PRAMANA
— journal of
physics

© Indian Academy of Sciences

Vol. 77, No. 5
November 2011
pp. 827–832

Matrix models of RNA folding with external interactions: A review

I GARG and N DEO*

Department of Physics & Astrophysics, University of Delhi, Delhi 110 007, India
*Corresponding author. E-mail: ndeo@physics.du.ac.in

Abstract. The matrix model of (simplified) RNA folding with an external linear interaction in the action of the partition function is reviewed. The important results for structure combinatorics of the model are discussed and analysed in terms of the already existing models.

Keywords. RNA folding-structure combinatorics; random matrix models; external perturbation; structural properties.

PACS Nos 87.14.gn; 02.10.Yn; 87.10.−e; 87.15.Cc; 87.15.−v

1. Installation

The improved understanding of the role of RNA in biological activities with discoveries and developments in the field of biophysics has highlighted the importance of studying their tertiary (folded 3D) conformations [1]. At the very base of understanding the different levels of structures of these biomolecules, lies the quest for understanding three fundamental problems (given in order): (i) to predict an RNA structure (enumeration), (ii) to find energetically viable structures from the enumerated structures and (iii) to determine kinetics of fold formation [2]. Therefore it is extremely essential to first enumerate and classify all possible types of structures (secondary and tertiary) with a given length, i.e., to know the combinatorics. Secondary structures of RNA have been studied successfully and widely using different statistical and computational models, particularly dynamic programming algorithms [3]. Tertiary structures, owing to their complexity, have been largely unaccessible [4]. Some models have captured the effects of pseudoknotted conformations [5] on the combinatorial and thermodynamic aspects [6]. In particular, a graph theoretic model by Haslinger and Stadler [7] considered bi-secondary structures (secondary structures with non-nested pseudoknots) in addition to the secondary structures and found that the total structures grow asymptotically as β^L (where L is the length of RNA chain and β is the constraint-dependent combinatoric factor). Hence, depending upon the different constraining conditions faced by an RNA chain under physiological conditions, the total possible conformations may vary. This very idea springs up the thought of interesting outcomes in the study of the effects of different kinds of perturbations/external interactions/constraints

DOI: 10.1007/s12043-011-0193-1; *e*Publication: 31 October 2011

on the structural combinatorics of RNA in statistical modelling. With this hindsight, we choose a statistical model of RNA folding [8] that computes and sequentially arranges all the possible secondary and tertiary structures (according to the structural complexity, genus) that can exist for a given length of RNA chain (here length is defined as the number of nucleotide bases in the chain, not the geometric end-to-end distance). The enumeration is made possible by some simplifying assumptions: the RNA chain is considered to be infinitely flexible (i.e., adjacent base pairings can take place) with non-complementary base pairings allowed and all the base pairing interactions happen with the same probability. To this model, an external (interaction) term is added in the action of the partition function that defines an RNA chain of length L (discussed in §2). The effect of the externally introduced interaction [9] on the structural properties such as enumeration, distribution functions (with respect to the length and genus of the structures), asymptotic (large L) behaviour [10] and thermodynamic properties [11] such as (a) free energy and specific heat as a function of length, temperature and external interaction parameter and (b) distribution of structures with respect to temperature and different number of pairings (possible for a given length structure) are studied in §2.

2. The model

The partition function equation of the matrix model of RNA folding with a linear external interaction [9] is given by

$$Z_L(N) = \frac{1}{A_L(N)} \int \prod_{i=1}^{L} \mathrm{d}\phi_i \mathrm{e}^{(-N/2)\sum_{i,j=1}^{L}(V^{-1})_{i,j}\mathrm{Tr}\,\phi_i\phi_j} \times \mathrm{e}^{-N\sum_{i=1}^{n}(W^{-1})_i \mathrm{Tr}\,\phi_i} \frac{1}{N}\mathrm{Tr}\prod_{i=1}^{L}(1+\phi_i), \tag{1}$$

where the linear interaction term is $\mathrm{e}^{-N\sum_{i=1}^{n}(W_i)^{-1}\mathrm{Tr}\,\phi_i}$. The interaction acts on $i = 1, ..., n$ bases of the chain where $n \leq L$ and $W_i = w$ gives the strength of the external perturbation which acts uniformly on each base. All the simplifying assumptions discussed in the Introduction for the model in [8] hold. The ϕ_is are $i = 1, ..., L$ independent $(N \times N)$ random Hermitian matrices placed at each of the L bases in the chain. The interaction between different ϕ_is (bases) is contained in an $(L \times L)$ interaction matrix V_{ij} with $V_{ij} = \exp[-\beta\epsilon_{ij} v_{ij}(r_{ij})]\theta(|i-j| > 4)$ as the elements [8]. Here $\beta = 1/T$ gives the temperature, $v_{ij}(r_{ij})$ is the short-range space-dependent part of the attractive interaction between nucleotides at base positions i and j, ϵ_{ij} is a (4×4) symmetric matrix giving base-specific pairing energies between i and j, $\theta(|i-j| > 4)$ is the Heaviside function which ensures finite flexibility of the chain and preserves the fact that only those bases which are separated by at least four positions can interact. All the diagonal elements $V_{ii} = 0$, i.e., no self-pairing is allowed. The inter-base interaction matrix in view of the simplifications becomes $V_{ij} = v = \exp(-\beta\epsilon)$. The observable $\prod_i(1+\phi_i)$ is an ordered matrix product over ϕ_i which also ensures that V_{ii} do not appear in the partition function $Z_L(N)$. $A_L(N)$ is the normalization constant, $A_L(N) = \int\prod_{i=1}^{L}\mathrm{d}\phi_i \mathrm{e}^{(-N/2)\sum_{i,j=1}^{L}(V^{-1})_{ij}\mathrm{Tr}\,\phi_i\phi_j}\mathrm{e}^{-N\sum_{i=1}^{n}(W^{-1})_i\mathrm{Tr}\,\phi_i}$.

In this construction of the partition function, the observable plays an important role. If one adds a linear term in the Gaussian matrix model partition function (without the observable) then it is just a generalization of the partition function of the scalar Gaussian field theories to Gaussian matrix field theories. The special form of the observable incorporates the specific properties of RNA and pulls down the V_{ij}s from the quadratic term in the action to give partition functions for different lengths in terms of the base-pairing interaction v and N (using the Wick theorem). The importance of adopting this method for enumeration of RNA structures (though simplified) lies in that the different genus structures are arranged sequentially in a series (of powers of $1/N^2$) with proper distinction between planar (terms with $(1/N^2)^0$) and non-planar structures (terms with $(1/N^2)^k$ where $k \geq 1$), adopted in analogy with [12]. This facilitates the study of the distribution functions of secondary and tertiary structures separately. Since the random matrix models give averaged and universal properties of a system under consideration [13], this method may be very useful in getting a handle on the average characteristics of real RNA molecules.

The generating function of the partition function, following the mathematical steps in [9], is

$$G\left(t, N, \frac{n\alpha}{L}\right) = \sum_{L=0}^{\infty} Z_{L,\frac{n\alpha}{L}}(N)\frac{t^L}{L!} = \mathrm{e}^{\frac{vt^2}{2N}+t(1-\frac{n\alpha}{L})}\left[\frac{1}{N}\sum_{k=0}^{N-1}\begin{pmatrix} N \\ k+1 \end{pmatrix}\frac{(t^2v)^k}{k!N^k}\right], \tag{2}$$

where the effect of interaction appears in the exponent as $\mathrm{e}^{-t(n\alpha/L)}$. This equation can be used to find the partition function for interaction acting on different n bases of the chain with the strength α [9]. When the interaction is assumed to act on one ($n = 1$) and all ($n = L$) the bases uniformly, one gets the model and its properties were studied in [10]. These cases were studied by assuming that external linear interaction strength, in the beginning of the mathematical derivation, is the same on all bases in the chain ($W_{\mathrm{i}} = w$) but it is interesting to observe how this assumption translates into the results found for this model (discussed in the next paragraph). Further, it is also possible to calculate the effect of interaction when the n different bases are acted upon by different strengths of interaction [9]. In this case, $n\alpha/L$ is replaced by a more complicated function, $C(L, w_i)$, of length L of the chain and different external interaction strengths W_{i}. For instance, if the interaction acts on a single base only (base 1 in this case) with strength w_1, then $C(L, w_1) = 1/Lw_1$ and for two bases only (bases 1 and 2 with strengths w_1 and w_2 respectively) it is $C(L, w_1, w_2) = (1/Lw_1) + (1/Lw_2)$. Furthermore, if $W_1 = w_1$, $W_2 = w_2$ and $W_3 = W_4 = \cdots = W_L = w$, then $C = ((Lw_1w_2 + ww_1 + ww_2 - 2w_1w_2)/Lww_1w_2)$. Further, the asymptotic behaviour of the genus distribution functions for the matrix model of RNA with interaction ($n = 1$ and L) in [10] is found numerically. The numerical analysis shows that the term 3^L in $a_{L,g,\alpha}$ found in [8] changes to $(3-\alpha)^L$ when $n = L$ (which becomes $(3-(n\alpha/L))^L$ when the perturbation is on n bases). The RNA structure combinatorial problem (with pseudoknots: only bi-secondary structures) has been solved with a graph theoretic approach in [7]. It has been shown that the number of bi-secondary structures (secondary structures with non-nested pseudoknots) grows asymptotically as $(\beta)^L$ where β is a combinatorial factor that depends upon different constraints that an RNA chain is subject to and L is the length of the chain. The effect of α is therefore like an added constraint in the model which can be modelled as per requirement in different theoretical/experimental situations.

The general form of the partition function (obtained from (1) or (2)) is

$$\begin{aligned} Z_{L,\alpha}(N) = {} & \left(1-\frac{n\alpha}{L}\right)^{L} + \left(1-\frac{n\alpha}{L}\right)^{(L-2)} \sum_{i<j} V_{i,j} \\ & + \left(1-\frac{n\alpha}{L}\right)^{(L-4)} \sum_{i<j<k<l} V_{i,j}V_{k,l} \\ & + \left(1-\frac{n\alpha}{L}\right)^{(L-4)} \sum_{i<j<k<l} V_{i,l}V_{j,k} \\ & + \left(\frac{1}{N}\right)^{2} \left(1-\frac{n\alpha}{L}\right)^{(L-4)} \sum_{i<j<k<l} V_{i,k}V_{j,l} \\ & + \ldots \end{aligned} \tag{3}$$

which gives the structural regimes into which the RNA structures are distributed for the model with linear interaction: Regime 1: $0 \leq \alpha \leq 1, n < L$ and $0 \leq \alpha < 1, n = L$ and Regime 2: $\alpha = 1, n = L$. In the linear representation of the biomolecules, the effect of the external linear interaction appears as a factor $(1 - n\alpha/L)$ (which is the multiplicative term in the general partition function as compared to [8]) on each unpaired base of the chain with no such distinction or weighting of the paired bases in the chain. The structures in Regime 1 therefore consist of structures which have a combination of paired and unpaired sites in the chain and Regime 2 has only completely paired structures which do not have any unpaired sites at all. The matrix model formulation therefore provides a convenient mathematical method for arranging different structures according to genus in an asymptotic series of powers of $1/N^2$ and displaying the effect of interactions very clearly. The partition function obtained in this way characterizes the whole ensemble of structures possible for a chosen length according to the number of pairings in each structure (power of v), genus and the number of unpaired sites by the power of weight $(1 - n\alpha/L)$. The coefficients of v give the total number of conformations $a_{L,g,\alpha}$ for a fixed L, g and α. For a given length and α, the total possible structures for all genii are obtained by putting $v = 1$ and $N = 1$ and are labelled by $\mathcal{N}_\alpha$.

This general partition function exhibits a scaling relation with the model in [8] which can be observed by factoring out the term $(1 - (n\alpha/L))^L$ from (3) to get $Z_L(v, n\alpha/L, N) = (1 - (n\alpha/L))^L Z_L[(v/(1 - (n\alpha/L))^2), 0, N]$. The bracketed quantity on the right side is the partition function of the RNA matrix model in [8] where the base pairing strength parameter v is re-scaled by an amount $(1 - (n\alpha/L))^{-2}$. Such scaling forms have not been seen in the context of matrix theory and hence provides an interesting mathematical model construction, both from the point of view of RNAs and matrix models. For the RNAs, the re-scaling of v physically corresponds to changing of the base pairing interaction strength which may be caused by external conditions such as applied pressure or proximity with ions [14].

3. Conclusions

The matrix models of RNA folding address an important question in the basic understanding of biomolecules, i.e., the exhaustive enumeration of all possible conformations, secondary and tertiary, of a chain of any given length. Though the method adopted is simplified and gives a schematic picture of the RNA chain, it can be used to obtain the distribution functions and universal characteristics that are found for the real RNA molecules [15]. The study of external interactions in these matrix models of RNA has shown that the interactions introduced in such a way act as additional constraints on the parameters of the model. These constraints may physically imply (i) a change in the inter-base interaction strength which may happen due to presence of ions or applied external pressure or both or (ii) biasing the unpaired sites in the chain by assigning them a weight compared to paired ones.

The formalism gives a systematic analytical method for studying the effect of external interactions on the statistics and distribution of the planar and non-planar diagrams for a given L and genus. These models are matrix models of RNA which enumerate structures of all types that are possible for a given length of the chain. So the characteristics observed in the distribution functions may as well be expected to be found in some real RNAs such as the micro-RNA which are small with only 21–23 nucleotides. The results found here (scaling etc.,) are also interesting in their own right purely from the point of view of random matrix theory. The results found here may also be useful in understanding the structure and dynamics of experiments with ssDNA (single stranded DNA) because the model does not consider specificity of bases and may as well be used to understand and formulate synthetic sequences of desired use in many situations of material science and nanotechnology.

An important and possible area of extension in these hard crust problems is to solve the RNA heteropolymer within the same framework. Also, a useful study in solving RNA structure combinatorics completely will be to consider real RNA sequences, construct their base–base interaction V_{ij} and extract all the structure-related information (secondary and tertiary) from this matrix itself. This is a procedure employed in the graph theoretic approach used in complex system analysis.

Acknowledgements

The authors thank Prof. Henri Orland for his encouragement. This work was presented at Perspectives in Non-Linear Dynamics (PNLD 2010) held at Indian Institute of Science, Bangalore, India from 26 to 30 July 2010. The authors would like to thank CSIR Project No. 03(1019)/05/EMR-II for financial support.

References

[1] R Gillet and B Felden, *Mol. Microbiol.* **42**, 879 (2001)
G Ruvkun, *Science* **294**, 797 (2001)
G J Hannon, *Nature* **418**, 244 (2002)
E Masse and S Gottesman, *Proc. Natl. Acad. Sci. USA* **99**, 4620 (2002)
Y Dorsett and T Tuschl, *Nature Rev. Drug Discovery* **3**, 318 (2004)
H Tagawa and M Seto, *Leukemia* **19**, 2013 (2005)
M A Matzke and J A Birchler, *Nature Rev. Gen.* **6**, 24 (2005)
L Jaeger and A Chworos, *Curr. Opin. Struct. Biol.* **16**, 531 (2006)
K Whalley, *Nature Rev. Gen.* **7**, 331 (2006)
J Haasnoot and B Berkhout, *Handb. Exp. Pharmacol.* **173**, 117 (2006)

J Haasnoot, E M Westerhout and B Berkhout, *Nat. Biotechnol.* **25**, 1435 (2007)
P A Sharp, *Cell* **136**, 577 (2009)
G Zemora and C Waldsich, *RNA Biology* **7**, 1 (2010)

[2] R Blossey, *Computational biology: A statistical mechanics perspective* (Chapman & Hall/CRC, London, 2006) Chapter 3

[3] I Tinoco, P N Borer, B Dengler, M D Levine, O C Uhlenbeck, D M Crothers and J Gralla, *Nature New Biol.* **246**, 40 (1973)
R Nussinov and A B Jacobson, *Proc. Natl. Acad. Sci. USA* **77**, 6309 (1980)
M Zuker and P Stiegler, *Nucl. Acids Res.* **9**, 133 (1981)
R Lyngso, M Zuker and C Pedersen, *Bioinformatics* **15**, 440 (1999)
D H Mathews, J Sabina, M Zuker and D H Turner, *J. Mol. Biol.* **288**, 911 (1999)

[4] J Diaz, J Karhumki, A Lepist and D Sannella (eds), *Proceedings of the 31st International Colloquium on Automata, Languages and Programming* (ICALP) Vol. **3142** of *Lecture Notes in Computer Science*, pp. 919–931 (Springer, Berlin/Heidelberg, 2004), in the chapter Complexity of pseudoknot prediction in simple models by R B Lyngso

[5] B A Deiman and C W Pleij, *Virology* **8**, 166 (1997)
F H van Batenburg, A P Gultyaev and C W Pleij, *Nucl. Acids Res.* **29**, 194 (2001)
K Han and Y Byun, *Nucl. Acids Res.* **31**, 3432 (2003)
A Condon, B Davy, B Rastegari, S Zhao and F Tarrant, *Theor. Comput. Sci.* **320**, 35 (2004)
D W Staple and S E Butcher, *PLoS (Public Library of Science) Biol.* **3**, e213 (2005)
B Rastegari and A Condon, *J. Comput. Biol.* **14**, 16 (2007)

[6] J P Abrahams, M van den Berg, E van Batenburg and C Pleij, *Nucl. Acids Res.* **18**, 3035 (1990)
E Rivas and S R Eddy, *J. Mol. Biol.* **285**, 2053 (1999)
Y Uemura, A Hasegawa, Y Kobayashi and T Yokomori, *Theor. Comp. Sci.* **210**, 277 (1999)
T Akutsu, *Disc. Appl. Math.* **104**, 45 (2000)
R B Lyngso and C N Pedersen, *J. Comput. Biol.* **7**, 409 (2000)
R M Dirks and N A Pierce, *J. Comput. Chem.* **24**, 1664 (2003)
A Xayaphoummine, T Bucher, F Thalmann and H Isambert, *Proc. Natl. Acad. Sci. USA* **100**, 15310 (2003)
J Reeder and R Giegerich, *BMC Bioinformatics* **5**, 104 (2004)

[7] C Haslinger and P F Stadler, *Bull. Math. Biol.* **61**, 437 (1999)

[8] H Orland and A Zee, *Nucl. Phys.* **B620[FS]**, 456 (2002)
G Vernizzi, H Orland and A Zee, *Phys. Rev. Lett.* **94**, 168103 (2005)

[9] I Garg and N Deo, *Pramana – J. Phys.* **73**, 533 (2009)

[10] I Garg and N Deo, *Phys. Rev.* **E79**, 061903 (2009)

[11] I Garg and N Deo, *Eur. Phys. J.* **E33**, 359 (2010)

[12] G ’t Hooft, *Nucl. Phys.* **B72**, 461 (1974)

[13] E P Wigner, *SIAM Rev.* **9**, 1 (1967)
D J Gross and A A Migdal, *Phys. Rev. Lett.* **64**, 127 (1990)
M L Mehta, *Random matrices and the statistical theory of energy levels*, 2nd ed. (Academic, New York, 1991)
A Muller-Groeling and H A Weidenmuller, *Phys. Rep.* **299**, 189 (1998)
J J Verbaarschot and T Wettig, *Annu. Rev. Nucl. Part. Sci.* **50**, 343 (2000)
P Bleher and A Its, *Random matrices and their applications* (Cambridge University Press, London, 2001)
P J Forrester, N C Snaith and J J M Verbaarschot, *J. Phys.* **A36**, R1 (2003)

[14] R B Macgregor Jr, *Nucl. Acid. Sci.* **48**, 253 (1998)
R B Macgregor Jr, *Biochemica et Biophysica Acta* **1595**, 266 (2002)
M Giel-Pietraszuk and J Barciszewski, *Int. J. Biol. Macromol.* **37**, 109 (2005)

[15] M Bon, G Vernizzi, H Orland and A Zee, *J. Mol. Biol*. **379**, 900 (2008)