Fixation & Sequence Alignment

1. *Multi–allele model:* Consider a locus occurring in one of $s+1$ states $(A_1, A_2, \cdots A_s, A_{s+1})$ with frequencies $x_1, x_2, \cdots, x_s, x_{s+1} = 1 - \sum_{i=1}^{s} x_i$ in a (haploid) population of fixed size N. After 1 generation, random reproduction leads to new population in which the allele numbers $\{m_i\}$ occur with the multinomial probability

$$p(\{m_i\}) = N! \prod_{i=1}^{s+1} \frac{x_i^{m_i}}{m_i!}.$$

(a) Show that the changes in frequency after one generation satisfy (for $i = 1, \cdots, s$)

$$\langle \Delta x_i \rangle = 0, \quad \left\langle \Delta x_i^2 \right\rangle = \frac{x_i(1-x_i)}{N}, \quad \text{and } \langle \Delta x_i \Delta x_j \rangle = -\frac{x_i x_j}{N}, \text{ for } i \neq j.$$

(b) Obtain the forward Kolmogorov equation for the probability $p(\vec{x}, t)$, where $\vec{x} = (x_1, \cdots, x_s)$ is a vector with s components.

(c) Obtain the backward Kolmogorov equation for the probability $p(\vec{x}, \vec{y}, t)$, given an initial state $\vec{y}$.

(d) The probability that an initial polymorphism characterized by $\vec{y}$ disappears (by either loss or mutation) at time t is given by

$$p_\times(t|\vec{y}) = -\int d^s\vec{x} \frac{dp(\vec{x}, t|\vec{y})}{dt}.$$

Find an equation (similar to the backward Kolmogorov equation) satisfied by the mean time for fixation $\langle \tau(\vec{y}) \rangle_\times$, and show that a simple generalization of the result for two alleles satisfies this equation.

*****

2. *Splitting probability and mean first passage time:*

(a) Consider a particle diffusing without a drift ($v = 0, D = const$) on an interval $[0, 1]$ that has adsorbing boundaries. The particle starts at $0 < x < 1$ and diffuses until it gets adsorbed by either of the boundaries. Using backward Kolmogorov equations calculate the probabilities of adsorption (splitting probabilities) by either boundary $\Pi_0(y)$ and $\Pi_1(y)$.

(b) Obtain the expression for the mean first passage time $\overline{t}(y)$ through either boundary (survival time). Find the location $y_{\max}$ that provides the longest mean life-time for the particle.

(c) Consider a generalization of the same problem to two dimensions, with two concentric circles of radii $R_0 < R_1$ centered at the origin. A particle starts a 2D diffusion (without a

drift) at distance R ($R_0 < R < R_1$) from the origin. Use the general form of the equations for slitting probabilities, $\nabla^2 \Pi = 0$, and mean exit time, $D\nabla^2 \overline{t} = -1$, to calculate $\Pi_0(R)$, $\Pi_1(R)$ and $\overline{t}(R)$. Find the initial distance $R_{\rm max}$ that provides the longest life-time. Compare to result obtained for one dimension.

(d) Consider the same problem in three dimensions with two concentric spheres. Compare results in 1D, 2D and 3D. How do the split of probabilities between the inner and the outer boundaries, and $R_{\rm max}$ change as the dimensionality of the system increases from 1 to 3?

*****

3. *Fibonacci's Rabbits:* Consider a generalization of Fibonacci's model for rabbit populations, in which some of the rabbits (whether young or adult) die in each generation, and not all adults reproduce. *(i)* Assume the same mortality rate for young and adult rabbits, such that a fraction p of young rabbits matures to adulthood, and the same fraction p of adult rabbits survives to the next generation. *(ii)* Assume that a fraction f of adults reproduce the next generation of young rabbits.

(a) Write the 2×2 (transfer) matrix that relates the populations of young and adult rabbits from one generation to the next, i.e. find the elements of the matrix T, such that

$$\begin{pmatrix} Y_{N+1} \\ A_{N+1} \end{pmatrix} = T \begin{pmatrix} Y_N \\ A_N \end{pmatrix},$$

where Y_N and A_N are the (average) numbers of young and adult rabbits in generation N.

(b) Show that for large N the population grows (or decays) exponentially, and find the ratio $\lambda(p, f)$ between numbers of successive generations. What is the condition that separates growing and decaying populations?

(c) Find the asymptotic ratio of young to adult rabbits, and note its connection to an eigenvector of the transfer matrix.

*****

4. *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.

*****

5. *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?

*****

6. *Extremal Gaussian distribution:* Let $x = \max\{r_1, r_2, \cdots, r_N\}$ be the largest of N independent, identically distributed Gaussian variables. Specifically, each r is distributed according to

$$p_1(r) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{r^2}{2\sigma^2}\right), \quad \text{and for any PDF } p_1 \text{ set } \quad \overline{P}_1(x) \equiv \int_x^\infty dr p_1(r).$$

(a) Find the cumulative probability, $P_N(x)$ that the maximum is less than or equal to x.

(b) Show that in the limit of $N \gg 1$, x^*, the most likely value of x, can be obtained from either of the following expressions

$$\overline{P}_1(x^*) = \frac{1}{N}, \qquad \text{or} \qquad p_1'(x^*) + N p_1(x^*)^2 = 0,$$

and behaves as $x^* \simeq \sqrt{2\sigma^2 \ln N}$.

(c) By expanding the solution to part (a) around x^*, and expressing the first order expansion as an exponential (i.e. replacing $1 + \delta$ with e^δ), show that the probability distribution for x approaches a Gumbel form, and identify the corresponding parameters.

*****

7. *Gapless alignment significance:* Consider a scheme for aligning DNA sequences in which a score $s = 1$ is assigned to a match, while $s = 0$ for a transversion (A↔G or T↔C) and $s = -\mu$ for a transition (e.g. A↔C). (Assume all four nucleotides occur with equal frequency.)

(a) Find the parameter $\lambda(\mu)$ that appears in the Gumbel distribution for the statistics of such random gapless alignments.

(b) Show that alignments are possible only for $\mu > \mu_c$, and plot λ as a function of μ.

*****

8. *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 on the course web-page 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.

(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)
MET 2.370 (M)
PHE 4.100 (F)
ILE 5.810 (I)
LEU 9.430 (L)
VAL 6.580 (V)
TRP 1.240 (W)
TYR 3.190 (Y)
ALA 7.580 (A)
GLY 6.840 (G)
THR 5.670 (T)
SER 7.130 (S)
GLN 3.970 (Q)
ASN 4.440 (N)
GLU 6.360 (E)
ASP 5.270 (D)
HIS 2.240 (H)
ARG 5.160 (R)
LYS 5.940 (K)
PRO 4.920 (P)

Some suggested sequences

Myoglobin: MAKRRGSVPGRVREYWLPSPCWKCHMLHQGKWWGRRSQGMGGAE GFMEHGSTTLQRKPGASSELGILQVR DLSWLVQPQAQTCCGSFVPLSAGLRASAK

Histon H2B: MTDKITKKKRNETYSIYIYKVLRQVHPKIGVSSKAMNIMNSFVNDLFE RLVS-
ESYNLSNSSRSKTLTARE IQTSVRLVIPGELAKHSVSEGTKAVAKYRSSI
SH3 Domain: MDETGKELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKV
EVNDRQGYVPAAYVKKLD

*****