README

Carly Geissler

4/30/2021

Overview

This program outputs linguistic complexty for each sequence in a file of sequences. The linguistic complexity of a string is defined as the number of kmers that are observed for all possible k-mer lengths, divided by the total number that are theoretically possible. The code is broken down into functions to complete tasks.

countKmersPossible

This function counts possible kmers. Possible kmers include repeats. It takes a string and k as parameters and returns a count.

countKmersObserved

This function counts observed kmers. Observed kmers are unique and do not include repeats. The function takes a string and k as parameters and returns a count.

createDF

This function creates a dataframe for a string that includes k, possible, and observed kmers. The function takes a string as a parameter and returns two integers, totalO and totalP which are the sum totals of the possible and observed lists. It also returns a dataframe. The function works by calling countKmersObserved and countKmersPossible.

calcLinguistic

This function calculates linguistic complextity. It takes totalO and totalP as parameters and returns the linguistic complexity for these two integers.

main

The main function of this program opens and reads a csv file. Then for each sequence in the file an output file is opened and the dataframe for that sequence is written to the output file. Linguistic comlexity for each sequence is also calculated.