

1. Assume that the length of the human genome is 3×10^9 base pairs, and that each of the 4 base pairs occurs with probability $1/4$.
 - (a) How long in base pairs does a motif have to be to occur approximately once per genome? A fractional result is fine.
 - (b) Suppose a motif occurs once on average in the genome. You can model this as a binomial distribution with 3×10^9 attempts and a success rate of p per attempt. What is p ?
 - (c) The binomial distribution reduces to a Poisson distribution. From the Poisson distribution for a motif with λ occurrences on average per genome, what is the probability of exactly k occurrences? This question is really just asking you to write down the Poisson distribution.

2. Coding length and information content. The Shannon entropy for a discrete random variable with n states $i \in \{1, 2, \dots, n\}$ is $-\sum_{i=1}^n p_i \log_2(p_i)$, where p_i is the probability of state i and the entropy is in bits.
 - (a) If all 4 base pairs are equally likely, what is the Shannon entropy in bits for a specific position in the genome?
 - (b) Suppose that x_1, x_2, \dots, x_T are T total random variables that are independent and identically distributed. Each random variable considered on its own has entropy H . Prove that the joint distribution has entropy $T \times H$.
 - (c) Suppose an alien form of life has 6 possible base pairs instead of 4. In fact, Steve Benner is working on creating unnatural nucleotides that would increase the coding capacity of the genome. What is the entropy per position in bits? What genome size would have the same entropy as the human genome?
 - (d) The malaria parasite, *Plasmodium falciparum*, has a 23 megabase genome that is AT-rich: AT and TA base pairs have about 40% frequency, whereas CG and GC have 10% frequency. How many bits of information are encoded at each position? What genome size would have the same entropy if the 4 base pairs had equal frequency?

3. Related problems.
 - (a) The human genome has about 20,000 protein-coding genes. One method used in the 1990's to analyze expressed genes was to sequence the 3' terminus of a transcript immediately upstream of the poly-A tail, termed an expressed sequence tag (EST). Assuming that each nucleotide in a transcript is equally likely, how long must a tag be to occur once on average among the 3' ends of 20,000 genes?

- (b) Suppose that the human population is 10 billion, and each gene has 10 equally likely variants. How many genes must be examined to identify a person by giving a pattern that occurs on average once among humans? What fraction of the genome is this?
4. In class we showed that the probability distribution for a geometrically distributed random variable x is $p_n = (1 - \theta)\theta^n$ for $n \in \{0, 1, 2, \dots, \infty\}$. Calculate $\tilde{p}(s)$, $\langle n \rangle$, and $\langle n^2 \rangle - \langle n \rangle^2$. For this question, the Laplace transform for a discrete domain is

$$\mathcal{L}[p_n] = \tilde{p}(s) \equiv \sum_{n=0}^{\infty} \exp(-sn)p_n.$$

Hint: the result for $\tilde{p}(s)$ is an infinite series that you should sum to get a compact closed-form expression. Since p_n is normalized, you can test your result by checking that $\tilde{p}(s) = 1$ when $s = 0$. Then you can calculate the moments as $(-d/ds) \ln \tilde{p}(s)$ and $(-d/ds)^2 \ln \tilde{p}(s)$.