2.4 DNA structure

DNA molecules come in a wide range of length scales, from roughly 50,000 monomers in a λ-phage, 6×10^9 for human, to 9×10^{10} nucleotides in the lily. The latter would be around thirty meters long if fully stretched. If we consider DNA as a random (non-self avoiding) chain of persistence length $\xi_p \approx 50$ nm, its typical size would be $R_g \approx \sqrt{L \cdot R_p}$, coming to approximately 0.2 mm in human. Excluded volume effects would further increase the extent of the polymer. This is much larger than the size of a typical cell, and thus DNA within cells has to be highly compactified. Eukaryotes organize DNA by wrapping the chain around histone proteins (nucleosomes), which are then packed together.

At the microscopic level the double helix is held together through Watson–Crick pairs, G–C and A–T, the former (with a binding energy of around $4k_BT$) being roughly twice as strong as the latter. At finite temperatures, this energy gain competes with the loss of entropy that comes with braiding the two strands. Indeed at temperatures of around 80°C the double strand starts to unravel, denaturing (melting) into bubbles where the two strands are apart. Regions of DNA that are rich in A–T open up at lower temperatures, those with high G–C content at higher temperatures. These events are observed as separate blips in ultraviolet absorption as a function of temperature for short DNA molecules, but overlap and appear as a continuous curve in very long DNA.

There are software packages that predict the way in which a specific DNA sequence unravels as a function of temperature. The underlying approach is the calculation of free energies for a given sequence based on some model of the binding energies, e.g. by adding energy gains from stacking successive Watson-Crick pairs. Another component is the gain in entropy upon forming a bubble, which is observed experimentally to depend on the length l of the denatured fragment as

$$S(l) \approx bl + c \log l + d\,, \quad \text{with} \quad c \approx 1.8k_B\,. \tag{2.71}$$

The logarithmic dependence is a consequence of *loop closure*, and can be justified as follows. The two single-stranded segments forming a bubble can be regarded as a loop of length $2l$. Let us consider an open polymer of length $2l$, and the probability that the two end-points are separated by a distance $\vec{r}$. For non-interacting random walks the number of configurations of length $2l$ and end-to-end separation $\vec{r}$ is easily obtained by appropriate extension of Eq. (2.40) to

$$W(\vec{r}, 2l) = W(\vec{r}, l)^2 = g_1^{2l} \exp\left[-\frac{dr^2}{2l\xi_p}\right] \frac{1}{(4\pi l\xi_p/d)^d}\,, \tag{2.72}$$

where we have further generalized to the case of random walks in d space dimensions. The number of configurations of a bubble is now obtained by integrating over all positions of the intermediate point as

$$\Omega(l) = \int d^d\mathbf{r}\, W(\vec{r}, 2l) = \left(\frac{d}{8\pi\xi_p}\right)^{d/2} \frac{g^l}{l^c}\,, \tag{2.73}$$

with $g = g_1^2$ and $c = d/2$.

For the more realistic case of self-avoiding polymers, scaling considerations suggest

$$W(\vec{r}, 2l) = \frac{g^l}{R^d}\,\Phi\left(\frac{\vec{r}}{R}\right)\,, \quad \text{with} \quad R \sim l^\nu\,, \quad \text{and} \quad \Omega(l) = \frac{g^l}{l^{d\nu}}\,. \tag{2.74}$$

We can understand this dependence by noting that in the absence of the loop closure constraint the end-point is likely to be anywhere in a volume of size roughly $R^d \propto l^{d\nu}$, and that brining the ends together reduces the number of choices by this volume factor. As we shall see shortly, the parameter g is important in determining the value of the denaturation temperature, while c controls the nature (sharpness) of the transition.

2.4.1   The Poland–Scheraga model for DNA Denaturation

Strictly speaking, the denaturation of DNA can be regarded as a phase transition only in the limit where the number of monomers N is infinite. In practice, the crossover in behavior occurs over an interval that becomes narrower with large N, so that it is sharp enough to be indistinguishable from a real singularity, say for $N \sim 10^6$. We shall describe here a simplified model for DNA denaturation due to Poland and Scheraga[1]. Configurations of partially melted DNA are represented in this model as as alternating sequence of double-stranded segments (rods), and single-stranded loops (bubbles).
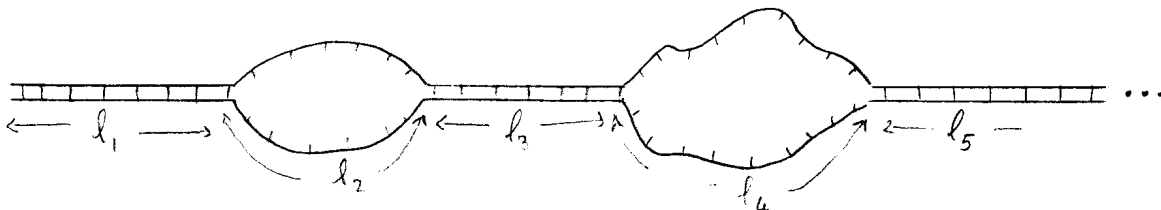
Ignoring any interactions between the segments, each configuration is assigned a probability

$$p\left(l_1, l_2, l_3, \cdots\right) = \frac{R(l_1)B(l_2)R(l_3)\cdots}{Z}\,, \tag{2.75}$$

where we have assumed that the first segment is a rod of length l_1, the second a bubble formed from two single strands of length l_2, and so on. The double stranded segments are energetically favored, but carry little entropy. To make analytical computations feasible, we shall ignore the variations in binding energy for different nucleotides, and assign an average energy $\epsilon < 0$ per double-stranded bond. (In this sense, this is a model for denaturation of a DNA homo-polymer.) The weight of a rod segment of length l is thus

$$R(l) = e^{-\beta\epsilon l} \equiv w^l, \text{ where } w = e^{-\beta\epsilon} > 1\,. \tag{2.76}$$

[1]D. Poland and H. A. Scheraga, "Phase transitions in one dimension and the helix-coil transition in polyamino acids," J. Chem. Phys. **45**, 1456 (1966).

The single-stranded portions are more flexible, and provide an entropic advantage that is modeled according to a weight similar to Eqs. (2.73-2.74), as

$$B(l) = \frac{g^l}{l^c}. \tag{2.77}$$

Clearly the above weight cannot be valid for strands shorter than a persistence length, but better describes longer bubbles.

For DNA of length L the segment lengths are constrained such that

$$l_1 + l_2 + l_3 + \cdots = L, \tag{2.78}$$

and the partition function, normalizing the weights in Eq. (2.75), is given by

$$Z(L) = \sum_{l_1, l_2, l_3, \ldots}' w^{l_1} \Omega(l_2) w^{l_3} \Omega(l_4) \cdots, \tag{2.79}$$

where the prime indicates the constraint in Eq. (2.78). As usual in statistical physics, such a global constraint can be removed by moving to a different ensemble in which the number of monomers is not fixed, but a DNA of length L is assigned an additional weight of z^L. (The quantity z is like a "fugacity," related to a chemical potential μ by $z = e^{\beta \mu}$.) In such an ensemble, the appropriate partition function is

$$\mathcal{Z}(z) = \sum_{L=1}^{\infty} z^L Z(L). \tag{2.80}$$

Since L can now take any value, we can sum over the $\{l_i\}$ independently without any constraint, to obtain

$$\mathcal{Z}(z) = \left(\sum_{l_1} z^{l_1} w^{l_1} \right) \left(\sum_{l_2} z^{l_2} \Omega(l_2) \right) \left(\sum_{l_3} z^{l_3} w^{l_3} \right) \left(\sum_{l_4} z^{l_4} \Omega(l_4) \right) \cdots. \tag{2.81}$$

The result is thus a product of alternating contributions from rods and bubbles. For each rod segment, we get a contribution

$$R(z) = \sum_{l=1}^{\infty} (zw)^l = \frac{zw}{1 - zw}, \tag{2.82}$$

while the contribution from a bubble is

$$B(z) = \sum_{l=1}^{\infty} z^l \Omega(l) = \sum_{l=1}^{\infty} \frac{z^l g^l}{l^c} \equiv f_c^+(zg). \tag{2.83}$$

The result for bubbles has been expressed in terms of the functions $f_n^+(x)$, frequently encountered in describing the ideal Bose gas in the grand canonical ensemble. We recall some

properties of these functions. First, note that taking the logarithmic derivative lowers the index by one, as

$$z\frac{df_c^+(zg)}{dz} = \sum_{l=1}^{\infty} \frac{(zg)^l}{l^{c-1}} = f_{c-1}^+(zg) \ . \tag{2.84}$$

Second, each $f_n^+(x)$ is an increasing function of its argument, and convergent up to $x = 1$, at which point

$$f_c^+(1) \equiv \zeta_c \ , \tag{2.85}$$

where ζ_n is the well-known Riemann zeta-function. The zeta-function is well behaved for $c > 1$, and indeed for $c < 1$, $f_c^+(x)$ diverges is $(1-x)^{c-1}$ for $x \to 1$.[2]

Next, we must sum over all possible numbers of bubbles in between two rod segments as end points, leading to

$$\mathcal{Z}(z) = R(z) + R(z)B(z)R(z) + R(z)B(z)R(z)B(z)R(z) + \cdots \ . \tag{2.86}$$

This is a just geometric series, easily summed to

$$\mathcal{Z}(z) = \frac{R(z)}{1 - R(z)B(z)} = \frac{1}{R^{-1}(z) - B(z)} = \frac{1}{(zw)^{-1} - 1 - f_c^+(zg)} \ . \tag{2.87}$$

The logarithm of the sum provides a useful thermodynamic free energy,

$$\log \mathcal{Z}(z) = -\log\left[\frac{1}{zw} - 1 - f_c^+(zg)\right] \ , \tag{2.88}$$

from which we can extract physical observables. For example, while the length L is a random variable in this ensemble, for a given z, its distribution is narrowly peaked around the expectation value

$$\langle L \rangle = z\frac{\partial}{\partial z}\log \mathcal{Z}(z) = \frac{\frac{1}{zw} + f_{c-1}^+(zg)}{\frac{1}{zw} - 1 - f_c^+(zg)} \ . \tag{2.89}$$

We can also compute the fraction of the polymer that is in the denatured state. Since each double-strand bond contributes a factor w to the weight, the number of bound pairs N_B has a mean value

$$\langle N_B \rangle = w\frac{\partial}{\partial w}\log \mathcal{Z}(z) = \frac{\frac{1}{zw}}{\frac{1}{zw} - 1 - f_c^+(zg)} \ . \tag{2.90}$$

Taking the ratio of N_B and L gives the fraction of the polymer in the native state as

$$\Theta = \frac{\langle L \rangle}{\langle N_B \rangle} = \frac{1}{1 + zwf_{c-1}^+(zg)} \ . \tag{2.91}$$

---

[2]Furthering the mathematical analogy between DNA melting and Bose-Einstein condensation, note that when the bubble is treated as a random walk, $c = d/2$, implying that $B(z)$ is only finite for $d \le 2$. Indeed, $d = 2$ is also a critical dimension for Bose-Einstein condensation.

Equation (2.91) is not particularly illuminating in its current form, because it gives Θ in terms of z, which we introduced as a mathematical device for removing the constraint of length in the partition function. For meaningful physical results we need to solve for z as a function of L by inverting Eq. (2.90). This task is simplified in the thermodynamic limit where $L, N_B \to \infty$, while their ratio is finite. From Eqs. (2.90-2.91), we see that this limit is obtained by setting the denominator in these expressions equal to zero, i.e. from the condition

$$f_c^+(zg) = \frac{1}{zw} - 1\,. \tag{2.92}$$

The type of phase behavior resulting from Eqs. (2.92-2.91), and the very existence of a transition, depend crucially on the parameter c, and we can distinguish between the following three cases:

(a) For $c < 1$, the function $f_c^+(zg)$ goes to infinity at $z = 1/g$. The right hand side of Eq. (2.92) is a decreasing function of z that goes to zero at $z = 1/w$. We can graphically solve this equation by looking for the intersection of the curves representing these functions. As temperature goes up, $1/w = e^{\beta\epsilon}$ increases towards unity, and the intersection point moves to the right. However, there is no singularity and a finite solution $z < 1/g$ exists at all temperatures. This solution can then be substituted into Eq. (2.91) resulting in a native fraction that decreases with temperature, but never goes to zero. There is thus no denaturation transition in this case.
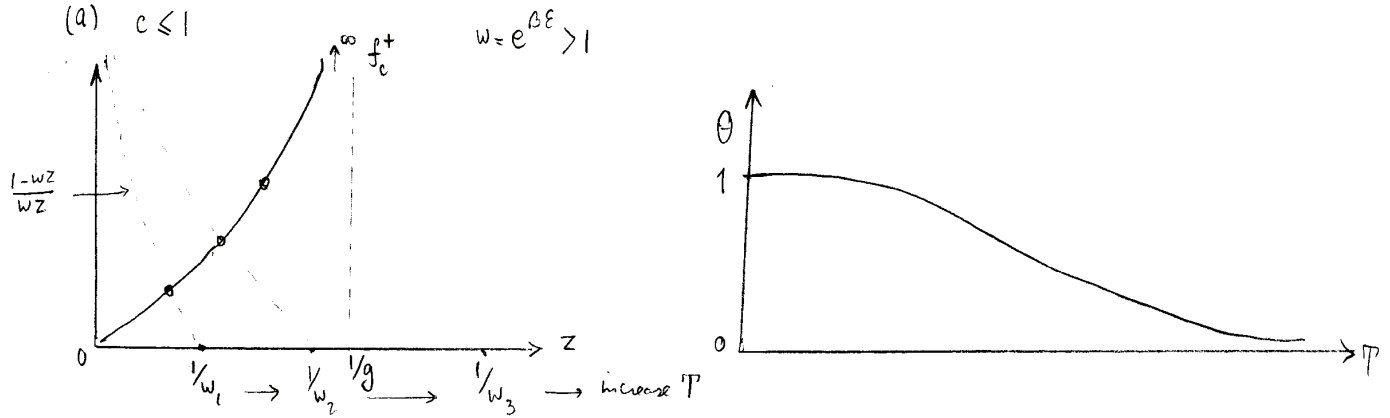
(b) For $1 \leq c \leq 2$, the function $f_c^+(zg)$ reaches a finite value of ζ_c at $zg = 1$. The two curves intersect at this point for $z_c = 1/g$ and $w_c = g/(1 + \zeta_c)$. For all values of $w \leq w_c$, z remains fixed at $1/g$. The derivative of $f_c^+(zg)$, proportional to $f_{c-1}^+(zg)$ from Eq. (2.84), diverges as its argument approaches unity, such that

$$f_c^+(zg) - \zeta_c \propto (1 - zg)^{c-1}\,. \tag{2.93}$$

From the occurrence of $f_{c-1}^+(zg)$ in the denominator of Eq. (2.91), we observe that Θ is zero for $w \leq w_c$, i.e. the polymer is fully denatured. On approaching the transition point from

the other side, Θ goes to zero continuously. Indeed, Eq. (2.93) implies that a small change $\delta w \equiv w - w_c$ is accompanied by a much smaller change in z, such that $\delta z \equiv (z_c - z) \propto (\delta w)^{\frac{1}{c-1}}$. Since $f_{c-1}^+(zg) \propto (1 - zg)^{c-2}$, we conclude from Eq. (2.91) that the native fraction goes to zero as

$$\Theta \propto (\delta z)^{2-c} \propto (w - w_c)^\beta, \quad \text{with} \quad \beta = \frac{2 - c}{c - 1}. \tag{2.94}$$

For a loop treated as a random walk in three dimensions, $c = 3/2$ and $\beta = 1$, i.e. the

(b) $\quad 1 \leqslant c \leqslant 2$

$(\zeta_{3/2} = 2.612\ldots)$
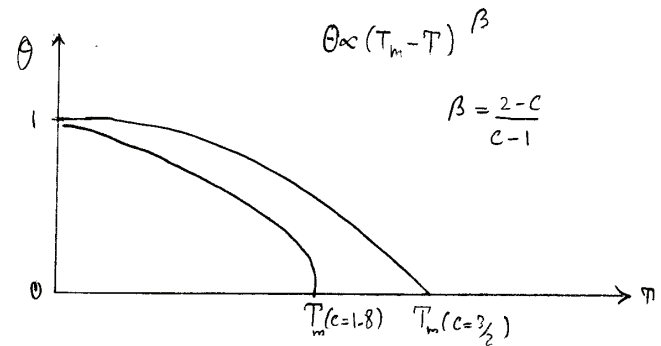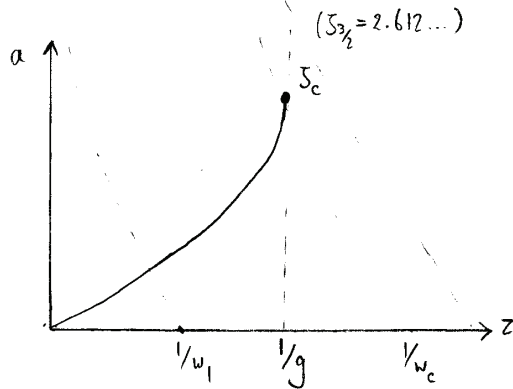
denatured fraction disappears linearly. Including self-avoidance with $c = 3\nu \approx 1.8$ leads to $\beta \approx 1/4$ and a much sharper transition.

(c) For $c > 2$, the function $f_{c-1}^+(zg)$ approaches a finite limit of ζ_{c-1} at the transition point. The transition is now discontinuous, with Θ jumping to zero from $\Theta_c = (1 + \zeta_c)/(1 + \zeta_c + \zeta_{c-1})$. Including the effects of self-avoidance within a single loop increases the value of c from 1.5 to

(c) $\quad c > 2$
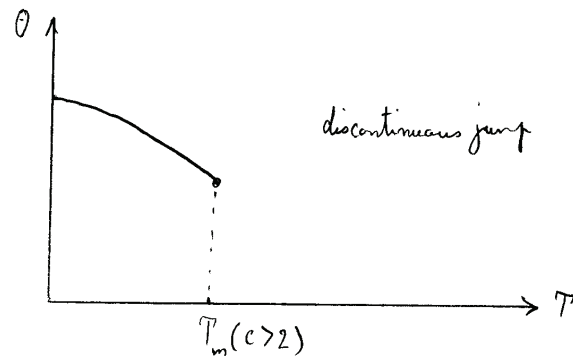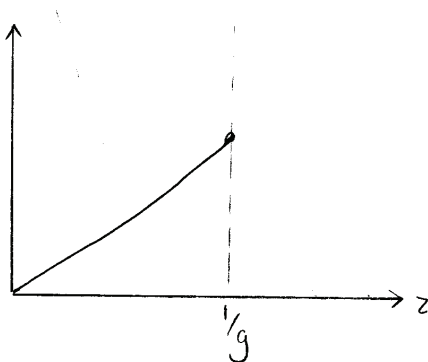
1.8. In reality there are additional effects of excluded volume between the different segments. It has been argued that including the interactions between the different segments (single and

53

double-strands) further increases the value of c to larger than 2, favoring a discontinuous melting transition.[3]

A justification of the role of the exponent c in controlling the nature/existence of the phase transition can be gleaned by examining the behavior of a single bubble. Examining the competition between entropy and energy suggests that the probability (weight) of a loop of length $\ell = 2l$ is proportional to

$$p(\ell) \propto \left(\frac{g}{w}\right)^{\ell} \times \frac{1}{\ell^c} . \tag{2.95}$$

The probability broadens to include larger values of ℓ as $(g/w) \to 1$.

(a) For $c < 1$, the above probability cannot be normalized if arbitrarily large values of ℓ are included. Thus at any ration of (g/w), the probability has to be cut-off at some maximum ℓ, and the typical size of a loop remains finite.

(b) For $1 \leq c \leq 2$ the probability can indeed be normalized including all values of ℓ (the normalization is $f_c^+(g/w)$), but the average size of the loop (related to $f_{c-1}^+(g/w)$) diverges as $(g/w) \to 1$ signaling a continuous phase transition.

(c) For $c > 2$, the probability is normalizable, and the loop size remains finite as $(g/w) \to 1$. There is a limiting loop size at the transition point suggesting a discontinuity.

[3]Y. Kafri, D. Mukamel, and L. Peliti, Phys. Rev. Lett. **85**, 4988 (2000).

8.592J / HST.452J Statistical Physics in Biology
Spring 2011