

Problem Set # 1: Due Monday, Sept. 26.

These are open book problem sets: any library or web resource is useable, but please **reference** those you use. I also don't mind if you talk with other students in a general way, but I want each group to work the problems out themselves.

1. Homeland Security. Professor Burns, fearing his flight will be sabotaged with a bomb, decides to bring a bomb on board himself. "The probability of a bomb being brought on board is small, but the probability of TWO being brought on board is so much smaller!" What is the error in his logic?

2. Testing a Statistical Hypothesis. Professor Burns got a ticket twelve times for illegal overnight parking. All twelve tickets were given either Tuesdays or Thursdays. Find the probability of this event. (Was his renting a garage for only Tuesdays and Thursdays justified?)

This is a famous exercise (up to the name of the character) from the most famous of probability texts – William Feller's "An Introduction to Probability Theory and Its Applications". The point of the problem is for you to say what you need to solve this problem! You have to construct a probability model of what is happening, and use that to answer my dilemma.

3. Independent acids model. (This is ex. 11.2 in Durbin, p. 307.)

Assume a model in which $p_i(a)$ is the probability of amino acid a in position i in a sequence of length l . The various amino acids are considered independent of one another. What is the probability $P(x)$ of any given sequence $x = x_1x_2\dots x_l$? Show that the average (expectation) of the log of the probability is the negative entropy $\sum P(x) \log P(x)$, where the sum is over all sequences x of length l . [Note that the length of the sequences is fixed at l .]

4. Entropy. In the previous example, substitute a nucleic acid model and assume that each of the four nucleotides is equally likely. What is the entropy of a sequence of length n ? Be explicit about your assumptions. Use logs base 2, so the answer is in bits.

5. Simple Entropy Calculation. What is the relative entropy of the two distributions on one coin flip, $P = (p, 1 - p)$, and $Q = (1 - p, p)$? What happens when p gets very small? What is the relative entropy of $P_0 = (0, 1)$ and $Q_0 = (1, 0)$?

6. Acceptor Sites and Entropy. Continue the exercise begun in lecture: interpret the two graphs at the top of the data sheet available at the class web site:

<http://www.math.lsa.umich.edu/dburns/547/entropy.pdf>

In particular: (1) be careful to read the description to see what the significance of the plot value over the position. Once that is clear, (2) interpret the period three behavior of the two graphs downstream of the acceptor site.

7. An Example From the Literature. In *Molecular Biology and Evolution* v. 22, no. 5 (May, 2005), available at

<http://mbe.oxfordjournals.org/cgi/content/full/22/5/1260>,

Graur and coworkers study the so-called *isochore theory*, a hypothesis about the CG bias in composition of nucleic acid sequence. This is outlined clearly in the first few paragraphs of the paper. In the first paragraph on “methods”, the authors introduce an entropy measure for a partition of a chromosome of length L into two subsequences of lengths ℓ_i (these are contiguous strings of nucleotides). Let F_{AT} be the likelihood of an A or a T in the whole chromosome, and similarly for $F_{CG} = 1 - F_{AT}$. Similarly, we have f_{AT}^i and $f_{CG}^i = 1 - f_{AT}^i$, for the distributions for the two partition pieces. The Jensen-Shannon entropic divergence is given as

$$D_{JS} = \max[H^{tot} - \sum \frac{\ell_i}{L} H^i].$$

where $H^{tot} = -F_{AT} \log F_{AT} - F_{CG} \log F_{CG}$, and similarly for the distributions for the two pieces. What is the expression $H^{tot} - \sum \frac{\ell_i}{L} H^i$ measuring? What does it mean to maximize this expression? Interpret this measure in terms of the different kinds of entropy we have already discussed.

8. Counting.

Consider two sequences, $\mathbf{a} = a_1 \dots a_n$ and $\mathbf{b} = b_1 \dots b_m$, where $1 \leq n \leq m$.

- (a) How many alignments are there of length m ?
- (b) How many alignments are there of length $m + 2$?
- (c) How many alignments *in general* are there between \mathbf{a} and \mathbf{b} ?

[This does not have a closed formula. Just make an attempt. For a hint on the general case, see p. 19 of Durbin (available in the Ctools site). It is important to note here that for purposes of sequence alignment, one does not allow aligning a gap against a gap, and that one *does not distinguish* between cases such as

$$\begin{array}{cccc} \dots & a_i & - & \dots \\ \dots & - & b_j & \dots \end{array}$$

and

$$\begin{array}{cccc} \dots & - & a_i & \dots \\ \dots & b_j & - & \dots \end{array}$$

which is justified biologically, and makes the count easier. You should also try problem 2.7 of Durbin (available in the Ctools site) to get a feel for how large these numbers are!]