1. Get the files **orf\_coding.fasta** and **NotFeature.fasta** from ftp://genome-ftp.stanford.edu/pub/yeast/ (you will find the files in two different subdirectories).

For each sequence in each of the two files, compute the Fourier measure

\[
F(k) = \frac{1}{L} \sum_{p=1}^{150} \sum_{m=1}^{L} \text{EQ}(X_m, X_{m+p}) e^{2\pi ip/k},
\]

for \( k = 2, \ldots, 9 \), \( \text{EQ}(X, Y) = 1 \) if \( X = Y \), \( \text{EQ}(X, Y) = 0 \) if \( X \neq Y \), and \( L \) is the length of the sequence. Then determine the modulus \( G(k) = \sqrt{\text{Re}(F(k))^2 + \text{Im}(F(k))^2} \). You can ignore sequences in either file that are shorter than 200nt.

(a) For each \( k = 2, \ldots, 9 \), carry out a t-test to see how strongly \( G(k) \) distinguishes the coding set from the non-coding set. You can use the simplified t-statistic:

\[
t = \frac{\bar{Y} - \bar{X}}{S},
\]

\[
S^2 = \frac{1}{N_X + N_Y} \left( \sum_i (Y_i - \bar{Y})^2 + \sum_i (X_i - \bar{X})^2 \right),
\]

where \( X_1, \ldots, X_{N_X} \) and \( Y_1, \ldots, Y_{N_Y} \) are the two lists of numbers being compared.

T-statistics:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8030</td>
</tr>
<tr>
<td>2</td>
<td>-0.3672</td>
</tr>
<tr>
<td>3</td>
<td>1.1669</td>
</tr>
<tr>
<td>4</td>
<td>0.3968</td>
</tr>
<tr>
<td>5</td>
<td>-0.1993</td>
</tr>
<tr>
<td>6</td>
<td>-0.1911</td>
</tr>
<tr>
<td>7</td>
<td>0.5199</td>
</tr>
<tr>
<td>8</td>
<td>0.3364</td>
</tr>
<tr>
<td>9</td>
<td>0.4419</td>
</tr>
</tbody>
</table>

(b) For each of the three pairs selected from the top three \( G(k) \), produce a scatterplot, where the non-coding sequences have a different color or symbol from the coding sequences.
Coding
Non-Coding

G(3)

G(1)

Coding
Non-Coding

G(7)

G(1)
(c) Using all 8 $G(k)$ values, construct a nearest neighbor predictor to predict whether a sequence is coding or non-coding (use the majority vote of the three nearest neighbors in Euclidean distance). Report the sensitivity and specificity.

I get 0.74 for the sensitivity and 0.93 for the specificity.