The end-to-end distance of RNA as a randomly self-paired polymer

Li Tai Fang

Mt. Zion Campus, Box 1351, University of California, San Francisco, CA 94115, USA

1. Introduction

Determining the end-to-end distance of a linear polymer is a classic problem in statistical physics. Neglecting excluded volume interactions, specifically for ideal linear polymers, leads to the well-known result that the root-mean-square end-to-end distance, \( d \), scales as the 1/2 power of the number of monomers, \( N \) (i.e., \( d \propto N^{1/2} \)). For ideal randomly branched polymers, the situation is considerably more complicated, with \( d \) increasing more slowly with \( N \), i.e., as the 1/4 power of \( N \) (Redner, 1979; Zimm and Stockmayer, 1949; de Gennes, 1968; Gutin et al., 1993). In general, however, the scaling depends on the nature of the branching and of the monomer interaction (Redner, 1979).

Single-stranded RNA is a linear polymer that presents special statistical physics challenges because of the difficult-to-characterize branching that develops as a consequence of its self-complementarity. More explicitly, hydrogen-bonding between complementary pairs of nucleotides (G–C, A–U, and G–U) and base-stacking between the adjacent base pairs lead to the formation of duplex portions of RNA separated by single-stranded loops, as shown in Fig. 1 for an arbitrary 115-nucleotide RNA.

Recently, there has been considerable interest regarding the close proximity of the 5' and 3' ends of some RNAs. In fact, the close proximity of the ends of many viral RNAs, specifically the apparent “circularization” of these RNAs, is required for the efficient replication of the viral RNAs (Hsu et al., 1987; Frey et al., 1979; Ooms et al., 2007; Karetnikov and Lehto, 2008; Fabian and White, 2004). In some viruses, there are unusually long complementary sequences strategically located at the ends of the RNA to ensure the robustness of this proximity (Hsu et al., 1987; Frey et al., 1979; Ooms et al., 2007). In addition, the “circularization” of messenger RNA, achieved by protein binding, is also required for translation (Gallie, 1991). Questions arise whether the close proximity of the ends of these RNAs is a result of evolutionary pressure, or a property of RNA molecules in general.

As demonstrated in the recent work of Yoffe et al. (2011), this proximity of the two ends of a RNA molecule is a general property, remarkably independent of its length or sequence. Yoffe et al. have provided a systematic computational check of this result for a wide range of lengths and sequences (both biological and random), and have presented a statistical mechanical theory for random sequences of arbitrary lengths. Their analysis includes an estimate of the universal proximity of the two ends and an explanation of its invariance with respect to length and sequence. Biologically, the close proximity of the RNA’s ends need not be a result of evolutionary pressure. Rather, due to the statistical likelihood of this close proximity, evolved mechanisms may be necessary to take the ends of an RNA apart should there be a need to suppress such biological processes.

In the present paper, we propose the randomly self-paired polymer (RSPP) model to explain the proximity of the ends of an RNA. We define a randomly self-paired polymer as a linear polymer each of whose monomers has a probability, \( f \) (\( 0 < f < 1 \)), of pairing with any other one monomer. The RSPP model is relevant to linear polymers whose monomers associate with one another either because of direct interaction between the monomers (as in the case of RNA) or through an additional component that simultaneously binds to two monomers (e.g., RNA in the presence of proteins, where each protein has two RNA binding sites).

Fig. 2 shows a circle diagram representation of the secondary structure (i.e., set of base pairings) depicted in Fig. 1. After examining a large number of circle diagrams associated with the secondary structures of many RNA molecules (of different lengths and
2. Defining the randomly self-paired polymer model

The particular model we consider is the randomly self-paired polymer (RSPP). In this polymer (→→ RNA), each monomer (→ nucleotide or base) has equal probability of pairing with any other one monomer, creating a “monomer-pair.” One restriction imposed to our RSPP which makes it distinct from many other self-attracting polymers is that the monomer of an RSPP can pair with only one other monomer or none at all (Nieuwenhuizen, 1989; Meirovitch, 2002). Consider a case where monomer i pairs with monomer j, and monomer k pairs with monomer l, then j = l if and only if i = k. The pairings are allowed to cross (i.e., creating pseudoknots, kissing hairpins, etc.), as opposed to numerous RNA secondary structure prediction algorithms that prohibit pseudoknot formation to reduce computational time, justified by the rarity of pseudoknots (Hofacker et al., 1994; Zuker, 2003; Nussinov and Jacobson, 1980). Another property that makes RSPPs different from other self-attracting polymer is that the pairing fraction (i.e., the number of paired monomers in each polymer) is fixed. Each configuration in the ensemble corresponds to a unique set of monomer pairs, and each configuration is of equal statistical weight. Thus, there is no rule whatsoever that favors one pairing over another. We are not concerned with how or why the monomers pair with one another, but rather only with the fact that a fixed fraction of them does. Therefore, for an RSPP of N total monomers, of which \( N_p \) monomers are paired, the size of the ensemble, \( \Omega_f \) (i.e., total number of possible structures), is

\[
\Omega_f(N, N_p) = \frac{N!}{(N - N_p)! \cdot \left(\frac{N_p}{2}\right)!},
\]

because the number of ways to “choose” \( N_p \) paired monomers from \( N \) total monomers is \( \binom{N}{N_p} \). The number of ways to make the pairs after determining the positions of the paired monomers is \( \frac{N_p!}{\left(\frac{N_p}{2}\right)! \cdot (N_p/2)!} \).

For a simple calculation that illustrates our point, consider an RSPP of \( N = 1000 \), of which 60% of them are paired. Thus, for any given monomer, there is a 0.6 chance that it is involved in pairing. Out of the remaining 999 monomers, it has equal probability (1/999) of pairing with any one of them. It follows that 0.6 × 1/999 = 6 × 10^{-4} is the probability of any particular pair. We can calculate the probability that at least one pair of monomers will exist that connects one end of the polymer to the other, effectively creating a bridge that brings the ends to no more than, say, 50 unpaired monomers apart. We define a bridge as a monomer pair responsible for bringing the ends closer together. A monomer pair that plays no role in bringing in the ends closer is not considered a bridge. The following pairs all satisfy this condition: the 1st monomer pairs with the 950th, 951st, 952nd, ... or 1000th monomer; the 2nd monomer pairs with the 951st, 952nd, 953rd, ... or 1000th monomer; ... and at last, the 51st monomer pairs with the 1000th monomer. The total number of ways to create such a bridge is 51 + 50 + 49 + ... + 1 = 1326. Thus, the probability that at least one such bridge will form is 1 - (1 - 6 × 10^{-4})^{1326} = 0.55. We will use this logic to calculate the mean end-to-end distance of RSPP.

Note that 0.55 is the probability that the two ends will be separated by not more than 50 unpaired monomers due to a single bridge. The ends can be brought close without such a bridge if there are enough other bridges in the middle that bring the ends together through a chain (Fig. 3). Thus, 0.55 is a lower bound on the probability of finding the ends within no more than 50 unpaired monomers of one another.

2.1. Mapping RNA onto the model

One common characteristic of RNA secondary structures, as evident in Figs. 1 and 2, is that no lone base pair is present in the predicted structure, because the hydrogen bonds between the base pairs are insufficient to overcome the entropic penalty of duplex formation. Base stacking energies (i.e., hydrophobic interactions and enhancement of van der Waals interactions caused by
3. Results

3.1. Defining the end-to-end distance

We define the end-to-end distance, $X$, as the number of unpaired monomers (or nucleotides) associated with the shortest linear path between the ends after the RSPP (or RNA) is folded into a secondary structure (Figs. 3 and 4). It is important to stress that $X$ is not the physical distance between the first and the last monomers (i.e., $X \neq |r_1 - r_N|$), where $r_1$ and $r_N$ represent the spacial positions of the first and the last monomers. Issues regarding the stiffness of the chain, the rigidity of the base pairing bonds, and the excluded volume affecting the physical distance between the ends are beyond the scope of this work.

3.2. Probability distribution of the end-to-end distances

The derivations of the equations presented in this section are worked out in detail in Appendix A. In this section, we only discuss the results and their consequences. Following the logic presented in the previous section, the probability, $P$, that the end-to-end distance of a RSPP is less than or equal to $X$ unpaired monomers, given the effective total number of monomers, $N_{T,\text{eff}}$, and effective paired number of monomers, $N_{p,\text{eff}}$, is

$$P(N_{T,\text{eff}}, N_{p,\text{eff}}, X) = 1 - \prod_{i=1}^{N_{T,\text{eff}}/2} (1-p)^{(i-1)}.$$  

where $p = [1/(2^{N_{\text{unp}}}) - 2n]/N_{T,\text{eff}}$ is the probability of any particular set of bridges occurring in the RSPP. $B$ is the number of bridges comprising such a set. $Q = 2^{B-1} \sum_{X} X^{B} \cdot e^{-X}$ is the number of such sets. (Again, the derivations for $P$, $p$, and $Q$ are presented in Appendix A.) It follows that the probability, $\rho(X)$, of the end-to-end distance being exactly $X$ is

$$\rho(N_{T,\text{eff}}, N_{p,\text{eff}}, X) = P(N_{T,\text{eff}}, N_{p,\text{eff}}, X) - P(N_{T,\text{eff}}, N_{p,\text{eff}}, X-1),$$  

from which the probability distribution of the end-to-end distances can be calculated and plotted, as shown in Fig. 5 for an RNA of length $N=1000$ and pairing fraction $f=0.6$ mapped as an RSPP (i.e., $N_{T,\text{eff}}=550, N_{p,\text{eff}}=0.27$). The probability distribution of the end-to-end distances appears roughly Gaussian. The mean end-to-end distance, $\langle X \rangle$, is 14.4 unpaired bases of separation, with the width of the distribution being approximately 5. As evident in Fig. 5, from a statistical view, it is highly unlikely (less than 0.5% chance) to find the ends of a 1000-nucleotide long RNA more than 25 unpaired bases apart. Thus, to find the $5'$ and $3'$ of an RNA being in close proximity is a statistical inevitability.

3.3. End-to-end distance vs. sequence length

The mean end-to-end distance, $\langle X \rangle$, given the probability distribution, $\rho(X)$, is calculate as follows:

$$\langle X(N_{T,\text{eff}}, N_{p,\text{eff}}) \rangle = \sum_{X=0}^{N_{\text{unp}}} \rho(N_{T,\text{eff}}, N_{p,\text{eff}}, X) \cdot X.$$  

A plot of $\langle X \rangle$ vs. $N$ for RNAs with pairing fraction, $f$, fixed at 0.6 (again, $f_{\text{eff}}=0.27$) is shown in Fig. 6. From Fig. 6, it appears that $\langle X \rangle$ increases weakly with $N$. Recall that the $\langle X \rangle$ of an RNA with
Fig. 6. Predicted values of end-to-end distances, $\langle X \rangle$, for a pairing fraction of 0.6 ($f_{\text{eff}}=0.27$), as a function of total chain length, $N$.

Fig. 7. The log-log plot of the mean end-to-end distance vs. total length. The range of $N$ in this plot is from 50 to 10,000. From top to bottom, the plots are for $f=0.4$, 0.6, and 0.7 (or $f_{\text{eff}}=0.12$, 0.27, and 0.37). The linear regression values, $R^2$, are in the 0.998 range for all three. The slopes on the log-log plots are in the range of 0.249–0.255 for all three.

Fig. 8. Predicted values of end-to-end distances, $\langle X \rangle$, for RNA of $N=1000$, as a function of pairing fraction, $f$.

$N=1000$ and $f=0.6$ is predicted to be 14.4 unpaired bases apart; doubling the total length ($N=2000$) yields $\langle X \rangle = 17.1$, merely 19% greater. For an RNA with a sequence length of an order of magnitude greater ($N=10,000$), $\langle X \rangle = 24.9$.

Eq. (4) provides a simple approximate relationship between $\langle X \rangle$ and $N$, which roughly follows $\langle X \rangle \propto N^{1/4}$ for large $N$ (i.e., $N \geq 50$ when $f=0.6$), when the pairing fraction is held constant (Fig. 7). The exponent 1/4 is found to be invariant with RNA pairing fractions. Fig. 7 shows the log-log plots of RNAs with “realistic” pairing fractions, with $f$ ranging from 0.4 to 0.7 (or $f_{\text{eff}}$ from 0.12 to 0.37). Fig. 7 shows that the log-log plots yield excellent linear fits with a slope of 1/4 for these values of $f$. It shows, perhaps unsurprisingly that $\langle X \rangle$ of an RSPP is considerably smaller than what would be expected for an ideal linear polymer. We emphasize again, that $\langle X \rangle$ is calculated as the shortest path going from one end of the RSPP to the other, not the physical root-mean-square distance, $d$, i.e., $\langle X \rangle \neq \sqrt{\langle f^2 N \rangle - \langle f^2 \rangle^2}$. However, if we assume ideal behavior for RSPP, i.e., assuming perfect flexibility of the vertices and neglecting excluded volume interactions, the path going from one end of the RSPP to the other can be regarded as an ideal linear chain in itself. Then, the physical root-mean-square end-to-end distance, $d_{\text{RMS}} \propto \langle X \rangle^{1/2} \propto N^{1/8}$.

Due to the discretized nature of the theory, the probability distribution of $X$ for small $N$ no longer appears Gaussian as in Fig. 5, and the $\langle X \rangle \propto N^{1/4}$ power law breaks down for small $N$ (e.g., for $N \leq 20$ when $f=0.6$). In addition, the approximation of $X \propto N$ (i.e., the assumption for Eq. (6) in A.2) is invalid for small $N$. Hence, $\langle X \rangle$ calculated from Eq. (4) deviates from this simple power law for small $N$ (which turns out to be less than what would be predicted by linear interpolation of the linear fits in Fig. 7).

3.4. End-to-end distance vs. pairing fraction

The probability distribution of end-to-end distances, $\rho(N_{\text{eff}},N_{p,\text{eff}},X)$, can also be used to calculate how $\langle X \rangle$ varies with the pairing fraction, $f$. In reality, the pairing fraction of RNA molecules, of biological or random sequences, very rarely deviates significantly from 0.6 (Yoffe et al., 2011; Fang et al., 2011). Nevertheless, the exact pairing fraction of each sequence is subject to statistics, and it is possible to find sequences which yield pairing fractions lower or higher, say, between 0.5 and 0.7. For an RNA of $N=1000$, the $\langle X \rangle$ can vary by almost a factor 2 within this range, from 19.5 for $f=0.5$ (or $f_{\text{eff}}=0.2$) to 10.5 for $f=0.7$ (or $f_{\text{eff}}=0.37$). As shown in Fig. 8. However, as emphasized in the previous section, the pairing fraction does not change the $\langle X \rangle \propto N^{1/4}$ scaling law.

4. Discussion

We have used the RSPP model to demonstrate a fundamental reason why the ends of an RNA always appear close to each other: it is a statistical inevitability that they will be brought together by at least one set of duplexes. The RSPP is a very simple and extremely approximate model to explain the end-to-end proximity of RNA molecules. Yoffe et al. (2011) have demonstrated that the end-to-end distance of RNAs approaches an asymptotic value of 12 ± 5 for RNA of $N \geq 500$. In the RSPP of $f=0.6$ (or $f_{\text{eff}}=0.27$), the expected end-to-end distance, $\langle X \rangle \approx 12.0, 14.4$, and 24.9, for $N=500, 1000$, and 10,000.

Yoffe et al. (2011) have also shown that for tymovirus, with a viral RNA of $N=6300$ and an unusually low pairing fraction of $f=0.45$, the end-to-end distance is considerably higher at 26 ± 5. The prediction of RSPP that $\langle X \rangle$ increases as $f$ decreases agrees qualitatively with that findings. For the length of tymovirus RNA ($N=6300$), the RSPP predicts that $\langle X \rangle$ increases from 22 to 35 as $f$ decreases from 0.60 to 0.45.

4.1. Summarizing the RSPP model: assumptions and limitations

There is only one basic assumption for RSPP that each monomer has equal probability of pairing with any other one monomer in the entire chain, i.e., a monomer can pair with one other monomer or none at all. This assumption of the RSPP make the model very simple and easy to understand. However, it does not provide an entirely accurate picture of RNA secondary structures.
Due to the fact that RNA base pairs very rarely cross each other, duplex formations, even for random sequences, are distinctly non-random. Every existing base pair places base pairing restrictions on the rest of the RNA molecule. In fact, the probability of base pairing in an RNA decreases asymptotically as the contour distance between them increases (David et al., 2008). In other words, bases in an RNA tend to pair with other bases nearby, and long-distance “global” base pairs are statistically rarer.

4.2. An important difference between RSPP and RNA: frequency of pseudoknots

When two RNA base pairs cross each other, a pseudoknot is created (Fig. 10b). Pseudoknots are uncommon structures in actual RNA molecules, but the assumption of equal pairing probability means that pseudoknots are common in the RSPP model. In fact, most RNA folding algorithms (e.g., mfold, Vienna RNA, etc.) simply prohibit the formation of pseudoknots for computational simplification (Hofacker et al., 1994; Zuker, 2003; Nussinov and Jacobson, 1980), whereas the overwhelming majority of structures allowed in our model contain pseudoknots. The number of allowed secondary structures in our model is given in Eq. (1). For \( N \geq 1 \), the number of possible secondary structures allowed in our model scales roughly as \( N^4 \). On the other hand, the number of allowed secondary structures in most RNA structure prediction algorithms (prohibiting pseudoknot) scales as \( \sim 1.86^N \) (Hofacker et al., 1998).

The abundance of pseudoknots in RSPP yields a qualitative difference in the scaling relationships between the \( \langle X \rangle \) and \( N \). In RNA structures where pseudoknots are prohibited, \( \langle X \rangle \) converges to an asymptotic value of \( \sim 12 \) for RNAs of \( N > 500 \) (Yoffe et al., 2011). The reason for this invariance is analyzed in detail by Yoffe et al. (2011). In RSPP, on the other hand, \( \langle X \rangle \) increases weakly with its total length, i.e., \( \langle X \rangle \propto N^{1/4} \), and \( \langle X \rangle \) does not converge to a constant value. It appears that in the “mission” to bring the ends of a polymer into close proximity, many of the crossed-over base pairs are “wasted,” as an RNA of equal pairing fraction prohibiting pseudoknots yields an end-to-end distance of exactly \( \langle X \rangle \), even smaller than that of an RSPP. Nevertheless, the point we have emphasized that the close proximity of the ends is a statistical inevitability, is still valid despite the overrepresentation of pseudoknots in the RSPP. The close proximity of the two ends is underscored by both kinds of models, which further demonstrates that the phenomenon of end-to-end proximity does not depend on details of the RNA secondary structure. It is a universal property of all linear polymers whose monomers pair singly with one another.

Acknowledgements

The work described in this paper was partially supported by the U.S. National Science Foundation (Grant CHE 0714411). The author thanks William M. Gelbart and Avinoam Ben-Shaul for their advice and support, and also Charles M. Knobler, Ajaykumar Gopal, Peter Princen, and Aron M. Yoffe for valuable discussions.

Appendix A. Deriving the equations

First of all, let

- \( N_{\text{eff}} = \text{effective total number of monomers,} \)
- \( N_{\text{p,eff}} = \text{effective number of paired monomers,} \)
- \( N_{\text{unp}} = \text{number of unpaired monomers,} \)
- \( B = \text{number of bridges involved in bringing together the ends,} \)
- \( X = \text{end-to-end distance associated with the shortest linear path between the ends.} \)

The approach we take is to first calculate the probability, \( p(X) \) that the ends of an RSPP will be less than or equal to \( X \) unpaired bases apart. The probability that the ends will be separated by a distance of exactly \( X \) is then given by \( p(X) = p(X) - p(X-1) \), or \( dp(X)/dX \), with \( X \) taking on values of 0, 1, 2, …, \( N_{\text{unp}} \).

To calculate \( p(X) \), we need to know two things:

1. the probability that a particular set of bridges will form that can bring together the ends of an RSPP to a distance of less than or equal to \( X \) unpaired monomers; and
2. the number of ways to make such sets of bridges.

A.1. The probability that a set of bridges will form

Now, consider the probability that the set of four bridges (\( B=4 \)) illustrated in Fig. 3 will form. The probability that monomer \( j \) is involved in pairing is \( N_{\text{p,eff}}/N_{\text{eff}} \). The probability that \( j \) is paired specifically with monomer \( k \) is \( 1/(N_{\text{eff}}-1) \), due to the equal probability of pairing monomer \( j \) with any of the remaining \( (N_{\text{eff}}-1) \) monomers. Hence, the probability of forming the \( j-k \) bridge is the product of these two terms. Similarly, the probability that monomer \( l \) is involved in pairing is \( (N_{\text{p,eff}}-2)/(N_{\text{eff}}-2) \). The probability that monomer \( m \) is specifically paired with monomer \( n \) is \( 1/(N_{\text{eff}}-3) \). So it follows that the probability, \( p \), that this set of four bridges will form is the product of the probabilities of the individual bridges, i.e.,

\[
p = \frac{N_{\text{p,eff}}}{N_{\text{eff}}} \left( \frac{1}{N_{\text{eff}}-1} \right) \left( \frac{1}{N_{\text{eff}}-2} \right) \left( \frac{1}{N_{\text{eff}}-3} \right) \cdots \left( \frac{1}{N_{\text{eff}}-\text{unp}} \right).
\]

Therefore, \( p(N_{\text{eff}}, N_{\text{p,eff}}, B) = \prod_{j=1}^{B-1} \frac{N_{\text{p,eff}}-j}{N_{\text{eff}}-j} \) (5)

A.2. Number of sets of bridges that can bring together the ends

As shown in Fig. 3, if \( B=4 \) is the number of bridges that bring together the ends to no more than \( X \) unpaired monomers, then \( X \) is separated into \( (B+1) \) regions of unbridged bases, such that \( X = x_1 + x_2 + \cdots + x_{B+1} \). The number of ways to form \( B \) bridges such that the \( (B+1) \) regions of unbridged bases sum to a total of \( X \) is a product of the following two factors:

1. (i) the number of integer sets \( \{x_1, x_2, \ldots, x_{B+1}\} \), such that \( X = x_1 + x_2 + \cdots + x_{B+1} \); and
2. (ii) the number of ways to rearrange the \( (B+1) \) regions of unbridged bases on the polymer, such that these regions maintain their sequential order \( \{x_1, x_2, \ldots, x_{B+1}\} \). If their orders are scrambled, it would be a different set of \( \{x_i\} \), resulting in a different structure.

The factor (i) is simply the number of ways \( X \) indistinguishable objects (i.e., monomers in these regions) can be distributed into \( (B+1) \) distinguishable boxes (i.e., regions of unbridged bases): 
\[
\binom{X+B}{B} = \frac{(X+B)!}{(X-B)! B!}.
\]

For factor (ii), we need to know the number of ways \( (B+1) \) regions of unbridged bases can move around and rearrange themselves on the polymer while maintaining their sequential order. The two regions at the two ends (i.e., \( x_1 \) and \( x_{B+1} \)) are fixed at the ends,
so only \((B-1)\) regions can change their positions on the polymer. Thus, the relevant combinatoric factor has the form \(\binom{N_{\text{unp}}}{B-1}\). When considering the number of “available positions” for \((B-1)\) “objects” to choose from, we must subtract from \(N_{\text{eff}}\) the 2 “positions” occupied by each bridge (i.e., \(-2B\)), as well as \(X\) “positions” reserved as the end-to-end distance (i.e., \(-X\)). We must also give back one “position” for each of the \((B-1)\) “objects” in the middle of the polymer, i.e., \(+B\). Therefore, considering all possible \(X\)'s, i.e., \(X=1,2,\ldots,N_{\text{unp}}\), the number of ways to rearrange the regions of unbridged bases is

\[
\sum_{X=0}^{X} \frac{(N_{\text{eff}}-X+B-1)!}{X!(B-1)!}.
\]

However, this is not yet the total number of possible sets of bridges, because the bridges can cross over each other without changing the \(X\)'s or their positions (Fig. 9), and still maintain the same end-to-end distances, \(X\) (Fig. 10), as long as \(X_2 \leq j-i\) and \(X_2 \leq i-k\). Since we are interested primarily in “small" \(X\) and “large" \(N\), i.e., \(X < N\), these conditions are true for the overwhelming majority of the cases. Thus, we neglect cases when these conditions are not met. For each of the \((B-1)\) regions of unbridged bases in the middle of the polymer (i.e., excluding the two regions at the ends), there are two ways it can be connected, thus we would have under-counted by a factor of \(2^{B-1}\) without considering bridge crossovers. It follows that,

\[
\Omega(N_{\text{eff}},X,B) = \sum_{X=0}^{X} \left(\frac{(N_{\text{eff}}-X+B-1)!}{X!(B-1)!}\right)
\]

A.3. The probability that the ends will be no more than \(X\) unpaired bases apart

For every \(N_{\text{eff}},P_{\text{eff}}\) and \(B\), the probability that a particular set of bridge will form is \(p(N_{\text{eff}},P_{\text{eff}},B)\), as given by Eq. (5). The probability that this set of bridges will not form is \(1-p(N_{\text{eff}},P_{\text{eff}},B)\). Since there are \(\Omega(N_{\text{eff}},X,B)\) sets of qualifying bridges (Eq. (6)), the probability that none of those sets will form is

\[
[1-p(N_{\text{eff}},P_{\text{eff}},B)]^{\Omega(N_{\text{eff}},X,B)}.
\]

We must also consider all possible values for \(B=1,2,3,\ldots,N_{\text{eff}}/2\). Thus, the probability that \textbf{no} bridges will form to make the ends close enough is

\[
\frac{N_{\text{eff}}}{2} \prod_{B=1}^{B} \left(1-p(N_{\text{eff}},P_{\text{eff}},B)\right)^{\Omega(N_{\text{eff}},X,B)}.
\]

from which it follows that the probability, \(P\), that the ends will be bridged to a distance of no more than \(X\) unpaired bases apart in Eq. (2) (previously shown in Section 3.2):

\[
P(N_{\text{eff}},P_{\text{eff}},X) = 1 - \sum_{B=1}^{B} \left(1-p(N_{\text{eff}},P_{\text{eff}},B)\right)^{\Omega(N_{\text{eff}},X,B)}.
\]

The probability that the ends will be exactly \(X\) (as opposed to less than or equal to \(X\)) as in Eq. (3) (previously shown in Section 3.2):

\[
\rho(N_{\text{eff}},P_{\text{eff}},X) = P(N_{\text{eff}},P_{\text{eff}},X) - P(N_{\text{eff}},P_{\text{eff}},X-1).
\]

Eq. (3) can be used to predict the probability distribution of end-to-end distances given the total length, \(N_{\text{eff}}\) and the pairing fraction, \(P_{\text{eff}}\) of an RSPP, using the mapping discussed in Section 2.1. A sample plot was shown in Fig. 5. The expected end-to-end distance, \(<X>\), given \(N_{\text{eff}}\) and \(P_{\text{eff}}\), as in Eq. (4) (previously shown in Section 3.3):

\[
<X(N_{\text{eff}},P_{\text{eff}})> = \sum_{X=0}^{X} \rho(N_{\text{eff}},P_{\text{eff}},X) \cdot X,
\]

where \(N_{\text{unp}} = N_{\text{eff}} - P_{\text{eff}}\) is the total number of unpaired bases in an RSPP. This is the key equation that shows how the expected end-to-end distance of RNA, \(<X>\), varies with its total length, \(N\), and pairing fraction, \(f\), while fixing \(f\) and \(N\), respectively (Figs. 6 and 8). Fig. 7 demonstrates the key finding of this paper, \(<X> \propto N^{1/4}\).

References


Karetnikov, A., Lehto, K., 2008. Translation mechanisms involving long-distance base pairing interactions between the 5′ and 3′ non-translated regions and internal ribosomal entry are conserved for both genomic RNAs of blackcurrant reversion nepovirus. Virology 371 (2), 292–308.


