2(b):

Some sequences are misclassified because of:

- The models are not perfect: The models are based on the probability of a codon changing from one codon to another, but there are always some exceptions to the rule.
- Limited number of sequences in dataset: The models are trained on a dataset of 40 sequences. So they may not be able to capture all of the variation that exists in real-world sequences.

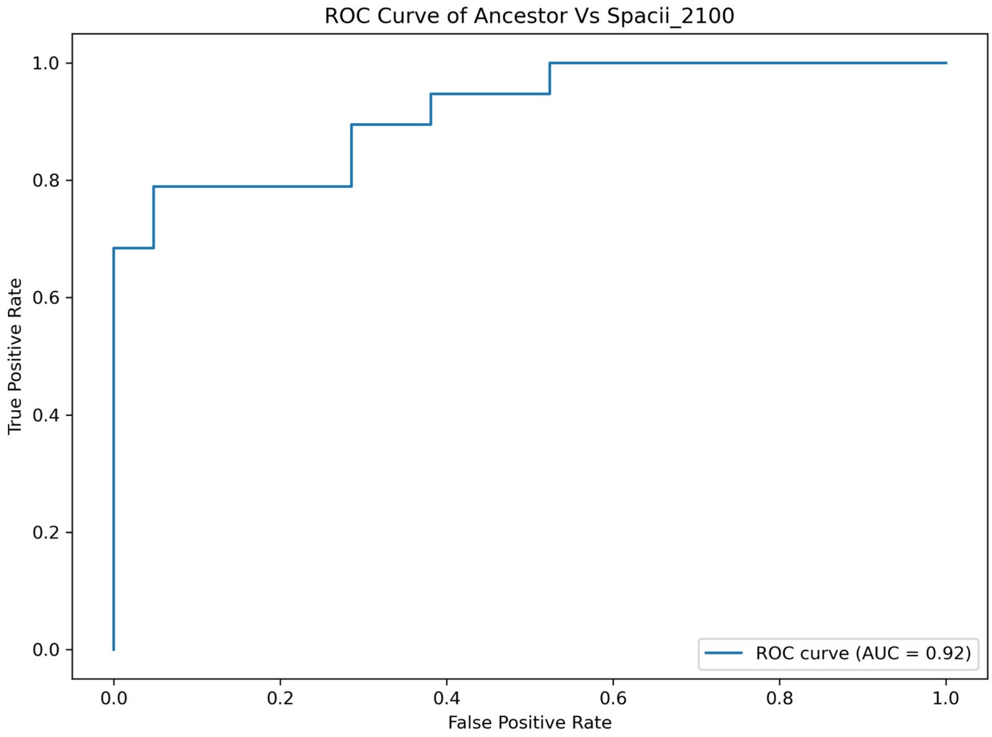
2©:

The ROC curve shows the trade-off between specificity and sensitivity for various threshold values. Specificity is the proportion of negative cases that are correctly identified as negative, while sensitivity is the proportion of positive cases that are correctly identified as positive.

We can choose cutoff like the following:

- Cutoff closer to 1: When we are more concerned with avoiding false positives (i.e., misclassifying non-coding sequences as coding), we can choose a cutoff closer to 1. This would result in a lower sensitivity, but a higher specificity.
- Cutoff closer to 0: When we are more concerned with avoiding false negatives (i.e., misclassifying coding sequences as non-coding), we can choose a cutoff closer to 0. This would result in a higher sensitivity, but a lower specificity.

The ROC curve of Spacii_2100 is given below:



3(a):

The coding matrix at time 2t will be the product of the coding matrix at time t with itself. This is because the probability of a codon changing from one codon to another is independent of the probability of that codon changing to another codon.

The probability of TTT staying TTT after time 2t will be affected by the each codon transition probability with TTT.

The following array contains TTT to other 64 codons transition probability.

```
[8.87119805e-01, 8.44223379e-02, 4.94706357e-03, 4.85484205e-03,
1.34155708e-02, 1.47330746e-03, 4.03780394e-04, 3.93691270e-04,
2.92617943e-03, 2.79923670e-04, 2.55114556e-04, 1.74349793e-04,
4.02501497e-03, 3.83980532e-04, 2.70325066e-04, 1.80825922e-04,
7.05352437e-03, 1.15154582e-03, 5.78102157e-04, 5.06491533e-04,
2.54029837e-04, 1.77607275e-05, 9.09061525e-04, 1.90701651e-04,
2.75102083e-05, 1.06333206e-05, 3.59933329e-04, 5.24105091e-05,
1.40920493e-05, 3.84118001e-05, 3.13997949e-04, 6.23719576e-05,
3.54707411e-05, 5.92031259e-05, 8.12958172e-03, 9.65805765e-04,
5.31513134e-04, 1.26771872e-04, 4.33809904e-05, 1.77816414e-05,
1.04567509e-04, 2.19975216e-05, 1.41332369e-05, 1.07350273e-05,
1.42561610e-04, 1.36985926e-05, 9.58331736e-06, 9.98543790e-06,
3.10996193e-03, 2.42343461e-04, 2.72796712e-04, 5.97502440e-05,
1.26324832e-05, 2.86718620e-05, 6.84407334e-05, 6.21264646e-06,
1.79271869e-04, 9.23343983e-06, 4.47579490e-06, 1.24968237e-05,
2.73098024e-04, 2.83286119e-04, 1.37504297e-03, 1.61121405e-04]
```

We can square each probability, sum them and take mean to get the probability of TTT staying TTT after time 2t.

After calculating we get probability at time 2t = 0.0124

The calculation described above provides an estimate of the probability at time 2t, not the exact probability. This is because the calculation assumes that the probabilities at time t are independent of each other. In reality, the probabilities at time t may be correlated, which means that the calculation will not be accurate.

3(b):

If the sequences had longer to diverge, the coding matrix would become more diagonal. This is due to the transition probability of each codon would become smaller as the sequences diverge. This could have a significant impact on the classification success. The model could be less capable to distinguish between coding and non-coding sequences which might result in a decrease in both specificity and sensitivity.

The ROC curve would change as t increases. The curve would move to the left, indicating a decrease in specificity. The curve would also become less steep, indicating a decrease in sensitivity.