Extender Implementation

1 Introduction

Extender converts each of the 256 kmers with the smallest signatures into a 256-base wide DNA fragment, by extending such kmer left and right. To accomplish that, Extender reads the DNA fragments directly from one of the FM buffers, using the index it receives from the Sorter to calculate the FM row pointer.

2 Component Description

1. Calculate the fragment position on the genome sequence $frag\_idx = \left\lfloor \frac{(kmer\_idx - \lfloor \frac{frag\_len - kmer\_len}{2} \rfloor)}{frag\_len} \right\rfloor$.
2. Read the DNA fragment pointed to by $frag\_idx$.
3. Read the DNA fragment from the next FM buffer row, addressed by $frag\_idx + 1$.
4. Render the extended DNA fragment by concatenating the relevant parts of those two consecutive DNA fragments.

The 256 extended DNA fragments comprise the GFMs that become the output of ACMI.

An example of extension is shown in Figure 1. Suppose the kmer length is 4 (a 4-mer), the fragment length is 8, and the size of the genome is 32. The Extender has to extend the 4-mer with index 9 (i.e., the 4-mer TAAG marked in red in Figure 1). In this case,

$$frag\_idx = \left\lfloor \frac{(kmer\_idx - \frac{frag\_len - kmer\_len}{2})}{frag\_len} \right\rfloor = \left\lfloor \frac{(9 - \frac{8-4}{2})}{8} \right\rfloor = \left\lfloor \frac{7}{8} \right\rfloor = 0.$$

Hence the Extender will read from FM buffer fragments 0 and 1. It will then concatenate the relevant bases using shift operations, render a single 8-base wide DNA fragment, encode the bases using one-hot encoding, and append the encoded extended DNA fragment to the GFMs.

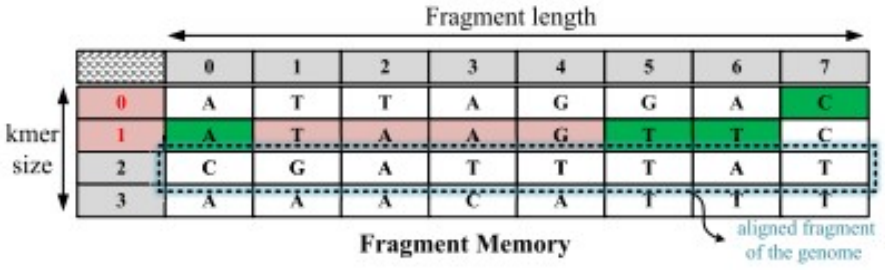

Figure 1: Example of extension. The fragment length is 8, the kmer size is 4. Each line represents an aligned fragment of the genome. The bases in pink are the bases of the 4mer at position 9. The extended fragment includes two bases (marked in green) to the left and to the right of the original 4mer.