1. **Directed Polymer:** A simplified model polymer is constructed by joining together elementary bonds of length $\ell$. There is usually a stiffness that tends to align consecutive bonds. For the purposes of this problem, consider a simpler situation in which all bonds have a tendency to align to the same direction, denoted by the $z$ axis. The direction of the bond is then indicated by the angle $\theta$ with respect to the $z$ axis, which is taken independently for each bond from the probability density $p(\theta) = 2\cos^2(\theta/2)/\pi$; and the polar angle $\phi$ which is uniformly distributed between 0 and $2\pi$. (Note that the solid angle factor of $\sin \theta$ is already included in the definition of $p(\theta)$, which is correctly normalized to unity.) One end of this $z$-directed polymer is attached to the origin, and the other end, after including a larger number of bonds $N$, is located at the point $\vec{r} = (x, y, z)$.

(a) Calculate the expectation values $\langle z \rangle$, $\langle x \rangle$, $\langle y \rangle$, $\langle z^2 \rangle$, $\langle x^2 \rangle$, and $\langle y^2 \rangle$, and the covariances $\langle xy \rangle$, $\langle xz \rangle$, and $\langle yz \rangle$.

(b) Use the central limit theorem to estimate the probability density $p(x, y, z)$ for the particle to end up at the point $(x, y, z)$.

**********

2. **Optimal Selections:** In many specialized populations, there is little variability among the members. Is this a natural consequence of optimal selection?

(a) Let $\{r_\alpha\}$ be $n$ random numbers, each independently chosen from a probability density $p(r)$, with $r \in [0, 1]$. Calculate the probability density $p_n(x)$ for the largest value of this set, i.e. for $x = \max\{r_1, \ldots, r_n\}$.

(b) If each $r_\alpha$ is uniformly distributed between 0 and 1, calculate the mean and variance of $x$ as a function of $n$, and comment on their behavior at large $n$.

**********

3. **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 finding a sequence of $N$ triplets without a stop codon, i.e. an open reading frame (ORF) of length $N$?
(c) The genome of E-coli has roughly $5 \times 10^6$ 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) 4. 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.

********

5. 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?

********