ComS 331 Spring 2024 Name: Aren Ashlock

1. 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 $\approx 3 \cdot 10^9$. Because of the double-helix nature of DNA, we really should 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) called *reads*. Sequencing technology is rapidly evolving, but let's assume for simplicity that it is possible to generate large number (e.g., 10^8 or 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 $\{A, C, G, T\}$?
   Choice from 4 options (with replacement) 100 times $\to 4^{100} \approx 1.61 \times 10^{60}$.

   (b) Assuming that the human reference DNA has length exactly equal to $3 \cdot 10^9$, what fraction of the μ possible reads is present in the human DNA?
   We need to find the total number of reads in the reference DNA $\to 3 \cdot 10^9/100 = 3 \cdot 10^7$. Then, we take that and find the percentage by dividing by the total number μ of possible reads $\to 3 \cdot 10^7/4^{100} \approx 1.87 \times 10^{-51}\%$.

   (c) Describe how one could use an MDD to encode all the reads present in the human reference DNA, and then efficiently determine if 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).
   You can **encode** all the reads using an MDD by doing the following: Feed each human reference DNA read into the MDD. Start with a set that contains the first read letter, then have it point to another set that contains all the letters read after it (adding the letter if the set doesn't already contain it). Continue that step until the read is complete. This results in a "tree" with the "leaves" each containing a set of final letters that essentially indicate if the read is in the human reference DNA or not (whether the read makes it to the end).
   In order to **efficiently determine** if a sample read is present in the human reference DNA, you just use the MDD to trace through the pointed sets and see if the letter that is read is present. So, the first set would be the starting point. If the letter that was read first is in the set, look at the set that the letter is pointing to and repeat. If the read makes it all the way to the end following this pattern, it is present in the human reference DNA. At any point, if the letter is not in the set that's currently being pointed to, then the read is not in the human reference DNA.

   (d) IDEA FOR A POSSIBLE PROJECT (AFTER THE COURSE IS OVER) Implement an efficient algorithm that takes in input a (very large) set of reads and builds the MDD encoding them.