292–299 Nucleic Acids Research, 2011, Vol. 39, No. 1 Published online 1 September 2010
doi:10.1093/nar/gkq642

The ends of a large RNA molecule are necessarily close

Aron M. Yoffe[1], Peter Prinsen[2], William M. Gelbart[1,*] and Avinoam Ben-Shaul[3,*]

[1]Department of Chemistry and Biochemistry, University of California, Los Angeles, 607 Charles E. Young Drive East, Los Angeles, CA 90095-1569, USA, [2]Instituut-Lorentz, Leiden University, P.O. Box 9506, 2300 RA Leiden, The Netherlands and [3]Department of Physical Chemistry, and the Fritz Haber Institute, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

Received March 23, 2010; Revised July 3, 2010; Accepted July 5, 2010

ABSTRACT

We show on general theoretical grounds that the two ends of single-stranded (ss) RNA molecules (consisting of roughly equal proportions of A, C, G and U) are necessarily close together, largely independent of their length and sequence. This is demonstrated to be a direct consequence of two generic properties of the equilibrium secondary structures, namely that the average proportion of bases in pairs is ~60% and that the average duplex length is ~4. Based on mfold and Vienna computations on large numbers of ssRNAs of various lengths (1000–10 000 nt) and sequences (both random and biological), we find that the 5′–3′ distance—defined as the sum of H-bond and covalent (ss) links separating the ends of the RNA chain—is small, averaging 15–20 for each set of viral sequences tested. For random sequences this distance is ~12, consistent with the theory. We discuss the relevance of these results to evolved sequence complementarity and specific protein binding effects that are known to be important for keeping the two ends of viral and messenger RNAs in close proximity. Finally we speculate on how our conclusions imply indistinguishability in size and shape of equilibrated forms of linear and covalently circularized ssRNA molecules.

INTRODUCTION

There are many situations in which it is biologically important for the two ends of a large RNA molecule to be close to each other. In animal viruses with single-stranded (ss) RNA genomes, for example, efficient replication of the genome has been shown to depend on its effective 'circularization'. More explicitly, complementary sequences have been identified at or near the 5′- and 3′-ends that are responsible for forming 'panhandles' that keep the two ends close together. These panhandles are duplexes that are 21 bp in the case of yellow fever virus (1), and 15 bp in the case of influenza A (2), thereby according them unusual robustness. Another example where RNA genome circularization of this kind has been implicated in RNA replication is sindbis virus; here an 18 bp 5′–3′ panhandle has been shown to survive denaturing conditions sufficient to eliminate much of the remaining secondary structure, leaving the genome with a circular appearance in electron micrographs (3). In dengue, also (like yellow fever, influenza A and sindbis) a positive-sense RNA virus, minus-strand synthesis involves long-distance 5′–3′ base pairing that facilitates the transfer of the RNA-dependent RNA polymerase from its binding site at the 5′-end to the initiation site at the 3′-end (4). Similarly, circularization of HIV-1 has been shown to arise from base pairing between the 5′- and 3′-ends of the RNA genome (5); these interactions are found to occur as well in different HIV-1 subtypes with large sequence variation, suggesting they share an evolutionary basis.

It has also long been known that effective circularization of messenger RNA molecules is important for efficient translation. The 5′- 'capping' and 3′-polyadenylation of mRNAs—through a variety of specific protein-binding events—result in the association of the two ends of the molecules and subsequent formation of translation initiation complexes (6). In eukaryotes, for example, the 3′-poly(A) 'tail' interacts with the poly(A)-binding protein, the 5′-G-cap binds a eukaryotic initiation factor, and these two bound proteins—with the full length of mRNA intervening—simultaneously bind a 'bridging' protein. This effective circularization of the molecule

*To whom correspondence should be addressed. Tel: +1 310 825 2005; Fax: +1 310 267 0319; Email: gelbart@chem.ucla.edu
Correspondence may also be addressed to Avinoam Ben-Shaul. Tel: +972 2 6585271; Fax: +972 2 6513742; Email: abs@fh.huji.ac.il

results in recruitment of the 40S ribosomal subunit (via binding of still another protein) and initiation of translation.

Because circularization of mRNA is so important for its translation, mechanisms that co-localize the ends have evolved even in cases where the molecules are not capped or polyadenylated. Plant viruses, for example, often lack both of these special sequences and yet are translated efficiently (7,8). The effective circularization is enhanced by direct base pairing between sub-sequences in the untranslated regions (UTRs) at the 5′- and 3′-ends; the UTRs functionally replace the G-cap and poly(A) tail. Further, the RNAs of many positive-sense (mRNA) viruses have internal ribosome entry sites (IRESs) at their 5′-ends, i.e. subsequences that recruit ribosomes and initiate translation (9,10).

In all of the above examples—involving both direct interaction between 5′- and 3′-ends or interaction mediated by binding proteins—particular, evolved, subsequences are involved in effective circularization. But in all of these scenarios, an even more fundamental requirement is that the two ends of the fluctuating molecule must spend enough time near each other in order for there to be a high probability for the special elements—RNA subsequences or binding proteins—to find one another. More explicitly, we will argue here that effective circularization of large RNA molecules is achieved through generic properties of secondary structure that are essentially independent of sequence. The specific evolved subsequences mentioned above are not needed so much for circularization as for facilitating the binding of particular proteins—e.g. RNA replicases and ribosome initiation factors—that are important for biological function of the circularized RNA.

Consider the analogous situation of double-stranded (ds) DNA with 'sticky' ends arising from complementary ss overhangs (generated, say, by a restriction enzyme). Here the probability of the two ends being covalently bound by a ligase is directly determined by—and ultimately limited by—the likelihood that they are close enough to each other to bind, i.e. that the double helix can twist and bend enough for its two ends to get close together (11). This classic problem is informed by the well-known statistical mechanical result giving the likelihood of the ends of a linear, semiflexible, polymer being within a monomer distance of one another. For sufficiently long molecules ($L \gg \xi$), this probability is of order $1/(L/\xi)^{3/2}$, where L and ξ are the contour and persistence lengths, respectively, of the linear polymer; the contour length is the number of monomers times the average inter-monomer distance, and the persistence length is the distance along the chain contour beyond which the polymer can bend almost freely (12). Thus, the circularization probability of long DNA is small because L/ξ is large, i.e. the molecule is long compared to its persistence length (50 nm, for DNA): maximization of configurational entropy requires that the ends be far apart. The small probability of finding them close, decreasing as $L^{-3/2}$, reflects directly the fact that the root-mean-square distance between the ends of the molecule is increasing as $L^{1/2}$.

To understand the basis for effective circularization of ssRNA, then, it is natural to ask: is there, in analogy with dsDNA, a generic result for the probability of finding the two ends of an RNA molecule close to one another, and how different is it from that for a linear polymer? In this article we argue that there is indeed a universal distribution of end-to-end distances in large RNA molecules, and furthermore that it is essentially independent of overall sequence and length. We show in particular that the distance between ends is necessarily small, because of generic features of the secondary structure, notably that the percentage (f) of paired nucleotides (nt) is ∼60% and that the average duplex length (k) is ∼4. Using an early variant of the RNA folding algorithm developed by Zuker *et al.* (13,14), Fontana *et al.* (15) have calculated various characteristics of the minimum free energy (MFE) structure corresponding to several different types of short (20–100) nucleotide sequences. Averaging over many sequences of the same length (number of nucleotides, N) and base composition (φ), they found that f and k approach a constant value with increasing N. They also calculated a property (the number of unpaired bases in 'joints' and 'free ends') that is closely related to our definition of the 5′–3′ distance (see next section), finding that for the short chains analyzed this number increases, yet with a gradually decreasing slope, as N increases. The constancy of f and k has been confirmed for a wide range of biological (viral and yeast) ssRNA sequences (16) by application of the mfold and Vienna codes for predicting thermally accessible secondary structures.

For certain models of polynucleotide chains, the N-independence of f and k has been proven analytically, using a variety of powerful theoretical tools. Hofacker *et al.* (17), applying an elegant graph-theoretic approach, derived exact results for these properties (see their Table 3) and various other secondary structure attributes of RNA-like heteropolymers. Their results apply to an idealized ensemble where all possible secondary structures have equal statistical weight, resulting in low values of f and k. More recently, Clote *et al.* (18), using the Nussinov–Jacobson ('maximum base pairing') model (19) have shown that, for an ssRNA chain with Watson–Crick pairing rules, f approaches a constant value slightly exceeding 90% for N large (>1000). Earlier, de Gennes had noted (20) that, for a random sequence of two complementary nucleotides, the distance between chain ends remains finite even as N approaches infinity. Based on this notion he also concluded that '…many properties of a large, open, strand are not very different from those of a cyclic strand of equal molecular length' (20). We elaborate on this idea in the next section.

Our goal in the present work is to emphasize the generality of the proximity of the 5′- and 3′-ends of large RNA molecules of arbitrary length and sequence. Based on the general findings noted above for large ssRNA chains, we derive a simple expression for the 5′–3′ distance that can be evaluated numerically for sequences of given f and k. We also calculate this distance using the RNAsubopt (21,22) and mfold (23,24) folding algorithms. A further consequence of our analyses is that the

secondary—and hence tertiary—structures of linear and covalently-circularized RNA molecules are practically identical. These conclusions are tested against several systematic calculations of secondary structures for specific linear and circular sequences, both random and viral.

METHODS

Figure 1A displays the MFE secondary structure of a rather short (200 nt) random-sequence ssRNA molecule, composed of equal numbers of A, C, G and U, as predicted by the mfold algorithm (23,24). The duplexes are represented in the usual way by straight 'ladders' and the loops by circles of different sizes. The same secondary structure is visualized slightly less schematically in Figure 1B, with more realistic scaling of duplex dimensions, using the jViz.Rna drawing program (25). This latter representation illustrates that the dangling ss segments in the 'exterior loop'—the one including the 5′- and 3′-ends—are independent flexible chains. In Figure 1C the secondary structure is mapped into a tree graph, where each edge (bond) represents a duplex and the vertices represent the loops (15,17,26); the interior loops are denoted by solid circles, and the exterior loop by an open circle. The term 'interior loop' is conventionally defined as the chain of bases, both paired and unpaired, comprising a closed loop, excluding its closing ('downstream') base pair. In the following we slightly depart from this definition and include the closing base pair as part of the (hence closed) loop. Our definition of the exterior loop, which lacks a closing base pair, is identical to the conventional one, namely, it includes all bases (paired and unpaired) along the shortest connected (covalently or H-bonded) path from the 5′- to the 3′-end.

5′–3′ Distance

As a simple intuitive measure of the 5′–3′ distance (in a given secondary structure of a given sequence) we use the total number of nucleotide links comprising the exterior loop, i.e.

$$D = l_{\mathrm{ext}} + d_{\mathrm{ext}} \qquad (1)$$

Here l_{ext} is the number of covalent (phosphodiester) bonds (hereafter also referred to as ss links) in the exterior loop and d_{ext} is the number of base-paired (H-bonded, ds) links in the exterior loop or, equivalently, the number of duplexes emanating from the exterior loop. As it is the total number of (ss and ds) links in the nucleotide chain constituting the exterior loop, we shall refer to D as the 'effective contour length' of this loop. Expressing D in the form $D = n_{\mathrm{ext}} - 1 = (s_{\mathrm{ext}} + 2d_{\mathrm{ext}}) - 1$, where n_{ext} is the total number of nucleotides in the exterior loop, and noting that $2d_{\mathrm{ext}}$ is the total number of paired bases in the exterior loop, it follows from Equation (1) that $s_{\mathrm{ext}} = l_{\mathrm{ext}} - d_{\mathrm{ext}} + 1$ is the number of unpaired bases in this loop. Figure 2 illustrates an exterior loop where $D = l_{\mathrm{ext}} + d_{\mathrm{ext}} = 11 + 2 = 13$, whereas in Figure 1 $D = l_{\mathrm{ext}} + d_{\mathrm{ext}} = 14 + 1 = 15$. It should be emphasized that

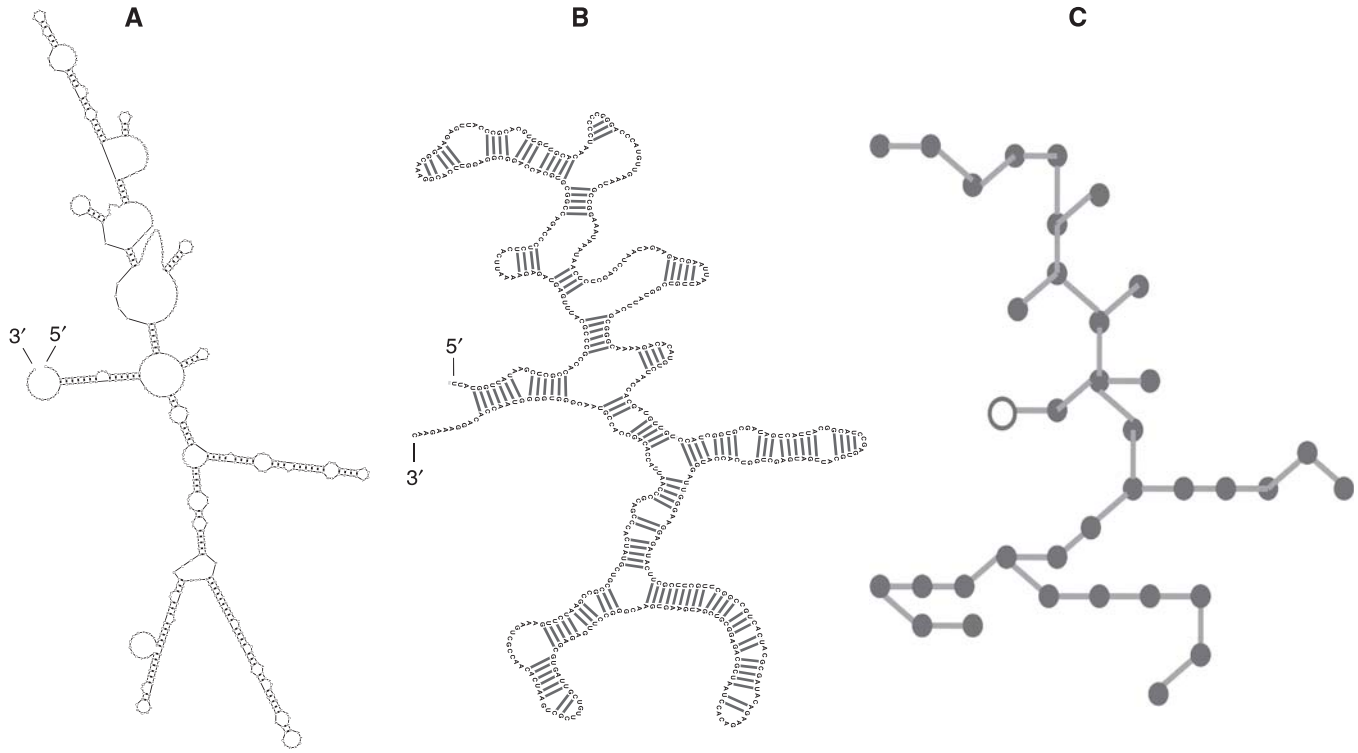

Figure 1. Three different representations of the mfold-predicted minimum free energy secondary structure of a random 200 nt ssRNA of uniform composition (25% A, C, G, U). (**A**) Conventional schematic, drawn with mfold, showing base-paired regions (duplexes) and single-stranded loops. (**B**) jViz.Rna drawing (16), emphasizing the flexibility of single-stranded loops and scaled dimensions of duplexes. (**C**) Graph-theoretic mapping of this secondary structure, reducing duplexes to edges (bonds) and loops to vertices (filled circles); the single 'exterior' loop is depicted by an open circle.

the average physical distance between the 5′- and 3′-ends depends not only on D but also on the specific sequence of the loop, as well as the number of duplexes branching from the loop. In fact the lengths of the covalent and H-bonded links are different (the latter are about three times larger). If all links were of equal length b, and their joints were fully flexible, then the physical 5′–3′ distance would be roughly $b\sqrt{D}$, where we have neglected excluded volume effects because of the shortness of the exterior loop (12). It follows that small, N-independent, D-values imply small, N-independent physical distances between the two chain ends.

Four simple observations will guide our calculation of the 5′–3′ distance:

(i) The MFE secondary structures of a given linear ssRNA molecule and that of the circular RNA obtained by linking the 5′- and 3′-ends of the linear chain are very similar, and their energies practically identical. This is because the presence or absence of a covalent (phosphodiester) bond between the terminal nucleotides does not significantly alter overall base pairing. Its small influence on the configurational free energy of the molecule enters only through the entropy difference between the open exterior loop in the linear RNA and the corresponding closed (interior) loop in the circular analog. Actually, for any secondary structure of the linear ssRNA, not only the one of minimum free energy, the corresponding circular structure has essentially the same energetic and structural characteristics. Conversely, any secondary structure of a linear RNA can be regarded as derived from 'cutting' a specific covalent bond in one of the interior loops of the corresponding circular RNA. We thus expect that secondary structure characteristics of long RNA molecules, such as the pairing fraction or average duplex length, are practically the same for the linear and circularized 'isomers'. These conclusions have been confirmed by numerical analyses of a large number of linear and circular RNA sequences of different lengths and compositions, as reported below and in Supplementary Figure S1 and Supplementary Table S1.

(ii) As noted in the Introduction, for long chains (say $N > 1000$) composed of comparable proportions of A, C, G and U ($25 \pm 5\%$), we find that $f = 60 - 65\%$ for randomly-permuted sequences and for most viral RNAs (Tables 1 and 2).

(iii) For long chains, we also know that the average length of (i.e. number of base pairs in) a duplex, k, is independent of N and rather insensitive to φ (for compositions involving $25 \pm 5\%$ of the four bases). For nearly all the sets of sequences examined in this study—randomly-permuted, viral and yeast-derived—k is between 4 and 5 (Tables 1 and 2; Supplementary Table S1).

(iv) As is well known, every secondary structure can be represented by a tree graph (26), as illustrated in Figure 1C.

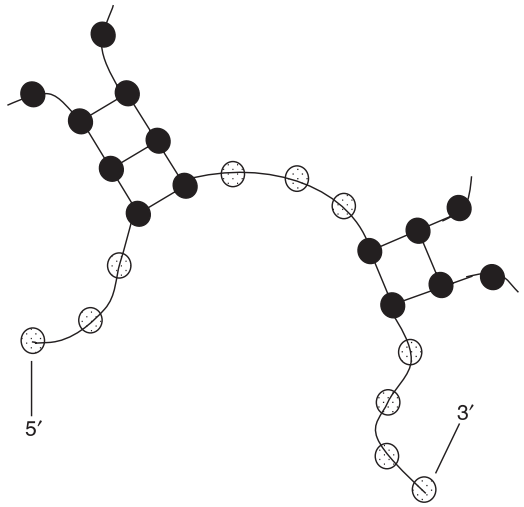

Figure 2. Detailed view of an exterior loop consisting of $l_{ext} = 11$ covalent links and $d_{ext} = 2$ H-bonded links of nucleotides. The effective contour length of the loop is $D = l_{ext} + d_{ext} = 11 + 2 = 13$.

Two simple and important results can easily be proved from the tree graph analogy. First, the number of vertices, L, and the number of bonds, S, of a circular RNA are related by the equality $L = S + 1$. This relation is also valid for linear RNAs provided the exterior loop is also represented by a vertex (possibly differently labeled, as in Fig. 1C). Second, on average (over all loops in any given structure), each loop (vertex) is connected to $\langle d \rangle = 2 - 2/L$ duplexes (edges). For long ($N \gg 1$) sequences we also find $L \gg 1$ (see below), in which case we can safely set $\langle d \rangle = 2$, which (unless otherwise stated) will

Table 1. Composition (φ)-dependence of the average percentage of bases paired (f), the average duplex length (k) and the average 5′–3′ distance (D), for different sets of random and yeast-derived sequences of length 3000 nt; each set consists of 500 sequences

Type of ssRNA	Folding program	φ (%)[a]				f (%)	k (bp)	D, links	D, from Equation (2)
		G	C	A	U				
Random, viral-like φ	RNAsubopt	24	22	26	28	62 ± 1	4.0 ± 0.1	12 ± 4	11.6
Random, uniform φ	RNAsubopt	25	25	25	25	61 ± 1	3.9 ± 0.1	12 ± 5	12.6
Yeast-derived[b]	RNAsubopt	19	19	31	31	58 ± 2	4.1 ± 0.1	14 ± 5	11.9
Random, viral-like φ	mfold	24	22	26	28	61 ± 1	4.5 ± 0.1	14 ± 7	12.8

Values following the ± symbols are standard deviations.
[a]The randomly-permuted ssRNAs of each type are of identical composition; for the yeast ssRNAs, the mean composition is listed.
[b]These are ssRNA transcripts of successive 3000 bp sections of yeast (*S. cerevisiae*) chromosomes XI and XII.

Table 2. Values of f, k and D for viral ssRNAs, determined with RNAsubopt

Viral taxon	No. of seq.[a]	Host	N (nt)	f (%)	k (bp)	D, links
Bromoviridae RNA3	8	Plant	2210	63 ± 1	4.2 ± 0.1	19 ± 6
Bromoviridae RNA2	8	Plant	2891	63 ± 2	4.3 ± 0.1	18 ± 4
Bromoviridae RNA1	8	Plant	3265	64 ± 2	4.3 ± 0.1	15 ± 3
Leviviridae	9	Bacterium	3780	68 ± 2	4.3 ± 0.1	15 ± 9
Sobemovirus	9	Plant	4199	66 ± 2	4.2 ± 0.2	17 ± 4
Luteovirus	17	Plant	5725	62 ± 1	4.2 ± 0.1	16 ± 7
Tymovirus	9	Plant	6300	45 ± 4	3.9 ± 0.1	26 ± 5
Tobamovirus	22	Plant	6425	64 ± 1	4.2 ± 0.1	19 ± 5
Astroviridae	6	Animal	6719	63 ± 1	4.3 ± 0.1	16 ± 8
Caliciviridae	18	Animal	7713	62 ± 1	4.1 ± 0.1	20 ± 19

Values following the ± symbols are standard deviations.
[a]Number of sequences analyzed.

be the value used in our calculations. Note that the averaging here is over all loops in a given structure. The same holds, of course, after averaging over any number of structures and/or sequences. Note also that we always have $d \geq 1$, with $d=1$ corresponding to a 'hairpin' loop, $d=2$ to a 'bubble' or 'bulge,' and $d \geq 3$ to a 'multi loop'.

Among the numerous possible secondary structures of long RNA sequences, there are often thousands whose free energies are just marginally higher ($k_B T$ or less) than that of the MFE configuration, and under equilibrium conditions all these structures are nearly equally likely. Consequently, any property of the molecule that depends on its secondary structures should be averaged over their full thermal (Boltzmann) distribution. Suppose that, using RNAsubopt or a similar program, we have stochastically sampled the thermal ensemble of structures corresponding to a certain circular ssRNA sequence of given N and φ. As argued in (i), above, all the linear ssRNA molecules derived by cutting any covalent (ss) bond in any interior loop of any member of the above ensemble will fold into ensembles of structures that are practically identical both to each other, and to the ensemble of the original circular molecule. The only difference is the appearance of an exterior loop, which now contains the 5′- and 3′-ends. For every given circular structure containing L interior loops, this cutting procedure yields M linear ssRNA sequences, where $M = \sum_{i=1}^{L} l_i$ is the total number of ss (covalent) bonds in all loops of the given structure, l_i denoting the number of covalent bonds in loop i. Noting that the total number of nucleotides in the closed loop i, namely $n_i = s_i + 2d_i$, is equal to the total number of bonds in this loop ($d_i + l_i$), we find $l_i = s_i + d_i$, with s_i and $2d_i$ denoting the number of unpaired and H-bonded nucleotides in loop i, respectively, and d_i the number of duplexes emerging from this loop. This yields $M = \sum_{i=1}^{L} s_i + \sum_{i=1}^{L} d_i = N(1-f) + 2S$. We have used the fact that the first sum is the total number of unpaired nucleotides, $N - Nf$, and the fact that because every duplex is connected to two loops, the second sum is twice the total number (S) of duplexes in the structure. But S can be expressed in the form $S = Nf/2k$ so that $M = N(1 - f + f/k)$. Here, and in all subsequent analytical expressions involving f, its numerical value will be understood to be the fraction of bases in pairs, rather than the percentage. As before, k denotes the average duplex length in the particular sequence considered. For $f=0.6$ and $k=4$ we find $M \sim 0.55N$.

In the next section we present numerical calculations of the average 5′–3′ distance D for two types of ssRNA molecules, biological (yeast-derived and viral) and randomly-permuted sequences. The random sequences were included both for direct comparison to the biological sequences, and for general theoretical interest. In each case, a Boltzmann-weighted average D-value is determined for the thermal ensemble of structures associated with each sequence. We then report the mean of these ensemble-average D-values for each set of sequences.

For the random sequences a simple theoretical prediction of D (showing good agreement with the numerical calculation) can be derived based on two reasonable approximations, as argued in the Appendix 1. We show there that, for any given secondary structure of a very long ($N \gg 1$) ssRNA molecule, the 5′–3′ distance is given by

$$D = \frac{2\langle s \rangle^2 + 4\langle s \rangle + 10}{\langle s \rangle + 2}, \qquad (2)$$

with $\langle s \rangle$ denoting the average number of ss covalent bonds per interior loop in the structure considered. In terms of the pairing fraction, f, and duplex length, k, of this structure we obtain $\langle s \rangle = N(1-f)/L \sim N(1-f)/S = 2k(1-f)/f$. For both the MFE structure and the canonical ensemble averages of secondary structures of random (but also viral) sequences containing roughly equal proportions of the four bases it is found that $f \sim 0.6$ and $k \sim 4$, yielding $\langle s \rangle = 5.33$, and hence $D = 12$. See also Table 1.

Numerical computations

RNA sequences. Randomly-permuted ssRNA sequences were generated with a Fisher–Yates shuffle driven by a Mersenne Twister random number generator (27) implemented in C++ (by R. Wagner, University of Michigan, available at: www-personal.umich.edu/~wagnerr/MersenneTwister.html). Viral ssRNA sequences were obtained from the National Center for Biotechnology Information Genome Database (www.ncbi.nlm.nih.gov). Yeast (*Saccharomyces cerevisiae*)

genomic sequences were obtained from the *Saccharomyces* Genome Database (www.yeastgenome.org).

Folding programs. Secondary structure predictions were made with two RNA folding programs, RNAsubopt, a program in the Vienna RNA Package, Version 1.7 (21,22), and mfold, Version 3.1 (23,24). These programs employ detailed empirically-based energy models to estimate the free energies of the non-pseudoknotted secondary structures that are formed by a specified ssRNA sequence. With RNAsubopt, it is possible to sample stochastically from the ensemble of secondary structures, with a sampling probability in proportion to each structure's Boltzmann weight. Thus, sampling a sufficient number of structures (we use 1000), and averaging the D-values for this set, gives a close approximation to the ensemble-average predicted value of the end-to-end distance for that sequence. In earlier work (16) we demonstrated that the average properties of subsets of 1000 structures are not significantly different from those of the complete ensemble of structures. More generally, for any property X, its RNAsubopt-predicted ensemble-average value is calculated as $\sum_{\alpha=1}^{1000} X_\alpha/1000$, where X_α is its value in the α^{th} member of the stochastically-generated subset of the Boltzmann ensemble of secondary structures. In mfold, by contrast, an algorithm is used to generate a structurally diverse representation of the ensemble, rather than a thermally-representative average. We configured mfold to generate the 1000 lowest-energy structures from such a set, measured D for each, and averaged them in proportion to their Boltzmann weights, to give an mfold-averaged D-value. For any property X, its mfold-predicted average value is $\sum_{\alpha=1}^{1000} X_\alpha \exp(-\Delta G_\alpha/k_B T)/\sum_{\alpha=1}^{1000} \exp(-\Delta G_\alpha/k_B T)$ with ΔG_α the free energy of the α^{th} secondary structure relative to the MFE for that sequence.

RESULTS

While there can be significant inter-taxon variation, the average composition, φ, of the viral RNAs in this study is ~24% G, 22% C, 26% A and 28% U (16). With this 'viral-like' φ, we generated 2000 random sequences of lengths 50, 100, 200 and 400; 1000 of lengths 800 and 1500; 500 of lengths 2000, 2500, 3000 and 4000; 300 of lengths 5000, 6000 and 7000; and 1000 of length 8000. These sequences were folded with RNAsubopt. Figure 3 shows the mean D and standard deviation for each length of RNA, and a regression line fitted to sequences of length 400 and greater. Except for the very short sequences, D is ~12, independent of sequence length; in addition, it is relatively insensitive to small changes in φ. That this D-value is identical to the estimate obtained above, through the theoretical calculation, is coincidental, because the latter is based on the somewhat approximate expression given in Eq. (2) (the approximations are explained in Appendix 1). But it is nevertheless very striking, and highly significant, that the simple theory predicts a D-value that is of the correct magnitude and that is independent of length and sequence.

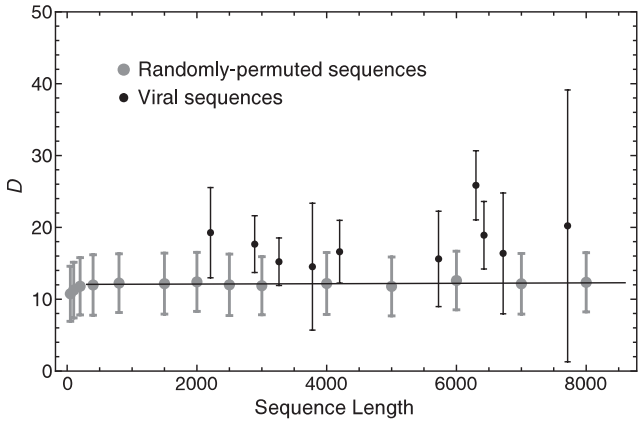

Figure 3. Mean ensemble-averaged 5′–3′ distances, D, from Equation (1), for random and viral sequences. Standard deviations are shown with vertical bars. The small black points represent the 10 groups of viral sequences listed in Table 2. The large gray points represent the 14 different lengths of randomly-permuted RNAs (50–8000 nt), of viral-like composition, described in the text. The line is a least-squares fit to the D values for random sequences with $N \geq 400$. The asymptotic value of D for the random sequences is very close to the theoretically predicted one, $D \sim 12$ [see Equation (2)].

Table 1 shows the results for 500 3000-nt ssRNAs of viral-like and uniform φ, as well as 500 ssRNAs that are the transcripts of consecutive 3000 bp sections on yeast (*S. cerevisiae*) chromosomes XI and XII. In these sets, the values of D, f and k (averaged over the 500 sequences) were 12–14, ~60% and ~4, respectively. The last column in the table lists the values of D calculated according to Equation (2), and these results are seen to agree closely with those from the detailed numerical calculations (especially for the random sequences, as expected).

The viral taxa analyzed are listed in Table 2. All are non-enveloped ssRNA viruses and, except for the rod-shaped Tobamoviruses, have $T = 3$ icosahedral capsids. The Leviviridae infect bacteria, the Astroviridae and Caliciviridae are animal viruses, and the remainder infect plants. The Bromoviridae are, in addition, tripartite: the genome consists of three ssRNAs, divided among three separate capsids. The number of sequences analyzed in each case corresponds to the number of species considered.

From Figure 3 it can be seen that the values and standard deviations of D for the viral RNAs are higher, but overlap those of the random sequences for all taxa except the Tymoviruses. The latter can be understood from the fact that small D-values are an inherent consequence of base pairing; all non-pathological secondary structures with a sufficiently high percentage of bases in pairs, f, will have a low D. The Tymoviruses show a relatively larger D (although still small relative to sequence length) because they have a significantly smaller f.

We note that current RNA folding programs have been shown to be limited in their ability to correctly predict individual base pairs in long ssRNA sequences (28). Consistent with this, RNAsubopt and mfold (which use slightly different energy models to generate their ensembles of secondary structures, and different algorithms to

sample from these ensembles), when given long sequences to fold, output structures that often show significant differences in the details of base pairing, as well as overall appearance. However, our simple theoretical model predicts that D depends only on the values of f and k, which we have previously found to be robust with respect to the details of the folding program used (16). Consequently, D should likewise be robust to the details of the folding program, and thus insensitive to low-level inaccuracies in specific predictions of base pairing. To test this, we compared predictions of D made using mfold and RNAsubopt. As expected, we found that the values do not differ significantly between the two folding programs, and can thus be considered broadly robust to the specific characteristics of the energy model used (Table 1).

There is currently no published experimental work that directly measures the 5′–3′ distance of large (10^3–10^4 nt) ssRNAs in their native state (i.e. not complexed with proteins). However, based on a combination of experimental and computational approaches, Filomatori *et al.* (4) have proposed a model for the secondary structure of the exterior loop of native dengue ssRNA. Their proposed loop has a D-value of 25, which is of the same magnitude as both the theoretical predictions in Table 1, and the numerical predictions in Table 2.

DISCUSSION

We have made two predictions in the current work, both of which can be tested experimentally. First, we have predicted with general theoretical arguments—and demonstrated with numerical computations involving the equilibrated secondary structures of a large number of different lengths and sequences—that the distance between ends of an ssRNA (or ssDNA) should be ~10–15 nt links. This corresponds to a 3D physical distance of a few nm, which is far smaller than the contour lengths of large ssRNA molecules. As mentioned earlier, a crude estimate of the 3D distance between ends may be obtained in terms of the root-mean-square (RMS) end-to-end distance ($b\sqrt{D}$) associated with a flexible linear polymer defined by the string of covalent and H-bonded links shown in Figure 2. With an average link size, b, of ~3/4 nm, and a D of 12, one obtains an RMS end-to-end distance of ~3 nm. This is approximately an order of magnitude less than the 37 nm average distance between nucleotides (radius of gyration) that has been measured by small-angle X-ray scattering for a 6400 nt viral ssRNA (29). Our estimate of 3 nm could be confirmed by fluorescence resonance energy transfer (FRET) measurements, or still more directly by cryo-EM imaging of large ssRNA molecules whose ends have been labeled by small gold particles (for example, 1 nm particles conjugated to oligonucleotides that are complementary to the 5′- and 3′-ends).

Second, we have predicted that all the linearized ssRNAs obtained by making a single cut in a long circular ssRNA molecule should have secondary (and hence) tertiary structures that are essentially identical to that of the parent circular form. Accordingly, they should have the same size and shape. And because they necessarily have the same charge, they should show virtually indistinguishable band positions in native gels, even though the linear and circular forms can be easily distinguished in denaturing gels where the secondary structure needed to effectively circularize the linear molecule has been destroyed. Similarly, under native conditions, small-angle X-ray scattering experiments, cryo-EM, and measurements of diffusion coefficients/hydrodynamic radii should show no difference between the circular and linearized molecules. The only caveat here, as well as for the measurements of 5′–3′ distance described earlier, is that the secondary structures of the molecules be equilibrated, since this is explicitly assumed in the theoretical arguments leading to all of these predictions [for a critical discussion of the equilibration/renaturation (and the lack thereof) of ssRNA, see Uhlenbeck (30)].

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

We thank Li Tai Fang and Charles M. Knobler for many helpful discussions.

FUNDING

US National Science Foundation (grant number CHE07-14411 to W.M.G.); the Israel Science Foundation (grant number 695/06 to A.B.-S.); the US–Israel Bi-National Science Foundation (grant number 2006-401 to A.B.-S.); The Netherlands Organization for Scientific Research, Rubicon grant (to P.P.); and the University of California, Los Angeles, a Dissertation Year Fellowship (to A.M.Y.). Funding for open access charge: Research grant of A.B.-S. (grant number ISF 695/06).

Conflict of interest statement. None declared.

REFERENCES

1. Corver,J., Lenches,E., Smith,K., Robison,R.A., Sando,T., Strauss,E.G. and Strauss,J.H. (2003) Fine mapping of a cis-acting sequence element in yellow fever virus RNA that is required for RNA replication and cyclization. *J. Virol.*, **77**, 2265–2270.
2. Hsu,M.-T., Parvin,J.D., Gupta,S., Krystal,M. and Palese,P. (1987) Genomic RNAs of influenza viruses are held in a circular conformation in virions and in infected cells by a terminal panhandle. *Proc. Natl Acad. Sci. USA*, **84**, 8140–8144.
3. Frey,T.K., Gard,D.L. and Strauss,J.H. (1979) Biophysical studies of circle formation by sindbis virus 49S RNA. *J. Mol. Biol.*, **132**, 1–18.
4. Filomatori,C.V., Lodeiro,M.F., Alvarez,D.E., Samsa,M.M., Pietrasanta,L. and Gamarnik,A.V. (2006) A 5′ RNA element promotes dengue virus RNA synthesis on a circular genome. *Genes Dev.*, **20**, 2238–2249.
5. Ooms,M., Abbink,T.E.M., Pham,C. and Berkhout,B. (2007) Circularization of the HIV-1 RNA genome. *Nucleic Acids Res.*, **35**, 5253–5261.
6. Gallie,D.R. (1991) The cap and poly(A) tail function synergistically to regulate mRNA translational efficiency. *Genes Dev.*, **5**, 2108–2116.

7. Kneller,E.L.P., Rakotondrafara,A.M. and Miller,W.A. (2006) Cap-independent translation of plant viral RNAs. *Virus Res.*, **119**, 63–75.
8. Miller,W.A. and White,K.A. (2006) Long-distance RNA-RNA interactions in plant virus gene expression and replication. *Annu. Rev. Phytopathol.*, **44**, 447–467.
9. Karetnikov,A. and 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**, 292–308.
10. Fabian,M.R. and White,K.A. (2004) 5′–3′ RNA-RNA interaction facilitates cap- and poly(A) tail-independent translation of tomato bushy stunt virus mRNA: a potential common mechanism for Tombusviridae. *J. Biol. Chem.*, **279**, 28862–28872.
11. Cloutier,T.E. and Widom,J. (2005) DNA twisting flexibility and the formation of sharply looped protein-DNA complexes. *Proc. Natl Acad. Sci. USA*, **102**, 3645–3650.
12. Grosberg,A.Y. and Khokhlov,A.R. (1994) *Statistical Physics of Macromolecules*. AIP Press, New York.
13. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.*, **9**, 133–148.
14. Zuker,M. and Sankoff,D. (1984) RNA secondary structures and their prediction. *Bull. Math. Biol.*, **46**, 591–621.
15. Fontana,W., Konings,D.A.M., Stadler,P.F. and Schuster,P. (1993) Statistics of RNA secondary structures. *Biopolymers*, **33**, 1389–1404.
16. Yoffe,A.M., Prinsen,P., Gopal,A., Knobler,C.M., Gelbart,W.M. and Ben-Shaul,A. (2008) Predicting the sizes of large RNA molecules. *Proc. Natl Acad. Sci. USA*, **105**, 16153–16158.
17. Hofacker,I.L., Schuster,P. and Stadler,P.F. (1998) Combinatorics of RNA secondary structures. *Discr. Appl. Math.*, **88**, 207–237.
18. Clote,P., Kranakis,E., Krizanc,D. and Stacho,L. (2007) Asymptotic expected number of base pairs in optimal secondary structure for random RNA using the Nussinov–Jacobson energy model. *Discr. Appl. Math.*, **155**, 759–787.
19. Nussinov,R. and Jacobson,A.B. (1980) Fast algorithm for predicting the secondary structure of single stranded RNA. *Proc. Natl Acad. Sci. USA*, **77**, 6309–6313.
20. de Gennes,P.G. (1968) Statistics of branching and hairpin helices for the dAT copolymer. *Biopolymers*, **6**, 715–729.
21. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
22. Wuchty,S., Fontana,W., Hofacker,I.L. and Schuster,P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.
23. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
24. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
25. Wiese,K.C., Glen,E. and Vasudevan,A. (2005) JViz.Rna—a Java tool for RNA secondary structure visualization. *IEEE T. Nanobiosci.*, **4**, 212–218.
26. Waterman,M.S. (1978) Secondary structure of single-stranded nucleic acids. *Adv. Math. Suppl. Stud.*, **1**, 167–212.
27. Matsumoto,M. and Nishimura,T. (1998) Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM T Model. Comput. Sci.*, **8**, 3–30.
28. Mathews,D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178–1190.
29. Muroga,Y., Sano,Y., Inoue,H., Suzuki,K., Miyata,T., Hiyoshi,T., Yokota,K., Watanabe,Y., Liu,X., Ichikawa,S. *et al.* (2000) Small angle X-ray scattering studies on local structure of tobacco mosaic virus RNA in solution. *Biophys. Chem.*, **83**, 197–209.
30. Uhlenbeck,O.C. (1995) Keeping RNA happy. *RNA*, **1**, 4–6.

APPENDIX 1: DERIVATION OF *D*

Consider a particular secondary structure α of a given circular ssRNA molecule, containing N nucleotides and with base composition φ. Let $L_\alpha(s,d)$ denote the number of s,d-loops (i.e. loops composed of s unpaired nucleotides and d duplexes) in this structure. Each s,d-loop can be cut through any of its $l=s+d$ covalent bonds, yielding open exterior loops of $s+2d-1$ links. The average effective contour length D_α resulting from this cutting procedure is

$$D_\alpha = \frac{\sum_{s,d}(s+2d-1)(s+d)L_\alpha(s,d)}{\sum_{s,d}(s+d)L_\alpha(s,d)} = \frac{\langle s^2\rangle_\alpha + 3\langle sd\rangle_\alpha + 2\langle d^2\rangle_\alpha - \langle s\rangle_\alpha - \langle d\rangle_\alpha}{\langle s\rangle_\alpha + \langle d\rangle_\alpha}, \quad \textbf{(A1)}$$

where the averages after the second equality are over all loops belonging to the particular structure. This follows from the fact that $D = n_{\mathrm{ext}} - 1 = (s_{\mathrm{ext}} + 2d_{\mathrm{ext}}) - 1$ is the effective contour length of the exterior loop in a particular secondary structure, and $(s+d)L_\alpha(s,d)$ is the statistical weight of s,d-loops containing $s+d$ covalent bonds. $\langle s\rangle_\alpha = \sum_{s,d} sP_\alpha(s,d) = \sum_s sP_\alpha(s)$, with $P_\alpha(s,d) = L_\alpha(s,d)/L_\alpha$ denoting the fraction of s,d-loops in this structure and $L_\alpha = \sum_{s,d} L_\alpha(s,d)$ denoting the total number of loops in this structure. The 'marginal' probability distribution $P_\alpha(s) = \sum_d P_\alpha(s,d) = L_\alpha(s)/L_\alpha$ is the fraction of loops containing s unpaired nucleotides, regardless of the number of duplexes connected to these loops. Similarly, $\langle d\rangle_\alpha = \sum_{s,d} P_\alpha(s,d)d = \sum_d P_\alpha(d)d$, etc. The sums over s include all $s \geq 1$ ($s=1$ corresponds to a bulge) yet we also note that, in the case of a hairpin ($d=1$), energetic considerations generally imply $s \geq 3$. The sums over d include all $d \geq 1$.

For long random sequences a simplified expression for D_α [see Equation (2)], involving only $\langle s\rangle_\alpha$, can be derived based on two reasonable approximations. The first is to assume there are no correlations between the distributions of unpaired and paired nucleotides in loops, i.e. $P_\alpha(s,d) = P_\alpha(s)P_\alpha(d)$, from which it follows that $\langle sd\rangle_\alpha = \langle s\rangle_\alpha\langle d\rangle_\alpha$. Small deviations from this approximation may occur because, for hairpins, we generally have $s \geq 3$, whereas for other loops we have $s \geq 1$. The second approximation serves to relate $\langle s^2\rangle$ to $\langle s\rangle$ and $\langle d^2\rangle$ to $\langle d\rangle$. Here we assume that the distributions $P_\alpha(s)$ and $P_\alpha(d)$ of, respectively, [the $(1-f_\alpha)N$] unpaired nucleotides and $(f_\alpha N/2k_\alpha)$ duplexes among the L_α loops of structure α, are random. These distributions (analogous to those of indistinguishable balls randomly distributed among boxes) are determined by maximizing the (entropy) functional $-\sum_{i\geq i^*} P_\alpha(i)\ln P_\alpha(i)$ $(i=s,d)$, subject to the normalization $\left[\sum_{i\geq i^*} P_\alpha(i) = 1\right]$ and conservation $\left[\sum_{i\geq i^*} iP_\alpha(i) = \langle i\rangle_\alpha\right]$ constraints. In this way we find $P_\alpha(s) = (1-\lambda_\alpha)\lambda_\alpha^{s-s^*}$, with a similar expression for $P_\alpha(d)$. For concreteness and simplicity we set $s^*=1$ and $d^*=1$ for the minimum values of s and d, thus obtaining $\langle s\rangle_\alpha = \lambda_\alpha/(1-\lambda_\alpha)+1$ and $\langle s^2\rangle_\alpha = 2\langle s\rangle_\alpha^2 - \langle s\rangle_\alpha$. Similarly, $\langle d^2\rangle_\alpha = 2\langle d\rangle_\alpha^2 - \langle d\rangle_\alpha = 6$, with the second equality following from the fact that, for all structures, $\langle d\rangle = 2$. Equation (A1) now yields Equation (2) of the main text.