Ground state and glass transition of the RNA secondary structure

Hui, S.; Tang, L.-H.

Published in:
European Physical Journal B

DOI:
10.1140/epjb/e2006-00347-x

Published: 01/09/2006

Document Version:
Peer reviewed version

Ground state and glass transition of the RNA secondary structure

Sheng Hui and Lei-Han Tang*
Department of Physics, Hong Kong Baptist University, Kowloon Tong, Hong Kong SAR, China
(Dated: June 30, 2005)

RNA molecules form a sequence-specific self-pairing pattern at low temperatures. We analyze this problem using a random pairing energy model which mimics the specificity in stem formation. A number of interesting properties are investigated, including the branching structure, the influence of temperature, etc.

PACS numbers: 75.50.Lk, 64.60.Ak, 74.78.-w, 75.10.Hk

I. INTRODUCTION

A central dogma in modern biology is the generally one-to-one correspondence between the spatial structure (i.e. conformation) of a biomolecule and its sequence information (i.e., the genetic code). Yet this link remains at an empirical level due to the hitherto unyielding computational complexity in predicting the shape of a heterogeneous polymer. At the heart of the problem is the lack of a general understanding on the energetics of a collapsed polymer in the presence of sequence-specific contact energies. Such a situation has been compared with the low temperature behavior of the spin glass model,[1] although the chain constraint and the unknown nature of sequence specificity may invalidate the analogy.

In the present paper we focus on the secondary structure of RNA molecules. RNA, like DNA, is a long chain molecule made of four different types of nucleotides adenine (A), uracil (U), guanin (G) and cytosine (C). Under normal physiological conditions, an RNA molecule folds into a relatively compact shape which can be loosely described as a branching tree of double-stranded helical segments (known as stems) with occasional single-stranded bulges and hairpins. Unlike a ds-DNA molecule, however, each helical segment is made of two accidentally complementary strands from different parts of the same chain, running in opposite directions. The matching of bases to form the Watson-Crick A-U and G-C pairs, and the energetically less favorable wobble G-U pairs defines the secondary structure of an RNA molecule. The problem of RNA secondary structure prediction is then to find the map of optimal pairings for a given sequence of the nucleotides (the primary structure). At finite temperatures, one has to consider structures that are not necessarily optimal in energy, but are nevertheless important due to their configurational entropy.

Compared to protein folding, RNA secondary structure prediction is a simpler problem due to the saturation of base-pairing. In particular, for RNA molecules without the so called "pseudoknots", pairing of bases in an RNA molecule may be represented by one-dimensional, non-intersecting rainbow diagrams.[7] Due to this topological constraint, the partition function of a chain of N bases can be determined through an exact dynamic programming algorithm whose computational complexity scales as N^3.[9,10] Consequently, chains of length up to a few thousand bases can be readily investigated on a workstation. Even so, the statistics of the branching pattern and the mechanism through which sequence information is expressed have not been thoroughly understood.

The homopolymer version of the RNA problem has been solved analytically for a long time.[11] It was shown that the branching entropy of the polymer is extensive in the chain length, but it contains a logarithmic term when the two ends of the polymer are pinched together.[7] We shall argue that this logarithmic contribution is of entropic origin. For long chains, the size of the polymer is found to scale as $N^{1/2}$.

The problem becomes much more complicated when the bases on the chain are treated as of distinct types. To gain a general understanding of the resulting energy landscape, Higgs introduced a random heteropolymer model of RNA secondary structure formation.[4] In his model, only Watson-Crick pairing is allowed and each such pair is assigned a negative energy. Through numerical simulations of random sequences, he observed that the ground state is highly degenerate and the system at low temperatures exhibits a broad distribution of the overlap function, characteristic of a weak glass. The same conclusion was reached in a recent work by Pagnani *et al.* who also studied the molten-to-glass transition[5]. The existence of a spin-glass type ground state (which is a open topic by itself) is however disputed by Hartmann.[6]

Bundschuh and Hwa have recently carried out extensive analytical and numerical studies of the RNA secondary structure problem. Their work revealed a rich phase diagram.[7] At low temperatures, sequence mutation brings the chain into a glass state where the size of the folded structure scales as $N^{0.69}$. As temperature increases, the glass state transforms into a molten state with similar characteristics as the homopolymer chain. A final transition into the denatured state takes place when temperature is further increased. One interesting finding of their work is that the pinching free energy in the glass state grows as a power-law with the chain length, but the exponent is small and nonuniversal. We shall demonstrate that the seemingly power-law behavior can be well-fitted by a logarithm-squared term with a temperature dependent coefficient, which vanishes at the glass transition.

Mention the work by Mèzard et al.[12]

The aim of the present work are as follows.

(i) Quantify the statistics of the branching process and the resulting secondary structure of the RNA molecule for random and biological sequences;

(ii) Develop a theory for the statistics of the pinching energy (i.e., the excess energy for forcing a particular pair). This will provide the basis for understanding various peculiar features of the low temperature phase and the transition to the molten state;

(iii) Investigate the stability of the ground state against sequence mutations.

II. THE MODEL AND DYNAMIC PROGRAMMING

The statistical mechanics of the secondary structure of random RNAs is presented in detail in Ref.[7]. An RNA molecule is defined by its nucleotide sequence. A secondary structure of the molecule is a pairing pattern of bases on the sequence, where each base (indexed by its position i in the sequence) has at most one partner. As in most previous studies, we consider here only secondary structures that obey the "noncrossing" constraint, i.e., if base i pairs with base $j > i$, and another base $k > i$ pairs with base $l > k$, then either $i < j < k < l$ (independent) or $i < k < l < j$ (nesting). This class of structures, which are the most common in nature, form the configuration space of the RNA molecule.

Realistic prediction of the thermodynamically favored RNA secondary structure involves a large parameter set. Its main purpose is to differentiate accurately local pairing alternatives. This complication, we believe, is not necessary for a statistical characterization of the scaling properties in the low temperature phase and around the glass transition in the random sequence ensemble. Instead, we consider here a much simpler model where the energy of a secondary structure S is given by,

$$E[S] = \sum_{(i,j)\in S} \epsilon_{ij}, \tag{1}$$

where ϵ_{ij} is the pairing energy of base i with base j. The sum is over all base pairings (i,j) of S.

To complete the description of the model, we need to assign values to the pairing energies ϵ_{ij} for a given nucleotide sequence. The standard choice is to make ϵ_{ij} dependent on the two nucleotides involved. A look-up table for the ten possible combinations can be constructed for this purpose. For the random sequence ensemble, an alternative approach is to choose ϵ_{ij} as independent random variables, as suggested in Ref.[7]. This was motivated at first by analytical considerations and supported by numerical evidence. In fact, the two approaches become quite identical when the alphabet size exceeds sequence length, as then every possible pair has a different combination of partners for a typical random sequence. Considering that, for real RNA, each duplex typically contains a consecutive stack of five or more paired bases (with more than $4^5 = 1024$ possible sequences on each side), one may view the second approach as defining a coarse-grained model on the scale of a duplex. Previous work on sequence alignment has shown that the matching energy of two randomly selected sequences follows a distribution with an exponential tail. Thus, as a coarse-grained model of RNA secondary structures in the sense described above, we choose $\epsilon_{ij} < 0$ to be independent random variables satisfying the distribution,

$$P(\epsilon) = \epsilon_0^{-1} \exp(\epsilon/\epsilon_0), \tag{2}$$

where $\epsilon_0 = 1$ sets the only energy scale of the problem.

Due to the noncrossing constraint on the pairing patterns, the partition function

$$Z(N) = \sum_S \exp(-E[S]/T) \tag{3}$$

of an RNA molecule of N bases at temperature T can be calculated using a dynamic programming algorithm.[9,10] This is done based on the recursive relation

$$Z_{i,j} = Z_{i,j-1} + \sum_{k=i}^{j-1} Z_{i,k-1} e^{-\epsilon_{k,j}/T} Z_{k+1,j-1}. \tag{4}$$

Here $Z_{i,j}$ denotes the partition function of a contiguous segment of the molecule from position i to position j. Starting from the shortest segments of one base each with $Z_{i,i} = 1, i = 1, 2, \ldots, N$, one obtains the partition function $Z(N) \equiv Z_{1,N}$ of the longest segment in $O(N^3)$ elementary computations. At $T = 0$, the following equation can be used instead to calculate the ground state energies $E_{i,j}$,

$$E_{i,j} = \min_{i\le k<j}\{E_{i,k-1} + E_{k+1,j-1} + \epsilon_{k,j}\}, \tag{5}$$

where as a convention we set $\epsilon_{i,i} = 0$ for all i, and $E_{i,j} = 0$ for $i \ge j$.

III. NUMERICAL RESULTS

A. Ground state properties

In this section we present numerical results regarding the ground state of an RNA molecule in the random sequence ensemble. Sequences upto $N = 2048$ bases are investigated, with a minimum of 1000 realizations of random sequences. Results for shorter sequences are obtained as a byproduct in the computation.

Pairing statistics – The statistical characteristics of ground state secondary structures have been studied in detail in Ref.[7] using the "mountain diagram". In this representation, a given secondary structure is mapped to a height profile following a simple rule: starting from one end of the chain, say $i = 0$ with $h_0 = 0$, one proceeds successively to the right, setting $h_i = h_{i-1}+1$ ($h_i = h_{i-1}-1$) if if base i is paired with base $j > i$ ($j < i$), and $h_i = h_{i=1}$

if base i is unpaired. Bundschue and Hwa have shown that the average value of h_i as defined above grows as a power-law of the sequence length N, $\overline{h} \sim N^{\zeta}$, where the "roughness exponent" $\zeta = \zeta_g = 0.67 \pm 0.02$, considerably larger than its value $\zeta_0 = 0.5$ in the molten phase.

Here we consider another measure of the pairing pattern, i.e., the distribution $P(d)$ of pairing distance d along the sequence. The black curve in Fig. 1 shows $P(d)$ against d in the ground state. Apart from the finite-size effect at d close to $N = 2048$, the data can be well-fitted by a power-law $P(d) \sim d^{-4/3}$. In comparison, $P(d)$ in the molten phase decays as $d^{-3/2}$.

When two bases at positions i and $j > i$ form a pair, their heights $h_i = h_j$, and there is no other site inbetween i and j at the same height. Therefore the distance $d = j - i$ can be considered as the first return time of the height function h_i. Let x be the exponent for the power-law decay of $P(d)$, the scaling analysis as presented in the Appendix yields,

$$\zeta + x = 2 \tag{6}$$

in agreement with the data.

We can have a direct argument linking the two exponents. We consider base pairs in the size range $L/2$ to L. The probability that a given base on the sequence is paired with another base at a distance in this size range is proportional to L^{1-x}. In an interval of size L, there are thus of the order of $L \times L^{1-x} = L^{2-x}$ pairs in this size range. Since such pairs extend at least half the interval size, they all contribute to the height difference between the site at the middle of the interval to either end. This again suggests $\zeta = 2 - x$.

For a sequence running from site i to site j, we may consider the size of the "first split" where, in the ground state, the base at the right end pairs with a site k to split the sequence into two halves. The distribution of the distance between k and j, which is readily obtained when computing the minimum of Eq. (5), is shown in Fig. 1. Again, the data can be fitted well to a power-law with an exponent $-4/3$.

Ground state energy – In addition to the scaling of the height profile, Bundschuh and Hwa proposed to examine the free energy cost for imposing a pairing (termed "pinching") to distinguish the molten and glass phases. In the molten phase, the pinching free energy ΔF grows with the pair size N as $\frac{3}{2}T \ln N$, and hence is purely entropic. Based on an estimate of the energy gain for the best matched duplex forbidden by the pinch, Bundschuh and Hwa argued that this behavior cannot continue below a certain temperature, and hence a glass transition is expected to take place. Their numerical results suggest that the disorder averaged ΔF grows as a small power of N, though the exponent cannot be determined conclusively.

A pinch isolates one part of an RNA molecule from the rest, and hence $\Delta F(N)$ can be identified with finite-size corrections to the free energy of a sequence of length N. Figure 2 shows $\overline{\Delta E}(N) \equiv 2\overline{E}(N) - \overline{E}(2N)$ against

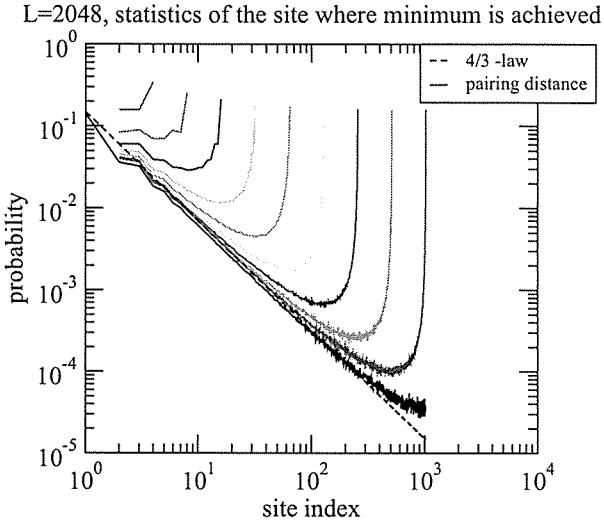

FIG. 1: Branching statistics in the ground state

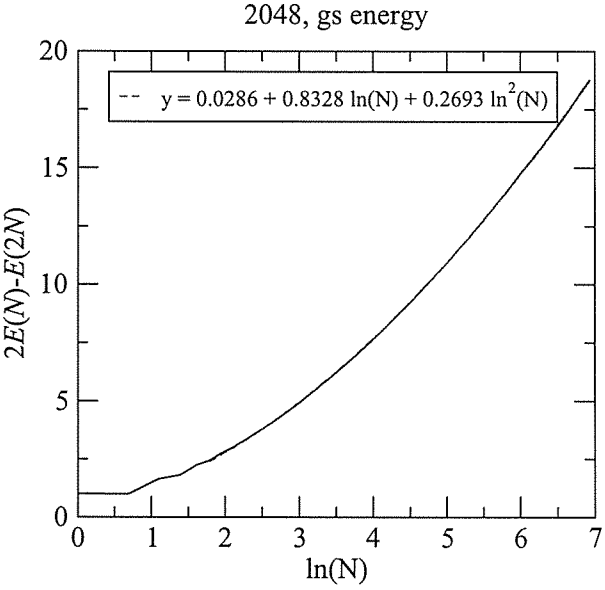

FIG. 2: Finite-size correction

$\ln(N)$ in the ground state. The data can be fitted nearly perfectly to the form

$$\Delta E(N) = a + b \ln N + c \ln^2 N. \tag{7}$$

Therefore the leading order finite-size correction to $\overline{E}(N)$ takes the form $\ln^2 N$. As we shall show below, the $\ln^2 N$ correction is present throughout the low temperature phase, but the coefficient c decreases with increasing T, and vanishes at the transition.

Another interesting property of the ground state is the dependence of pair energy on the pair distance, as illustrated in Fig. 3.

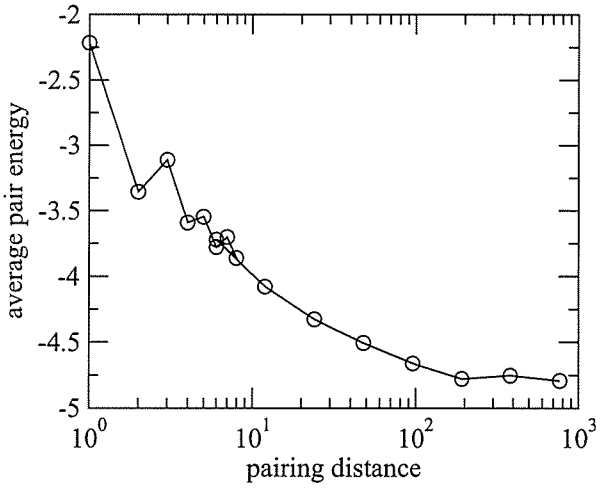

FIG. 3: Finite-size correction

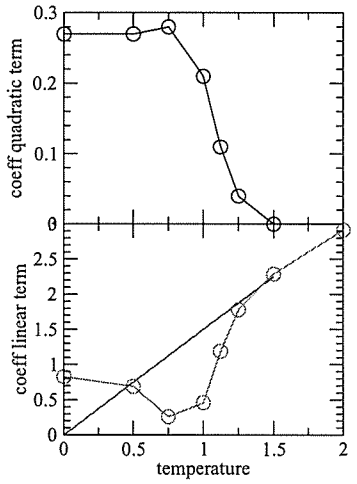

FIG. 4: coefficients of (a) ln^2N and (b) lnN terms against temperature

B. finite temperature properties and the glass transition

Statistics of the pinching energy holds the key to understanding the molten-to-glass transition and the nature of the glass state at low temperatures. In the molten state, the energy for pairing any two bases on the chain is expected to be finite. Pairing becomes sequence-specific only when this energy develops sufficient fluctuations. A localization of the base pairing pattern then takes place.

The system undergoes a glass transition at some finite temperature. Above the glass transition, $\delta F \sim \frac{3}{2}kT \ln N$, and $h \sim N^{1/2}$.

Research is supported in part by the Research Grants Council of the Hong Kong SAR under grants HKBU 2061/01P.Computations were carried out at HKBU's High Performance Cluster Computing Centre Supported by Dell and Intel.

* Electronic address: lhtang@hkbu.edu.hk

[1] For a review, see T. Garel, H. Orland, and E. Pitard, in *Spin Glasses and Random Fields*, A. P. Young Ed. (World Scientific, 1998), p. 387.

[2] *RNA Structure and Function*, Ed. by R. W. Simons and M. Grunberg-Manago (Cold-Spring Harbor 1998).

[3] *Nucleic Acids in Chemistry and Biology*, Ed. by G. M. Blackburn and M. J. Gait (IRL Press, Oxford, 1990); D. Voet and J. G. Voet, *Biochemistry*, 2nd Ed, (Wiley, 1995).

[4] P. G. Higgs, Phys. Rev. Lett. **76**, 704 (1996).

[5] A. Pagnani, G. Parisi and F. Ricci-Tersenghi, Phys. Rev. Lett. **84**, 2026 (2000).

[6] A. K. Hartmann, Phys. Rev. Lett. 86, 1382 (2001).

[7] R. Bundschuh and T. Hwa, Phys. Rev. Lett. **83**, 1479 (1999); Phys. Rev. E 65, 031903 (2002).

[8] M. L. M. Anderoson, Nucleic Acid Hybridization, Springer, New York (1998).

[9] M. Zuker and D. Sankoff, Bull. Math. Biol. 46, 591-621 (1984); M. Zuker, Science 244, 48-52 (1989).

[10] J. S. McCaskill, Biopolymers 29, 1105 (1990).

[11] P-G. de Gennes, Biopolymers 6, 715 (1968).

[12] F. Krzakala, M. Mèzard, and M. Müller, Europhys. Lett. 57, 752 (2002); M. Müller, F. Krzakala, and M. Mèzard, Eur. Phys. J. E 9, 67 (2002).

[13] T. Hwa and M. Lssig, Phys. Rev. Lett. 76, 2591 (1996).

[14] T. Hwa, Nature 399, 17 (1999).

[15] P. G. Higgs, Q. Rev. Biophys. 33, 199 (2000); R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison, Biological Sequence Analysis, (Cambridge University Press, Cambridge, England, 1998).