Section 1:
The length of a unique sequence in a genome

Nico Gort-Freitas
MCB
Harvard University

9/9/2022

Genome sizes range 1 million to 100 billion bp

10^6 10^{10}

Species	*T2 phage*	*Escherichia coli*	*Drosophila melanogaster*	*Homo sapiens*	*Paris japonica*
Genome Size	170,000 bp	4.6 million bp	130 million bp	3.2 billion bp	150 billion bp
Common Name	 Virus	 Bacteria	 Fruit fly	 Human	 Canopy Plant

For this section, we'll work with an equiprobable genome

We will assume

Each base is equally probable at every position of the genome

$$p_A = p_C = p_G = p_T = \frac{1}{4}$$

The probability of observing each base at any position is independent

$$P(\textbf{motif}) = \prod_{i=1}^{l} P(s_i = m_i)$$

$$P(\textbf{motif}) = P(s_1 = m_1) \times p(s_2 = m_2) \times \ldots p(s_l = m_l)$$

Motivating questions

- How rare is a motif?
- How many times would we observe it?
- Where would be observe it most often?
- How long is the shortest unique motif?

Which of these sequences would you expect to be rarer?

ACATCC

AAAAAA

In an equiprobable genome, all l-long motifs are equally frequent

$$P(\mathbf{AAAAAA}) = P(s_1 = A) \times p(s_2 = A) \times P(s_3 = A) \times p(s_4 = A) \times P(s_5 = A) \times p(s_6 = A)$$

$$P(\mathbf{AAAAAA}) = \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4}$$

$$P(\mathbf{AAAAAA}) = \left(\frac{1}{4}\right)^6$$

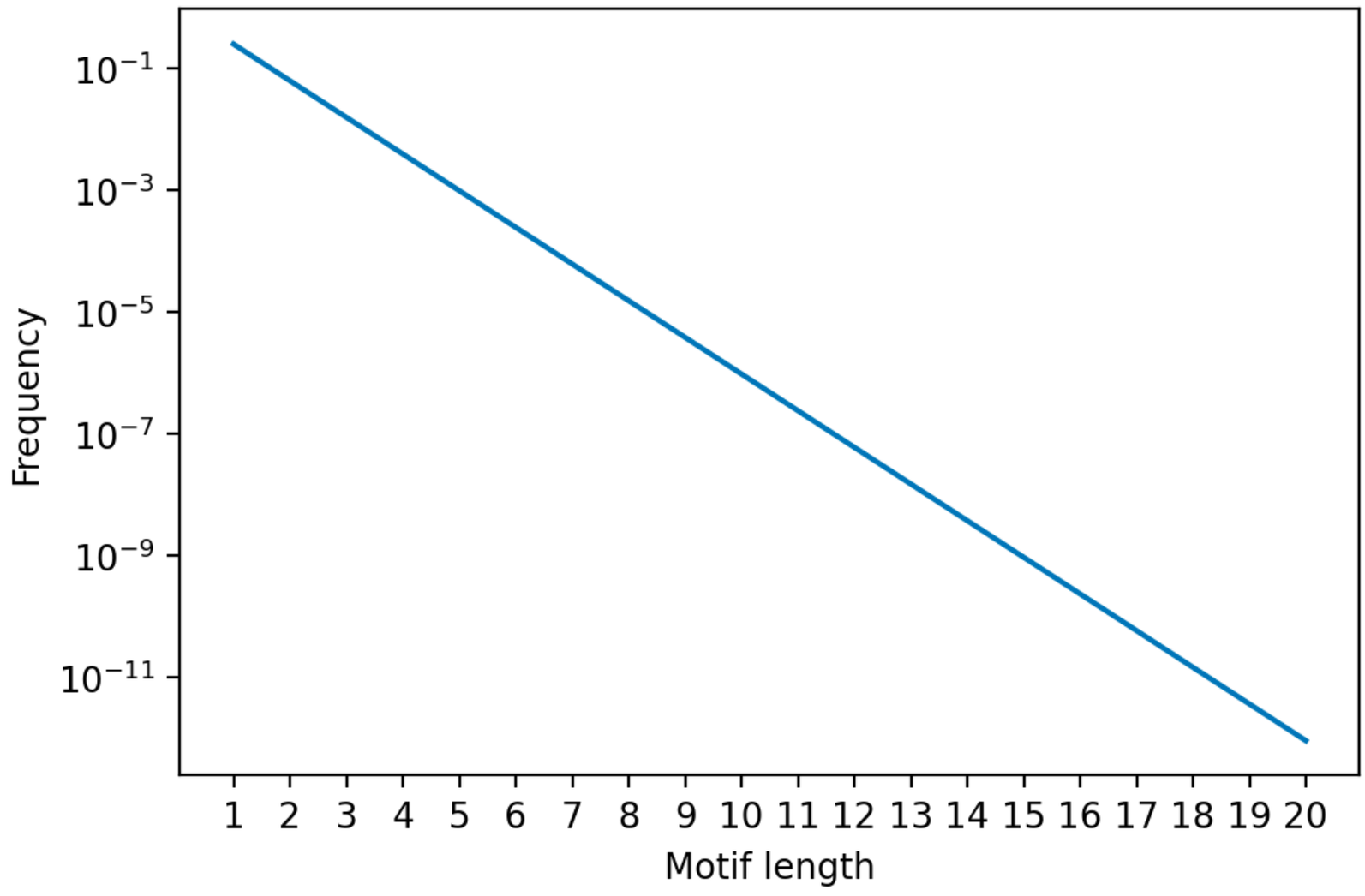
$$P(\mathbf{ACATCC}) = P(s_1 = A) \times p(s_2 = C) \times P(s_3 = A) \times p(s_4 = T) \times P(s_5 = C) \times p(s_6 = C)$$

$$P(\mathbf{ACATCC}) = \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4}$$

$$P(\mathbf{ACATCC}) = \left(\frac{1}{4}\right)^6$$

The probability of a motif depends on its length alone

$$P(\textbf{motif}) = \left(\frac{1}{4}\right)^{l}$$



The probability of a motif depends on its length alone

$$P(\textbf{motif}) = \left(\frac{1}{4}\right)^{l}$$

How many times do we expect to see a motif with such probability?

$$\mathbb{E}[n_{motif}] = L \times P(\textbf{motif}) = L \cdot \left(\frac{1}{4}\right)^{l}$$

$$\mathbb{E}[n_{motif}] = \frac{L}{4^{l}}$$

At what motif length do we expect a motif to be, at most, unique?

$$L \times \frac{1}{4^l} \leq 1$$

$$L \leq 4^l$$

$$\log_2(L) \leq \log_2(4^l) = 2l$$

$$l \geq \frac{1}{2}\log_2(L)$$

Now let's revisit this question through information theory

Information content of a single base

$$I(X = x_0) = \log_2\left(\frac{1}{P(X = x_0)}\right)$$

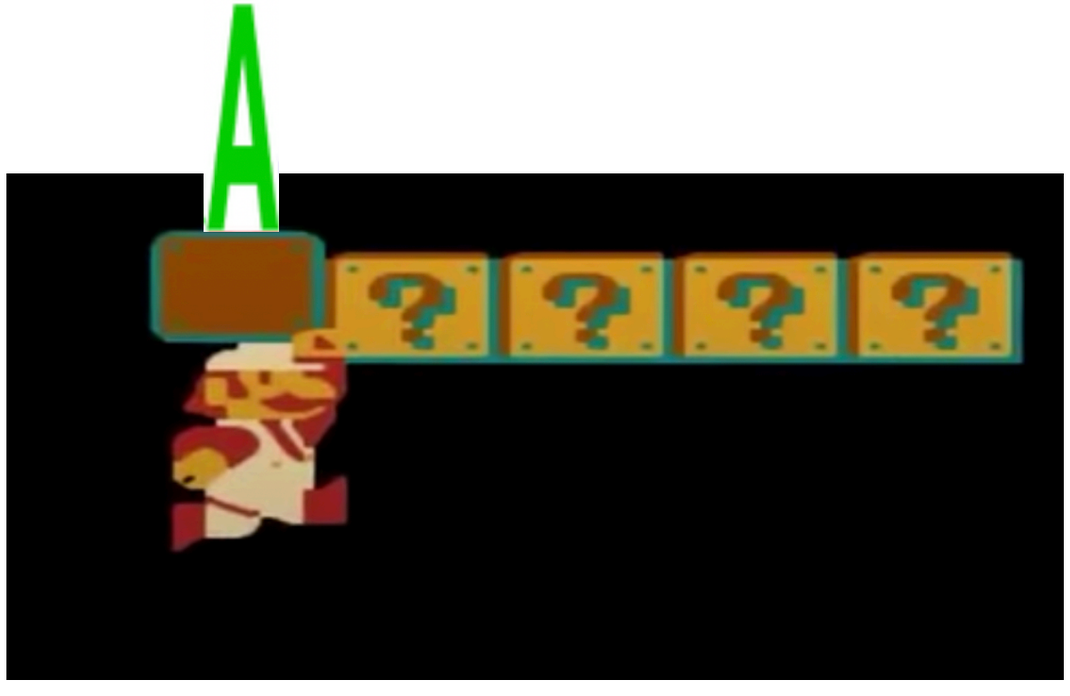
If we consider the probability of observing the first base of a motif:

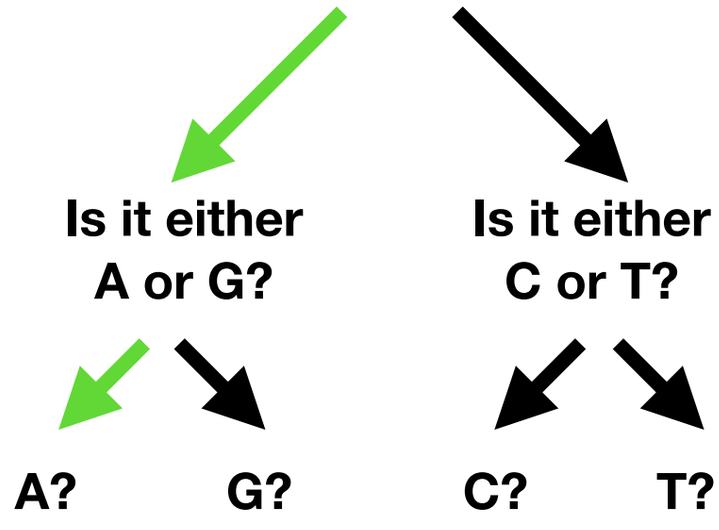
$$I(m_1) = \log_2\left(\frac{1}{P(s_1 = m_1)}\right)$$

$$I(m_1) = \log_2(\frac{1}{\frac{1}{4}}) = \log_2(4)$$

$$I(m_1) = \log_2(4) = 2$$

Bits of information can be understood as answers to yes/no questions





Learning each base answers two yes/no questions i.e. we learn 2 bits per position

Information content of a motif

$$I(m_1) = \log_2(4) = 2$$

$$I(m_{1,2}) = I(m_1) + I(m_2) = 2\log_2(4) = 4$$

$$I(\textbf{motif}) = \sum_{i=1}^{l} I(m_i) = \sum_{i=1}^{l} \log_2(4)$$

$$I(\textbf{motif}) = l\log_2(4) = 2l$$

What's the information content of a motif's position?

What's the information of a motif's position?

$$I(X = x_0) = \log_2\left(\frac{1}{P(X = x_0)}\right)$$

$$I(C = c_0) = \log_2\left(\frac{1}{P(C = c_0)}\right)$$

Since we assume every position is equally likely*:

$$I(\textit{position of a motif}) = \log_2\left(\frac{1}{1/L}\right) = \log_2(L)$$

***(and ignore the edges)**

Entropy in equiprobable random variables

$$\mathrm{H}(X) = -\sum_{i=1}^{n} I(X = x_i)$$

$$\mathrm{H}(X) = -\sum_{i=1}^{n} P\left(X = x_i\right) \log_2 P\left(X = x_i\right)$$

If every outcome is equally likely:

$$\mathrm{H}(X) = I(X = x_0) = \log_2 P(x_0)$$

Uniqueness in an equiprobable genome

$$H(\textit{presence of a motif}) = I(\textit{any } l\textit{-long motif}) = 2l$$

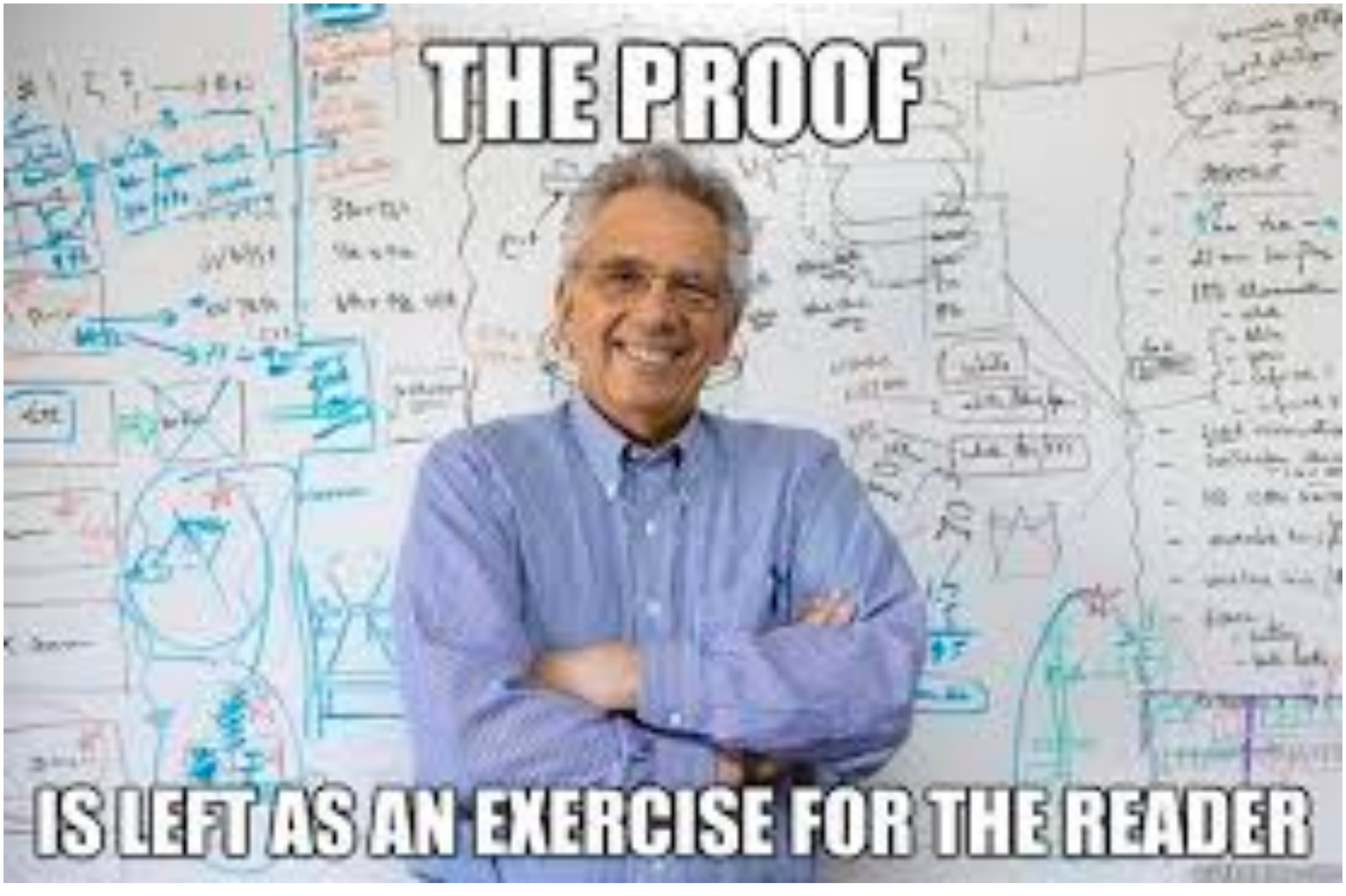
$$H(\textit{position of a motif}) = I(\textit{any position}) = \log_2(L)$$

We can prove that for a sequence to be unique, it must hold that:

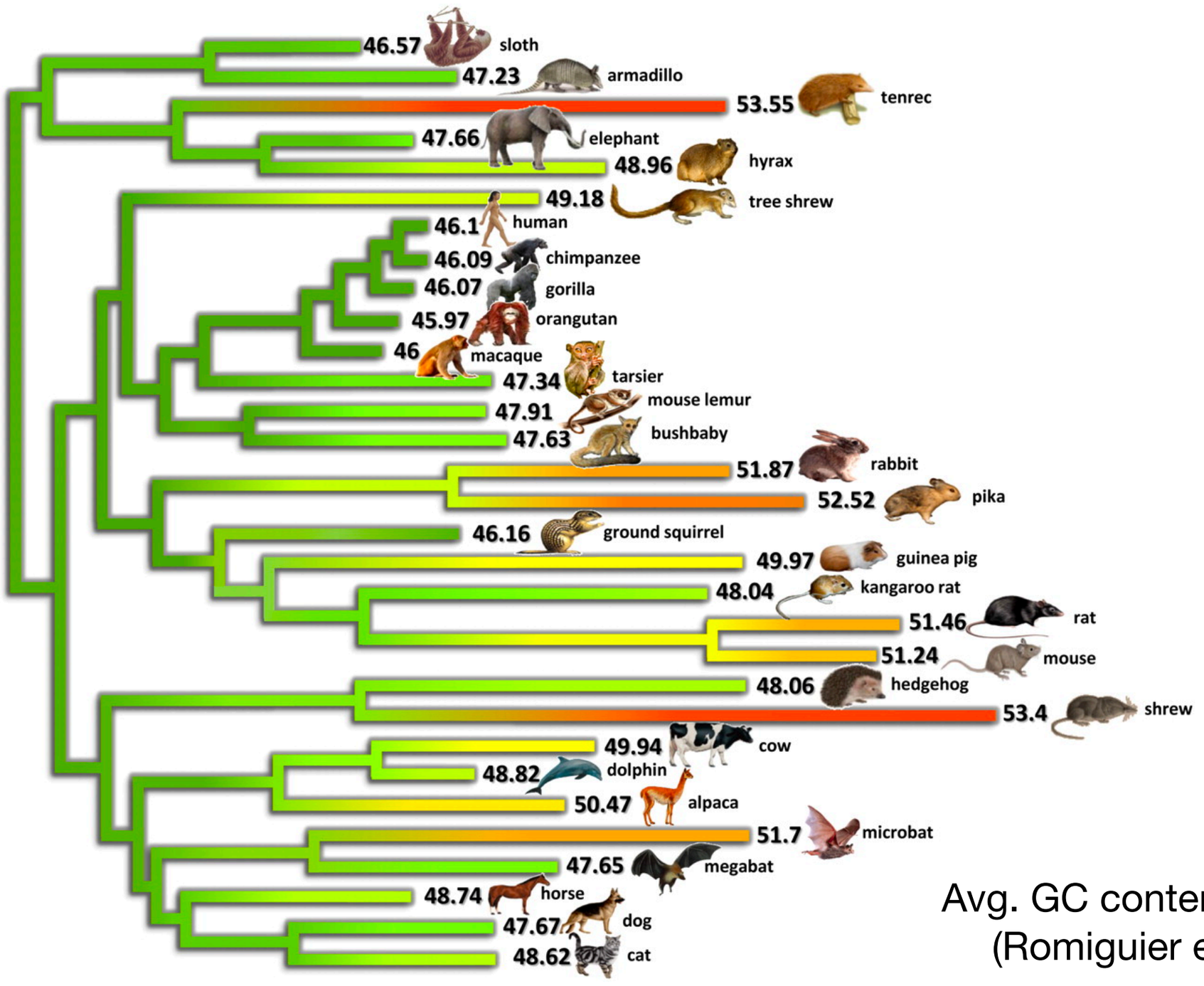
$$H(\textit{presence of a motif}) \geq H(\textit{position of a motif})$$

Hence, in our equiprobable genome:

$$l \geq \frac{1}{2}\log_2(L)$$


THE PROOF
IS LEFT AS AN EXERCISE FOR THE READER

Using entropy allows us to answer the uniqueness question for non-equiprobable genomes



Avg. GC content per species
(Romiguier et al., 2010)

Nobody:

Math teachers at the first day of school year: