8.592–HST.452J: Statistical Physics in Biology Assignment # **1** Due: **2/14/05**

Probability

1. *Open Reading Frames:* Assume that the nucleotides A, G, T, C occur with equal probability (and independently) along a segment of DNA.

(a) From the genetic code calculate the probability p_s that a randomly chosen triplet of bases corresponds to a stop signal.

(b) What is the probability for an open reading frame (ORF) of length N, i.e. a sequence of N non-stop triplets followed by a stop codon?

(c) The genome of E-coli has roughly 5×10^6 bases per strand, and is in the form of a closed loop. If the bases were random, how many ORFs of length 600 (a typical protein size) would be expected on the basis of chance. (Note that there are six possible reading frames.)

****

(Optional) 2. *ORFs in E. coli:* To compute the actual distribution of ORFs in *E. coli* you will need to download the complete sequence of its genome from ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Escherichia_coli_K12/U00096.fna .

(a) Write a program that goes through all consecutive (non-overlapping) triplets looking for stop codons. (Make sure you use the genetic code for DNA in the 5'-3' direction.) Record the distance L between consecutive stop codons. Repeat this computation for the 3 different reading frames (0, +1, +2) in this direction. (You may skip calculations for the reverse strand, that is complementary to the given one and proceeding in the opposite direction.)

(b) Plot the distribution for the ORF lengths L calculated above, and compare it to that for random sequences.

(c) Estimate a cut-off value L_{cut}, above which the ORFs are statistically significant, i.e. the number of observed ORFs with $L > L_{cut}$ is much greater than expected by chance.

****

3. *Mutual information:* Consider random variables x and y, distributed according to a joint probability $p(x, y)$. The mutual information between the two variables is defined by

$$M(x, y) \equiv \sum_{x,y} p(x, y) \ln \left(\frac{p(x, y)}{p_x(x) p_y(y)} \right),$$

where p_x and p_y denote the *unconditional* probabilities for x and y.

(a) Relate $M(x, y)$ to the entropies $H(x, y)$, $H(x)$, and $H(y)$ obtained from the corresponding probabilities.

(b) Calculate the mutual information for the joint Gaussian form

$$p(x, y) \propto \exp\left(-\frac{ax^2}{2} - \frac{by^2}{2} - cxy\right).$$

****

4. *Information content of the genetic code:* We would like to quantify the observation that the first two letters of the triplet code carry most of the information about the resulting amino acid. For the purpose of such a calculation, the stop signal and the 20 amino acids shall be regarded as 21 equivalent possible outcomes.

(a) In the absence of any other information, $\ln_2(21) \approx 4.39$ bits of information are needed to specify one of the 21 possible outcomes. If the first letter of the code is A, there are still 7 possible outcomes with probabilities that can be read from the genetic code. Calculate the Shannon entropy associated with the latter (conditional) probability, and hence deduce how many bits of information have been gained by the knowledge that the first letter of the code is A.

(b) Repeat the above calculation for the three other choices of the first letter, and hence compute the *average* information gained by knowledge of the first letter of the code. What fraction of the total information is this?

(c) How would you go about calculating the information content of the second and third letters of the code?

****

5. *Selection and mutation:* Consider a very large population of individuals characterized by a fitness parameter f, which is assume to be Gaussian distributed with a mean m and variance σ. The population undergoes cyclic evolution, such that at each cycle: *(i)* one half of the population with lower fitness f is removed without creating progeny; *(ii)* the remaining half (with f values in the upper half) reproduces before dying; *(iii)* because of mutations the f values of the new generation is again Gaussian distributed, with mean value and variance reflecting the parents (i.e. coming from the upper half of the original Gaussian distribution).

(a) Relate the mean m_n and variance σ_n of fitness values of the n-th generation to those of the previous ones (m_{n-1} and σ_{n-1}).

(b) What happens to the distribution of fitness after many generations?

****