

Pseudo Code and Intermediate Step Snapshots

Read each aligned sequence into big list

Create dictionary with key for every possible character for counting

Create 2D dictionary with key for each possible character in each row and column

Use 2D dictionary to count number of occurrences of each amino acid pair

For each amino acid in all rows of list:

Increment corresponding value in count list, increment total

For each pair between all amino acids (n^2):

Increment corresponding value in 2D dictionary, increment total

For each item in list of counts calculate probability:

Value = value / total

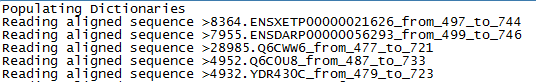
For each item in 2D dictionary calculate probability:

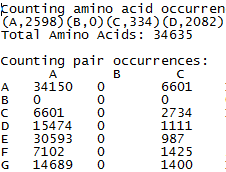
Value = value/total

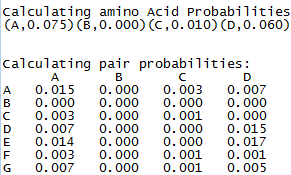
For each item in 2D dictionary calculate substitution score:

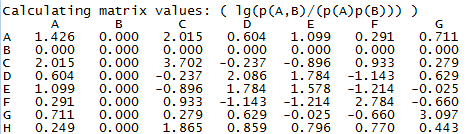
Value = lg(value/(probA \* probB))

Output matrix to file

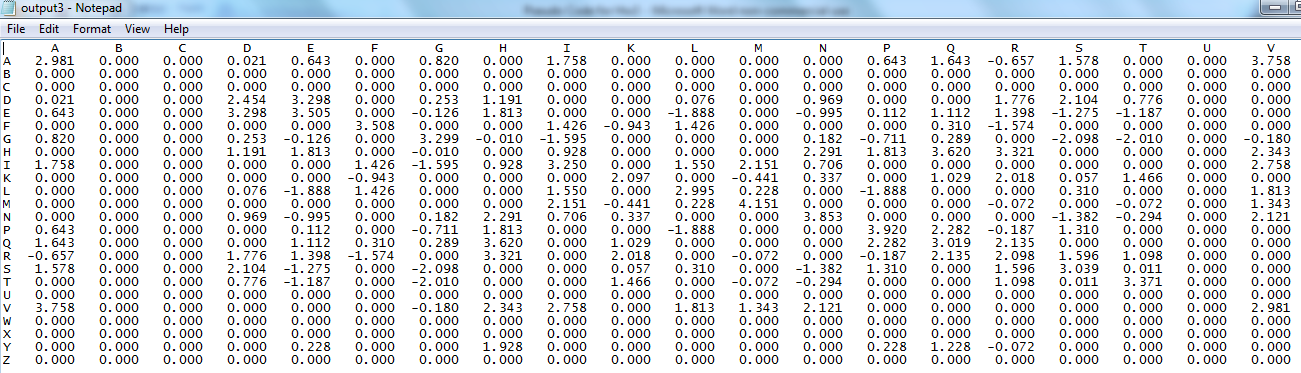
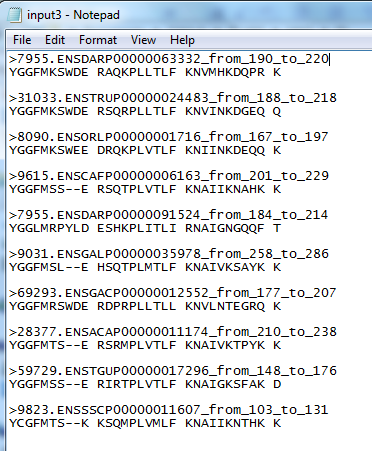


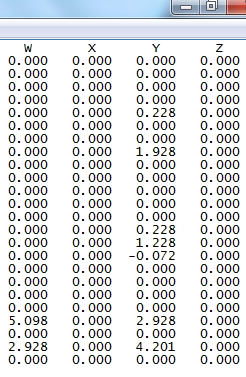




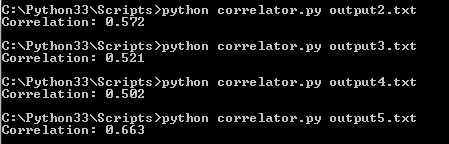


Example Input and Output Snapshots





Correlation between Our Matrix and BLOSUM Matrix



We believe that the correlation coefficient between our matrix and the BLOSUM matrix could be so low for three reasons. One, each matrix in the four different outputs were created using only one protein and the BLOSUM matrix was created using thousands of different proteins. Two, our matrix contains precise numbers whereas the BLOSUM matrix only contains whole numbers. Three, we used the BLOSUM62 matrix to compare to our matrix and the proteins we used might not have been 62% similar.