Kenneth Geiler 431388 Homework 3 Question 3

a.

1	2	3	4
acta	gaac	gaac	aagc
gaac	ctag	aagc	acta
aagc	tcat	tcat	ctag
ctag	ttct	gcgt	gaac
tcat	aagc	acta	gcgt
gcgt	gcgt	ctag	tcat
ttct	acta	ttct	ttct

b. the size k-mer substring can be calculated to be n-k+1. The time it takes to check if the k-mer substring is in the corpus takes time $\theta(m - k + 1) = \theta(m)$ + $\theta(n \log m) = \theta(m + nlogm)$ because the k and 1 are irrelevant as the function grows towards infinity, and the nlogm comes from a tree constructed of k-mer substrings with the worst case scenario being log m, which is then multiplied by the n elements resulting in n*log m.

c. The problem in b can be solved more efficiently by using radix sort on the k-mer substrings and the corpus. This results in time n-k+1 for the k-mer substring and m-k+1 to check the corpus string. Therefore the worst case time would be $\theta(m-k+1) + \theta(n-k+1) = \theta(m + n)$ due to the fact the k and 1s become irrelevant as the function grows towards infinity.