2.3 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 $R_g \approx \sqrt{2L \cdot \xi_p}$, 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 a double helix is formed by formation of Watson–Crick pairs, G–C and A–T. There are 3 hydrogen bonds in a GC pairing and two per an AT pair. While GC rich portions of DNA are more strongly bonded, it is not because of the above difference in number of hydrogen bonds, but due to increase in stacking energies (with a binding energy of around $4k_BT$ for AT stacks, roughly twice that for GC stacking). At finite temperatures, this energy gain competes with the loss of entropy that comes with the braiding of 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 separated. Regions of DNA that are rich in A–T open up at lower temperatures, those with high G–C content at higher temperatures. Such unbinding 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

$$S(l) \approx bl + c \log l + d\,, \quad \text{with} \quad c \approx 1.8 k_B\,. \tag{2.72}$$

The leading linear term in l is a measure of the gain in entropy per base pair, while the subleading logarithmic dependence is a consequence of *loop closure*, and can be justified as follows: A bubble is composed of two single stranded segments of length l, with start and end positions on the double strand. First we sum over all configurations of these two segments, *assuming* that the two end points are separated by a distance $\vec{r}$. Regarding each segment as a non-interacting random walk of length l and end-to-end separation $\vec{r}$, the number of configurations is easily obtained by appropriate extension of Eq. (2.40) to

$$W_{\text{loop}}(\vec{r}, 2l) = W(\vec{r}, l)^2 = g_1^{2l} \exp\left[-\frac{2dr^2}{4l\xi_p}\right] \frac{1}{(4\pi l \xi_p/d)^d}\,, \tag{2.73}$$

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

of the intermediate point as

$$\Omega(l) = \int d^d\mathbf{r} W_{\rm loop}(\vec{r}, 2l) = \left(\frac{d}{8\pi\xi_p}\right)^{d/2} \frac{g^l}{l^c}\,, \tag{2.74}$$

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

For the more realistic case of self-avoiding polymers, a naive scaling argument (ignoring interactions between segments) suggests

$$W_{\rm loop}(\vec{r}, 2l) = \frac{g^l}{R^d}\,\Phi\left(\frac{\vec{r}}{R}\right)\,, \quad \text{with} \quad R \sim l^\nu\,, \quad \text{and} \quad \Omega(l) \propto \frac{g^l}{l^{d\nu}}\,. \tag{2.75}$$

We can justify 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.

2.3.1 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 form fully bound to unbound occurs over a temperature interval that becomes narrower for 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[6]. Configurations of partially melted DNA are represented in this model as an alternating sequence of double-stranded segments (rods), and single-stranded loops (bubbles).

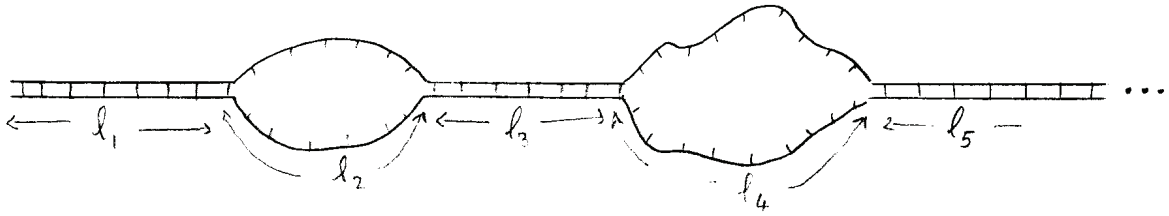

Figure 15: Partially denatured DNA as a sequence of bubbles and rods.

Ignoring any interactions between the segments, each configuration is assigned a probability

$$p\,(l_1, l_2, l_3, \cdots) = \frac{R(l_1)B(l_2)R(l_3)\cdots}{Z}\,, \tag{2.76}$$

[6]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).

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

$$R(l) = e^{-\beta\epsilon l} \equiv w^l, \text{ where } w = e^{-\beta\epsilon} > 1\,. \tag{2.77}$$

The single-stranded portions are flexible, and provide an entropic advantage that is modeled according to a weight similar to Eqs. (2.74-2.75), as

$$B(l) = a\frac{g^l}{l^c}\,, \tag{2.78}$$

where the parameter a incorporates the normalization of loop probability, as in Eq. (2.74), and more importantly energetic costs associated with opening up a bubble the edges joining double strands. Clearly the above weight cannot be valid for strands shorter than a persistence length, but better describes longer bubbles. As we shall see, a is irrelevant to the sharpness of the denaturation transition, although it does determine its temperature.

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

$$l_1 + l_2 + l_3 + \cdots = L\,, \tag{2.79}$$

and the partition function, normalizing the weights in Eq. (2.76), is given by

$$Z(L) = {\sum_{l_1,l_2,l_3,\ldots}}' w^{l_1}\Omega(l_2)w^{l_3}\Omega(l_4)\cdots\,, \tag{2.80}$$

where the prime indicates the constraint in Eq. (2.79). The passage from canonical to grand canonical ensemble exemplifies a typical transformation from statistical physics in which a global constraint (the number of particles) is removed by introducing a conjugate variable (chemical potential). It is similarly convenient here to consider an ensemble of DNA of variable length L, obtained by assigning weights z^L to segments of length L. (The quantity z, sometimes called a "fugacity" is related to a chemical potential μ for basepairs by $z = e^{\beta\mu}$.) In such an ensemble, the appropriate (grand) partition function is

$$\mathcal{Z}(z) = \sum_{L=1}^{\infty} z^L Z(L)\,. \tag{2.81}$$

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

$$\mathcal{Z}(z) = \left(\sum_{l_1} z^{l_1}w^{l_1}\right)\left(\sum_{l_2} z^{l_2}\Omega(l_2)\right)\left(\sum_{l_3} z^{l_3}w^{l_3}\right)\left(\sum_{l_4} z^{l_4}\Omega(l_4)\right)\cdots\,. \tag{2.82}$$

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

$$R(z) = \sum_{l=1}^{\infty} (zw)^l = \frac{zw}{1-zw}\,, \tag{2.83}$$

while the contribution from a bubble is

$$B(z) = a\sum_{l=1}^{\infty} z^l \Omega(l) = a\sum_{l=1}^{\infty} \frac{z^l g^l}{l^c} \equiv a f_c^+(zg)\,. \tag{2.84}$$

The result for bubbles has been expressed in terms of the special 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

$$z\frac{df_c^+(zg)}{dz} = \sum_{l=1}^{\infty} \frac{(zg)^l}{l^{c-1}} = f_{c-1}^+(zg)\ . \tag{2.85}$$

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

$$f_c^+(1) \equiv \zeta_c\,, \tag{2.86}$$

where ζ_c is the well-known Riemann zeta-function. The zeta-function is well behaved only for $c > 1$, and indeed for $c < 1$, $f_c^+(x)$ diverges is $(1-x)^{c-1}$ for $x \to 1$.[7]

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

$$\mathcal{Z}(z) = R(z) + R(z)B(z)R(z) + R(z)B(z)R(z)B(z)R(z) + \cdots\,. \tag{2.87}$$

This is a just geometric series, easily summed to

$$\mathcal{Z}(z) = \frac{R(z)}{1-R(z)B(z)} = \frac{1}{R^{-1}(z) - B(z)} = \frac{1}{(zw)^{-1} - 1 - a f_c^+(zg)}\,. \tag{2.88}$$

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

$$\log \mathcal{Z}(z) = -\ln\left[\frac{1}{zw} - 1 - a f_c^+(zg)\right]\,, \tag{2.89}$$

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

$$\langle L\rangle = z\frac{\partial}{\partial z}\ln \mathcal{Z}(z) = \frac{\frac{1}{zw} + ag f_{c-1}^+(zg)}{\frac{1}{zw} - 1 - a f_c^+(zg)}\,. \tag{2.90}$$

[7]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 $d \leq 2$. Indeed, $d = 2$ is also the lower critical dimension for occurrence of Bose-Einstein condensation.

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

$$\langle N_B \rangle = w \frac{\partial}{\partial w} \ln \mathcal{Z}(z) = \frac{\frac{1}{zw}}{\frac{1}{zw} - 1 - a f_c^+(zg)} \,. \tag{2.91}$$

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

$$\Theta = \frac{\langle N_B \rangle}{\langle L \rangle} = \frac{1}{1 + zwag \; f_{c-1}^+(zg)} \,. \tag{2.92}$$

Equation (2.92) 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 fixed length in the partition function. For meaningful physical results we need to solve for z as a function of L by inverting Eq. (2.91). This task is simplified in the thermodynamic limit where $L, N_B \to \infty$, while their ratio is finite. From Eqs. (2.91-2.92), we see that this limit is obtained by setting the denominator in these expressions equal to zero, i.e. from the condition

$$a f_c^+(zg) = \frac{1}{zw} - 1 \,. \tag{2.93}$$

The type of phase behavior resulting from Eqs. (2.93-2.92), and the very existence of a transition, depends 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. (2.93) 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. (2.92) 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 $w_c = g/(1 + a\zeta_c)$. 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. (2.85), diverges as its argument approaches unity, such that

$$f_c^+(zg) - \zeta_c \propto (1 - zg)^{c-1} \,. \tag{2.94}$$

From the occurrence of $f_{c-1}^+(zg)$ in the denominator of Eq. (2.92), 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. (2.94) 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. (2.92) that the native fraction goes to zero as

$$\Theta \propto (\delta z)^{2-c} \propto (w - w_c)^{\beta} \,, \quad \text{with} \quad \beta = \frac{2-c}{c-1} \,. \tag{2.95}$$

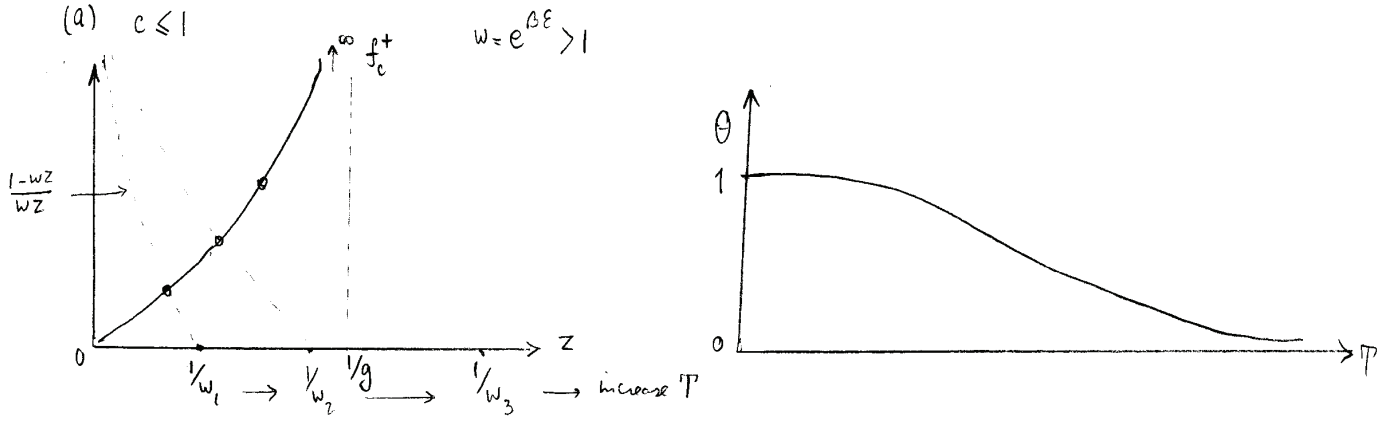

Figure 16: Graphical solution for $c \leq 1$.

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

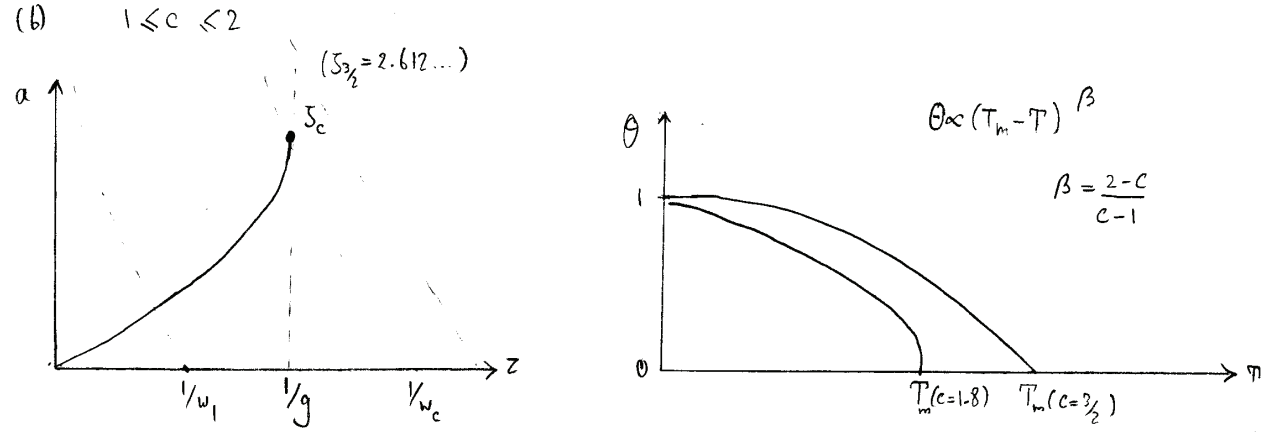

Figure 17: Graphical solution for $1 < c < 2$.

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 a finite value of $\Theta_c = (1+a\zeta_c)/(1+a\zeta_c+ag\zeta_{c-1})$. Including the effects of self-avoidance within a single loop increases the value of c from 1.5 to 1.8. In reality there are additional effects of excluded volume between the different segments. It has been argued that including interactions between the different segments (single and double-strands) further increases the value of c to larger than 2, favoring a discontinuous melting transition.[8]

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

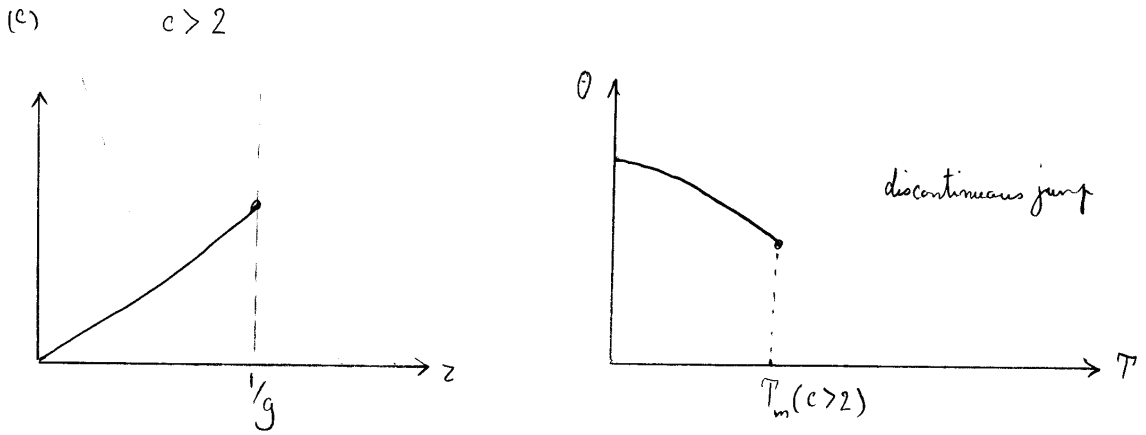

Figure 18: Graphical solution for $c \geq 2$.

A justification of the role of the exponent c in controlling the nature/existence of the phase transition can be gleaned by considering 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

$$p(\ell) \propto \left(\frac{g}{w}\right)^{\ell} \times \frac{1}{\ell^c}\,. \tag{2.96}$$

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 ratio 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 discontinuous jump.

Note that the the loop initiation factor a does not affect the argument.