Challenge 1 - High density storage in DNA

ATGCXYWZ

Background: In lecture, we saw the development of a few different types of encoding schemes that can be used to convert binary information into nucleotides (and back). The density of information stored depends on the encoding scheme chosen as well as any constraints to reading and writing. Given constraints, we find that the actual information density is lower than what one would calculate from assuming a uniform probability since not all sequences are possible.

Goal: Approximate how much information can be stored, per nucleotide, in a DNA encoding system that uses 8 letters: A, T, G, C, X, Y, W, Z with the following constraints:

- Avoid A, T, G, C homopolymers of 4 in a row (AAAA, TTTT, GGGG, CCCC)
- Avoid W, Z homopolymers of 3 in a row (WWW, ZZZ)
- Avoid X, Y, homopolymers of 2 in a row (XX, YY)

If you are assuming any other constraint, or architecture for information storage, list it as an additional assumption. Provide an answer in units of bits/nt with reasonable precision (hundredths place: e.g. 2.02 bits/nt). *Hint: The simplest solution to this problem will likely involve simulating possible sequences.*

Note: This question is asking for raw information stored per nucleotide, and does not need to consider error correction or fountain code solutions for robustness.

Note: To receive credit, you will need to provide proof that you did this work (and not copied the answer from another individual). The format of this proof can be a writeup of how you went about solving this problem alongside code you generated.