Zero-temperature properties of RNA secondary structures

Enzo Marinari,* Andrea Pagnani,† and Federico Ricci-Tersenghi‡

Dipartimento di Fisica, SMC and UdR1 of INFM, INFN, Università di Roma "La Sapienza," Piazzale Aldo Moro 2, I-00185 Rome, Italy
(Received 9 November 2001; published 8 April 2002)

We analyze different microscopic RNA models at zero temperature. We discuss both the most simple model, which suffers a large degeneracy of the ground state, and models in which the degeneracy has been removed in a more or less severe manner. We calculate low-energy density of states using a coupling perturbing method, where the ground state of a modified Hamiltonian, that repels the original ground state, is determined. We evaluate scaling exponents starting from measurements of overlaps and energy differences. In the case of models without accidental degeneracy of the ground state we are able to clearly establish the existence of a glassy phase with $\theta \simeq 1/3$. 87.15.Aa, 64.60.Fr

I. INTRODUCTION

RNA plays a fundamental role in the biochemistry of all living systems [1], and it is commonly believed to be at the origin of the pre-Darwinian epoch of life [2]. Much like for DNA, the primary structure of RNA can be described in terms of strings of the four letter alphabet composed by Adenosine, Citosine, Guanine, Uracile (ACGU). Since RNA is usually found in the single stranded pattern, formation of double-helix regions is accomplished by the molecule folding back onto itself to form Watson-Crick base pairs $G\equiv C$ and $A=U$, or the slightly less stable G-U pair. One of the most intriguing features of RNA folded secondary structures is that in most cases the connectivity graph is planar: This property greatly reduces the computational efforts needed for calculating the ground state structure.

It might be asked, whether secondary structures provide an adequate level of description for RNA real molecules [3]. It is believed that secondary structure description is biologically relevant for a number of reasons: Base pairing and base pair stacking provide the major part of the free energy of folding [4]; secondary structures have been used successfully by biologists in the interpretation of RNA function and activity [2]; and structures are conserved in evolutionary phylogeny. At the same time computer scientists find this level of description rather appealing since secondary structures are discrete and, therefore, easy to compare. Moreover, thanks to the planarity condition, efficient recursive algorithms for the computation of the native (ground state) structure are easily implemented [5,6].

Besides the genuine biological interest in RNA models, recently this subject has raised considerable attention as an intriguing problem in statistical mechanics of disordered systems. The focus is now set on the presence and the nature of a low-temperature phase in ensembles of random sequences. In a series of recent papers [7–13] different authors have presented evidence (mostly numerical) supporting the existence of a transition to a glassy phase (for a review see Ref.

[14]). While a careful study of the equilibrium thermodynamics of the model suggests a smoother than second-order phase transition [8], there is still much debate about the nature of the low-temperature phase, since finite size scaling corrections are very hard to keep under control. It has been shown in [8,10] that, at least for systems having up to 1000 bases, a broad overlap distribution characterizes the low-temperature phase, but a safe extrapolation to the thermodynamic limit from the available data is still out of control [9,10].

Bundschuh and Hwa [11,12] have presented an extensive study on a variety of similar RNA models supporting the existence of a low-temperature glassy phase. They were able to show analytically, via a two replica calculation, that weak quenched sequence disorder is equivalent to a high-temperature phase in which all replicas are independent (molten phase), and that there must be a finite temperature below which replicas start feeling themselves as in a strong coupling regime. Numerically the authors established this glassy transition measuring the free energy cost of imposing a pinch between two bases. They observed that the energy of the pinching excitation (with respect to the ground state) increases with the sequence length following a logarithmic law (even if a power law with small exponent was not excluded).

In this paper we study the scaling regime of the lowest energy excitations in different models of random RNA secondary structures. Following an idea put forward in [10], we use a perturbing method [15,16] that has been very valuable in the study of low-temperature properties of disordered systems [17]. In the following, we will call it the ε-coupling method. Very recently the same procedure has been followed by Krzakala, Mézard, and Müller [13]. We will comment on their results in the concluding section after having presented our data.

The goal of the ε-coupling method is to calculate the energy cost of typical excitations above the ground state involving a finite fraction of the system. As in the droplet model [18] these energy excitations are assumed to scale as $\Delta E(L) \propto L^{\theta}$, L being the length of the molecule, and θ being a relevant exponent we would like to determine. It is well possible that also the local pinching of coupled replicas imposed in [11,12] would generate "typical" configurations,

*Electronic address: Enzo.Marinari@roma1.infn.it
†Electronic address: Andrea.Pagnani@roma1.infn.it
‡Electronic address: Federico.Ricci@roma1.infn.it

but we believe that our method based on a bulk perturbation will surely do so (in this sense we find its application very illuminating). We apply a perturbation that is simply a repulsion term from the ground state structure, which forces the system to find new low-energy structures far away from the one of the original ground state, without any other constraint.

The paper is organized as follows: in Sec. II we introduce the formalism for describing the RNA secondary structure and the different ways we have used to remove the ground state degeneracy intrinsic to the original model [8]. In Sec. III we sketch the method we have applied for calculating the low-energy spectrum of the model. We also discuss the measurable observables. In Sec. IV we present our results, focusing on the differences and similarities among the models we have introduced. Finally, in Sec. V, we summarize our findings, compare ours with previous work, and comment on further developments.

II. MODELS

The secondary structure of RNA is a set of base pairs that occur in its three-dimensional structure. Let us define a sequence of basis as $\mathcal{R} \equiv \{r_1, r_2, \ldots, r_n\}$, r_i being the ith base of the chain and $r_i \in \{A, C, G, U\}$. A secondary structure on $\mathcal{R}$ is now defined as a set $\mathcal{S}$ of (i, j) pairs (with the convention that $1 \leq i < j \leq L$) satisfying the following rules.

(1) $j - i \geq 4$. This restriction permits flexibility of the chain in its three-dimensional arrangement.

(2) Two different base pairs $(i, j), (i', j') \in \mathcal{S}$ if and only if (without loss of generality we can assume that $i < i'$) $i < j < i' < j'$, i.e., the pair (i, j) precedes (i', j'), or $i < i' < j' < j$, i.e., the pair (i, j) includes (i', j'). This rule, called *planarity condition*, excludes the occurrence of the so-called *pseudoknots*, which are very unlikely in real RNA.

We consider a simplified model for RNA folding, very similar to the one studied in [7,8,14,19]. The model is described in terms of the Hamiltonian

$$\mathcal{H} = \sum_{(i,j) \in \mathcal{S}} e_{ij} = \sum_{(i,j)} e_{ij} \ell_{ij}, \qquad (1)$$

where e_{ij} is the pairing energy between bases i and j and the variable ℓ_{ij} takes value 1 if $(i, j) \in \mathcal{S}$ and 0 otherwise. On a first approximation one can assume that the pairing energies depend only on the paired bases, $e_{ij} = e(r_i, r_j)$. Reasonable values for the energies $e(r_i, r_j)$ of the allowed base pairs (C-G, A-U and G-U) at room temperature are of $O(1)$ kcal/mole [23]. One could consider other phenomenological parameters in order to take into account the whole complexity of a realistic energy function [4,20].

We have assumed a drastic approximation in order to get a tractable model both from a numerical and analytical point of view. We consider sequences made up of four symbols (A, C, G, and U) and we assume that only Watson-Crick base pairs may occur: we use a strong C-G coupling of energy -2 (in arbitrary units) and a weak A-U coupling of energy -1. All the other possible couplings increase the energy, so that the system avoids these links. One of the advantages of this model is that the role of the disorder (encoded in the

random sequence $\mathcal{R}$) is clearly separated from that of the frustration (induced by the planarity condition on the structure $\mathcal{S}$).

This four-letter model has an exponentially large ground state degeneracy, which gives a finite $T = 0$ entropy (as already found in [8] for the two-letter model). For this reason we refer to it as the degenerate model (the D model).

The large ground state degeneracy that occurs in this D model is a pathology of a frustrated models with simple discrete interactions: Since the couplings can take only the two negative values -2 and -1, the same exact energetic situation can be realized in many ways. This *accidental degeneracy* probably will not play a relevant role in the physical RNA: Since real RNA energy function is far more complex than that, ground state degeneracy is unlikely to occur. Because of that we define two new models with modified pairing energies, in order to remove the degeneracy. In both models this aim is accomplished by adding a small random perturbing term η_{ij} to the pairing energies: $e_{ij} \rightarrow e_{ij} + \eta_{ij}$.

In the quasidegenerate model (the QD model) the η_{ij} are independent and identically distributed (iid) variables extracted from a Gaussian distribution of zero mean $\langle \eta \rangle = 0$ and variance $\langle \eta^2 \rangle = \eta_0^2 / L$, with η_0 a small and finite constant of the order of 0.1: When $L \rightarrow \infty$ the pairing energies are modified of an infinitesimal amount. The variance is chosen such that the energies of the ground states (which are degenerate for $\eta_0 = 0$) are split over an $O(\eta_0)$ range. In this way we preserve somehow the structure of the original energy spectrum and the sequence still plays a key role, but the unphysical degeneracy is lifted and the ground state is now unique.

In the nondegenerate model (the ND model) the variance of the η_{ij} variables is finite, $\langle \eta^2 \rangle = \eta_0^2$ (we use $\eta_0 = 0.1$). This variance induces an $O(\sqrt{L})$ splitting of the degenerated ground states, which has to be considered as a strong reshuffling of the original energy spectrum, since the energy gaps among levels in the original model were of $O(1)$. The resulting energy landscape now depends very little on the sequence. Because of that the ND model is very similar to the "Gaussian disorder model" (the GD model), already discussed in [12], where the e_{ij} are iid Gaussian variables of zero mean and unitary variance. In this model the sequence plays no role.

III. METHODS

The ε-coupling method that we use to calculate low-energy excitations is the one already used in Refs. [17] and [13]. It works as follows: First of all one calculates the ground state structure $\ell_0 = \{\ell_{ij}^{(0)}\}$, which minimizes $\mathcal{H}$. Then one adds a perturbation to the Hamiltonian, $\mathcal{H}' = \mathcal{H} - \varepsilon(1 - q)$, where $q \equiv 1/L \sum_{ij} \ell_{ij} \ell_{ij}^{(0)} = (1/L) \ell \ell_0$ is the overlap with the ground state structure (note that, with this definition, the overlap is always positive). The perturbation term penalizes the structures that are close to the ground state ℓ_0 and thus acts as a repulsive term in the space of structures. Finally one calculates the ground state structure of $\mathcal{H}'$ for many values of ε. Let us call these new structures ℓ_ε.

By definition, for any disorder realization $\mathcal{J} = \{\mathcal{R}, \boldsymbol{\eta}\}$, both

the distance $d_{\mathcal{J}}(\varepsilon,L)=1-(1/L)\ell_{\varepsilon}\ell_0$ as well as the energy difference $\Delta E_{\mathcal{J}}(\varepsilon,L)=\mathcal{H}(\ell_{\varepsilon})-\mathcal{H}(\ell_0)$ between ℓ_{ε} and ℓ_0 are nondecreasing functions of ε. Moreover $\Delta E_{\mathcal{J}}(\varepsilon,L)<\varepsilon$, since the Hamiltonian has been perturbed by a term whose absolute value is less than ε, and the structures ℓ_{ε} are then low-energy excited states of the original Hamiltonian. We will indicate without the $\mathcal{J}$ subscript the observables averaged over the quenched disorder $\mathcal{J}=\{\mathcal{R},\boldsymbol{\eta}\}$: $d(\varepsilon,L)$ $=\overline{d_{\mathcal{J}}(\varepsilon,L)}$ and $\Delta E(\varepsilon,L)=\overline{\Delta E_{\mathcal{J}}(\varepsilon,L)}$.

The algorithm for finding the new ground states of ε-coupled system is exactly the same one used for the original Hamiltonian: The repulsion from the first ground state is included by modifying the values of the original pairing energies e_{ij}.

In the thermodynamical limit, structures differing by a finite ΔE have the same intensive energy, and one could try to understand how they are organized in the configurational space. An interesting question is whether, in the large L limit, these structures are extremely close together or spread over finite distances. The answer to this question can be given in terms of the asymptotic quantity

$$d_{\infty}(\varepsilon)=\lim_{L\to\infty}d(\varepsilon,L), \qquad (2)$$

which is again a nondecreasing function of ε. If $d_{\infty}(\varepsilon)=0$ for any finite ε then structures with the same energy are close together, while if $d_{\infty}(\varepsilon)>0$ for $\varepsilon>\varepsilon^*\sim O(1)$ then structures with the same intensive energy may have a broad probability distribution function of their distances and overlaps.

In the case where $d_{\infty}(\varepsilon)=0$ we can derive a relation describing the way $d(\varepsilon,L)$ vanishes. We assume, as in the droplet model [18], that the energy cost of a typical excitation involving a finite fraction of the system (i.e., having finite d) scales with the system size as

$$\Delta E_{\mathrm{typ}}\propto L^{\theta}. \qquad (3)$$

We call $\Pi(\Delta E,d,L)$ the probability distribution (over the disorder) of excitations with energy ΔE and size dL in systems of size L. For any fixed and finite $d\in(0,1]$, we assume that $\Pi(\Delta E,d,L)$ has a finite weight in $\Delta E=0$, and so, for normalization reasons, we must have $\Pi(0,d,L)=c(d)L^{-\theta}$ for large L (unless there is a δ function in $\Delta E=0$ as in the D model), where $c(d)$ is a smooth function in the scaling region.

Once we add the perturbing term $-\varepsilon d$ to the Hamiltonian, an excitation of size d will be activated only if its energy satisfies $\Delta E<\varepsilon d$. Thus the average distance of the new ground state structure is given by

$$d(\varepsilon,L)=\int_0^1 y\,dy\int_0^{\varepsilon y}\Pi(x,y,L)dx=\varepsilon L^{-\theta}\int_0^1 y^2 c(y)dy, \qquad (4)$$

for small ε and large L [24].

Then we can evaluate the θ exponent by two independent ways: (1) from Eq. (4), $d(\varepsilon,L)\propto\varepsilon L^{-\theta}$, by measuring the average distance $d(\varepsilon,L)$ for a fixed small ε as a function of
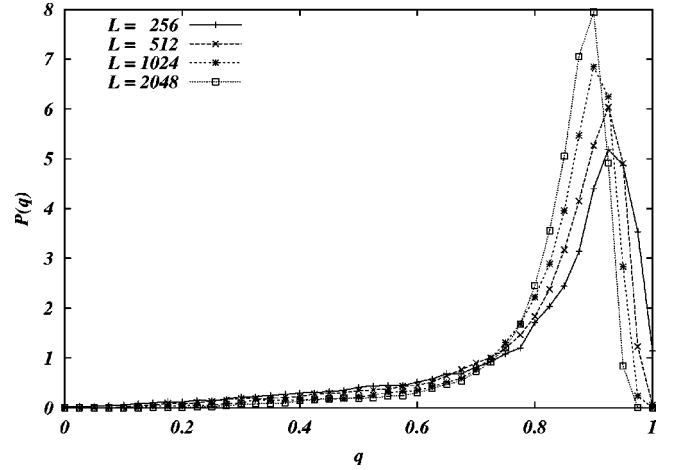

FIG. 1. Probability distribution of overlaps between all the pairs of ground state structures in the degenerate D model.

the system size L; (2) from Eq. (3), which can be equivalently rewritten for the average energy difference as

$$\Delta E(d,L)\propto L^{\theta}, \qquad (5)$$

by measuring the average energy difference for a fixed distance (not fixed ε) as a function of the system size L.

IV. RESULTS

We study zero-temperature properties of the models described above, i.e., we analyze ground states structures (GSS) of the original and of the perturbed Hamiltonians. We start showing the data for the $T=0$ overlap distribution in the D model (the only one with many different ground states). After that we present the results obtained with the ε-coupling method for all the models defined in Sec. II.

A. The D model

The D model possesses an exponentially large number of GSS, which form a set that we call $\mathcal{G}$. In order to understand how they are distributed in the space of structures, one can calculate the probability distribution function of the overlap, which is defined, for any pair of structures, as q $\equiv 1/L\Sigma_{ij}\ell_{ij}^{(1)}\ell_{ij}^{(2)}$.

Unfortunately the zero-temperature entropy of the D model is too large in order to list all the ground state structures for values of L large enough to be interesting. Because of that we have added to the D model a further constraint, suggested by observations on biological RNA, which strongly reduces the entropy: We avoid structures where a single base pairs is surrounded by nonpaired bases, that is, a structure with $\ell_{i-1,j+1}=0$, $\ell_{i,j}=1$, and $\ell_{i+1,j-1}=0$ is forbidden. The resulting $P(q)$ will be, by definition, narrower than the one for the unconstrained model, since the only effect of this new constraint is to select a subset of the original GSS.

The overlap distribution, averaged over 1000 samples, is shown in Fig. 1. It is worth noticing that the tail for small values of q is disappearing very slowly with increasing sys-
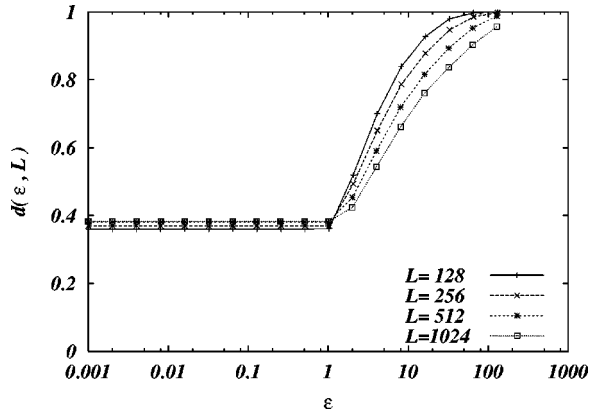
FIG. 2. Average distance $d(\varepsilon,L)$ versus ε for different chain lengths L in the degenerate D model.

FIG. 3. Probability distribution $P(q_{min})$ of the minimum overlap among any two GSS in the degenerate D model.

tem size. The peak location as well as the mean overlap converge somewhere around $q \simeq 0.87$, while the variance decreases with the system size approximately as $\sigma^2 \propto L^{-0.45}$. So the $P(q)$ seems to converge, in the thermodynamic limit, to a δ function centered on a value of q smaller than 1. Such a value is compatible with the observation that in a typical ground state structure the paired bases are a finite fraction (smaller than 1) of all bases. Nevertheless, as already explained in [10,21], the triviality of the $P(q)$ at zero temperature does not imply a trivial behavior of the whole low-temperature phase, and so we resort to the study of low-energy density of states.

We have calculated the ground state of the Hamiltonian $\mathcal{H}'$ for 18 values of $\varepsilon \in [0.001,131.072]$ (equally spaced on a logarithmic scale), and many L values. We have analyzed a minimum of 500 disorder realizations for the largest chain ($L=4096$), and a maximum of $\sim 5 \times 10^4$ samples for the smallest one ($L=128$).

The first GSS ℓ_0 is chosen with uniform probability in $\mathcal{G}$ (the set of all the degenerate ground states of the D model). When we switch on the perturbation the new GSS will be the one in $\mathcal{G}$ having the smallest overlap with ℓ_0. Nothing else will change as long as $\varepsilon \lesssim 1$, that is, as long as ε is not large enough to make an excited state with $\Delta E = 1$ to become the new GSS. This explains the plateau for $\varepsilon \lesssim 1$ in Fig. 2, where we show the average distance between ℓ_0 and ℓ_ε as a function of ε. The main information we get from Fig. 2 is the value of the plateau distance $d \simeq 0.38$, corresponding to an overlap $q = 1 - d \simeq 0.62$. This distance can be viewed as the radius of a sphere containing the set $\mathcal{G}$ of all the GSS. Note that the GSS are not uniformly distributed in this sphere [otherwise the $P(q)$ would be peaked on a much smaller overlap value], but they are very dense in the central region and very sparse on the boundaries. This means that if one chooses two GSS at random they will typically be very close in the dense region, giving a value of $q \simeq 0.87$, but if one forces the two GSS to be as far as possible the resulting minimum overlap q_{min} will be much smaller, and will depend strongly on the specific disorder realization [see Fig. 3, where we plot its probability distribution $P(q_{min})$].

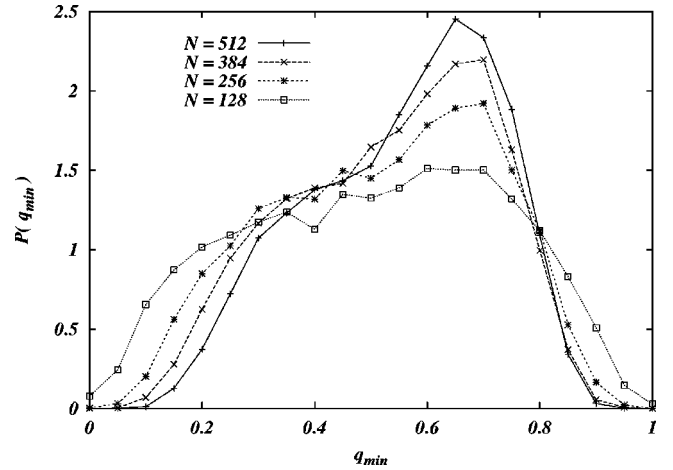As it is made clear by the results shown in Fig. 2, for models with high degeneracy the perturbing method does not

work properly in the interesting region of small ε, giving only information about the minimal overlap among GSS. Now we remove the degeneracy and analyze the other models.

B. The QD model

We have defined the QD model in a way as to keep as much as possible of the degenerate D model, even after removal of the accidental degeneracy. Here we can still distinguish two different regimes [see Fig. 4, where we plot the average distance $d(\varepsilon,L)$ as a function of ε for different chain lengths L]: For $\varepsilon > 1$ the data coincide with those for the (unconstrained) D model, while for $\varepsilon \lesssim 1$ they have now a nontrivial behavior.

In the interesting region of small ε the scaling of the data is very subtle and good results can be obtained either with $\theta = 0$ or with $\theta > 0$. Our finite size scaling analysis does not allow us to reach a quantitative estimate, and we cannot distinguish in a statistically significant way among a power law (dropletlike) scaling and a logarithmic scaling. Further and in depth studies are needed to understand better this model.

Despite the difficulties in the data analysis, we believe that the QD model has a large interest and relevance. Indeed
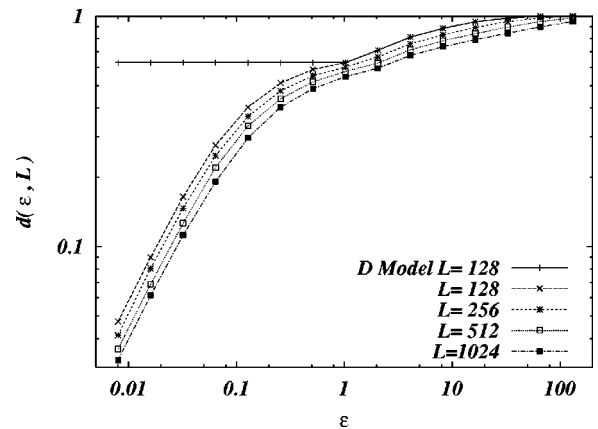

FIG. 4. As in Fig. 2 for the QD model. For comparison we also plot $L=128$ data for the unconstrained D model.
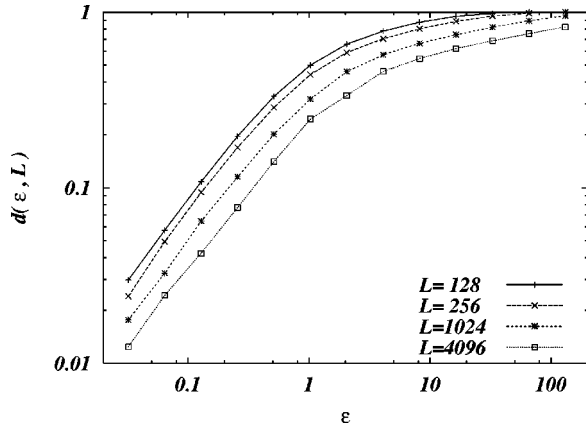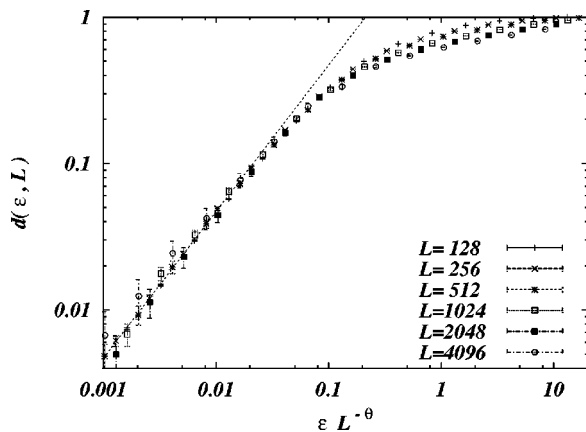
FIG. 5. As in Fig. 2 for the ND model.

it has the great advantage of a single nondegenerate ground state, but still the perturbation added in order to remove the original degeneracy modifies the energies of the structures by a quantity of order $\eta_0 \simeq 0.1$, thus keeping a large amount of information about the original energy landscape of the degenerate D model of RNA.

C. The ND model

In the ND model the ground state degeneracy has been removed by a random term that strongly reshuffles the energy levels.

In Fig. 5 we show the average distance $d(\varepsilon,L)$ for some values of ε and L (the error is of the order of the symbol size). Data are now smooth functions of ε, with no singular point at $\varepsilon \simeq 1$, and a finite size scaling analysis can be performed in an easier way.

Following Eq. (4) we have rescaled the data, plotting them versus $\varepsilon L^{-\theta}$. The results are shown in Fig. 6, where we have included all data points. The best collapse is achieved when using $\theta \simeq 0.33$. The dotted line has a unitary slope and clearly shows that $d(\varepsilon,L) \propto \varepsilon$ for small ε and any fixed L. We notice that we are looking for a finite size scaling that works well only up to a given value of $\varepsilon L^{-\theta}$: The method we are using is based on the idea of having a "small" per-

turbation that acts as a probe. Very large values of ε drastically change the Hamiltonian and the energy landscape. It is interesting to note that the scaling seems to work well up to $d \simeq 0.4$, that is the radius of the ball of degenerate ground states in the original D model.
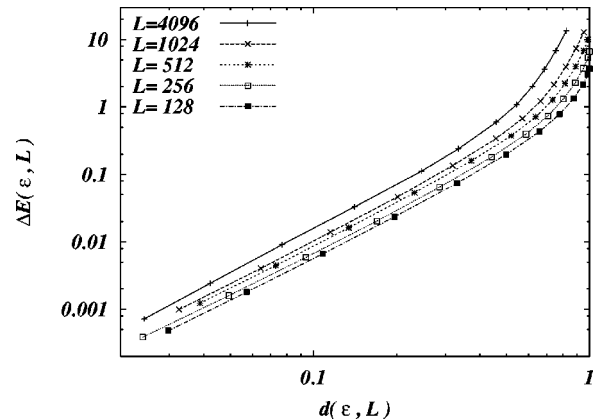
The physical interpretation of this result is the following. For any fixed ε the scaling variable $\varepsilon L^{-\theta}$ vanishes in the thermodynamical limit, which implies that the unperturbed ground state is stable against this kind of perturbation. In order to have a different and more complex behavior one should perturb the original Hamiltonian by a term whose amplitude increases with the size at least as fast as L^{θ}. The energy scale L^{θ} can thus be interpreted as the energy cost for reaching the first excited state.

A still clearer picture of this phenomenon is given in Fig. 7, where we plot the average energy difference $\Delta E(\varepsilon,L)$ as a function of the average distance $d(\varepsilon,L)$ between the unperturbed and the perturbed ground states. It is evident that, for any fixed distance, the energy difference is growing with the system size, according to the argument given above. We have rescaled the data following Eq. (5) and the results are shown in Fig. 8. Again the best collapse is achieved for $\theta \simeq 0.33$, in perfect agreement with the previous analysis. Also here the scaling region extends over distances up to $d \simeq 0.5$. For small distances the average energy increases as $\Delta E \propto d^2$ (see the dotted line in Fig. 8). The same behavior has been observed also in the QD model and we will present a simple explanation in the Sec. IV D, dedicated to the GD model.

Since the value of the θ exponent is small we also tried to fit the data under the assumption $\Delta E_{\text{typ}} \propto (\ln L)^a$, that is, $\theta = 0$. This behavior is suggested by mean-field solutions of disordered models [22] and more particularly from previous findings on similar models [8,11]. The conclusion is that a logarithmic fit with $a = 1.85 \pm 0.15$ still works rather well, but definitely the power law fit with $\theta = 0.33 \pm 0.01$ is more accurate and always has a smaller χ^2 value.

D. The GD model

Under the application of the ε-coupling perturbation, the GD model behaves very similarly to the ND model in that

FIG. 6. Average distance $d(\varepsilon,L)$ as a function of the rescaled variable $\varepsilon L^{-\theta}$ for different values of L in the ND model.

FIG. 7. The average energy difference ΔE between the unperturbed and the perturbed ground state structures as a function of their average distance d for different L values in the ND model.
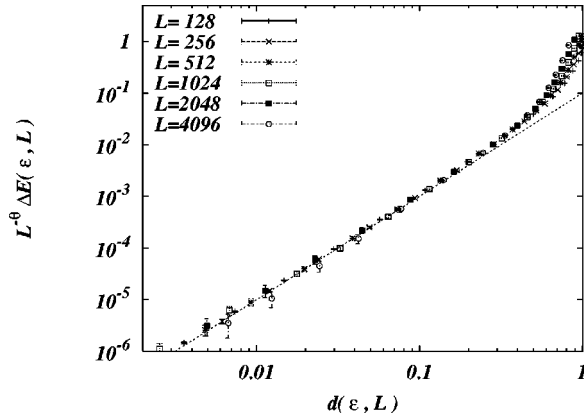
FIG. 8. Data of Fig. 7 rescaled according to Eq. (5).

(1) all the data perfectly collapse with $\theta \simeq 1/3$; (2) for small ε, $d \propto \varepsilon$; (3) for small d, $\Delta E \propto d^2$.

Moreover the results listed here do not depend on the presence of the constraint $\ell_{ij} = 0$ for $|i-j| < 4$ (see Sec. II). These findings imply that we have at hand a very simplified model that shares the same phenomenology with more realistic models and which is more amenable to an analytical treatment. For sake of clarity we recall its definition: We have a Hamiltonian of the form given in Eq. (1), where the pairing energies $e_{ij} = e_{ji}$ are $L(L-1)/2$ independent Gaussian variables with zero mean and unitary variance, and the ℓ_{ij} satisfy the planarity condition.

Within this model it is easier, for example, to understand the behavior $\Delta E \propto d^2$ for small d. First of all we observe from numerical simulations that in a typical GSS ℓ_0 the fraction of paired bases is $f < 1$ and the distribution of the pairing energies e_{ij} of the active links, the ones with $\ell_{ij}^{(0)} = 1$, can be very well approximated by a Gaussian of negative mean and finite width (the distribution is truncated since positive pairing energies are forbidden in the GSS). Let us call the distribution of the pairing energies absolute values $P_e(e)$. The only property we need for the proof is a *finite* weight in zero, $P_e(0) > 0$, and this is the case for the GD model (and also for the ND model).

Now we construct a sequence of structures ℓ_k such that $\Delta E \propto d^2$ for small d. ℓ_k is obtained from ℓ_0 removing the k weakest links, i.e., those with the smallest (in absolute value) pairing energies. So the distance between ℓ_0 and ℓ_k is $d = k/L$ and the energy difference ΔE is the sum of the smallest (in absolute value) k pairing energies. For large L we can write

$$\Delta E = \int_0^u P_e(e') e' \, de' = P_e(0) \frac{u^2}{2}, \qquad (6)$$

where the last equality only holds for small u. The upper integration limit u is chosen such that k pairing energies, or equivalently a fraction $2k/(fL)$ of pairing energies, are summed, that is

$$\frac{2k}{fL} = \frac{2d}{f} = \int_0^u P_e(e') \, de' = P_e(0) u, \qquad (7)$$

where the last equation is valid for small u. Combining the above equations we obtain $\Delta E \propto d^2$.

Since we have chosen a sequence of structures that are not guaranteed to have the lowest possible energies, we can only argue that $\Delta E \propto d^\alpha$ with $\alpha \geq 2$. Nevertheless from numerical simulations the exponent turns out to be exactly 2 (see Fig. 8).

In this very simplified GD model we can make one more analytical prediction, regarding the fraction of paired bases in the GSS. The number of planar structures with a fraction f of paired bases can be easily calculated with the help of generating functions and turns out to be given by $\exp[Ls(f)]$, with an intensive entropy

$$s(f) = -f \ln f - (1-f)\ln(1-f) + f \ln 2. \qquad (8)$$

$s(f)$ has a maximum for $f = \frac{2}{3}$, with $s(\frac{2}{3}) = \ln 3$.

Let us now fix f and see how the energies of the $\exp[Ls(f)]$ structures are distributed. They look random, but actually, since the independent Gaussian random variables are only $L(L-1)/2$, there must be many correlations among them. Since any of these energies is the sum of $fL/2$ random Gaussian pairing energies, we make the approximation that the distribution of structure energies is also Gaussian with a variance proportional to fL, i.e., $\mathcal{P}(E) \propto \exp[-E^2/(bfL)]$. The evaluation of the coefficient b is out of our present scope. Given, for any fixed f, the number of structures and the distribution of the energies, we can estimate the most probable lowest energy $E_{\min}(f)$ through

$$e^{-Ls(f)} = \int_{-\infty}^{E_{\min}} \mathcal{P}(E') dE' \simeq \exp[-E_{\min}^2/(bfL)], \qquad (9)$$

where the last equality holds because $E_{\min}$ is negative and large. The above equation implies $E_{\min}(f) = -L\sqrt{bfs(f)}$. In order to find the fraction of paired bases corresponding to the GSS one has to minimize $E_{\min}(f)$, or equivalently maximize $fs(f)$, over f. Such an extremum is achieved for $f = 0.86$ to be compared with the fraction of paired bases found numerically $f = 0.856$. The rather small discrepancy tells us that the approximation made on the form of the structure energies distribution $\mathcal{P}(E)$ is not so bad.

The apparent correctness of such a simple approximation could suggest that the GD model has a trivial energy landscape. We have checked for this possibility with the following method: In a trivial energy landscape any reasonably smart greedy algorithm should be able to reach, or at least to closely approach, the ground state energy. We have used a greedy algorithm that builds up the structure in the following way: It starts with a structure with no links, at each step it chooses the lowest negative pairing energy (largest in absolute value) among the set of those allowed by the planarity condition, and adds the corresponding link to the growing structure. Using this greedy algorithm we can reach the energies shown in Fig. 9, which are more than 10% higher than the corresponding ground state energies. So, it seems that finding a structure with low energy in linear time is not an easy task. This suggests a complex energy landscape. A
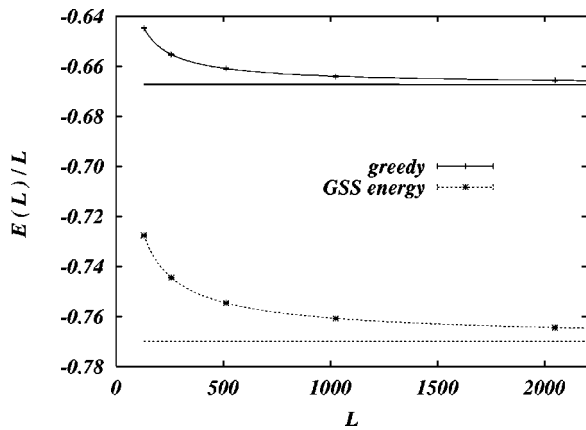
FIG. 9. As a function of the system size L we compare the ground state energy of the GD model (below) with the one reached by a simple greedy algorithm (above). The horizontal lines correspond to the infinite size extrapolations.

deeper analysis is obviously needed in order to say how much complex the energy landscape of the GD model is.

V. CONCLUSIONS

Our results allow one to describe a clear and simple physical picture for the RNA-inspired models studied here. All of them possess a glassy phase at low-enough temperatures (since we have analyzed very low energy density of states we cannot make predictions on the location of the critical temperature). We can claim that our study clearly

selects a positive θ exponent value for all cases but for the quasidegenerate QD model, where our numerical results are not precise enough to allow us quantitative statements.

At variance with the results of [11,12] we find that in the nondegenerate ND model the broken phase does not look marginal, but a standard droplet glassy phase with $\theta \simeq 1/3 > 0$. Our way of analyzing the data allows us to exclude (with good confidence) a simple logarithmic divergence of the energy difference between the ground state and the excited states. On this issue we agree with the results of [13]: The difference in the estimate we give for θ, as compared to the $\theta \simeq 0.23$ of [13], is probably due to the different fitting procedure and the number of free parameters used in the fit.

The θ exponent we find is perfectly compatible with that for directed polymers in random media in $1+1$ dimensions [16], $\theta_{DPRM} = \frac{1}{3}$. Since the two models have some similarities, this relation could indeed hide a deep connection.

The degenerate D model and the quasidegenerate QD model we have defined above are maybe the less trivial and the most intriguing from the theoretical point of view. Unfortunately we were not able to determine accurately the asymptotic scaling behavior in the latter.

It is probable, on the contrary, that the most part of the analytic developments will be obtained for the Gaussian disorder GD model, which is by far the simplest among all the models with a nontrivial behavior.

[1] C. Branden and J. Tooze, *Introduction to Protein Structure* (Garland Publishing, New York, 1991).

[2] R. F. Gesteland and J. F. Atkins, *The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA World* (Cold Spring Harbor Laboratory Press, New York, 1993).

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

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

[5] R. Nussinov and A. B. Jacobson, Proc. Natl. Acad. Sci. U.S.A. **77**, 6309 (1980).

[6] M. Zuker and P. Stiegler, Nucleic Acids Res. **125**, 133 (1981).

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

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

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

[10] A. Pagnani, G. Parisi, and F. Ricci-Tersenghi, Phys. Rev. Lett. **86**, 1383 (2001).

[11] R. Bundschuh and T. Hwa, e-print cond-mat/0106029.

[12] R. Bundschuh and T. Hwa, e-print cond-mat/0107210.

[13] F. Krzakala, M. Mézard, and M. Müller, e-print cond-mat/0108374.

[14] P. G. Higgs, Q. Rev. Biophys. **33**, 199 (2000).

[15] S. Caracciolo, G. Parisi, S. Patarnello, and N. Sourlas, Europhys. Lett. **11**, 783 (1990).

[16] M. Mézard, J. Phys. (Paris) **51**, 1831 (1990).

[17] F. Krzakala and O. C. Martin, Phys. Rev. Lett. **85**, 3013 (2000); M. Palassini and A. P. Young, *ibid.* **85**, 3017 (2000); E. Marinari and G. Parisi, Phys. Rev. B **62**, 11 677 (2000); Phys. Rev. Lett. **86**, 3887 (2001).

[18] W. L. McMillan, J. Phys. C **17**, 3179 (1984); A. J. Bray and M. A. Moore, Phys. Rev. Lett. **58**, 57 (1987); D. S. Fisher and D. A. Huse, Phys. Rev. B **38**, 373 (1988).

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

[20] D. H. Mathews, J. Sabina, M. Zucker, and H. Turner, J. Mol. Biol. **288**, 911 (1999).

[21] A. K. Hartmann and F. Ricci-Tersenghi, e-print cond-mat/0108307; F. Krzakala and O. C. Martin, Europhys. Lett. **53**, 749 (2001).

[22] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1986).

[23] In the models studied here we do not try a quantitative, detailed comparison to experimental results, and so we only select pairing energies of the correct experimental order of magnitude.

[24] The D model has $\theta = 0$, since, for a finite range of d values, its $\Pi(\Delta E, d, L)$ has a δ function in $\Delta E = 0$ that implies the presence of a gap (energy levels are discretized), so that the integral in Eq. (4) is finite for any L and ε small enough.