Additional Tools

A few tools remain to be explained before we can jump into the algorithm. These tools have a variety of reasons for being used that will be explained individually.

k-mers

In order to create the sets that $\mathbf{Mm}\frac{1}{2}$ and **Mm** sketches are built with, we shall be using k-mer to partition each sequence into subsets. The k-mer of a sequence string are defined as follows:

The k-mer of a sequence string s is the set of all the substrings of size k of s.

The 1-mer of a sequence string, will therefore be the set of all characters in the sequence string. It is therefore sensible to consider the size of k when partitioning the sequence string.

MapReduce

Given a sizeable amount of sequences per file, it was quite essential to have a parallel and distributed programming model. For this purpose, MapReduce is a popular programming model. It works by distributing its task to a multitude of workers. Worker can be computers or cores. It expresses its computation as two functions:

1. Map: Runs a function over each element of a list and returns an intermediate value.
2. Reduce: Merges the intermediate values to form a potentially smaller set of values.

As explained in [**?**], MapReduce processes input by the following steps

1. Map step: Splits the input into M splits. Each split is then distributed to a worker who will perform a Map function on the given split and saves the result into a temporary storage.
2. Shuffle step: The results from the Map calls are then written to a local disk, partitioned into R regions.
3. Reduce step: For each region, a worker is set to run a Reduce job on it, in parallel.

MapReduce has been shown to scaler better than other parallel programming tools for input sizes that surpass 100 Mb, which is why it chosen.[**?**] Apache Hadoop MapReduce was used, as it is free source.

Pig

Pig Latin is a high level language for compiling and executing MapReduce jobs over Hadoop. Advantageous when performing a pipeline of MapReduce jobs[**?**], it also demands very few lines of code compared to Hadoop MapReduce code. Additionally, users may write User Defined Functions (abbr. UDF), completely eliminating the need to write any map or reduce function, as they are on the lower level.