

## Problem Set # 1: Due Wednesday, Feb. 5.

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. 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  $l$ .]

### 2. Gap penalties as probabilities. (This is ex. 2.2 in Durbin, p. 17.)

Show that the probability distributions  $f(g)$  that correspond to the linear and affine gap schemes given in equations (2.4) and (2.5) [of Durbin] are both geometric distributions, of the form  $f(g) = ke^{-\lambda g}$ .

[The *geometric distribution* represents the *waiting time* until the first head in a sequence (of *indefinite length*) of independent coin flips. If the bias of the coin is  $p$ , then the probability that the first head will come on the  $n$ -th flip is  $p(1-p)^{n-1}$ . Thus the linear and affine gap penalties model gaps as continuing to a random length, the randomness treated as being an independent trial, or coin flip, at each new alignment position. In these terms, what is the "bias" of the coin being "flipped", if the gap penalty is linear,  $\gamma(g) = -dg$ ?

### 3. Another model.

Consider Durbin, p. 17, the second paragraph. Describe a model which takes into account a difference between residue distributions in gapped and ungapped regions. What kind of data would you need to set up a dynamic programming method of comparison?

### 4. 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  $\mathbf{b}$ ?

[For a hint on the general case, see p. 19 of Durbin. 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 to get a feel for how large these numbers are!]

### 5. Scoring an example by hand.

Consider two DNA sequence fragments,  $\mathbf{a} = CAGTATCGCA$ , and  $\mathbf{b} = AAGTTAGCAG$ .

(a) Compute an optimal global alignment between the sequences, using a scoring function which gives +1 for a match, -1 for a mismatch, and has a linear gap penalty with  $d = 2$ . Try  $d = 1$  and  $d = 0$ , also.

(b) Compute an optimal local alignment with the same scoring parameters.

(c) How many solutions are there to (a) and (b)?

(d) Go to <http://www-hto.usc.edu/software/seqaln/seqaln-query.html> on the web, and follow the directions there to check your calculations above. [You will not get a trace-back matrix as part of the output, but a particular choice of the trace-back.]