Dependence of RNA secondary structure on the energy model

Bernd Burghardt* and Alexander K. Hartmann†

Institut für Theoretische Physik, Universität Göttingen,
Friedrich-Hund-Platz 1, D–37077 Göttingen, Germany

(Dated: September 7, 2004)

We analyze a microscopic RNA model, which includes two widely used models as limiting cases, namely it contains terms for bond as well as for stacking energies. We numerically investigate possible changes in the qualitative and quantitative behaviour while going from one model to the other; in particular we test, whether a transition occurs, when continuously moving from one model to the other. For this we calculate various thermodynamic quantities, both at zero temperature as well as at finite temperatures. All calculations can be done efficiently in polynomial time by a dynamic programming algorithm. We do not find a sign for transition between the models, but the critical exponent ν of the correlation length, describing the phase transition in all models to an ordered low-temperature phase, seems to depend continuously on the model. Finally, we apply the ε-coupling method, to study low excitations. The exponent θ describing the energy-scaling of the excitations seems to depend not much on the energy model.

PACS numbers: 64.60.Fr, 87.15.Aa

I. INTRODUCTION

RNA plays an important rule in the biochemistry of all living systems [1, 2]. Similar to the DNA, it is a linear chain-molecule build from four types of bases, i.e., adenine (A), cytosine (C), guanine (G), and uracil (U). It does not only transmit pure genetic information, but, e.g., works as a catalyst. While for the former the primary structure, i.e., the sequence of the bases, is relevant, for the later the kind of higher order structures, i.e., secondary and tertiary structures, are relevant.

Like in the double helix of the DNA, in RNA complementary bases can build hydrogen bonds between each other. Compared to DNA, where the bonds are built between two different strands, RNA builds bonds between bases of the same RNA strand. The information, which bases of the strand are paired, gives the secondary structure, and the spatial structure is called the tertiary structure. The tertiary structure is stabilized by a much weaker interaction than the secondary structure. This leads to a separation of energy scales between secondary and tertiary structure, and gives the justification to neglect the later [3]. Therefore we deal here with the secondary structure only.

One crucial point in calculating the secondary structure is the used energy model: On the one hand, if one aims to get minimum structures close to the experimentally observed one, one uses energy models that take into account structure elements [4–7], e.g., hair pin loops. On the other hand, if one is interested in the qualitative behaviour, one uses models as simple as possible that keep the general behaviour, e.g., only one kind of base [8] or using energies depending only on the number and on the type of paired bases [9–12]. Here we will consider only models with the later kind of interaction energy. In recent years several authors examined this kind of models with regard to the thermodynamic behaviour, i.e., searching for phase transitions and describing the type of phase involved [9, 11–13]; Liu and Bundschuh [8, 14] recently discussed, whether native RNA is already in the regime of the thermodynamic limit or finite size effect have to be taken into account. In this paper we numerically investigate a hybrid model of two well known energy models [15, 16], i.e., a *pair energy model*, where only base pairs are considered regardless of their neighbourhood, and a *stacking energy model*, where only consecutive paired bases, i.e., forming a stack, gives an energy contribution. It has been claimed that the stacking energy is more relevant than just the pair energy in real RNA [17]. Our model contains terms for *both* types of interactions and allows to move continuously from one model to the other. We are interested, whether the two limiting models are qualitatively different, in particular, whether a phase transition occurs, when moving from one model to the other.

The paper is organized as follows. In section section II, we define our model, i.e., we formally define secondary structures and introduce our energy model. In section II B, we explain how to calculate the partition function with a dynamic programming algorithm. In section III, we introduce the observables which we investigate in the following section IV. While in section IV B and section IV C we do finite temperature calculations, in section IV D we use the ε-coupling method at zero temperature.

II. THE MODEL

Because RNA molecules are linear chains of bases, they can be described as a (quenched) sequence $\mathcal{R} =$

*Electronic address: burghard@theorie.physik.uni-goettingen.de

†Electronic address: hartmann@theorie.physik.uni-goettingen.de

$(r_i)_{i=1,\ldots,L}$ of bases $r_i \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{U}\}$, where L is the length of the sequence. Within this single stranded molecule some bases can pair and build a secondary structure. Typically Watson-Crick base pairs, i.e., A-U and C-G have the strongest affinity to each other, they are also called complementary base pairs. Each base can be paired at most once. For a given sequence $\mathcal{R}$ of bases the secondary structure can be described by a set $\mathcal{S}$ of pairs (i,j) (with the convention $1 \le i < j \le L$), meaning that bases r_i and r_j are paired. For convenience of notation we further define a Matrix $(S_{i,j})_{i,j=1,\ldots,L}$ with $S_{i,j} = 1$ if $(i,j) \in \mathcal{S}$, and $S_{i,j} = 0$ otherwise. Two restriction are used: (i) Here we exclude so called *pseudo-knots*, that means, for any $(i,j),(i',j') \in \mathcal{S}$, either $i < j < i' < j'$ or $i < i' < j' < j$ must hold, i.e., we follow the notion of pseudo knots being more an element of the tertiary structure [17].

(ii) Between two paired bases a minimum distance is required: $|j-i| \ge s$ is required, granting some flexibility of the molecule (here $s = 2$).

A. Energy models

Every secondary structure $\mathcal{S}$ is assigned a certain energy $E(\mathcal{S})$; note that this energy in general depends on the $\mathcal{R}$ as well, so it is more precisely to write $E(\mathcal{S},\mathcal{R})$, but we assume that the structure also includes the information about the sequence. With such an energy model it is possible to calculate the canonical partition function Z of a given sequence $\mathcal{R}$ by summing over all possible structures $Z = \sum_{\mathcal{S}} \mathrm{e}^{-\beta E(\mathcal{S})}$, but it is computationally more efficient to compute it by using the partition functions of the subsequences, i.e., by a dynamic programming approach.

Motivated by the observation that the secondary structure is due to building of numerous base pairs where every pair of bases is bound by hydrogen bonds, one assigns each pair (i,j) a certain energy $e(r_i,r_j)$ depending only on the kind of bases. The total energy is the sum over all pairs

$$E_p(\mathcal{S}) = \sum_{(i,j)\in\mathcal{S}} e(r_i,r_j)\,, \tag{1}$$

e.g., by choosing $e(r,r') = +\infty$ for non-complementary bases r, r' pairings of this kind are suppressed.

Another possible model is to assign an energy E_s to a pair $(i,j) \in \mathcal{S}$ iff also $(i+1,j-1) \in \mathcal{S}$. This *stacking energy* can be motivated by the fact that a single pairing gives some gain in the binding energy, but also reduces the entropy of the molecule, because through this additional binding it looses some flexibility. Formally the total energy of a structure can be written as

$$E_s(\mathcal{S}) = \sum_{(i,j)\in\mathcal{S}} E_s S_{i+1,j-1}\,, \tag{2}$$

assuming that for every pair $(i,\,j) \in \mathcal{S}$ the bases r_i and r_j are complementary bases. The total number t of consecutive base pairs is called the *stacking size.* Single base pairs are not considered as stacks, therefore $t \ge 2$ for any stack.

Both types of energy models are discussed in the literature [15, 16], but, to our knowledge, so far no one has discussed a hybrid model. We examine the sum of both models at once.

$$E(\mathcal{S}) := E_p(\mathcal{S}) + E_s(\mathcal{S}) \tag{3}$$

where the parameters E_s and $e(r,r')$ can be adjusted freely, including both models discussed above. Here we use

$$e(r,r') = \begin{cases} E_p & \text{if } r \text{ and } r' \text{ are compl. bases} \\ +\infty & \text{otherwise} \end{cases} \tag{4}$$

with a pair energy $E_p \le 0$ independent of the kind of bases.

Due to the simple structure of the energy model, e.g., the energies depend not on the position of the bases within the sequence or whether the paired bases include some structure elements like hairpins, the ground state is highly degenerated [9, 13].

B. Calculating the partition function

Due to the fact that pseudo knots are excluded from our model (see section II A), the calculation of the partition function can be done recursively. The algorithm is similar to that of Nussinov [18, 19]. The algorithms calculates the elements of two upper-triangular matrices $(Z_{i,j})_{1\le i\le j\le L}$ and $(\hat{Z}_{i,j})_{1\le i\le j\le L}$, where $Z_{i,j}$ is the partition function of the subsequence from base r_i to r_j under the boundary condition that bases r_{i-1} and r_{j+1} are not paired, and $\hat{Z}_{i,j}$ the partition function under the boundary condition that bases r_{i-1} and r_{j+1} are complementary; $\hat{Z}_{i,j}$ is only used as an auxiliary matrix. Then $Z_{i,j}$ can be computed from the partition functions Z and $\hat{Z}$ of smaller subsequences in the following way (remember that s denotes the minimum distance between two bases of a pair)

$$\begin{aligned} Z_{i,j} &= Z_{i,j-1} + \sum_{k=i}^{j-s-1} Z_{i,k-1}\,\mathrm{e}^{-e(r_k,r_j)/k_{\mathrm{B}}T}\,\hat{Z}_{k+1,j-1} \\ \hat{Z}_{i,j} &= Z_{i,j-1} + \mathrm{e}^{-(e(r_i,r_j)+E_s)/k_{\mathrm{B}}T}\,\hat{Z}_{i+1,j-1} \\ &\quad + \sum_{k=i+1}^{j-s-1} Z_{i,k-1}\,\mathrm{e}^{-e(r_k,r_j)/k_{\mathrm{B}}T}\,\hat{Z}_{k+1,j-1} \\ Z_{i,j} &= \hat{Z}_{i,j} = 1 \quad \text{for } i \ge j \\ \hat{Z}_{i,j} &= 0 \quad \text{for } 0 < j-i < s-2 \end{aligned} \tag{5}$$

which is schematically explained in Fig. 1. Because both

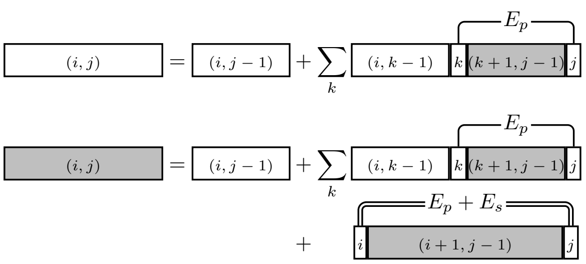

FIG. 1: Schematical explanation of eq. (5) and eq. (6), e.g., white boxes represent Z, gray boxes $\hat{Z}$.

matrices depend on each other, they must be calculated simultaneously, starting along the diagonal and continuing along the off-diagonals. The calculation of the partition function can be done in $\mathcal{O}(L^3)$ steps, where L is the length of the sequence. The partition function Z of the entire sequence is $Z_{1,L}$, but also the other matrix elements are useful for generating ensembles of structures according to the Boltzmann distribution (see section III).

A similar algorithm can be used to calculate the ground state energy:

$$
\begin{aligned}
N_{i,j} &= \min(N_{i,j-1}, \min_{k=i}^{j-s-1} (N_{i,k-1} + e(r_k, r_j)) + \hat{N}_{k+1,j-1}) \\
\hat{N}_{i,j} &= \min(N_{i,j-1}, e(r_i, r_j) + E_s + \hat{N}_{i+1,j-1}, \\
&\qquad \min_{k=i+1}^{j-s-1} (N_{i,k-1} + e(r_k, r_j) + \hat{N}_{k+1,j-1})) \\
N_{i,j} &= \hat{N}_{i,j} = 0 \quad \text{for } i \geq j \\
\hat{N}_{i,j} &= +\infty \quad \text{for } 0 < j - i < s - 2
\end{aligned}
\tag{6}
$$

In comparison to eq. (5) additions are replace by min-operations, multiplications by additions and the exponentials of the energies by the energies them self.

III. OBSERVABLES

After calculating the partition function Z for a given random sequence, we want to measure some quantities to compare the members of the ensemble. In principle most quantities could be calculated by a similar dynamic programming algorithm introduced above, but in general the running time behaviour would be of higher order (than three) in the sequence length. For this reason we use a different method, where an ensemble of structures is generated due to its Boltzmann weight [10]. The procedure to build a sequence is essentially a backtracking algorithm: Starting with the entire sequence a partner for an outermost base, e.g., base L, is chosen with the appropriate probability, e.g., base k, after this the procedure is applied to the subsequences 1 to $k-1$ and $k+1$ to L. If base L is chosen not to be paired at all, one uses the sequence shortened by base L. Considering a subsequence from base k to l, the probability $p_{i,j;k,l}$ that bases i and j ($k \leq i < j \leq l$) are paired is given by

$$
p_{i,j;k,l} = \begin{cases} Z_{k,l}^{-1}\, Z_{k,i-1} \mathrm{e}^{-(e(r_k,r_j)+E_0)/k_\mathrm{B}T}\, \hat{Z}_{j+1,l} \\ \qquad \text{bases } i-1 \text{ and } j+1 \text{ paired} \\ \\ Z_{k,l}^{-1}\, Z_{k,i-1} \mathrm{e}^{-e(r_k,r_j)/k_\mathrm{B}T}\, \hat{Z}_{j+1,l} \\ \qquad \text{bases } i-1 \text{ and } j+1 \text{ unpaired} \end{cases} \tag{7}
$$

For each member of this ensemble $\mathcal{E}$ the quantity of interest X is calculated and the average $\langle X \rangle = \frac{1}{|\mathcal{E}|}\sum_{\mathcal{S}} X(\mathcal{S})$ is used as an approximation to the expectation value of the Gibbs-Boltzmann-ensemble; for large enough ensembles $\mathcal{E}$ this average approaches the true expectation value.

A simple observable is the energy E and its fluctuations $(\Delta E)^2$, the later is directly connected with the specific heat $c_V = (\Delta E)^2 / L\, k_\mathrm{B} T^2$.

Of particular importance is the overlap q between two structures $\mathcal{S}$ and $\mathcal{S}'$

$$
q(\mathcal{S}, \mathcal{S}') := \frac{2}{L} \sum_{1 \leq i < j \leq L} S_{i,j}\, S'_{i,j} \tag{8}
$$

that is the number of bases paired to the same base in both structures normalized such that q lies always between zero and one. This is a measure of how similar two structures are. Note however, that with this definition the overlap of one structure with itself is $q(\mathcal{S}, \mathcal{S}) \lneqq 1$ unless all bases are paired, where it is equal to one. A definition of q where also bases unpaired in both structures are counted is used in [10] resulting in an overlap definition that is normalized, however this similarity measurement has the drawback that the less bases are paired the more two structures get similar. We further remark that for any two structures $\mathcal{S}$, $\mathcal{S}'$ the Cauchy-Schwarz inequality $(q(\mathcal{S}, \mathcal{S}'))^2 \leq q(\mathcal{S}, \mathcal{S})\, q(\mathcal{S}', \mathcal{S}')$ holds. With this quantity q two ensembles $\mathcal{E}$ and $\mathcal{E}'$ can be compared, e.g., looking at the distribution of $q(\mathcal{S}, \mathcal{S}')$ of all $\mathcal{S} \in \mathcal{E}, \mathcal{S}' \in \mathcal{E}'$.

The ensemble averages $\langle \cdot \rangle_{\mathcal{E}}$ in general depend on the chosen sequence, therefore a further averaging over several (random) sequences is required. This sequence average is denoted by $[\cdot]$. We again approximate this average by summing over a finite set of sequences.

Because of this two stage averaging, it is probably preferable instead of looking at $[\langle q \rangle]$ directly to use functions of the first and higher moments of q, e.g., the Binder cumulant [20–22], which is defined by

$$
B := \frac{1}{2}\left(3 - \frac{[\langle q^4 \rangle]}{\left[\langle q^2 \rangle^2\right]}\right) \tag{9}
$$

where $\langle q^n \rangle$ is either the average over all pairs of one ensemble

$$
q_{\mathcal{E}} := \frac{1}{|\mathcal{E}|\,(|\mathcal{E}|-1)} \sum_{\substack{\mathcal{S},\, \mathcal{S}' \in \mathcal{E} \\ \mathcal{S} \neq \mathcal{S}'}} q(\mathcal{S}, \mathcal{S}') \tag{10}
$$

or the average over all pairs of two ensembles

$$q_{\mathcal{E},\,\mathcal{E}'} := \frac{1}{|\mathcal{E}|\,|\mathcal{E}'|} \sum_{\mathcal{S}\in\mathcal{E},\,\mathcal{S}'\in\mathcal{E}'} q(\mathcal{S},\,\mathcal{S}')\,. \tag{11}$$

The later choice is appropriate when one is looking for a change in the behaviour of the models when one varies the parameters in comparison to a reference model, while the former is the better choice, if an external parameter, e.g., the temperature, is varied.

The Binder cumulant B vanishes at high temperatures, while for low temperatures it approaches a finite value in the thermodynamic limit.

A similar quantity has been used in [9]:

$$A := \frac{\left[\chi_{\mathcal{R}}^2\right] - \left[\chi_{\mathcal{R}}\right]^2}{\left[\chi_{\mathcal{R}}\right]^2} \tag{12}$$

where

$$\chi_{\mathcal{R}} := L\,\left(\left\langle q^2\right\rangle - \left\langle q\right\rangle^2\right) \tag{13}$$

is the variance of the q distribution. This parameter A measures how the probability distribution of q varies from sequence to sequence. A value close to zero indicates a self-averaging behaviour.

IV. NUMERICAL RESULTS

In order to find some possible differences in the behaviour of the energy model eq. (3) for different parameters E_p, E_s, as introduced in eq. (4) and eq. (2), respectively, we numerically calculated the quantities mentioned in section III above. In all our examples we averaged over randomly generated sequences $\mathcal{R} = (r_i)$, where the probability for a specific base r_i at position i is $\frac{1}{4}$ for all base types independent of the other bases $r_{j\neq i}$. The size of the sequence varied from $L = 128$ up to $L = 1024$, for the disorder average 2000 up to 6000 random sequences were used. Pairing of bases are only allowed for complementary bases and the minimum distance between to bases was chosen as $s = 2$.

In section IV A and IV B, we vary continuously the energy parameters between the two extreme cases ($E_p = 0, E_s = -1$) and ($E_p = -1, E_s = 0$). In section IV A we shortly discuss the averages of the stacking size t and of the q for different energy parameters. In section IV B we examine the cross overlap $q_{\mathcal{E},\,\mathcal{E}^{\mathrm{ref}}}$, see eq. (11), between a reference ensemble $\mathcal{E}^{\mathrm{ref}}$ generated for fixed energy parameters, and ensembles generated for different energy parameters. In the following section IV C we look at the temperature variation of various quantities without using any reference ensemble to estimate some critical parameters. In the last section IV D we apply the ε-coupling method at $T = 0$ to determine the critical exponent θ describing the behavior of low-lying excitations.

A. Basic observables

In section II A, where we introduced the energy model, we opposed the pair energy to the stacking energy model. In Fig. 2 we show the average size t of a stacking as a function of the energy parameter E_p at temperature $T = 0$. We keep $E_p + E_s = -1$ constant, to fix the overall energy scale. For all fixed energy parameters the

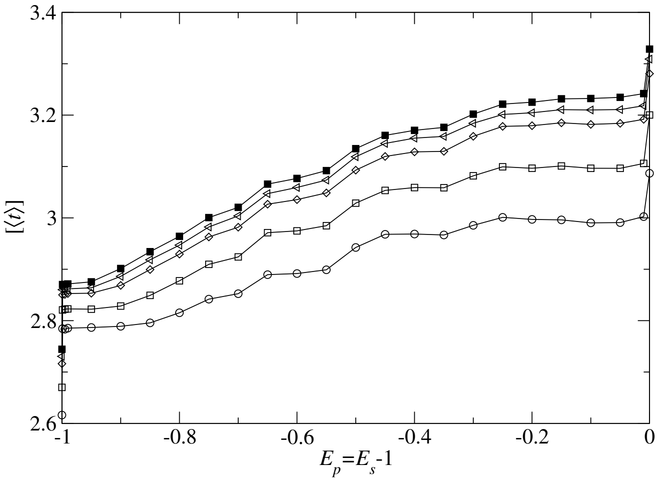

FIG. 2: Average stacking size as a function of energy parameter E_p for different system sizes. At $E_p = -1$, i.e., $E_s = 0$, and $E_p = 0$ the curves are discontinuous. For explanation of symbols see caption of Fig. 7. Missing error bars are of the size of the symbols or smaller, and omitted for legibility.

average stacking size t increases with increasing system size, which is expected as with increasing length the probability for longer stacks increases. Also as expected is the overall increase in the average stacking size with the energy parameter E_s, because the stacking energy prefers to build stacks. The large increase of t while changing E_s from zero to a nonzero value can be explained as following: The ground state for $E_p = -1.0, E_s = 0.0$ is highly degenerated, while changing to a nonzero E_s only those states stay ground states which have a high stacking contribution to the energy. This selection increases the average of t. A similar argument applies at the other end, where E_p changes from a nonzero value to zero.

Similarly the overlap q jumps to a larger value when changing from $E_p = 0$ to a nonzero value (Fig. 3). In addition at positions, where E_p/E_s are fractions with small numerator and denominator, e.g., $\frac{1}{1}$ or $\frac{1}{3}$, for this energy parameters the ground states in configuration space are more broadly distributed than for slightly different parameters. This can be seen in the right inset of Fig. 3, where q-Distribution in the symmetric case ($E_s = E_p = -0.5$) is broader than in the slightly asymmetric case ($E_s = -0.49, E_p = -0.51$). The width of this minima in the main plot, as well as that in Fig. 4 and Fig. 5, is below $\Delta E = 0.001$, as one can estimate from the left inset of Fig. 3.

For both the average stacking size and the average

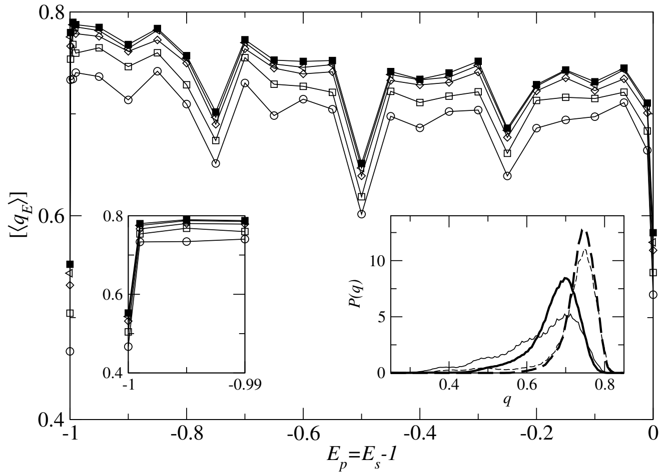

FIG. 3: Average overlap $q_{\mathcal{E}}$ as a function of the energy parameter E_p for different system sizes. The local minima at $E_p = -0.75, -0.5, -0.25$ are due to the commensurability of E_p and E_s and indicate a broad distribution of the ground states in configuration space. For explanation of symbols see caption of Fig. 7. The left inset is an enlargement to show the discontinuity. The right inset is the q-distribution for $E_s = E_p = -0.5$ (solid lines) and $E_s = -0.49, E_p = -0.51$ (broken lines) for sequence lengths $L = 512$ (thin lines) and $L = 1024$ (thick lines).

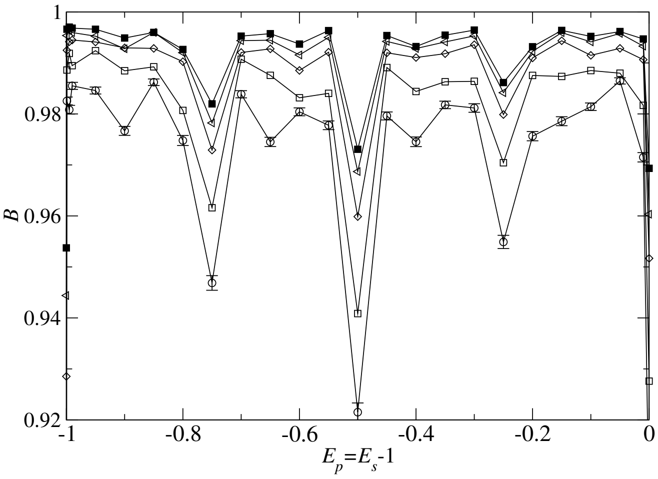

FIG. 4: The Binder Cumulant B of eq. (9) with $q = q_{\mathcal{E}}$ (see eq. (10)) at temperature $T = 0$. The curves for different system sizes do not cross, and therefore B gives no hint for a phase transition. For explanation of symbols see caption of Fig. 7.

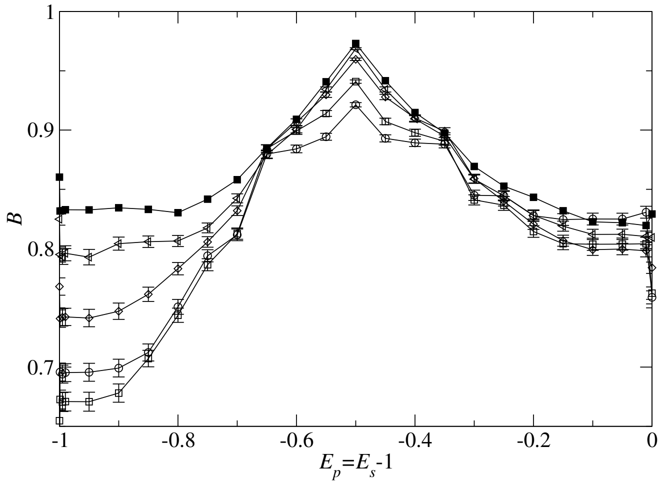

FIG. 5: The Binder Cumulant B of eq. (9) with $q = q_{\mathcal{E},\mathcal{E}'}$ (see eq. (10)) at temperature $T = 0$. For explanation of symbols see caption of Fig. 7.

overlap, the behavior changes smoothly when moving from one model to the other, with the exception of the highly degenerate points, where we can observe the jumps in the overlap q. Hence, there is no sign of a transition in between the two extremal models. To confirm this, we next study the Binder cumulant.

B. Binder cumulant

Since we introduced in eq. (3) a whole class of energy models depending on the pair energy E_p and the stacking energy E_s, we examined the behaviour of the Binder cumulant depending on this two energy parameters and the sequence length L. Second order phase transitions are characterized by crossing of the Binder cumulant for different system sizes at the transition point.

In Fig. 4 the Binder cumulant eq. (9) is shown at $T = 0$ using the "self-overlap" eq. (10). Again, the energy parameter E_p and E_s are varied such that always $E_p + E_s = -1$ holds. The value of B increases with increasing system size for E_p, i.e., the curves do not cross each other and therefore give no evidence of a phase transition. The local minima are – as the minima in Fig. 3 – due to the commensurability of the energy parameter E_p and E_s.

Further we used a reference ensemble generated for energy parameters $E_p = E_s = -0.5$ and used the "cross overlap" of eq. (11). First, the values of B at $E_p = E_s = -0.5$ coincide with the values shown in Fig. 4. Similar to the observation above, B roughly increases with the system size, although the separation of the curves is not as clear as in Fig. 4, especially in the range from $E_p = -0.4$ to $E_p = 0.0$, where the curves coincide within the error-bars. To summarize, the dependence of the Binder cumulant on the energy parameters does not indicate a phase transition.

C. Temperature dependence of the energy models

Another possible method to discriminate between the different energy parameters is to examine the temperature dependence of some quantities, especially the behaviour at a critical temperature. In Ref. 9 it was shown for a similar model that below a critical temperature, almost all sequences fold to a compact structure, but for most sequences not into a single structure. They point out that this kind of low temperature behaviour is well known from spin glass and other disordered systems. In

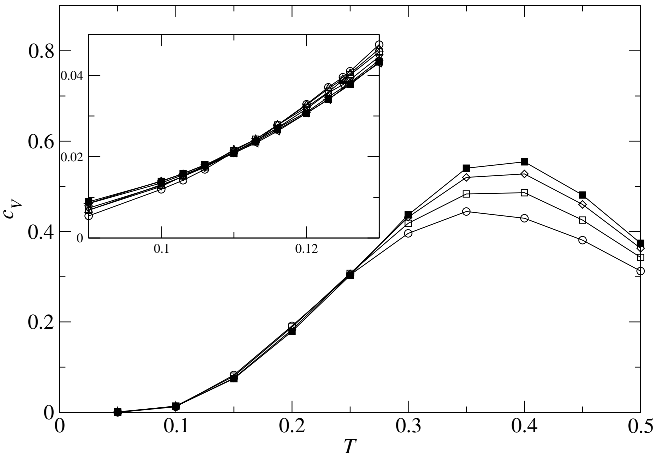

FIG. 6: Specific heat c_V as a function of temperature for parameters $E_s = -1.0$ and $E_p = 0.0$. The curves for different system sizes crosses at $T \approx 0.25$ and $T \approx 0.11$. The inset is an enlargement of the main plot. For explanation of symbols see caption of Fig. 7.

Fig. 6 we plotted the specific heat for different system sizes as a function of temperature for $E_s = -1.0$ and $E_p = 0.0$. All curves cross each other close to $T = 0.11$ and $T = 0.25$, which is an evidence for a phase transition at this temperature region. The data for other energy parameters (E_s, E_p) look similar, but the curves cross at different temperatures.

To determine the critical behaviour quantitatively we investigated the width $\chi_{\mathcal{R}}$ of the overlap distribution. As can be seen from Fig. 7 all curves have a maximum and the position of this maximum is decreasing with increasing sequence length. We assume that the maximum position has the form $T_c(L) = T_c + a\, L^{-1/\nu}$ and fit the data to this form to get the critical parameters (see Fig. 8). The results for three different pairs of energy parameters are shown in Tab. I. The critical exponent ν for the parameter pair $(E_s = 0, E_p = -1)$ is clearly different from that of the parameter pair $(E_s = -1, E_p = 0)$, showing that the quantitative behaviour changes when varying from one limiting case to the other. On the other hand for the intermediate parameter set the critical exponent is consistent with both within the error range.

Energy Model		$1/\nu$	T_c	θ
$E_s = 0$	$E_p = -1$	0.93(15)	0.086(3)	0.229(38)
$E_s = -0.5$	$E_p = -0.5$	0.70(36)	0.109(7)	0.237(50)
$E_s = -1$	$E_p = 0$	0.36(17)	0.125(21)	0.194(67)

TABLE I: Critical parameter of a $\chi_{\mathcal{R}}$-maximum fit. Comparing the two limiting cases $(E_s, E_p) = (-1, 0)$ and $(E_s, E_p) = (0, -1)$ the parameters ν and T_c are different and indicate a quantitative change. The last column belongs to the ε-coupling method in section IV D.

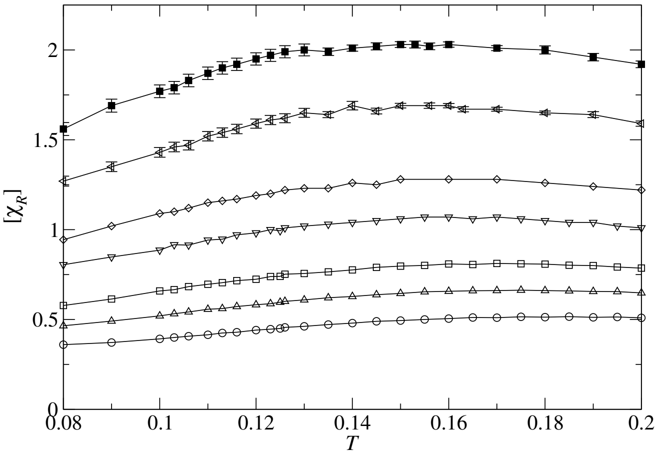

FIG. 7: $\chi_{\mathcal{R}}$ as a function of temperature for energy parameter $E_s = -1.0$, $E_p = 0.0$. For system size L the following symbols are used: ○ 128, △ 192, □ 256, ▽ 384, ◇ 512, ◁ 768, ■ 1024. Calculated data points are indicated by symbols. Lines are drawn to guide the eye. Missing error bars are of the size of the symbols or smaller, and omitted for legibility.

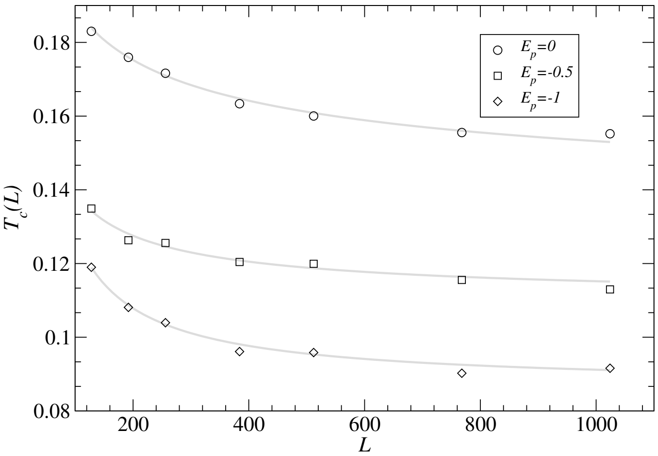

FIG. 8: The position of the maxima of $\chi_{\mathcal{R}}$ for different energy parameter. The curves are fitted to the form $T_c(L) = T_c + a\, L^{-1/\nu}$, and gives the critical parameters of Tab. I. Missing error bars are of the size of the symbols or smaller, and omitted for legibility.

Finally, we also found (not shown) that the behaviour of A of eq. (12) is in agreement with this observation and with the results [9] for a two-letter RNA model. For all three cases studied here, the width of the q–distribution varies only slightly from realization to realization at high temperatures, while for low temperatures the self averaging behaviour disappears.

D. ε-coupling

Previously the *ε-coupling method* has been used[12, 23] to investigate low-energy excitations of RNA secondary structures. The basic idea is to add another term to the energy model in eq. (3), which depends on the ground state of the original problem: Assume $\mathcal{S}_0$ is the unique ground state of $E(\mathcal{S})$, then a new energy function is defined as following

$$E_\varepsilon(\mathcal{S}) = E(\mathcal{S}) + \varepsilon\, q(\mathcal{S}, \mathcal{S}_0) \tag{14}$$

with $\varepsilon > 0$. The additional term penalizes structures similar to $\mathcal{S}_0$, where $\varepsilon\, q(\mathcal{S}, \mathcal{S}_0)$ is largest for $\mathcal{S} = \mathcal{S}_0$. In general the ground state $\mathcal{S}_\varepsilon$ of the new energy model E_ε will be different from $\mathcal{S}_0$. The difference $\Delta E(\varepsilon) := E(\mathcal{S}_\varepsilon) - E(\mathcal{S}_0)$ is an increasing function of ε and $\Delta E(\varepsilon) \leq \varepsilon$ holds. The later implies that $\mathcal{S}_\varepsilon$ is for small enough ε a low lying excitation of the original energy model, and has the smallest overlap with $\mathcal{S}_0$.

The average distance $d(\varepsilon, L) := 1 - q(\mathcal{S}_\varepsilon, \mathcal{S}_0)$ between the new and the old ground state scales as $d(\varepsilon, L) \propto \varepsilon\, L^{-\theta}$, ε constant, while the average energy difference scales as $\Delta E(d, L) \propto L^{\theta}$, d constant, with the critical exponents θ. For details see Refs. 12 or 23.

The assumption of a unique ground state used above does not hold for our model used so far: in general the ground state is highly degenerated because only two energy parameters (E_s and E_p) are used, many structures will have the same energy. The degeneracy of the ground state renders the ε-coupling method as described above almost useless. Since in natural RNA the contributions to the energy are more complex different structure will never be degenerated. This justifies to change the energy model slightly by adding a random energy $\eta_{i,j}$ to the pair energies introduced in eq. (1): $e(r_i, r_j) \to e(r_i, r_j) + \eta_{i,j}$. There are several possibilities to choose the distributions of the ηs (see [12]), here we use identical independently distributed Gaussian random number with zero mean $\langle \eta \rangle = 0$ and variance $\langle \eta^2 \rangle = \eta_0^2/L$ with $\eta_0 = 0.1$ (the QD model in Ref. 12).

The raw result for different values of $\varepsilon \in [0.01, 100]$ is shown in Fig. 9. A scaling plot of the data for $\varepsilon < 1$ according the scaling form $\Delta E L^{-\theta} = f(d)$ is shown in the inset of Fig. 9. The scaling parameters θ leading to the best data collapse for different energy parameters are shown in the right most column of Tab. I. They are equal within the error margins and thus does not give us a further hint of a quantitative different behaviour for different energy parameters. This difficulties in doing a good scaling of the data in the QD model were already pointed out [12]. However, for a different model using a scaling function with finite-size corrections a critical exponent $\theta = 0.23 \pm 0.05$ was obtained [23], which is close to our findings.

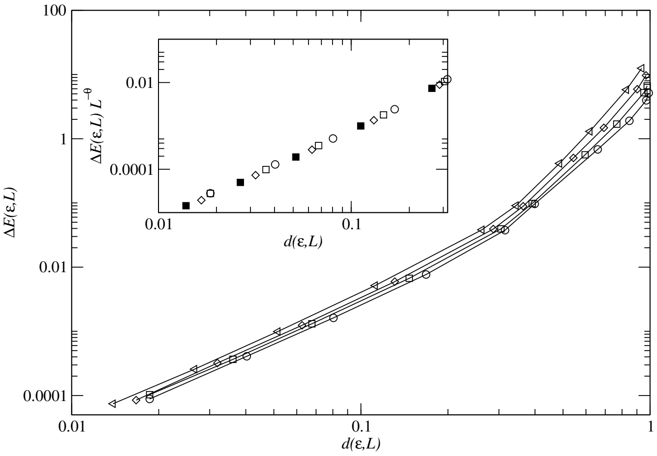

FIG. 9: ε-coupling results for $E_s = -0.5$, $E_p = -0.5$. The energy difference $\Delta E(\mathcal{S})$ between the original ground state structure $\mathcal{S}_0$ and the ground state structure $\mathcal{S}_\varepsilon$ of the disturbed model is plotted as a function of the distance between this structures. The inset is a scaled plot of the data with $\varepsilon < 1$ of the main plot. ($\theta = 0.24$, see Tab. I) For explanation of symbols see caption of Fig. 7.

V. SUMMARY

We have introduced a RNA model which continuously interpolates between two well known models. We sought for the answer to the question, whether there is any phase transition of the thermodynamic behaviour. We used both zero as well as finite temperature data.

Zero temperature results give no evidence for a phase transition apart from trivial transitions, e.g., discontinuity of t at points with $E_s = 0$ or $E_p = 0$. The curves of the Binder cumulant do not cross at a certain point, which would be an indication of a phase transition. Similar, the critical exponent θ derived from the ε-coupling method seems to be independent of the energy parameters E_s and E_p, and therefore gives no evidence for a quantitative difference in the thermodynamic properties. But as stated in [12], the determination of the critical parameter is rather difficult in the quasi-degenerated case studied here.

The finite temperature results show – in contrast to the zero temperature data – a quantitative dependence on the energy parameters. The critical exponent ν for the correlation length seems to depend on the energy model and may vary continuously while going from one limiting model to the other.

Acknowledgments

The authors have obtained financial support from the *Volkswagenstiftung* (Germany) within the program “Nachwuchsgruppen an Universitäten”. The simulations were performed at the Paderborn Center for Parallel Computing in Germany and on a workstation cluster at the Institut für Theoretische Physik, Universität Göttingen, Germany. We thank E. Yewande for helpful remarks.

[1] R. F. Gesteland, T. R. Cech, and J. F. Atkins, eds., *The RNA World* (Cold Spring Harbor Laboratory Press, New York, 1999), 2nd ed.

[2] P. G. Higgs, Quarterly Review of Biophysics **33**, 199 (2000).

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

[4] M. Zuker, Science **244**, 48 (1989).

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

[6] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, Monatsh. Chemie **125**, 167 (1994).

[7] R. Lyngsø, M. Zuker, and C. N. S. Pedersen, Bioinformatics **15**, 440 (1999).

[8] T. Liu and R. Bundschuh, Phys. Rev. E **69**, 061912 (pages 10) (2004), URL `http://link.aps.org/abstract/PRE/v69/e061912`.

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

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

[11] R. Bundschuh and T. Hwa, Phys. Rev. E **65**, 031903 (2002).

[12] E. Marinari, A. Pagnani, and F. Ricci-Tersenghi, Phys. Rev. E **65**, 041919 (pages 7) (2002), URL `http://link.aps.org/abstract/PRE/v65/e041919`.

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

[14] T. Liu and R. Bundschuh, *Large finite size effects in RNA secondary structures*, physics/0304108 (2003).

[15] K. Han and Y. Byun, Nucleic Acids Research **31**, 3432 (2003), http://nar.oupjournals.org/cgi/reprint/31/13/3432.pdf, URL `http://nar.oupjournals.org/cgi/content/abstract/31/13/3432`.

[16] R. Mukhopadhyay, E. Emberly, C. Tang, and N. S. Wingreen, Phys. Rev. E **68**, 041904 (2003).

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

[18] R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman, SIAM Journal of Applied Mathematics **35**, 68 (1978).

[19] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis* (Cambridge University Press, 1998).

[20] K. Binder, Z. Phys. B - Condensed Matter **43**, 119 (1981).

[21] R. N. Bhatt and A. P. Young, Phys. Rev. Lett. **54**, 924 (1985).

[22] R. N. Bhatt and A. P. Young, Phys. Rev. B **37**, 5606 (1988).

[23] F. Krzakala, M. Mézard, and M. Müller, Europhys. Lett. **57**, 752 (2002).