Math/Stat 547, Winter 2004: Biological Sequence Analysis

Problem Set # 1: Solution Remarks.

Probs. 1-3. didn't cause many problems.

Prob. 4. The argument is that $Prob(\text{sequence} = x_1 x_2 \ldots x_g) = \Pi_{j=1}^{g} P(x_j)$, and the probability of the completely gap alignment is $f(g) \times \Pi_{j=1}^{g} P(x_j)$. So the log of the ratio of probabilities is just

$$\log \frac{f(g) \times \Pi_{j=1}^{g} P(x_j)}{\Pi_{j=1}^{g} P(x_j)} = \log f(g).$$

You can solve in the linear case (affine is similar) $f(g) = e^{-gd} = \{e^{-d}\}^g$, which looks like a geometric distribution, but doesn't model the length of the sequence (doesn't explicitly terminate the sequence).

Prob. 5. This is the counting one, which had problems n the formulation. Part a. didn't pose any problem. For parts b. and c., we can define a *trace* of an alignment as the class of the alignments which only differ as in the statement of the homework problem, that is, we can change an alignment by a succession of replacements of the form:

$$\begin{pmatrix} \ldots & a_i & - & \ldots \\ \ldots & - & b_j & \ldots \end{pmatrix}$$

is replaced by

$$\begin{pmatrix} \ldots & - & a_i & \ldots \\ \ldots & b_j & - & \ldots \end{pmatrix}.$$

For every $k = 0, \ldots, n$, let's count the number of traces of length $m + k$ between a and b: as in part a., for $k = 0$, we get $\binom{m}{n}$. For a general k, we take $m + k$ positions and then choose k positions for the gaps in the b-sequence. Now, to fill the a_i's in the first line, we have to choose exactly k of the a_i's to match against the k gaps in the b-sequence. Thus, there are $\binom{n}{k}$ possibilities. Note that there are still many ways to fill out the rest of the a_i's and b_j's, but they will all be aligned against gaps, and the only difference would be the order in which you arranged the gaps top or bottom, that is, all these possibilities represent the same *trace*. So, there are

$$\binom{m+k}{k \times \binom{n}{k}}$$

possible traces of length $m + k$. Adding we get

$$\text{Number of traces } = \sum_{k=0}^{n} \binom{m+k}{k} \times \binom{n}{k} = \binom{m+n}{n}.$$

I don't know of an exact formula for the number of alignments (as opposed to traces). It can be estimated. The intercalation argument in the hint is *not* correct, but it does give a lower estimate on the size of the set of alignments.

Prob. 6-7. Not much problem here. Some couldn't get the USC server to work, but we can have a look at that in the lab 2/20/04.