A Model for Folding and Aggregation in RNA Secondary Structures

(Dated: April 20, 2005)

We study the statistical mechanics of RNA secondary structures designed to have an attraction between two different types of structures as a model system for heteropolymer aggregation. The competition between the branching entropy of the secondary structure and the energy gained by pairing drives the RNA to undergo a '*temperature independent*' second order phase transition from a molten to an *aggregated phase*. The aggregated phase thus obtained has a macroscopically large number of contacts between different RNAs. The partition function scaling exponent for this phase is $\theta \approx 1/2$ and the crossover exponent of the phase transition is $\nu \approx 5/3$.

PACS numbers: 87.15.Aa, 87.15.Cc, 64.60.Fr

RNA secondary structures are an excellent model system to study the folding phenomenon in heteropolymers. Unlike the protein folding problem where a large number of different monomers are needed to be taken into account to understand folding [1], an RNA has just four bases A, U, C and G. The interaction schemes are simpler due to the separable energy scales of the secondary and the tertiary structure. These features, which result in an algorithm to calculate the partition function of folding in polynomial time [2], make the RNA secondary structures a both analytically and numerically amenable model to study various thermodynamic properties of heteropolymer folding [3–8].

The thermodynamic phases of such secondary structures generally depend on the temperature and the monomer specific binding free energies (which could in turn depend on the temperature themselves). At low enough temperatures, where the monomer specific binding energies and the sequence heterogeneity are important, the resulting (frozen) phase is *glassy* [4, 6]. At high temperatures, the large thermal fluctuations lead to an unbound *denatured phase*, where the secondary structure is randomly coiled (without any binding) analogous to a self avoiding random walk. At temperatures in between, where an effective attraction between short segments is important, the molecules are expected be in the so called *molten phase* [5, 6]. The molten phase corresponds to a large number of different secondary structures all having comparable energies (within $O(k_BT)$) coexisting in the configuration space. Another important phase of the secondary structure is the *native phase*, which is a certain specific folded structure favored by evolution [5, 8]. Many important questions have been raised with regard to these phases, e.g. their stability, characteristics and the phase transitions in the context of both protein folding and RNA folding [1, 3–8]. In this Letter we shall try to understand another important aspect of heteropolymers, the misfolding leading to *aggregation*, using the RNA secondary structure formulation.

The function of a biological molecule such as a protein or an RNA is dependent on how a given sequence of monomers folds. The failure of protein molecules to fold correctly is believed to be associated with various diseases such as Alzheimer's, Mad Cow and Parkinson's [9, 10]. The importance of this phenomenon has led to various studies in Protein misfolding and aggregation in various contexts [10, 11]. Here, we consider a toy model to study the phase transition of an RNA secondary structure from the molten to a suitably defined aggregated phase. Our focus here is on the thermodynamic properties of the system. Thus, we solve the model exactly in the thermodynamic limit and calculate the critical exponents relevant to the phase transition.

RNA is a biopolymer with four different monomers A, U, C and G in its sequence. The Watson-Crick pairs A-U and C-G are energetically the most favorable pairs while G-U is marginally stable and the other combinations are prohibited. By an RNA secondary structure, we mean a sequence of binding pairs (i, j) with $1 \leq i < j \leq N$, where N is the number of bases in the sequence. Any two pairs (i_1, j_1) and (i_2, j_2) are either nested, i.e. $i_1 < i_2 < j_2 < j_1$ or are independent i.e. $i_1 < j_1 < i_2 < j_2$. The above restriction means we are not allowing pseudo-knots, which are generally energetically not as favorable [12]. Such a secondary structure can be represented abstractly by a helix diagram, noncrossing arch diagram or a mountain representation as shown in Fig. 1.

Let the free energy associated with the pairing of bases i and j in an RNA be ϵ_{ij}. This free energy would have contributions from the gain in the energy due to binding and the associated configurational entropy loss. In addition to these, in principle there are also entropic and/or energetic effects due to loop formation, stacking, etc. Even though the accurate parameters as determined by the experiments are essential to calculate the exact secondary structure, such microscopic details as well as the exact values of the energies ϵ_{ij} do not affect the asymptotic properties of the phases and the critical exponents. Hence we ignore them in our model calculations.

If we denote the partition function for a sequence of bases from i to j as Z_{ij}, it can be evaluated exactly using the recursive relation [2, 13]:

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

with $j > i$ and the initial conditions $Z_{i,i} = Z_{i,i-1} = 1$ $\forall\ i$. This recursive relation can be computed in $O(N^3)$ time using a dynamic programming algorithm [2].

To understand the phase transition from the molten

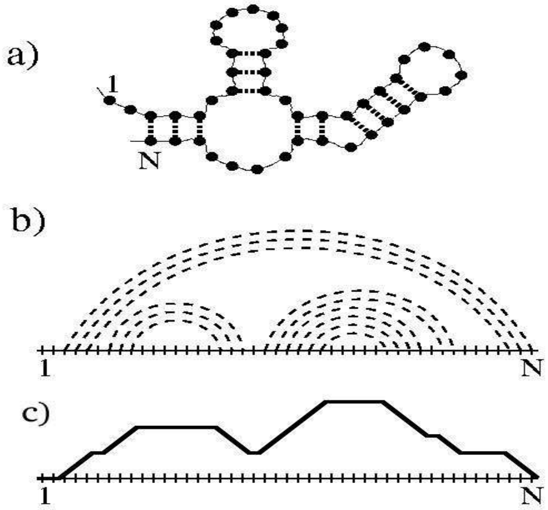

FIG. 1: Abstract representations of RNA secondary structures. (a) Helix representation (b) Non-crossing Arch diagram. Here, the solid line corresponds to the backbone of the RNA. The dashed arches correspond to the base pairs. The absence of pseudo-knots implies that the arches never cross. (c) Mountain Representation. Here, as we go along the backbone of the RNA from base 1 to N (represented by the base line), we go one step up for the beginning of a pair, one step down for the closing of a pair and a horizontal step for no pairing. Such a mountain never crosses the baseline and always returns to the baseline at the end.

to the aggregated phase, we first define the aggregated phase as an ensemble of RNA secondary structures in which a macroscopically large number of contacts occur between two different RNAs. We consider a dual RNA biomolecule system consisting of two types of RNA in a solution. We refer to them as RNA-1 and RNA-2. Individually, RNA-1 and RNA-2 are in the molten phase. However, when they are together in a solution, there is an effective attraction between the bases belonging to different RNAs. We study the phases of this dual RNA system, as the bias strength is varied.

To do so, we assume a simple pairing energy model with the free energy of pairing between bases i and j defined as:

$$\epsilon_{i,j} = \begin{cases} \epsilon_1 & \text{if } i,j \in \text{RNA-1} \\ \epsilon_2 & \text{if } i,j \in \text{RNA-2} \\ \epsilon_3 & \text{if } i \in \text{RNA-1}, j \in \text{RNA-2 or vice-versa} \end{cases} \tag{2}$$

Here, the intra RNA base pairing energies ϵ_1 and ϵ_2 could be of comparable magnitude in a realistic RNA molecule. The inter RNA base pairing energy, or the bias, ϵ_3 is the parameter which can in principle be controlled by sequence mutation. Note that in the spirit of the molten phase [5], the base as we call it here could be understood as a short segment consisting of several bases.

Denote the Boltzmann factors corresponding to the pairing energies by q_1, q_2 and q_3 respectively. We show that this simple model predicts a molten to an aggregated phase transition, as we tune the parameter q_3.

To keep the analytical calculations simple, we assume each RNA to be of equal length, containing $N-1$ bases [14]. We now consider the joint folding of these two RNAs and denote its partition function by $Z_d(N; q_1, q_2, q_3)$. As explained before, the free energy of pairing for the bases belonging to a given RNA has contributions from the energy gain due to the pairing and the entropy loss associated with the loop formation. This holds true even for pairing across the bases belonging to different RNAs. But when the first pairing between the bases belonging to different RNAs occur, there is an additional entropic loss due to the breakdown of translational invariance symmetry. Thereafter, only the free energy ϵ_3 plays a role in the inter RNA base pairing. In the thermodynamic limit, this additional entropic loss has no effect on the phase of the system, but it is the energetics of pairing that drives the phase transition. Hence, we ignore this additional entropic term. This essentially reduces the problem to the folding of a single sequence with $2N-2$ bases. The aggregated secondary structure can now be interpreted as having a macroscopically large number of contacts between the two halves of the concatenated RNA.

Let us first consider two special cases. Setting $q = q_1 = q_2 = q_3$ corresponds to the well known molten phase of the RNA secondary structure, whose partition function can be calculated exactly in the asymptotic form

$$Z_d(N; q, q, q) = Z_0(2N; q) = A(q)(2N)^{-\theta} z_c(q)^{2N} \tag{3}$$

with the characteristic scaling exponent $\theta = 3/2$ [5]. This exponent is characteristic in the sense that it is insensitive to various microscopic details of the RNA secondary structure such as the cost of a hairpin loop, weak sequence disorder and heterogeneity, etc. The other simple case is $q_3 = 0$. This case corresponds to having two RNAs in the molten phase which do not know of each other's presence. The partition function of such a dual RNA would then be just the product of individual partition functions, i.e. $Z_d(N; q_1, q_2, 0) \equiv Z_0(N, q_1)Z_0(N, q_2)$. Hence the scaling exponent is $\theta = 3$.

We now want to understand the case of general q_1, q_2 and q_3. To this end we calculate the partition function of the dual RNA as follows. Let the base pairings within a given RNA be called primary and those across different RNAs be called secondary. Any given secondary structure thus obtained has a series of secondary pairings $(i_1,j_1),\ldots,(i_k,j_k)$ such that $1 \le i_1 < \ldots < i_k \le N-1$ and $1 \le j_1 < \ldots < j_k \le N-1$. Note that we have labeled the RNA-1 by i and the RNA-2 by j indices. The bubbles thus formed between any two consecutive secondary pairings are allowed to have only the primary pairings. If all the secondary structure configurations are enumerated according to the number of the inter-RNA (or the secondary) contacts k, then the total partition function of this dual RNA system, in the Z-Transform representation can be written as:

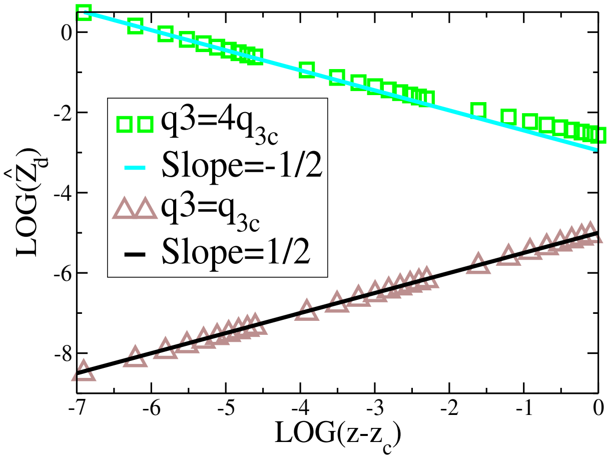

FIG. 2: (color online) The behavior of the partition function $\hat{Z}_d(z; q_1 = 4, q_2 = 9, q_3)$. For $q_3 = q_{3c}$, we observe a square root behavior. For $q_3 > q_{3c}$, we see an inverse square root behavior.

$$\hat{Z}_d(z; q_1, q_2, q_3) = \sum_{k=0}^{\infty} q_3^k \hat{Z}_0(z; q_1)^{k+1} * \hat{Z}_0(z; q_2)^{k+1} \quad (4)$$

$$= \oint \frac{dz'}{z'} \frac{\hat{Z}_0(z'; q_1)\hat{Z}_0(z/z'; q_2)}{1 - q_3\hat{Z}_0(z'; q_1)\hat{Z}_0(z/z'; q_2)} \quad (5)$$

where $\hat{Z}_d(z; q_1, q_2, q_3)$ and $\hat{Z}_0(z; q)$ are the Z-Transforms of $Z_d(N; q_1, q_2, q_3)$ and $Z_0(N; q)$ respectively. The symbol $*$ indicates the convolution in z-space defined as $f * g = \oint \frac{dz'}{z'} f(z')g(z/z')$. Eq.(5) is obtained by summing up the geometric series in Eq.(4). The convolution integration can be done numerically to obtain the singularities of $\hat{Z}_d$ and hence, the asymptotic behavior of $Z_d(N; q_1, q_2, q_3)$.

The results are shown in Fig. 2. For $q_3 = q_{3c} = \sqrt{q_1 q_2}$, we find a square root singularity and hence $\theta = 3/2$ [15], the characteristic exponent of the molten phase. For $q_3 > q_{3c}$, $\hat{Z}_d$ has an inverse square root singularity, indicating a new phase. We interpret the new phase with the partition function scaling exponent $\theta \approx 1/2$ as the aggregated phase. For $q_3 \gtrsim q_{3c}$, its hard to extract the singularity, but our claim that it remains the aggregated phase is verified by the numerical calculations shown below. Similarly, we claim that for all $q_3 < q_{3c}$, the dual RNA system is just the phase corresponding $q_3 = 0$ in the asymptotic limit, hence $\theta = 3$. This claim is again verified by numerical calculations of the exact partition function for finite length and the calculation of an asymptotic macroscopic quantity (the order parameter) to be defined below. The resulting simple phase diagram is shown in Fig. 3.

In order to verify that the phase transition indeed happens at $q_{3c} = \sqrt{q_1 q_2}$, we calculate the order parameter of the phase transition. Here, the order parameter Q is defined as the fraction of secondary pairings in a secondary structure. For arbitrary q_3 the order parameter can be calculated exactly from $Q = -\lim_{N\to\infty}(d \ln Z_d/d \ln q_3)/N$. The inset of Fig. 4 clearly shows $Q = 0$ for $q_3 \leq q_{3c}$ and continuously increasing with q_3 thereafter saturating to $Q = 1$ for $q_3/q_{3c} \gg 1$. From this behavior of the order parameter we can conclude that the phase transition indeed occurs at $q_{3c} = \sqrt{q_1 q_2}$ and that the phase transition is of second order. Physically, we can understand the behavior of the order parameter by using the mountain representation of RNA (see Fig. 1c). Between any two consecutive secondary pairings, the contribution of primary pairs to the height of the mountain is zero. Hence, the total number of secondary pairings is equal to the height $\langle h \rangle$ of the mountain at its midpoint. Using the random walk analogy [6, 16], we find that $\langle h \rangle \sim O(N^{1/2})$, hence $Q \sim O(N^{-1/2})$. For $q_3 < q_{3c}$, the secondary pairings are even less likely, and hence in the thermodynamic limit $Q = 0$ for $q_3 \leq q_{3c}$, consistent with what we have obtained by exact expression.

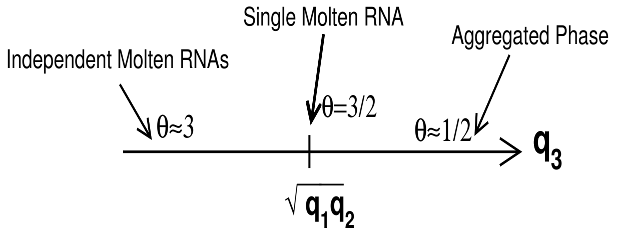

FIG. 3: The phase diagram for the dual RNA system.

To further verify our claims about the phase for $q_3 < q_{3c}$ and to calculate the scaling exponents corresponding to the second order phase transition, we iterated the recursion relation (Eq.(1)) to calculate the exact partition function for RNA of finite length N. The results of the numerical calculations are in complete agreement with the phase diagram of Fig. 3 when extrapolated to the thermodynamic limit, thus verifying our claim. Next we calculate the free energy per length $f(q_1, q_2, q_3) = -\ln Z_d(N)/N$, taking into account the finite size effects. We assume the usual scaling function for the order parameter $Q(N) = N^{-1/2}g[(q_3 - q_{3c})N^{1/\nu}]$ close to the critical point. Fig. 4 shows the result of scaling plot, with the best fit value for the crossover critical exponent $\nu \approx 5/3$.

Throughout this study our focus has been the thermodynamic properties of the transition. Yet it would be interesting to check its applicability to realistic sequences. To do so, we performed numerical simulations of two selected sequences. In the first case, we take $(ACU)_n$ and $(AGU)_n$ as our two RNA sequences. We assume $\epsilon_{CG} = 2\epsilon_{AU} = -2$ and $\epsilon_{GU} = 0$ which is approximately true at the temperature 37 °C. If we consider ACU and AGU as short segments of our coarse graining approximation, we observe that there is a relatively larger effective attraction towards the segment of the other RNA. As expected, our numerical simulations does indeed show an

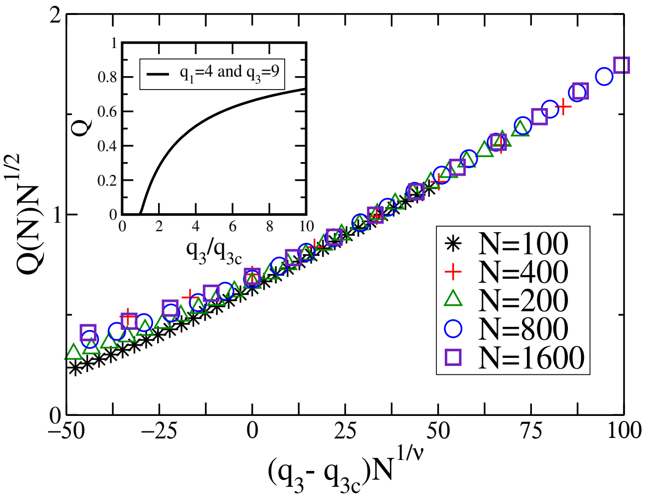

FIG. 4: (color online) Scaling plot for the order parameter. Inset shows the order parameter of the phase transition. In both the plots $q_1 = 4$ and $q_2 = 9$, hence $q_{3c} = 6$.

aggregated phase, i.e $\theta \approx 1/2$, in the asymptotic limit. The phase corresponding to $\theta = 3$ is observed in designed sequences such as $(AU)_n$ and $(CG)_n$ as RNA-1 and RNA-2 respectively.

This model has some similarities with Go-like model studied by Bundschuh and Hwa [5] which shows a molten-native transition. The physics behind the phase transition in their model as well as our model is the same, i.e., the competition between the energetic gain of the secondary contacts (or native contacts of Go-like model) and the branching entropy. But, contrary to the native phase where the ground state is unique, the aggregated phase has degenerate ground states. On the other hand, both these models can 'melt' from their (aggregated or native) ground state to any of the molten, glassy or denatured phase, depending on the temperature and the strength of the bias. The differences in the behavior of these models arises from the fact that for the Go-like model the bias is site specific where as for the model we have presented, the bias is towards a macroscopically large number of sites.

In summary, we have presented a simple model for heteropolymer folding using the RNA secondary structure formulation, which shows a second order phase transition from an *independently molten* to an *aggregated phase*. The behavior exactly at the criticality turns out to be the molten phase for the concatenated molecule. The transition is completely driven by the energetics of pairing and is temperature independent. For the case where the free energy of pairing is temperature dependent, our model would imply that at a given temperature, when the average attraction between pairs of different molecules exceeds a certain threshold, the aggregation would occur. Proteins are known to undergo a folding transition from native to an aggregated phase instead of from a molten to an aggregated phase [10]. It should be interesting to see if this study can be extended to understand the thermodynamics of such a phase transition. It should also be interesting to study the role of kinetics of RNA folding in this phase transition.

We gratefully acknowledge useful discussions with Tsunglin Liu. RB is supported by the National Science Foundation through grant no. DMR-0404615.

[1] K. A. Dill *et al.*, Protein Sci. **4**, 561 (1995); J. N. Onuchic *at al*, Annu. Rev. Phys. Chem. **48**, 545 (1997); T. Garel *et al.*, J. Phys. I **7**, 27 (1997); E. I. Shakhnovich, Curr. Opin. Struct. Biol. **7** 29(1997).

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

[3] P. G. Higgs, Q. Rev. BioPhys. **33**, 199 (2000).

[4] P. G. Higgs, Phys. Rev. Lett. **76**, 704 (1996); A. Pagnani, G. Parisi and F. Ricci-Tersenghi, Phys. Rev. Lett. **84**, 2026 (2000); A. K. Hartmann, Phys. Rev. Lett. **86**, 1382 (2001); A. Pagnani, G. Parisi and F. Ricci-Tersenghi, Phys. Rev. Lett. **86**, 1383 (2001); F. Krzakala, M. Mézard and M. Müller, Europhys. Lett. **57**, 752 (2002); E. Marinari, A. Pagnani and F. Ricci-Tersenghi, Phys. Rev. E. **65**, 041919 (2002).

[5] R. Bundschuh and T. Hwa, Phys. Rev. Lett. **83**, 1479 (1999).

[6] R. Bundschuh and T. Hwa, Europhys. Lett. **59** 903 (2002); R. Bundschuh and T. Hwa, Phys. Rev. E, **65**, 031903 (2002).

[7] H. Orland and Z. Lee, Nucl. Phys. B, **620**, 456 (2002); R. Mukhopadhyay *et al.*, Phys. Rev. E. **68**, 041904 (2003); M. Baiesi *et al.* Phys.Rev. Let **91**, 198102 (2003); P. Leoni and C. Vanderzande, Phys. Rev. E. **68**, 051904 (2003).

[8] P. G. Higgs, J. Phys. I (France), **3**, 43 (1993).

[9] S.B. Prusiner, in *Prion Biology and Diseases*, ed. S.B. Prusiner (Cold Spring Harbor Laboratory Press, Cold Spring Harbor NY, 1999), p. 1.

[10] for a review see: C. M. Dobson, Nature, **446**, 884 (2003).

[11] A. Slepoy, R.R.P. Singh, F. Pazmandi, R.V. Kulkarni, D.L. Cox , Phys. Rev. Lett. **87**, 058101 (2001).

[12] I. Tinoco Jr. and C. Bustamante, J. Mol. Biol. **293**, 271 (1999) and references therein.

[13] M. S. Waterman, *Advances in Mathematics, Supplementary studies*, edited by G.-C. Rota (Academic, New York, 1978), pp.167-212.

[14] The equal length approximation is not necessary. In general (at least) as long as the RNAs are of the order of same length, all the subsequent results hold.

[15] If the Z-transform of the partition function shows a power law singularity at z_c, say $\hat{Z}(z) \sim (z - z_c)^{\alpha}$, then the corresponding partition function scaling exponent $\theta = \alpha + 1$. See App. A of Ref T. Liu and R. Bundschuh, Phys. Rev. E, **69**, 061912 (2004) for derivation.

[16] See, eg. W. Feller, *An Introduction to Probability and its Applications* (Wiley, New York, 1950).