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.

   n =number of chance occurrences of a motif per genome $= 1$

   G =genome size in base pairs $= 3 \times 10^9$

   p =probability of observing motif of interest among all motifs of same length =?

   L =motif length

   $$n = G \times p$$

   $$p = \frac{1}{4^L} \text{ (assuming equal nucleotide probability)}$$

   Solve

   $$p = \frac{1}{3 \times 10^9}$$

   $$L = \frac{1}{2} log_2 \left(3 \times 10^9\right) = 15.74$$

   Motifs of length 15, and 16 will have an average occurrence of 2.79 and 0.70 per genome, respectively.

   (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? Modeling this process as binomial distributions assumes that we have 3×10^9 attempts; i.e. 3×10^9 observations of motifs of length L. The parameter p is the probability of success in each observation, which is the probability of a random motif of length L being equal to our motif of interest. Therefore,

   $$p = \frac{1}{4^L}$$

   (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.

   n =number of occurrences of a motif per genome

   $$n \sim \text{Poisson}(\lambda)$$

   $$Pr\left(n = k\right) = \frac{e^{-\lambda} \lambda^k}{k!}$$

2. Coding length and information content. The Shannon entropy for a discrete random variable with n states $i \in \{1,2,\ldots n\}$ 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?

   $H =$Shannon entropy per position

   $$H = -\sum_{i=1}^{n} p_i \log_2(p_i) = -4\left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) = 2 \text{ bits}$$

   Notice that this is also $\log_2 4$. The entropy of a uniform discrete random variable over an alphabet of size n is $\log_2 n$

   (b) Suppose that $x_1, x_2, \ldots, x_T$ 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 $T \times H$.
   Under independence assumption, we can write:

   $$p(x_1, x_2, \ldots, x_T) = p(x_1) \times p(x_2) \times \cdots \times p(x_T)$$

   Thus, we can simplify the definition of entropy as:

   $$\begin{aligned}
   H(x_1, x_2, \ldots, x_T) = & \sum_{x_1, x_2, \ldots, x_T \in X} -p(x_1, x_2, \ldots, x_T)\, log_2 p(x_1, x_2, \ldots, x_T) \\
   = & \sum_{x_1, x_2, \ldots, x_T \in X} (-1) p(x_1) \ldots p(x_T) log_2 (p(x_1) \times \cdots \times p(x_T)) \\
   = & \sum_{x_1, x_2, \ldots, x_T \in X} (-1) p(x_1) \ldots p(x_T) (log_2 (p(x_1)) + \cdots + log_2 (p(x_T))) \\
   = & \sum_{x_2, \ldots, x_T \in X} p(x_2) \ldots p(x_T) \sum_{x_1 \in X} p(x_1) log_2 (p(x_1)) \\
   & + \ldots \\
   & + \sum_{x_1, \ldots, x_{(T-1)} \in X} p(x_1) \ldots p(x_{(T-1)}) \sum_{x_T \in X} p(x_T) log_2 (p(x_T)) \\
   = & H(x_1) \sum_{x_2, \ldots, x_T \in X} p(x_2) \ldots p(x_T) + \cdots + \\
   & H(x_T) \sum_{x_1, \ldots, x_{(T-1)} \in X} p(x_1) \ldots p(x_{(T-1)})
   \end{aligned}$$

   Also, for any number of random variables x_i, we know that:

   $$\sum_{x_i, \ldots, x_j \in X} p(x_i) \ldots p(x_j) = \sum_{x_i \in X} p(x_i) \left(\sum_{x_{(i+1)} \in X} p(x_{(i+1)}) \left(\ldots \left(\sum_{x_j \in X} p(x_j) \right) \right) \right) = 1$$

Therefore,

$$H(x_1, x_2, \ldots, x_T) = \sum_{i=1}^{T} H(X_i)$$

And, because x_i are all identically distributed, $H(x_i) = H$ for all i, and

$$H(x_1, x_2, \ldots, x_T) = T \times H$$

(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?

G_j =genome size in bp for organism j
H_j =per position entropy in genome of organism j
I_j =genome entropy of organism j

$$I_j = G_j \times H_j$$

For the alien with 6 equally probable bases:

$$H_{alien} = -\sum_{i=1}^{n} p_i \log_2(p_i) = -6\frac{1}{6}\log_2\left(\frac{1}{6}\right) = \log_2 6 = 2.585 \text{ bits}$$

In order to have the same genome entropy as human genome, we should have:

$$G_{human} \times H_{human} = G_{alien} \times H_{alien}$$

$$G_{alien} = \text{genome size of alien in bp} = \frac{G_{human} \times H_{human}}{H_{alien}}$$

$$G_{alien} = \frac{2 \times 3 \times 10^9}{2.585} = 2,321,116,843 \text{ bp}$$

$$\approx 2.3 \times 10^9 \text{bp}$$

(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?
The subscript pf indicates *Plasmodium falciparum*.

$$H_{pf} = -\sum_{i=1}^{n} p_i \log_2(p_i) = -2 \times 0.4 \times \log_2(0.4) - 2 \times 0.1 \times log_2(0.1) = 1.722 \text{ bits}$$

If the 4 bases had equal probability, the entropy per position would be:

$$H_{pf}^{eq} = -\sum_{i=1}^{n} p_i \log_2(p_i) = -4 \times \frac{1}{4} \log_2(4) = 2 \text{ bits}$$

In that case, the same genome entropy can be encoded using:

$$\frac{G_{pf} \times H_{pf}}{H_{pf}^{eq}} = \frac{23 \times 10^6 \times 1.722}{2} \approx 19.8 \text{ megabase}$$

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?

$$N_g = \text{the number of genes in human genome } = 20,000$$
$$p = \text{probability of occurrence of ETS per gene}$$
$$n = \text{the average number of occurrences of ETS}$$

$$n = N_g p$$

Assuming that all tags are of the same length L, the probability of observing a given tag being equal to the ETS of a sample gene is equivalent to the probability of a random sequence of length L being equal to a given sequence of same length. Therefore:

$$p = \frac{1}{4^L}$$

and we want:

$$n = N_g p = 20,000 \frac{1}{4^L} \approx 1$$

Which results in $L = 7.14$. This indicates that tags of length 7 will on average be observed in more than 1 gene, and tags of length 8 will on average appear in less than 1 gene.

   (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? If N is the number of genes, and p is the probability of a specific variant, we want

$$p^N \times \#\text{ people} = 1$$
$$\left(\frac{1}{10}\right)^N \times 10^{10} = 1$$
$$N = 10$$

4. In class we showed that the probability distribution for a geometrically distributed random variable x is $p_n = (1-\theta)\theta^n$ for $n \in \{0,1,2,\ldots,\infty\}$. 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

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

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 $(-d/ds)^2\ln\tilde{p}(s)$.

$$p_n = \theta^n(1-\theta)$$

We can find the generating function $\tilde{p}(s)$ as,

$$\begin{aligned}\tilde{p}(s) &= \sum_{n=0}^{\infty} p_n e^{-sn} \\ &= \sum_{n=0}^{\infty} \theta^n(1-\theta)e^{-sn} \\ &= (1-\theta)\sum_{n=0}^{\infty}(\theta e^{-s})^n \\ &= (1-\theta)\frac{1}{1-\theta e^{-s}}\end{aligned}$$

In order to find $\langle x\rangle$ and $\langle x^2\rangle - \langle x\rangle^2$, we need to take first and second derivative of log of the generating function with respect to s.

$$ln\,(\tilde{p}(s)) = ln(1-\theta) - ln(1-\theta e^{-s})$$

The mean of the random variable can be found using:

$$\langle x\rangle = -\frac{d}{ds} ln\,(\tilde{p}(s))\,|_{s=0}$$
$$\langle x\rangle = \frac{\theta e^{-s}}{1-\theta e^{-s}}|_{s=0} = \frac{\theta}{1-\theta}$$

The variance of the random variable can be found similarly by:

$$\begin{aligned}\langle x^2\rangle - \langle x\rangle^2 &= \frac{d^2}{ds^s}\left(ln\left(\tilde{p}(s)\right)\right)|_{s=0} \\ &= \frac{\theta e^{-s}}{\left(1-\theta e^{-s}\right)^2}|_{s=0} \\ &= \frac{\theta}{\left(1-\theta\right)^2}\end{aligned}$$