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.)

*****

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_K_12_substr__MG1655_uid225/.
This file is also posted on the *Assignments* web-page.

(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. *Point mutations in DNA:* Since the four nucleotides in DNA have different chemical compositions and energetics, they could mutate at different rates. We shall explore whether, without natural selection at work, such preferential mutation may lead to different compositions of nucleotides.

(a) Consider a simple model in which all *transitions* (i.e. mutations between purines A and G, or between pyramidines T and C) occur with probability q, while *transversions* (i.e. any mutation from a purine to a pyrimidine or vice versa) occur with probability p, in each generation. Write down the 4×4 (Markov) transition matrix, Π_1, that relates the frequencies of nucleotides ($p_{\rm A}$, $p_{\rm G}$, $p_{\rm T}$, $p_{\rm C}$) from one generation to the next. (Make sure that the normalization condition $p_{\rm A} + p_{\rm G} + p_{\rm T} + p_{\rm C} = 1$ is preserved.)

(b) Find the eigenvalues of the transition matrix Π_1. (**Hint:** You should be able to simply guess the eigenvectors by considering the symmetries of the matrix.)

(c) Find the matrix $\Pi_t = \Pi_1^t$, describing the evolution of probabilities after t generations.

(d) Show that in steady state (after many duplications), all nucleotides occur with the same frequency. Estimate the number of generations (as a function of p and q) needed to reach such a steady state.

(e) You should be able to convince yourself that for any model in which mutation rates between pairs of bases are the same in the forward and backward directions, all nucleotides are equally likely in the steady state. However, in the human genome the nucleotides C and G occur less often than A and T. This is partly due to methylation of successive CG pairs which makes them more susceptible to mutations. To mimic this asymmetry, consider an unrealistic model in which transversions from A to C and T to G occur with probability p_+, while the reverse transversions (from C to A or G to T) occur at a lower probability of p_-. (The other transversions occur at rate p, and transitions at rate q as before.) Write the modified transfer matrix corresponding to this model, and obtain the resulting frequencies of nucleotides in steady state.

*****

4. *Correlations in the E. coli genome:* In the models examined in the previous problem, point mutations at each position on the DNA occur at rates independent of other locations. Consequently, they predict $p_{\rm XY} = p_{\rm X} p_{\rm Y}$, where $p_{\rm XY}$ is the *joint probability* of finding nucleotides X and Y, at different locations. Test this hypothesis on the genome of E. coli (available on the course web-page) as follows:

(a) Calculate the frequencies of the four nucleotides in the genome.

(b) Write a program to count all 16 possible pairs of neighboring bases (e.g AT); hence obtain the joint probabilities $p_{\rm XY}$, and construct the 4×4 matrix of correlations $c^{(1)}_{\rm XY} = p_{\rm XY}/(p_{\rm X} p_{\rm Y})$.

(c) Repeat the above calculation for nucleotides that are further neighbors, and find the corresponding matrices $c^{(n)}_{\rm XY}$ (e.g. consider next nearest neighbor locations j and $j+2$ to calculate $c^{(2)}_{\rm XY}$). How do correlations decay as a function of the separation n?

*****

5. *Selection and mutation:* Consider a very large population of individuals characterized by a fitness parameter f, which is assumed 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?

*****

6. *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 μ_1 (μ_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?

*****

7. *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

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

*****

8. *Activation/deactivation reaction:* Many molecules in biology can be made active or inactive through the addition of a phosphate group. The enzyme that adds the phosphate group is usually termed a kinase, while a phosphatase removes this group. Let us consider a case where a finite number N of such molecules within a cell can be exchanged between the two forms at rates a and b, i.e.

$$\mathrm{A} \rightleftharpoons_b^a \mathrm{B} \quad ,$$

where we have folded the probabilities to encounter the enzymes in the reaction rates.

(a) Write down the Master equation that governs the evolution of the probabilities $p(N_A = n, N_B = N - n, t)$.

(b) Assuming that initially all molecules are in state A, i.e. $p(n, t=0) = \delta_{n,N}$, find $p(n,t)$ at all times. You may find it easier to guess the solution, but should then check that it satisfies the equations obtained before.

*****

9. *Human polymorphisms:* Explore the data on gene polymorphism in human population from the Seattle SNP database, one of the largest collections of polymorphisms in humans. Each entry reports a polymorphism (e.g A/G) and its frequency x in individuals of European or African descent (ED and AD). Some information from this data is provided on the course

web-page in the form of the file cSNPsAfricanEuropean.dat, which contains these frequencies (each line has information about one polymorphism ordered as AD-freq, ED-freq).

(a) 1. Make the histogram of x and compare it to the steady–state distribution $f(x)$ obtained in the class. Make conclusions about the value of $N\mu$. Does the theoretical formula fit these distributions? (**Hint:** You may want to consider only alleles with $x < 0.5$ by replacing $x > 0.5$ with $1 - x$).

(b) Compare $f(x)$ for African and European populations. Do you see any difference? Assuming $\mu = 10^{-8}$ estimate the effective population size N for the two populations by fitting $f(x)$ to the data. (**Hint:** use roughly 20-30 bins for $0 < x < 0.5$, and ignore very rare polymorphisms). Although this method is not the best for estimation of the population size, it gives you an idea about the relative size of the population size in the two populations.

(c) Compare the frequencies of synonymous and non-synonymous mutations. Do they fit the theoretical $f(x)$? Which ones are more frequent and why? (The corresponding data are provided in the file cSNPsAfricanEuropeanType.dat.)

*****