1) The rarest amino acid in rice is 'W' or ‘Tryptophan’ which has a frequency of 181701 in the whole sequence.

2) Most abundant amino acid here is Alanine, which has a frequency of 1335216.

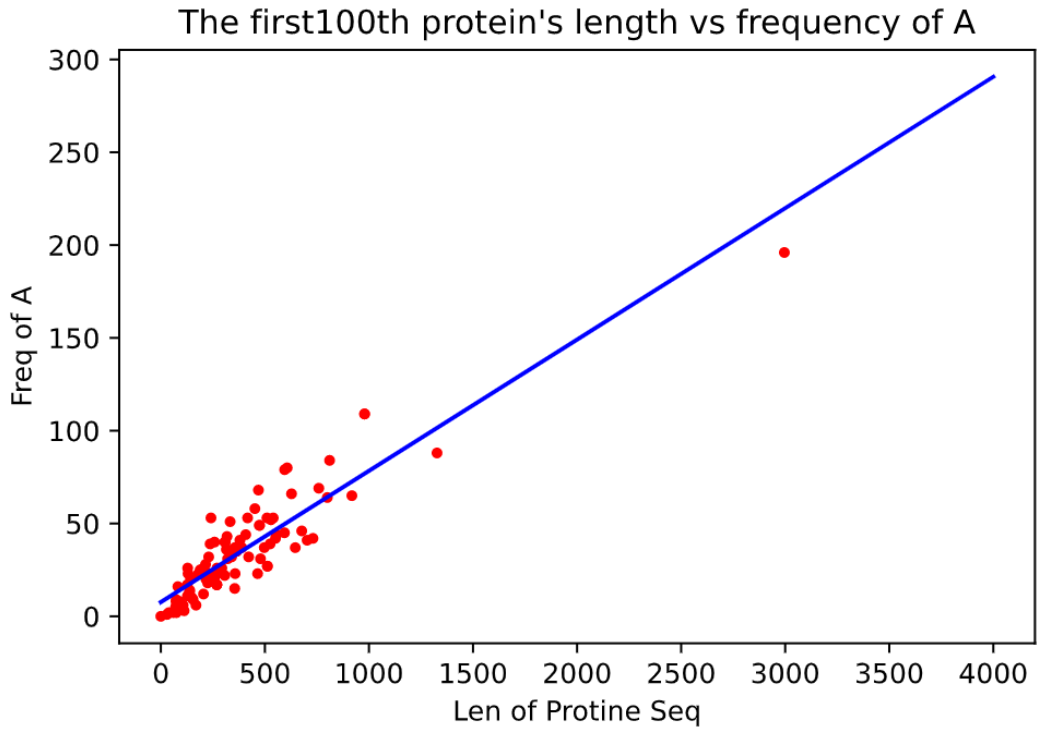
Now the model for linear regression, $y_i = a_o + a_1 * x_i$

$$a_1 = \frac{n\,\Sigma xiyi - \Sigma xi\,\Sigma yi}{n\Sigma xi^2 - (\Sigma xi)^2}$$

$$a_o = \overline{y} - \overline{x}.a_1$$

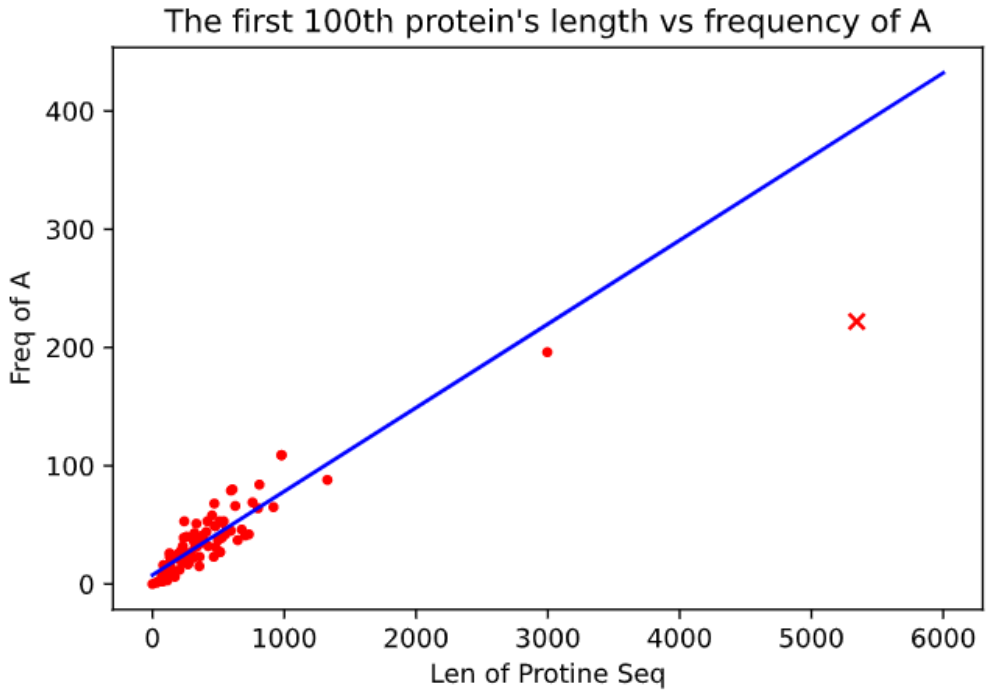
Using 100 protein’s length and frequency of Alanine in them as training data for the model, we get,
a_o , a_1 = (7.636320635929838 0.07074419670774298)

The following graph shows the prediction line and the actual data points.



3) Finding the outlier protein:
We can see that the 1922th protein is the outlier protein that has the largest discrepancy between the prediction and the actual number. (actual count 222 predicted count = 386)
It is >Os01t0356800-00 Domain of unknown function DUF3406, chloroplast translocase domain containing protein.

Here the outlier is marked as X in the previous graph.



4) Finding the amino acid that yields the most robust linear model:
We can use the linear model to predict data and find the residual (R_i = Actual Data - Predicted Data). The amino acid with the least average of absolute residuals will have the most robust model.

Based on this criteria Vanaline has the most robust model when we use the first 100 proteins as training data.

Submitted by,
Mahdee Mushfique Kamal