HW4

Jacob Russell

Due: 28 Feb 2025

NOTE: This is a somewhat open-ended assignment. Read the questions carefully. To derive your answers, use your imagination, and state your assumptions.

A DNA strand can be represented as a (very long) string w over the alphabet $\{A, C, G, T\}$. For example, the human DNA has length $\approx 3 \times 10^9$. Because of the double-helix nature of DNA, we should really be talking about the *base pairs* A-T and G-C, in the sense that DNA is made of base-paired sequences: for example, instead of $w = ACTGGACT$, we could instead look at its *reverse complement* $\overline{w^R} = AGTCCAGT$, obtained by reversing w and then applying to it the "complement" homomorphism $A \to T$, $T \to A$, $C \to G$, $G \to C$.

To match DNA from a sample to a *reference* DNA w, or even to build *de novo* a reference DNA w, a **sequencer** can be used to generate a large number of relatively short substrings appearing in w (or in $\overline{w^R}$). The sequencer has no way to tell in which direction the piece of DNA is oriented when it reads it, called **reads**.

Sequencing technology is rapidly evolving, but let's assume for simplicity that it is possible to generate a large number (e.g., 10^9) reads of length 100 each, in a reasonable time (e.g., hours). In reality, the length of these reads may vary a little, sometimes we may have reads over $\{A, C, G, T, N\}$, where "N" indicates that the sequencer was not able to determine the exact value being read, and sometimes the sequencer may even misread a value; let's ignore these possibilities.

(a) What is the number μ of possible reads of length 100 over $\{A, C, G, T\}$?

Each time we need to pick a letter, we can pick 1 of 4 options. These are always replaced back to the selection pool, and we pick from these 4 options 100 times. This means that $\mu = 4^{100}$ possible reads of length 100.

(b) Assuming that the human reference DNA has length exactly equal to 3×10^9, what fraction of the μ possible reads is present in the human DNA?

To calculate this, we need to simply figure out how many 100 character long strings we can fit into the human reference DNA by dividing. However, because these strings need to stay at exactly 100 characters long, we cannot fit the string within any of the closing 99 character positions in the human reference DNA. So, our equation is $\frac{3 \times 10^9 - 100 + 1}{4^{100}}$.

(c) Describe how one could use an MDD to encode all the reads present in the human reference DNA, and then efficiently (question: how efficiently?) determine whether a sample read is present in the human reference DNA. (application: a CSI technician collects some genetic material at a crime scene and wants to determine whether it may be of human origin.)

The Multi-Dimensional Dictionary in this problem will act as a tree. Starting from the first character in the human genome, we will read in characters. Starting at an empty root node in our tree, we will create a node for the corresponding $\{A, C, G, T\}$ we just read. As we read in more characters, we continue inserting the nucleotides in order, creating new nodes for each if the sequence is not already present. Once a full 100 characters has been read, we mark the end of the sequence with an end node and then repeat the steps for the next 100 character string. As we process new sequences, we reuse previously inserted prefixes to avoid redundant storage and ensure efficient lookup. Because our reads are always 100 characters long, our look-up times will be O(100) as it will always take the same number of character reads to check if a given 100-character string is in the DNA genome. We can search by starting at the root. When we read a nucleotide in a sequence, we traverse to the corresponding child node. Continuing through the entire 100 character string, if the node does not exist, the read is not in the DNA. If we reach an end node it is present.