New Alphabet-Dependent Morphological Transition in Random RNA Alignment

O. V. Valba,[1,3] M. V. Tamm,[2] and S. K. Nechaev[3,4]

[1]*Moscow Institute of Physics and Technology, 141700, Dolgoprudny, Russia*
[2]*Physics Department, Moscow State University, 119992, Moscow, Russia*
[3]*LPTMS, Université Paris Sud, 91405 Orsay Cedex, France*
[4]*P.N. Lebedev Physical Institute of the Russian Academy of Sciences, 119991, Moscow, Russia*

We study the fraction f of nucleotides involved in the formation of a cactuslike secondary structure of random heteropolymer RNA-like molecules. In the low-temperature limit, we study this fraction as a function of the number c of different nucleotide species. We show, that with changing c, the secondary structures of random RNAs undergo a morphological transition: $f(c) \rightarrow 1$ for $c \leq c_{cr}$ as the chain length n goes to infinity, signaling the formation of a virtually perfect gapless secondary structure; while $f(c) < 1$ for $c > c_{cr}$, which means that a nonperfect structure with gaps is formed. The strict upper and lower bounds $2 \leq c_{cr} \leq 4$ are proven, and the numerical evidence for c_{cr} is presented. The relevance of the transition from the evolutional point of view is discussed.

PACS numbers: 87.14.gn, 87.15.bd, 87.23.Kg

Genetic information in all life cells is kept within the primary sequences of DNA and RNA molecules. Both of them are heteropolymers consisting of four different nucleotide types. Why does nature use exactly four aminoacid bases? Could one find any properties of systems containing DNAs or RNAs sensitive to the number of different "letters" (i.e., different nucleotide types) used in construction of these heteropolymers? Typically, the attempts to answer this question are based on the chemistry of interacting nucleotides [1], or deal with the conjectures lying in the general information theory [2]. Here, we present a statistical observation concerning the dependence of the RNA secondary structures on the number of nucleotide types (alphabet size) c, which, to the best of our knowledge, was never discussed before. (However, see the Note added in proof.)

We demonstrate the existence of a morphological transition in the statistics of the secondary structure of a random RNA-like chain as a function of the alphabet size c. Namely, for small c, $c \leq c_{cr}$ chains which are long enough can form a perfect secondary structure, i.e., a structure in which the fraction of paired nucleotides (i.e., connected to the complementary ones via hydrogen bonds) approaches one as the chain length goes to infinity, while for $c > c_{cr}$, even the best possible secondary structure includes a finite fraction of gaps (i.e., nucleotides that have nobody to connect with).

Note that this problem belongs to the class of satisfiability ones, such as the celebrated k-satisfiabilty problem [3–5]. Indeed, we are looking for a transition from a situation when some problem (in our case, a search for a perfect secondary structure) is almost surely solvable for any random initial conditions (nucleotide sequences) to the situation when it is almost surely unsolvable. This transition occurs with a change of alphabet size, i.e., c plays a role analogous to M/N (a ratio of the number of equations to the number of variables) in k satisfiabilty.

As for the particular value of the critical alphabet size at which the transition occurs, we prove that it lies in the interval $c_{cr}^{min} \leq c_{cr} \leq c_{cr}^{max}$, where the lower, $c_{cr}^{min} = 2$, and the upper, $c_{cr}^{max} = 4$, bounds can be computed exactly. The numerical estimates of c_{cr} are rather restrictive since c takes integer values only. However, we argue below that with some minor modification, the problem under consideration can be naturally generalized to noninteger c. Numerical evaluation of this generalized problem leads to the value $c_{cr}^{num} \approx 2.7$.

The prediction of RNAs' secondary structure deals with a search for the structure with the lowest value of the free energy among all allowed cactuslike structures. Numerous dynamic programming algorithms are developed to that end [6]. In the simplest possible case, one supposes that a given chain consists of n monomer units, each unit chosen from a set of c different types (letters) A, B, C, D,. ... These units can form noncovalent bonds with each other, at most one bond per unit. The energy of a bond depends on which letters are bonded; for example, one can assign an attraction energy u to the bonds between similar letters (A-A, B-B, etc., we call them "matches") and zero energy to the bonds between different letters (A-B, A-D, etc., "mismatches") [7]. The topology of secondary structures is supposed to be "cactuslike" (see Fig. 1), i.e., hierarchically folded and topologically isomorphic to a tree (we suppose here that structures which are not treelike, known as "pseudoknots," are suppressed). To simplify the model as much as possible, we do not allow here for any constraints on the minimal size of loops in the structure, the variation in the energies of different types of matchings, nor for the contribution of loop factors to the partition function, or the stacking interactions (the cooperativity in the formation of bonds between adjacent pairs of monomers). Despite these essential simplifications, the
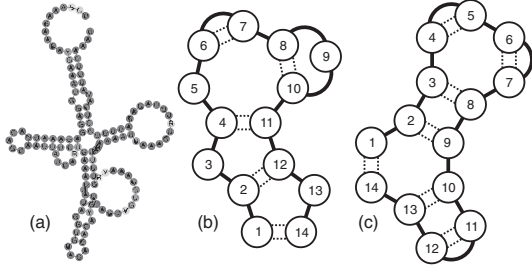
FIG. 1.   (a) Secondary structure of an RNA gene *HAR1F* (see, for example, [23]); (b), (c) Schematic cactuslike structures of RNA-like chains represented in Fig. 2 by Motzkin and Dyck paths, respectively.

considered model is known to be a common "firing ground" for the theoretical consideration of secondary structures formed in the ensemble of messenger RNAs [8].

The partition function of the random RNA-type heteropolymer is known (see, for example, [9] in the context of matching models) to satisfy the recursion:

$$g_{1,n} = 1 + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \beta_{i,j} g_{i+1,j-1} g_{j+1,n};$$

$$g_{i,i} = 1. \tag{1}$$

These equations (the analogues of which are used all over the place in the RNA-folding theory, see, for example, [10–13]) generate the hierarchical cactuslike RNA topology. The term $g_{i,j}$ describes the statistical weight of the part of the sequence between monomers i and j. The Boltzmann weights $\beta_{i,j}$ ($1 \le i \le j \le n$) are the statistical weights of bonds: $\beta_{i,j} = e^{u/T}$ if i and j match, and $\beta_{i,j} = 1$ otherwise. Now, the energy of the ground state is just a limiting value of the free energy as the temperature approaches zero: $E_{1:n} = \lim_{T \to 0} T \ln g_{1:n}$. After some algebra (see [9] for details) one can reduce the expression for E to the following form:

$$E_{i,i+k} = \lim_{T \to 0} T \ln g_{i,i+k} = \max_{s=i+1,\ldots,i+k} \{E_{i+1,i+k}, \\ [\varepsilon_{i,s} + E_{i+1,s-1} + E_{s+1,i+k}]\}, \tag{2}$$

where $\varepsilon_{i,j}$ is the interaction energy between monomers i and j; it equals u if these monomers match and 0 if they do not.

For random RNA-type sequence made of c letters, the average energy of a ground state, $E_{1,n}$, for $n \gg 1$ behaves as $\langle E \rangle \simeq \frac{un}{2} f(c)$, where $f(c)$ is the fraction of nucleotides that have formed bonds. We argue that for c less than a certain value c_{cr} there exists a perfect match, i.e., $f(c) = 1$, and the fraction of connected monomers converges to unity, while for $c \le c_{cr}$ a finite fraction of monomers $1 - f(c)$ remains unmatched even in the best match, i.e., $f(c) < 1$. Below we compute the exact lower and upper bounds for c_{cr}, derive the upper bound for $f(c)$ in the $c \ge c_{cr}$ region, and discuss the numerical evidence of our conjecture.

Consider $c_{cr}^{min} = 2$. It turns out that matching with $f(c = 2) \to 1$ as $n \to \infty$ is possible not only on average but also for any given primary structure. Indeed, consider a random heteropolymer RNA constituted of A and B monomers, forming saturating bonds of type A-A and B-B and construct the optimal structure as follows. Take the left end of the chain as a starting point, and move along a sequence until meeting the first pair of two sequential letters AA or BB. Connect these two letters with a bond and erase them from the sequence. Iterating this procedure, one arrives finally to an alternating sequence of the type ABAB... (assume, with no loss of generality, that the starting letter is A). Connect now the first letter A from this sequence to the last one, the next B to the B before the last A, etc. It is clear that this algorithm results in a nested secondary structure which leaves unmatched at most two letters (one in the middle of the ABAB... sequence and, possibly, another one in the very end). The fraction of mismatched letters decreases as n^{-1} with n, proving the conjecture. A similar algorithm for alternating (A-B) bonding can be easily constructed (though the fraction of mismatches decreases as $n^{-1/2}$ in this case). Note that this lower bound is already nontrivial: in the celebrated "longest common subsequence" problem, whose results are used for the comparison of two *linear* DNA sequences, the fraction of matches equals (see, for example, [14]) $f_{lin}(c = 2) = 2(\sqrt{2} - 1) < 1$, and the "critical" alphabet size, at which $f_{lin} = 1$, is $c_{cr}^{lin} = 1$.

To construct the upper bound for c_{cr}, recall the one-to-one mapping between cactuslike RNA secondary structures and discretized Brownian excursions, known as Motzkin paths [15]. Under this mapping, shown in Fig. 2, the gapless (perfect) secondary structures correspond to excursions with no horizontal steps, the Dyck paths. The total number $D(n)$ of Dyck paths of even length n is given by a Catalan number $C_{n/2}$:

$$D(n) = C_{n/2} \equiv \frac{\Gamma(n+1)}{\Gamma(\frac{n}{2}+1)\Gamma(\frac{n}{2}+2)} \sim \frac{2^n}{n^{3/2}}, \tag{3}$$

where $\Gamma[n]$ is the Γ function and the asymptotic expression is valid for $n \gg 1$.

Consider a set of random sequences of length n. Each of these sequences (there are c^n of them) must correspond to a certain perfect match, i.e., a Dyck path. Meanwhile, if one
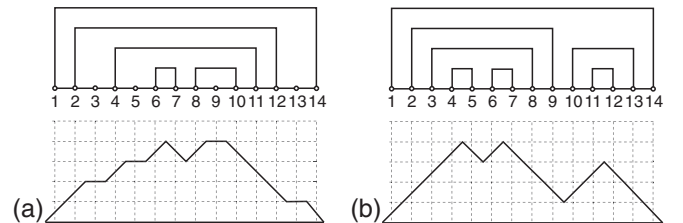

FIG. 2.   The secondary structure of RNAs: (a) With gaps represented by Motzkin path; (b) Gapless represented by Dyck path.

particular Dyck path corresponds to a perfect match of some particular sequence, it simultaneously corresponds to perfect matches of many others. Indeed, each "up-down" pair of steps in a Dyck path can be realized in c different ways (A-A, B-B, etc.) independently of all others, leading to a degeneracy of order $c^{n/2}$. Thus, the number of different primary sequences that can have perfect secondary structures is at most

$$W(c, n) = c^{n/2}D(n) \sim \frac{(2\sqrt{c})^n}{n^{3/2}}. \qquad (4)$$

One primary sequence can be represented by several Dyck paths. Therefore, (4) is an estimate from above. Comparing the value $W(c, n)$ to the total number of primary sequences, $W_0(c, n) = c^n$, we have for $n \gg 1$:

$$\lim_{n\to\infty} \frac{\ln W(c, n)}{n} > \lim_{n\to\infty} \frac{\ln W_0(n)}{n} \quad \text{for } c < c_{cr}^{max} = 4,$$

$$\lim_{n\to\infty} \frac{\ln W(c, n)}{n} < \lim_{n\to\infty} \frac{\ln W_0(n)}{n} \quad \text{for } c > c_{cr}^{max}. \qquad (5)$$

One can follow this reasoning to develop the upper bound also for $f(c)$ at $c > c_{cr}^{max} = 4$. In this case, the fraction of random primary sequences admitting a perfect match among all $W_0(c, n)$ of them is exponentially small. Therefore, the ground states of almost all sequences should correspond to matchings with gaps, i.e., to Motzkin paths. The Motzkin paths with a finite fraction of gaps (horizontal steps) produce much more possibilities for the RNA ground states than Dyck paths of the same length. The number of n-step Motzkin paths with m gaps is $M(m, n) = \frac{n!}{m!(n-m)!}D(n - m)$ and

$$\frac{\ln M(f, n)}{n} = -(1 - f)\ln(1 - f) - f\ln\frac{f}{4} + o\left(\frac{\ln n}{n}\right), \qquad (6)$$

where $f = \frac{n-m}{n}$ and (6) works for $n \gg 1$ and even $m - n$.

How many different primary structures can have a given Motzkin path as a ground state? Each pair of "up-down" steps is bound to belong to the same species, as for the Dyck paths, while each horizontal step can be chosen independently. The total degeneracy Z is thus

$$Z(c, n, f) = c^{(fn)/2}c^{(1-f)n} = c^{n(2-f)/2}. \qquad (7)$$

As f decreases, the total number of structures that can have ground states with the fraction of matches more than f increases and is given by

$$W(c, n, f) = \sum_{j=0}^{(1-f)n} Z(c, n, j/n)M(j, n). \qquad (8)$$

At some $\bar{f}$ it becomes equal to the total number of possible primary structures $W_0(c, n) = c^n$, giving the estimate for the typical value of $f(c)$. For $n \gg 1$ the sum in (8) can be evaluated up to the leading order using the saddle-point approximation. One has for $\Delta w(f, c) = \lim_{n\to\infty} \frac{1}{n} \ln\frac{W(c,n,f)}{W_0(c,n)}$:

$$\Delta w(f, c) = \begin{cases} -f\ln\frac{\sqrt{c}f}{2} - (1 - f)\ln(1 - f); & f < f_m \\ \ln\left(1 + \frac{\sqrt{c}}{2}\right) > 0; & f < f_m, \end{cases} \qquad (9)$$

where $f_m = \frac{2}{2+\sqrt{c}}$. For $f < f_m$ the sum in (8) is dominated by the contribution from the upper boundary, while for $f < f_m$ it is given by the maximum at f_m and is, therefore, independent of the upper summation limit. The desired value of $\bar{f}(c)$ is defined by the solution of the equation $\Delta w(f, c) = 0$ and is plotted in Fig. 3 with a dotted line [16].

We have analyzed numerically the statistical properties of the ground state free energy $E_n(c)$, applying (2) to the random sequences with different numbers of letters (nucleotide types) c. In Fig. 3 we show the numeric results for the average value of $f(c)$ for $c = 2, \ldots, 8$ for sequences of finite length $n = 200$, as well as the limiting values extrapolated to $n \to \infty$. The experimental values of $f(c)$ lay lower than the upper bound given by the theory. To go beyond the integer values of c one can use the analogue with the linear Bernoulli matching problem [17] and replace the correlated matrix $\varepsilon_{i,j}$ with the uncorrelated random matrix, whose entries are independent randomly distributed variables taking the value u with the probability c^{-1} and 0 otherwise. The numerical results seem to show that this simplified model belongs to the same universality class, and the change in $f(c)$ due to the removal of correlations in the adjacency matrix is lower than 1%. Moreover, in the Bernoulli case, the generalization to the noninteger values of c is straightforward and the numerical simulations show that the transition from a perfect to nonperfect match occurs at $c \approx 2.7$. More details on the Bernoulli RNA-like matching will be provided elsewhere [18].
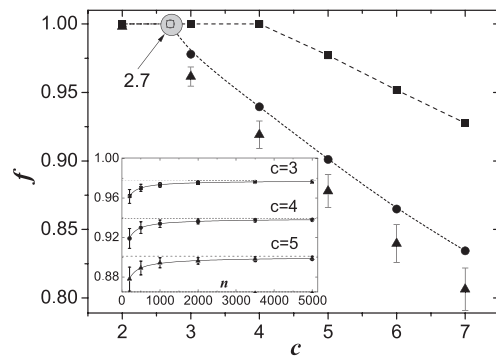

FIG. 3. The average fraction of connected nucleotides in the ground state secondary structure, f, as a function of the number of nucleotide types, c. ▲, the numeric dependence obtained as an average value for 500 randomly generated sequences of length $n = 200$ each; ●, the approximation to chains of infinite length; ■, upper analytic estimation (9) (the line connecting the points are a guide for the eyes); the inset shows the finite-size scaling used to obtain the data at infinite lengths.

Summing up, we demonstrate here that alphabets with different number of letters c are nonequivalent if one considers the matching problem of long random RNA. This nonequivalence is tightly coupled to the restrictions on the morphology of allowed secondary structures. Indeed, the existence of two regimes (for $c \leq c_{cr}$ and $c > c_{cr}$) is a peculiarity of RNAs and is due to the additional freedom in the formation of the complex cactuslike secondary structures typical for messenger RNAs. For the linear matching problem used in the DNA comparison, the fraction of nucleotides in the optimal alignment is less than 1 for any alphabet with $c > 1$. In our model, the transition between two regimes occurs at $2 < c_{cr} < 4$. The exact value of the critical alphabet size should be sensitive to the microscopic details of the model, and one can enumerate factors which are neglected in our model and which could shift the transition point to the right or to the left from the observed critical value. On the one hand, the presence of stacking energies and minimal loop sizes in real RNA leads to the bonds being effectively formed not by single nucleotides, but by blocks of them, increasing the effective alphabet size for given c, thus, decreasing c_{cr} in terms of the size of a "bare" alphabet. On the other hand, one would not expect any real-life random RNA to have a completely random structure with exactly equal concentrations of letters and no short-range correlations between them. Any such correlations reduce the information entropy of the sequence, and, therefore, lead to the decrease of the effective alphabet size, and thus push c_{cr} to higher values. The exact value of c_{cr} is nonuniversal. However, our analysis shows (i) the existence of two different morphological regimes, depending on the number of nucleotide types in the alphabet; and (ii) the fact that this transition point can plausibly be rather close to 4 (see Note added in proof).

This particular number, obviously, sounds suggestive since it is exactly the number of nucleotide types in the alphabet used in real-world RNAs. The criticality on alphabet size, observed only for RNAs thus nicely rhymes with the modern opinion that the life originates from the template-directed replication of random RNA molecules (the so-called "RNA world" hypothesis) [19,20]. Can it be, indeed, advantageous to have the alphabet of critical or close-to-critical size? For RNA to have a biological function it should (i) fold predictably and (ii) form a robust structure not too sensitive to thermal noise. Short nucleotide alphabets with $c < c_{cr}$ tend to produce structures which have many different ground states [see (5), also compare with similar reasoning for proteins [21,22]]. On the other hand, long alphabets correspond to loosely bound ground states with many unpaired nucleotides, which is disadvantageous in terms of the stability of the structure. The critical alphabets, thus, seem to be optimal for biological purpose.

Note added in proof.—Recently, Ref. [24] was brought to our attention. The authors of Ref. [24] have considered numerically the secondary structures of RNAs in a model somewhat different to ours, and have observed the qualitative difference in the degeneration of RNA ground states for two- and four-letter alphabets. We identify this result with the transition described here and see it as a confirmation of the robustness of the transition with respect to the concrete details of the model.

[1] E. Szathmary, Nat. Rev. Genet. **4**, 995 (2003).
[2] A. Patel, J. Biosci. **26**, 145 (2001).
[3] S. Kirkpatrick and B. Selman, Science **264**, 1297 (1994).
[4] R. Monasson, R. Zeccina, S. Kirkpatrick, B. Selman, and L. Troyansky, Nature (London) **400**, 133 (1999).
[5] M. Mézard, G. Parisi, and R. Zecchina, Science **297**, 812 (2002).
[6] See, for example, http://rna.tbi.univie.ac.at/, http://9mfold.rit.albany.edu/, and http://kinefold.curie.fr.
[7] In real RNAs, matches are the interactions between *complementary* nucleotides rather than *similar* ones, which give rise to a slightly different matrix of interactions. However, at least for random RNAs, this difference is irrelevant: it is important that the fraction of possible matches is $\frac{1}{c}$, the rest corresponding to mismatches.
[8] R. Bundschuh and T. Hwa, Phys. Rev. E **65**, 031903 (2002).
[9] S. K. Nechaev, M. V. Tamm, and O. V. Valba, J. Phys. A **44**, 195001 (2011).
[10] R. Bundschuh and T. Hwa, Phys. Rev. Lett. **83**, 1479 (1999).
[11] M. Müller, Phys. Rev. E **67**, 021914 (2003).
[12] M. Müller, F. Krzakala, and M. Mezard, Eur. Phys. J. E **9**, 67 (2002).
[13] M. V. Tamm and S. K. Nechaev, Phys. Rev. E **75**, 031904 (2007).
[14] S. N. Majumdar and S. Nechaev, Phys. Rev. E **72**, 020901 (2005).
[15] S. K. Lando *Lectures on Generating Functions*, (AMS Providence, RI, 2003).
[16] It may seem that $\bar{f}$ is an estimate for the "typical smallest," not an average value of f. However, since $\bar{f} > f_m$, it belongs to the regions where the sum in (8) is dominated by the upper bound and thus the average and "typical largest" values of f converge in the thermodynamic limit.
[17] J. Boutet de Monvel, Eur. Phys. J. B **7**, 293 (1999); Phys. Rev. E **62**, 204 (2000).
[18] O. Valba, M. Tamm, and S. Nechaev (to be published).

[19] W. Gilbert, Nature (London) **319**, 618 (1986).
[20] G. F. Joyce, Nature (London) **338**, 217 (1989).
[21] T. M. A. Fink and R. C. Ball, Phys. Rev. Lett. **87**, 198103 (2001).
[22] A. Yu. Grosberg and A. R. Khokhlov, *Giant Molecules: Here, There and Everywhere* (World Scientific Publishing, Hackensack, NJ, 2011), 2nd ed..
[23] K. S. Pollard, S. R. Salama, N. Lambert, M.-A. Lambot, S. Coppens J. S. Pedersen, S. Katzman, B. King, C. Onodera, A. Siepel *et al.*, Nature (London) **443**, 167 (2006).
[24] R. Mukhopadhyay, E. Emberly, C. Tang, and N. S. Wingreen, Phys. Rev. E **68**, 041904 (2003).