Steepest descent calculation of RNA pseudoknots

M. Pillsbury,[1] Henri Orland,[2] and A. Zee[1,3]
[1]*Department of Physics, University of California, Santa Barbara, California 93106, USA*
[2]*Service de Physique Théorique, CEA Saclay, 91191 Gif-sur-Yvette Cedex, France*
[3]*Institute for Theoretical Physics, University of California, Santa Barbara, California 93106, USA*
(Received 30 July 2002; revised manuscript received 22 August 2003; published 19 July 2005)

We enumerate possible topologies of pseudoknots in single-stranded RNA molecules. We use a steepest-descent approximation in the large N matrix field theory, and a Feynman diagram formalism to describe the resulting pseudoknot structure.

DOI: 10.1103/PhysRevE.72.011911 PACS number(s): 87.15.Cc

An RNA molecule is a heteropolymer strand made up of four types of nucleotides, uracil (U), adenine (A), guanine (G), and cytosine (C). The sequence of these nucleotides, or bases, makes up the molecule's primary structure. Bases form hydrogen bonds with each other to give the molecule a stable shape in three dimensions, with U bonding to A, and C to G. Calculating the shape a given primary structure will fold into is important in molecular biology.

We can associate $-U_{ij}$ with the energy of forming a hydrogen bond between the ith and jth bases, and let $V_{ij} = \exp(U_{ij}/T)$ where T is the temperature. This is a minimalist model: we make no attempt to account for loop penalties or stacking interactions. There is some rigidity in the chain of nucleotides, as well as steric constraints, which prevent hydrogen bonding between nucleotides that are within four bases of each other, so we let $V_{i,i+k}=0$ if $k<4$. The partition function associated with this bonding is given by

$$Z_{L,1} = 1 + \sum_{i<j} V_{ij} + \sum_{i<j<k<l} V_{ij}V_{kl} + \sum_{i<j<k<l} V_{ik}V_{jl} + \sum_{i_1<i_2<\cdots<i_n} V_{i_1i_2}V_{i_3i_4}\cdots V_{i_ni_{n+1}} + \cdots . \quad (1)$$

Evidently, $Z_{L,1}$ is the combinatorial heart of the RNA folding problem [1]. While $Z_{L,1}$ appears very simple at first glance, it contains a term for every possible configuration of bonds on the chain. Finding the folded state could involve searching through $\sim L!$ terms, which is a daunting task for even the shortest RNAs.

Fortunately, in RNA, there is a hierarchical separation between primary, secondary, and tertiary structures that reduces the number of configurations that must be considered. One can find the secondary structure by drawing the chain of nucleotides around the circumference of a circle, with the first nucleotide next to the last, and finding a bond structure that minimizes the free energy with the constraints that all bonds are drawn as arcs within the circle, and no bonds cross. Another representation is to draw the bond structures as systems of parallel arches which do not cross. This planar configuration (in the sense used in Refs. [2,3], though other usages are common in the RNA folding literature) is made up of the secondary structure's characteristic loops and bulges. Bonds between distinct parts of the secondary structure are called pseudoknots, and are typically considered part of the molecule's tertiary structure. For instance, the contributions from the third sum in (1) come from pseudoknot configurations. The formation of the tertiary structure is believed not to alter the more stable secondary structure [4,5].

Secondary and tertiary structures are usually stable at biological temperatures, which are typically well below the RNA molecule's melting point. This makes certain very efficient algorithms for determining RNA secondary structure at zero temperature possible and useful. These "dynamic programming" methods involve recursively calculating $Z_{L,1}$ and then backtracking to find the dominant terms, and thus determine which bonds are present in the folded RNA. There are also dynamic programming techniques that try to account for pseudoknots, but they are necessarily slower [2,6,7].

The distinction between secondary and pseudoknot structure has a topological flavor. One powerful tool for dealing with topological considerations is the large N expansion used in matrix field theories. Originally proposed by 't Hooft to represent quantum chromodynamics with N colors, it predicts that nonplanar Feynman diagrams have amplitudes proportional to negative powers of N, and are thus suppressed when N is large [3,8]. Two of the authors applied a similar technique to the problem of RNA folding, leading to the same sort of suppression of nonplanar configurations; we summarize the results below, and refer the reader to Ref. [1] for details.

One can perform a series of manipulations to find that a chain of L bases has

$$Z_{L,1} = \frac{1}{C}\int dA\, e^{-(N/2)[\mathrm{tr}\, A^2 + 2\,\mathrm{tr}\ln M(A)]} M^{-1}(A)_{L+1,1}, \quad (2)$$

where the integral is taken over all Hermitian $(L+1)\times(L+1)$ matrices A. C is an unimportant normalization constant and M is a matrix function of A given by

$$M_{ij} = \delta_{ij} - \delta_{i,j+1} + i\sqrt{V_{i-1,j}}A_{i-1,j}. \quad (3)$$

Here, N is used to keep track of topology. as mentioned above. Thus we can expand in powers of $1/N$ and evaluate the integral by steepest descent. We need to find the stationary point of the "action"

$$S(A) \equiv \frac{1}{2}\mathrm{tr}\, A^2 - \mathrm{tr}\ln M(A) \quad (4)$$

which requires solving $\delta S(A)/\delta A = 0$. This occurs at the point $\tilde{A}$, which is defined by

 ©2005 The American Physical Society

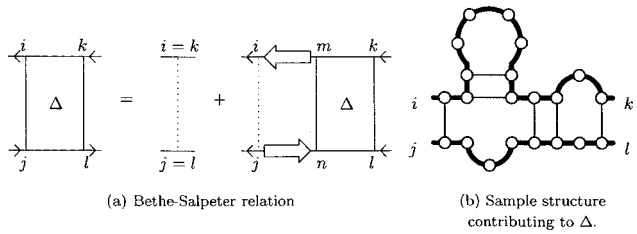

FIG. 1. The Δ propagator.

$$\tilde{A}_{lk} = i\sqrt{V_{lk}}(M^{-1})_{l,k+1}. \tag{5}$$

We define a new matrix in terms of M^{-1} at the stationary point

$$G_{ij} = (M^{-1})_{i+1,j} \tag{6}$$

and use the trivial identity $\Sigma_j M_{ij}(M^{-1})_{jk} = \delta_{ik}$ to derive the Hartree equation

$$G_{i+1,k} = \delta_{i+2,k} + G_{ik} + \sum_j V_{i+1,j} G_{i,j+1} G_{j-1,k}. \tag{7}$$

This equation is recursive, and we need to impose the boundary condition that $G_{i,i+l}=0$ for $l \geqslant 2$ to solve it. Then, G_{ij} is the partition function of the helical secondary structure of a chain that starts with the jth base and ends with the ith base. This form is precisely that used in existing dynamic programming algorithms [2,4,6]. Since it carries two indices, G_{ij} is analogous to the quark propagator in large N QCD, which carries two indices for color. The recursion relation ensures that it is a "dressed" propagator.

We can then introduce the fluctuation x_{ij}, defined by $A_{ij} = \tilde{A}_{ij} + x_{ij}/\sqrt{N}$, and expand $\mathrm{tr}\log(M^{-1}(A))$ and $M^{-1}(A)$ as power series in x. Then we collect powers of $N^{-1/2}$ to find corrections to the steepest descent approximation of $Z_{L,1}$. We are left with Gaussian integrals in x_{ij} that can be evaluated by applying Wick's theorem, with contractions given by the inverse of the quadratic form in the exponential. This inverse is a propagator which satisfies the Bethe-Salpeter equation

$$\Delta_{kl,mn} = \delta_{km}\delta_{nl} + \sum_{ij} V_{kl}^{1/2} V_{ij}^{1/2} G_{k-1,i+1} G_{j-1,l+1} \Delta_{ij,mn}. \tag{8}$$

While the Hartree equation gave the partition function for a single contiguous chain of RNA interacting with itself, the Bethe-Salpeter relation gives the contribution from two separated segments.

Physically, Δ represents the resummation of all ladder diagrams between anti-parallel segments, where each segment is itself dressed by secondary structure elements. Equation (8) can be represented pictorially, as in Fig. 1(a). There are four indices on $\Delta_{ij,kl}$, indicating where the segments begin and end, so we call it a "gluon propagator," in analogy to gluon propagators in QCD, which carry four color indices. A typical structure contributing to Δ is shown in Fig. 1(b). Each single line propagator is dressed by any system of arches. There can be any number of parallel interactions between the two strands. The only constraint is that no interaction lines should cross. Note that a system of arches along one line (RNA strand) typically represents a piece of helix on this strand, whereas a system of parallel interactions between the two lines can represent a helical fragment between the two strands.

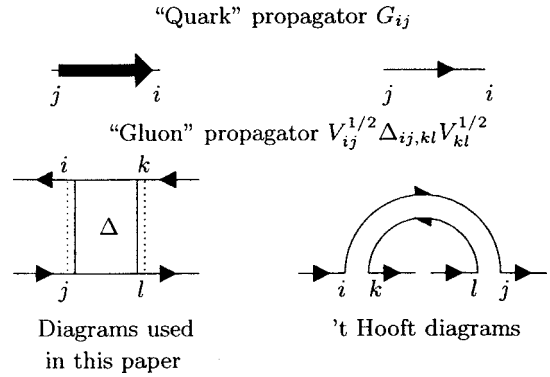

FIG. 2. Propagators.

There are two ways of drawing Feynman diagrams for these propagators. The first was introduced in Ref. [1], and is useful for visualizing the RNA's topology. The second is the double-line formalism of 't Hooft, which makes it very easy to find a graph's order in $1/N$, by assigning appropriate powers of N to loops, edges, and vertices [3,8]. It follows from Eq. (8) that the Δ propagator contains powers of $V_{ij}^{1/2}$, but the partition function (1) contains only whole powers of V_{ij}. Thus, all Δ's in the expansion appear with factors of $V_{ij}^{1/2}$, as $V_{ij}^{1/2}\Delta_{ij,kl}V_{kl}^{1/2}$. This is reflected in the diagrams in Fig. 2. We can then expand $Z_{L,1}$ to order N^{-2}, getting the secondary structure as well as the tertiary correction to it. Then

$$Z_{L,1} = G_{L,1} \tag{9}$$

$$+ \frac{1}{N^2}\left\langle\left[\left(B_4 - \frac{1}{4}B_2T_4 - \frac{1}{3}B_3T_3 - \frac{1}{5}B_1T_5 + \frac{1}{12}B_1T_3T_4 + \frac{1}{18}B_2T_3^2 - \frac{1}{162}B_1T_3^3\right)M^{-1}\right]_{L+1,1}\right\rangle, \tag{10}$$

where we use the value of M^{-1} at the stationary point from Eqs. (5) and (6). We have also introduced some convenient shorthand for matrices and traces that contain powers of x,

$$c_{ij} = \sqrt{V_{i-1,j}}x_{i-1,j},$$

$$D_{mn} = \sum_{m'} (M^{-1})_{mm'}c_{m'n},$$

$$(B_p)_{kl} = (D^p)_{kl},$$

$$T_p = \mathrm{tr}\, B_p.$$

The angle brackets in Eq. (10) mean the included terms should be integrated over x_{ij} with the Gaussian weight $\exp\{-[\mathrm{tr}\, x^2 + \mathrm{tr}(M^{-1}c)^2]/2\}$. These integrals are simple in principle, as the x_{ij}'s can be contracted with the Bethe-Salpeter propagator (8). Each power of x introduces a vertex for gluon lines.

The multiplication implicit in the definition of B_p is matrix multiplication, so many indices must be summed over

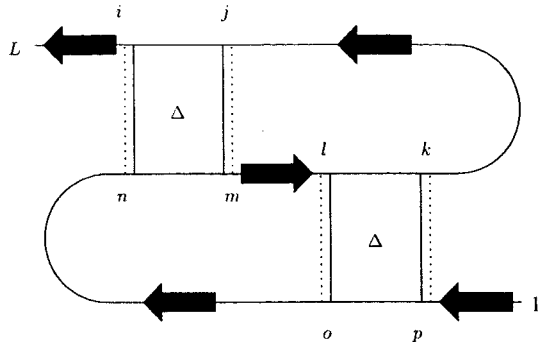

FIG. 3. Diagram for $\langle [B_4 M^{-1}]_{L+1,1} \rangle$.

when evaluating the terms in Eq. (10). For instance, evaluating one of the contractions of $\langle B_4 M^{-1} \rangle$ produces the sum

$$\langle (B_4 M^{-1})_{L+1,1} \rangle = \sum_{\substack{i,j,k,l,\\ m,n,o,p}} G_{L,i+1} G_{j,k+1} \times G_{l,m+1} G_{n,o+1} G_{p,1} \times V^{1/2}_{i,n+1} \Delta_{i,n+1,j+1,m} V^{1/2}_{j+1,m} \times V^{1/2}_{k,p+1} \Delta_{k,p+1,l+1,o} V^{1/2}_{l+1,0}. \tag{11}$$

Looking at the diagram associated in the contraction in Fig. 3, and using the condition that $G_{a,a+b}=0$ for $b \geq 2$, we deduce the proper constraint for the indices, $L \geq i > j \geq k > l \geq m > n \geq o > p \geq 0$.

The B_m and T_n terms have simple 't Hooft diagrams, as shown in Fig. 4. The ellipses in the diagram represent the string of m or n gluon vertices associated with those terms. The graph for T_n closes on itself, reflecting the trace's cyclic symmetry.

These diagrams make it simple to pick out the Wick contractions that actually contribute to $Z_{L,1}$. One can draw Feynman diagrams for the contractions of the seven terms in Eq. (10), and find that 25 of them are distinct (many contractions are equivalent under the cyclic symmetry of the traces T_n). However, most of these vanish, as they contain closed G loops. Diagrams involving closed loops will depend on a factor of $G_{i,i+l}$ for $l>2$, and therefore vanish. This can also be understood in terms of the diagrams from Ref. [1], where G's represent segments of RNA, and Δ's represent interactions between two segments. A closed G loop with both ends connected to the same side of a Δ propagator describes a closed loop of RNA interacting with the main strand. We have specifically excluded this possibility from our definition of $Z_{L,1}$, so such configurations must vanish. This is the reason why there is no graph of order $1/N$ in Eq. (9).

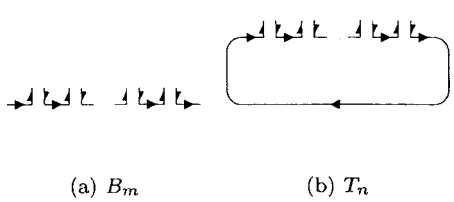

FIG. 4. Matrix products

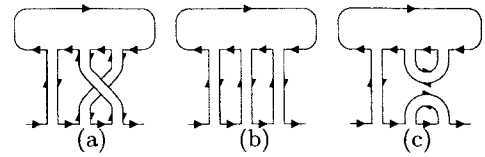

FIG. 5. Contractions for $\langle B_3 T_3 M^{-1} \rangle$

As an example, consider $\langle B_3 T_3 M^{-1} \rangle$, which can be contracted in the three distinct ways shown in Figs. 5(a), 5(b), and 5(c). Each of these occurs with a symmetry factor of 3, since an x_{ij} from the B_3 can be contracted with any of the (cyclically equivalent) x_{ml}'s in T_3. Only the diagram in Fig. 5(a) can be traced with an unbroken line—the other diagrams contain closed loops. Thus, only one of the three sorts of contractions contributes to the partition function.

When all the contractions have been carried out, there remain eight nonvanishing graphs, which are shown in Fig. 6. The contractions associated with each diagram are shown in Table I. The alphabetic notation, common in the biochemical literature, shows the order in which sites pair with each other. For example, "*ABAB*" indicates that the first and third vertices (both denoted by "*A*") are paired, and that the vertex between them is linked to the fourth vertex (both denoted by "*B*").

Since the pseudoknots we consider contribute to order $1/N^2$, only one pseudoknot may be present at a time. This problem can be solved by noting that all the pseudoknot diagrams are one particle irreducible (1PI) (i.e., they cannot be disconnected by opening a single quark line), and can thus be re-summed by a Dyson equation. Define Σ_{mn} as the sum of all the amputated pseudoknot diagrams defined above [i.e., the sum of all $O(N^{-2})$ 1PI diagrams with their external G propagators removed]. Then the partition function Z_{mn} satisfies the usual Dyson equation

$$Z_{mn} = G_{mn} + \sum_{m>k>l>n} Z_{mk} \Sigma_{kl} G_{ln}. \tag{12}$$

Once the eight diagrams for Σ have been calculated, the full partition function (with any number of pseudoknots) can be calculated using the above recursion relations. Present knot-prediction algorithms use dynamic programming allow knots which have bonds drawn inside and outside of the disc, as

TABLE I. Contractions associated with the diagrams.

Figure	Contraction	Pseudoknot
(a)	$B_4 M^{-1}$	*ABAB*
(b)	$B_2 T_4 M^{-1}$	*ABACBC*
(c)	$B_3 T_3 M^{-1}$	*ABCABC*
(d)	$B_1 T_5 M^{-1}$	*ABCBCA*
(e)	$B_1 T_3 T_4 M^{-1}$	*ABCBDCDA*
(f)	$B_1 T_3 T_4 M^{-1}$	*ABCDBCDA*
(g)	$B_2 T_3^2 M^{-1}$	*ABCADBCD*
(h)	$B_1 T_3^3 M^{-1}$	*ABCDBECDEA*

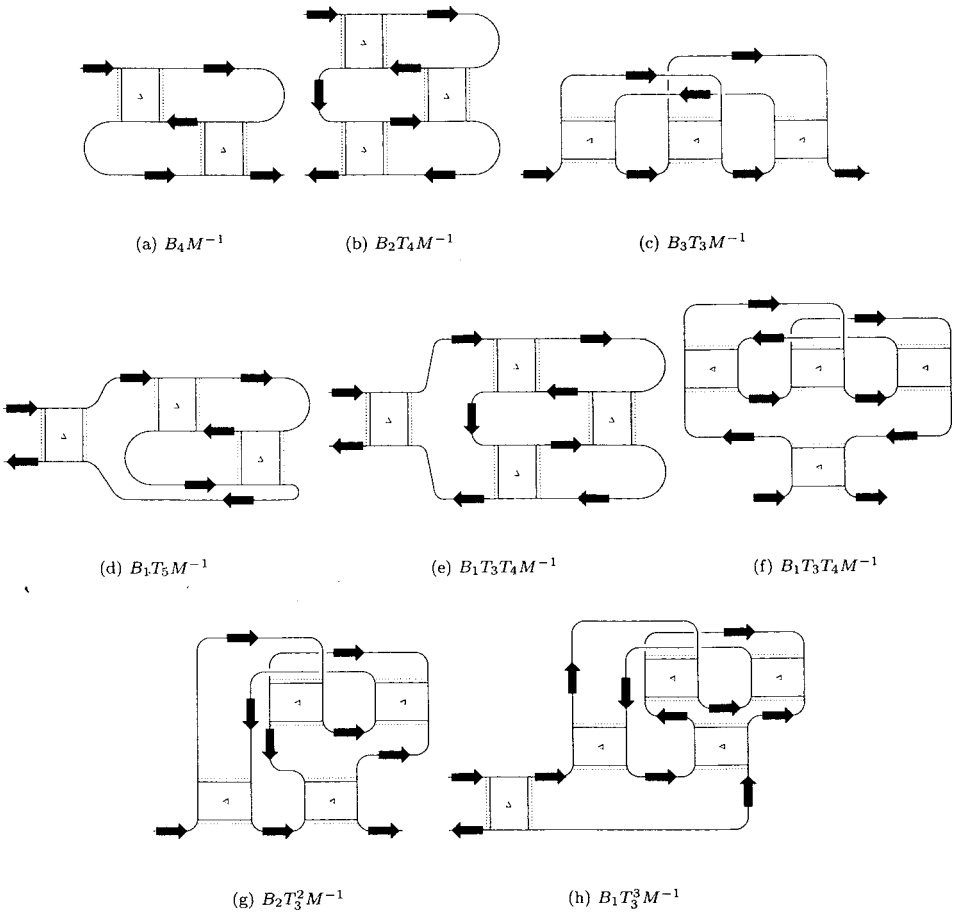

FIG. 6. Nonvanishing contractions.

long as they are no crossings [7]. This excludes certain topologies our algorithm provides for, such as *ABCABC* pseudoknots. On the other hand, these algorithms do provide for some topologies that we have excluded as $\mathcal{O}(N^{-4})$.

The method presented allows us to calculate the partition function in $O(L^6)$ time, so it can be used for folding, by backtracking to pick out the largest term in the partition function. The strategy for doing so is the following: (i) solve for the Hartree partition function (7), (ii) solve the Bethe-Salpeter recursion equation (8) to get $\Delta_{kl,mn}$, (iii) calculate the eight amputated diagrams of Fig. 5 making up the 1PI function Σ_{mn}, (iv) solve the Dyson equation (12) by recursion to obtain the full partition function with any number of pseudoknots, (v) and then backtrack to find the largest term in this partition function. Some numerical calculations are under way and we hope to present those results in a future paper, along with an explicit calculation for the order N^{-2} folding of a short ($L \simeq 10$) RNA [9–15].

[1] H. Orland and A. Zee, Nucl. Phys. B **620**, 456 (2002).
[2] R. Nussinov and A. B. Jacobson, Proc. Natl. Acad. Sci. U.S.A. **77**, 6309 (1980).
[3] S. Coleman, *Aspects of Symmetry* (Cambridge University Press, New York, 1985), Chap. 9.
[4] P. G. Higgs, Q. Rev. Biophys. **33**, 199 (2000).
[5] I. Tinoco, Jr. and C. Bustamante, J. Mol. Biol. **293**, 271 (1999).
[6] M. S. Waterman and T. F. Smith, Adv. Appl. Math. 7 (1986).
[7] E. Rivas and S. R. Eddy, J. Mol. Biol. **285**, 2053 (1999).
[8] G. 't Hooft, Nucl. Phys. B **72**, 461 (1974).
[9] M. Zuker, Science **244**, 48 (1989).
[10] D. K. Lubensky and D. R. Nelson, Phys. Rev. Lett. **85**, 1572 (2000).
[11] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, Monatsch. Chem. **125**, 167 (1994).
[12] A. Montanari and M. Mézard, Phys. Rev. Lett. **86**, 2178 (2001).
[13] R. Bundschuh and T. Hwa, Phys. Rev. Lett. **83**, 1479 (1999).
[14] J. S. McCaskill, Biopolymers **29**, 1105 (1990).
[15] H. Zhou, Y. Zhang and Z-C. Ou-Yang, Phys. Rev. Lett. **86**, 356 (2001).