



where noting that  $N_B = (z - 2)N$  (of the  $z$  nearest neighbors for each site of the lattice, two connect subsequent monomers along the chain), we have folded the proportionality constant into the definitions of  $\varepsilon_0$  and  $\sigma^2$ .

For large  $N$ , the probability distribution for the energy will take the Gaussian form

$$p(E) = \frac{1}{\sqrt{2\pi N\sigma^2}} \exp\left[-\frac{(E - N\varepsilon_0)^2}{2N\sigma^2}\right]. \quad (2.61)$$

Since the total number of states is  $g'^N$ , the density of states is  $\Omega(E) = g'^N p(E)$ , and the entropy of this *random energy model* (REM) is given by

$$S(E) = k_B \ln \Omega(E) = k_B \left[ N \ln g' - \frac{(E - N\varepsilon_0)^2}{2N\sigma^2} \right] - \frac{k_B}{2} \ln(2\pi N\sigma^2). \quad (2.62)$$

The last term is not extensive (proportional to  $N$ ) and can be safely ignored.

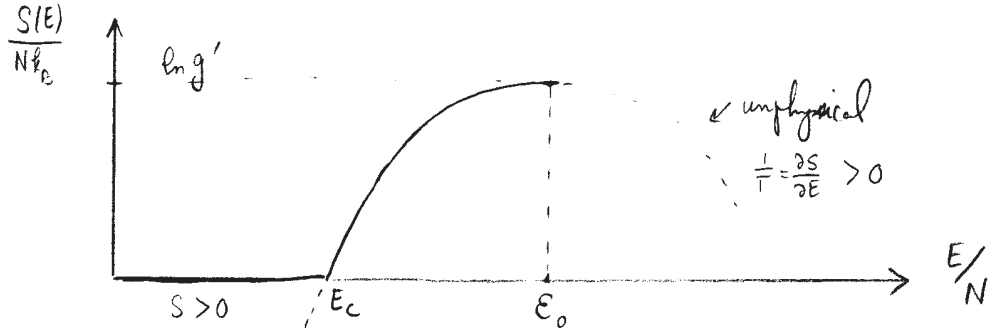


Figure 12: Entropy of a random energy model.

According to Eq. (2.62),  $S(E)$  is shaped like a parabola, but thermodynamic constraints imply that only a certain portion of this curve is physical. First, the temperature  $T$  is obtained from the slope of the curve via  $T^{-1} = dS/dE$ . Positive temperatures require the entropy to increase with temperature, and thus only the states with  $E < N\varepsilon_0$  are physically accessible. Second, the entropy cannot be negative, and  $S(E)$  should thus stick to zero for  $E < E_c$ , where  $E_c$  is easily obtained as

$$S(E_c) = 0 \implies \frac{E_c}{N} = \varepsilon_0 - \sigma \sqrt{2 \ln g'}. \quad (2.63)$$

(Note the connection to the extreme value problem studied earlier:  $E_c$  is also the mean value of the lowest of  $g'^N$  energies randomly selected from  $p(E)$ .) The singularity of entropy at  $E_c$  signifies a phase transition into a glassy state, at a temperature  $T_c$  given by

$$\frac{1}{T_c} = \left. \frac{dS}{dE} \right|_{E_c} = -k_B \left( \frac{E_c - N\varepsilon_0}{N\sigma^2} \right) = k_B \frac{\sqrt{2 \ln g'}}{\sigma} \implies k_B T_c = \frac{\sigma}{\sqrt{2 \ln g'}}. \quad (2.64)$$

There are presumably a few low energy states with energy close to  $E_c$ , and the system freezes into one of these for  $T \leq T_c$ .

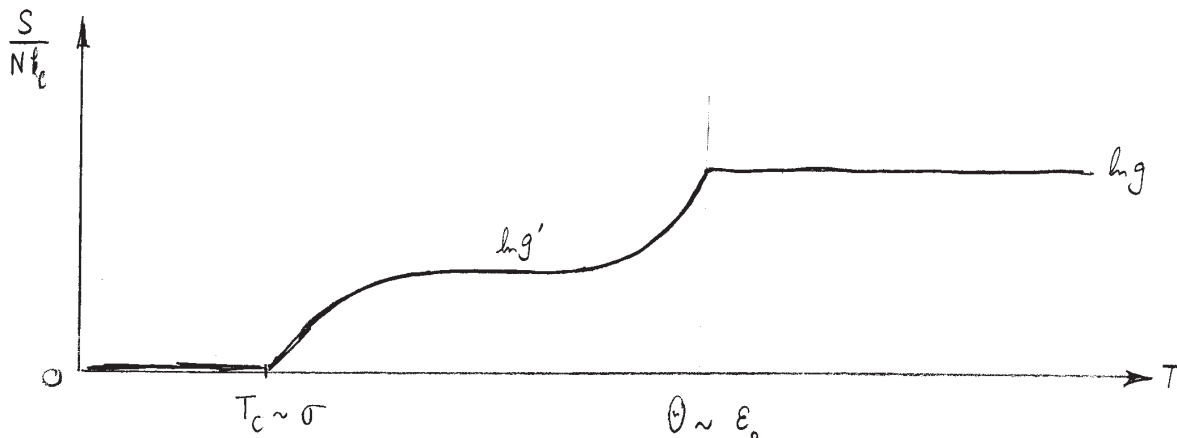


Figure 13: Schematic depiction of entropy for a random heteropolymer.

### 2.3.5 Designed REM for protein folding

It is tempting to equate the freezing of the heteropolymer with the folding transition separating denatured and folded states of a protein. There is, however, a problem with such an interpretation: As the temperature is lowered towards  $T_c$ , the number of states decreases drastically. It is unlikely that the lower energy configurations of the REM polymer in the vicinity of  $E_c$  have much in common. To change its state, the polymer will likely have to rearrange many of its monomers, running into high energy barriers in the process. Thus we expect that the kinetics of the REM polymer will slow down significantly on approaching  $T_c$ . This contradicts the observation that most proteins fold easily and in a short time. Of course proteins are not typical random heteropolymers, and are presumably “designed” through evolution for both function and ease of folding. Fortunately, we can mimic such “design” by a small modification of the REM; we only need to add to the continuum of random energy states, a single state with low energy ( $E_n < E_c$ ) representing the native configuration.

With the added state at  $E_n$ , the system makes a transition to the native configuration (i.e. folds) at a temperature  $T_f$ , high enough that there are still many equivalent states to explore. The location of  $T_f$ , and the corresponding energy  $E_f$ , can be obtained by equating free energies or Boltzmann weights, and leads to the “tangent construction” whereby  $T_f$  and  $E_f$  are related to  $E_n$  via

$$\beta_f = \frac{S(E_f)/k_B}{E_f - E_n} = \frac{N \ln g' - (E_f - N\epsilon_0)^2/(2N\sigma^2)}{E_f - E_n}. \quad (2.65)$$

As depicted in the figure, the above result equates the slope of the tangent line from the point at  $E_n$  computed in two different ways.

To justify the above result, note that in the canonical ensemble, the probability of finding

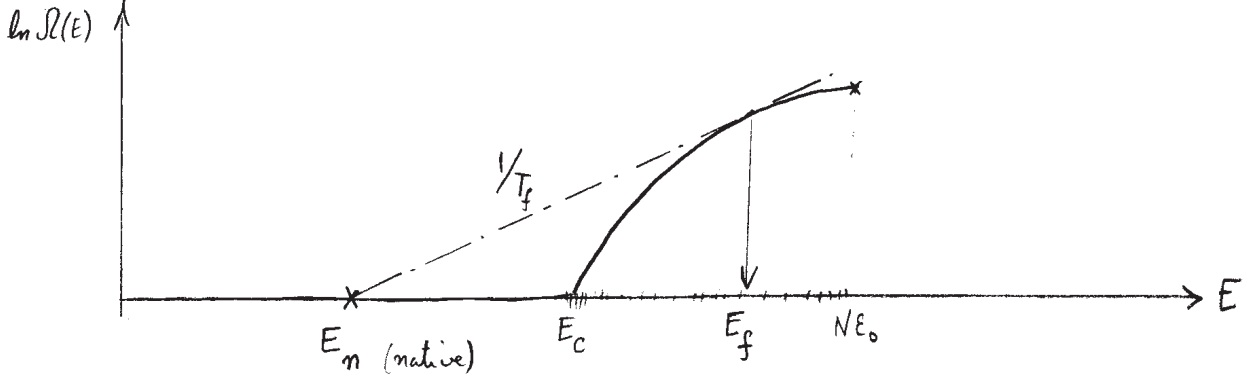


Figure 14: Tangent construction for a designed random energy model.

the system in the native state is

$$p_n = \frac{e^{-\beta E_n}}{Z(\beta)}, \quad \text{with} \quad Z(\beta) = e^{-\beta E_n} + \int dE \Omega(E) e^{-\beta E}. \quad (2.66)$$

A phase transition in which  $p_n$  changes discontinuously from zero to one occurs only in the thermodynamic limit of  $N \rightarrow \infty$ . For the system to have a well-behaved thermodynamic limit (in which case various thermodynamic identities involving entropy and temperature can be safely used), we must insist that the range of energies as well as  $\ln \Omega(E)$  should be proportional to  $N$ ; the former implies that  $E_n \propto N$ . If so, then at a particular value of  $\beta$  a single value of energy  $E$  completely dominates the partition function  $Z(\beta)$ . For the partition function in Eq. (2.66), the dominant value occurs for some  $E \geq E_f$  for  $\beta \leq \beta_f$ , and for  $E = E_n$  for  $\beta > \beta_f$ . The probability to find the system in its native state then jumps discontinuously from 0 to 1 at the point when the corresponding contributions to the partition function are equal, i.e. at

$$e^{-\beta_f E_n} = \Omega(E_f) e^{-\beta_f E_f}, \quad (2.67)$$

which after taking the logarithm leads to the tangent rule in Eq. (2.65).

We can eliminate  $E_f$  in terms of  $\beta_f$  by noting that  $E = N\varepsilon_0 - N\sigma^2\beta$ , and  $\ln g' = (\beta_c\sigma)^2/2$ . Using these expressions and defining a quantity  $\beta_n = (E_n - N\varepsilon_0)/(N\sigma^2)$ , the above equation reduces to

$$\beta_f = \frac{\beta_c^2 - \beta_f^2}{-2\beta_f + 2\beta_n}. \quad (2.68)$$

This can be rearranged as a quadratic equation with solution

$$\beta_f = \beta_n - \sqrt{\beta_n^2 - \beta_c^2}. \quad (2.69)$$

The ratio of the folding temperature to the REM freezing temperature is thus

$$\frac{T_f}{T_c} = \frac{\beta_c}{\beta_f} = \frac{\beta_n}{\beta_c} + \sqrt{\left(\frac{\beta_n}{\beta_c}\right)^2 - 1}. \quad (2.70)$$

Faster folding to the native state can be achieved at higher temperatures by increasing the energy difference between  $E_n$  and  $E_c$ .

## 2.4 DNA structure

DNA molecules come in a wide range of length scales, from roughly 50,000 monomers in a  $\lambda$ -phage,  $6 \times 10^9$  for human, to  $9 \times 10^{10}$  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  $4k_B T$ ) 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  $80^\circ\text{C}$  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

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

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

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

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

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

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

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

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 bringing 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<sup>1</sup>. Configurations of partially melted DNA are represented in this model as an alternating sequence of double-stranded segments (rods), and single-stranded loops (bubbles).

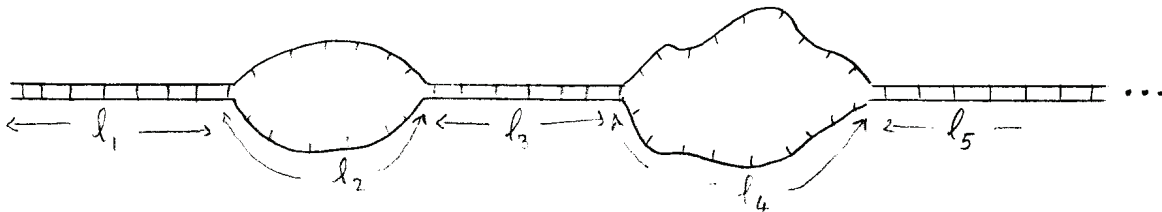


Figure 15: Partially denatured DNA as a sequence of bubbles and rods.

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

$$p(l_1, l_2, l_3, \dots) = \frac{R(l_1)B(l_2)R(l_3)\dots}{Z}, \quad (2.75)$$

<sup>1</sup>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

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

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

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

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 individual segment lengths are constrained such that

$$l_1 + l_2 + l_3 + \dots = L, \quad (2.78)$$

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

$$Z(L) = \sum'_{l_1, l_2, l_3, \dots} w^{l_1} \Omega(l_2) w^{l_3} \Omega(l_4) \dots, \quad (2.79)$$

where the prime indicates the constraint in Eq. (2.78). The passage from canonical to grand canonical ensemble in a typical example 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  $L$ , distinguished by assigning a weights of  $z^L$ . (The quantity  $z$ , sometimes called a ‘‘fugacity’’ is related to a chemical potential  $\mu$  for basepairs by  $z = e^{\beta\mu}$ .) In such an ensemble, the appropriate (grand) partition function is

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

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

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

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

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

while the contribution from a bubble is

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

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

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

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

$$f_c^+(1) \equiv \zeta_c, \quad (2.85)$$

where  $\zeta_n$  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 as  $(1-x)^{c-1}$  for  $x \rightarrow 1$ .<sup>2</sup>

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

$$\mathcal{Z}(z) = R(z) + R(z)B(z)R(z) + R(z)B(z)R(z)B(z)R(z) + \dots. \quad (2.86)$$

This is a just geometric series, easily summed to

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

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

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

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

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

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

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

---

<sup>2</sup>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 a critical dimension for Bose-Einstein condensation.



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

$$\Theta = \frac{\langle L \rangle}{\langle N_B \rangle} = \frac{1}{1 + zwgf_c^+(zg)}. \quad (2.91)$$

Equation (2.91) is not particularly illuminating in its current form, because it gives  $\Theta$  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. (2.90). This task is simplified in the thermodynamic limit where  $L, N_B \rightarrow \infty$ , 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

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

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.

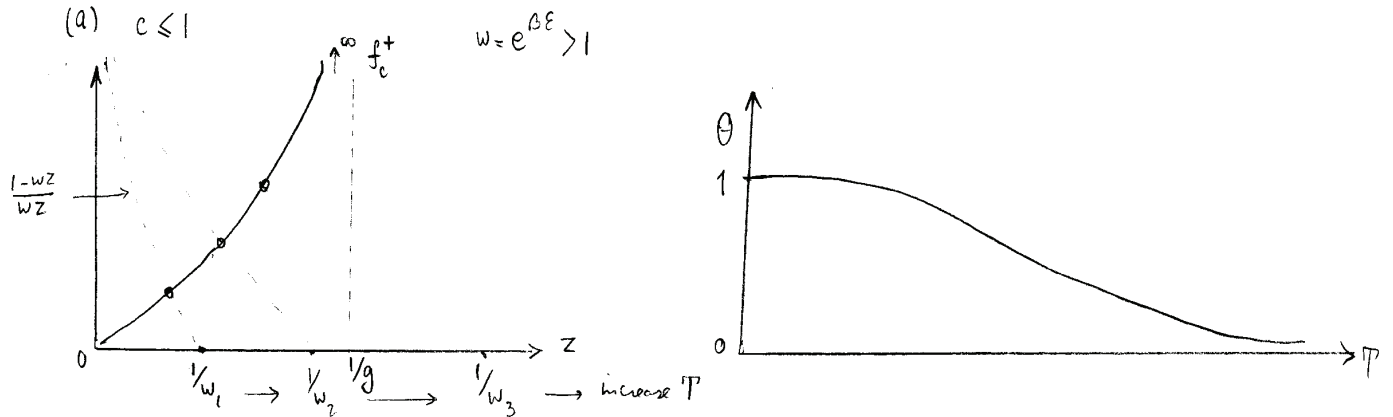


Figure 16: Graphical solution for  $c \leq 1$ .

**(b)** For  $1 \leq c \leq 2$ , the function  $f_c^+(zg)$  reaches a finite value of  $\zeta_c$  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

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

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

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

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

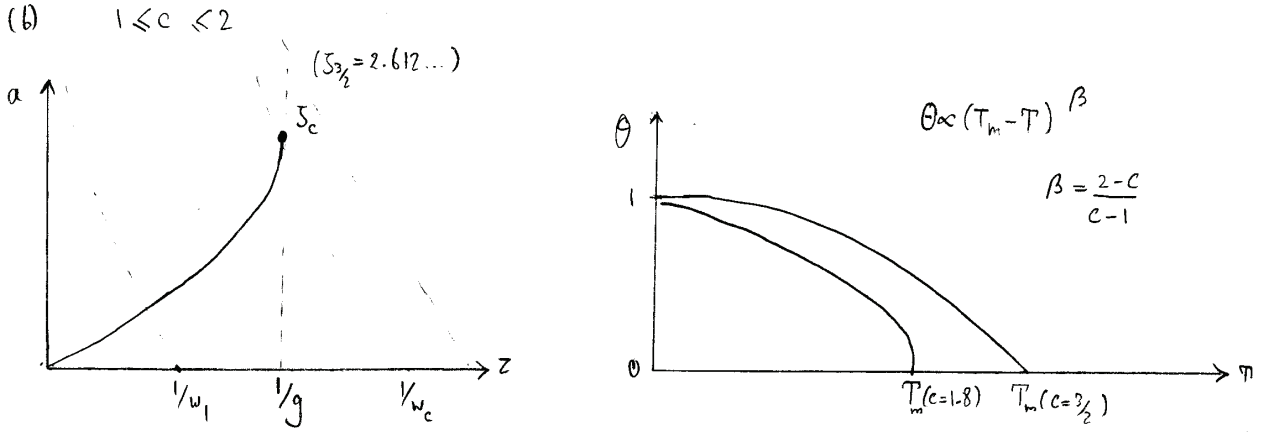


Figure 17: Graphical solution for  $1 < c < 2$ .

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  $\zeta_{c-1}$  at the transition point. The transition is now discontinuous, with  $\Theta$  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 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.<sup>3</sup>

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

<sup>3</sup>Y. Kafri, D. Mukamel, and L. Peliti, Phys. Rev. Lett. **85**, 4988 (2000).

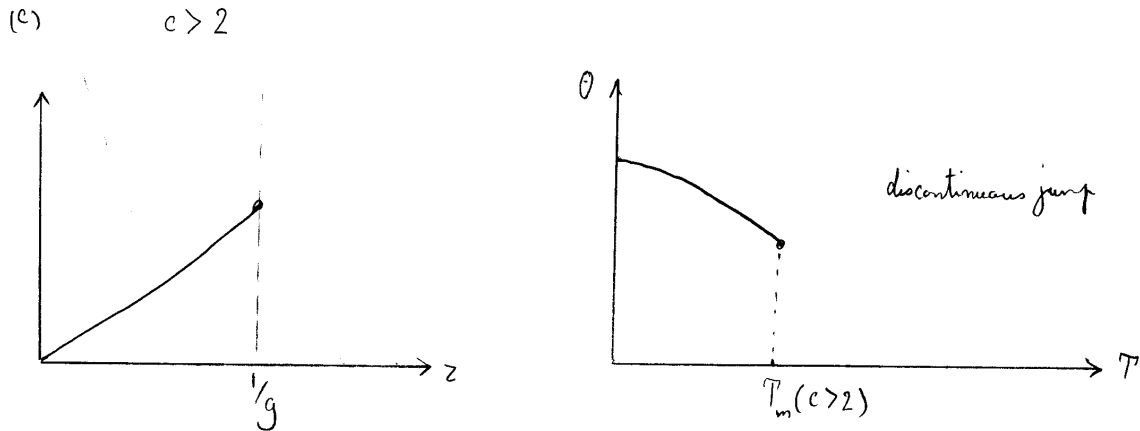


Figure 18: Graphical solution for  $c \geq 2$ .

the competition between entropy and energy suggests that the probability (weight) of a loop of length  $\ell = 2l$  is proportional to

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

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

**(a)** For  $c < 1$ , the above probability cannot be normalized if arbitrarily large values of  $\ell$  are included. Thus at any ratio of  $(g/w)$ , the probability has to be cut-off at some maximum  $\ell$ , 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  $\ell$  (the normalization is  $f_c^+(g/w)$ ), but the average size of the loop (related to  $f_{c-1}^+(g/w)$ ) diverges as  $(g/w) \rightarrow 1$  signaling a continuous phase transition.

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

## 2.5 RNA structure

Like DNA, RNA is a hetero-polymer composed from nucleotides, each consisting of sugar, a phosphate and a nucleic acid base. The sugar ribose in RNA has one more OH group, compared to deoxyribose in DNA. The four distinct bases in DNA are adenine (A), guanine (G), cytosine (C) and thymine (T), while in RNA uracil (U) takes the place of thymine. While the central role of DNA is storage of genetic information, RNA molecules carry a variety of roles from structural (as in the protein building machinery of ribosome) to information transfer (in messenger RNA). Concomitant with their diverse roles, the structure of RNA molecules is more complicated and they can assume a variety of shapes. An important distinction to DNA that enables such diversification is that RNA is a single stranded molecule. While two complementary strands of DNA wrap around each other to form a stable and

relatively rigid molecule, the single strand of RNA is more flexible. The molecule can fold upon itself bringing bases far apart along the backbone of the molecule close enough to form complimentary Watson-Crick pairs. While the primary structure refers to the sequence of bases along RNA, its secondary structure indicates the bases that come into contact to form complimentary bonds. The thus connected macromolecule then assumes particular shape(s) in three dimensions, known as its tertiary structure.

Given the sequence of RNA, can one predict its secondary structure? In principle one should list all possible pairings, compute their energies say by adding specified energies for the different Watson-Crick pairings), and select the lowest energy ones (or with appropriate Boltzmann weights). For  $N$  base pairs there a maximum of  $N/2$  base pairing, which (ignoring constraints) can occur in  $(N-1)(N-3) \cdots = (N-1)!!$  possible ways. Including the possibility that some bases are unpaired will increase the above number of states even further. However, for large  $N$ ,  $(N-1)!! \approx (N/e)^{N/2}$  which far exceeds the total number of possible arrangements of a polymer, which as we have seen grows at most as  $g^N$ . Thus a large fraction of pairings is excluded by steric constraints of foldability into a viable three dimensional structure. Although the number of foldable states still grows quite rapidly for large  $N$ , there are a number of algorithms that perform this computational task *for specific subsets of pairings* in polynomial time. A particularly convenient subset of pairing is that of planar graphs for which the RNA backbone, and all secondary connections can be drawn on a two dimensional plane, without any two lines crossing. Secondary connections that violate planarity lead to three dimensional structures containing elements called *pseudoknots* which are very rare (though not impossible) in actual RNAs. Thus limiting the search to this subset is not too severe a restriction.

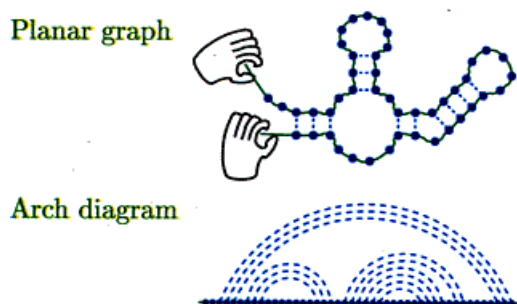


Figure 19: RNA secondary structure without pseudoknots represented as a planar graph or arch diagram.

The advantage of the planar subset of pairings is that it can be represented in multiple ways, and importantly enables finding the optimal configuration in polynomial time. One simple representation, indicated above, is obtained by stretching the RNA along a straight line and connecting the paired monomers by arches. Two arches are either disconnected, or one is entirely enclosed by the other— the arches will not intersect for planar graphs. Another representation is in terms of parentheses: Starting from one end of the RNA sequence, an open parenthesis is placed when the first nucleotide of pair is encountered, the parenthesis

is closed when its partner is encountered. A planar diagram will then correspond to a grammatically correct string. The latter representation then yields a useful graphical prescription as a random walk: Moving along the sequence an up step indicates a parenthesis opened, a down step one that is closed. The planar diagram is now depicted as an island or mountain landscape with no segments where the height is negative.

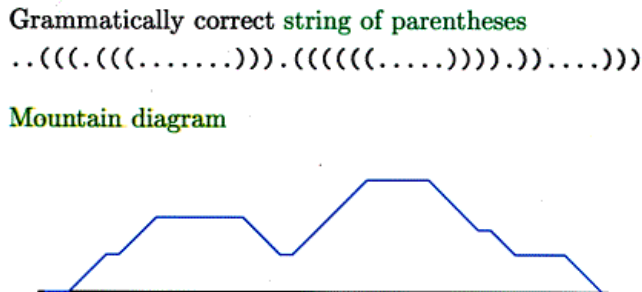


Figure 20: RNA secondary structure without pseudoknots represented with parenthesis, and as a mountain landscape.

Let us consider a simple model for secondary structures in which *all pairings* without pseudo-knots are allowed (i.e. without consideration of bending or steric constraints). For each configuration  $C$ , the energy is the sum over energies assigned to all bonded pairs, i.e.  $E[C] = \sum_{\langle ij \rangle} \varepsilon_{ij}$ , where  $\varepsilon_{ij}$  is the energy of the bond between monomers  $i$  and  $j$ ; naturally the sum includes only the subset of indices paired in configuration  $C$ . The configuration of minimal  $E$  can be obtained recursively as follows. Suppose we have found optimal configurations (and energies) for all sub-sequences of length  $n$  and shorter. The optimal energy for a sub-sequence of length  $(n + 1)$ , say spanning sites  $i$  to  $j = i + n + 1$  is obtained by considering the following  $(n + 2)$  possibilities:  $j$  is unpaired in the optimal configuration, or  $j$  is paired to a site  $i \leq k \leq j - 1$ . In any one of the latter  $(n + 1)$  cases the arch between  $j$  and  $k$  creates two segments (from  $i$  to  $k - 1$ , and from  $k + 1$  to  $j - 1$ ) which are independent due to the planarity restriction. The best energy is thus given by

$$E_{i,j} = \min [E_{i,j-1}, \varepsilon_{kj} + E_{i,k-1} + E_{k+1,j-1}] \quad \text{for } i \leq k \leq j - 1. \quad (2.96)$$

Starting from segments of length  $n = 1$ , where  $E_{i,i+1} = \varepsilon_{i,i+1}$ , the above equation can be used to generate optimal energies for longer segments. The optimal configuration can then be obtained by tracing back.

The above procedure is easily extended to finite temperatures where considerations of entropy may be relevant. We can then assign free energies to segments of the RNA, obtained from corresponding partition functions which may be computed recursively by appealing to Eq. (2.96) as

$$Z_{i,j} = Z_{i,j-1} + \sum_{k=i}^{j-1} e^{-\beta \varepsilon_{kj}} Z_{i,k-1} Z_{k+1,j-1}. \quad (2.97)$$

### 2.5.1 Free energy of molten RNA

Using variants of Eq. (2.97), it is possible to follow how the secondary structure denatures as a function of temperature. Presumably at some temperature  $T_m$  the native structure disappears in favor of a molten state resembling a branched polymer. The nature of the melting process should depend strongly on the RNA sequence and its native structure. In the next section we shall explore this melting for the simple case of an RNA hairpin. In its molten phase, RNA can explore a variety of structures reflecting the competition between energy gain of pairing and the resulting loss of entropy. To estimate the fraction of bound pairs in the molten phase, we can neglect variations in binding energy, setting  $\varepsilon_{ij} = \varepsilon$  and a corresponding Boltzmann weight of  $q \equiv e^{-\beta\varepsilon} \geq 1$ . Once sequence variations are removed, the constrained partition function will depend only on segment length, i.e.  $Z_{i,j} = Z_m(|j-i|+1)$ , and Eq. (2.97) simplifies to

$$Z_m(N+1) = Z_m(N) + q \sum_{k=1}^N Z_m(k-1)Z_m(N-k), \quad \text{with } Z_m(0) = 1. \quad (2.98)$$

It is possible to solve the above recursion (by changing to an ensemble of variable length  $N$ ). However, a more informative solution is obtained by considering the ‘‘mountain’’ representation of planar graphs. The correct weight for each graph is obtained by assigning a factor of 1 for each horizontal step, and  $\sqrt{q}$  to a vertical step (up or down). Each configuration can then be regarded as a Markovian random walk with these weights, and the additional requirement that it never goes below the starting point. The constraint (for an island/mountain landscape, or correct formulation of parentheses) is thus equivalent to a barrier to the random walk at a position one step below the starting point. The problem of a random walk with a so-called absorbing barrier can be solved in several ways— a quite elegant solution is presented by Chandrasekhar in *Rev. Mod. Phys.* **15**, 1 (1943). Let us first ignore the constraint: The partition function for all paths of  $N$  steps starting at the origin is simply  $(1 + 2\sqrt{q})^N$ , accounting for all three possibilities in each step. Similarly, adding the uncorrelated fluctuations in each step leads to a variance of  $\sigma^2 = N(2\sqrt{q})/(1 + 2\sqrt{q})$ . In the limit of large  $N$ , and appealing to the central limit theorem, the net weight of the subset of walks ending at a height  $h$  after  $N$  steps is obtained as

$$W(N, h) = (1 + 2\sqrt{q})^N \exp \left[ -\frac{(1 + 2\sqrt{q})h^2}{4\sqrt{q}N} \right] \times \sqrt{\frac{1 + 2\sqrt{q}}{4\pi\sqrt{q}N}}. \quad (2.99)$$

If we were to ask the question of what fraction of these random walks return to the origin ( $h = 0$ ), we would obtain the expected result of  $\Omega(N) \propto g^N/N^c$  with the ‘loop closure’ exponent of  $c = 1/2$  for our one-dimensional random walks. Of course, for counting planar graphs we need the smaller subset of walks that return to the origin without ever passing to  $h < 0$ , and need to subtract all undesired walks from our sum.

Chandrasekhar’s solution to this problem is closely related to the method of images in electrostatics. The image of the starting point ( $h = 0$ ) with respect to the forbidden state

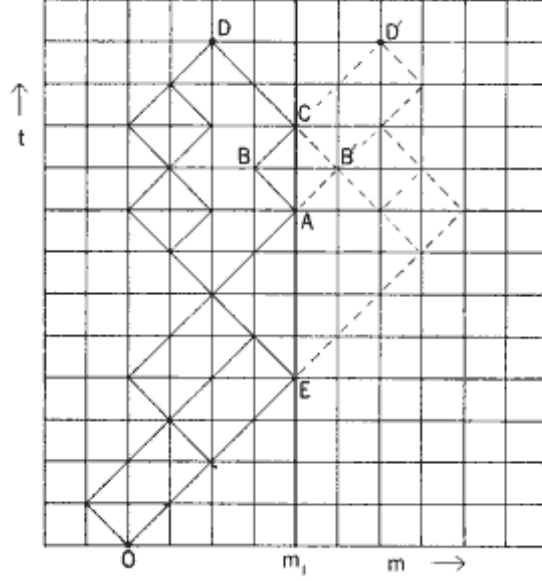


Figure 21: Copied from S. Chandrasekhar, Rev. Mod. Phys. **15**, 1 (1943).

( $h = -1$ ) is located at  $h = -2$ . Consider all walks that start at this image point and end at  $h = 0$ . Each such walk  $W^*$  must cross the ‘mirror’ plane at  $h = -1$  at least once. We can construct a related ensemble of walks  $W'$  which are the reflection of these walks in the mirror-plane (thus starting at the original point  $h = 0$ ) up to the point of first intersecting the forbidden state at  $h = -1$ , and after which following they follow the path of  $W^*$ . We note that the ensemble  $W'$  consists of precisely the paths starting and ending at  $h = 0$  which violate the non-crossing condition. As these paths are in one to one correspondence to  $W^*$ , we simply need to subtract them from the sum in Eq. (2.99) to get the correct number of non-crossing paths. Since  $W^*$  is the ensemble of walks with an end to end excursion of  $h = 2$ , we obtain

$$Z_m(N + 1) = W(N, 0) - W(N, 2) = (1 + 2\sqrt{q})^N \left[ 1 - \exp\left(-\frac{(1 + 2\sqrt{q})}{\sqrt{q}N}\right) \right] \times \sqrt{\frac{1 + 2\sqrt{q}}{4\pi\sqrt{q}N}}. \quad (2.100)$$

The Gaussian approximation is only valid for large  $N$ , and we should similarly expand the difference in brackets above to get the final form

$$Z_m(N + 1) = A(q) \frac{g(q)^N}{N^c}, \quad \text{with } A(q) = \left(\frac{1 + 2\sqrt{q}}{64\pi^3\sqrt{q}}\right)^{3/2}, \quad g(q) = (1 + 2\sqrt{q}), \quad \text{and } c = \frac{3}{2}. \quad (2.101)$$

The most important consequence of the constraint is the change of the exponent  $c$  from  $1/2$  to  $3/2$ . The fraction of bound pairs is merely determined by the probability of going up or

down at any step, and thus given by

$$\frac{\langle N_B \rangle}{N} = \frac{2\sqrt{q}}{1 + 2\sqrt{q}}, \quad (2.102)$$

which changes continuously from 1 at large  $q$  (low temperatures) to  $2/3$  as  $q \rightarrow 1$  at high temperatures.

### 2.5.2 Melting of a hairpin

A particularly simple native RNA structure is a hairpin. For long hairpins the transition from the native form to the molten state can be described analytically using a so-called G $\ddot{o}$  model<sup>4</sup>. In the native configuration monomers  $k$  and  $2N - k + 1$  are paired together, while in

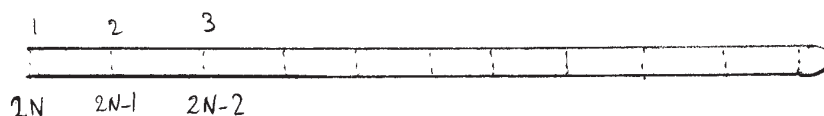


Figure 22: The native state of a hairpin

intermediate configurations partially melted segments alternate with segments that maintain the original bonding.

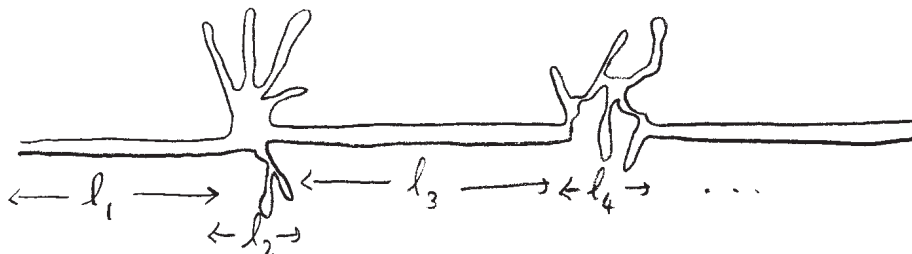


Figure 23: The molten state of a hairpin

A partition function is obtained by summing over all partially melted configurations, and ignoring any interaction between the segments, takes the form

$$Z_n(N) = \sum_{l_1, l_2, l_3, \dots} R(l_1)Z_m(2l_2 + 1)R(l_3)Z_m(2l_4 + 1) \dots, \quad (2.103)$$

with the constraint  $l_1 + l_2 + l_3 + \dots = N$ . The contribution of the molten segments comes from Eq. (2.101). For the native segments, we should add the binding energies of the segments.

<sup>4</sup>R. Bundschuh and T. Hwa, Phys. Rev. Lett. **83**, 1479 (1999).



To make the problem analytically tractable, we assign to each native bond an energy  $\bar{\varepsilon} < \varepsilon$ , and a corresponding Boltzmann weight  $\bar{q} = e^{-\beta\bar{\varepsilon}} > q$ .

With these simplifications, the problem becomes *identical* to the Poland–Scheraga model in Eq. (2.79) with  $w = \bar{q}$ ,  $g = 1 + 2\sqrt{\bar{q}}$  and  $c = 3/2$ . It is thus possible to obtain a melting transition at a finite temperature at which the native fraction goes to zero linearly ( $\beta = 1$ ) from Eq. (2.94)).