1. (1) Scoring scheme 5. A perfect alignment would be all matches. A match should grant more points than anything else. This is the only scheme that does this.

   (2) Scoring scheme 4. This could be somewhat reasonable if the only goal was to identify sequences of the same length.

   (3) Scoring scheme 1. This should align the sequences as far apart as possible, but where they meet should be as much matches as possible. Doesn't make sense, but it's better than the others.

   (4) Scoring scheme 3. Same as scheme 1, but more points for further apart sequences.

   (5) Scoring scheme 2. This will just try to make gaps. It doesn't care about anything else.

   (6) Scoring scheme 6. Avoid matches, prefer gaps. This makes no sense.

2. 2 cases. The first is when there is no gap (or the gap is size 0, or it is to the left of the first or the right of the last base). In this case, there are 8 ways to place the bases on their row. The second case is when the gap is at least size 1 and between two bases. In the second case, there are 4 places to put the gap, and $8 - g$ ways to place the sequence on the grid for a gap of size g. There are $8 + 4(7 + 6 + 5 + 4 + 3 + 2 + 1) = 120$ arrangements for each row, leading to a grand total of $120^6 = 2\,985\,984\,000\,000$ total possible grid arrangements. Some of these will be effectively the same alignments.

3. RAGM $\rightarrow 4, 4, 4, 1$ possible base pair sequences for each, respectively. Total would be $4 \times 4 \times 4 \times 1 = 64$ possible DNA substrands.

4. (a) 19 possible substitutes for any amino acid, with 5 choices of amino acid, is $19 \times 5 = 95$ possible mutants.

   (b) Same number of possible first mutations, times 19×4 choices for the second mutation, for $95 \times 19 \times 4 = 7220$ possible mutations. Divide that by 2 because there is no difference based on which was chosen first, for 3610 total possible mutations.

   (c) 19 replacement choices for each amino acid mutated is 19^k. 'Order' the n amino acids $n!$ different ways, select the first k, divide by the number of ways those could be ordered and divide by the number of ways those not selected could be ordered.
   $$19^k \frac{n!}{k!(n-k)!}$$

Introduction & Motivation

The goal of this paper is to explore the effects of *splicex* on the wild type of the protein best fitting a certain protein sequence. The protein sequence was found to best match Barnase. The purpose of the drug is to disrupt the function of the protein through mutating a Phenylalanine in the wild type protein to either an Alanine or a Lysine.

Methods

First I used protein BLAST [1] to try to identify the protein, which matched very well to Barnase and many of its mutants. Unfortunately, the PDB entry for Barnase was not included in the BLAST result, so I had to perform a PDB search and use my best judgment to determine the wild type protein entry. Because there multiple entries, I decided to incorporate the results from two difference PDB entries into my results, each referring to wild type Barnase. From here I simply used SDM (Site Directed Mutator) [2] to predict change in stability ($\Delta\Delta G$) for each mutation. I additionally cross referenced the results with a Barnase PDB entry that represented wild type Barnase interacting with a ligand, as that would provide insight into Barnase while functioning.

Swap Rankings

Boxplot of ΔΔG

Residue	Mutation	1BNR	1BNS	1A2P
82	Alanine	+0.78	+0.78	+0.78
7	Lysine	+0.23	+0.26	+0.14
82	Lysine	-0.02	-0.14	-0.14
7	Alanine	-2.08	-0.34	+0.97
56	Lysine	-1.50	-0.07	-0.27
56	Alanine	-3.79	-1.55	-0.60
106	Lysine	-1.51	-1.53	-1.50
106	Alanine	-3.80	-3.80	-3.73
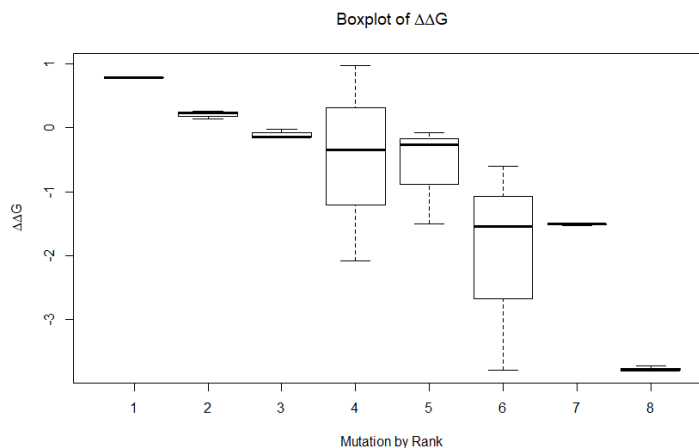
Justification

1. 82A has consistent structural disruption. Clear winner.

2. 7K has little structural disruption.

3. 82K simply has nearly 0 effect, but consistently negative.

4. 7A is high variance. It's difficult to determine what the true effect of this mutation most likely is, so I rank it poorly. Minimizing my chance of a false positive ranking.

5. 56K is relatively near 0, but is high variance.

6. 56A is inconsistent, so while it does have very low $\Delta\Delta G$, there is better chance the prediction is off compared to lower ranked predictions.

7. 106K is consistently stabilizing. Seems mutating residue 106 is unlikely to disrupt the protein.

8. 106A is clearly predicted to stabilize the protein. All samples contain very significant negative $\Delta\Delta G$ scores.

Conclusions

Phenylalanines 7 and 82 seem to have the most significant impact on the protein's structure, with mutations in those positions predictively causing the most disruption when mutated. I conclude the most likely choice for *splicex* to disrupt the protein is to replace the Phenylalanine at residue 82 with an Alanine.

References

[1] Stephen F. Altschul et al. "Basic local alignment search tool". In: *Journal of Molecular Biology* 215.3 (1990), pp. 403 –410. ISSN: 0022-2836. DOI: `https://doi.org/10.1016/S0022-2836(05)80360-2`. URL: `http://www.sciencedirect.com/science/article/pii/S0022283605803602`.

[2] Arun Prasad Pandurangan et al. "SDM: a server for predicting effects of mutations on protein stability". In: *Nucleic Acids Research* 45.W1 (2017), W229–W235. DOI: `10.1093/nar/gkx439`. eprint: `/oup/backfile/content_public/journal/nar/45/w1/10.1093_nar_gkx439/5/gkx439.pdf`. URL: `http://dx.doi.org/10.1093/nar/gkx439`.