Biophysics Midterm – Spring 2009

Gel-less DNA electrophoresis

Though it's rarely done, it is possible to separate DNA by electrophoresis without using a gel. In this problem, I ask you to predict how long it will take to resolve two bands of DNA under different models for the "size" of DNA. The DNA moves because of an eternally applied electric field. At the same time that it migrates the DNA diffuses. You may assume that:

i. Two bands can be resolved when the difference in their centers exceeds their width.
ii. DNA has a fixed charge / length
iii. The electric field is constant inside the fractionation apparatus.
iv. The DNA is running in a buffer which is essentially water (i.e. it has the same viscosity as water).
v. Everything is occurring at room temperature.
vi. Electrophoresis is typically run at an electric field of 10 V/cm.

If necessary, use the variables:

L: DNA length (contour length)
N_{bp}: number of base pairs in the DNA
η: viscosity of water
λ: charge/length of DNA
R_H: hydrodynamic radius of the DNA
ξ: persistence length of DNA
l: effective link length of DNA treated as a random coil.
D: diffusion coefficient
e: charge on the electron
v: speed of the DNA

Useful constants:

Thermal energy k_BT = 25 meV = 4.1 pN nm.
Viscosity of water: η = 1 cP; P = g cm^{-1} s^{-1}.
Charge on DNA : λ = 1 e / 4 nm

To guide you through the calculation, work through the following steps:

1. Write an expression for the propulsive force (due to E acting on the DNA's charge) in terms of λ, L and E.
2. Write an expression for the drag force on the DNA in terms of η, R_H and v.
3. Equate the propulsion and drag to find the terminal (drift) velocity of DNA.
4. Write an expression for the diffusion coefficient of DNA in terms of k_BT, η and R_H.

Three models for the hydrodynamic size of DNA are

random coil: $R_H \propto L^{1/2}$

solid sphere: $R_H \propto L^{1/3}$

free-draining coil: $R_H \propto L$

Don't worry about the origin of these expressions for now: just set $R_H = \beta L^{\alpha}$. For this part, we're only going to calculate a scaling law for fractionation time, so you can drop all constants except α and L for the moment, if you'd like.

5. Calculate the difference (Δv) between the drift velocities of two bands of DNA whose lengths differ by a *small* amount ΔL.
6. Using your expression for Δv, calculate the distance between the two bands' centers as a function of migration time t.
7. Calculate the width of one of the bands of DNA as a function of time t. (You may assume that ΔL is small enough that the bands will have the same diffusion coefficient)
8. Find the *resolving time* T by setting the center-to-center distance equal to the band width.
    a. How does T scale with DNA length?
    b. Your expression should blow up for α=1. Why?

Put in actual numbers for DNA, treating it as a random coil and assuming that $R_H = R_g$. That is, go back and insert all the constants if you dropped them before. *Assume you need to achieve single bp resolution*.

9. Work out an expression for the resolving time T (in seconds) as a function of DNA size (Nbp, in base pairs).
10. What is the resolving time T for a 1 kb strand of DNA?
11. How far will the DNA move in that time?

HP model

Take the following data for degeneracies of an HP model run on an unbounded lattice:

sequence	W(0)	W(1)	W(2)	W(3)	W(4)	W(5)	W(6)
hhhhhphphhph	5969	5506	2539	808	184	29	1
phhphppphpphp	10059	4385	573	18	1	0	0
hphphphphhhh	9823	3700	1218	270	24	1	0

This is an exhaustive list of all the possible conformations of the sequences, tabulated by number of H-H contacts n. Each W(n) is the number of occurrences of n-contact conformations for the given sequence. This particular H-P model gives a fixed negative (binding) energy to every H-H contact and zero energy to all other contacts. Note that this is a different definition from the one used in PBoC, but it gives equivalent results. (Penalizing H-P contacts or rewarding H-H contacts amounts to the same thing).

Consider each column to be a different state. That is, from left to right we have the zero-contact state (completely unfolded), the one-contact state (slightly folded), ... down to the single, unique folded state. Depending on the sequence, the folded state has 4, 5 or 6 H-H contacts. Call the binding energy $-\varepsilon$, so all temperatures are expressed in units of ε / k_B.

For a given sequence, compare the free energies of two states i and j. Define the *transition temperature* $T_{i,j}$ as the temperature at which $G(i,T) = G(j,T)$.

1. Show that $k_B T_{i,j} / \varepsilon \ \ln W(i)/W(j) = -(i-j)$
2. Calculate $T_{i-1,i}$ (in units of ε / k_B) for each of the sequences' folded states. This could plausibly be called the unfolding temperature.
3. Show that $k_B T_{i,j} / \varepsilon = (i-j) / (\ln W(i) - \ln W(j))$. This means that on a plot of n versus W(n), the slope of the line connecting data points i and j is equal to $T_{i,j}$ in units of (ε/k_B).

We expect that $T_{i,i+1}$ will decrease as we approach the folded state, because as we lower the temperature the binding energy should increasingly win over entropy; the lowest energy states should become progressively more favored with decreasing T, until we finally reach the native (W=1) state. From part 3, if $T_{i,i+1}$ decreases monotonically, the n versus ln W(n) curve will be everywhere *concave*.

4. One of the three sequences has a *convex* region of the n versus ln W(n). (You may need to zoom in on the graph).
   a. Which sequence, and between which values of n?
   b. Calculate transition temperatures in the neighborhood of this region and find the temperature range over which the folded state is metastable. (By metastable, I mean that the folded state is a local minimum, but not a global minimum, of the free energy surface.)
   c. Check your numbers by plotting $G(n,T)$ for three temperatures: just below the metastable range, in the middle of the metastable range, and just above the metastable range. Zoom in on the region of interest if necessary, and show that the free energy curve has the appropriate shape at each temperature.

Villin headpiece subdomain folding

Read the article by Lei, Wu, Liu and Duan, "Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations", PNAS 104:4925-4930 (2007). The article is accessible through a link in the E-Reserves section of the course website.

Answer the following questions. A couple of sentences per answer is probably enough.

1. p.1: "When heavy-atom ... common" What is the difference between "heavy-atom" rmsd and C_α-rmsd? Why would you expect the former to be larger than the latter?
2. p.1: "HP35 is arguably ... understanding of protein folding". What are the authors particularly interested in finding the *smallest* such protein?
3. p.1: "Nonetheless, the best C_α-sampled ... higher". What is the authors' point in explicitly referring to both the "best" and "most-sampled" conformations here? That about what defined the "best" conformation and how it relates to the goal of MD folding simulations: the prediction of protein structure completely *in silico*.
4. p.1: Expand on the authors' rather sketchy description of REMD. What is the advantage of REMD over conventional molecular dynamics simulations?
5. p.2: "Subangstrom folding ... temperatures < 360 K ... excluding terminal residues Leu1 and Phe35".
    1. What does "subangstrom folding" mean?
    2. In the introduction, the authors state this protein has a melting temperature of 342 K. Why are they able to get it to fold at 360 K?
    3. Why are the authors excluding the terminal residues from their rmsd calculation?
6. p.2: "Segment A encompasses ... segment B encompasses ...". Apparently the two segments overlap.
    1. How many residues are shared by segments A and B?
    2. Extra credit: Why would you want them to overlap? ("Extra" because I don't understand the authors' explanation of this, and I wonder if any of you do).
7. p.2: "At 300K ... marginally populated denatured state". Identify these states on a printout of the article. Why do the authors say the intermediate state is "widely distributed"?
8. p.2: Estimate where the initial conformation (used to start the folding simulations) would appear on figure 2 or 3. You can't guess what the population (or free energy) of that state would be, but you should be able to place it (roughly) on the rmsd A - rmsd B plane.
9. p.2: What is the difference between the top left panel of Fig 2 and Fig 3? I.e. what property of the conformations is included in Fig 3 but omitted in Fig 2?
10. p.2-3: "The minimum in the I1 region ... in the D region ... in the I2 region..." All of these are minima. Shouldn't they all have negative free energies?
11. p.3: "This is consistent with ... enthalpy and entropy contributions". Explain.
12. p.3: "This is consistent with ... two-step folding process". Why is this *qualitatively* consistent? Is it *quantitatively* consistent?
13. p.4: "Therefore, the major intermediate (I1) ... I2 was an off-pathway intermediate". What do "on-pathway" and "off-pathway" intermediate mean?
14. p.4: "Therefore, the rigid Pro21 ... segment B." Why do they say Pro21 "restrict[s] movement of the two helices"?
15. p.4: "The calculated $\Delta G_{folding}$ was smaller than ... -3.1 kcal/mol" What *is* the authors' calculated $\Delta G_{folding}$?
16. p.5. "For example, the CD signal ... partial helical structures". This doesn't strike me as such a serious problem in this protein. Explain why CD measurements could conflate the two intermediates I1 and I2, but why (if the protein really folds the way these authors claim it does) this is unlikely to cause problems in real CD experiments that fold this protein.