

## C'mon K-mers!

A useful operation on strings, specially when analyzing certain types of data, like genomic data, is to find all the substrings of length  $k$  in a string. These substrings are called the string's  $k$ -mers and we are typically interested in finding out how many times each  $k$ -mer appears within the string. For example, consider the following string:

aaabaab

This string has three unique 2-mers:

aaabaab

aa

aa

ab

ba

aa

ab

More specifically, the 2-mers are **aa** (appearing three times), **ab** (appearing two times), and **ba** (appearing once). Notice how counting the  $k$ -mers considers *overlapping* occurrences of each substring. For example, the first two occurrences of **aa** above overlap with each other.

We could do a similar analysis for the string's 3-mers:

aaabaab

aaa

aab

aba

baa

aab

In this case, there are four unique 3-mers: **aaa** (appearing once), **aab** (appearing twice), **aba** (appearing once), and **baa** (appearing once).

In this problem, you will write a function `kmers` that, given a potentially large string, a value for  $k$ , and some  $k$ -mers, will determine how many times each  $k$ -mer appears in the string. The function should return a list of integers where the  $i$ th integer is the number of times the  $i$ th mer in the list of mers occurs in the sequence.

Here are some sample inputs and outputs:

| Sequence | K-Mers   | Results |
|----------|----------|---------|
| aaabaab  | ab ba aa | 2 1 3   |

| Sequence | K-Mers              | Results   |
|----------|---------------------|-----------|
| aaabaab  | aaa aab aaa baa xyz | 1 2 1 1 0 |

Here is the skeleton code for this task:

```
import java.util.List;
import java.util.HashMap;

public class Problem4 {
    /**
     * kmers: Takes in a sequence of letters, a kmer length, and a list of k-mers and
     * returns a dictionary of k-mers and their counts in the sequence.
     *
     * Arguments:
     *   seq: A string of letters
     *   k: The length of the k-mers represented as an integer
     *   k-mers: An array of the k-mers of interest
     *
     * Returns: an array where the ith entry is the number
     *   of times that the ith k-mer in kmers occurs on the sequence
     */
    public static int[] kmers(String seq, int k, String[] mers) {
        // Return included to allow the skeleton code to compile
        return null;
    }
}
```

When you take the placement exam, you will be expected to copy the skeleton code into a file and then complete the function. For the practice problems, we have provided a file named Problem4.java that includes the above header for your convenience.