

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.

One hot encoding of a fragment refers to the operation of transforming each base in the fragment as follows: $A \rightarrow [1, 0, 0, 0]$; $C \rightarrow [0, 1, 0, 0]$; $G \rightarrow [0, 0, 1, 0]$; $T \rightarrow [0, 0, 0, 1]$; $N \rightarrow [0, 0, 0, 0]$.

Translation of each base to hexadecimal format is as follows: $A \rightarrow 0x8$; $C \rightarrow 0x4$; $G \rightarrow 0x2$; $T \rightarrow 0x1$; $N \rightarrow 0x0$.

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.

In hexadecimal format, the extended DNA fragment can be represented as follows: Suppose the original 4-mer is "TAAG", with bases "T", "A", "A", "G". Using the encoding scheme, we have $T \rightarrow 0x1$, $A \rightarrow 0x8$, and $G \rightarrow 0x2$. If the bases extended to the left and right are "C" and "T" respectively, with $C \rightarrow 0x4$ and $T \rightarrow 0x1$, the extended fragment would be "CATAAGTT" and its hexadecimal representation would be 0x48188211.

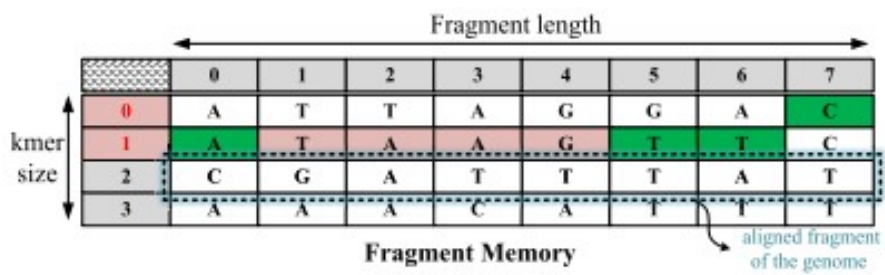


Figure 1: Example of extension. The fragment length is 8, the kmer size is 4. Each line represents an aligned fragment of the genome. The bases in pink are the bases of the 4mer at position 9. The extended fragment includes two bases (marked in green) to the left and to the right of the original 4mer.