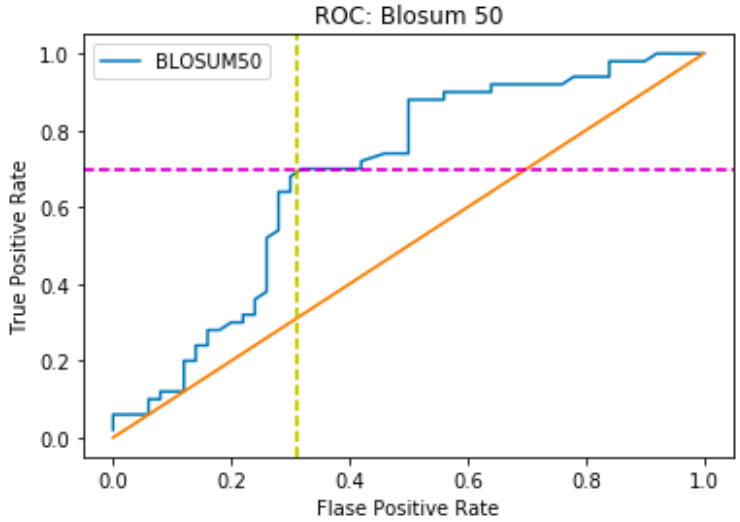
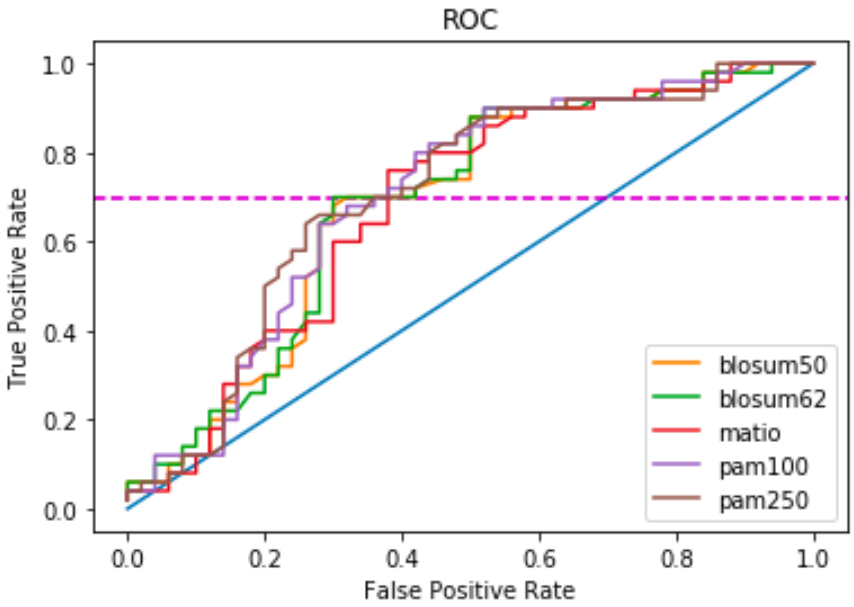
Snow Naing
BMI203
HW-3: Smith-waterman algorithm implementation and optimization algorithm

All my codes can be found in smithwaterman.ipynb.

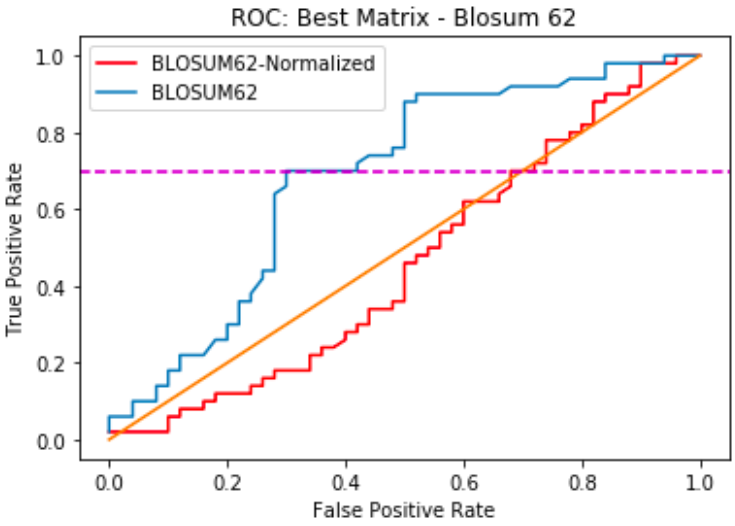
1. I used scikit bio package to do smith-waterman alignment and ran every possible gap penalties. I found that at true positive rate of 0.7, the best false positive rate achieved is 0.31 at best open and extension gap penalty scores at 1 and 5, respectively. ROC graph drawn using best gap penalty combination with BLOSUM50 matrix is shown below.



2. Using the best gap penalties combination, Blosum62 matrix performs the best as it has the lowest false positive rate at a true positive rate of 0.7. The false positive rates at a true positive rate of 0.7 for blosum50, blosum62, matio, pam100 and pam250 are 0.31, 0.30,0.38, 0.36 and 0.36 respectively. This information can also be observed from the graph below.



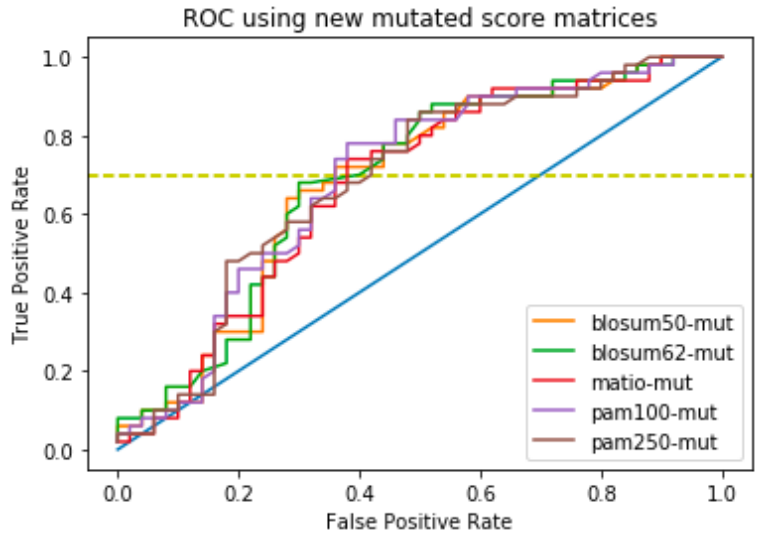
3. False positive rate performance is worse when the Smith-Waterman scores are normalized by the length of the shorter sequence in a pair. As seen in the ROC graph below, the best performing scoring matrix – BLOSUM62 outputs very high false positive rates when normalized. Local alignment scores show how similar two sequences are, which is independence of the length. When normalizing the score by dividing the similarity score by the shorter length of the sequence, the score just gets minimized and hence, poorer performance.



Optimization

I randomly switched 5 indices of the scoring matrices, giving a new score for those switched amino acids. When calculating the false positive rates using the new matrices, I found that it performs worse than the original matrix. To optimize, I should think more about the chemical

property of the amino acid instead of randomly switching and assigning new values. All the new matrices are stored in matrices_mut folder.



My optimization algorithm needs a lot of work. For the optimized matrix to be of general utility, the score output has to make sense biologically. For example, giving more score values to certain amino acids with higher active site functionality.