Melting of Branched RNA Molecules

Ralf Bundschuh[1] and Robijn Bruinsma[2]

[1]*Department of Physics, Ohio State University, Columbus, Ohio 43210-1117, USA*
[2]*Department of Physics and Astronomy, The University of California at Los Angeles, Los Angeles, California 90049, USA*

In this Letter we show that the melting thermodynamics of RNA molecules is very sensitive to the branching geometry. We find that, when pairing interactions are described by a Gō model, unbranched RNA molecules with a linear geometry melt via a conventional continuous phase transition with classical exponents while RNA molecules with the branching geometry of a Cayley tree, with coordination number three, have a free energy that shows no thermodynamic singularity within numerical precision. Nevertheless, we provide an analytical proof that the free energy does have a mathematical singularity at the stability limit of the ordered structure. The correlation length appears to diverge but only on the high-temperature side of this singularity.

A fundamental principle of statistical mechanics states that phase transitions are not possible for one-dimensional systems unless long-range interactions are present. It thus came as a surprise when Poland and Scheraga (PS) showed [1] that an infinite, linear molecule composed of two flexible polymer strands bound together by a local attractive interaction *does* undergo a true phase transition when the two strands separate. The required long-range correlations are due to the fact that the partition function of a strand separation "bubble" has a power-law dependence on size [2]. This PS mechanism is encountered in the context of the denaturation of RNA molecules [3]. RNA molecules usually operate in a single-stranded mode. Pairing between complementary bases of the strand produces the "secondary structure" [4]: a treelike graph of unpaired "bubbles" linked by paired double-helical segments. The minimum-energy secondary structure can be predicted from the primary sequence of nucleotides [5]. Melting of the secondary structure of an RNA molecule produces a "molten-globule" state with the molecule fluctuating over a range of different secondary structures [6]. In his pioneering paper of 1968 [7], de Gennes showed that the partition function $G(L)$ of a large RNA molecule fluctuating over *all* possible secondary structures, with identical pairing energies, has a power-law dependence on size of the form z_0^L/L^θ with $\theta = 3/2$. Subsequently, Bundschuh and Hwa [8] showed that if the ground state secondary structure of an RNA molecule is a long, uniform hairpin, then the molecule undergoes a continuous phase transition to the molten-globule state with the same melting thermodynamics as that of the PS model.

Actual RNA secondary structures have a heterogeneous, branched, treelike form. In this Letter we will discuss the effects of branching on the melting thermodynamics. The ground state secondary structure will be assumed to be a *Cayley tree* (see Fig. 1) with coordination number three. The RNA strand traces out the perimeter of the tree, starting and ending at the root of the tree, with each branch of

the tree occupied by a complementary base pair. The total size of the molecule is indexed by the *level k* of the tree, which is related to the total sequence length $N(k)$ by $N(k) = 2^{k+2} - 2$ bases (a $k = 1$ tree is here a three-armed star with one base pair per arm). If one numbers the bases of the strand then the "designed" ground state will be denoted by the set $S = \{(i_1, j_1), (i_2, j_2), \ldots, (i_M, j_M)\}$ of complementary pairs.

The folding energy of the molecule will be described by a "Gō model" [9] where one assigns a binding energy $-\tilde{\varepsilon}$ to a pair belonging to the set S and a weaker binding energy $-\varepsilon$ if the pair does not belong to S (the "nonspecific" binding energy). We will allow in general *any* secondary structure provided it does not have circuits (or "pseudoknots"). In the Gō model, the finite temperature partition
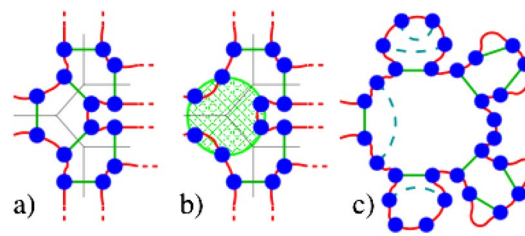

FIG. 1 (color online). Single-stranded RNA molecule having a branched secondary structure that follows the outline of a Cayley tree with coordination number three. Nucleotides are schematically indicated by circles, bonds between nucleotides and complementary pairing by solid lines around the perimeter or across the structure, respectively. (a) Ground state structure with pairing restricted to a complementary "native" pair for each branch of the Cayley tree. (b) In a molten-globule bubble (hatched) all possible pairing interactions are permitted. (c) For a fixed set S' of specific base pairs (solid lines) the set $C(S')$ is defined as all possible sets of base pairs compatible with the base pairs in set S' (dashed lines are one example of such a set). Note that the additional base pairs are constrained to the loops of the structure S' so the sum over all possible base-pairing configurations factorizes into a factor from each such loop.

function is given by

$$Z(q, \tilde{q}, S) = \sum_{S' \subset S} \tilde{q}^{|S'|} \sum_{S'' \in C_{NS}(S')} q^{|S''|}, \quad (1)$$

where $q = \exp(\beta\epsilon)$ and $\tilde{q} = \exp(\beta\tilde{\epsilon})$ are the Boltzmann weights of nonspecific and specific pairs, respectively. The first sum in Eq. (1) is over all subsets S' of S. Once such a subset S' is fixed, $C_{NS}(S')$ denotes the collection of all sets S'' of nonspecifically paired bases such that the union $S' \cup S''$ defines an acceptable, circuit-free secondary structure for the molecule as a whole [see Fig. 1(c)]. The difficulty in evaluating the partition function resides in the fact that in the second sum we cannot allow pairing between specific pairs that do not belong to S'. In order to simplify the second sum, we apply a variant of the binomial theorem to the first sum over S' by replacing it with a *double* sum over all possible ways to divide S' into two parts S_1 and S_2 whose union $S_1 \cup S_2$ equals S':

$$\sum_{S' \subset S} \tilde{q}^{|S'|} f(S') = \sum_{S_1 \subset S} \sum_{S_2 \subset S\S_1} (\tilde{q} - q)^{|S_1|} q^{|S_2|} f(S_1 \cup S_2).$$

$$(2)$$

Here, $S\S_1$ denotes the complement of S_1, and $f(T)$ is any function defined on the set T. Replacing the sum over S' in Eq. (1) by a double sum over S_1 and S_2 gives

$$Z(q, \tilde{q}, S) = \sum_{S_1 \subset S} (\tilde{q} - q)^{|S_1|} \sum_{S_3 \in C(S_1)} q^{|S_3|}, \quad (3)$$

where $C(S_1)$ is defined in the same way as $C_{NS}(S')$, except that the restriction excluding specific pairing has been lifted. The specific pairs were included through the sum over S_2 in Eq. (2). The statistical weight $\sum_{S_3 \in C(S_1)} q^{|S_3|}$ now can be written as the product of the statistical weights for the individual loops $L(S_1)$ linking the clusters of specific pairs belonging to S_1:

$$Z(q, \tilde{q}, S) = \sum_{S_1 \subset S} (\tilde{q} - q)^{|S_1|} \prod_{\{L(S_1)\}} G(L(S_1)). \quad (4)$$

Here, $G(L) \propto z_0(q)^L / L^\theta$ is the familiar "molten-globule" partition function for a strand of length L, which depends only on q. The partition function can be viewed as a sum over all possible bubble configurations, with $G(L)$ the partition function of the bubble [10].

Equation (4) has a form for which we can construct explicit recursion relations. The first step is to remove from the sum the open loop located at the root of the tree (see Fig. 1):

$$Z(q, \tilde{q}, S) = \sum_{n=0}^{N(k)/2} G(2n) W(k, n), \quad (5)$$

with n the number of base pairs in the root bubble. Here, $W(k, n)$ is a *restricted* partition function, i.e., the partition function of a molecule with n accessible bases in the open bubble at the root, but *not* including the configurations of

the open bubble. We cut the tree into two equal sized subtrees with level index $k - 1$. The number of accessible base pairs of the two subtrees together must add to $n - 1$, as we removed one pair by the cutting operation. Because we permit no circuits, the restricted partition function of a level k tree for $n > 0$ can be expressed in terms of a product of the restricted partition functions of two $k - 1$ level subtrees:

$$W(k, n) = \sum_{m=0}^{n-1} W(k - 1, m) W(k - 1, n - 1 - m), \quad (6)$$

with $W(k - 1, m) = 0$ if $m > 2^k - 1$. The $n = 0$ case, i.e., a tree with no bubble at the root, must be treated separately. Take the first complementary pair at the root of the tree out of the partition function, and then sum over all possible sizes for the bubble that immediately follows this pair (including bubbles of zero size) and treat *those* bubbles as the bubble at the root of a new tree that can again be cut into two equal parts in the same way as before. Using Eq. (4), we obtain a second recursion relation:

$$W(k, 0) = (\tilde{q} - q) \sum_{n_1=0}^{2^k-1} \sum_{n_2=0}^{2^k-1} W(k - 1, n_1) W(k - 1, n_2)$$
$$\times G[2(n_1 + n_2)]. \quad (7)$$

Equations (6) and (7) together constitute a complete set of recursion relations from which $W(k, n)$ can be obtained by iterative solution. The initial conditions for the recursion relations are $W(1, 0) = (\tilde{q} - q)[1 + 4q + q^2 + 2\tilde{q} + \tilde{q}^2]$, $W(1, 1) = (\tilde{q} - q)^2$, $W(1, 2) = 2(\tilde{q} - q)$, and $W(1, 3) = 1$, as follows by inspection. We carried out this iteration procedure numerically, up to level $k = 19$, for different values of $\tilde{q} = \exp(\beta\tilde{\epsilon})$ and for fixed $q = 4$. In Fig. 2 we show the second derivative of the free energy per site with respect to $\tilde{q}$, which effectively correspond to the heat capacity. As one increases the value of k, a maximum develops near $\tilde{q} = 80$. However, within the numerical precision, the free energy per site does *not* develop a thermodynamic singularity in the large N limit. This must be contrasted with the Bundschuh and Hwa case, where the molecule had a uniform, linear ground state, in which case the heat capacity very clearly develops such a singularity already for much smaller system sizes (left-hand inset of Fig. 2).

In order to examine *subleading* contributions to the free energy, i.e., terms that are small compared to the leading term proportional to N, we also computed the "pinching free energy"

$$\Delta F(k)/k_B T \simeq \ln Z(k + 1) - 2 \ln Z(k). \quad (8)$$

For example, in a molten-globule phase the partition function should have the asymptotic scaling form $a^+ z_0^N / N^{3/2}$ for large N. The pinching free energy $\Delta F(k)/k_B T \simeq \frac{3}{2} \times (k + 2)\ln 2 - \ln a^+$ then should have a *linear* dependence
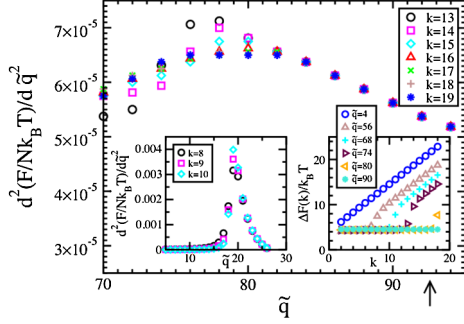
FIG. 2 (color online). Second derivative of the free energy with respect to the Boltzmann weight $\tilde{q}$ of specifically paired bases plotted as a function of $\tilde{q}$ for different values of the level k of the Cayley tree ground state. The free energy was computed numerically from the recursion relations Eqs. (6) and (7) and expressed in units of Nk_BT with N the sequence length of the RNA strand. The arrow denotes the location of the mathematical singularity associated with melting of the root of the Cayley tree. Left-hand inset: Same except that the ground state is a linear hairpin. A thermodynamic singularity develops near $\tilde{q}_c = 18.4$ with mean-field critical exponents. Right-hand inset: Numerically computed "pinching" free energy $\Delta F(k)$ [see Eq. (8)] versus the level k of the Cayley tree. For $\tilde{q}$ larger then 90, $\Delta F(k)/k_BT$ is independent of k, consistent with the ordered ground state. For $\tilde{q}$ less than 20, $\Delta F(k)/k_BT$ can be fitted by the scaling relation $\Delta F(k)/k_BT \simeq \frac{3}{2}(k+2)\ln 2 - \ln a^+$ associated with the molten-globule state. The crossover point between these two regimes for intermediate values of $\tilde{q}$ marks the size of the ordered, correlated regions in the molten-globule state. For $\tilde{q}$ above 80, the size of these correlated regions exceeds the system size.

on k, with slope $3/2$. In an ordered phase, the partition function should scale as $a^- z_0^N$ for large N, in which case $\Delta F(k)$ should be a *constant* independent of k. The right-hand inset of Fig. 2 shows that, for $\tilde{q}$ values up to 80, $\Delta F(k)$ indeed has a linear dependence on k, for large k, with a slope close to $3/2 \ln 2$. This indicates that, for $\tilde{q}$ values below 80, the tree is in the molten-globule phase. Since for the corresponding case of a linear ground state the melting point is as low as $\tilde{q}_c = 18.4$ for $q = 4.0$, we are forced to conclude that branching has a powerful *destabilizing* effect on the ordered state.

For smaller k values, the pinching free energy is a constant, which indicates that the ordered ground state dominates over shorter length scales. The crossover point separating the two regimes can be interpreted as a *correlation length* ξ whose physical meaning would be that of the typical size of smaller ordered Cayley tree-type structures imbedded in a larger molten-globule state. The value of ξ increases with $\tilde{q}$ and beyond $\tilde{q} = 80$ it exceeds our maximum system size ($N = 10^6$). A fit to a power law $\xi \propto (\tilde{q}_c - \tilde{q})^{-\nu}$ produces a correlation length exponent $\nu \simeq 2.1$ and a critical $\tilde{q}_c \sim 80$.

Can we really be sure that the ordered phase is thermodynamically stable at any finite temperature? The re-

stricted partition function is expected to have the scaling form $W(N, n) \sim w(n)z_0^n$ in the ordered phase, with $w(n)$ the fraction of configurations that have an open bubble at the root of size n. If we insert this ansatz into the recursion relation Eq. (6), we obtain, for positive n, the following *fixed-point condition*:

$$w(n) = \sum_{m=0}^{n-1} w(m)w(n-1-m). \qquad (9)$$

This equation can be solved by applying the discrete Laplace transform $\hat{w}(z) = \sum_{m=0}^{\infty} w(m)z^{-m}$. The solution $\hat{w}(z) = \frac{z}{2} - \sqrt{\frac{z^2}{4} - zw(0)}$ has a branch cut starting at $z = 4w(0)$, with $w(0)$ an undetermined constant. After applying an inverse Laplace transform, one finds that $w(n)$ actually has the same scaling form as the partition function of a molten globule:

$$w(n) \propto \frac{\exp\{-n \ln[1/4w(0)]\}}{n^{3/2}}. \qquad (10)$$

The mathematical origin of the $n^{-3/2}$ factor is here a combinatorial factor that reflects the different ways one can partition the open bubble between the two subtrees. We may interpret $\xi \sim 1/\ln[1/4w(0)]$ as the characteristic size of a molten-globule bubble at the root of the tree in the ordered phase. Numerical iteration of the recursion relations for $W(k, n)$ for $\tilde{q} = 150$ and $q = 4$ were found to be consistent with Eq. (10). If we now use $W(N, n) \sim w(n)z_0^N$ in the remaining recursion relation Eq. (7), together with Eq. (9), then one obtains the following self-consistency relation for the remaining unknown constant $w(0)$:

$$w(0) = \frac{\tilde{q} - q}{2\pi i} \oint \frac{1}{z}\hat{G}(z)\hat{w}(1/z)^2 dz, \qquad (11)$$

where $\hat{G}(z)$ is the discrete Laplace transform of $G(L)$ [11]. $\hat{G}(z)$ has a branch cut that terminates at $z = (1 + 2\sqrt{q})^2$. The integration contour in Eq. (11) must run inside an annulus in the complex plane that surrounds the origin passing the real axis *outside* the branch cut of $\hat{G}(z)$ but *inside* the branch cut of $\hat{w}(1/z)$ that starts at $z = 1/4w(0)$. That means that the contour integral only can be carried out as long as $w(0) \leq 1/4(1 + 2\sqrt{q})^2$. The partition function develops a mathematical singularity when the two branch cuts merge. At that point, the partition function $w(n) \propto [1/4w(0)]^n/n^{3/2}$ of the root bubble has the same form as the partition function $G(n) \propto [(1 + 2\sqrt{q})^2]^n/n^{3/2}$ for a molten globule of the same size. We thus identify $w(0) = 1/4(1 + 2\sqrt{q})^2$ as the stability limit of a finite temperature ordered phase. Note that the (low-temperature) correlation length $\xi \sim 1/\ln[1/4w(0)]$ *cannot* diverge at the stability limit. The critical value $\tilde{q}_c$ for $\tilde{q}$ at the stability limit is now easily obtained by noting that $w(0)$ is small compared to 1. Expanding the argument of the contour integral in powers of $w(0)$ leads to

$$(\tilde{q} - q)^{-1} \cong w(0) + 2(1 + q)w(0)^2$$
$$+ 5(1 + 6q + 2q^2)w(0)^3 + \cdots. \quad (12)$$

If Eq. (12) is combined with $w(0) = 1/4(1 + 2\sqrt{q})^2$ one finds that for $q = 4$ the singularity is located at $\tilde{q}_c \cong 92.6$. The numerically computed free energy per site exhibits no singular dependence on $\tilde{q}$ in that range (see Fig. 2). This is not inconsistent because $w(n)$ only contributes a subleading term to the total free energy. On the other hand, the correlation length obtained from the pinching free energy appears to diverge near $\tilde{q}_c$. We encountered, however, strong *finite-size effects* in the numerical solution of the recursion relations for $\tilde{q}$ values in the range between 80 and 90, which make it difficult to numerically explore the critical properties in more detail.

In summary, a branched RNA molecule in the form of a Cayley tree does have a mathematical singularity in the free energy at a temperature where the ordered ground state becomes unstable. This singularity does not correspond to a conventional phase transition, however, as the numerically computed specific heat does not exhibit an anomaly at the singularity. On the high-temperature side, the numerically computed correlation length appears to diverge yet the mathematical singularity itself is not associated with a divergence of the correlation length. We conclude that branching (i) ''smears out'' the melting transition and (ii) destabilizes the ordered phase but without suppressing it altogether.

Experimental studies comparing the melting characteristics of large, branched RNA molecules with that of linear, unbranched molecules that could probe this exotic form of melting would be very interesting. An important question in this respect would be the role of excluded-volume interactions and of ''tertiary'' pairing interactions, i.e., pairing interactions that introduce, for example, pseudo-knots. Excluded-volume interactions in general tend to suppress thermal fluctuations and possibly could restore the thermodynamic singularity in the free energy per site that was encountered for linear molecules. Tertiary interaction could have the effect of turning a branched, secondary template into a three-dimensional gel-like structure, in which case the transition to the molten-globule state could resemble the melting transition of a bulk solid material.

[1] D. Poland and H. A. Scheraga, J. Chem. Phys. **45**, 1464 (1966).
[2] The order of the phase transition depends delicately on the role of excluded-volume interactions.
[3] See, e.g., B. Alberts *et al.*, *Molecular Biology of the Cell* (Garland, New York, 1994); P. B. Moore in *The RNA World*, edited by T. T. Cech, J. F. Atkins, and R. F. Gesteland (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 2005).
[4] For a review, see P. G. Higgs, Q. Rev. Biophys. **33**, 199 (2000).
[5] M. Zuker and P. Stiegler, Nucleic Acids Res. **9**, 133 (1981).
[6] P. G. Higgs, J. Phys. I (France) **3**, 43 (1993).
[7] P. G. de Gennes, Biopolymers **6**, 715 (1968).
[8] R. Bundschuh and T. Hwa, Phys. Rev. Lett. **83**, 1479 (1999).
[9] N. Gō, J. Stat. Phys. **30**, 413 (1983).
[10] For an expanded version of the proof of Eq. (4), see R. Bundschuh, ''Unified Approach to Partition Functions of RNA Secondary Structures,'' in Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore (Singapore University Press and World Scientific, Singapore, to be published).
[11] The exact form is $\hat{G}(z) = \frac{z}{4q} - \frac{\sqrt{z}}{4q}[\sqrt{(\sqrt{z} - 1)^2 - 4q} + \sqrt{(\sqrt{z} - 1)^2 + 4q}]$.