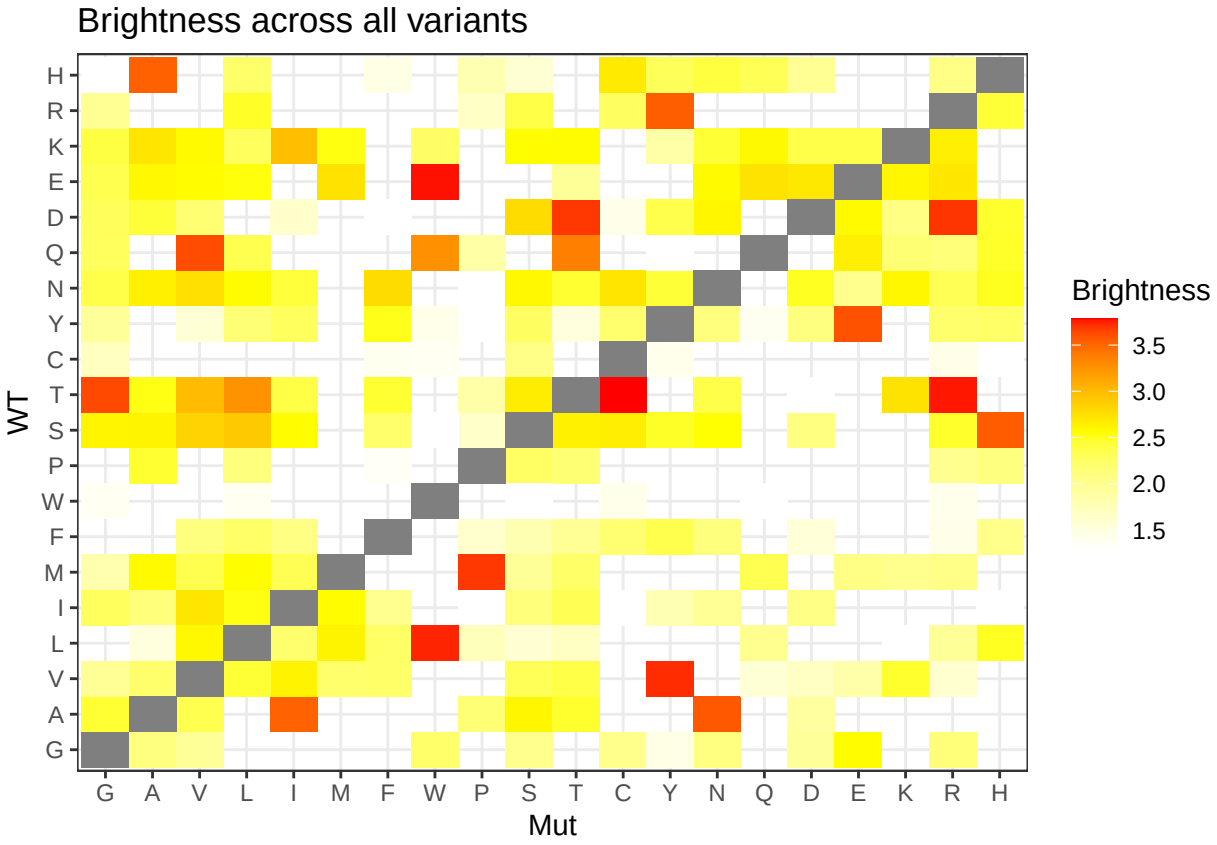
The red dashed lines rapresent the minimum and the maximum brightness of protein sequences resulting from synonymous mutations, while the blue dashed line rapresent their average brightness. The left side of the plot shows the distribution of the brightness measured in variants where the amino acid serine (S, pink) and proline (P, purple) mutated to each other amino acids. The right side shows the distribution of the brightness in variants where each amino acid mutated to serine (S, orange) and proline (P, green). Identical single nucleotide variants were averaged between protein sequences, and the brightness was also averaged across all contexts.

We can see that different peaks are present, which might rapresent the effect of mutations where the amino acid changed into another one within the same or a different group. In fact, if an amino acid mutates into another one within the same group, since they share similar properties, we expect that the stability of the protein is less affected than what we would observe if the amino acid is mutated into another one of a different group. In particular, apart from mutations involving proline as wild-type amino acid, it is possible to observe that three peaks are present in each distribution. The central peak might rapresent mutations to amino acids with partially shared properties whose effect was small, while the peak on the left rapresent significantly deleterious mutations, probably involing buried residues. Lastly, the peak on the right rapresent mutations to amino acids that did not have a deleterious effect on the fluorescence. In fact, it corresponds to the same brightness of proteins resulting from synonymous mutations.

Regarding the mutations involing proline, we can see that they are more deleterious than mutations involving serine. This might be explained by the special structure of the proline amino acid, which affect the protein secondary structure. In fact, cause of the distinctive cyclic structure of its side chain, proline is the least flexible between the amino acids and thus it gives conformational rigidity to the protein. Also, if inserted in the middle of apha helices and beta sheets, it can disrupt the secondary structure of the protein. Therefore, mutations from and to proline usually have a large impact on protein stability. Meanwhile, serine is a slightly polar amino acid which is consider neutral in regard to mutations. It has a small size and it can reside both within the interior of a protein, or on the protein surface.

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.



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.

```
## [1] "DNA alignment:"

## [1] 1
## [1] "AGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGC"
## [1] "-----------------------------------------------------------------------------------"
##
##
## [1] 81
## [1] "CGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGT"
## [1] "-----------------------------------------------------------------------------------"
##
##
## [1] 161
## [1] "TGCCCTGGCCCACCCTCGTGACCACCCTGTCGTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACG"
```

```
## [1] "-------------------------------------------------------------------------------------"
##
##
## [1] 241
## [1] "GACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACC"
## [1] "-------------------------------------------------------------------------------------"
##
##
## [1] 321
## [1] "CCGCGCCGAGGTGAAGTTCGAGGGCGACACACTAGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAA"
## [1] "-----------------------------------------------------------------------------GGTAA"
##
##
## [1] 401
## [1] "ACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATCA"
## [1] "ACATCTTGGGCCACAAACTTGAATATAATTACAACAGCCATAACGTCTATATTATGGCAGACAAACAAAAAAATGGAATCA"
##
##
## [1] 481
## [1] "AAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATC"
## [1] "AAAGTGAACTTTAAAATTCGCCACAATATTGAGGATGGCAGTGTGCAGCTGGCGGACCACTACCAGCAGAATACCCCAATC"
##
##
## [1] 561
## [1] "CGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCG"
## [1] "CGGTGACGGTCCGGTGCTTCTGCCCGATAACCATTATCTGTCAACTCAGAGCGCTCTATCTAAAGATCCTAACGAGAAACG"
##
##
## [1] 641
## [1] "GCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCACGGCATGGACGAGCTGTACAAGTGA"
## [1] "GTGATCACATGGTACTGCTCGAATTTGTTACGGCTGCCGGCATTACCTA------------------------A"
##
##
## [1] "Alignment starting at position 397 of the pattern"

## [1] "Protein alignment:"

## [1] 1
## [1] "SKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLSYGVQCFSRYPDHMKQHD"
## [1] "-------------------------------------------------------------------------------------"
##
##
## [1] 81
## [1] "DFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIK"
## [1] "-----------------------------------------------------GNILGHKLEYNYNSHNVYIMADKQKNGIK"
##
##
## [1] 161
## [1] "KVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDELYK*"
## [1] "KVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGIT--------*"
##
##
## [1] "Alignment starting at position 133 of the pattern"
```

We can see that some mismatches are present in both, the alignment of DNA and protein sequences. The

DNA alignment start at position 397 of the native DNA used as reference for the Sarkisyan dataset, while the alignment of the proteins start at position 133. This information is important to compare the variants present in Sarkisyan and GFP mutagenesis datasets. In order to do that, we can add 132 to the positions of the GFP mutagenesis variants.

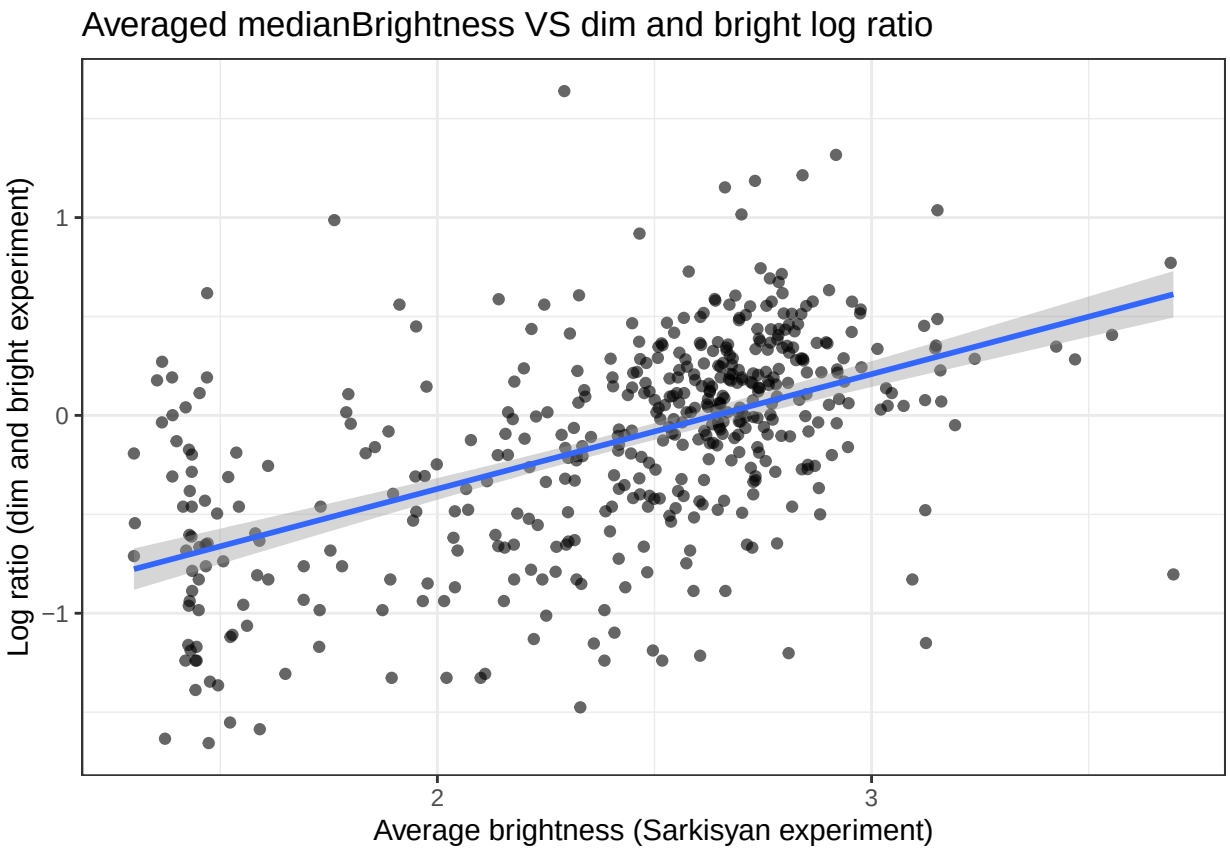
Task 4.2

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?

- Variants present in both GFP and Sarkisyan datasets: 465.
- Variants present only in Sarkisyan datasets 1346.
- Variants present only in GFP datasets: 59.

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.



We can observe that there is a positive linear relationship between the log ratio of the GFP experiment (dim and bright dataset) and the average brigthness of the Sarkisyan one. This was expected because the variants that caused a fitness reduction in the GFP experiment were more likely to show the same deleterious effect in

the Sarkisyan one, and the same was expected for variants that shown a neutral effect. We can see that this is not always true, and this might be due to noise and to the presence of different combination of mutations (Sarkisyan proteins are longer), which might have a very different effect on the fitness of the genotypes.