Part 1: Analysis

Q2, Q3. Python was used to compute the overall amino acid composition in TMH and TMB. The table below displays the composition and Fisher Discriminant Ratio (FDR) to six decimal places.

Amino Acid	**Composition (TMH)**	**Composition (TMB)**	**Fisher Discriminant Ratio**
A	0.088315	0.080537	0.029069
C	0.011713	0.004354	0.209947
D	*0.036240*	*0.065945*	*1.416455*
E	0.044570	0.048207	0.016416
F	0.057735	0.043620	0.268398
G	0.074383	0.095614	0.277331
H	0.022347	0.022626	0.000158
I	0.067272	0.039876	0.790544
K	0.044578	0.048739	0.018414
L	0.118257	0.080344	0.829171
M	0.030375	0.016019	0.582138
N	0.033895	0.058643	0.975783
P	0.046126	0.036355	0.140816
Q	0.030716	0.042411	0.253831
R	0.044323	0.048234	0.020958
S	0.064703	0.075902	0.145255
T	0.052418	0.064472	0.238494
V	0.075926	0.061454	0.235982
W	0.019926	0.017853	0.013828
Y	0.035460	0.048796	0.289638

Q4. The residue pair preferences were then computed. The table below shows the five most preferred pairs for Alpha and Beta, and the five pairs with the highest FDR. For the five most preferred pairs of Alpha and Beta, the number in brackets refers to their abundance. For the five most important pairs based on FDR, the number in brackets is their FDR.

Top	Alpha (TMH)	Beta (TMB)	FDR
1	LL (0.014868)	AG (0.009422)	LL (0.666240)
2	AL (0.011773)	LG (0.009218)	LI (0.536511)
3	LA (0.011469)	GG (0.009137)	IL (0.518132)
4	AA (0.009789)	GL (0.008769)	DN (0.479397)
5	GL (0.009781)	GA (0.007997)	VL (0.473729)

Part 2: Discrimination

Q5, Q6, Q7. The steps in the assignment instruction were performed, resulting in the following confusion matrix.

	Alpha	Beta
Predicted as Alpha	614	133
Predicted as Beta	86	11

We therefore have 614 true positive, 86 false negative, 11 true negative, and 133 false positives. We can therefore compute the sensitivity, specificity, and accuracy.

Sensitivity	0.877143
Specificity	0.076389
Accuracy	0.764218

Q9. We now repeat the above using only 50% of TMH and 50% of TMB sequences. We only take the first half of each set of sequences. Our calculations result in the following confusion matrix.

	Alpha	Beta
Predicted as Alpha	298	64
Predicted as Beta	52	8

We therefore have 298 true positive, 52 false negative, 8 true negative, and 64 false positives. We can therefore compute the sensitivity, specificity, and accuracy.

Sensitivity	0.851429
Specificity	0.125
Accuracy	0.725118

Although the sensitivity and accuracy deproves, specificity improves. However, compared to the machine learning models from Assignment 1, this classifier is still not good enough.

Part 3: Comparison of different features

(a) We first repeat the work done in Part 2 using the residue pair preference (all 400 of them) to predict TMB and TMH, yielding the following confusion matrix.

	Alpha	Beta
Predicted as Alpha	630	135
Predicted as Beta	70	9

We therefore have 630 true positive, 70 false negative, 9 true negative, and 135 false positives. We can therefore compute the sensitivity, specificity, and accuracy.

Sensitivity	0.9
Specificity	0.0625
Accuracy	0.757109

We then repeat this using only 50% of TMH and 50% of TMB data, yielding the following confusion matrix.

	Alpha	Beta
Predicted as Alpha	317	66
Predicted as Beta	33	6

We therefore have 317 true positive, 33 false negative, 6 true negative, and 66 false positives. We can therefore compute the sensitivity, specificity, and accuracy.

Sensitivity	0.905714
Specificity	0.083333
Accuracy	0.765403

We see that using 50% of the data set improves the sensitivity, specificity, and accuracy of the classification is better than using 100% of the data. Although this is the most sensitive and accurate classifier so far, all the machine learning models from Assignment 1 still outperform it.

(b) We now combine both the amino acid composition and residue pair preference, obtaining the following confusion matrix.

	Alpha	Beta
Predicted as Alpha	617	133
Predicted as Beta	83	11

We therefore have 617 true positive, 83 false negative, 11 true negative, and 133 false positives. We can therefore compute the sensitivity, specificity, and accuracy.

Sensitivity	0.881429
Specificity	0.076389
Accuracy	0.744076

(c) Finally, we choose to use only selected features. For amino acids, we only use “D” and “N”, which are the two amino acids with the highest FDR. For residue pair preferences, we only use “LL”, “LI”, “IL”, “DN”, and “VN,” which have the 5 highest FDR values. This produces the following confusion matrix.

	Alpha	Beta
Predicted as Alpha	601	129
Predicted as Beta	99	15

We therefore have 601 true positive, 99 false negative, 15 true negative, and 129 false positives. We can therefore compute the sensitivity, specificity, and accuracy.

Sensitivity	0.858571
Specificity	0.104167
Accuracy	0.729858

Although the specificity and accuracy for this model is very low, it also uses much fewer features than the other models. This model uses a total of 7 features (2 amino acids and 5 residue pairs), compared to the other classifiers which used 20, 400, and 420 features respectively.

In Assignment 1, we concluded that only D and N amino acids have any meaningful predictive power. This is mimicked here, as D and N have the highest FDR for amino acids (1.416155 and 0.975783 respectively). As classification using residue pairs was not performed in Assignment 1, we cannot compare whether the classifier we built here is as effective as the machine learning models in Assignment 1.

Finally, the worst performing machine learning models in Assignment 1 had an accuracy of 85.50%, which is still higher than our best performing model (which uses all residue pair preferences and was trained on 50% of the data), which had an accuracy of 76.54%. This is actually in disagreement with Assignment 1 – in Assignment 1, we found that training the models on 50% of the data instead of 100% of the data resulted in lowered sensitivity, specificity, and accuracy. However, for this classifier, training with only 50% of the data produced the best classifier.

Link to code: https://github.com/chengjunyuan/cs2220-a2