An additional 31 proteins in our dataset were identified in a logistic regression model to predict MSI. We included these 62 proteins in a model to distinguish sex in the test data. At an adjusted p-value threshold of 0.05, we discovered 31 proteins strongly associated with MMR dysfunction. In total, we used 62 protein expression profiles to build a logistic regression model predicting the MSI status for unlabelled patients.

\[
\ln \left( \frac{p_0}{1-p_0} \right) = \beta_0 + \beta_{\text{protein}} \cdot x + \cdots + \beta_{\text{protein}} \cdot x^{22}
\]

Determining Microsatellite Instability

A key aspect of the challenge was identifying the correct microsatellite instability (MSI) status for patients using only proteomic data. Recent work shows that the mutator phenotype caused by mismatch repair deficiency (MMRD) exerts broad and predictable effects on the transcriptome, allowing RNA expression to predict the mutator phenotype with high specificity and sensitivity. As an extension of this work, our hypothesis was that high MSI will lead to a greater variance in protein expression in exons with accumulating SNVs. We applied the following formal hypothesis to each protein:

\[
H_0: \sigma_{\text{high MSI}}^2 = \sigma_{\text{low MSI}}^2 \\
H_1: \sigma_{\text{high MSI}}^2 \neq \sigma_{\text{low MSI}}^2
\]

To test this hypothesis we calculated the F-statistic for each protein, which is a ratio of the expression variance:

\[
F = \frac{\overline{\text{SSD}_{\text{high MSI}}} + \overline{\text{SSD}_{\text{low MSI}}}}{\overline{\text{SSD}_{\text{high MSI}}} + \overline{\text{SSD}_{\text{low MSI}}}}
\]

At an adjusted p-value threshold of 0.05, we discovered 31 proteins from the total set of 4,117 for which we rejected the null hypothesis. In addition, we did a literature search that yielded an additional 31 proteins that are dysregulated in unstable tumor genomes.

Determining Sex

We examined several proteins coded on the Y-chromosome to identify men from women, since women should have zero expression of these genes. However, we found these to have high specificity but low sensitivity due to incomplete measurements in men. We decided to look for the genes with the highest prediction value by calculating Cohen's kappa coefficient for these sex-linked genes:

\[
\kappa = \frac{p_o - p_e}{1 - p_e}
\]

Where \(p_o\) is the prediction accuracy of the protein, and \(p_e\) is the expected accuracy due to random chance. The two highest scoring proteins were those coded by DDX3Y and RPS4Y1. We used both in order to distinguish sex in the test data.

Results

<table>
<thead>
<tr>
<th>Ground Truth</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>58</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

Sensitivity = 0.833
Specificity = 0.853
F1 = 0.625
Accuracy = 0.850

Conclusions

When comparing the prediction of sex and MSI by our model to the actual prediction, any mismatches were determined to be mislabelled data. Results on the unseen test set yielded sensitivity of 0.833 and a specificity of 0.853. These results demonstrate the potential of machine learning-based techniques to predict sample mislabeling. The next steps would be to test the current model on real tumor research data and then improve the model with a larger training data of naturally occurring mislabels. With larger sample sizes more advanced techniques can be applied, such as neural networks and random forest to improve performance and generalizability and, ultimately, reduce the quantity of human errors in research.

References


