Protein, DNA, and RNA

1. *Designed Random Energy Model (REM):* Consider a protein model in which for a given sequence and structure, the energy is randomly taken from the Gaussian probability density

$$p(E) = \frac{1}{\sqrt{2\pi\Sigma^2}} \exp\left(-\frac{E^2}{2\Sigma^2}\right).$$

The total number of structures is Ω_{str}, while the number of sequences is $\Omega_{seq} \gg \Omega_{str}$.

(a) A particular *sequence* has a (unique) native structure of energy E_N. Calculate and plot the energy $E(T)$ of this sequence as a function of temperature T.

(b) For a particular *structure*, we attempt to design a good sequence by Monte Carlo sampling of representative sequences at a 'temperature' τ. Calculate and plot the designed native energies $E_N(\tau)$ as a function of the design temperature τ.

*****

2. *Charged Random Energy Model:* Use the random energy model to investigate the freezing of a charged heteropolymer. Assume that there are g^N possible globular states of the polymer, whose energies are randomly selected from a Gaussian distribution of mean zero, and variance

$$\sigma^2 = u^2 N + c^2 \left(\frac{Q^2}{R}\right)^2.$$

The second term in the above formula is a rough estimate of the variations in Coulomb energy from different ways of distributing a charge Q over a volume of size R.

(a) Find the energy E_c at which the entropy vanishes, and the corresponding freezing temperature T_c.

(b) For compact globular states, how should Q^2 scale with N for the freezing temperature to be asymptotically independent of N?

*****

3. *Amino-acid interactions:* What can we learn by combining the Random Energy Model with commonly used interaction potentials between amino acids?

(a) Find a 20×20 matrix of interactions $U(a, a')$ amongst amino acids, and calculate the mean $\langle U \rangle$ and variance $\langle U^2 \rangle_c$ of its elements. The commonly used Miyazawa–Jernigan (MJ) interaction matrix can be found in S. Miyazawa and R.L. Jernigen, J. Mol. Biol. **256**, 623 (1996). (Table 3 of this publication is available on the web-page for assignments.)

(b) Model the possible configurations of a protein by the ensemble of compact self-avoiding walks on a cubic lattice. (All lattice sites are visited by compact walks.) Calculate the number n of non-polymeric nearest neighbor interactions for such configurations on an $N = L \times L \times L$ lattice, and deduce the ratio n/N for large N.

(c) The number of compact walks on a cubic lattice asymptotically grows as g^N, with $g \approx 1.85$. Use this in conjunction with the results from parts (a) and (b) to estimate the folding temperature T_c of a random sequence of amino-acids, and the corresponding energy E_c.

(d) Select a protein, find its amino-acid sequence and construct a contact matrix corresponding to its structure. Use the interaction matrix from part (a) to estimate the energy of the native structure, and calculate the ratio E_N/E_c.

*****

4. *Kinetics of protein folding:* [Adapted from Gutin *et al.*, J. Chem. Phys. **108**, 6466 (1998).] Assume protein folding proceeds through a folding nucleus which has the free energy $F^\ddagger = E^\ddagger - k_B T \log M^\ddagger$. The folding nucleus serves as a transition state for the folding reaction. The typical folding time needed to climb over this free energy barrier is

$$t = \tau_0 \exp\left(\frac{F^\ddagger - F}{k_B T}\right),$$

where T is the temperature, and τ_0 is an elementary time step.

(a) Use a random energy model to calculate F as a function of temperature T, and calculate the folding time $t(T)$ for two regimes $T > T_c$ and $T < T_c$. Plot $\ln t(T)$ as a function of $1/T$.

(b) Consider a limit of $T \to \infty$ and express the folding time as a function of the total number of conformations $M = g'^N$ and the number of states in the folding nucleus $M^\ddagger$. Interpret your result.

(c) Find a temperature T_{opt}, which provides the fastest folding, compare it to T_c. Compare the optimal folding time with the folding time from "non-designed" REM at T_c. Make conclusions about folding kinetics for random sequences (REM) and designed sequences (designed REM).

*****

5. *Denaturing DNA by force:* Obtain the phase diagram of DNA pulled by a force $\vec{F}$, by generalizing the Poland–Scheraga model as follows:

(a) By integrating over the position vectors, show that the (Gibbs) partition function of DNA of length N can be decomposed into products of contributions from double-stranded rods and single stranded bubbles, as

$$Z(N,F) = \sum_{\ell_1,\ell_2,\ell_3,\cdots} R(\ell_1)B(\ell_2)R(\ell_3)\cdots, \qquad \text{with} \qquad \ell_1 + \ell_2 + \ell_3 + \cdots = N.$$

(b) Treat the double stranded segments as rigid rods of fixed length $a\ell$. By integrating over all orientations in three dimensions show that

$$R(\ell) = w^\ell \times \frac{\sinh(\beta F a \ell)}{\beta F a \ell},$$

where $w = e^{-\beta\varepsilon}$, and ε is the energy gain of forming the double strand.

(c) Treat the double stranded loop as two random walks of length ℓ connected at the two end points. Integrating over all separations of the two end points show that

$$B(\ell) = \frac{s}{\ell^{3/2}} \left[g^2 \exp\left(\frac{\beta^2 F^2 a^2}{12} \right) \right]^{\ell}.$$

(d) Examine the problem in a (grand canonical) ensemble with variable DNA lengths N, additionally weighted by a factor of z^N. Give the expressions for the (Laplace) transformed $\tilde{B}(z)$ and $\tilde{R}(z)$ in this ensemble in terms of the (Bose) sums $f_m^+(x) = \sum_{\ell=1}^{\infty} x^\ell/\ell^m$.

(e) Show that the strands become fully separated at a critical point satisfying $\tilde{R} = \tilde{B}^{-1} = \left(s\zeta_{3/2}\right)^{-1}$, where $\zeta_{3/2} \equiv f_{3/2}^+(1) \approx 2.612$.

(f) For $s = 1$, plot the phase diagram of the model in the coordinates (w/g^2) and $(\beta F a)$.

*****

6. *Denaturing RNA by force:* By pulling on the ends of RNA, the hydrogen bonds can be broken to yield a stretched polymer. Let us model the partially denatured state as a sequence of linear segments with no hydrogen bonds and 'blobs' which are hydrogen bonded (opposite to the case of DNA). Assume that the force carrying backbone of the molecule is made up of the linear segments, and that the RNA blobs carry no force (similar to the loop in problem 2). After integrating over the position vectors, the (Gibbs) partition function of an RNA of length N can be written as

$$Z(N,F) = \sum_{\ell_1,\ell_2,\ell_3,\cdots} P(\ell_1)R(\ell_2)P(\ell_3)\cdots, \qquad \text{with} \qquad \ell_1+\ell_2+\ell_3+\cdots = N.$$

The contributions of linear and blob segments are respectively

$$P(\ell) = g^\ell \exp\left(\frac{F^2 a^2 \ell}{6k_B^2 T^2}\right), \quad \text{and} \quad R(\ell) = f^\ell \frac{A}{\ell^{3/2}}.$$

(a) Exploit the mathematical similarity to the Poland–Scheraga model to evaluate the grand partition function of the model.

(b) Identify the force F_c at which denaturation starts.

(c) Sketch the fraction of denatured sites as a function of force, clearly indicating the nature of the singularity at F_c.

*****

7. *Pulling RNA:* The server on http://bioinfo.ucsd.edu/rna/ (or the pulling server at http://bioserv.mps.ohio-state.edu/rna/) gives force extension curves for RNA based on secondary structure calculations. Use this server to examine force extension curves for: (a) a uniform sequence; (b) an alternating sequence of G and C; (c) an alternating sequence of A and U; (d) an actual RNA sequence. (Choose sequences of roughly the same length.)

Comment on the general characteristics of these curves. Does any of them resemble the theoretical result from the previous problem?

*****

8. *Analysis of protein structures:* Calculate ϕ and ψ torsion angles in `Rasmol` for a given protein (see the commands below). Make (ϕ, ψ) "Ramachandran" diagrams by plotting ϕ along the x and and ψ along the y axis; one (ϕ, ψ) point for each amino acid.

(a) Do amino acids that are part of different secondary structure elements (helices, sheets) land in the same or different islands on the (ϕ, ψ) diagram? You can find secondary structure elements in fields `HELIX` and `SHEET` of the protein structure file (aka PDB file). Explain your observations.

(b) Find amino acids that have unusual (ϕ, ψ) angles (i.e. deviate from the many clouds of points). What types of amino acids tend to have "unusial" (ϕ, ψ) conformation? Discuss.

(c) Visualize protein structure in `Rasmol`, following the sequence of commands below, and select those with "unusual" (ϕ, ψ) conformation. Do they tend to be close to the ligand?

Some sample proteins to explore (PDB files provided on the Assignment page):
Hemoglobin (alpha chain) 4HHB_A.PDB
Immunoglobulin domain 1TEN.PDB

You can use the following sequence of `Rasmol` commands to generate a good view of a protein, and the `fipsi.dat` file of (ϕ, ψ) angles

```
set background white
wireframe off
ribbons
color structure
select ligand
cpk
color green
select protein
write RDF fipsi.dat
```

To select a particular set of amino acids, (e.g. 128 and 156) you can do the following

```
select 128,156
cpk
color red
```

*****