THE JOURNAL OF PHYSICAL CHEMISTRY B

ARTICLE

pubs.acs.org/JPCB

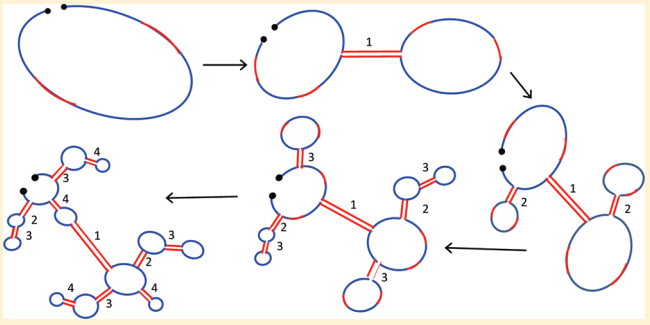
A Sequential Folding Model Predicts Length-Independent Secondary Structure Properties of Long ssRNA

Li Tai Fang,[†] Aron M. Yoffe,[‡] William M. Gelbart,[‡] and Avinoam Ben-Shaul*[,†]

[†]Institute of Chemistry and the Fritz Haber Research Center, The Hebrew University of Jerusalem, Givat Ram—Safra Campus, Jerusalem 91904, Israel

[‡]Department of Chemistry and Biochemistry, University of California, Los Angeles, 607 Charles E. Young Drive East, Los Angeles, California 90095-1569, United States

ABSTRACT: We introduce a simple model for folding random-sequence RNA molecules, arguing that it provides a direct route to predicting and rationalizing several average properties of RNA secondary structures. The first folding step involves identifying the longest possible duplex, thereby dividing the molecule into a pair of daughter loops. Successive steps involve identifying similarly the longest duplex in each new pair of daughter loops, with this process proceeding sequentially until the loops are too small for a viable duplex to form. Approximate analytical solutions are found for the average fraction of paired bases, the average duplex length, and the average loop size, all of which are shown to be independent of sequence length for long enough molecules. Numerical solutions to the model provide estimates for these average secondary structure properties that agree well with those obtained from more sophisticated folding algorithms. We also use the model to derive the asymptotic power law for the dependence of the maximum ladder distance on chain length.



1. INTRODUCTION

As a result of partial complementarity (base pairing) between the nucleotides (nt) constituting single-stranded (ss) RNAs, these molecules develop secondary structures composed of double-stranded (ds) "duplexes" of contiguous base pairs (bp) connected by ss loops of unpaired nt. These components in turn determine the tertiary structure of RNA molecules and thereby their biological function. Accordingly, a great deal of theoretical and experimental work has been devoted to predicting and measuring the secondary structures of RNAs (see, e.g., refs 1−12).

In particular, several dynamical programming algorithms have been developed and used widely to predict the secondary structures and associated free energies of ssRNA.[9,10] While sometimes at variance with respect to their prediction of detailed structural or energetic properties, such as the exact configuration of the minimum free energy (MFE) state of a given sequence of nt, the predictions of *coarse-grained* properties obtained using the different folding algorithms are generally in good agreement with each other. Among these properties are the fraction of bases in pairs, f, the average loop size, $\langle l \rangle$, and the average length of base-pair duplexes, $\langle k \rangle$, in the MFE structure or the Boltzmann-weighted ensemble of secondary structures. Moreover, upon further averaging over large sets of random sequences, many features arise that are even less sensitive to the folding algorithm or energy model used. Notably, upon increasing the length of the RNA chain (keeping the base composition constant), it is found that $\overline{f}$, $\overline{\langle l \rangle}$, and $\overline{\langle k \rangle}$ approach constant values. (The angled brackets are used, here and throughout the paper, to indicate within-structure averages. When overbars are used, they indicate further averaging over sets of sequences of the same length and base composition.) Further, these and several other generic attributes of the secondary structure have been treated previously, both theoretically and computationally, for several simple models of self-complementary chains.[11−18]

In the present work, we show that similar conclusions regarding coarse-grained properties of ssRNA can be derived from a particularly simple model of RNA folding, in which a single secondary structure is derived by means of a sequence of successive folding stages. This *sequential folding model* (SFM) explains, both qualitatively and quantitatively, why and how $\overline{f}$, $\overline{\langle l \rangle}$, and $\overline{\langle k \rangle}$, and related secondary structure attributes (of random sequences of a given base composition), approach constant values with increasing number of nt, N. On the basis of several approximations, we show that the model can be solved analytically to obtain simple closed-form expressions for these quantities. Numerical realizations of the SFM are presented, confirming the asymptotic N-independence of $\overline{f}$, $\overline{\langle l \rangle}$, and $\overline{\langle k \rangle}$; further, their values show reasonable agreement with those derived from the more accurate, and far more detailed, folding algorithms that explicitly introduce energies and entropies of duplex and loop

Received: November 8, 2010
Revised: December 27, 2010
Published: March 03, 2011

ACS Publications © 2011 American Chemical Society

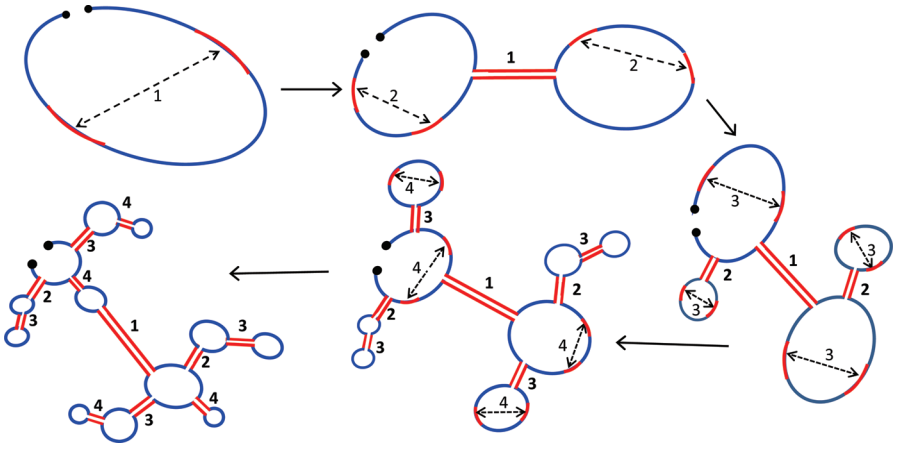


Figure 1. Schematic illustration of the sequential folding model. First, the longest duplex is formed (labeled "1"), resulting in the formation of two smaller daughter loops. These are then divided by their respective longest duplexes (each labeled "2"), and so on. Only four "generations" are depicted here. The process ends when none of the loops can be divided by a viable duplex. As bases may be apportioned unevenly between daughter loops, some loops may reach the end of the process in fewer generations than others. The analytical solution described in section 2 incorporates the simplifying assumption that all divisions are equal (hence all loops undergo the same number of divisions).

formation.[9,10] Other predictions of the model, such as the scaling with N of the maximum ladder distance (MLD),[5,17] show similarly good agreement, suggesting the reasonableness of the approximations on which the SFM is based. We emphasize, however, that the model is not proposed as an alternative to any of the (theoretically and numerically) more rigorous algorithms designed for RNA secondary structure prediction. Rather, its purpose is to provide, through a simple analysis of the folding of random sequences, insights into how generic asymptotic properties of secondary structure arise.

We open the discussion with a general description of the SFM, in section 2. In section 3 we present a highly simplified, closed-form, analytical solution of the SFM for long ssRNAs with random sequences. In section 4 we elaborate on the numerical solution of the SFM and compare its predictions to those derived using the Vienna RNA folding package.[10]

2. SEQUENTIAL FOLDING MODEL (SFM)

The basic premise of our SFM is that the secondary structure of an RNA molecule can be regarded as the final state in a sequence of folding events consisting of the successive division of ss loops by their longest possible duplexes. The procedure is schematically illustrated in Figure 1. The first folding stage involves the formation of two smaller daughter loops. In the case of a linear ssRNA, one of the daughter loops is open—containing the 5′ and 3′ ends—while the other is closed. Both daughter loops are closed if the chain is circular. In the second folding stage, both daughter loops are divided by their respective longest duplexes, each yielding a pair of still smaller loops, and so on. Clearly, successive folding generations yield successively smaller loops and concomitantly smaller maximal duplexes. The folding process ends when none of the loops is large enough to enable its further division by a viable duplex, typically assumed to consist of two bp. If the RNA chain is circular, all loops are closed; if it is linear, one of the loops in the final structure is open—this is the "exterior loop", which contains the 5′ and 3′ ends of the molecule.

In the course of this virtual folding process, certain loops may give rise to two (and very rarely three) duplexes of the same maximal length. When this happens in our numerical realization of the model (see section 3), the first of the maximal duplexes found is chosen. Note that, once a duplex is formed, it survives throughout the entire folding process; i.e., the model does not allow the structure to anneal. Moreover, every folding sequence yields a single final secondary structure. Alternative ("degenerate") final structures corresponding to a given nt sequence will generally arise upon repeating the folding process, because, as noted above, certain loops may yield more than one maximal duplex, which upon random sampling will generally lead to different final structures. These degeneracies are (indirectly) accounted for by sampling many (typically 100) random ssRNA sequences for any given molecular length and base composition of interest.

On the basis of detailed numerical calculations, we found that elaborations of the model—involving, say, the assignment of duplex energies—produce results that differ little from those derived using the simple version of the SFM outlined above. Explicit inclusion of loop entropies appears similarly unwarranted. In neglecting both loop entropies and duplex-specific energies, the SFM resembles the Nussinov−Jacobson model, where the MFE structure is the one of maximal base pairing.[12] However, whereas the Nussinov−Jacobson model identifies the secondary structure of global maximum pairing, the SFM follows one particular folding path whose every step is prescribed by the formation of the longest possible duplex remaining in each of the unfolded loops. In a comprehensive theoretical paper, David et al.[18] have recently described a hierarchical folding model for random N-base RNA sequences, in which the $(1/2)N(N-1)$ potential base pairs are ordered according to the magnitudes of their binding affinities. The energetically most favorable base pair is formed first, the second base pair is chosen to be the strongest among the remaining *allowed* pairs (i.e., those that are either independent of, or nested within, the first base pair), and the process is repeated until all possible base pairs have been formed. The SFM resembles this approach, but the hierarchy of structures generated involves the successive formation of duplexes rather than individual base pairs.

Although the SFM does not explicitly account for loop entropies, the formation of the longest duplex corresponding to a given loop provides the most efficient energetic means for overcoming

the entropy loss associated with the division of this loop into two smaller daughter loops. Hence, physically, the SFM may be regarded as mimicking the gradual cooling of a structureless high-temperature ssRNA chain. Indeed, in such a process, the longest duplex of the entire sequence will be likely to appear first, yielding two smaller loops that are then divided by their own longest duplexes, and so on, until the loops become too small to enable the formation of even the minimal stable duplex. Clearly, however, this simple and appealing "kinetic" scheme is necessarily approximate because it does not allow the evolving secondary structure to anneal and relax toward its optimal (MFE) configuration.

3. APPROXIMATE ANALYTICAL SOLUTION TO THE SFM

Let $i = 1, \ldots, N$ denote the chain of N nt constituting a linear ssRNA molecule, numbered from the 5′ end to the 3′ end. The same notation applies to a circular chain, but with $i = 1$ and N denoting an arbitrary pair of covalently bound nt. We shall use p $(0 < p < 1)$ to denote the average base-pairing probability between any two nt along the chain. For example, $p = 1/4 = 0.25$ if the ssRNA chain is comprised of equal proportions (25%) of the four bases G, C, A, and U, randomly distributed along the chain, with only equal-energy Watson−Crick (WC) pairs allowed, i.e., G with C and A with U. Similarly, $p = 3/8 = 0.375$ if G−U pairs are included with equal probability. We have performed detailed numerical calculations for both cases, obtaining similar qualitative conclusions regarding the N-independence of $\overline{f}$, $\overline{\langle l \rangle}$, and $\overline{\langle k \rangle}$. In the numerical calculations described in section 4, we shall therefore present the results only for $p = 3/8$.

Consider two arbitrary subsequences of the entire chain, $i, i+1, i+2, \ldots, i+k-1$ and $j, j+1, j+2, \ldots, j+k-1$ (with $j \geq i+k+m$), each containing k bases. Suppose that these are aligned against each other, with i facing $j+k-1$, $i+1$ facing $j+k-2$, etc.; m is some small number, corresponding to the minimal number of nt in a hairpin loop, often set to $m = 3$. The overall number, $A(k;N)$, of possible alignments of subsequences of length k is, to a very good approximation, given by $A(k;N) = (1/2)(N - 2k - m)^2$. A fraction of these alignments involves a succession of perfectly matched bases, enabling the formation of a duplex comprising k bp. For random sequences, if k is considerably smaller than N (and more generally, if the two subsequences are short compared to the entire sequence), the pairing probabilities of adjacent nt pairs are uncorrelated. Hence, the probability that two subsequences, each comprising k nt, will perfectly match, giving rise to a duplex of length k (or a part of a longer duplex), is given by p^k. Using $D(k;N)$ to denote the number of *matching* alignments of length k, it follows that—on average—for random sequences $D(k;N) = A(k;N)p^k$, which is a rapidly decreasing function of k.

Let k_1, generally not an integer, denote the solution of the equation $D(k_1;N) = A(k_1;N)p^{k_1} = 1$. In the ensemble of random sequences of length N, there is—on average—less than one potential duplex of length $k > k_1$ and more than one potential duplex of length $k < k_1$. Accordingly, we shall identify $k_1 = k_1(N)$ as the "most probable maximal duplex" in the ensemble. For sufficiently long sequences, i.e., ones for which $N \gg k, m$, we have $A(k;N) = (1/2)(N - 2k - m)^2 \approx (1/2)N^2$ so that $(1/2)N^2p^{k_1} = 1$, indicating that k_1 increases logarithmically with sequence length. More explicitly,

$$k_1 = a \ln N - b = \text{1st maximal duplex} \qquad (1)$$

For $p = 3/8$, we have $a = 2/\ln(1/p) \approx 2.04$ and $b = \ln(2)/\ln(1/p)$ ≈ 0.71. A plot of k_1 vs N calculated using this relationship is shown in Figure 2, revealing excellent agreement with the results obtained by the numerical simulations detailed in the next section, especially for large values of N.

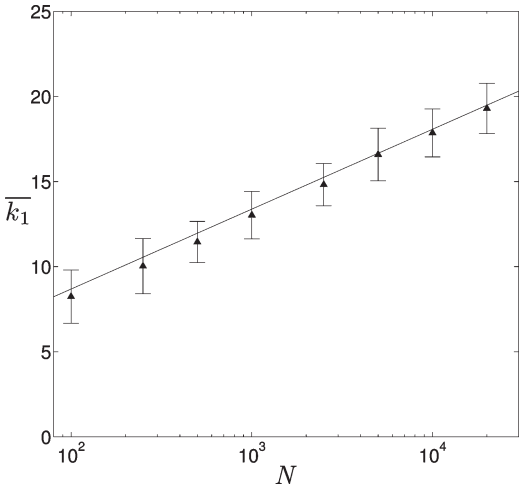


Figure 2. The longest duplex length as a function of sequence length. Triangles and error bars correspond to the numerical calculations discussed in section 4. The solid line is a plot of the theoretical expression $k_1 = a \ln N - b$, with $a = 2.04$, $b = 0.71$, as presented above in the text (see eq 1). All calculations are performed for $p = 3/8$.

In order to derive a closed-form expression for $\langle k \rangle$, we now introduce two rather drastic approximations. First, we assume that in each stage ("generation"), s, of the folding process all loops are divided into two daughter loops of equal length. In general, as evidenced by our numerical solution of the SFM in section 4, and illustrated schematically in Figure 1, each loop divides according to the sequence-dependent placement of its maximal duplex, which can occur in most positions. The symmetric division invoked here, however, enables a simple derivation of approximate, closed-form, results. It may be noted, as is also apparent from Figure 1, that eventually, as the loops get progressively smaller, the duplexes (which also become shorter) involve nt which are not far from each other along the chain. It follows that bases tend to pair with close-by bases because, as shown below, most base pairs are present in short duplexes which are, in turn, present in small loops. The probability of two bases being paired is thus an asymptotically decreasing function of the contour distance between them (whether loop divisions in the SFM are symmetric or not). Quantitative mathematical analysis of the scaling behavior of this pairing function, for an alternate folding model, has been presented by David et al.[18]

Because we are treating random sequences, loops containing the same number of nt possess, on average, equally long maximal duplexes. Thus the first loop, whose length is $N \equiv N_0$ nt, is divided by its longest duplex (of k_1 bp) into two loops, each containing $N_1 = (N_0 - 2k_1)/2 \approx N_0/2$ unpaired bases. Each of the two daughter loops is then divided by its own maximal duplex (of length k_2) into "granddaughter loops" of size $N_2 = [(N_0 - 2k_1)/2 - 2k_2]/2 \approx N_0/2^2$, and so on, until the loops become too small to enable the formation of a stable duplex and the folding sequence ends. We shall use $\hat{s}$ to denote the last stage of loop division, so that $k_{\hat{s}}$ is the shortest duplex in the structure and $N_{\hat{s}}$ is the size of the last and hence smallest loops formed in the process.

Our second approximation will be to extend the long-chain approximations $N_1 \approx N_0/2$, $N_2 \approx N_1/2 \approx N_0/2^2$, ... to all folding

stages, so that the loops in the sth generation of the folding sequence contain $N_s = N_0/2^s = N/2^s$ nt. In the same approximation, the length, k_s, of the maximal duplex in the sth generation is determined by $(1/2)(N_{s-1})^2 p^{k_s} = 1$, yielding $k_s = a \ln N_{s-1} - b$, with a and b as given above. Note that the long-chain approximation $N_s = N_{s-1}/2$ becomes poorer toward the end of the folding process, especially in the last stage when $2k_{\hat{s}}$ is not much smaller than its "mother loop" of size $N_{\hat{s}-1}$ (see additional remarks below).

From the approximate scheme outlined above, it follows that the maximal duplex length in the sth folding generation

$$k_s = a \ln N_{s-1} - b = k_1 - (a \ln 2)(s-1) \qquad (2)$$

decreases linearly with s. From eqs 1 and 2, we also conclude that the total number of loop division generations, $\hat{s}$, is a logarithmically increasing function of N. Setting $s = \hat{s}$ in eq 2, we find that $\hat{s} = \ln N/\ln 2 + 1 + (k_{\hat{s}} - b)/(a \ln 2)$, and since $k_{\hat{s}}$ is a constant dictated by the particular energy model used, it follows that $\hat{s} \sim \ln N$. In the standard folding algorithms (such as mfold[9,19] or RNAfold[10]), a stable duplex typically consists of at least 2 bp. According to eq 2, if for instance $p = 3/8$, then a duplex of length $k = 2$, say, can in principle be formed from a loop containing only $\exp[(2 + 0.71)/2.04] \approx 4$ nt. Note, however, that eq 2 is a poor approximation for *small* loops, such as those encountered in the final stages of our folding scheme. This is because N_{s-1}, which appears in this equation, does not account for the (minimally 4) nt contained in the duplex that defines the splitting of a mother loop, nor for the additional nt (at least two but typically more) appearing in the two smaller daughter loops. Finally, it also ignores the severe constraints on base pairing imposed by the "interference" of previous-generation duplexes (two on average) emanating from the loop. Among the consequences of this interference is that not all loops formed in a given generation undergo the same number of successive divisions, and not all loops in the ultimate structure are of the same size. Removing the simplifying assumption that loops are always divided equally would further add to differences in the final loop sizes and the number of divisions those loops underwent. All of these factors are taken into account in our numerical calculations in the next section, where indeed we find that the loops of the penultimate generation (which are not all of the same size) typically contain ~20 nt. The average duplex size in the last loop division is found to be nearly 3 bp (just bigger than the minimal allowed duplex length of 2 bp). Accordingly, the average number of unpaired bases in the loops of the final structure is, roughly, $(20 - 2 \times 3)/2 = 7$. In principle, we could improve the applicability of eq 2 to even the smallest loop sizes by a perturbative treatment that takes into account all the factors mentioned above. This would improve the numerical results predicted by the analytical approach but would not affect our qualitative conclusions, in particular the asymptotic (N-independent) behavior of $\langle k \rangle$, $\langle l \rangle$, and f, as clearly revealed by numerical solution of the SFM.

On the basis of the assumption that the number of duplexes is doubled upon any successive generation of loop division, an approximate expression for the average duplex length in the secondary structure resulting from the SFM can be derived as follows:

$$\langle k \rangle = \sum_{s=1}^{\hat{s}} 2^{s-1} k_s \Big/ \sum_{s=1}^{\hat{s}} 2^{s-1} \approx k_1 - (a \ln 2)(\hat{s} - 2)$$
$$= k_{\hat{s}} + a \ln 2 = a \ln N_{\hat{s}} + (2a \ln 2 - b) \qquad (3)$$

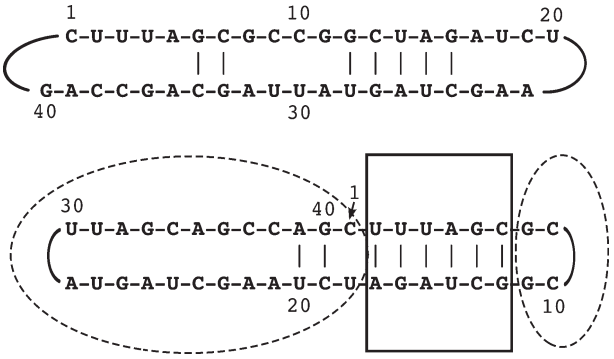


Figure 3. Sequence alignments used to identify the longest duplex. Shown here are two (of the 40 possible) alignments of a 40 nt chain containing equal numbers of randomly distributed G, C, A, and U, allowing G−C, A−U, and G−U pairing with equal probability. Two viable duplexes (of 2 and 5 bp) are found in the upper alignment. In the lower alignment, the longest duplex associated with this sequence (comprising 6 bp) is shown surrounded by a square box. The resulting daughter loops are encircled by dotted lines.

From this equation, it follows that $\langle k \rangle \approx k_{\hat{s}} + a \ln 2$, so that the average duplex length is not much larger than the smallest duplex length, e.g., $\langle k \rangle \approx k_{\hat{s}} + 1.41$ for $p = 3/8$. The N-independence of $\langle k \rangle$ now follows from the fact that the minimal duplex length, $k_{\hat{s}}$, is a local property dictated by the composition and pairing energies of the polynucleotide chain and is therefore independent of N. The exponential generational increase of bases in loops explains the small difference between $\langle k \rangle$ and $k_{\hat{s}}$. In addition, from the relationship $k_{\hat{s}} = a \ln N_{\hat{s}} + (a \ln 2 - b)$ in eq 3, we conclude that the smallest loop size, $N_{\hat{s}}$, is independent of N and thus, like $k_{\hat{s}}$, also a local property.

Let S denote the total number of duplexes in the final secondary structure and L the corresponding number of loops. For circular RNA, $L = S + 1$, which also holds for linear RNA if we add the exterior loop to the loop count (otherwise $L = S$). For the long molecules of interest here (i.e., for $N \gg 1$), we can safely put $L = S$, which holds for all secondary structures, not only the ultimate one. In terms of the pairing fraction, f, and the average duplex length, $\langle k \rangle$, we can express S in the form $S = Nf/2\langle k \rangle$. Analogously, the total number of loops can be expressed as the ratio, $L = N(1 - f)/\langle l_{ss} \rangle$, between the total number of unpaired nt in the molecule, $N(1 - f)$, and the average number of unpaired nt per loop in the final structure, $\langle l_{ss} \rangle$. In the approximate closed-form solution to the SFM presented above, all loops in the final structure comprise the same number of unpaired nt, corresponding to the approximation $\langle l_{ss} \rangle \approx N_{\hat{s}}$. Equating L and S, we find $f/(1 - f) = 2\langle k \rangle/\langle l_{ss} \rangle \approx 2\langle k \rangle/N_{\hat{s}}$, or

$$f = \frac{2\langle k \rangle}{\langle l_{ss} \rangle + 2\langle k \rangle} \approx \frac{2\langle k \rangle}{N_{\hat{s}} + 2\langle k \rangle} \qquad (4)$$

where it should be noted that the equality in eq 4 is always true, whereas the near-equality corresponds to the approximate closed-form solution to the SFM. Similarly general is the expression $\langle l \rangle = \langle l_{ss} \rangle + 2\langle d \rangle \approx \langle l_{ss} \rangle + 4$, where $\langle l \rangle$ is the average number of nt (paired or unpaired) per loop in any arbitrary secondary structure and $\langle d \rangle$ is the average number of duplexes connected to one loop. The above near-equality arises from the fact that, in every secondary structure, $\langle d \rangle = 2 - 2/L$, implying $\langle d \rangle \approx 2$ for large L (see, e.g., ref 20). In agreement with eq 4, our numerical solutions of the SFM (for long sequences of uniform base composition) yield $\overline{\langle k \rangle} \approx 5$, $\overline{\langle l_{ss} \rangle} \approx 6.5$, and $\overline{f} \approx 0.6$.

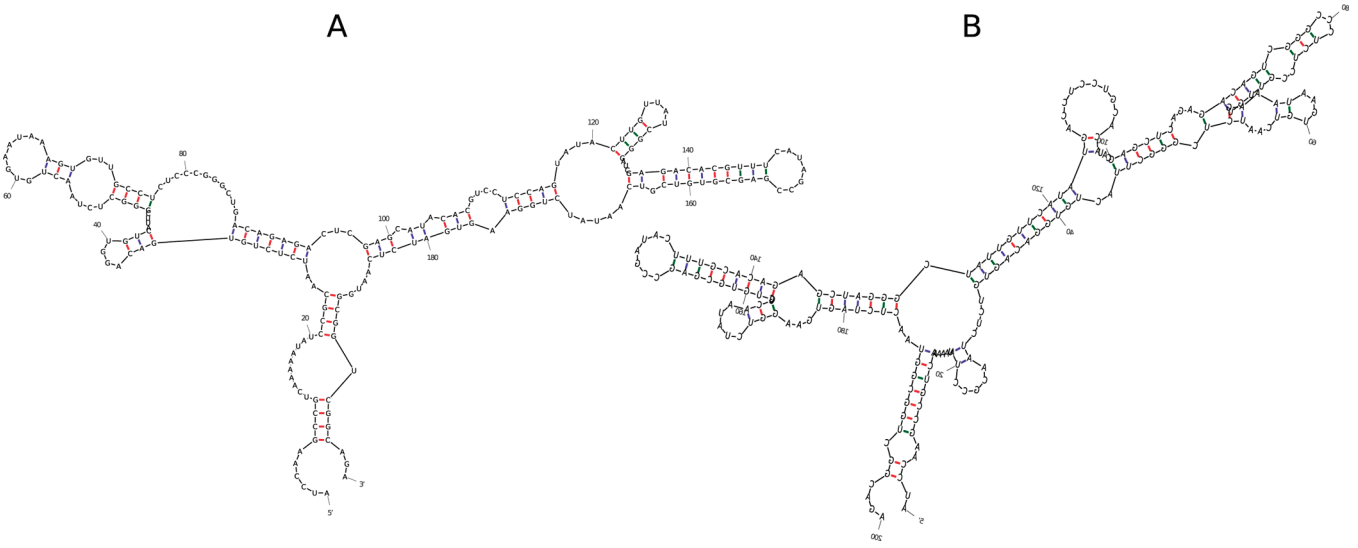


Figure 4. Secondary structures predicted for an arbitrary 200 nt RNA sequence of uniform composition. (A) The MFE structure predicted by RNAfold. (B) The SFM structure. Both structures were drawn using the mfold graphing utilities.[9]

From eqs 3 and 4 it follows that if, as argued above, $k_{\hat{s}}$ (equivalently $N_{\hat{s}}$) is independent of N, then, for asymptotically long sequences, $\langle k\rangle$, f, and $\langle l\rangle$ are also independent of N. In other words, these properties may be interpreted as "intensive" properties of random RNA molecules in the long-chain (or thermodynamic) limit. The conclusion that f is an intensive property also follows directly from the ansatz that, in this limit, the free energy of folding of a long RNA molecule is an extensive property.[21] Equation 4 implies that from this ansatz $\langle k\rangle$ and, equivalently, $\langle l\rangle$ are intensive properties as well. The derivation of these qualitative conclusions has been the main goal of the approximate analytical solution presented here for the SFM. However, given the many simplifications made along the way, we do not expect the relations given by eqs 3 and 4 to be *numerically* accurate. Accordingly, in the next section, we present a numerical solution to the SFM for several secondary structure properties of random sequences of varying length, and compare them to MFE structures predicted by RNAfold, a program in the Vienna RNA Package, version 1.83.[10] We show there that the asymptotic independence of $\langle k\rangle$, f, and $\langle l\rangle$ from N, predicted above, is clearly exhibited by our numerical solutions of the SFM. Finally, we use these numerical solutions to derive the large-N scaling behavior of the maximum ladder distance,[5,17] obtaining good agreement with results from the RNAfold calculations.

4. NUMERICAL RESULTS AND DISCUSSION

Computations were performed based on the scheme outlined in Figure 1. Namely, the entire chain is first split by its longest possible duplex into two daughter loops (generally of different lengths); these are then further divided by their own longest duplexes into still smaller loops, etc. The process ends when none of the loops sustains a viable duplex, which in all calculations was assumed to consist of at least two contiguous base pairs. The longest duplex of any given loop is found by scanning all possible intrachain alignments, as illustrated in Figure 3. This duplex is chosen to split the given loop. If two or more duplexes are found to have the same maximal length, we pick the first one found; for sufficiently large sets of random sequences, this is equivalent to a random selection. Calculations were performed for randomly permuted sequences of varying lengths (N = 100−20 000) of uniform composition (i.e., containing equal proportions [25%] of the four bases). We have analyzed sequences allowing G−C, A−U, and G−U bp, all with equal weight (corresponding to p = 3/8).

In Figure 4, we show the MFE structure predicted by RNAfold (see A), and the structure predicted by the SFM (see B), for an arbitrary 200 nt RNA sequence of uniform composition. The two structures appear qualitatively similar in terms of their distributions of duplexes and loops of varying size and complexity. For instance, one MLD (see definition below) is 39 (A) and the other 41 (B); the central loop of each is composed of approximately the same nucleotides (involving nt numbers around 25, 95, and 185); and the external loop in each is separated from this central loop by a single bubble. The average duplex length ($\langle k\rangle$) is 4.1 in A and 4.7 in B, and the pairing fractions (f) are 0.53 and 0.61, respectively. More comprehensive, statistically significant, comparisons between the two numerical procedures are reported in Figures 5−7.

Figures 5 and 6 show, respectively, the set-average pairing fraction and the set-average duplex length as a function of sequence length. With the SFM we analyzed sets of 100 randomly permuted sequences, of uniform composition, of each of the following lengths: N = 100, 250, 500, 1000, 2500, 5000, 10 000, and 20 000. The same sequence lengths were analyzed with RNAfold, except the set sizes were reduced to 20, and the longest chain length (N = 20 000) was omitted, because of RNAfold's increased computation time; these sequences constituted a subset of those analyzed by the SFM. The values of $\overline{f}$ predicted by the SFM and by RNAfold reveal surprisingly, possibly fortuitously, good agreement. The SFM predictions of the average duplex length, $\overline{\langle k\rangle}$, are higher, though not by much, than those predicted by RNAfold. This difference is not surprising since the SFM is biased toward long duplexes, whereas many other factors, such as loop entropy, are incorporated into RNAfold. The most significant result conveyed by the SFM calculations in Figures 5 and 6 is a qualitative one: the convergence of $\overline{\langle k\rangle}$ and $\overline{f}$ to their respective constant values for $N \gg 1$, as predicted by the simple analytical scheme outlined in the previous section, and in agreement with results derived using RNAfold.

dx.doi.org/10.1021/jp110680e |*J. Phys. Chem. B* 2011, 115, 3193–3199

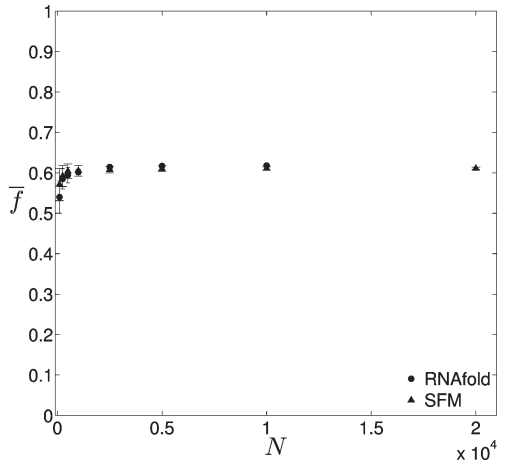


Figure 5. The average fraction of nucleotides in base pairs, $\overline{f}$, as a function of sequence length, N, for randomly permuted ssRNAs of uniform composition, calculated using RNAfold (circles) and the SFM (triangles).

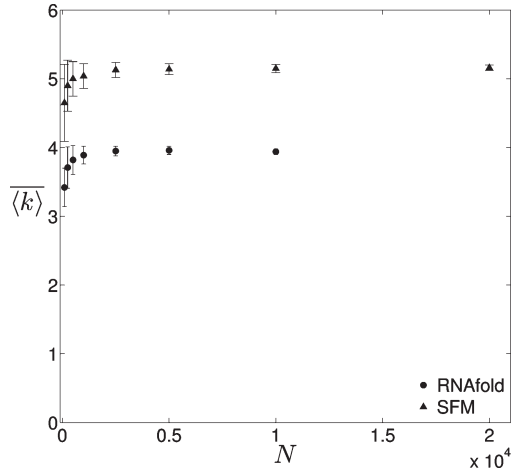


Figure 6. The average duplex lengths, $\overline{\langle k \rangle}$, as a function of sequence length, N, for randomly permuted ssRNAs of uniform composition, calculated using RNAfold (circles) and the SFM (triangles).

Originally introduced by Bundschuh and Hwa,[5] the ladder distance (LD) between two arbitrary bases i and j in a given secondary structure, ζ_{ij}, is defined as the number of duplex rungs that are crossed by the (unique and shortest) path connecting these pairs of bases. The *maximum* ladder distance (MLD), $\zeta = \mathrm{Max}\{\zeta_{ij}\}$, is the longest ladder distance within the structure.[17] It provides a convenient measure for the size of the secondary structure. In analogy to $\overline{\langle k \rangle}$, $\overline{f}$, etc., we denote by $\overline{\zeta}$ the average MLD corresponding to a given set of nt sequences, all of the same base composition and chain length. In Figure 7, we show $\overline{\zeta}$ as a function of N for the same sets of structures used to derive the results shown in Figures 5 and 6.

The results in Figure 7 reveal a nearly linear dependence of $\ln \overline{\zeta}$ on $\ln N$, indicating the power law behavior $\overline{\zeta} \sim N^{\alpha}$, with $\alpha \approx 0.67$ for the SFM and $\alpha \approx 0.7$ for RNAfold. These exponents are in similarly good agreement with those from previous, more extensive, MLD calculations. Specifically, in a recent study,[20] several thousand randomly permuted sequences were sampled and analyzed using RNAsubopt (a program in the Vienna RNA folding package[22]), yielding the 2-fold average $\overline{\zeta}_{\mathrm{th}}$. Here, the "th" (for "thermal") subscript denotes the average over the ensemble of Boltzmann-weighted secondary structures associated with each sequence, and the overbar denotes the further averaging over many randomly permuted sequences of the same length and base composition. The scaling behavior was approximately the same as that seen above: $\overline{\zeta}_{\mathrm{th}} \sim N^{\alpha}$ with $\alpha = 0.69 \pm 0.01$. Similar calculations have previously[17] been performed for randomly permuted sequences with a slightly nonuniform composition (24% G, 22% C, 26% A, and 28% U), in which case it was found that $\alpha = 0.67 \pm 0.01$.

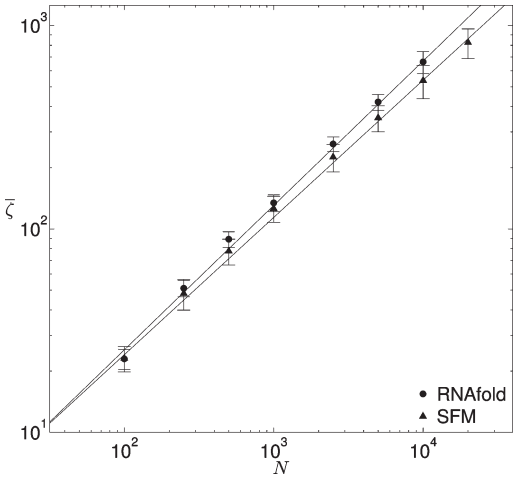


Figure 7. The average maximum ladder distance, $\overline{\zeta}$, as a function of sequence length, N, for randomly permuted ssRNAs of uniform composition, calculated using RNAfold[10] (circles) and the SFM (triangles). The vertical bars indicate the range of $\overline{\zeta}$ values corresponding to the 100 sequences analyzed with the SFM and the 20 sequences analyzed using RNAfold. Both models exhibit a linear dependence of $\ln \overline{\zeta}$ on $\ln N$, with slopes of $\alpha \approx 0.67$ for the SFM and $\alpha \approx 0.70$ for RNAfold.

5. CONCLUDING REMARKS

The sequential folding model (SFM), presented here for randomly permuted RNA sequences, provides a simple basis for understanding several generic features of the average properties of their secondary structures. In particular, we are able to obtain closed-form analytical solutions for these properties by introducing two approximations: (1) in each successive folding step, the identification of a maximal duplex leads to daughter loops of equal size; (2) the number of duplexes generated by the sth step is equal to 2^{s-1}, for all steps in the folding scheme. The length-independence of the average duplex size ($\langle k \rangle$), loop size ($\langle l \rangle$), and fraction of bases paired (f) follows straightforwardly from these analytical solutions. Furthermore, numerical solutions to the SFM, for these average properties, give reasonably good agreement with those obtained from RNAfold, which is a significantly more sophisticated folding algorithm. The numerical SFM prediction for the large-N scaling behavior of the maximum ladder distance also agrees with that obtained from RNAfold.

AUTHOR INFORMATION

Corresponding Author

*E-mail: abs@fh.huji.ac.il.

■ ACKNOWLEDGMENT

We thank the support of the Israel Science Foundation (grants 695/06 and 1448/10 to L.T.F. and A.B.-S.), the Archie and Marjorie Sherman Chair (A.B.-S.), and the U.S. National Science Foundation (grant CHE07-14411 to W.M.G.).

■ REFERENCES

(1) Capriotti, E.; Marti-Renom, M. A. *Curr. Bioinf.* **2008**, *3*, 32.
(2) Onoa, B.; Tinoco, I. *Curr. Opin. Struct. Biol.* **2004**, *14*, 374.
(3) Deigan, K. E.; Li, T. W.; Mathews, D. H.; Weeks, K. M. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 97.
(4) Lilley, D. M. J. *Biopolymers* **1998**, *48*, 101.
(5) Bundschuh, R.; Hwa, T. *Phys. Rev. E* **2002**, *65*, 031903.
(6) Holbrook, S. R. *Ann. Rev. Biophys.* **2008**, *37*, 445.
(7) Thirumalai, D; C. Hyeon, C. *Biochemistry* **2005**, *44*, 4957.
(8) Mathews, D. H. *RNA* **2004**, *10*, 1178.
(9) Zuker, M. *Nucleic Acids Res.* **2003**, *31*, 3406.
(10) Hofacker, I. L.; Fontana, W.; Stadler, P. F.; Bonhoeffer, L. S.; Tacker, M.; Schuster, P. *Monatsh. Chem.* **1994**, *125*, 167.
(11) Hofacker, I. L.; Schuster, P.; Stadler, P. F. *Discr. Appl. Math.* **1998**, *88*, 207.
(12) Nussinov, R.; Jacobson, A. B. *Proc. Natl. Acad. Sci. U.S.A.* **1980**, *77*, 6309.
(13) Müller, M. *Phys. Rev. E* **2003**, *67*, 021914.
(14) de Gennes, P. G. *Biopolymers* **1968**, *6*, 715.
(15) Clote, P.; Kranakis, E.; Krizanc, D.; Stacho, L. *Discr. Appl. Math.* **2007**, *155*, 759.
(16) Nussinov, R.; G. Pieczenik, J. R.; Griggs, R.; Kleitman, D. J. *SIAM J. Appl. Math.* **1978**, *35*, 68.
(17) Yoffe, A. M.; Prinsen, P.; Gopal, A.; Knobler, C. M.; Gelbart, W. M.; Ben-Shaul, A. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 16153.
(18) David, F.; Hagendorf, C.; Wiese, K. J. *J. Stat. Mech.* **2008**, P04008.
(19) Mathews, D. H.; Sabina, J.; Zuker, M.; Turner, D. H. *J. Mol. Biol.* **1999**, *288*, 911.
(20) Yoffe, A. M.; Prinsen, P.; Gelbart, W. M.; Ben-Shaul, A. *Nucleic Acids Res.* **2011**, *39*, 292.
(21) Klopper, A. V.; Bois, J. S.; Grill, S. W. *Phys. Rev. E* **2010**, *81*, 030904(R).
(22) Wuchty, S.; Fontana, W.; Hofacker, I. L.; Schuster, P. *Biopolymers* **1999**, *49*, 145.