RNA secondary structure formation: a solvable model of heteropolymer folding

R. Bundschuh and T. Hwa
Department of Physics, University of California at San Diego, La Jolla, CA 92093-0319, U.S.A.
(June 16, 1999)

The statistical mechanics of heteropolymer structure formation is studied in the context of RNA secondary structures. A designed RNA sequence biased energetically towards a particular native structure (a hairpin) is used to study the transition between the native and molten phase of the RNA as a function of temperature. The transition is driven by a competition between the energy gained from the polymer's overlap with the native structure and the entropic gain of forming random contacts. A simplified Gō-like model is proposed and solved exactly. The predicted critical behavior is verified via exact numerical enumeration of a large ensemble of similarly designed sequences.

PACS numbers: 87.15.Aa, 05.40.-a, 87.15.Cc, 64.60.Fr

A biopolymer such as a DNA or protein is a heteropolymer. It consists of different types of monomers connected linearly in a specific order. Interactions among the monomers give each polymer a robust three dimensional structure on which its biological function depends. This sequence-to-structure relation is rather simple in the case of complementary DNA strands, but can be very complex in the case of proteins. The latter has been intensively studied in the last decade using many different approaches; see e.g., Refs. [1–4].

A number of important ingredients are involved in determining the structure of a heteropolymer. They include (i) thermal fluctuations which "denature" the polymer into a random coil at high temperatures, (ii) monomer-specific binding which freezes a random heteropolymer into a "glass" at low temperatures, and (iii) sequence correlation which biases the polymer into a certain specific (nonrandom) structure, commonly referred to as the "native" structure. The native structures are selected in nature by evolution, but can also be obtained artificially through *sequence design* [5]. The interplay of these ingredients leads to a number of phases depending on the environment (e.g., the temperature) and the extent of sequence correlations or design. The nature of these phases and the transitions among them have been discussed in the context of protein folding [2,3]. However, protein-like models are *not* ideal systems to study phase-related issues because proteins are rather short (typically under 500 monomers). Thus, the thermodynamic limit of long proteins (e.g., strands of $10,000$ amino acids) is not very meaningful.

Here we will study the molecule RNA, an interesting biopolymer which has a mixture of protein-like and DNA-like properties [6]. Due to the nature of the physical interaction [7] between the monomers of a RNA, aspects of the RNA structure formation problem are considerably easier to treat than protein folding. Also, the thermodynamic limit is more meaningful for the RNA, which can contain as many as 10,000 monomers. In this paper, we will describe the simplest effect of sequence bias to the formation of RNA *secondary structures* (defined below). We will focus on the transition between a designed native structure and the RNA's "molten phase", a thermalized, collapsed phase which exists in an intermediate temperature regime in between denaturation and freezing. Such a transition was suggested previously [5] based on numerical studies of short RNA sequences. Here, we will elucidate the physics of this transition by introducing a simplified *two-state* model, analogous to the approach taken by N. Gō for proteins [8]. We will solve the model exactly, and derive the critical properties of the transition between the native and the molten phase. The applicability of our two-state model to the native-molten transition of designed heterogeneous sequences is verified by direct numerical enumeration and finite-size scaling analysis.

RNA is a polynucleotide chain consisting of the four "bases" A, U, G, and C. Energy of the order of several k_BT's can be gained by forming a complementary pair (i.e., $A-U$ and $C-G$) and then stacking them in a double helical structure similar to a double-stranded DNA. In order to form these base pairs, the RNA will need to bend back onto itself at various locations, resulting in a number of helical segments. These helices are then arranged in a three dimensional "tertiary" structure, stabilized by the much weaker interaction between the helices. Due to this crucial separation of energy scales for the RNA, it is possible to distinguish between the formation of "secondary" and tertiary structures: a RNA secondary structure is a collection of base pairings, with the restriction that any two base pairs (i_1, j_1) and (i_2, j_2) have to be either "nested" (i.e., $i_1 < i_2 < j_2 < j_1$) or "independent" (i.e., $i_1 < j_1 < i_2 < j_2$) [9–11]. Base pairings violating these rules lead to structural elements which typically cannot form simple double-helices. Thus, they are energetically or kinetically suppressed and deemed part of the tertiary structure. Each secondary structure can be represented by a non-crossing arch diagram [see Fig. 1(a)], where a pairing between the bases i and j is indicated by a dashed line connecting i and j on a stretched back-

bone. Fig. 1(b) shows an alternative representation of the same structure; here the backbone is bent and the dashed lines are short, in order to convey a sense of the backbone topology. The regions with consecutive base pairings form the above mentioned double-helices.

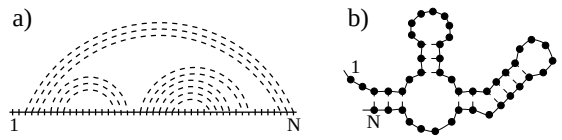

FIG. 1. Representations of the secondary structure of a RNA: (a) a non-crossing arch diagram; (b) a helix diagram. The dashed lines indicate base pairings.

To study the thermodynamic ensemble of all possible secondary structures of a given RNA molecule, the energy of each structure needs to be specified. Here we shall take the simplest energy function, with $v(b,b')$ for each pairing of the bases (b,b'), and $v_0 = -s_0 T$ for each unpaired base mimicking the entropy gained from unbinding [12]. Accurate energy parameters including the effects of stacking, loops, etc. [11] should be used for predicting actual secondary structures of real RNA molecules. As we will argue below, they are *irrelevant* in regards to the asymptotic properties of the phase transition, and will therefore be neglected here for simplicity. In fact, our parameters $v(b,b')$ should be viewed as *coarse-grained* quantities describing the energy of pairing two short *segments* of bases.

The class of RNA secondary structures is clearly hierarchical and belongs to the class of Hartree diagrams widely used in the self-consistent treatment of many-body quantum systems. The recursive nature of the diagrams allows efficient computation of the exact partition function of an arbitrary sequence: Consider a segment of bases from the positions i to $j > i$ inclusive. The base at j can be either unbound or bound to any base $k \in \{i, \ldots, j-1\}$. For the simple energy function we have adopted here, the partition function $Z_{i,j}$ for this segment of bases then obeys

$$Z_{i,j} = Z_{i,j-1} + \sum_{k=i}^{j-1} Z_{i,k-1} \cdot e^{-\varepsilon_{k,j}/T} \cdot Z_{k+1,j-1}, \qquad (1)$$

where we take $\varepsilon_{i,j} \equiv v(b_i, b_j) - 2v_0$. The partition function $Z_{1,N}$ of a strand of length N can be computed recursively using (1) (with $Z_{i,i} = Z_{i,i-1} = 1$) in $O(N^3)$ time [11].

Before we discuss the effect of structure formation due to sequence bias, we first give a qualitative description of the behavior of an uncorrelated random RNA sequence [15]. The energetics of this system is determined by the mean $\overline{\varepsilon}(T)$ and standard deviation $\delta\varepsilon$ of the pairing energies $\varepsilon_{i,j}$. Denaturation occurs at a temperature where $\overline{\varepsilon}(T_d) \approx 0$, since for large and positive ε_0, the unbound state is preferred [14]. Below the denaturation temperature T_d, the bases of the RNA are mostly paired together. There, the system can take on two possible phases: At very low temperatures ($T \ll \delta\varepsilon$), heterogeneity of the sequence is important, forcing the polymer to adopt the optimal base-pairings which minimize the total energy; this is the glass phase [15,16]. At intermediate temperatures ($T_d > T \gtrsim \delta\varepsilon$), differences in the binding energy are less important while an average attraction between the monomers still exists. There, the entropy of forming different pairings becomes dominant, resulting in the molten phase. In a separate study [15], we will demonstrate the irrelevance of weak sequence inhomogeneity in the intermediate temperature regime, thereby establishing the stability and self-consistency of the molten phase. Furthermore, as sequence heterogeneity is irrelevant in the molten phase, statistical properties in this phase can be obtained from (1) by simply taking $\varepsilon_{i,j} = \varepsilon_0 \lesssim \overline{\varepsilon}$, where $\varepsilon_0 < 0$ can be interpreted as an effective mean attraction. Here, we will take the existence of such a molten phase as a conjecture, and examine the effect of sequence bias.

To do so, we need to construct a sequence with a dominant native structure. For simplicity, we will take the native structure to be a hairpin with a long stem, a structure which has been studied numerically and experimentally for short RNAs [5,17] and oligopeptides [18]. We will consider here the limit where the stem is long in order to elucidate asymptotic properties. In the native structure, the bases $(1,2L), (2,2L-1), \ldots, (L,L+1)$ of a length $N = 2L$ sequence are paired. We call these the "native pairs" or "native contacts". Bias towards this structure can be "designed" into the sequence by choosing a *random* sequence for the bases 1 to L of the molecule and then taking the second half of the molecule ($L+1$ to $2L$) to be the *exact reverse complement* of the first half [5]. The perfectly complementary native pairs then make the native structure the "ground state" of the system. Upon increasing temperature, the entropy of forming non-native pairings will compete with and hence weaken the effective bias of the native structure. Alternatively, this bias can be weakened by random "mutations" of the designed sequence. For sufficiently weak effective bias, the RNA can "melt" from its native structure into any of the denatured, molten, or glass phase, depending on the temperature and the strength of the bias.

To study the native-molten transition of the designed sequence analytically, we shall describe the pairing energies and the bias by a simple two-state model,

$$\varepsilon_{i,j} = \widetilde{\varepsilon}\, \delta_{i+j,2L+1} + \varepsilon_0, \qquad (2)$$

for designed sequences of length $2L$. The first term in (2) (with $\widetilde{\varepsilon} < 0$) mimics the *additional* attraction of native pairs due to sequence design; $|\widetilde{\varepsilon}|$ characterizes the "strength" of design which can be controlled by the "mutation" process mentioned above. The second term de-

scribes the average attraction of the "background" characteristic for the molten phase. The two-state model (2) is conceptually similar to the one introduced by N. Gō in the context of protein folding [8]. While Gō proposed this model to simplify the numerical simulation of lattice protein models, we will show that this model actually gives a quantitative description of the native-molten transition of the RNA secondary structure problem (1).

Our task now is to study the system defined by Eqs. (1) and (2). We begin with a description of the molten phase, with $\widetilde{\varepsilon}$ in (2) set to zero. This phase is described by a single parameter, $q \equiv e^{-\varepsilon_0/T}$. Due to the *translational invariance* of the uniform interaction, the partition function can be written as $Z_{i,j} = Z_0(j-i+2;q)$. In terms of the Laplace transform $\widehat{Z}_0(\mu;q) = \sum_{\ell=1}^{\infty} Z_0(\ell;q)e^{-\mu\ell}$, the recursion relation (1) takes on the simple form $\widehat{Z}_0^{-1} = e^{\mu} - 1 - q\widehat{Z}_0$. Inverse Laplace transforming this solution in the limit of large ℓ using the saddle point method, one finds the asymptotic form [9,13]

$$Z_0(\ell;q) = A(q)\,\ell^{-\theta} e^{\mu_0(q)\ell} \tag{3}$$

with $\theta = 3/2$ and $\mu_0(q) = \log(1+2\sqrt{q})$. Physically, the partition function describes the configurational entropy of forming different secondary structures. Each such structure can be viewed as a configuration of an annealed (and rooted) *branched polymer* [14,19], as Fig. 1 suggests. Note that microscopic effects such as the energetic gain of base stacking, cost of hairpin loop formation, and the exclusion of very short loops correspond to the different fugacities for the stem and end points of the branched polymer; they are *irrelevant* to the asymptotic scaling properties of the branched polymer [19]. In particular, they do not change the value of the exponent $\theta = 3/2$, which is a defining characteristic of the molten phase and will play a key role in what follows.

We now include the additional energetic bias $\widetilde{\varepsilon}$ in (2) due to sequence design. For a RNA sequence of $2L$ bases, observe that each secondary structure consists of a series of native pairings, e.g., $(i_1, 2L-i_1+1)$, $(i_2, 2L-i_2+1)$, etc., separated by "bubbles" of lengths $\ell_k = i_{k+1} - i_k$ containing *only* non-native pairings of the intervening bases. Let the Boltzmann weight of each native pairing be $\widetilde{q} \equiv e^{-\widetilde{\varepsilon}/T}$, and let the restricted partition function describing all possible non-native pairings in a molten bubble of length ℓ be $W(\ell;q)$. The total partition function $Z(L+1;\widetilde{q},q)$ for the model (2) can then be conveniently written in Laplace space as

$$\widehat{Z}(\mu;\widetilde{q},q) = \widehat{W}(\mu;q) \sum_{n=0}^{\infty} \left(q\widetilde{q}\,\widehat{W}(\mu;q)\right)^n, \tag{4}$$

where $\widehat{W}(\mu;q)$ and $\widehat{Z}(\mu;\widetilde{q},q)$ are respectively the Laplace transform of $W(\ell;q)$ and $Z(L;\widetilde{q},q)$. Note that $\widehat{W}(\mu,q)$ is completely specified by (4) and the "boundary condition"

$$Z(L+1;\widetilde{q}=1,q) = Z_0(2L+1;q). \tag{5}$$

Before we proceed with an analytical solution of this system, let us observe that Eq. (4) is mathematically very similar to the equation derived by Poland and Scheraga [20] describing the *thermal denaturation* of perfectly complementary DNA double strands. [Note however that while *only* the interaction of native pairs is considered in the treatment of DNA denaturation, the RNA folding problem considered here includes interactions between bases far apart along the backbone of the chain.] The mathematical similarity can be made clearer if one assumes (as it will turn out to be the case) that the the statistics of the molten bubbles can be approximated simply by that of a molten RNA of appropriate length, i.e., $W(\ell;q) \approx Z_0(2\ell-1;q)$. Then the form (3) suggests that one can think of $W(\ell)$ as the Boltzmann weight of a "Gaussian polymer loop" of length ℓ in d-dimensions, with the fictitious dimension d given by 2θ. Thus, the molten bubbles described by W are analogous to the denaturation bubbles in the standard denaturation problem. They both represent the entropic excitations from the native phase, but the physical origin of these entropies is quite different: The denaturation bubbles are driven by by the configuration entropy gained by the unconstrained single strands, while the molten bubbles are driven by the "branching entropy" of secondary structures within the (collapsed) molten phase. Nevertheless, we expect that the native-molten transition belongs to the same universality class as the denaturation transition of Ref. [20], with $d = 2\theta = 3$.

The partition function $Z(L;\widetilde{q},q)$ can actually be computed *exactly* for all q and $\widetilde{q}$ in the limit of large L [14]. First, by using Eqs. (4) and (5), an *exact* expression for $\widehat{Z}(\mu;\widetilde{q},q)$ can be derived. It is then straightforward to extract the reduced free energy $f(\widetilde{q},q) \equiv -(\ln Z)/L$ from the singularity in $\widehat{Z}$. Close to the critical point $\widetilde{q}_c = (3\sqrt{1+2\sqrt{q}}-1)/(\sqrt{1+2\sqrt{q}}-1) > 1$, one finds $f(q,\widetilde{q}) = -2\mu_0(q)$ [with $\mu_0(q)$ as given in (3)] for $\widetilde{q} \le \widetilde{q}_c$, while $f(q,\widetilde{q}) = -2\mu_0(q) + B(q)\,(\widetilde{q}-\widetilde{q}_c)^2$ with a known regular function $B(q)$ for $\widetilde{q} \gtrsim \widetilde{q}_c$. This form of the free energy implies a *continuous* phase transition with a *finite jump* in the specific heat at the critical point; thus, the specific heat exponent is $\alpha = 0$. From the free energy, we can easily compute the average *fraction of native contacts*, $Q = -\mathrm{d}f/\mathrm{d}\ln\widetilde{q}$, which constitutes the *order parameter* of the phase transition. In the thermodynamic limit, $Q = 0$ for $\widetilde{q} \le \widetilde{q}_c$ and $Q = 1$ for $\widetilde{q} \gg \widetilde{q}_c$. Close to the critical point, $Q \sim (\widetilde{q}-\widetilde{q}_c)$.

For strands of *finite* length L, this length always enters the saddle point equation involved in the inverse Laplace transform of $\widehat{Z}(\mu)$ in the combination $L(\widetilde{q}-\widetilde{q}_c)^{\nu}$ with $\nu = 2$ [14]. The finite-size result can be cast into the form $Q(L) = L^{-1/2}\,g\left[(\widetilde{q}-\widetilde{q}_c)L^{1/2}\right]$ in the vicinity of the critical point. The scaling function $g[y]$ can be computed

exactly, with $g[y] \sim y$ for $y \gg 1$, $g[y] \sim 1/|y|$ for $y \ll -1$, and $g[y] \sim O(1)$ for $|y| \ll 1$.

In order to verify whether the above critical behaviors of the Go-like model describes the native-molten transition of designed *heterogeneous* sequences, we numerically iterated Eq. (1) for perfectly designed sequences [21]. The pairing energies $\varepsilon_{i,j}$'s were chosen to be -1 for complementary pairs and 0 otherwise.

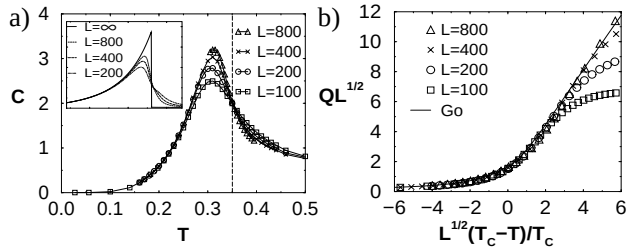

FIG. 2. Numerical results on RNA sequences with bias for a hairpin. (a) Specific heat for different sequence lengths. The vertical line indicates the critical temperature T_c. The inset depicts exact results obtained from the Gō-like model. (b) Scaling plot of the fraction Q of native contacts; the solid line is the exact solution of the Gō-like model.

Fig. 2(a) shows the specific heat for perfectly designed sequences of 200 to 1600 bases, averaged over 100 realizations of randomness. Direct extraction of critical exponents from this data is difficult due to the strong correction-to-scaling effects of the expected discontinuity ($\alpha = 0$). However, for $\alpha = 0$, a good numerical estimate of the critical temperature T_c can be obtained from the common intersection point of the curves at different lengths. This is more clearly seen from the result of the Gō-like model (2); see inset of Fig. 2(a). The fraction Q of native contacts can then be used for a detailed scaling analysis. As shown in Fig. 2(b), the scaling plot of $QL^{1/2}$ versus $L^{1/2}(T_c - T)/T_c$ collapses the data and approaches the predicted critical behavior $g[y]$ represented by the solid line.

To summarize, we analyzed the heteropolymer structure formation problem in the context of RNA secondary structures. The native-molten structural transition results from a competition between the energetic gain of native contacts and the "branching entropy" of the molten phase. Critical properties can be obtained exactly after introducing an approximate two-state model à la Gō; the validity of the approximation is verified by direct numerical calculation of designed sequences. Our findings are in qualitative agreement with the earlier numerical study on short RNA sequences with more realistic energy parameters [5]. Aside from depicting a concrete physical picture of the phase transition and the molten phase, our analytical study provides a quantitative description of thermodynamic properties of the system, e.g., the *continuous* nature of the phase transition contrary to what was claimed in Ref. [5]. Throughout this study, we have neglected the effect of the excluded-volume interaction [22]. This effect changes the value of the exponent θ [23]; hence it is expected to change the universality class or even the order of the phase transition in 3 dimensions as will be discussed elsewhere [14]. However, it should not change the qualitative physics of the competing interactions discussed here. This physics should also be relevant to the statistics of heteroduplex formation which occurs in the hybridization of partially complementary heterogeneous DNA strands [24]. In this regard, the structural transition discussed here resembles the physics of similarity detection investigated previously in the context of DNA sequence alignment [25].

We are grateful to D. Cule who was involved in the early stages of this study, and to discussions with M. Zuker, J.N. Onuchic and J.D. Moroz. RB is supported by a Hochschulsonderprogramm III fellowship of the DAAD and TH by a Beckman Young Investigator Award.

[1] K.A. Dill *et al.*, Protein Sci. **4**, 561 (1995).
[2] J.N. Onuchic, Z. Luthey-Schulten, and P.G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).
[3] T. Garel, H. Orland, and E. Pitard, J. Phys. I **7**, 1201 (1997).
[4] E.I. Shakhnovich, Curr. Op. Struct. Biol. **7**, 29 (1997).
[5] P.G. Higgs, J. Phys. (France) **3**, 43 (1993).
[6] See, e.g., *The RNA World*, R.F. Gesteland and J.F. Atkins ed. (Cold Spring Harbor Laboratory Press, 1993).
[7] W. Sanger, *Principles of Nucleic Acid Structure* (Springer Verlag, New York, 1984).
[8] N. Gō, J. Stat. Phys. **30**, 413 (1983).
[9] M.S. Waterman, Adv. Math. Suppl. Studies **1**, 167 (1978).
[10] M. Zuker and D. Sankoff, Bull. Math. Biol. **46**, 591 (1984).
[11] J.S. McCaskill, Biopolymers **29**, 1105 (1990).
[12] We have neglected here the additional logarithmic entropy cost associated with the formation of single-stranded *loops*. This loop entropy is crucial to denaturation [13,14], but is not relevant to this study of the native-molten transition away from the denaturation phase.
[13] P.G. de Gennes, Biopolymers **6**, 715 (1968).
[14] J.D. Moroz, R. Bundschuh, and T. Hwa (unpublished).
[15] R. Bundschuh and T. Hwa (unpublished).
[16] P. Higgs, Phys. Rev. Lett. **76**, 704 (1996).
[17] E. Zuo *et al.*, Biochemistry **29**, 4446 (1990).
[18] V. Muñoz *et al.*, Nature **390**, 196 (1997).
[19] T.C. Lubensky, J. Isaacson, and S.P. Obukhov, J. Physique **42**, 1591 (1981).
[20] D. Poland and H.A. Scheraga, J. Chem. Phys. **45**, 1464 (1966).
[21] The sequence is chosen to have the strongest bias so that it is far away from the native-glass transition.
[22] S.J. Chen and K.A. Dill, J. Chem. Phys. **103**, 5802 (1995).
[23] G. Parisi and N. Sourlas, Phys. Rev. Lett. **46**, 871 (1981).
[24] See, e.g., J.G. Wetmur, Crit. Rev. Biochem. Mol. Biol. **26**, 227 (1991).
[25] T. Hwa and M. Lässig, Phys. Rev. Lett. **76**, 2591 (1996).