**8.592:** Statistical Physics in Biology Assignment # **2** Due: **2/28/05**

## Sequence Alignment

**1.** *Number of gapped alignments:* The following problem is taken from Chapter 2 of *Durbin et. al.*, which provides the needed background on sequence alignments.

(a) Show that the number of ways of intercalating two sequences of lengths $n$ and $m$ to produce a sequence of length $n + m$, while preserving the order of symbols is $\binom{n+m}{m}$ For example $(B_1, A_1, B_2, B_3, A_2)$ is a possible intercalation of $(A_1, A_2)$ with $(B_1, B_2, B_3)$. *(Note a similarity to number of ways of distributing $m$ quanta of energy between $n + 1$ harmonic oscillators.)*

(b) By taking alternating symbols from the upper and lower sequences in an alignment, then discarding the gap characters, show that there is a one-to-one correspondence between gapped alignments of two sequences and intercalated sequences of the type described in part (a). The example in part (a) thus corresponds to the alignment $\begin{pmatrix} - & A_1 & - & A_2 \\ B_1 & B_2 & B_3 & - \end{pmatrix}$. Hence obtain the number of possible gapped alignments between two sequences of length $n$.

(c) Use Stirling's approximation ($x! \approx \sqrt{2\pi} x^{x+1/2} e^{-x}$) to simplify the expression for the number of alignments of two sequences of length $n$.

********

**2.** *Alignments:* Calculate the dynamic programming matrices and the optimal *global* and *local* alignments for the DNA sequences `GAATTC` and `GATTA`, scoring +1.5 for a match, -1 for a mismatch, and with a penalty of 2 for each gap. Do you get different alignments? Do you notice ambiguity in the global alignment?

********

**3.** *Random Sequences and BLAST:* Sequence alignments can be performed online using the programs and data provided by the National Center for Biotechnology Information (NCBI). To better understand the underlying program, read the paper on BLAST, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*, by Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, Nucleic Acids Res. **25**, 3389-3402 (1997). (This paper is available in the Assignment section.)

(a) Generate a random amino acid sequence and run it against a database of non-redundant sequences employing BLAST (http://www.ncbi.nlm.nih.gov/BLAST/); use the standard protein-protein BLAST [blastp]. Repeat your runs for several times and for sequences of different length (10-1500 amino acids). Did you find any “false homologous” in the database?

(b) Generate a random amino acid sequence with amino acid frequencies from the table below and with PERIODICITY of hydrophobic and non-hydrophobic residues. (Hint: the first 10 amino acids in the table below are hydrophobic.) Run these sequences using BLAST. Interpret your results.

(c) Take an amino acid sequence (a protein of your choice, or one of proteins suggested below) and introduce $X\%$ of random mutations. Run mutated proteins against the database using BLAST or PSI-BLAST (same web page). Try different frequency of mutations ($X = 0, \cdots 100\%$). What level of mutations is tolerated by BLAST? by PSI-BLAST? Interpret your results.

**(Optional)** (d) Introduce $X\%$ of mutations such that a hydrophobic amino acid is substituted by a random hydrophobic one and a polar is substituted by a polar one. Run using BLAST. Did the threshold level for $X$ change?

*Table of Amino Acid Frequencies, and their single letter designation (in parenthesis)*

CYS 1.660 (C)<br>
MET 2.370 (M)<br>
PHE 4.100 (F)<br>
ILE 5.810 (I)<br>
LEU 9.430 (L)<br>
VAL 6.580 (V)<br>
TRP 1.240 (W)<br>
TYR 3.190 (Y)<br>
ALA 7.580 (A)<br>
GLY 6.840 (G)<br>
THR 5.670 (T)<br>
SER 7.130 (S)<br>
GLN 3.970 (Q)<br>
ASN 4.440 (N)<br>
GLU 6.360 (E)<br>
ASP 5.270 (D)<br>
HIS 2.240 (H)<br>
ARG 5.160 (R)<br>
LYS 5.940 (K)

PRO 4.920 (P)

*Some suggested sequences*

*Myoglobin:* MAKRRGSVPGRVREYWLPSPCWKCHMLHQGKWWGRRSQGMGGAE GFMEHGSTTLQRKPGASSELGILQVR DLSWLVQPQAQTCCGSFVPLSAGLRASAK

*Histon H2B:* MTDKITKKKRNETYSIYIYKVLRQVHPKIGVSSKAMNIMNSFVNDLFE RLVSESYNLSNSSRSKTLTARE IQTSVRLVIPGELAKHSVSEGTKAVAKYRSSI

*SH3 Domain:* MDETGKELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKV EVNDRQGYVPAAYVKKLD

********

## Population Genetics

**4.** *Steady State:* Consider a population of $N$ diploid individuals, containing an allele with two alternate forms of $A_1$ and $A_2$, in a proportion $x \equiv [A_1]/([A_1]+[A_2])$. Mutations occurs at rate $\mu_1$ ($\mu_2$) per generation for changes $A_2 \to A_1$ ($A_1 \to A_2$), and there is no selection.

(a) Write down the equation governing the probability $p(x,t)$, and find its steady state solution $p^*(x)$.

(b) Find the mean, $\langle x \rangle$, and variance, $\langle x^2 \rangle_c$, in the steady state population.

(c) What is the most likely value (mode) of $x$ as a function of the population size $N$?

********