ELSEVIER

Nuclear Physics B 620 [FS] (2002) 456–476

NUCLEAR PHYSICS B

www.elsevier.com/locate/npe

RNA folding and large N matrix theory

Henri Orland [a,b,*], A. Zee [a]

[a] Institute for Theoretical Physics, University of California, Santa Barbara, CA 93106, USA
[b] Service de Physique Théorique, CEA-Saclay, 91191 Gif-sur-Yvette Cedex, France

Received 19 June 2001; accepted 11 October 2001

Abstract

We formulate the RNA folding problem as an $N \times N$ matrix field theory. This matrix formalism allows us to give a systematic classification of the terms in the partition function according to their topological character. The theory is set up in such a way that the limit $N \to \infty$ yields the so-called secondary structure (Hartree theory). Tertiary structure and pseudo-knots are obtained by calculating the $1/N^2$ corrections to the partition function. We propose a generalization of the Hartree recursion relation to generate the tertiary structure. © 2002 Published by Elsevier Science B.V.

PACS: 87.10; 87.15; 11.15-q; 11.15.Pg

---

1. Introduction

Over the last decade, RNA has transformed itself from being a relatively minor player in the central dogma of Watson and Crick to being one of the central players in molecular biology. Indeed, it has been recently demonstrated that in addition to its "information carrier" role in protein synthesis, some types of RNA's, known as ribozymes, have an enzymatic activity which is crucial to the functioning of the cell [1]. As a consequence of this new prominent role of RNA, the search for the three-dimensional structure of RNA has become an important problem in biology. This view was expressed forcefully by Tinoco and Bustamante [2].

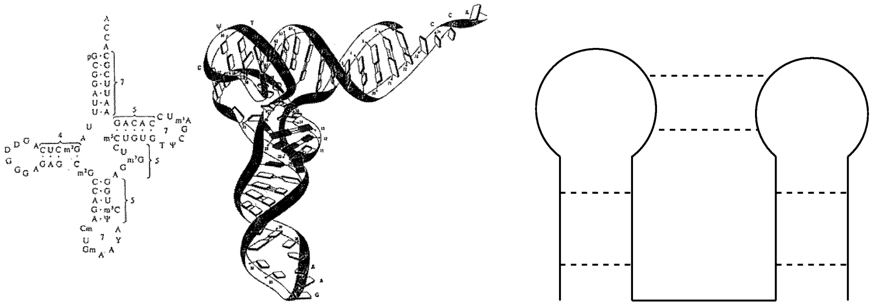
As this paper is addressed to theoretical physicists, we begin with a schematic review. A very thorough review on RNA folding can be found in Ref. [3].

RNA is a heteropolymer constructed out of a four-letter alphabet, C, G, A, and U (for the four bases or nucleotides cytosine, guanine, adenine, and uracil). The length of an RNA chain ranges typically from 76 for tRNA to a few thousand base pairs for mRNA. In solution, there is an attraction between C and G and between A and U, with energies

---

[*] Corresponding author.
E-mail address: horland@cea.fr (H. Orland).

PII: S0550-3213(01)00522-3



(a) (b)

Fig. 1. (a) Secondary (left) and tertiary (right) structure of a tRNA. (From I. Tinoco, with permission.) (b) A “kissing hairpin”.

$\varepsilon(C,G) \simeq -3$ kCal/mole and $\varepsilon(A,U) \simeq -2$ kCal/mole respectively. There is also a weaker attraction between G and U, with energy $\varepsilon(G,U) \simeq -1$ kCal/mole. Note the correspondence 300 K $\simeq$ 0.6 kCal/mole $\simeq 1/40$ eV.

Consider an RNA sequence $\{s\} = \{s_1, s_2, \ldots, s_L\}$ (where s_i takes on one of the four possible values C, G, A, and U). For example, we might be given the sequence $\{s\} = \{CCCGAAAUUCGUAG\ldots\}$. The attraction between the nucleotides folds the RNA heteropolymer into a 3-dimensional structure referred to as a shape. Biological functions depend largely on the shape assumed by a particular RNA. Thus, the map from sequence space to shape space is of great importance in molecular biology and has been much discussed in the biophysical literature. As mentioned above, this has been even more true since the discovery of the enzymatic activity of some RNA.

In the molecular biology of biopolymers, it is conventional to define three levels of structures. The primary structure is just the chemical sequence, or sequence of nucleotides. The secondary structure is the local short-range pairing of complementary bases, leading to segments of helices separated by loops and bulges (“clover-leaf” structure). Finally, the tertiary structure is the spatial arrangement of these secondary motifs, in which the loops and bulges themselves can partially pair, leading to the so-called pseudo-knots (see Fig. 1(a)).

An example of pseudo knot is the “kissing hairpin” (Fig. 1(b)).

In contrast to the problem of protein folding [4,5], RNA folding is hierarchical in that its secondary structure is much more stable than its tertiary structure, which can be treated as a perturbation [2]. Experimentally, the two levels of folding (secondary and tertiary) can be separated by varying the concentration of Mg^{++} ions [6]. In addition, the attractive force between nucleotides saturates. Once a given nucleotide C has paired with a nucleotide G, it cannot be paired with yet another G. In contrast, the attraction between amino acids do not saturate. Thus, the problem of RNA folding is considerably simpler than the problem of protein folding.

The determination of secondary structure has reached a very high level of sophistication based on dynamic programming algorithms [7–9].

The problem of RNA folding is clearly topological in flavor and is thus not easily amenable to dynamic programming methods, although some algorithm has been proposed recently [10]. On the other hand, we know from the field theoretic literature that topological considerations also play an important role in such subjects as matrix theory or M-theory. In this paper, we propose that matrix theory may be useful to the problem of RNA folding. We develop a matrix theoretic representation of the topological aspect of RNA folding.

In Section 1, we formulate the RNA folding problem more precisely. In Section 2, we show how it can be formulated as an $N \times N$ matrix field theory. In Section 3, we show that the N dependence of the field theory can be made explicit in the functional integral formulation of the problem. As a result, the natural way to compute the $1/N$ expansion is through a steepest descent method which is described in Section 4. As this expansion is very complicated to perform at higher order, we resort in Section 5 to recursion relations which allow us to approximately incorporate the higher order powers in $1/N$.

For a simple introduction to this work, one can go for instance to the website http://online.itp.ucsb.edu/online/infobio01/zee/.

2. RNA folding

Given an RNA sequence $\{s\} = \{s_1, s_2, \ldots, s_L\}$ of L bases, let us write down the partition function $\mathcal{Z}$ at temperature $1/\beta$. We will proceed in steps.

First, construct the matrix

$$V_{ij} = e^{-\beta|\varepsilon(s_i,s_j)|v(|\vec{r}_i-\vec{r}_j|)}\theta(|i-j|>4), \quad i \neq j; \qquad V_{ii}=0, \tag{1}$$

where $\varepsilon(a,b)$ denotes the 4 by 4 real symmetric matrix giving the attractive energy between nucleotides, $\varepsilon(A,U)$ etc. We set the diagonal elements V_{ii} to 0 to indicate the fact that a nucleotide does not attract itself. The Heaviside function $\theta(|i-j|>4)$ incorporates the fact that the RNA molecule is not infinitely flexible and we cannot pair nucleotides separated by less than 4 sites. The attractive potential can be taken to be $v(r) = -w\theta(R-r)$ with w and R the strength and range of the attraction respectively.

Now construct

$$Z = 1 + \sum_{\langle ij\rangle} V_{ij} + \sum_{\langle ijkl\rangle} V_{ij}V_{kl} + \cdots + \sum_{\langle ijkl\rangle} V_{ik}V_{jl} + \cdots, \tag{2}$$

where $\langle ij\rangle$ denotes all pairs with $j > i$, $\langle ijkl\rangle$ all quadruplets with $l > k > j > i$, and so on. Then the partition function is given by

$$\mathcal{Z} = \int \prod_{k=1}^{L} d^3\vec{r}_k \prod_{i=1}^{L-1} f\big(|\vec{r}_{i+1}-\vec{r}_i|\big)Z. \tag{3}$$

The function $f(r)$ can be taken to be, for example, $\delta(r-l)$ for a model in which the nucleotides are connected along the RNA heteropolymer by rigid rods of length l, or $e^{-(r-l)^2/6\sigma^2}$ for a model with elastic rods. Note that the saturation of the hydrogen bond has been incorporated by the requirement $l > k > j > i$, and so on. Once the nucleotide at i has interacted with the nucleotide at j it cannot interact with the nucleotide at k. Note that in (2), only the enthalpy and combinatorics of pairings are included. The integration over the atomic coordinates in (3) accounts for the actual topological feasibility of a given pairing and also for the entropic factor associated with loop formation.

Biologists are interested in the folded configuration essentially at room temperature. Since room temperature is substantially less than the melting temperature (of order 80 °C, in other words, the characteristic energy scale of the problem), we want to determine the ground state configuration of the RNA heteropolymer. In other words, once we have obtained Z we would like to extract the term in Z that dominates as $\beta\varepsilon$ tends to infinity in (1).

We have given a simplified quantitative framework for the RNA folding problem. From a chemical point of view, it would be appropriate to include also the stacking energies of couples of complementary base pairs, instead of energies of single pairs of bases. However, in the following, we will stick with the latter. We will also concentrate on the evaluation of the "pairing" partition function (2). We expect that the various effects we have ignored, such as stacking, etc., can be added later as "bells and whistles" to our approach. The stacking energies for instance can be taken into account by utilizing a 16×16 interaction matrix between pairs of bases instead of the 4×4 matrix $\varepsilon(s_i, s_j)$ we use here.

3. Matrix theory

What is the connection with matrix theory?

Consider pulling apart the folded RNA structure given in Fig. 2(a).

We obtain the structure of Fig. 2(b) which to physicists are reminiscent of Feynman diagrams in a variety of subjects: matrix theory, quantum chromodynamics, and so on.

For the sake of definiteness, let us borrow the terminology of quantum chromodynamics. The dotted lines are known as gluon propagators, and the solid line as a quark propagator. The secondary structure corresponds to diagrams in which the gluon lines do not cross over each other, while the tertiary structure corresponds to diagrams in which the gluon lines do cross.

The crucial observation, originally made by 't Hooft [11], is that there is a systematic relation between the topology of a graph and its corresponding power of $1/N^2$. For

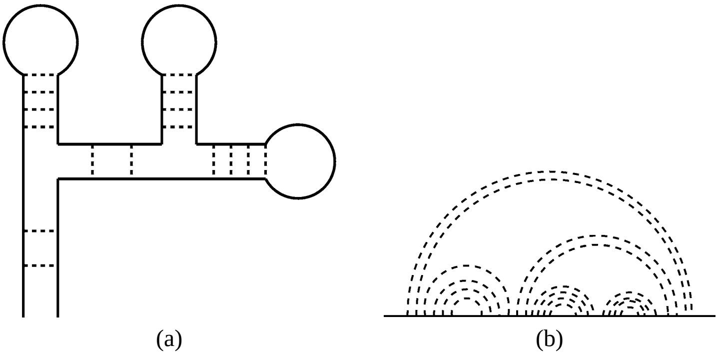


Fig. 2. (a) Representation of the secondary structure of an RNA. (b) Representation of the same RNA stretched.

instance, planar diagrams are of order $1/N^0$, and diagrams in which gluon lines cross are of higher order. We merely have to go to the large N expansion, and the diagrams are classified by powers of $1/N^2$. Note that a somewhat similar formulation in terms of matrix theory has been used for the meander problem [12].

Consider the quantity

$$Z(1,L) = \frac{1}{A(L)} \int \prod_{k=1}^{L} d\varphi_k \, e^{-\frac{N}{2}\sum_{ij}(V^{-1})_{ij}\,\mathrm{tr}(\varphi_i\varphi_j)} \, \frac{1}{N}\,\mathrm{tr}\prod_{l=1}^{L}(1+\varphi_l). \tag{4}$$

Here φ_i $(i = 1,\ldots,L)$ denote L independent N by N Hermitian matrices and $\Pi_l(1+\varphi_l)$ represents the ordered matrix product $(1+\varphi_1)(1+\varphi_2)\cdots(1+\varphi_L)$. All matrix products will be understood as ordered in this paper. The normalization factor $A(L)$ is defined by

$$A(L) = \int \prod_{k=1}^{L} d\varphi_k \, e^{-\frac{N}{2}\sum_{ij}(V^{-1})_{ij}\,\mathrm{tr}(\varphi_i\varphi_j)}. \tag{5}$$

Let us refer to the row and column indices a and b of the matrices $(\varphi_i)_a^b$ as color indices, with $a, b = 1, 2, \ldots, N$. The matrix integral (4) defines a matrix theory with L matrices. We can either think of it as a Gaussian theory with a complicated observable $\frac{1}{N}\,\mathrm{tr}\prod_l(1+\varphi_l)$, or alternatively, by raising $\frac{1}{N}\,\mathrm{tr}\prod_l(1+\varphi_l) = e^{\log[\frac{1}{N}\,\mathrm{tr}\prod_l(1+\varphi_l)]}$ into the exponent, as a complicated matrix theory with the action $(\frac{N}{2}\sum_{ij}(V^{-1})_{ij}\,\mathrm{tr}(\varphi_i\varphi_j) - \log[\frac{1}{N}\,\mathrm{tr}\prod_l(1+\varphi_l)])$. Another trivial remark is that we can effectively remove $\frac{1}{N}\,\mathrm{tr}$ from (4).

The important remark is that the matrix theory [13] defined by (4) has the same topological structure as ’t Hooft’s large N quantum chromodynamics. There are L types of gluons, and the gluon propagators are given by $\frac{1}{N}V_{ij}$. As in large N quantum chromodynamics, each gluon propagator is associated with a factor of $\frac{1}{N}$ and each color loop is associated with a factor of N. The reader familiar with matrix theory or large N quantum chromodynamics can see immediately that the Gaussian matrix integral (4) evaluates precisely to the infinite series

$$Z(1,L) = 1 + \sum_{\langle ij\rangle} V_{ij} + \sum_{\langle ijkl\rangle} V_{ij}V_{kl} + \cdots + \frac{1}{N^2}\sum_{\langle ijkl\rangle} V_{ik}V_{jl} + \cdots \tag{6}$$

Some “typical” terms in this series correspond to the diagrams in Fig. 3.

This differs from (2) only in that the terms with different topological character are now classified by inverse powers of $\frac{1}{N^2}$. Thus, the use of the large N expansion allows us to separate out the tertiary structure, represented in (6) for example by the term $\frac{1}{N^2}\sum_{\langle ijkl\rangle} V_{ik}V_{jl}$, from the secondary structure.

Note that the ordered product $\prod_l(1+\varphi_l)$ ensures that the diagonal elements V_{ii} of the matrix V do not appear in $Z(1,L)$. We have nevertheless already set V_{ii} to 0.

Fig. 3. Graphical representation of a few terms of the partition function.

The program proposed in this paper is thus to evaluate $Z(1, L)$ with V an arbitrary matrix. Once $Z(1, L)$ is known we can then insert it into (3) to evaluate $\mathcal{Z}$. The parameter $\frac{1}{N}$ serves as a convenient marker to distinguish the tertiary structure from the secondary structure. What we offer here is a systematic way of generating refinements to the calculation of Z, and hence the free energy F, to any desired accuracy in a well controlled approximation.

Since in $Z(1, L)$ the quantities 1 and L represent arbitrary labels we can just as well define

$$Z(m,n) = \frac{1}{A(m,n)} \int \prod_{k=1}^{L} d\varphi_k \, e^{-\frac{N}{2} \sum_{i,j=m}^{n} (V^{-1})_{ij} \operatorname{tr}(\varphi_i \varphi_j)} \frac{1}{N} \operatorname{tr} \prod_{l=m}^{n} (1 + \varphi_l), \tag{7}$$

where again the normalization is given by

$$A(m,n) = \int \prod_{k=m}^{n} d\varphi_k \, e^{-\frac{N}{2} \sum_{ij=m}^{n} (V^{-1})_{i,j} \operatorname{tr}(\varphi_i \varphi_j)}.$$

As we shall see in the following, we will construct recursion relations to evaluate (7) approximately. These recursion relations can be easily programmed to calculate the free energy of the RNA chain.

4. Large *N*

In the matrix representation (4) N appears implicitly in the size of the matrices φ_i. In order to study the large N limit, we need to extract the N dependence explicitly, for which we have developed the following method. Define $G_{l'l}$ by $G_{l'l} = \prod_{i=l}^{l'-1} (1 + \varphi_i)$ for $l' - 1 \geqslant l$, $G_{l-k,l} = 0$ for all $k > 0$, and $G_{ll} = 1$. Then $G_{l'l}$ satisfies the equation

$$G_{l'l} - (1 + \varphi_{l'-1}) G_{l'-1,l} = \delta_{l'l}. \tag{8}$$

Thus, if we define $M_{l'l} = \delta_{l'l} - (1 + \varphi_{l'-1}) \delta_{l'-1,l}$ then we see that $G_{ll'}$ is the inverse of the matrix $M_{ll'}$ and thus

$$\begin{aligned} Z(1, L) &= \frac{1}{A(L)} \int \prod_k d\varphi_k \, e^{-N \frac{1}{2} \sum_{ij} (V^{-1})_{ij} \operatorname{tr}(\varphi_i \varphi_j)} M^{-1}_{L+1,1} \\ &= \frac{1}{A(L)} \int \prod_k d\varphi_k \, e^{-N \frac{1}{2} \sum_{ij} (V^{-1})_{ij} \operatorname{tr}(\varphi_i \varphi_j)} \\ &\quad \times \int \prod_l d\psi_l^* \, d\psi_l \, e^{-\psi_l^* M_{ll'} \psi_{l'}} \psi_{L+1} \psi_1^*. \end{aligned} \tag{9}$$

We have used the standard representation of the inverse of a matrix by an integral over Grassmanian fermionic variables ψ_l and ψ_l^*. Note the felicitous fact that $\det M = \int d\psi^* \, d\psi \, e^{-\psi^* M \psi} = 1$ which allows us to write (9) without a denominator.

To compactify this representation of Z further we introduce $M(h)_{ij} = M_{ij} + h\delta_{i,1}\delta_{j,L+1}$ and write

$$Z(1,L) = \frac{1}{N}\frac{\partial}{\partial h}\frac{1}{A(L)}\int\prod_k d\varphi_k\, e^{-N\frac{1}{2}\sum_{ij}(V^{-1})_{ij}\operatorname{tr}(\varphi_i\varphi_j)}\int d\psi^*\, d\psi\, e^{-\psi^* M(h)\psi}. \tag{10}$$

Henceforth, it is understood that after differentiation with respect to h we set h to 0.

We can now perform the Gaussian integration over φ_k, thus obtaining

$$Z(1,L) = \frac{1}{N}\frac{\partial}{\partial h}\int d\psi^*\, d\psi\, e^{-S_0(\psi^*,\psi)-S_1(\psi^*,\psi)} \tag{11}$$

with the free fermion action

$$S_0(\psi^*,\psi) = \sum_j\bigl(\psi_j^* - \psi_{j+1}^*\bigr)\psi_j + h\psi_1^*\psi_{L+1} \tag{12}$$

and the interacting fermion action

$$S_1(\psi^*,\psi) = -\frac{1}{2N}\sum_{j,j'}\sum_{a,b}\psi_{a,j+1}^*\psi_j^b V_{jj'}\psi_{b,j'+1}^*\psi_{j'}^a. \tag{13}$$

Note that in (13) we have displayed the color indices a and b explicitly.

We next rewrite

$$S_1(\psi^*,\psi) = +\frac{1}{2N}\sum_{j,j'}K_{jj'}K_{j'j} = \frac{1}{2N}\operatorname{tr}K^2 \tag{14}$$

in terms of the color singlet variable

$$K_{jj'} = \sum_a (V_{jj'})^{\frac{1}{2}}\psi_{a,j+1}^*\psi_{j'}^a. \tag{15}$$

Now use the Gaussian representation

$$e^{-\frac{1}{2N}\operatorname{tr}K^2} = \frac{1}{C}\int dA\, e^{-\frac{N}{2}\operatorname{tr}A^2 + i\operatorname{tr}AK} \tag{16}$$

with the normalization factor $C = \int dA\, e^{-\frac{N}{2}\operatorname{tr}A^2}$. Note that even though K is complex we can take A to be Hermitean. (Equivalently, the anti-Hermitean part of A drops out.) Putting it together we obtain

$$Z(1,L) = \frac{1}{N}\frac{\partial}{\partial h}\frac{1}{C}\int dA\, e^{-\frac{N}{2}\operatorname{tr}A^2}\int d\psi^*\, d\psi\, e^{-\sum_{ij}\sum_a \psi_{a,i}^* M_{ij}\psi_j^a} \tag{17}$$

where

$$M_{ij} = \delta_{ij} - \delta_{i,j+1} + h\delta_{i,1}\delta_{j,L+1} + i(V_{i-1,j})^{\frac{1}{2}}A_{i-1,j} \tag{18}$$

or in matrix form

$$M_L = \begin{pmatrix} 1 & 0 & 0 & \cdot & \cdot & 0 & h \\ -1 & 1+a_{12} & a_{13} & \cdot & \cdot & a_{1L} & 0 \\ a_{12}^* & -1 & 1+a_{23} & \cdot & \cdot & a_{2L} & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & -1 & 1+a_{L-1L} & 0 \\ a_{1L}^* & a_{2L}^* & \cdot & a_{L-2L}^* & a_{L-1L}^* & -1 & 1 \end{pmatrix} \tag{19}$$

where we have used the convenient notation

$$\begin{aligned} i\sqrt{V_{ij}}\,A_{ij} &= a_{ij} \quad \text{for } i < j, \\ i\sqrt{V_{ij}}\,A_{ij} &= a_{ji}^* \quad \text{for } j < i. \end{aligned} \tag{20}$$

The point of these manipulations is that in (17) we have now isolated the color index a so that the integral over ψ^* and ψ factors into N copies of the same integral, thus giving

$$\begin{aligned} Z(1,L) &= \frac{1}{N}\frac{\partial}{\partial h}\frac{1}{C}\int dA\, e^{-\frac{N}{2}\operatorname{tr}A^2}\bigl(\det M(A)\bigr)^N \\ &= \frac{1}{N}\frac{\partial}{\partial h}\frac{1}{C}\int dA\, e^{-\frac{N}{2}\operatorname{tr}A^2+N\operatorname{tr}\log M(A)}. \end{aligned} \tag{21}$$

At this point, we can differentiate with respect to h and set h to 0, obtaining the alternative form

$$Z(1,L) = \frac{1}{C}\int dA\, e^{-\frac{N}{2}\operatorname{tr}A^2+N\operatorname{tr}\log M(A)} M^{-1}(A)_{L+1,1}. \tag{22}$$

In this expression,

$$M_{ij} = \delta_{ij} - \delta_{i,j+1} + i(V_{i-1,j})^{\frac{1}{2}} A_{i-1,j}. \tag{23}$$

Let us introduce the action

$$S(A) = \frac{1}{2}\operatorname{tr}A^2 - \operatorname{tr}\log M(A), \tag{24}$$

and define the average of an “observable” O by

$$\langle O\rangle = \frac{1}{C}\int dA\, e^{-NS(A)} O. \tag{25}$$

(Note the nonstandard normalization used here.) Then, our result can be summarized elegantly as

$$Z(1,L) = \bigl\langle M^{-1}(A)_{L+1,1}\bigr\rangle. \tag{26}$$

At this point, as remarked earlier, we note that the quantity $Z(1,L)$ can obviously be generalized to $Z(i,j)$: after all, the site labels 1 and L are arbitrary. Then we have the appealing result that

$$Z(i,j) = \bigl\langle M^{-1}(A)_{j+1,i}\bigr\rangle \quad \text{for } j > i. \tag{27}$$

It is also useful to introduce the free action

$$S_0(A) = \frac{1}{2}\operatorname{tr}A^2 \tag{28}$$

and to define

$$\langle O\rangle_0 = \frac{1}{C}\int dA\, e^{-NS_0(A)}O. \tag{29}$$

Then we can also write our result as

$$Z(1,L) = \frac{1}{N}\frac{\partial}{\partial h}\big\langle(\det M(A))^N\big\rangle_0. \tag{30}$$

Remarkably, it turns out that we will need both the representations (26) and (30) later in a single calculation.

Incidentally, our formulation of the RNA folding problem can be immediately adapted to the marriage problem (or bipartite matching problem) [14–17], one of the classic problems in combinatorial optimization. We will mention only the simplest version here. Label L (with L even) individuals by the index $i = 1, \ldots, L$ where the individual is male for i odd and female for i even. Define a matrix $V_{ij} = \frac{1}{2}(1 - (-1)^{i+j})e^{-\beta E_{ij}}$ where E_{ij} represents the energy cost of a marriage between i and j and its negative provides a measure of happiness. Referring back to (2) we see that we want to extract in $Z(1,L)$ all the terms with $L/2$ powers of V, for example the term $V_{14}V_{38}\cdots V_{L-1,2} = e^{-\beta E_{\text{Total}}}$ with $E_{\text{Total}} = E_{14} + E_{38} + \cdots + E_{L-1,2}$. Since we now want to include possible crossings in the Feynman diagram language we can set the number of colors N to 1. Thus, from (21), we have immediately $Z(1,L) = \frac{\partial}{\partial h}\frac{1}{C}\int dA\, e^{-\frac{1}{2}\operatorname{tr}A^2}\det M(A)$. Referring to (19) we see that the differentiation with respect to h and setting h to 0 amounts to replacing the $L+1$ by $L+1$ matrix $M(A)$ by the L by L matrix obtained by deleting the first row and last column. Furthermore, since we want the terms with $L/2$ powers of V, that is, with L powers of $V^{\frac{1}{2}}$, we can set the 1's and -1's in this matrix to 0. Denoting the resulting matrix by $\mathcal{M}(A)$, we obtain the following representation for the marriage problem

$$Z_m(L) = \frac{1}{C}\int dA\, e^{-\frac{1}{2}\operatorname{tr}A^2}\det\mathcal{M}(A). \tag{31}$$

Clearly, the representation given here can be generalized in a number of directions, for example, by including individuals who remain single.

It is easy to see how this representation works: the Gaussian integration insures that in $\det\mathcal{M}(A)$ only the appropriate terms are picked out.

5. Steepest descent

The fact that we have been able to display explicitly the N dependence is crucial and allows us in principle to carry the $1/N$ expansion to any order. The standard strategy to evaluate integrals such as (22) is of course to use the method of steepest descent [18,19].

To leading order the steepest descent approximation is easy enough to carry out. The stationary point is determined by $\frac{\delta S(A)}{\delta A} = 0$, that is

$$A_{lk} = i(V_{lk})^{\frac{1}{2}}G_{l-1,k+1} \tag{32}$$

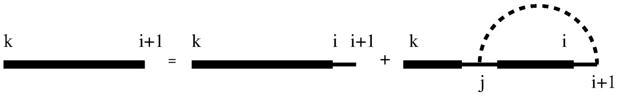


Fig. 4. Graphical representation of the Hartree recursion relation. The thick line represents the propagator G.

where we find it useful to define

$$G_{ij} = \left(M^{-1}\right)_{i+1,j}. \tag{33}$$

Notice that with this definition G_{ij} is defined for i from 0 to $L-1$ and for j from 2 to $L+1$. The identity $\sum_j M_{ij}(M^{-1})_{jk} = \delta_{ik}$ can now be written as

$$G_{i+1,k} - G_{ik} - \sum_j V_{i+1,j} G_{i,j+1} G_{j-1,k} = \delta_{i+2,k}. \tag{34}$$

Referring to (22) and (27) we see that to leading order in steepest descent, $Z(i,j)$ is just $M^{-1}(A)_{j+1,i} = G_{ji}$ evaluated at the stationary point.

Eq. (34) has already been written down in the literature [7–9,20–25] and is known as the "Hartree approximation". It has the obvious interpretation (see Fig. 4) that to lowest order the additive effect of including one extra nucleotide labelled by $L+1$ to the RNA heteropolymer can be described by pairing that nucleotide to the nucleotide labeled by j, which separates the heteropolymer into two segments, one from 1 to j and the other from $j+1$ to $L+1$. We then sum over all possible j of course.

In principle, steepest descent gives a systematic expansion of $Z(1,L)$ to any desired power of $\frac{1}{N}$ by expanding the exponent and the observable around the saddle-point. In the present context, this implies that the full three dimensional structure of the RNA can be obtained by expanding around the secondary structure. In particular, the higher order terms do not disrupt the secondary structure, but merely add new interactions, in addition to the existing secondary pairing. This is in marked contrast with protein folding, where it is known that there is a strong correlation between secondary and tertiary structure.

In practice, however, it proves to be quite tedious to calculate the $\frac{1}{N^2}$ terms explicitly. In the integral in (22) we are now to replace A_{ij} by $A_{ij} + x_{ij}/\sqrt{N}$ where A_{ij} is determined by (32) and (34). A straightforward calculation shows that

$$\begin{aligned} Z(1,L) = \int dx \exp\left(-\frac{1}{2}\operatorname{tr} x^2 - \frac{1}{2}\operatorname{tr}\left(M^{-1}c\right)^2 - \sum_{p=3}^{\infty} \frac{(-1)^p}{pN^{p/2-1}} \operatorname{tr}\left(M^{-1}c\right)^p\right) \\ \times \left\{\left(1 + \sum_{p=1}^{\infty} \frac{(-1)^p}{N^{p/2}}\left(M^{-1}c\right)^p\right) M^{-1}\right\}_{L+1,1} \end{aligned} \tag{35}$$

where M^{-1} is related to G through Eq. (33), and $c_{ll'} = \sqrt{V_{l-1,l'}}\, x_{l-1,l'}$. The systematic corrections to Z are obtained by expanding (35) in powers of $1/N^{1/2}$. By symmetry, no half-integer powers of N remain in the expansion of Z.

The first thing to evaluate is the propagator of the fluctuation fields x_{ij}. This is just the inverse of the kernel of the quadratic form appearing in the exponent of (35). This

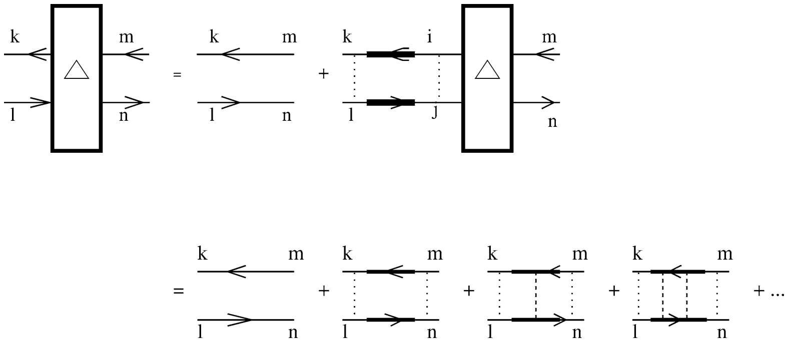


Fig. 5. Graphical representation of the Bethe–Salpeter recursion relation. The dotted lines represent factors $\sqrt{V_{ij}}$ while the dashed lines represent factors V_{ij}. The solid thick lines represent Hartree propagators G. The Hartree propagators being directed, the arrows denote the direction of increasing spatial index.

propagator $\Delta_{ij,kl}$ is in fact a scattering amplitude and satisfies a form of the Bethe–Salpeter equation [26]

$$\Delta_{kl,mn} = \delta_{km}\delta_{nl} + \sum_{ij} V_{kl}^{1/2} V_{ij}^{1/2} G_{k-1,i+1} G_{j-1,l+1} \Delta_{ij,mn} \tag{36}$$

where G are the Hartree propagators (34). In Fig. 5, we show a graphical representation of this recursion equation, as well as the series of graphs it resums. It is clear that this equation resums all the possible ladder (or rainbow) diagrams to this order.

This equation is to be solved for the particular sequence studied. The scattering amplitude Δ defines the contractions of the x fields, and thus its knowledge allows us in principle to calculate (35) to any order. Note that as usual in field theory, only contractions which are linked to the operator that we calculate are to be included. (This reduces considerably the number of contraction.)

A fairly simple calculation allows us to show that the $1/N$ correction vanishes identically (see Appendix A). This result appears true by drawing a few graphs, but this gives an algebraic proof.

It is easy to see that we have to expand (35) to $\mathrm{O}(x^6)$ in order to calculate the free energy to order $\frac{1}{N^2}$. The calculation, although cumbersome, is straightforward. The free energy reads

$$\begin{aligned} Z(1,L) = G_{1L} + \frac{1}{N^2}\Bigg\langle \Bigg\{ \Bigg(& -\frac{1}{5} B_1 T_5 + \frac{1}{12} B_1 T_3 T_4 - \frac{1}{162} B_1 T_3^3 \\ & - \frac{1}{4} B_2 T_4 + \frac{1}{18} B_2 T_3^2 - \frac{1}{3} B_3 T_3 + B_4 \Bigg) M^{-1} \Bigg\}_{L+1,1} \Bigg\rangle, \end{aligned} \tag{37}$$

where we have used the notation

$$D_{mn} = \sum_{m'} M_{mm'}^{-1} \sqrt{V_{m'-1,n}}\, x_{m'-1,n},$$

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

$$T_p = \operatorname{tr} B_p. \tag{38}$$

In (37), the bracket means that the Wick theorem should be applied to contract the fields $x_{ll'}$ which appear in this expression, their contraction being given by the kernel Δ.

The calculation of the correction to the free energy is possible numerically for not too long RNA sequences. Work in this direction is in progress.

Because of the complexity of the (exact) order $1/N^2$ obtained in this approach, we found it simpler to generalize the Hartree recursion equation to incorporate some residual interactions between the loops and bulges.

6. Recursion approach

Two approaches can be used to derive recursion relations for the partition functions. One is detailed in the following, whereas the other one is described in Appendix B.

A possible approach is to take the expression in (30)

$$Z(1, L) = \frac{1}{N}\frac{\partial}{\partial h}\big\langle \big(\det M(A)\big)^N \big\rangle_0 \tag{39}$$

and try to relate $Z(1, L+1)$ to $Z(1, L)$. In other words, we would like to relate $\langle (\det M_{L+1}(A))^N \rangle$ to $\langle (\det M_L(A))^N \rangle$ where the subscript on M keeps track of the different matrices in the discussion. Note that M_L is an $L+1$ by $L+1$ matrix. Explicitly, as noted before, the $L+2$ by $L+2$ matrix M_{L+1} has the form

$$M_{L+1} = \begin{pmatrix} 1 & 0 & 0 & \cdot & \cdot & 0 & h \\ -1 & 1+a_{12} & a_{23} & \cdot & \cdot & b_1 & 0 \\ a_{12}^* & -1 & \cdot & \cdot & \cdot & b_2 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & -1 & 1+b_L & 0 \\ b_1^* & b_2^* & \cdot & b_{L-1}^* & b_L^* & -1 & 1 \end{pmatrix} \tag{40}$$

where for convenience we have denoted

$$\begin{aligned} i\sqrt{V_{ij}}\,A_{ij} &= a_{ij} \quad && \text{for } i < j \leqslant L, \\ i\sqrt{V_{i,L+1}}\,A_{i,L+1} &= b_i \quad && \text{for } i \leqslant L, \\ i\sqrt{V_{ij}}\,A_{ij} &= a_{ji}^* \quad && \text{for } j < i \leqslant L, \\ i\sqrt{V_{L+1,j}}\,A_{L+1,j} &= b_j^* \quad && \text{for } j \leqslant L. \end{aligned} \tag{41}$$

Our strategy is to first perform the Gaussian integration over the b_j's in evaluating $\langle (\det M_{L+1}(A))^N \rangle$, keeping in mind that we need the terms of order h. This method of integrating out a row and a column has also been used in random matrix theory [27].

We briefly outline the procedure. Write $M_{L+1} = M_{L+1}(b = 0) + B$ where B is the matrix extracted from (40) upon keeping only the entries which depend on the b's and b^*'s. Expand $(\det M_{L+1}(A))^N$ in powers of B and then perform the Gaussian average over the

b's and b^*'s, using $\langle b_i b_j^*\rangle = \frac{1}{N}\delta_{ij}V_{j,L+1}$. After some arithmetic, we obtain

$$Z(1,L+1) = Z(1,L) + \sum_{j=1}^{L} V_{j,L+1}\left\langle (\det M)^N \left[\left(\frac{\partial}{\partial h}M^{-1}_{j,L+2}\right)M^{-1}_{L+1,j+1} - \frac{1}{N}\left(\frac{\partial}{\partial h}M^{-1}_{L+1,L+2}\right)M^{-1}_{j,j+1}\right]\right\rangle_0. \tag{42}$$

We have suppressed the subscript $L+1$ on the matrix M on the right hand side. It is understood that this expression is to be evaluated at $h=0$. Noting that the matrix $\frac{\partial M}{\partial h}$ is particularly simple and that $(M^{-1})_{L+2,L+2} = 1$, we find that

$$Z(1,L+1) = Z(1,L) + \sum_{j=1}^{L} V_{j,L+1}\left\langle M^{-1}_{L+1,j+1}M^{-1}_{j,1} - \frac{1}{N}M^{-1}_{L+1,1}M^{-1}_{j,j+1}\right\rangle. \tag{43}$$

Using the definition of the connected expectation value

$$\langle AB\rangle = \langle A\rangle\langle B\rangle + \langle AB\rangle_C$$

we note, as is well-known, that the connected part is of order $1/N^2$ [11] and we can thus write

$$\begin{aligned} Z(1,L+1) = Z(1,L) &+ \sum_{j=1}^{L} V_{j,L+1}\langle M^{-1}_{L+1,j+1}\rangle\langle M^{-1}_{j,1}\rangle \\ &+ \sum_{j=1}^{L} V_{j,L+1}\langle M^{-1}_{L+1,j+1}M^{-1}_{j,1}\rangle_C \\ &- \frac{1}{N}\sum_{j=1}^{L} V_{j,L+1}\langle M^{-1}_{L+1,1}M^{-1}_{j,j+1}\rangle_C. \end{aligned} \tag{44}$$

Recalling (27) we recognize that the quantities $\langle M^{-1}_{L+1,j+1}\rangle$ and $\langle M^{-1}_{j,1}\rangle$ appearing in the second term on the right hand side of (44) are nothing but $Z(j+1,L+1)$ and $Z(1,j)$ respectively. Thus, if we keep only the first two terms on the right hand side of (44) we obtain the closed recursion relation

$$Z(1,L+1) \simeq Z(1,L) + \sum_{j=1}^{L} V_{j,L+1} Z(1,j) Z(j+1,L+1). \tag{45}$$

This is precisely the recursion relation in the Hartree approximation (45) mentioned earlier.

As announced in the introduction, the formulation given here offers a systematic way to go beyond the Hartree approximation. We expect the third and fourth term on the right hand side of (44), when evaluated to leading order in $\frac{1}{N}$ to give the corrections of order $\frac{1}{N^2}$. It is intriguing then that the superficially similar objects $\langle M^{-1}_{L+1,j+1}M^{-1}_{j,1}\rangle_C$ and $\langle M^{-1}_{L+1,1}M^{-1}_{j,j+1}\rangle_C$ must be of order $\frac{1}{N^2}$ and order $\frac{1}{N}$ respectively. We note however that a "backward-propagating object" which we define as M^{-1}_{jk} with $k > j$ makes its first

Fig. 6. A few graphs corresponding to the $1/N^2$ term.

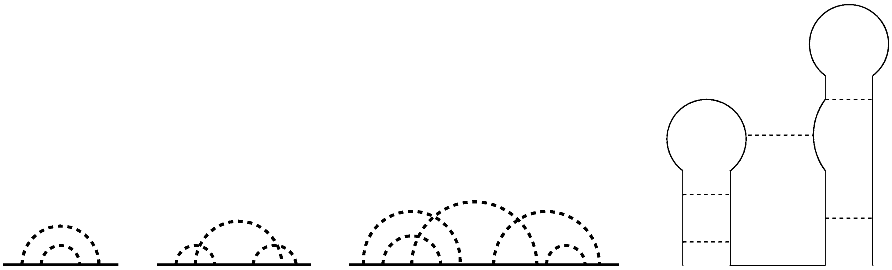


Fig. 7. (a) A few one particle irreducible graphs. (b) The third graph of Fig. 7(a).

appearance in $\langle M^{-1}_{L+1,1} M^{-1}_{j,j+1}\rangle_C$. All other terms in (44) involve only forward-propagating objects.

We can of course calculate (44) explicitly for small L in order to check our formulation and the order of the various terms in $\frac{1}{N}$. The result for $L = 5$ is shown graphically in Fig. 6. While the recursion relation (44) has an appealing structure, we are not able to evaluate the two objects $\langle M^{-1}_{L+1,j+1} M^{-1}_{j,1}\rangle_C$ and $\langle M^{-1}_{L+1,1} M^{-1}_{j,j+1}\rangle_C$ and express them in a simple form. Neither should we be able to do that. Our experience in field theory, for example the Dyson–Schwinger equation in quantum electrodynamics, indicates that recursion relations generically do not close: new objects appear in the right hand side. There is no reason why $\langle M^{-1}_{L+1,j+1} M^{-1}_{j,1}\rangle_C$ should be expressible in terms of $\langle M^{-1}_{ik}\rangle$. New objects, corresponding to vertex functions in field theory, must appear.

6.1. Recursion relation

At this point, all we can do is to inspect the set of Feynman diagrams to obtain an approximate recursion relation for $Z(i, j)$. We propose the following approximate recursion relation. We propose the following recursion relation. Given $Z(i, j)$ for all i and j satisfying $j - i \leqslant L - 1$, we obtain $Z(i, j)$ for all i and j satisfying $j - i \leqslant L$ as follows.

First, define $Z^{1\mathrm{PI}}(i, j)$ as the one-particle irreducible (1PI) part of $Z(i, j)$, that is the sum of all those diagrams in $Z(i, j)$ that do not fall apart into two separate pieces when a quark propagator is cut. Some examples are shown in Fig. 7(a).

In Fig. 7(b), we show a different representation of the third graph of Fig. 7(a).

The concept of, and the necessity of introducing, one-particle irreducibility is of course the same here as in field theory such as quantum electrodynamics.

Second, define the vertex function Γ^j_{mn} for $n > j > m$ by

$$\Gamma^j_{mn} = \left[1 - \sum_{n\leqslant k\neq j\leqslant m} V_{jk}\frac{\partial}{\partial V_{jk}}\right] Z^{1\mathrm{PI}}(m, n) \tag{46}$$

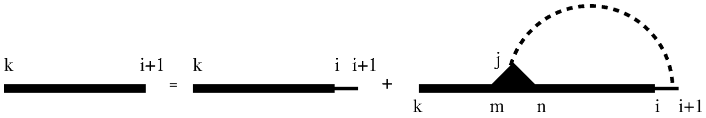


Fig. 8. Graphical representation of the recursion equation to order $1/N^2$. The black triangle represents the vertex function Γ^j_{mn}.

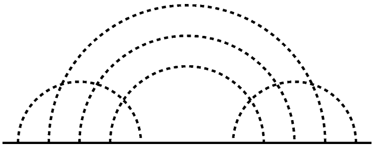


Fig. 9. A contribution of order $1/N^2$ generated by the modified recursion relation.

keeping terms to the desired order in $1/N$. Using the language of quantum chromodynamics, this equation is actually easy to describe in words. The vertex function Γ^j_{mn} describes a quark propagating from m to n and interacting with a gluon at site j. The operator $[1 - \sum_{k\neq j} V_{jk} \frac{\partial}{\partial V_{jk}}]$ simply insures that there is not already a gluon attached to the site j (see Fig. 7(a)). The relation between Γ^j_{mn} and $Z^{1\mathrm{PI}}(m,n)$ has the same form as the Ward identity in quantum electrodynamics.

Third, we calculate for $k+1 > i$

$$Z(i,k+1) = Z(i,k) + \sum_{j=1}^{k} V_{j,k+1} \sum_{m,n} Z(i,m-1)\Gamma^j_{mn} Z(n+1,k) \tag{47}$$

with the boundary condition $Z(i,i) = 1$, $Z(i,i-1) = 1$, and $Z(1,0) = 1$. The meaning of this equation is expressed graphically in Fig. 8.

The recursion is approximate because it introduces higher order terms in $1/N$ than the order we are considering.

We have checked this equation explicitly for L up to 6. A graph generated to order $1/N^2$ is displayed in Fig. 9.

These equations are adequate to order $1/N^2$, but not to order $1/N^4$.

We summarize the steps of the new recursion relation.

- Assume the partition functions $Z(i,j)$ are known for all pairs (i,j) such that $i - j < l$.
- Calculate all the one-particle irreducible functions $Z^{1\mathrm{PI}}(m,n)$ to the appropriate order in $1/N$; this can be done by cutting the quark line in $Z(m,n)$ at any site i $(m < i < n)$ between the extremities of two gluon propagators, and keeping only the graphs which do not fall apart into two separate pieces. If no gluon is connected to the site i, then this contributes to Γ^i_{mn}. When we attach the gluon to the site i, this may or may not disrupt an existing loop, and so the order in $1/N$ may or may not be changed by one power.
- Insert this function Γ and all the functions $Z(i,j)$ in (47) to calculate the partition functions with one more base.
- Iterate the process.

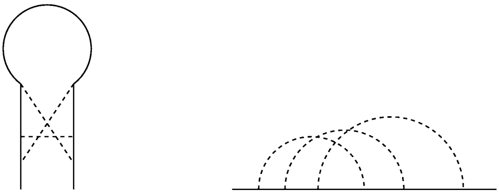


Fig. 10. A contribution of order $1/N^2$ sterically forbidden.

This procedure allows obviously to evaluate the free energy of a given RNA sequence recursively. Regard $Z(m,n)$ as the element in the mth row and nth column of a matrix. We impose the boundary conditions $Z(j,j)=1$ and $Z(j,j-1)=1$. We then use (47) to expand the matrix to ever larger size, moving "towards the northeast". In numerical evaluation, we no longer need to know the origin of the parameter $1/N^2$: we can simply take $N=1$. The factor $1/N^2$ has just allowed us to extract the most relevant diagrams beyond the Hartree theory.

To find the "ground state configuration" for a given RNA sequence we simply write (47) for $Z(1,L)$

$$Z(1,L)=Z(1,L-1)+\sum_{j=1}^{L-1}V_{jL}\sum_{m,n}Z(1,m-1)\Gamma_{mn}^{j}Z(n+1,L-1) \tag{48}$$

and evaluate it "backwards". We replace $Z(1,L)$ by the largest term on the right hand side

$$Z(1,L)\simeq\max_{j,m,n}\bigl\{Z(1,L-1),\ V_{jL}Z(1,m-1)\Gamma_{mn}^{j}Z(n+1,L-1)\bigr\}. \tag{49}$$

The largest term, in turn, comprises Z of lower order, for which we can apply this bactracking algorithm. Repeating this process, we obviously obtain the dominant configuration.

In fact, since the lowest energy configuration obtained in this way is not necessarily feasible in real space, a better strategy could be to use the backtracking algorithm to generate a set of lowest energy configurations, and check which one can be realized with real molecules with their rigidity and chemical constraints. For example, configurations such as the one of Fig. 10 with crossing "gluon" lines should be discarded, as they are forbidden by steric constraints.

7. Conclusion

We have shown that the RNA folding problem can be mapped onto a large N matrix field theory. The dominant term (N independent) is the usual Hartree theory, which is known to generate secondary structures. The $1/N$ correction term vanishes, and the correction of order $1/N^2$ generates the pseudo-knots or tertiary structure. The standard Hartree recursion relation is then replaced by a corrected recursion relation. The resulting three dimensional structure can be obtained by backtracking the recursion relation. The spatial feasibility of this tertiary structure (which remains to be checked) is due to the fact that the $1/N$ expansion classifies diagrams in terms of their topology. What remains to be done is to include the loop entropy, stacking energies and a numerical study of the recursion equations to order $1/N^2$, together with the backtracking algorithm. This will be presented in a forthcoming paper.

Acknowledgements

We wish to thank Dan Holz for help with computer programming and Walter Fontana, Paul Higgs, Terry Hwa and Luca Peliti for very useful comments.

This work was carried out during the "Statistical Physics and Biological Information" program at the ITP, Santa Barbara, and is supported in part by the NSF under grant number PHY 89-04035 at ITP.

Appendix A

In this appendix, we show that the $1/N$ correction to the free energy vanishes identically. We first note that Eq. (21) can be recast in the form

$$\begin{aligned} Z(1,L) &= \frac{1}{N}\frac{\partial}{\partial h}\int dA^*_{ll'}\,dA_{ll'}\,e^{-N\sum_{l<l'}\operatorname{tr}A_{ll'}A^*_{ll'}+N\operatorname{tr}\log M(A_{ll'})}\Big|_{h=0} \\ &= \int dA^*\,dA\,dA^*_{ll'}\,dA_{ll'}\,A^*\,e^{-N\sum_{l<l'}(\operatorname{tr}A_{ll'}A^*_{ll'}+\operatorname{tr}AA^*)+N\operatorname{tr}\log M(A_{ll'},A)} \end{aligned} \tag{A.1}$$

where

$$M(A_{ll'},A)=\begin{pmatrix} 1 & 0 & 0 & \cdot & \cdot & 0 & A \\ -1 & 1+a_{12} & a_{13} & \cdot & \cdot & a_{1L} & 0 \\ a^*_{12} & -1 & 1+a_{23} & \cdot & \cdot & a_{2L} & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & -1 & 1+a_{L-1L} & 0 \\ a^*_{1L} & a^*_{2L} & \cdot & a^*_{L-2L} & a^*_{L-1L} & -1 & 1 \end{pmatrix}.$$

The steepest descent method applied to (A.1) yields

$$A=0, \qquad A^*=M^{-1}_{L+1,1},$$

whereas the definition for all the other $A_{ll'}$ and $A^*_{ll'}$ are identical to those of Sections 4 and 5. The correction of order $1/N$ to Eq. (35) can be easily recast in the form

$$\begin{aligned} Z^{(1)} = \int da^*_{ll'}\,da_{ll'}\,da^*\,da\,\exp\left(-\operatorname{tr}a_{ll'}a^*_{ll'}-\operatorname{tr}aa^*-\frac{1}{2}\operatorname{tr}\left(M_0^{-1}c\right)^2\right) \\ \times\left\{A^*\left(\frac{1}{4}\operatorname{tr}\left(M_0^{-1}c\right)^4+\frac{1}{18}\left(\operatorname{tr}\left(M_0^{-1}c\right)^3\right)^2-\frac{1}{3}a^*\operatorname{tr}\left(M_0^{-1}c\right)^3\right)\right\}_{L+1,1} \end{aligned} \tag{A.2}$$

with the notations of Sections 4 and 5 and M_0 denotes the matrix M evaluated at the stationary point. It is clear that a^* occurs only in the term $\operatorname{tr}aa^*$ of the first line and in the term $a^* \operatorname{tr}(M_0^{-1}c)^3$ of the second line of (A.2). This second term can be integrated by part in favor of a, to remove all dependence on a^* except in the exponent. Once it is clear that a^* occurs only in the exponent, we recognize the holomorphic representation of the δ-function. Thus, the integration over a^* implies that we can set $a = 0$ everywhere. This being done, we see that all the terms like $a^*_{12}, \dots, a^*_{1L}$ and $a^*_{2L}, \dots, a^*_{L-1,L}$ are present only in the exponent (in the $\operatorname{tr} a_{ll'} a^*_{ll'}$ term). Therefore, we can integrate them out, and the result is again a set of δ-function which impose

$$a_{12} = \dots = a_{1L} = a_{2L} = \dots = a_{L-1,L} = 0.$$

This procedure can be carried out recursively to “eat up” all the a^* and a, leading to the vanishing of the $1/N$ correction (A.2).

Appendix B

B.1. Recursion

An alternative strategy to evaluating Z recursively is by integrating out φ_{L+1} in the expression for $Z(1, L+1)$. For notational simplicity, let us define $\mu^2 \equiv (V^{-1})_{L+1,L+1}$, $M \equiv \varphi_{L+1}$ and $T \equiv \sum_{i=1}^{L} (V^{-1})_{L+1,i}\varphi_i$. Evidently, we have to do two Gaussian integrals over M:

$$\int dM\, e^{-N \operatorname{tr}(TM + \frac{\mu^2}{2} M^2)} = C(\mu, N) e^{+\frac{N}{2\mu^2} \operatorname{tr} T^2} \tag{B.1}$$

and

$$\int dM e^{-N \operatorname{tr}(TM + \frac{\mu^2}{2} M^2)} M = -\frac{1}{\mu^2} C(\mu, N) e^{+\frac{N}{2\mu^2} \operatorname{tr} T^2} T, \tag{B.2}$$

where (B.2) is obtained by differentiating (B.1) with respect to the matrix T. Thus, after integrating out φ_{L+1} in $Z(1, L+1)$, we find that the “action” $\sum_{ij} (V^{-1})_{ij} \operatorname{tr}(\varphi_i\varphi_j)$ has been replaced by the effective action $\sum_{ij} (\widetilde{V}^{-1})_{ij} \operatorname{tr}(\varphi_i\varphi_j)$ where $(\widetilde{V}^{-1})_{ij} = (V^{-1})_{ij} - (V^{-1})_{i,L+1} \frac{1}{(V^{-1})_{L+1,L+1}} (V^{-1})_{L+1,j}$. It is easy to see that $\widetilde{V}$ is the L by L matrix obtained by crossing out the last row and column of the $L+1$ by $L+1$ matrix V, as we might have expected. Putting these steps together we obtain

$$Z(1, L+1) = Z(1, L) - \frac{1}{(V^{-1})_{L+1,L+1}} \sum_{l=1}^{L} \left(V^{-1}\right)_{L+1,l} \left\langle \frac{1}{N} \operatorname{tr}\left(\prod_{i=1}^{L} (1 + \varphi_i)\right) \varphi_l \right\rangle, \tag{B.3}$$

where $(\prod_{i=1}^{L} (1 + \varphi_i))$ is ordered as before. The expectation value of a matrix O constructed out of the φ_i's is defined by

$$\langle O \rangle \equiv \frac{1}{A(L)} \int \prod_k d\varphi_k\, e^{-N \frac{1}{2} \sum_{ij} (V^{-1})_{ij} \operatorname{tr}(\varphi_i\varphi_j)} O. \tag{B.4}$$

In other words, $Z(1, L) \equiv \langle \frac{1}{N} \operatorname{tr} \prod_i (1 + \varphi_l) \rangle$.

To evaluate $\langle \frac{1}{N}\operatorname{tr}(\prod_{i=1}^{L}(1+\varphi_i))\varphi_l\rangle$ we follow the standard procedure of replacing

$$\varphi_l e^{-N\frac{1}{2}\sum_{ij}(V^{-1})_{ij}\operatorname{tr}(\varphi_i\varphi_j)} \to -\frac{1}{N}\sum_{k=1}^{L} V_{lm}\frac{\delta}{\delta\varphi_k} e^{-N\frac{1}{2}\sum_{ij}(V^{-1})_{ij}\operatorname{tr}(\varphi_i\varphi_j)}.$$

Integrating by parts, we finally obtain

$$Z(1,L+1) = Z(1,L) + \sum_{k=1}^{L} V_{L+1,k}\left\langle \frac{1}{N}\operatorname{tr}\left(\prod_{i=1}^{k-1}(1+\varphi_i)\right)\frac{1}{N}\operatorname{tr}\left(\prod_{j=k+1}^{L}(1+\varphi_j)\right)\right\rangle. \tag{B.5}$$

In other words, in (B.3) we have Wick contracted φ_l with φ_k in the ordered product $\prod_{i=1}^{L}(1+\varphi_i)$. Evidently, $\frac{1}{N}\operatorname{tr}(\prod_{i=1}^{k-1}(1+\varphi_i))$ is to be interpreted as 1 for $k=1$. Similarly, $\frac{1}{N}\operatorname{tr}(\prod_{j=k+1}^{L}(1+\varphi_j))$ is to be interpreted as 1 for $k=L$.

In principle, we can extract what we need from this recursion relation (B.5). We emphasize that (B.5) is derived without taking the large N limit and holds for finite N, including $N=1$.

B.2. Large N expansion

We can now perform a large N expansion, giving us a systematic way of evaluating Z to any desired order of $1/N^2$. In the language of quantum chromodynamics, quantities in which the indices of the matrices φ_j are summed over such as $\frac{1}{N}\operatorname{tr}(\prod_{i=1}^{k-1}(1+\varphi_i))$ are known as color singlet operators. It is well known [11] that given two color singlet operators A and B, the expectation value factorizes to leading order in large N:

$$\langle AB\rangle = \langle A\rangle\langle B\rangle + \langle AB\rangle_C \tag{B.6}$$

with the connected correlation function $\langle AB\rangle_C$ suppressed by a factor of $O(1/N^2)$ relative to $\langle A\rangle\langle B\rangle$. It is easy to see the validity of (B.6) by drawing a few diagrams such as those in Fig. 8(b). Connected correlation functions $\langle AB\rangle_C$ have been intensively studied [13] in the matrix theory literature and a good deal is known about them. Thus, we can write in (B.5)

$$\begin{aligned}
&\left\langle \frac{1}{N}\operatorname{tr}\left(\prod_{i=1}^{k-1}(1+\varphi_i)\right)\frac{1}{N}\operatorname{tr}\left(\prod_{j=k+1}^{L}(1+\varphi_j)\right)\right\rangle \\
&\quad= \left\langle \frac{1}{N}\operatorname{tr}\left(\prod_{i=1}^{k-1}(1+\varphi_i)\right)\right\rangle\left\langle \frac{1}{N}\operatorname{tr}\left(\prod_{j=k+1}^{L}(1+\varphi_j)\right)\right\rangle \\
&\qquad+ \left\langle \frac{1}{N}\operatorname{tr}\left(\prod_{i=1}^{k-1}(1+\varphi_i)\right)\frac{1}{N}\operatorname{tr}\left(\prod_{j=k+1}^{L}(1+\varphi_j)\right)\right\rangle_C.
\end{aligned} \tag{B.7}$$

We immediately recognize that first term in (B.7) as $Z(1,k-1)Z(k+1,L)$. By definition, the connected correlation function

$$Z_C(1,k-1;k+1,L) \equiv \left\langle \frac{1}{N}\operatorname{tr}\left(\prod_{i=1}^{k-1}(1+\varphi_i)\right)\frac{1}{N}\operatorname{tr}\left(\prod_{j=k+1}^{L}(1+\varphi_j)\right)\right\rangle_C$$

is evaluated by contracting a matrix φ_i from one of the traces to a matrix φ_j from the other trace. Thus the exact recursion relation is given by

$$\begin{aligned} Z(1,L+1) = Z(1,L) &+ \sum_{k=1}^{L} V_{L+1,k} Z(1,k-1) Z(k+1,L) \\ &+ \sum_{k=1}^{L} V_{L+1,k} Z_C(1,k-1;k+1,L). \end{aligned} \tag{B.8}$$

This gives an alternative representation of (44). Evidently,

$$Z_C(1,k-1;k+1,L) = \langle M^{-1}_{L+1,k+1} M^{-1}_{k,1} \rangle_C - \frac{1}{N} \langle M^{-1}_{L+1,1} M^{-1}_{k,k+1} \rangle_C. \tag{B.9}$$

In principle, we can take the exact recursion relation (B.8) and evaluate the two terms on the right hand side to any desired order in $1/N$ and thus generate, given an RNA sequence, secondary structure, tertiary structure, ad infinitum.

References

[1] T. Pan, D.L. Long, O. Uhlenbeck, Divalent metal ions in RNA folding and catalysis, in: R.F. Gesteland, J.F. Atkins (Eds.), The RNA World, Cold Spring Harbor Laboratory Press, pp. 271–302.

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

[3] P.G. Higgs, Quarterly Reviews in Biophysics 33 (2000) 199.

[4] T.E. Creighton (Ed.), Protein Folding, W.H. Freeman, New York.

[5] T. Garel, H. Orland, E. Pitard, Protein folding and heteropolymers, in: A.P. Young (Ed.), Spin Glasses and Random Fields, World Scientific, Singapore, 1997.

[6] V.K. Misra, D.E. Draper, Biopolymers 48 (1998) 113.

[7] R. Nussinov, A.B. Jacobson, PNAS 77 (1980) 6309.

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

[9] I.L. Hofacker, W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, P. Schuster, Monatshefte für Chemie 125 (1994) 167, The Vienna RNA software package is available at http://www.tbi.univie.ac.at/~ivo/RNA.

[10] E. Rivas, S.R. Eddy, J. Mol. Biol 285 (1999) 2053.

[11] G. t' Hooft, Nucl. Phys. B 72 (1974) 461;
G. t' Hooft, Nucl. Phys. B 75 (1974) 461.

[12] P. Di Francesco, O. Golinelli, E. Guitter, Commun. Math. Phys. 186 (1997) 1.

[13] For example: E. Brézin, A. Zee, Nucl. Phys. B 453 (FS) (1995) 531.

[14] H. Orland, J. Phys. Lett. (Paris) 46 (1985) L763.

[15] M. Mézard, G. Parisi, J. Phys. Lett. (Paris) 46 (1985) L771.

[16] T.M. Nieuwenhuizen, Physica A 252 (1998) 178.

[17] Y.C. Zhang, cond-mat/0105186.

[18] E. Brézin, C. Itzykson, G. Parisi, J.B. Zuber, Commun. Math. Phys. 59 (1978) 35.

[19] See for instance: J. Zinn-Justin, Quantum Field Theory and Critical Phenomena, Oxford University Press, 1999.

[20] P.G. de Gennes, Biopolymers 6 (1968) 715.

[21] M.S. Waterman, T.F. Smith, Adv. Applied Maths. 7 (1986) 455.

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

[23] R. Bundschuh, T. Hwa, Phys. Rev. Lett. 83 (1999) 1479.
[24] D.K. Lubensky, D.R. Nelson, Phys. Rev. Lett. 85 (2000) 1572.
[25] A. Montanari, M. Mézard, Phys. Rev. Lett. 86 (2001) 2178.
[26] A recent use of the Bethe–Salpeter equation in the context of large N matrix theory can be found in: J. Moore, J. Sinova, A. Zee, cond-mat/0012341.
[27] E. Brézin, A. Zee, CR. Acad. Sci. II 317 (1993) 735.