BMI 203

Homework 3

Erin Gilbertson

**Part 1**

1. In order to save time, I first set an intermediate (-10) gap opening penalty and tested for the optimal gap extension penalty. Then, using that gap extension penalty I tested for the optimal opening penalty. I recognize that there are issues with this approach given that many possible pairs of penalties are not tested, but testing every possible pair was not feasible in the time I dedicated to this part of the assignment.

This method resulted in a lowest false positive rate of 0.239 using a gap opening penalty of 10 and a gap extension penalty of 1.

Code to achieve this result as well as a complete output of all tested combination can be viewed in part1.ipynb.

1. A close up of a map

   Description automatically generated Using the gap penalties determined above, the PAM100 matrix performs best at a TPF of 0.7. It also has the best AUC overall as shown in the ROC plot below. Code for generating this plot and doing calculations can be found in part1.ipynb.
2. The performance decreases markedly when Smith-Waterman score is normalized by shorted sequence length as shown in the ROC plot below. I think this is because the chances of a randomly occurring good alignment increase with sequence length.

A close up of a map

Description automatically generated

**Part 2:**

I got quite close to implementing a useful genetic algorithm to optimize a scoring matrix for alignment but did not quite get everything working properly. The code that I attempted, as well as comments for what when wrong are found in part2.ipynb. I am not yet sure how I might improve my algorithm beyond making it actually work right, and a few minor details mentioned in the commenting.