LARGE-N RANDOM MATRICES FOR RNA FOLDING*

Graziano Vernizzi, Henri Orland

Service de Physique Théorique, CEA/Saclay
F91191 Gif-sur-Yvette Cedex, France
vernizzi@cea.fr

(Received July 28, 2005)

We review a recent formulation of the RNA folding problem as an $N \times N$ matrix field theory. It is based on a systematic classification of the terms in the partition function according to their topological character. In particular large-N terms yield the secondary structures, whereas pseudoknots are obtained by calculating the $1/N^2$ corrections. We also describe a Monte Carlo algorithm for the prediction of RNA secondary structures with pseudoknots, based on this topological approach.

PACS numbers: 87.14.Gg, 82.39.Pj, 02.10.Yn

1. Introduction

It is well-known that the three-dimensional structure of a RNA molecule is intimately connected to its specific biological function in the cell (*e.g.* for protein synthesis and transport, catalysis, chromosome replication and regulation) [1]. For this reason, the quest for an algorithm which can predict the spatial structure of the RNA molecule given its chemical sequence has received considerable attention from molecular biologists in recent years. The primary structure of the RNA is the sequence of covalently linked nucleotides along the molecule from the 5' to the 3' end. The four basic types of nucleotides are adenine (A), cytosine (C), guanine (G) and uracil (U). At room temperature, different nucleotides can pair by means of saturating hydrogen bonds, *i.e.* the standard Watson–Crick pairs A–U, C–G, and the wobble pair G–U. Adjacent base pairs can stack, providing and additional binding energy which is actually the origin of the formation of stable A-form helices, one of the main structural characteristics of folded RNAs. Helices may embed unpaired sections of RNA, in the form of hairpins, loops and bulges.

* Presented at the Conference on Applications of Random Matrices to Economy and Other Complex Systems, Kraków, Poland, May 25–28, 2005.

It is all these pairings, stackings of bases and structural motifs which bring the RNA into its folded three-dimensional configuration. It is also possible to define secondary structures of RNA as structures in which the pairings between canonical base pairs do not cross in a certain representation (*e.g.* planar disk diagrams, see below). Finally, one defines the tertiary structure of RNA as the actual three-dimensional arrangement of the base sequence. This classification corresponds to the fact that the secondary structure of RNA carries the main contribution to the free energy of a fully folded RNA configuration, including also some of the steric constraints. For that reason one can attempt to describe the folding process hierarchically. Over the past twenty years several algorithms have been proposed for the prediction of RNA folding. It is fair to say that despite the large number of tools available for the prediction of RNA structures, no reliable algorithms exist for the prediction of the full RNA tertiary structure, and many of the exiting algorithms deal with the prediction of the secondary structure only. To describe the full folding it is important to introduce the concept of RNA pseudoknot [2]. One says that two base pairs form a pseudoknot when the parts of the RNA sequence spanned by those two base pairs are neither disjoint, nor have one contained in the other. Thus RNA secondary structures without pseudoknots can be represented by planar diagrams, whereas RNA with pseudoknots appear when two base pairs can "cross", leading to non-planar diagrams. Pseudoknots play important structural, regulatory and catalytic roles in natural RNAs [3]. However, pseudoknots are excluded in the definition of RNA secondary structure and many authors consider them as part of the tertiary structure. This restriction is due to the fact that RNA secondary structures without pseudoknots can be predicted easily. One should also note that pseudoknots very often involve base-pairing from distant parts of the RNA, and are thus quite sensitive to the ionic strength of the solution. It has been shown that the number of pseudoknots depends on the concentration of Mg^{++} ion, and can be strongly suppressed by decreasing the ionic strength (thus enhancing electrostatic repulsion).

2. Topology of RNA pseudoknots and large-N matrix theory

Rivas and Eddy noticed in [4] a correspondence between a graphical representation of RNA secondary structures with pseudoknots and Feynman diagrams. In [5] the correspondence between RNA secondary structures and Feynman diagrams is made more explicit and general by formulating a matrix field theory model whose Feynman diagrams give exactly all the RNA secondary structures with pseudoknots. The remarkable facts of this new approach is that it provides an analytic tool for the prediction of pseudoknots, and all the diagrams appear to be naturally organised in a series of

terms, called the topological expansion, where the first term corresponds to planar secondary structures without pseudoknots, and higher-order terms correspond to structures with pseudoknots. We explore here in more details this topological expansion and its potential predictive power.

First of all, it is important to recall a graphical representation of RNA secondary structures. Among the several ways to represent an RNA secondary structure, given the primary structure, we consider here the disk diagram representation. The RNA sequence is represented as an oriented circle (from 5' to 3') by virtually linking the first nucleotide to the last one, and each base pairing is represented as an arc inside the circle, connecting the two paired bases. Figure 1 shows a typical disk diagram.

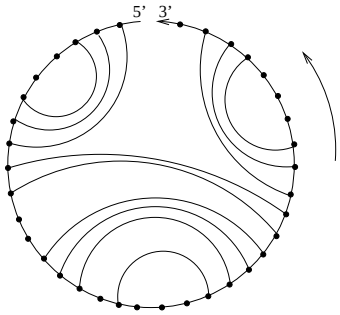

Fig. 1. Typical disk diagram representation of the RNA secondary structure without pseudoknots. The circle is anticlockwise oriented from 5' to 3'. Note that there are no crossing arcs.

In this representation, all secondary structures without pseudoknots are purely planar diagrams, *i.e.* diagrams that can be drawn without crossing arcs, whereas pseudoknots correspond to structures which are not planar (see figure 2).

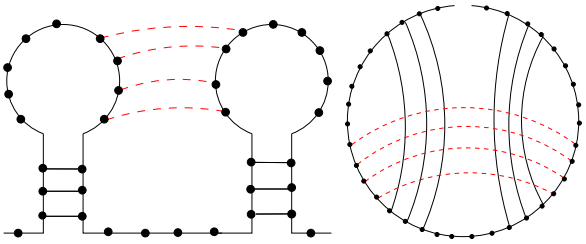

Fig. 2. A "kissing hairpin" pseudoknot, and the respective disk diagram (which has crossing arcs necessarily).

As we said, such a graphical representation has an analytical counterpart, corresponding to the Feynman diagrammatic expansion of the following matrix integral:

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

Here φ_i $(i = 1, \cdots, L)$ denote L independent N by N Hermitian matrices and $\Pi_l(1+\varphi_l)$ represents the ordered matrix product $(1+\varphi_1)(1+\varphi_2)\cdots(1+\varphi_L)$. The normalisation factor $A(L)$ is defined by

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

The crucial point is that the matrix theory defined by (1) has the same topological structure as 't Hooft's large N topological expansion. The reader familiar with matrix theory or large N field theory sees immediately that the Gaussian matrix integral (1) evaluates precisely to the infinite series

$$Z = 1 + \sum_{\langle ij\rangle} V_{ij} + \sum_{\langle ijkl\rangle} V_{ij}V_{kl} + \cdots + \frac{1}{N^2}\sum_{\langle ijkl\rangle} V_{ik}V_{jl} + \cdots\,, \qquad (3)$$

where $\langle ij\rangle$ denotes all pairs with $j > i$, $\langle ijkl\rangle$ all quadruplets with $l > k > j > i$, and so on. We identify the matrix:

$$V_{ij} = e^{-\beta\varepsilon_{ij}}\theta(|i-j| \geq 4)\,, \qquad (4)$$

where ε_{ij} is the matrix giving the attractive energy between the i-th and j-th nucleotides. The Heaviside function $\theta(|i-j| \geq 4)$ accounts the fact that the RNA molecule is not infinitely flexible and one cannot pair nucleotides separated by less than 3 bases. Note that the saturation of the hydrogen bond corresponds to the strict inequalities $l > k > j > i$, and so on. Once the nucleotide at i has interacted with the nucleotide at j it cannot interact with the nucleotide at k .

With the choice (4) for the matrix V_{ij}, the partition function of equation (3) is exactly the partition function of the RNA molecule where the entropic contribution is neglected. This is so, because only the energetic contribution from the contact structure of the folding is taken into account. A direct application of the matrix integral (1) in the limit of an infinitely flexible homopolymer chain can be found in [7]. In the following section we will introduce a model that includes also the entropic contribution effectively.

The topological expansion of the above matrix integral provides a very natural way for classifying the "degree of non-planarity" of any given RNA disk diagram. It is based on a topological analysis introduced long ago by Euler, and it has been already introduced in [5] for RNA secondary structures. The main idea is to draw the disk diagram on a surface with a sufficient number of "handles", such that crossing arcs can be avoided. The minimum number of handles is called genus of the surface. For instance, the surface corresponding to the pseudoknot in figure 2 would be a torus, since that disk diagram can be drawn there without crossing arcs. We therefore say that the "kissing hairpin" pseudoknot has genus 1. This correspondence is not one-to-one, and actually there are 8 fundamental types of pseudoknots with genus 1 (for a complete list see [6]). We propose to classify all RNA pseudoknots according to their genus. This idea can be exploited when formulating a statistical mechanics model for RNA structures with pseudoknots.

3. RNA pseudoknots from Monte Carlo simulations

The most popular and successful bioinformatics technique for predicting secondary structures (without pseudoknots) is "dynamic programming" (see *e.g.* [8]), for which the memory and CPU requirements scale with the sequence length L as $\mathrm{O}(L^2)$ and $\mathrm{O}(L^3)$ respectively. Recently, new deterministic algorithms that deal with pseudoknots have been formulated (*e.g.* [4, 9, 10]) but the memory and CPU requirements are generally very demanding, even for short RNA sequences. The increase of computational complexity does not come as a surprise. In fact the RNA-folding problem with pseudoknots has been proven to be NP-complete for some classes of pseudoknots [11]. For that reason, stochastic algorithms might be a better choice to predict secondary structures with pseudoknots in a reasonable time and for long enough sequences. In [12] stochastic Monte Carlo algorithms for the prediction of RNA pseudoknots have been proposed. In these stochastic approaches, the very irregular structure of the energy landscape (glassy-like) is the main obstacle: configurations with small differences in energy may be separated by high energy barriers, and the system may very easily get trapped in metastable states. Among the stochastic methods, the direct simulation of the RNA-folding dynamics (including pseudoknots) with kinetic folding algorithms [13] is the most successful. This technique allows to describe the succession of secondary structures with pseudoknots during the folding process. The approach we follow is close in spirit to that one, with a stronger emphasis on the topological character of the RNA pseudoknots. In fact one can control the topological character of pseudoknots by simply coupling the free energy of the RNA molecule with an additional pa-

rameter μ, which is a topological "chemical potential". Namely the standard free energy of the RNA configuration

$$\mathcal{F} = E - TS \tag{5}$$

is modified to

$$\mathcal{F} = E - TS + \mu g, \tag{6}$$

where E is the internal energy (from base pairs and stackings), S is the entropy (internal loops, bulges, hairpin loops), T is the temperature and g is the genus of the configuration. We perform a Monte Carlo simulation with the standard Metropolis method for generating a set of RNA configurations distributed according to the Boltzmann weight

$$P = \frac{1}{\mathcal{Z}} e^{-(E-TS+\mu g)/kT}, \tag{7}$$

where k is the Boltzmann constant, and $\mathcal{Z}$ is the partition function (including the entropic contribution this time). Assuming that at low temperature the RNA molecule assumes a configuration which corresponds to the minimum energy, we can also find such a configuration by using the so-called simulated annealing method [12]. The model without chemical potential, *i.e.* $\mu = 0$, corresponds to the case where there are no restrictions on the possible fluctuations of the topology. On the other hand when μ is very large, all the configurations with $g > 0$ are suppressed by the Boltzmann weight, and one recovers the planar limit (*i.e.* RNA secondary structures without pseudoknots). A first check of this method is whether we can reproduce the results produced by deterministic algorithms such as "mfold" or the "Vienna Package" [14]. For that purpose, it is sufficient to use the very same energy model and run our algorithm with a large value of the chemical potential μ. Our preliminary tests show that the minimum can be easily found for sequences with length up to around 300 bases. For longer RNA sequences, the simulation time increases and the minimum is harder to find. In these cases we use an additional feature of our model. In fact our approach offers also the interesting possibility of using the chemical potential for overcoming the energy barriers. It means that we can apply a "simulated annealing" method on μ rather than on T. Thus, starting with a low value of μ (where all the topologies with any genus are allowed) the Monte Carlo simulation can quickly explore regions which are very distant from each other in the energy landscape. Then by slowly increasing the value of μ we gradually constrain the simulation to select only planar configurations (*i.e.* secondary structures without pseudoknots), and the minimum energy configuration eventually. During this process, that is for intermediate values of μ, many configurations in thermal equilibrium are generated, and in general they correspond

to RNA configurations with pseudoknots. These configurations should be compared with the experimental data. It is at this level that the value of μ can be tuned, in order to fit the data. Unfortunately, to our knowledge there are no available experimental data about the dependence of the genus of RNA molecules on the temperature. Information and inputs from experiments would be highly desirable. At the moment we are able, by means of the algorithm described above, to predict correctly the pseudoknotted structures of short sequences of *real* RNA.

REFERENCES

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

[2] C.W. Pleij, K. Rietveld, L. Bosch, *Nucleic Acids Res.* **13**, 1717 (1985).

[3] E. Westhof, L. Jaeger, *Current Opinion Struct. Biol.* **2**, 327 (1992).

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

[5] H. Orland, A. Zee, *Nucl. Phys.* **B620**, 456 (2002).

[6] M. Pillsbury, H. Orland, A. Zee, `http://arXiv.org/physics/0207110`

[7] G. Vernizzi, H. Orland, A. Zee, *Phys. Rev. Lett.* **94**, 168103 (2005).

[8] M. Zuker, D. Sankoff, *Bull. Math. Biol.* **46**, 591 (1984).

[9] Y. Uemura, A. Hasegawa, S. Kobayashi, T. Yokomori, *Theor. Comput. Sci.* **210**, 277 (1999); Y.T. Akutsu, *Discr. Appl. Math.* **104**, 45 (2001).

[10] M. Pillsbury, J.A. Taylor, H. Orland, A. Zee, `http://arXiv.org/cond-mat/0310505`

[11] R.B. Lyngsø, C.N.S. Pedersen, *J. Comp. Biol.* **7**, 409 (2000).

[12] M. Schmitz, G. Steger, *J. Mol. Biol.* **255**, 254 (1996); J.P. Abrahams, M. van den Berg, E. van Batenburg, C.W.A Pleij, *Nucleic Acids Res.* **18**, 3035 (1990); A.P. Gultyaev, *Nucleic Acids Res.* **19**, 2489 (1991).

[13] H. Isambert, E.D. Siggia, *Proc. Natl. Acad. Sci. USA* **97**, 6515 (2000); A. Xayaphoummine, T. Bucher, F. Thalmann, H. Isambert, *Proc. Natl. Acad. Sci. USA* **100**, 15310 (2003).

[14] M. Zuker, *Nucleic Acids Res.* **31**, 3406 (2003). Also, I.L. Hofacker, *Nucleic Acids Res.* **31**, 3429 (2003).