Statistical mechanics of pseudoknot polymers

Adam Lucas[a)]
Department of Mathematics, Mills College, Oakland, California 94613

Ken A. Dill
Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94143-1204

(Received 28 January 2003; accepted 7 May 2003)

We present the theory for the conformational free energies of RNA molecules and other polymers that have pseudoknots. We derive an expression for the partition function by counting conformational loops that we call pseudoknot core units, using the theory of self-avoiding walks. We show that the thermal transitions of pseudoknot polymers to their denatured states are not two-state; they are continuous. We also find that the entropy of a pseudoknot loop depends not only on its chain length, as is assumed in most current RNA models, but also on the length of the adjacent stem, because stems are stiff and the stiffness determines the starting and ending points of the loop. © *2003 American Institute of Physics.* [DOI: 10.1063/1.1587129]

I. INTRODUCTION

Pseudoknots are an important class of conformations in RNA molecules. Similar conformations occur in proteins and in other compact polymers. Figure 1 shows a pseudoknot. It is a hairpin (stem) in which one end of the chain returns and forms additional contacts with the loop between the two strands of the stem. From a computational perspective, pseudoknots are interesting because they are the simplest conformations that can be regarded as being *complex* or *tertiary structures*, in the following sense. The next simpler class of conformations are simple hairpins. The number of simple hairpin conformations scales only as N^4, where N is the chain length. The addition of a single intrachain pseudoknot contact increases the conformational complexity, to N^6.[1] Folded molecules such as proteins are even more complex, involving an exponential scaling, a^N.

Two problems have been the focus of most of the theory on RNA molecules. First, there is the problem of predicting the native structure of a molecule based on its nucleotide sequence.[1–4,11] That is not our focus here. The second class of problems is the prediction of the folding thermodynamics. The pioneering work in this area was by McCaskill, who developed a simple random-flight theory of RNA folds.[5] This and several related models treat the chain using random flight statistics and neglect excluded volume.[6,7] Considerably more sophisticated and accurate is a model that treats the steric excluded volume of the chains explicitly, rather than neglecting it.[8,9] But the latter models are limited to hairpins and nested hairpins; they are not able to treat pseudoknots, due to the complexity of the conformations. This is the problem we now treat here.

II. THE MODEL

We use a three-dimensional cubic lattice model of a chain. Lattice models are currently among the best ways to insure uniform sampling of conformational spaces of polymers. Although not perfect representations of the chain, for present purposes they are sufficient to capture the essential features of the physics. A ''contact'' is defined as any pair of monomers that are not adjacent in the chain sequence and are located on spatially adjacent lattice sites. We define a *pseudoknot core unit*, which is a chain having a part of the pseudoknot conformation. It has two components, a double-stranded *hairpin*, or *stem*, and a loop. The hairpin stem contains $2n$ nucleotides, n monomers in each strand. The loop is a single-stranded stretch having m monomers. We define a pseudoknot core unit using the label $U(n,m)$. This name is intended to indicate that the stem and loop are the two most central parts of the overall pseudoknot conformation. After describing the pseudoknot core unit in the next section, we follow with a description of the full pseudoknot. Figure 2 shows one specific conformer of the 26 nucleotide pseudoknot core unit, $U(6,14)$. $L_{U(n,m)}$ is the number of ways $U(n,m)$ can be embedded in a three-dimensional cubic lattice, with the first nucleotide fixed at the origin. Equivalently, $L_{U(n,m)}$ is the number of ''neighbor-avoiding'' walks (NAW) through the $m+2n$ lattice sites that satisfy the specific hairpin and pseudoknot contact distance constraints. A neighbor-avoiding walk is more restrictive than a self-avoiding walk; neighbor avoidance means that the chain is not only self-avoiding (i.e., having no two monomers on the same lattice site), but also that no monomer is *adjacent* to any other monomer either. The central challenge in computing the partition function is in obtaining the density of states, the conformation count. So our focus in the next sections is a description of how we obtain the conformation counts.

Consider two limiting cases. In one limit, $n=0$, there is no stem, and $U(0,m)$ is a single stranded m-nucleotide loop with a single contact between the first and last nucleotides. Another limiting case is $U(n,0)$, for which there is no single-stranded loop segment; the double-stranded chain bends around to make two contacts between the top and the bottom of the hairpin stem. Examples of these two limiting

a)Electronic mail: alucas@mills.edu

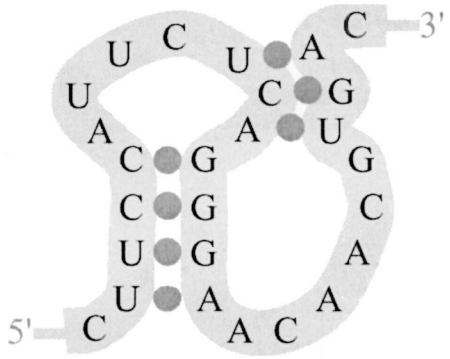


FIG. 1. A simple pseudoknot. In a pseudoknot, monomers inside a hairpin loop pair with monomers outside the stem-loop.

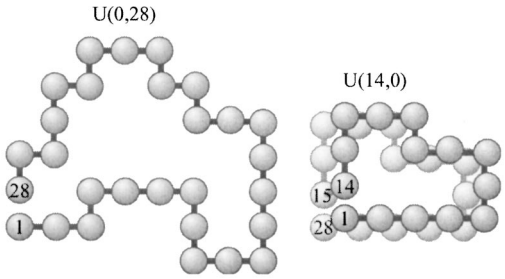


FIG. 3. The pseudoknot core units, $U(0,28)$ and $U(14,0)$ are examples of special limiting cases in which $n=0$ or $m=0$. Our theory can count conformations for these cases exactly.

situations are $U(0,28)$ and $U(14,0)$, as shown in Fig. 3.

We refer to an m-step neighbor avoiding walk in which only the first and last site are adjacent as an *m-step polygon*. Following the notation in the physics literature, an m-step polygon on a cubic lattice is often expressed as $U^{3D}(m)$.[10] Figure 3 shows a planar 28-step polygon. We denote the number of m-step polygons on a cubic lattice by $L_{U^{3D}(m)}$, so $L_{U(0,m)}=L_{U^{3D}(m)}$.

The other case, $U(n,0)$, is a planar n-step polygon on a cubic lattice. For example, $U(14,0)$ in Fig. 3 is a planar 14-step polygon. A double-stranded chain has an asymmetry in its flexibility. Think of a ribbon. A ribbon cannot readily bend around an axis that is embedded in its plane, but it can bend around an axis perpendicular to its plane. In a cubic lattice, the consecutive contacts of a double stranded chain is highly constraining and prevents the chain from twisting. Hence the flexibility of a lattice hairpin is limited to two dimensions and there are no nonplanar conformations in $U(n,0)$. An n-step polygon on a square lattice is denoted by $U^{2D}(n)$.[10] The number of n-step polygons on a square lattice is $L_{U^{2D}(n)}$. Because there are six first steps in a cubic lattice, $U(n,0)$ can have six planar orientations, so $L_{U(n,0)}=6L_{U^{3D}(n)}$. We have computed $L_{U^{3D}(n)}$ and $L_{U^{2D}(n)}$ by exact enumeration for $n<20$. For longer chains, we used asymptotic expressions previously developed in the literature for the conformation counts of self-avoiding polygons.[10]

In the more general case, where $n\neq 0$ and $m\neq 0$, the pseudoknot core unit, $U(n,m)$, has both a stem and loop.

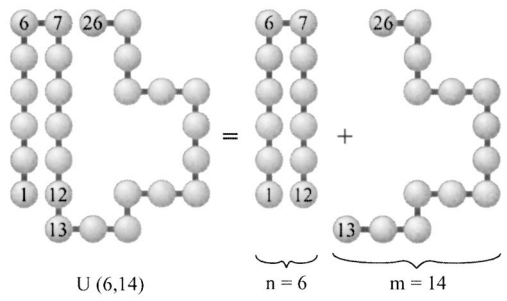


FIG. 2. The pseudoknot core unit, $U(6,14)$, is made up of a 12 monomer hairpin stem and a 14 nucleotide single stranded segment. $n=6$ is half the number of monomers in the stem and $m=14$ is the number of single stranded monomers.

The m-mer loop chain together with one strand of the stem gives a chain of length $m+n$. For example, in $U(6,14)$ of Fig. 2, there is a 20-mer loop consisting of residues 7–26. Hence, $L_{U^{3D}(m+n)}$ is an upperbound for $L_{U(n,m)}$. To approximate $L_{U(n,m)}$, we use

$$L_{U(n,m)}\approx \frac{2L_{U^{2D}(n)}}{L_{U^{3D}(n)}}L_{U^{3D}(m+n)}. \tag{1}$$

The quantity $L_{U^{2D}(n)}/L_{U^{3D}(n)}$ gives the fraction of three-dimensional loops of length n that lie in a plane, such as the one that is defined by the stem. This expression reduces to the appropriate limits. When $n=0$, it reduces to $L_{U(0,m)}=2L_{U^{3D}(m)}$ and when $m=0$, this equation reduces to $L_{U(n,0)}=2L_{U^{2D}(n)}$. In general, we find by tests against exact enumerations that this expression is also quite accurate for the more general case, for nonzero m and n.

In applying Eq. (1) we used exact values from 3D lattice exhaustive enumerations for $L_{U^{3D}(m+n)}$ for $m+n<20$. To find $L_{U^{2D}(n)}$ and $L_{U^{3D}(n)}$ in Eq. (1) we approximated conformation counts for self-avoiding polygons, rather than NAW polygons, because we expect errors to cancel out in the ratio $L_{U^{2D}(n)}/L_{U^{3D}(n)}$. For large n, the number of self-avoiding polygons of length n scales as $n^{-h}\mu^n$, where the exponent h and connective constant μ are positive and depend on the dimensionality of the lattice.[10] Self-avoiding polygons on square and cubic lattices have exponents given by $h=1.5$ and $h=1.75$, respectively, and connective constants given by $\mu=2.638$ and $\mu=4.684$, respectively.[10] For n between 16 and 22 we found by exact enumeration that the number of n step self-avoiding polygons in a square and cubic lattice is approximately $1.08n^{-1.5}(2.638)^n$ and $0.518n^{-1.75}(4.684)^n$, respectively. We used these expressions also to interpolate $L_{U^{3D}(n)}$ and $L_{U^{2D}(n)}$ for odd values of n (since square and cubic lattices artifactually do not permit odd values of n). To use $L_{U^{3D}(m+n)}$ in Eq. (1) for $m+n>=20$ we used the formula $L_{U^{3D}(m+n)}\sim 0.018(n+m)^{-1.75}(4.684)^{(n+m)}$. The scaling factor, 0.018, was chosen to agree with our exact enumerations for $10\leqslant m+n\leqslant 16$.

We summarize below our counting formulas for pseudoknot core units,

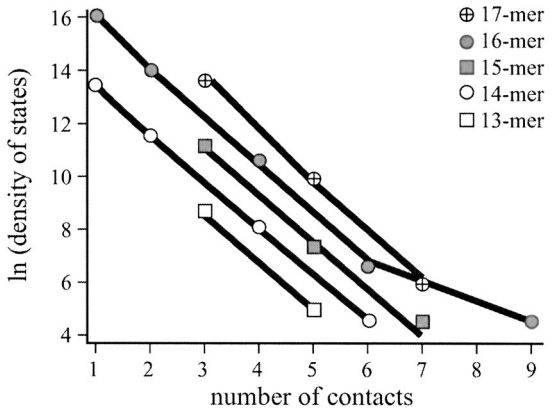


FIG. 4. The figure gives a test of theory against exact enumeration for 13-mer–17-mer pseudoknot core units, showing that the predicted density of states is almost exact. The solid lines are the theoretical prediction, and the points are the exact values. The energy function used is $E=-\epsilon\times$(number of contacts) with $\epsilon=1$.

$$L_{U(n,m)} = \begin{cases} \dfrac{2L_{U^{2D}(n)}}{L_{U^{3D}(n)}} L_{U^{3D}(m+n)} & m\neq 0,\ n\neq 0,\ m+n \text{ even} \\ 6L_{U^{2D}(n)} & m=0,\ n\neq 0,\ n \text{ even} \\ L_{U^{3D}(m)} & n=0,\ m\neq 0,\ m \text{ even.} \end{cases} \tag{2}$$

To test our theory we use a very simple model of energy, the homopolymer model in which all monomers stick to each other with an energy $-\epsilon<0$, and compute the density of states. For a fixed chain length, using Eq. (2), we can predict the number of lattice conformations different pseudoknot core units have. Keeping track of the number of contacts in each pseudoknot core unit, we get the density of states. We compare our theoretical density of states prediction with the precise density of states obtained by exact enumeration. Figure 4 shows that our model is nearly exact. The exact enumerations were computationally limited to 17-monomer pseudoknot core units. The advantage of the present theory is that it is much more efficient than exact enumeration; the complexity of our counts grow linearly with chain length. So our model computes the density of states in only a fraction of a second for $N<50$. The rate limiting step is the linear time computation of determining which pairs of m and n satisfy the constraint $N=m+2n$.

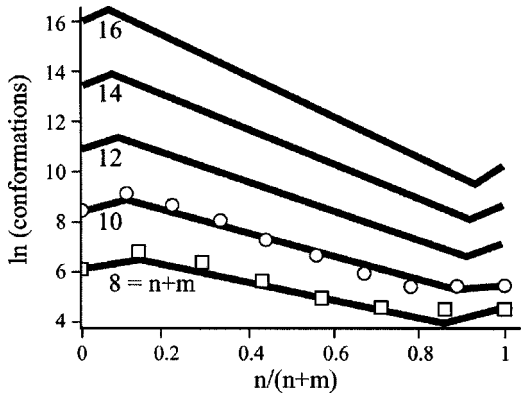
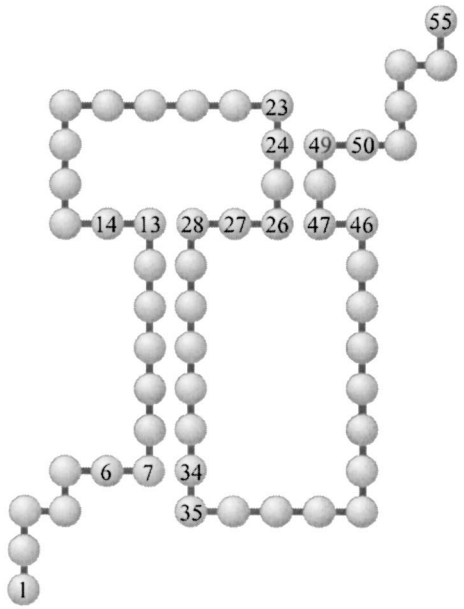


FIG. 5. In a pseudoknot core unit, $U(n,m)$, $n+m$ monomers form a closed loop (we only consider half the monomers in the stem). For fixed loop size, $n+m$, between 8 and 16 monomers ($n+m$ even), the conformation counts for pseudoknot core units, with different fractions of stem content, $n/(n+m)$, is predicted. The figure shows that conformation counts for pseudoknot core units decrease with increasing stem content. Solid lines represent theoretical counts and points ($n+m=8$ and $n+m=10$ only) represent exact counts.



t1 = 6 involves nucleotides {1-6}
t2 = 6 involves nucleotides {50-55}
o = 5 involves nucleotides {13, 26, 27, 28, 47}
n1 = 7 involves nucleotides {7-13 and 28-34}
m1 = 15 involves nucleotides {35-47 and 26-27}
n2 = 3 involves nucleotides {24-26 and 47-49}
m2 = 13 involves nucleotides {27-28 and 13-23}

N = t1 + t2 + 2n1 + 2n2 + m2 - o

FIG. 6. A planar representation of a pseudoknot in a cubic lattice. A pseudoknot is described by 7 parameters. The value of each parameter is given as well as the monomer numbers associated with each value. The final equation expresses the total number of monomers as a sum of the parameter values.

Here is a test that demonstrates the advantage of the present theory compared to the traditional random-flight method for treating loop conformations. Figure 5 shows that increasing the proportion of the hairpin stem in a pseudoknot core unit decreases the number of conformations accessible to the loop. The loop gets ''stretched'' as the stem gets longer, leading to fewer conformations. In random-flight theory, the loop conformation count is, incorrectly, independent of the stem length. Hence the present model improves upon this and related flaws of random-flight loop models.

III. MORE COMPLEX PSEUDOKNOTS

Our discussion above is limited to only the simplest possible pseudoknots, namely those that have only a stem of length n and loop of length m. To make the model more realistic, we now consider a pseudoknot chain that also has *a*

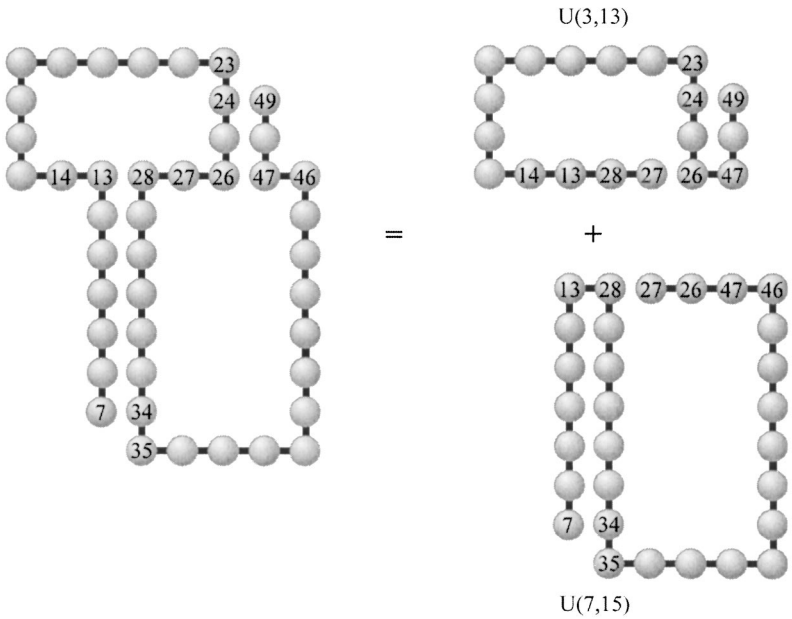


FIG. 7. A pseudoknot minus its tails can be decomposed as the sum of two pseudoknot core units. The figure shows the decomposition of the pseudoknot on the left as the sum of the pseudoknot core units $U(3,13)$ and $U(7,15)$. Monomers 13, 28, 27, 26, and 47, are superimposed in the two pseudoknot core units to form the pseudoknot on the left.

tail at each end, and for which there may be more than one tertiary contact. In this class of conformers, we partition the chain monomers into 7 nondisjoint groups. Figure 6 shows an example. The length of the two tails of the pseudoknot, i.e., the number of monomers free of any contacts at the beginning and the end of the pseudoknot, are labeled $t1$ and $t2$. For example, in Fig. 6, the tail monomers in $t1$ are 1–6 and in $t2$ are 50–55. The full pseudoknot can be assembled from the two tails and two pseudoknot core units $U(n1,m1)$ and $U(m2,n2)$. Let $m1$ be the number of nucleotides forming the single stranded segment, and $n1$ be half the number of nucleotides in the hairpin stem of $U(n1,m1)$. We use identical definitions for $n2$ and $m2$ in $U(n2,m2)$. In Fig. 6, the pseudoknot core units are $U(3,13)$ and $U(7,15)$ (shown explicitly in Fig. 7). The two pseudoknot core units have monomers in common. For example, in Fig. 7, nucleotides 13, 28, 27, 26, and 47, are common to both pseudoknot core units. The number of these overlapping nucleotides in the pseudoknot is denoted o. In this case $o=5$.

For the purpose of counting conformations to get the density of states, we need not distinguish which bonds are covalent and which are noncovalent. For example nucleotides 26 and 27 form a noncovalent bond in $U(3,13)$ yet a covalent bond in $U(7,15)$. Both types of interaction impose an identical distance constraint, for the purpose of counting conformers.

A. Generating contact maps for pseudoknots

To compute the density of states for a chain with N monomers, we generate all combinations of the parameters $t1$, $t2$, o, $m1$, $m2$, $n1$, and $n2$ subject to the following conditions:

$$m1 \geqslant 4,$$

$$m2 \geqslant 4,$$

$$n1 \geqslant 1,$$

$$n2 \geqslant 1,$$

$$t1 \geqslant 0,$$

$$t2 \geqslant 0,$$

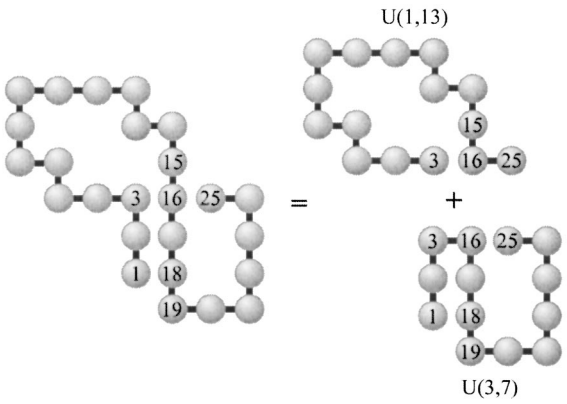
$$3 \leqslant o \leqslant \min((m1+5)/2,(m2+5)/2),$$

$$m1+n1 \quad \text{even},$$

$$m2+n2 \quad \text{even},$$

$$N = t1+t2+2n_2+2n_1+m_1+m_2-o.$$

Here are the reasons for these conditions. The conditions on $n1$ and $n2$ is so that the pseudoknot will have two hairpin stems. The conditions on $m1$ and $m2$ are required to form the head of the two hairpin stems in the pseudoknot. The single stranded tails of our pseudoknots may have any number of nucleotides. Figure 8 shows that the minimum possible overlap between two pseudoknot core units is three. The maximum number of overlaping nucleotides was chosen to be the closest integer less than or equal to $\min((m1+5)/2,(m2+5)/2)$. Figure 9 shows a pseudoknot with maximal overlap, o. The greater the overlap between two pseudoknot core units, the more compact the pseudoknot. Such pseudoknots usually do not have planar representations. As a result, it is likely that some of the pseudoknots with large overlap will violate excluded volume constraints in a cubic lattice and not be physically viable. We chose our range of overlaps to be so that 98% of our pseudoknots having $N=18$ nucleotides are physically viable in a cubic lattice. For larger pseudoknots, we expect an even smaller fraction of excluded volume violations, since they are less compact, on average. We require $m1+n1$ and $m2+n2$ to have even parity because of the constraints of the cubic lattice. The final equation constrains the parameters so that the pseudoknot has the correct number of nucleotides.



t1 = 0	involves no nucleotides
t2 = 0	involves no nucleotides
o = 3	involves nucleotides {13, 16, 25}
n1 = 3	involves nucleotides {1-3 and 16-18}
m1 = 7	involves nucleotides {19-25}
n2 = 1	involves nucleotides {16, 25}
m2 = 13	involves nucleotides {3-15}

FIG. 8. An example of a pseudoknot with minimal overlap, $o=3$, between its two pseudoknot core units.

For a chain of 30 nucleotides there are 14167 different sets of parameters $t1$, $t2$, o, $m1$, $m2$, $n1$, $n2$. Each set of parameters corresponds to a single contact map. So our procedure generates 14167 different contact maps representing a diverse assortment of 30 nucleotide pseudoknots. Each contact map represents the full ensemble of conformations that satisfy those distance constraints.

B. Generation of contact maps for hairpins and the open chain

To explore the transition between a pseudoknot and its denatured open conformations, we consider a zipping process in which the intermediate structures between the native and denature states can have any degree of zipping or unzipping of either of the two hairpins. Unzipping is only allowed to initiate from the top or bottom of the hairpins. Figure 10 shows an example of allowed conformations on the route between the pseudoknot of Fig. 6 and the open chain. We expect this to be a good approximation to the full ensemble of conformers that are significantly populated during the folding and unfolding processes.

To explore such transitions, we consider a class of hairpins where the monomers can be partitioned into 4 disjoint groups. The variables are $t1$, $t2$, n, and m. $t1$ and $t2$ are the length of the tails of the hairpin, i.e., the numbers of nucleotides that are free of any contacts at the beginning and the end of the hairpin. n is half the number of nucleotides in the hairpin stem and m is the number of nucleotides involved in the loop. Figure 10 gives an illustration.

For a chain with a fixed number of nucleotides N, we generate all combinations of the numbers $t1$, $t2$, m, and n subject to the following conditions:

$$m \geqslant 4 \quad \text{even},$$

$$n \geqslant 1,$$

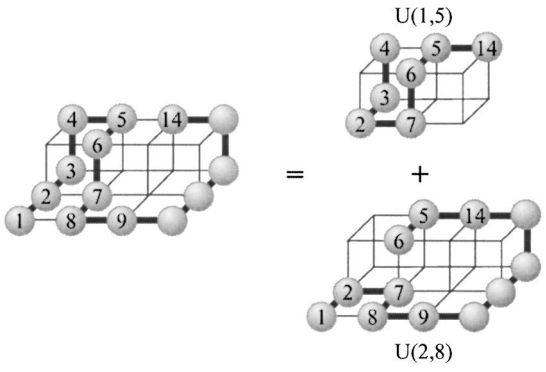
$$t1 \geqslant 0,$$

$$t2 \geqslant 0,$$

$$N = t1 + t2 + 2n + m.$$

The loop in the hairpin will have $m+2$ nucleotides, so we require $m \geqslant 4$ with m having even parity. The hairpin must have at least one contact in the stem so we require $n \geqslant 1$. We allow the tails of the hairpin to be any length.

For a chain of $N=30$ nucleotides there are 819 different sets of parameters $t1$, $t2$, m, n. Each set of parameters corresponds to a single contact map, so our procedure generates 819 different contact maps representing a diverse assortment of 30 nucleotide hairpins. All contact maps for hairpins are physically viable on a cubic lattice.



t1 = 0	involves no nucleotides
t2 = 0	involves no nucleotides
o = 5	involves nucleotides {14, 5, 6, 2, 7}
n1 = 1	involves nucleotides {5,14}
m1 = 5	involves nucleotides {2-4, 6, 7}
n2 = 2	involves nucleotides {1, 2, 7, 8}
m2 = 8	involves nucleotides {9-14, 5, 6}

FIG. 9. An example of a pseudoknot with maximal overlap, $\min((m1+5)/2,(m2+5)/2)=5$, between its two pseudoknot core units.

IV. THE PARTITION FUNCTION FOR A PSEUDOKNOT AND ITS UNFOLDING INTERMEDIATES

As a test of principle of the model, we consider a simple monomer sequence in which all monomers are identical, so that the energy is simply proportional to the number of intrachain contacts, with proportionality constant, ϵ. The partition function over the conformations of a chain can be computed either by summing over contact maps, or by summing over energy levels. We number all the possible contact maps. Letting E_i be the energy of the ith contact map, the partition function $Q(T)$ is

$$Q(T) = \Sigma_i \Omega_i e^{-E_i/RT}, \tag{3}$$

where R is the gas constant and T is the temperature. $\Omega_i e^{-E_i/RT}$ is the statistical weight for the ith contact map.

It remains to determine Ω_i, the number of conformations corresponding to the ith contact map. To do this by exact enumeration would require searching 5^N possible conformers for a chain of N nucleotides. We present below a more efficient algorithm for estimating Ω_i. To reduce the conformational complexity, we take advantage of the reduc-

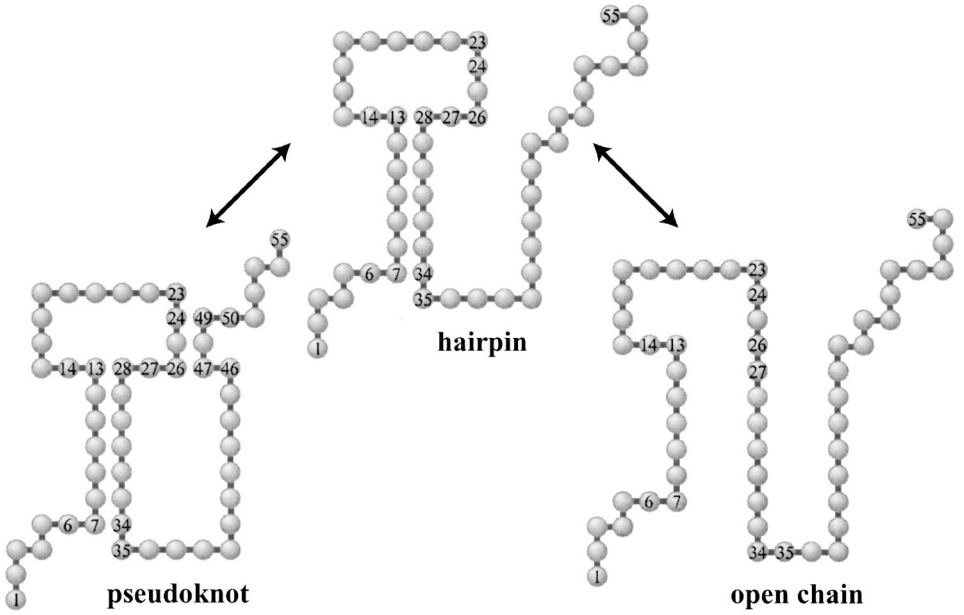


FIG. 10. The transition between a pseudoknot, a hairpin and an open chain involves unzipping at the ends of the pseudoknot stems. The 4 parameters describing the intermediate hairpin is given.

tion in dimensionality imposed by the constraint that the bending of a hairpin stem lies in a plane. For pseudoknots, we factor Ω_i into component counts from its two pseudoknot core units. We follow the definitions of Madras and Slade for self-avoiding walks except that we will be considering the more restrictive case of NAWs rather than SAWs.[10] An n step NAW has $n+1$ monomers.

Definition 1: $L_{C^{3D}(n)}$ is the number of n-step NAWs on a cubic lattice.

Definition 2: $L_{C^{2D}(n)}$ is the number of n-step. NAWs on a square lattice.

An open chain with $n+1$ monomers is an n-step NAW and hence has precisely $L_{C^{3D}(n)}$ conformations on a cubic lattice. We found $L_{C^{3D}(n)}$ and $L_{C^{2D}(n)}$ by exact enumeration for $n \leqslant 16$. For chains with $n>17$, $L_{C^{3D}(n)}$ was approximated by the formula $L_{C^{3D}(n)}=0.096(n)^{0.16}(4.684^n)$. In this formula we assume that the critical exponent, 0.16, and connective constant, 4.684, for NAWs is the same as for SAWs on a cubic lattice. The scaling factor for the longer chains, 0.096, was chosen to agree with our exact enumerations for $10 \leqslant n \leqslant 16$.

To discuss our counting algorithm for our class of hairpins we first handle the special case of a simple hairpin bend (i.e., a hairpin which is entirely double stranded) in the following simple theorem.

Theorem 1: A hairpin bend occupying $2n+2$ sites on a cubic lattice has exactly $6L_{C^{2D}(n)}$ conformations.

Proof: Given a hairpin bend of $2n+2$ nucleotides, every nucleotide in the chain, except the middle two forming the bend in the hairpin, are involved in a contact. This constrains the hairpin and prevents the chain from twisting. As a result the hairpin loses a degree of freedom and can only move in a two-dimensional plane. There are six possible first steps for the hairpin. Fixing the first step, a conformation of a hairpin is determined by an n step ($n+1$ monomer) neighbor avoiding walk in a square lattice. There are $L_{C^{2D}(n)}$ conformations of an n step neighbor avoiding walk in a square lattice, giving therefore a total of $6L_{C^{2D}(n)}$ conformations for a hairpin bend. □

To handle more general hairpins we approximate the tails as a single 3D NAW of length $t1+t2$. Hence two tails together have $L_{C^{3D}(t1+t2)}$ conformations. We introduce an excluded volume term, π_t, for the interaction of the tails with the double stranded part of the hairpin. We found by exact enumeration that without an excluded volume term we overcounted hairpin conformations by a factor of 5/3 so we let π_t be 0.60 when t_1 or t_2 are greater than 0. There is no need for an excluded volume term when t_1 and t_2 are both 0 and so $\pi_t=1$ in this case.

To handle hairpins where the bend at the head of the hairpin involves more than two monomers (i.e., $m>2$), we model the turn at the top of the hairpin as a $(m+1)$-step polygon, $U^{3D}(m+1)$. We found by exact enumeration that without an excluded volume term we overcounted hairpin conformations by a factor of 3 so we let π_h be 0.33 if $m>2$ and $\pi_h=0$ otherwise.

The configuration count for a hairpin with parameters, m, n, $t1$, and $t2$ is given by the equation,

$$L_h=\pi_h*\pi_t*L_{C^{2D}(n)}*L_{U^{3D}(m)}*L_{C^{3D}(t1+t2)}. \tag{4}$$

The quantity $L_{C^{2D}(n)}*L_{U^{3D}(m)}*L_{C^{3D}(t1+t2)}$ in Eq. (4) is the product of the number of conformations contributed by the stem, the head of the hairpin, and the tails. This is multiplied by excluded volume factors for the head and tails interacting with the stem of the hairpin. Figure 11 shows that Eq. 4 is almost exact for $11 \leqslant N \leqslant 15$, the chain lengths that are accessible by exact enumeration.

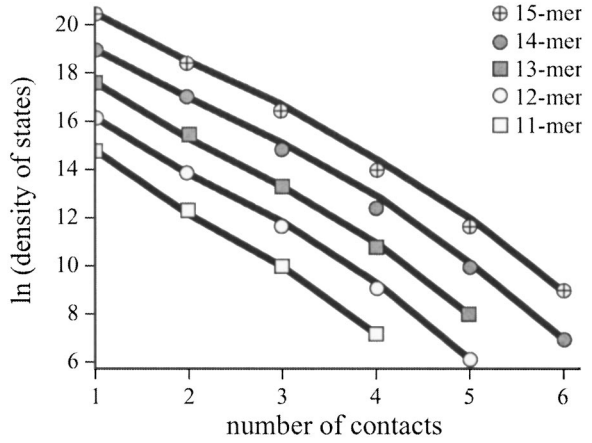


FIG. 11. Test of theory against exact enumeration for 11-mer–15-mer hairpin/open chain ensembles, showing that the predicted density of states is essentially exact. The energy function used is $E=-\epsilon\times$(number of contacts) with $\epsilon=1$.

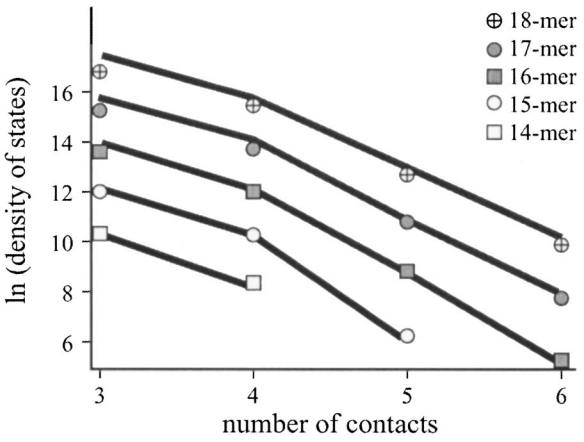


FIG. 12. Test of theory against exact enumeration for 14-mer–18-mer pseudoknots, showing that the predicted density of states is essentially exact. The energy function used is $E=-\epsilon\times$(number of contacts) with $\epsilon=1$.

To find the conformational redundancy of a contact map for a pseudoknot in a cubic lattice we begin with a pseudoknot without tails (i.e., $t_1=0$ and $t_2=0$). We have seen that such a pseudoknot is the sum of two pseudoknot core units, $U(n1,m1)$ and $U(n2,m2)$, with a single stranded segment having o nucleotides, double counted. It seems reasonable then to count such a pseudoknot as the product of $U(n1,m1)$ and $U(n2,m2)$ divided by $L_{C^{3D}(o)}$. We found empirically that the excluded volume, π, due to the local interaction of the two pseudoknot core units depends on the length of the stems as well as the length of the nonoverlapping single stranded section of the loops, $m1-o$ and $m2-o$. In the case where $n1>0$ and $n2>0$, if $m1-o>2$ and $m2-o>2$, then $\pi=\frac{1}{35}$ otherwise if $m1-o\leqslant 2$ or $m2-o\leqslant 2$ we let $\pi=\frac{1}{70}$. This reflects our observation that the excluded volume is greater between the two pseudoknot core units when one of the single stranded loops is very short.

We handle pseudoknots with $t1>0$ and $t2>0$ by modeling these single strands as three dimensional neighbor avoiding walks of length $t1-1$ and length $t2-1$, respectively, having $L_{C^{3D}(t1-1)}$ and $L_{C^{3D}(t2-1)}$ many conformations. We handle the local excluded volumes π_{t1} and π_{t2} by setting $\pi_{t1}=\frac{2}{3}$, when $t1>0$ and $\pi_{t2}=\frac{2}{3}$, when $t2>0$. Otherwise, these excluded volume factors are equal to 1. These factors were determined similarly to the excluded volume terms for the tails of hairpins discussed above.

The configuration count for a pseudoknot, with parameters, $n1$, $m1$, $n2$, $m2$, $t1$, $t2$, and o is given by the equation,

$$L_p=\pi*\pi_{t1}*\pi_{t2}*L_{C^{3D}(t1)}*L_{C^{3D}(t2)} *L_{U(n1,m1)}*L_{U(n2,m2)}/L_{C^{3D}(o)}. \quad (5)$$

The quantity $L_{U(n1,m1)}*L_{U(n2,m2)}/L_{C^{3D}(o)}$ in Eq. (5) is the contribution to the conformation count from the pseudoknot minus its tails. The quantity $L_{C^{3D}(t1)}*L_{C^{3D}(t2)}$ represents the contribution from the tails of the pseudoknot. The remaining terms are excluded volume factors. Figure 12 shows that Eq. (5) is nearly exact for $14\leqslant N\leqslant 18$. We notice a trend that our theory uniformly overestimates the exact counts starting with $N>16$.

V. PSEUDOKNOT HOMOPOLYMER SHOWS SINGLE STATE COOPERATIVE TRANSITION

We examined the free energy plot for the pseudoknot/hairpin/open chain homopolymer ensemble. Fixing the size of the chain at $N=20$ our pseudoknots/hairpin/open chain ensemble allows a maximum of 9 contacts. The free energy of the ensemble with energy E, and a fixed temperature, T, is given by the formula,

$$F(E,T)=E-kT*\ln g(E), \quad (6)$$

where k is Boltzmann's constant, and $g(E)$ is the density of states,

$$g(E)=\Sigma_i\Omega_i|_{E_i=E}. \quad (7)$$

Every contact in a contact map was given an energy $\epsilon=1$. $g(E)$ was determined from summing the conformation counts of all contact maps having energy E. Figure 13 shows that for a chain with $N=20$ monomers, the thermal transition between the pseudoknot native structure and its denatured state is 1-state, i.e., a continuous transition. We tested chains

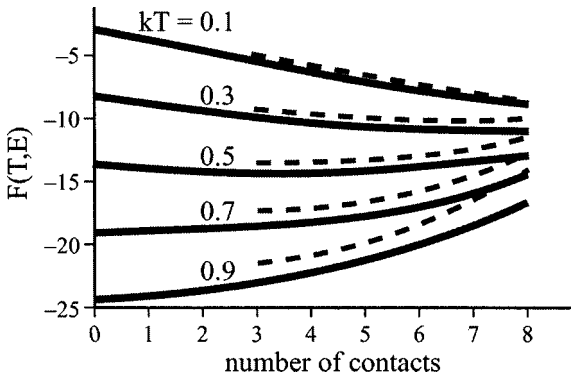


FIG. 13. A free energy plot for a 20-mer pseudoknot (dotted lines), and pseudoknot/hairpin/open chain ensemble (solid lines). The energy function used is $E=-\epsilon\times$(number of contacts) with $\epsilon=1$. The thermal transition for pseudoknot/hairpin/open chain ensembles is 1 state with the open chain stable at high temperature and pseudoknots favored at low temperatures. The zipping of pseudoknots is also shown to be continuous.

as long as $N=50$ and found the same result suggesting that in the limit of long chain length, the transition will be 1-state.

The transition for simple hairpins is also known to be 1-state. We confirmed this (figure not shown). Figure 13 shows that pseudoknots exhibit single state transitions. In contrast, Chen showed that a more general class of hairpin homopolymers, having multiple nucleation points along the chain for zipping and unzipping, have 2-state transitions.[8]

VI. CONCLUSION

We have developed an analytic theory for pseudoknot chain conformations that treats the excluded volume explicitly. Although many RNA structures of interest are much larger than what is studied here, these structures can contain pseudoknots of the form studied in this manuscript. This makes small structures worth studying. Our method is to decompose a pseudoknot into two pseudoknot core units, which are chains having a part of the pseudoknot conformation. We use the theory of self-avoiding walks to estimate the number of conformations for pseudoknot core units. To demonstrate the advantage of our model compared to the traditional random-flight method of treating loop conformations, we show that increasing the proportion of the hairpin stem in a pseudoknot core unit decreases the number of conformations accessible to the loop. In random-flight theory, the loop conformation count is, incorrectly, independent of stem length. The counts for a pseudoknot is essentially found by multiplying the counts for the two pseudoknot core units and dividing by a factor related to the length of the overlapping section. Our theory is shown to predict well the partition function in simple test cases that have been analyzed by exact enumeration on lattices. We show that transitions from pseudoknots to their unfolded states are not first-order in this model.

ACKNOWLEDGMENT

The authors thank Sarina Bromberg for help with the figures.

[1] E. Rivas and S. Eddy, J. Mol. Biol. **285**, 2053 (1999).
[2] E. Rivas and S. Eddy, Bioinformatics **16**, 334 (2000).
[3] M. Zucker and D. Sankoff, Bull. Math. Biol. **46**, 591 (1984).
[4] M. Zuker, Science **244**, 48 (1989).
[5] J. S. McCaskill, Biopolymers **29**, 1105 (1990).
[6] R. Bundschuh and T. Hwa, Phys. Rev. E **65**, 031903 (2002).
[7] H. Isambert and E. Siggia, Proc. Natl. Acad. Sci. U.S.A. **97**, 6515 (2000).
[8] S.-J. Chen and K. A. Dill, J. Chem. Phys. **109**, 4602 (1998).
[9] D. Chiang and A. Joshi, Proceedings of HLT 2002, San Diego, 2002.
[10] N. Madras and G. Slade, *The Self-Avoiding Walk* (Birkhauser, Boston, 1993).
[11] R. B. Lyngso and C. N. Pedersen, J. Comput. Biol. **7**, 409 (2000).