Topology of pseudoknotted homopolymers

Graziano Vernizzi,[1] Paolo Ribeca,[2] Henri Orland,[1] and A. Zee[3,4]
[1]*Service de Physique Théorique, CEA Saclay, 91191 Gif-sur-Yvette Cedex, France*
[2]*Humboldt Universität, Newtonstr. 15, 12489 Berlin, Germany*
[3]*Department of Physics, University of California, Santa Barbara, California 93106 USA*
[4]*Institute for Theoretical Physics, University of California, Santa Barbara, California 93106 USA*
(Received 11 September 2005; published 3 March 2006)

We consider the folding of a self-avoiding homopolymer on a lattice, with saturating hydrogen bond interactions. Our goal is to numerically evaluate the statistical distribution of the topological genus of pseudoknotted configurations. The genus has been recently proposed for classifying pseudoknots (and their topological complexity) in the context of RNA folding. We compare our results on the distribution of the genus of pseudoknots, with the theoretical predictions of an existing combinatorial model for an infinitely flexible and stretchable homopolymer. We thus obtain that steric and geometric constraints considerably limit the topological complexity of pseudoknotted configurations, as it occurs for instance in real RNA molecules. We also analyze the scaling properties at large homopolymer length, and the genus distributions above and below the critical temperature between the swollen phase and the compact-globule phase, both in two and three dimensions.

DOI: 10.1103/PhysRevE.73.031902 PACS number(s): 87.14.Gg, 87.15.Aa, 87.15.By

I. INTRODUCTION

One of the most exciting fields in modern computational molecular biology is the search for tools predicting the complex foldings of biopolymers such as RNA [1–3], when homologous sequences are not available. The prediction of the full tertiary structure of a RNA molecule is still an open issue [4], mainly because of its intrinsic high computational complexity [5]. It is known that the tertiary structure involves an important set of structural motifs, the so-called *pseudoknots* [6]. These are conformations such that the associated disk diagram (which represents all nucleotides along the RNA backbone as points on an oriented circle from the 5′ end to the 3′ end, and where each base-pair is represented by an arc joining the two interacting nucleotides, inside the circle; see Fig. 1) is not planar, i.e., it contains intersecting arcs. RNA pseudoknots have been identified in nearly every organism, and they proved to play important regulatory and functional roles [7,8]. Their ubiquity manifests in a large variety of possible shapes and structures [9], and their existence should not be neglected in structure prediction algorithms, as they account for 10%–30% on average of the total number of base pairs. Actually, several computer programs have been proposed for predicting RNA secondary structures including pseudoknots [10–16] (the list is not exhaustive), but the complexity of the problem and the approximations involved are usually such that the issue is far from being solved [17].

An analytical mathematical tool which can fully describe any RNA contact structure including all possible pseudoknots, appeared first in [18]. There, all RNA disk diagrams are considered as Feynman diagrams of a suitable field theory of $N\times N$ Hermitian matrices (a combinatorial tool borrowed from quantum field theory). The latter is known to organize all the diagrams according to an asymptotic $1/N^2$ topological expansion at large-N [19]. This provides in fact a rigorous way to classify nonplanar diagrams, and therefore it induces a natural topological classification of pseudoknots [12]. Namely, to any given pseudoknotted configuration (and more generally, to any contact structure of an heteropolymer with binary saturating interactions), one can associate an integer number g, the *genus*. It is defined as the topological genus of the associated disk diagram, i.e., by $\chi=1-2g$, where χ is the Euler characteristic number of the diagram. As reviewed in [20], the genus is the minimum number of handles the disk should have in order that all the cords are not intersecting (see Fig. 1). Other characterizations of pseudoknots have been proposed (e.g. [21–23]). The classification [18] is truly topological, meaning that it is independent from the way the diagram is drawn, and dependent only on the intrinsic complexity of the contact structure.

The large-N asymptotics of the analytical model in [18] is hard to obtain exactly. However, in [24] a special case of the general model [18] has been considered and solved. It was the simple case of an infinitely flexible and stretchable homopolymer, where there is no dependence on the primary sequence, and any saturating base pair between all the "nucleotides" is allowed. An analytical asymptotic expansion was evaluated and the distribution of the genus of pseudoknotted contact structures was obtained. One of the results is that an homopolymer with L nucleotides has an average genus close to the maximal one, that is $L/4$. Of course, real RNA molecules are not infinitely flexible and stretchable homopolymers. It is customary to assume that the bases i and j can interact only if they are sufficiently far apart along the chain (e.g., $|i-j|\geqslant 4$, [4]) because of bending rigidity. Moreover helices have a long persistence length ($\sim$200 base pairs) and this necessarily constrains the allowed pairings even more. We expect that including all steric and geometrical constraints should considerably decrease the genus of allowed pseudoknots, compared to the purely combinatorial case [24] where the actual three-dimensional conformation was neglected. The purpose of this work is to numerically analyze the effects of steric and geometric con-

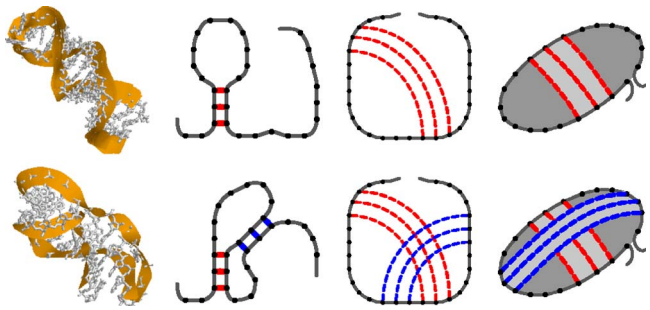

FIG. 1. (Color online) Top, from left to right: a hairpin loop (PDB number 1NA2), its squiggle-plot, and disk diagram representation which is of genus zero since it is planar. Bottom: *H*-type pseudoknot (PDB number 1RNK). In this case the disk diagram is not planar and has genus one.

straints on the genus distribution of pseudoknots topologies in homopolymers, in the same spirit of [24].

II. THE MODEL

We model the system by considering a polymer on a cubic lattice, i.e., a self-avoiding random walk with short-range attractive interaction [25]. A self-avoiding walk (SAW) is a sequence of neighboring lattice sites $i=0,1,\ldots,n$ with coordinates $\{\mathbf{r}_i\}$, such that the same lattice-site cannot be visited more than once. This is a standard approach in polymer physics (and RNA; see e.g. [21,23]). The attractive interaction is usually used to describe bad solvent quality, but in our case we insist more on the saturating nature of the hydrogen bond interactions. Such a requirement is crucial here, since the concept of the topological genus for a contact structure can be defined unambiguously only when the interactions are saturating. One of the most natural ways to model the interaction is by considering a "spin" model (see e.g. [26–28]). Strictly speaking, our model is a variation of the standard θ-polymer model, and similar interaction models for RNA on the lattice have been already proposed (e.g., [29]). To each vertex i we associate a unit spin $\mathbf{s}_i$ which represents the nucleotide direction with respect to the backbone. The only allowed directions for $\mathbf{s}_i$ are the lattice ones. Moreover, the spins cannot overlap with the backbone because of the excluded volume between the nucleotides and the backbone. The saturating nucleotide-nucleotide interaction occurs when two spins $\mathbf{s}_i,\mathbf{s}_j$ on neighboring sites, $|\mathbf{r}_i-\mathbf{r}_j|=1$, are pointing to each other. The energy of a configuration $\{\mathbf{r}_i,\mathbf{s}_i\}$ is thus defined by the Hamiltonian

$$\mathcal{H}=-\epsilon\sum_{i<j}\delta(\mathbf{r}_i+\mathbf{s}_i-\mathbf{r}_j)\,\delta(\mathbf{s}_i+\mathbf{s}_j)\,\delta(|\mathbf{r}_i-\mathbf{r}_j|-1), \qquad (1)$$

where $\epsilon\geqslant 0$ is an effective hydrogen-binding energy, the same for all monomers of the chain. Let us note that since we are not aiming to set up a realistic lattice model for RNA-folding, but rather to understand steric effects on the genus distributions of a homopolymer, we do not take into account stacking energies.

The basic features of our model are clear: At high temperatures, we expect the system to be in a swollen SAW state (entropy dominated coil state), whereas at lower temperatures we expect a kind of "compact globule"-like phase [25]. The transition temperature T_θ defines the so-called θ-point. However, details on the thermodynamics, kinetics, phase diagram, etc. can be rather complex [23,29]. We limit ourselves here only to the analysis of the genus distribution of pseudoknotted structures for comparing the effects of sterici-ty constraints versus the purely combinatorial model of [24]. All other considerations are postponed elsewhere.

III. THE METHOD

The numerical sampling of the statistical distribution $\mathcal{Z}=\Sigma_{\mathrm{SAW},\{\mathbf{s}_i\}}\exp(-\mathcal{H}/k_BT)$, where k_B is Boltzmann's constant, T is the absolute temperature, and the sum is restricted to SAWs and configurations of spins $\{\mathbf{s}_i\}$ satisfying the aforementioned constraints, is implemented by using the Monte Carlo Growth Method. It was originally proposed by Garel and Orland in [30] and has been applied to several statistical systems since then (see references in [31]). It consists in starting with an ensemble of chains at equilibrium and then growing each chain by adding one monomer at a time with a probability proportional to the Boltzmann factor for the energy of the chain. At each step the ensemble remains at equilibrium (a detailed description of the algorithm with applications can be found in [31]). It belongs to the family of so-called "population Monte Carlo algorithms" [32], where, contrary to the "dynamical" Markov Chain Monte Carlo methods, the population is fully grown and evolved, nondynamically. At high temperatures we considered populations with a variable number of chains in the range 10 000–40 000, and with a typical length of $L=500$ monomers (up to $L=1200$ in some cases). Accuracy and statistical averages were computed by taking several independent populations (of the order of 40). At low temperatures we considered populations of up to 100 000 chains. That is so because at low temperatures the chains are unavoidably trapped into local minima. This problem can be partly controlled and monitored by considering more populations of a larger size, and by analyzing the population-population correlations for *all* observables we estimate. That is also the reason why we do not explore the region very much below the critical temperature, i.e., at very low temperatures. In this work, we limit our analysis only in the region of temperatures where all our tests are statistically reliable and not biased by local minima trapping. Finally, all the simulations in three dimensions have also been performed on a square lattice in two dimensions.

IV. RESULTS AND DISCUSSION

We expect different genus distributions above and below T_θ. We therefore first determine T_θ, which can be done efficiently by computing the end-to-end distance $R_e^2=(\mathbf{r}_L-\mathbf{r}_0)^2$, and the radius of gyration $R_g^2=\Sigma_{i<j}(\mathbf{r}_i-\mathbf{r}_j)^2/L^2$. It is known that the ratio $\rho^2=\langle R_e^2\rangle/\langle R_g^2\rangle$ is universal in the limit $L\to\infty$ and converges to a step function as a function of T, with a universal critical value at T_θ [25,33,34]. In Fig. 2 we plot ρ which shows a transition temperature $T_\theta=0.39\pm0.01$

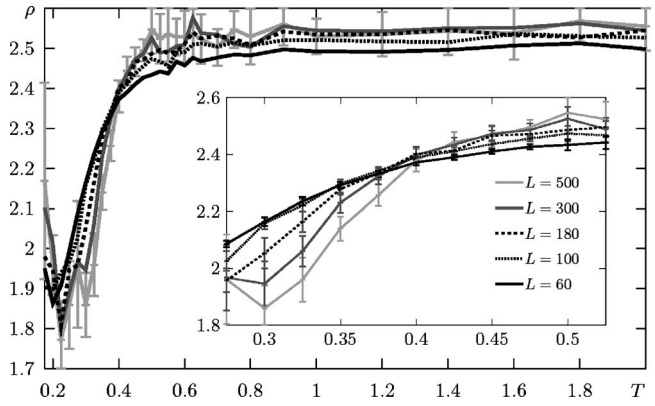

FIG. 2. The ratio ρ in 3D as a function of the temperature T, for several values of L (with error bars plotted only for $L=60$). At low temperatures the error bars are larger than the variations of the curves. The inset shows that as L increases, the curves approach a universal step function about $T_\theta \sim 0.39 \pm 0.01$.

($T_\theta^{2D}=0.48\pm0.02$ in two dimensions), in units where $k_B=1$ and $\epsilon=1$. We also verified that $\rho_\infty=2.5\pm0.05$ for $T \gg T_\theta$ asymptotically, and we find an intermediate value $\rho_\theta=2.35\pm0.05$ at T_θ (and $\rho_\infty=2.67\pm0.01$ and $\rho_\theta=2.39\pm0.01$ in 2D, respectively). At large-L we find the following scalings: for $T>T_\theta$, $\langle R_g \rangle \sim L^\nu$ with $\nu=0.59\pm0.01$ ($\nu=0.75\pm0.02$ in two dimensions), which is consistent with the critical exponent of a swollen SAW; for $T<T_\theta$, $\nu=0.32\pm0.02$ ($\nu=0.50\pm0.01$ in two dimensions) which is consistent with a compact phase. All these results are in agreement with high-accuracy simulations of similar models [35,36]. We then proceed with extracting the genus distributions in the two phases. The results are in Figs. 3 and 4. When comparing them with the combinatorial results of [24], we see that the genus at a fixed L is on the average much smaller. More precisely, below the θ-point the average genus scales like by $\langle g/L \rangle \sim 0.141\pm0.003$ and $\langle g/L \rangle \sim 0.1318\pm0.0025$, in 3D (at $T=0.2$) and 2D (at $T=0.225$), respectively. In both cases the scaling is at a lower rate (about 50% less) than the value $L/4$ computed in [24]. In the swollen-phase (e.g., $T=10>T_\theta$), the average genus is given by $\langle g/L \rangle \sim (585\pm8)10^{-6}$ in 3D, and $\langle g/L \rangle \sim (410\pm1)10^{-5}$ in 2D. Such a low rate comes from the tendency of a homopolymer to develop long rectilinear subchains in the swollen phase. In two dimensions the entropic factor is smaller than in three dimensions and the genus growth rate is therefore larger (see Fig. 4). Moreover, the genus distributions for $T>T_\theta$ are numerically consistent with Poissonian distributions (see Fig. 4), whereas at smaller temperatures they are closer to Gaussian ones.

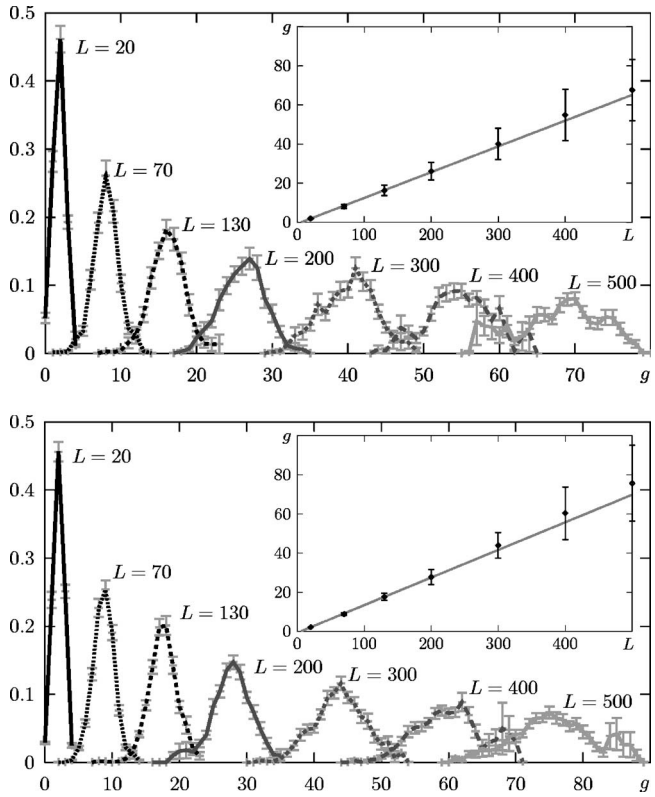

FIG. 3. The genus distributions of pseudoknots at fixed L, in 2D (top) and 3D (bottom), in the compact phase at $T=0.225$ and $T=0.2$, respectively. The insets represent the behavior of $\langle g \rangle$ at large L.

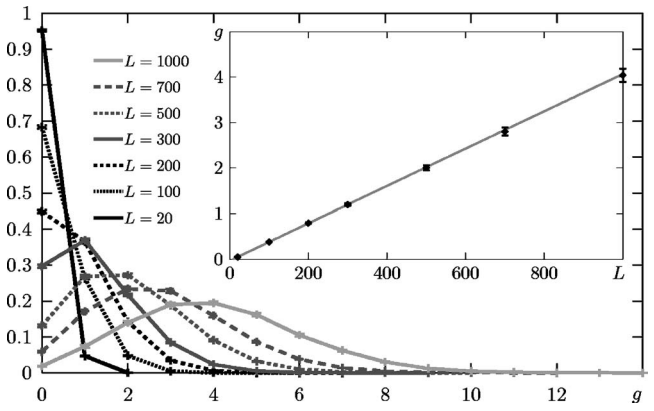

FIG. 4. The genus distribution of a two-dimensional homopolymer, in the swollen phase $T=10>T_\theta^{2D}$, at various values of L.

It turns out that the average genus of homopolymers described by the Hamiltonian Eq. (1) is an extensive quantity, like the energy, and their ratio is shown in Fig. 5. All these results confirm that the genus distribution behaves differently in the two phases, as expected. They also quantify how much the restrictions induced by the actual three-dimensional arrangement of the chain can limit the number and complexity of pseudoknots (compared to [24]). We find values closer to

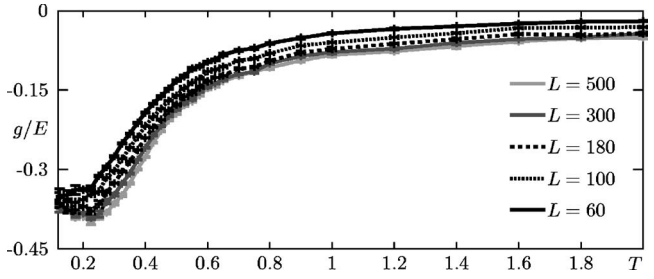

FIG. 5. The ratio (genus)/(energy) of an homopolymer on a cubic lattice, as a function of T, at different values of L.

what seems to happen in pseudoknots of real RNA molecules. In fact, real RNA molecules typically have small genus. For instance, a simple *H*-type pseudoknot (~20 bases or more) or the classical kissing-hairpins pseudoknot (~30 bases or more) both have genus 1, much less than the toy-model prediction in [24]. Even tRNAs (~80 bases) mostly contain 4 helices, two of them linked together by a kissing-hairpin pseudoknot, has still genus 1. Typical tmRNAs (~350 bases long) contain four *H*-type pseudoknots, and its total genus is 4, far below the theoretical upper bound $L/4$. Our numerical results would instead indicate, for instance for a 80 bases long homopolymer, a genus of about 11.2 in three dimensions (~10.5 in two dimensions). Even if it is smaller than the value suggested in [24] (because of the steric constraints), it is still too high when compared to real RNA molecules. The obvious reasons are that we neither included the primary sequence nor realistic stacking energies. We have nevertheless been able to quantify the general effect of steric constraints on the genus distribution of a pseudoknotted homopolymer on a lattice, as a first step towards a model which includes a more realistic energy function.

ACKNOWLEDGMENTS

We wish to thank T. Garel and R. Guida for discussions. This work was supported in part by the National Science Foundation under Grant No. PHY 99-07949, and by Sonderforschungsbereich-Transregio "Computational Particle Physics" (SFB-TR9). G.V. acknowledges the support within the European program, MEIF-CT-2003-501547.

[1] R. Schroeder, A. Barta, and K. Semrad, Nat. Rev. Mol. Cell Biol. **5**, 908 (2004).
[2] I. Tinoco Jr. and C. Bustamante, J. Mol. Biol. **293**, 271 (1999).
[3] P. G. Higgs, Q. Rev. Biophys. **33**, 199 (2000).
[4] M. Zuker, Nucleic Acids Res. **31**, 3406 (2003); see also I. L. Hofacker, *ibid.* **31**, 3429 (2003).
[5] R. B. Lyngsø and C. N. Pedersen, J. Comput. Biol. **7**, 409 (2000).
[6] C. W. Pleij, K. Rietveld, and L. Bosch, Nucleic Acids Res. **11**, 1717 (1985).
[7] L. X. Shen and I. Tinoco Jr., J. Mol. Biol. **247**, 963 (1995).
[8] P. L. Adams, M. R. Stahley, A. B. Kosek, J. Wang, and S. A. Strobel, Nature (London) **430**, 45 (2004).
[9] D. W. Staple and S. E. Butcher, PLoS Biol. **3**, 213 (2005).
[10] E. Rivas and S. R. Eddy, J. Mol. Biol. **285**, 2053 (1999).
[11] M. Pillsbury, H. Orland, and A. Zee, Phys. Rev. E **72**, 011911 (2005).
[12] M. Pillsbury, J. A. Taylor, H. Orland, and A. Zee, e-print cond-mat/0310505.
[13] A. Xayaphoummine, T. Bucher, and H. Isambert, Nucleic Acids Res. **33**, 605 (2005).
[14] A. P. Gultyaev, Nucleic Acids Res. **19**, 2489 (1991).
[15] J. P. Abrahams, M. van den Berg, E. van Batenburg, and C. W. A. Pleij, Nucleic Acids Res. **18**, 3035 (1990).
[16] J. E. Tabaska, R. B. Cary, H. N. Gabow, and G. D. Stormo, Bioinformatics **14**, 691 (1998).
[17] S. R. Eddy, Nat. Biotechnol. **22**, 1457 (2004).
[18] H. Orland and A. Zee, Nucl. Phys. B **620**, 456 (2002).
[19] G. 't Hooft, Nucl. Phys. B **72**, 461 (1974).
[20] G. Vernizzi, H. Orland, and A. Zee, http://arxiv.org/q-bio.BM/0405014; see also Acta Phys. Pol. B **36**, 2821 (2005); *ibid.* **36**, 2829 (2005).
[21] A. Kabakçioglu and A. L. Stella, Phys. Rev. E **70**, 011802 (2004).
[22] E. Rivas and S. R. Eddy, Bioinformatics **16**, 334 (2000).
[23] A. Lucas and K. A. Dill, J. Chem. Phys. **119**, 2414 (2003).
[24] G. Vernizzi, H. Orland, and A. Zee, Phys. Rev. Lett. **94**, 168103 (2005).
[25] P. G. De Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, 1979); C. Vanderzande, *Lattice Models of Polymers* (Cambridge University Press, Cambridge, 1998).
[26] T. Garel, H. Orland, and E. Orlandini, Eur. Phys. J. B **12**, 261 (1999).
[27] D. P. Foster and F. Seno, J. Phys. A **34**, 9939 (2001).
[28] J. Borg, M. H. Jensen, K. Sneppen, and G. Tiana, Phys. Rev. Lett. **86**, 1031 (2001).
[29] M. Baiesi, E. Orlandini, and A. L. Stella, Phys. Rev. Lett. **91**, 198102 (2003); P. Leoni and C. Vanderzande, Phys. Rev. E **68**, 051904 (2003).
[30] T. Garel and H. Orland, J. Phys. A **23**, L621 (1990).
[31] P. G. Higgs and H. Orland, J. Chem. Phys. **95**, 4506 (1991).
[32] Y. Iba, Trans. Jpn. Soc. Artif. Intell. **16**, 279 (2001), e-print cond-mat/0008226.
[33] V. Provman, P. C. Hohenberg, and A. Aharony, *Phase Transitions and Critical Phenomena*, edited by C. Domb and J. L. Lebowitz (Academic, New York 1991).
[34] B. G. Nickel, Macromolecules **24**, 1358 (1991).
[35] P. Grassberger, Phys. Rev. E **56**, 3682 (1997).
[36] B. Li, N. Madras, and A. D. Sokal, J. Stat. Phys. **80**, 661 (1995).