IRREDUCIBILITY IN RNA STRUCTURES

EMMA Y. JIN [*] AND CHRISTIAN M. REIDYS [*,†]

ABSTRACT. In this paper we study irreducibility in RNA structures. By RNA structure we mean RNA secondary as well as RNA pseudoknot structures. In our analysis we shall contrast random and minimum free energy (mfe) configurations. We compute various distributions: of the numbers of irreducible substructures, their locations and sizes, parameterized in terms of the maximal number of mutually crossing arcs, $k-1$, and the minimal size of stacks σ. In particular, we analyze the size of the largest irreducible substructure for random and mfe structures, which is the key factor for the folding time of mfe configurations.

1. INTRODUCTION AND BACKGROUND

In this paper we study irreducibility in RNA structures. Intuitively, an irreducible substructure over a subsequence is a configuration of bonds, beginning and ending with arcs of certain stack size, that cannot be written as a nontrivial concatenation of smaller configurations. Since any minimum free energy (mfe) folding algorithm depends at least polynomially (to a degree larger than one) on the sequence length, the size of the largest, irreducible substructure determines the folding time.

Let us begin by recalling some basic facts about RNA structures: an RNA structure is the helical configuration of its primary sequence, i.e. the sequence of nucleotides **A**, **G**, **U** and **C**, together with Watson-Crick (**A-U**, **G-C**) and (**U-G**) base pairs. One well-known class of RNA structures, are RNA secondary structures, pioneered three decades ago by Waterman [11, 16, 17, 2, 18]. Secondary

structures exhibit exclusively noncrossing bonds and are subject to specific minimum arc-length conditions. They can readily be identified with Motzkin-paths satisfying some minimum height and plateau-length, see Figure 1 [18]. The latter restrictions come from biophysical constraints due
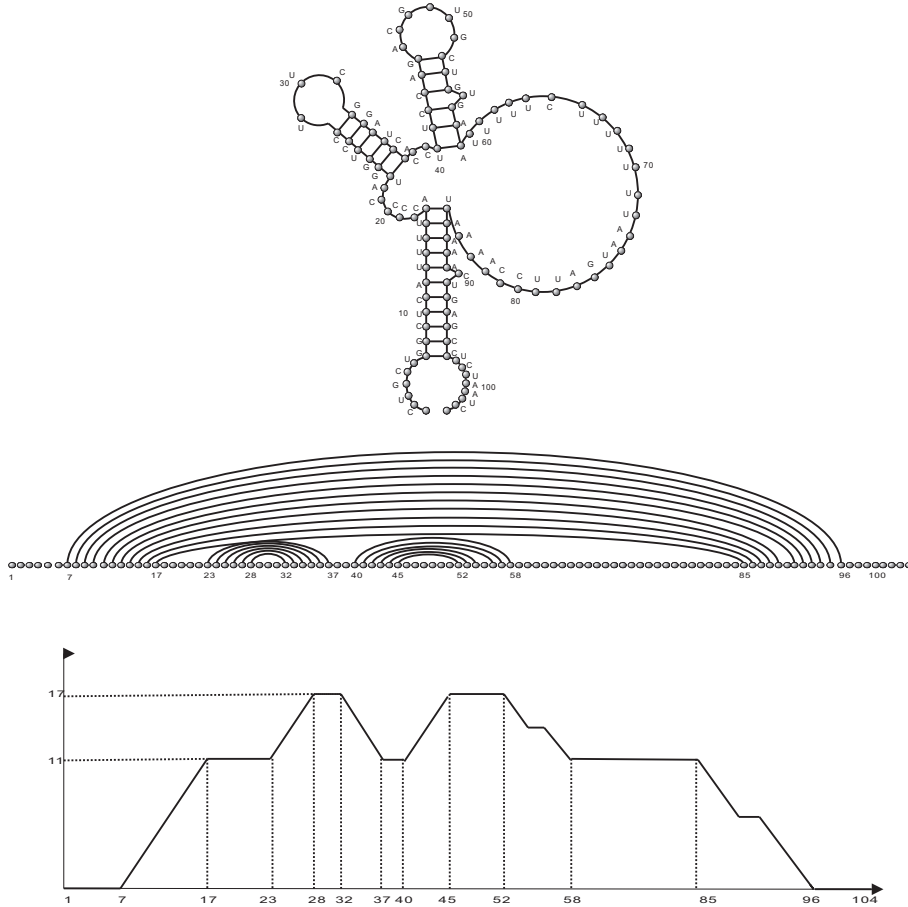

FIGURE 1. The RNA secondary structure of the 3'-UTR of subnuclei mRNA, represented as planar graph, diagram and Motzkin-path.

to mfe loop-energy parameters and limited flexibility of bonds. It is clear from the above bijection, that irreducible substructures in RNA secondary structures are closely related to the number of nontrivial returns, i.e. the number of non-endpoints, for which the Motzkin-path meets the x-axis. As a purely combinatorial problem this has been studied by [1, 7].

It is well-known that RNA configurations are far more complex than secondary structures: they exhibit additional, cross-serial nucleotide interactions [12]. These interactions were observed in natural RNA structures, as well as via comparative sequence analysis [19]. They are called pseudoknots, see Figure 2, and widely occur in functional RNA, like for instance, eP RNA [9] as well as ribosomal RNA [8]. RNA pseudoknots are conserved also in the catalytic core of group I introns. In plant viral RNAs pseudoknots mimic tRNA structure and in vitro RNA evolution [14] experiments have produced families of RNA structures with pseudoknot motifs, when binding HIV-1 reverse transcriptase.
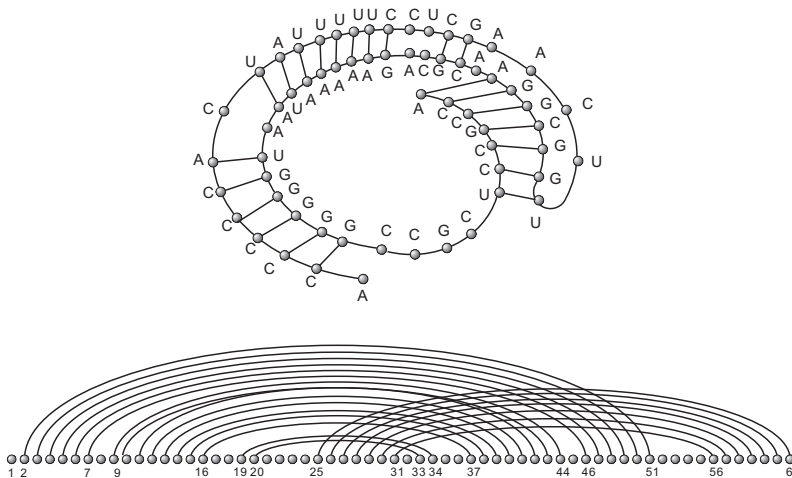

FIGURE 2. The mRNA plasmid pMU720 (IncB)-pseudoknot structure: its planar graph (top) and diagram representation (bottom).

Combinatorially, cross serial interactions are tantamount to crossing bonds. Therefore, RNA pseudoknot structures have been modeled as k-noncrossing (k-(nc)) diagrams [5, 6, 10], i.e. labeled graphs over the vertex set $[n] = \{1, \ldots, n\}$ with degree ≤ 1. Diagrams are represented by drawing their vertices $1, \ldots, n$ in a horizontal line and their arcs (i, j), where $i < j$, in the upper half-plane. Here the degree of i refers to the number of non-horizontal arcs incident to i, i.e. the backbone of the primary sequence is not considered. The vertices and arcs correspond to nucleotides and Watson-Crick (**A-U**, **G-C**) and (**U-G**) base pairs, respectively, see Figure 3. Diagrams are characterized via their maximum number of mutually crossing arcs, $k-1$, their minimum arc-length, λ, and their minimum stack-length, σ. A k-crossing is a set of k distinct arcs $(i_1, j_1), (i_2, j_2), \ldots (i_k, j_k)$ with the property $i_1 < i_2 < \ldots < i_k < j_1 < j_2 < \ldots < j_k$. A diagram without any k-crossings is
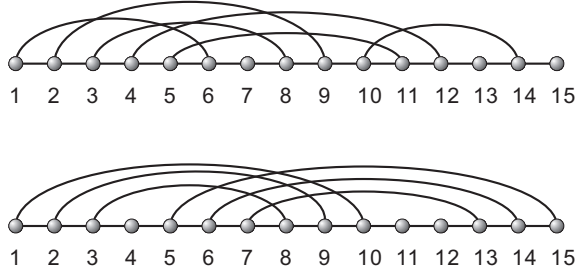
FIGURE 3. k-(nc) diagrams: we display a 4-(nc), arc-length $\lambda \geq 4$ and $\sigma \geq 1$ diagram (top), where the edge set $\{(1,6),(2,9),(4,12)\}$ is a 3-crossing, the arc $(10,14)$ has length 4 and $(1,6)$ has stack-length 1. Below, we display a 3-(nc), $\lambda \geq 5$ and $\sigma \geq 3$ (lower) diagram, where $(3,8)$ has arc-length 5 and the stack $((1,10),(2,9),(3,8))$ has stack-length 3.

called a k-(nc) diagram. The length of an arc (i,j) is $j-i$ and a stack of length σ is a sequence of "parallel" arcs of the form

$$((i,j),(i+1,j-1),\ldots,(i+(\sigma-1),j-(\sigma-1))).$$

A subdiagram of a k-(nc) diagram is a subgraph over a subset $M \subset [n]$ of consecutive vertices that starts with an origin and ends with a terminus of some arc. Let $(i_1,\ldots,i_m)$ be a sequence of isolated points, and (j_1,j_2) be an arc. We call $(i_1,\ldots,i_m)$ interior if and only if there exists some arc (j_1,j_2) such that $j_1 < i_1 < i_m < j_2$ holds and exterior, otherwise. Any exterior sequence of consecutive, isolated vertices is called a gap. A diagram and subdiagram is called irreducible, if it cannot be decomposed into a (nontrivial) sequence of gaps and subdiagrams, see Figure 4. As a result, any k-(nc) diagram can be uniquely decomposed into an alternating sequence of gaps and irreducible subdiagrams. We call a k-(nc), σ-canonical (σ-(ca)) diagram with arc-length ≥ 4 and stack-length $\geq \sigma$, a k-(nc), σ-(ca) RNA structure, see Figure 3. We accordingly adopt the notions of gap, substructure and irreducibility for RNA structures. A k-(nc), σ-(ca) RNA structure has return at position i if i is the endpoint of some irreducible substructure, see Figure 5. Unique large irreducible substructures are quite common for natural RNA pseudoknot structures, see Figure 6. The size of the largest irreducible substructure is typically very large: it contains almost all nucleotides, see Figure 7.

The paper is organized as follows: in Section 2 we recall some combinatorial framework due to [7]. In particular, we derive the probability generating function for the number of irreducible substructures. We remark that the framework presented in Section 2 can be generalized to RNA
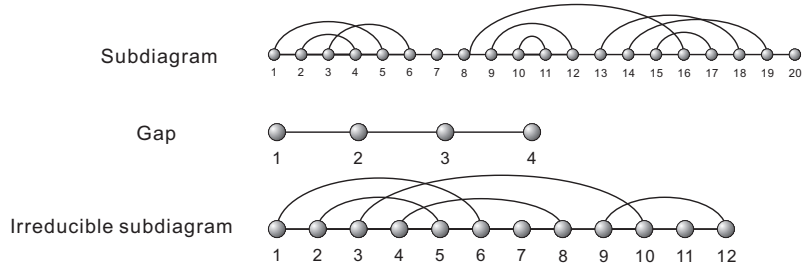
FIGURE 4. Subdiagrams, gaps and irreducibility: a diagram (top), decomposed into the subdiagram over $(1, 6)$, the gap 7 and the subdiagram over $(8, 19)$ and gap 20. A gap (middle) and an irreducible diagram over $(1, 12)$ (bottom).
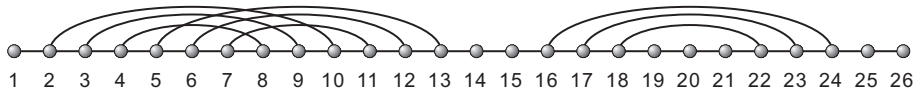

FIGURE 5. A 3-(nc), 3-(ca) RNA structure has returns at position 13 and 24, respectively.
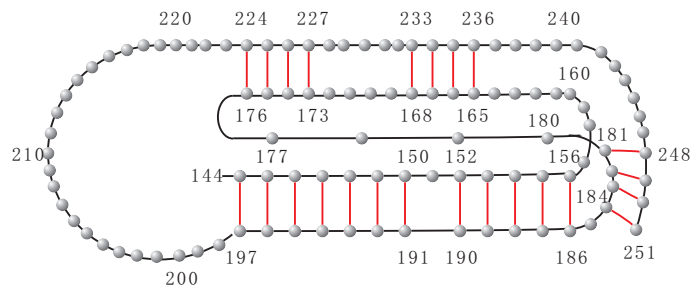

FIGURE 6. mRNA-Ec$_\alpha$: the irreducible pseudoknot structure of the regulatory region of the α ribosomal protein operon.

tertiary structures. In Section 3 we put these results to the test: we shall compare random and mfe structures. We begin by observing specific deviations of the distributions of mfe 2- and 3-(nc) structures for $n = 75$ and $\sigma = 3$ from that of random structures. The rest of the section we analyze these deviations and prove in the process (Proposition 1) a "shift"-result. The latter allows us to understand the effect of increasing the stack-size σ on irreducibility. In Section 4 we derive simple formulas to the probabilities of return locations, i.e. the endpoints of irreducible substructures and contrast random and mfe secondary and pseudoknot structures. In Section 5 we study the size of the largest irreducible components for k-(nc), σ-(ca) RNA structures.
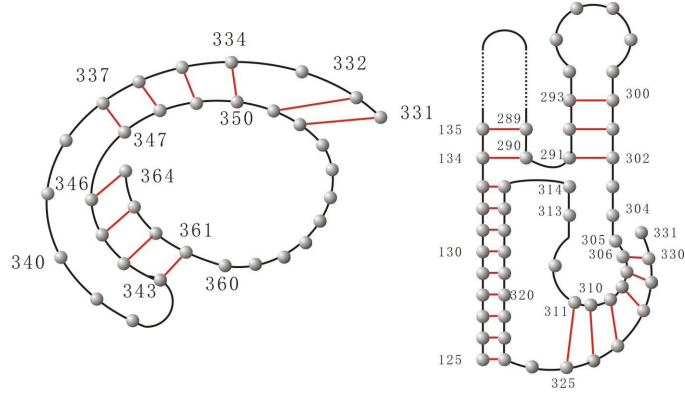
FIGURE 7. Hepatitis C Virus (left) and Hepatitis C Virus-IRES (right): the largest irreducible substructure is of size almost n.

2. SOME COMBINATORICS

Let $\delta_{n,j}^{(k,\sigma)}$ denote the number of k-(nc), σ-(ca) RNA structures, containing exactly j irreducible substructures and let $\delta_n^{(k,\sigma)} = \sum_{j \geq 0} \delta_{n,j}^{(k,\sigma)}$. That is, $\delta_n^{(k,\sigma)}$ denotes the number of k-(nc), σ-(ca) RNA structures. The bivariate generating function of the $\delta_{n,j}^{(k,\sigma)}$ indexed by j, the number of irreducible substructures and n, the sequence length is given by

$$(2.1) \qquad \mathbf{U}_{k,\sigma}(z,u) = \sum_{n \geq 0} \sum_{j \geq 0} \delta_{n,j}^{(k,\sigma)} u^j z^n.$$

Let furthermore $\mathbf{T}_{k,\sigma}(z) = \sum_{n \geq 0} \delta_n^{(k,\sigma)} z^n$ and $\mathbf{R}_{k,\sigma}(z)$ denote the generating function of irreducible RNA structures. The following lemma [7] derives the generating function $\mathbf{U}_{k,\sigma}(z,u)$:

Lemma 1. *The bivariate generating function of the number of k-(nc), σ-(ca) RNA structures, which contain exactly j irreducible k-(nc), σ-(ca) RNA substructures, is given by*

$$\mathbf{U}_{k,\sigma}(z,u) \quad = \quad \frac{\frac{1}{1-z}}{1 - u\left(1 - \frac{1}{(1-z)\mathbf{T}_{k,\sigma}(z)}\right)}.$$

Lemma 1 is the key for computing the limit distribution of the number of k-(nc), σ-(ca) RNA structures that have exactly j irreducible RNA substructures. For this purpose, let $\xi_n^{(k,\sigma)}$ be the

r.v. having the probability distribution

$$(2.2) \qquad \mathbb{P}(\xi_n^{(k,\sigma)} = j) = \frac{\delta_{n,j}^{(k,\sigma)}}{\delta_n^{(k,\sigma)}}.$$

The theorem below [7] shows that the probabilities $\mathbb{P}(\xi_n^{(k,\sigma)} = j)$ satisfy a discrete limit law:

Theorem 1. *Let $\alpha_{k,\sigma}$ be the real positive dominant singularity of $\mathbf{T}_{k,\sigma}(z)$ and*

$$\tau_{k,\sigma} = 1 - \frac{1}{(1 - \alpha_{k,\sigma})\mathbf{T}_{k,\sigma}(\alpha_{k,\sigma})}.$$

Then the r.v. $\xi_n^{(k,\sigma)}$ satisfies the discrete limit law

$$(2.3) \qquad \lim_{n \to \infty} \mathbb{P}(\xi_n^{(k,\sigma)} = i) = \frac{(1 - \tau_{k,\sigma})^2}{\tau_{k,\sigma}} i \tau_{k,\sigma}^i.$$

That is, $\xi_n^{(k,\sigma)}$ is determined by the density function of a $\Gamma(-\ln \tau_{k,\sigma}, 2)$-distribution. Furthermore, the probability generating function probability generating function of the limit distribution is given by

$$q(u) = \sum_{n \geq 1} \mathbb{P}(\xi_n^{(k,\sigma)} = i)u^i = \frac{u(1 - \tau_{k,\sigma})^2}{(1 - \tau_{k,\sigma}u)^2}.$$

The generating function of k-(nc), σ-(ca) RNA structures, $\mathbf{T}_{k,\sigma}(z)$ and its dominant singularities, $\alpha_{k,\sigma}$, have been studied in [5, 6, 10]. In particular the limiting probability of irreducible RNA structures is given by

$$(2.4) \qquad \lim_{n \to \infty} \mathbb{P}(\xi_n^{(k,\sigma)} = 1) = (1 - \tau_{k,\sigma})^2.$$

We observe that for fixed σ and increasing crossing number, k, the singularity $\tau_{k,\sigma}$ decreases. Therefore the limiting probability of RNA structures to be irreducible increases with increasing crossing number. However, for fixed k and increasing σ, the singularity $\tau_{k,\sigma}$ increases. Consequently, the limiting probability of RNA structures to be irreducible decreases with increasing σ. Theorem 1 allows to compute the characteristic function of the r.v. $\xi_n^{(k,\sigma)}$:

$$\mathbb{E}[e^{it\xi_n^{(k,\sigma)}}] = q(e^{it}) = \frac{e^{it}(1 - \tau_{k,\sigma})^2}{(1 - \tau_{k,\sigma}e^{it})^2}.$$

By Taylor expansion of the characteristic function we obtain the k-th moments of $\xi_n^{(k,\sigma)}$, i.e.,

$$(2.5) \quad \mathbb{E}[e^{it\xi_n^{(k,\sigma)}}] = 1 + (it)\mathbb{E}[\xi_n^{(k,\sigma)}] + \frac{(it)^2}{2!}\mathbb{E}[(\xi_n^{(k,\sigma)})^2] + \cdots + \frac{(it)^m}{m!}\mathbb{E}[(\xi_n^{(k,\sigma)})^m] + o(t).$$

Consequently, we can compute expectation and variance of $\xi_n^{(k,\sigma)}$ for varying k and σ, see Table 1 and Table 2: Table 1 shows, that for RNA secondary structures increasing the stack size σ

TABLE 1. $k = 2$

	$\tau_{2,\sigma}$	$\mathbb{E}[\xi_n^{(2,\sigma)}]$	$\mathbb{V}[\xi_n^{(2,\sigma)}]$
$\sigma = 3$	0.3201	1.9416	5.1548
$\sigma = 4$	0.3441	2.0492	5.7991
$\sigma = 5$	0.3615	2.1323	6.3203

TABLE 2. $k = 3$

	$\tau_{3,\sigma}$	$\mathbb{E}[\xi_n^{(3,\sigma)}]$	$\mathbb{V}[\xi_n^{(3,\sigma)}]$
$\sigma = 3$	0.0167	1.0340	1.1036
$\sigma = 4$	0.0208	1.0425	1.1302
$\sigma = 5$	0.0244	1.0500	1.1538

significantly increases reducibility. Table 2 indicates that 3-(nc) RNA pseudoknot structures are typically irreducible with rather subtle dependence on the minimum stack-size, σ.

3. RNA RANDOM STRUCTURES AND RNA MFE STRUCTURES

In this section we analyze irreducibility in random and mfe RNA secondary and pseudoknot-structures. As folding algorithms for the generation of the mfe RNA secondary and pseudoknot structures we employ Vienna RNA [15] and cross [3]. We shall begin by comparing in Figure 8 irreducibility of 2-(nc) and 3-(nc) random and mfe structures (of length $n = 75$) for minimum stack size $\sigma = 3$. Figure 8 shows that: (a) for 2-(nc) 3-(ca) structures the mfe structures are more irreducible than their random counterparts and (b) for 3-(nc) structures the contrary is being observed: 3-(nc) mfe structures are less irreducible than 3-(nc) random structures.

In order to understand the above observations, let us proceed by analyzing first the effect of increasing the minimum stack-size σ for 2- and 3-(nc) mfe structures. While we have shown in Section 2 that the limit distribution shifts towards *less* irreducibility when increasing σ, Figure 9 shows that for fixed σ, 2-(nc) as well as 3-(nc) mfe structures become *less* irreducible when the sequence length n increases. Intuitively, the increase in σ for fixed n implies that any irreducible substructures has to become larger. Therefore, in light of the fact that there are only a few irreducible substructures, disallowing for these small irreducible substructures implies the shift towards irreducibility. With this picture in mind, we shall proceed by quantifying this phenomenon: let

$$R_{k,\sigma}(n) = [z^n]\mathbf{R}_{k,\sigma}(z) = [z^n]\left[-\frac{1}{\mathbf{T}_{k,\sigma}(z)}\right],$$
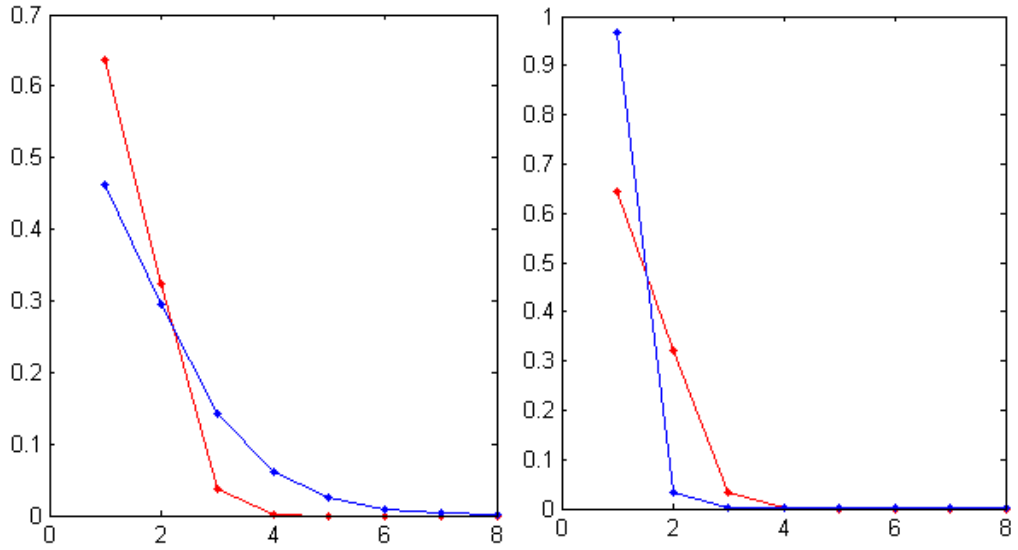
FIGURE 8. Random versus mfe: the lhs shows the distribution of irreducibles in 2-(nc), 3-(ca) mfe (red) and random structures (blue), for $n = 75$. The rhs showcases these distributions for 3-(nc), 3-(ca) mfe (red) and random structures (blue) for $n = 75$.

i.e. $R_{k,\sigma}(n)$ denotes the number of irreducible structures over n nucleotides and $\beta_{k,\sigma,j}(n)$ denotes the number of k-(nc), σ-(ca) RNA structures with exactly j irreducible substructures. Then, clearly, $R_{k,\sigma}(n) < \beta_{k,\sigma,1}(n)$. We shall prove that, for sufficiently large n, the scaling factor needed for passing from σ to $\sigma + 1$ is $\frac{\ln(\alpha_{k,\sigma})}{\ln(\alpha_{k,\sigma+1})}$.

Proposition 1. *For sufficiently large n and arbitrary σ, we have*

$$(3.1) \qquad R_{k,\sigma}(n) \sim \left(\frac{\ln(\alpha_{k,\sigma})}{\ln(\alpha_{k,\sigma+1})} \right)^{-\mu} R_{k,\sigma+1} \left(\left[\frac{\ln(\alpha_{k,\sigma})}{\ln(\alpha_{k,\sigma+1})} \right] n \right).$$

Furthermore, $\beta_{k,\sigma,j}(n)$ satisfies

$$(3.2) \qquad \beta_{k,\sigma+1,j} \left[\frac{\ln(\alpha_{k,\sigma})}{\ln(\alpha_{k,\sigma+1})} \cdot n \right] \approx \beta_{k,\sigma,j}(n).$$

In order to illustrate Proposition 1, we consider 2-(nc), 3-(ca) and 2-(nc), 4-(ca) structures. According to eq. (7.3)

$$x_{75}^{(2)} = \left\lfloor n \left(\frac{\ln(\alpha_{k,\sigma})}{\ln(\alpha_{k,\sigma} + 1)} - 1 \right) \right\rfloor \Bigg|_{n=75, k=2, \sigma=3} = \left\lfloor 75 \left(\frac{\ln(0.6053)}{\ln(0.6504)} - 1 \right) \right\rfloor = 12$$
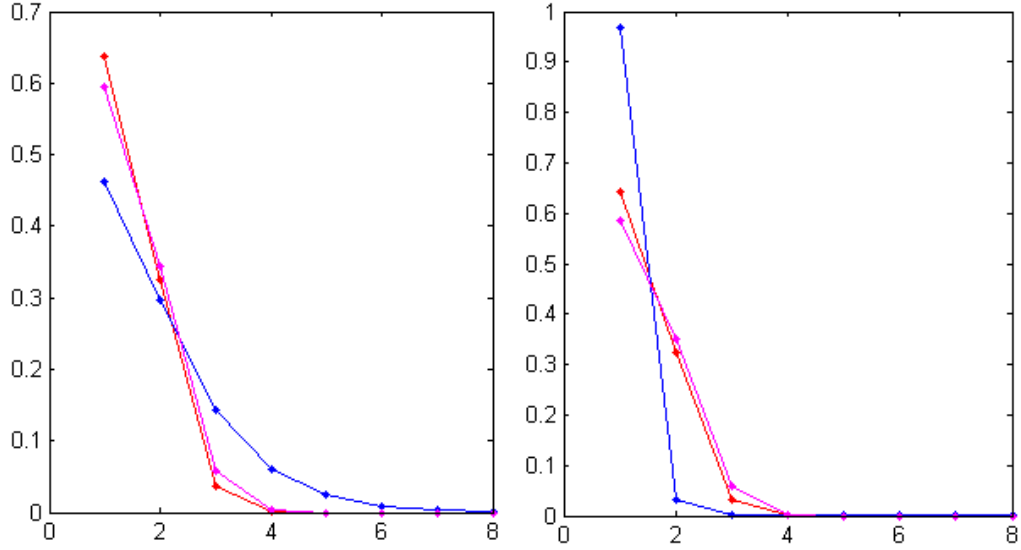
FIGURE 9. Finite size effect: here we insert the distribution of irreducibles of 2-(nc), 3-ca mfe structures for $n = 85$ (magenta) into Figure 8.

and for 3-(nc), 3-(ca) and 3-(nc), 4-(ca) structures we obtain

$$x_{75}^{(3)} = \left\lfloor n \left(\frac{\ln(\alpha_{k,\sigma})}{\ln(\alpha_{k,\sigma} + 1)} - 1 \right) \right\rfloor \bigg|_{n=75, k=3, \sigma=3} = \left\lfloor 75 \left(\frac{\ln(0.4914)}{\ln(0.5587)} - 1 \right) \right\rfloor = 16.$$

Figure 10 shows how well the "shifting" works for 2- and 3-(nc), 3-(ca) mfe RNA structures. Proposition 1 brings us now in the position to understand observation (a): the discrepancy of the irreducibility of 2-(nc), 3-(ca) mfe structures of length 75 and random structures. To this end we compare in Figure 11 the irreducibility of 2-(nc), 1-(ca) mfe structures [15] for $n = 75$ with that of random structures. In view of

$$\Delta_{75} = \left\lfloor 75 \left(\frac{\ln(\alpha_{2,1})}{\ln(\alpha_{2,3})} - 1 \right) \right\rfloor \approx \left\lfloor 75 \left(\frac{\ln(0.4369)}{\ln(0.6053)} - 1 \right) \right\rfloor = 48,$$

the former correspond to 2-(nc), 3-(ca) structures of length $75 + 48 = 123$. Accordingly, Proposition 1 and Figure 11 imply that 2-(nc), 3-(ca) mfe structures of length 123 exhibit an almost identical distribution as random structures of infinite sequence length. Finally, we remark upon the "paradox" that Proposition 1–being entirely based on the limit distribution of random structures– allows us to obtain information about mfe structures of length 75.
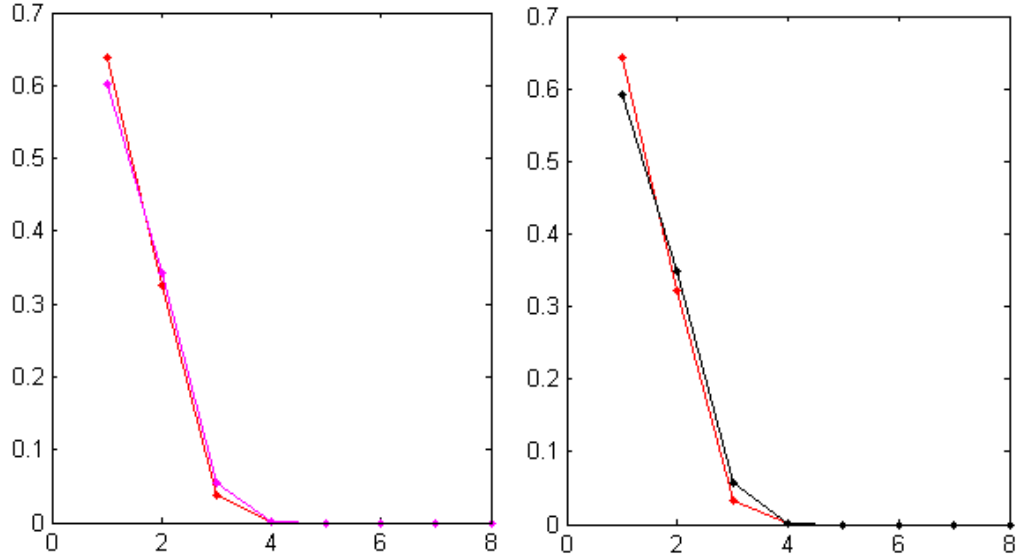
FIGURE 10. Proposition 1 at work: the lhs shows 2-(nc), 3-(ca) mfe structures for $n = 75$ (red) and 2-(nc), 4-(ca) mfe structures for $n = 85$ (magenta). The rhs displays 3-(nc), 3-(ca) mfe structures for $n = 75$ (red) and 3-(nc) 4-(ca) mfe structures for $n = 90$ (black).

k	2	3	4	5	6	7	8	9
$\sigma = 3$	1.6521	2.0348	2.2644	2.4432	2.5932	2.7243	2.8414	2.9480
$\sigma = 4$	1.5375	1.7898	1.9370	2.0488	2.1407	2.2198	2.2896	2.3523
$\sigma = 5$	1.4613	1.6465	1.7532	1.8330	1.8979	1.9532	2.0016	2.0449
$\sigma = 6$	1.4063	1.5515	1.6345	1.6960	1.7457	1.7877	1.8243	1.8569

TABLE 3. The exponential growth rates of k-(nc), σ-(ca), RNA structures where $\sigma \geq 3$.

As for observation (b), recall that 3-(nc) structures consist of two distinct combinatorial classes: 2-(nc) and 2-crossing RNA structures. The combinatorics of 2- and 3-(nc) RNA structures [10], implies that the number of 2-crossings is exponentially larger than the number of 2-(nc) RNA structures, see Table 3. To be concrete, the ratio of 2-(nc) over 2-crossing random 3-(nc) structures for $n = 75$ is $\approx 6 \times 10^{-5}$, while the ratio of 2-(nc) versus 2-(nc) RNA structures generated by cross is ≈ 1.7027. In other words, cross overrepresents 2-(nc) structures at a rate of approximately
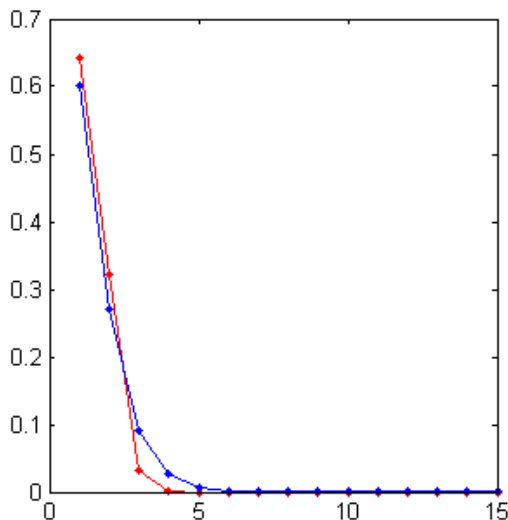
FIGURE 11. The distributions of irreducibles of 2-(nc), 1-(ca) mfe (red) and random structures (blue) for $n = 75$.

$300\,000 : 1$. In Figure 12 we illustrate the fact that 3-(nc), 3-(ca) mfe structures are more similar to 2-(nc) than to 3-(nc) random structures.

4. THE DISTRIBUTION OF RETURNS

In this section we study the distribution of returns, i.e. the endpoint locations of irreducible substructures in RNA random and RNA mfe structures. In other words, we compute the probability for a particular position to be the *endpoint* of an irreducible substructure. Let $\chi(s)$ denote the set of returns of a given structure s. Clearly, for each return at i there exists an irreducible substructure starting at $j+1$ and ending at i. Accordingly, a structure decomposes into 3 distinct segments: the first being an arbitrary substructure over the $[1, j]$, the second being an irreducible substructure over $[j+1, i]$ and the third being an arbitrary substructure over $[i+1, n]$, see Figure 13. We denote the n-th coefficient of the generating function of $\mathbf{R}_{k,\sigma}(z)$ by $R_{k,\sigma}(n)$, i.e., $R_{k,\sigma}(n)$ is the number of irreducible k-(nc), σ-(ca) RNA structures over length n. Let $T_{k,\sigma}(n)$ denote the coefficient $\mathbf{T}_{k,\sigma}(z)$, i.e., $T_{k,\sigma}(n)$ is the number of k-(nc), σ-(ca) RNA structures over length n. Finally, let $a_i(n)$ denote the number of RNA structures of length n containing i as a return.
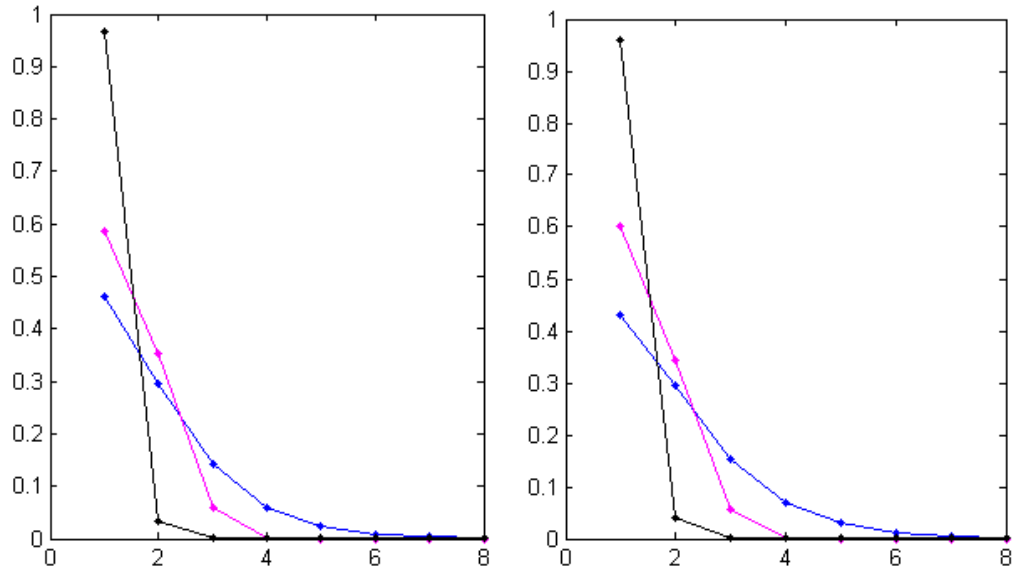
FIGURE 12. 2- and 3-(nc) random structures (blue/black) versus 3-(nc) mfe structures (magenta) for $n = 85$. The lhs and rhs display these curves for $\sigma = 3$ and $\sigma = 4$, respectively.
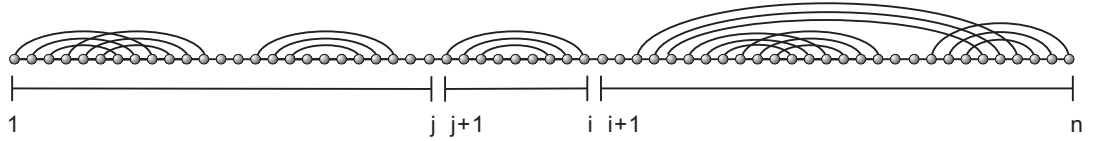

FIGURE 13. Three distinct segments: The first, $[1, j]$ and the last, $[i+1, n]$ contains an arbitrary RNA structure, respectively. While the second, $[j+1, i]$ contains an irreducible RNA structure.

Proposition 2. *Let $k \geq 2$ and $\sigma \geq 1$ be natural numbers, $\mu = (k-1)^2 + (k-1)/2$ and $\alpha_{k,\sigma}$ denote the unique dominant singularity of $\mathbf{T}_{k,\sigma}(z)$. Then*

$$\mathbb{P}[i \in \chi(s)] \quad \sim \quad \left(1 - \frac{i}{n}\right)^{-\mu} \left[i^{-\mu} - (i-1)^{-\mu}\alpha_{k,\sigma}\right]$$

and in particular for $i \to \infty$ and $(n - i) \to \infty$

$$\frac{\mathbb{P}[i + 1 \in \chi(s)]}{\mathbb{P}[i \in \chi(s)]} \sim \left(\frac{n - i}{n - i - 1}\right)^{\mu}.$$

Proposition 2 implies that returns are most likely to occur at the end of the sequence, see Figure 14. Furthermore the probability for the occurrence at the end of the sequence is exponential with exponent $\mu = (k-1)^2 + (k-1)/2$. Consequently, larger crossing numbers imply that "later" returns become more likely.
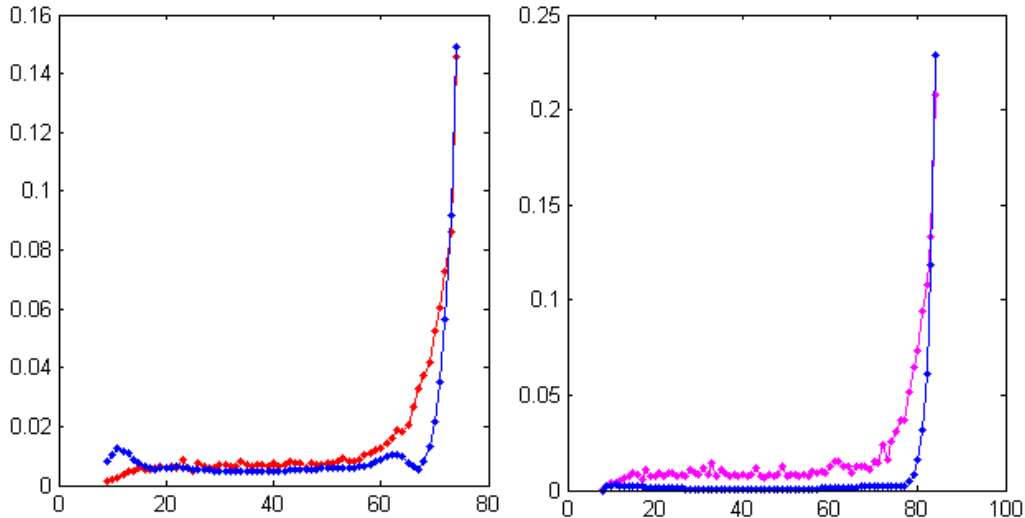

FIGURE 14. The distribution of returns for mfe and random structures of finite size: the lhs shows 2-(nc), 3-(ca), mfe and random structures of length 75 (red/blue). The rhs displays 3-(nc), 3-(ca), mfe and random structures of length 85 (magenta/blue).

5. IRREDUCIBLE SUBSTRUCTURES

In this section we compute the distribution of sizes of large irreducible substructures. In Section 4 we established that an irreducible substructure is typically "large", in the following we shall prove substantial improvements: the size of the largest, irreducible substructure is of size at least $n - O(1)$. Section 3 and Section 4 show, that RNA structures typically decompose into either one or two irreducible components. We therefore restrict our analysis to these two scenarios. We shall begin by studying the size, x_n, of an unique irreducible RNA substructure.

Lemma 2. *Suppose an RNA structure contains an unique, irreducible substructure, s, then*

$$(5.1) \qquad \mathbb{P}(|s| = x_n \mid s \text{ is unique}) \sim \frac{(n - x_n + 1)c \cdot x_n^{-\mu}\alpha_{k,\sigma}^{-x_n}}{c \cdot \left(\frac{1}{1-\alpha_{k,\sigma}}\right)^2 n^{-\mu}\alpha_{k,\sigma}^{-n}}.$$

In particular, any unique irreducible substructure has size of at least $n - O(1)$.

Furthermore for

$$(5.2) \qquad \frac{1}{\alpha_{k,\sigma}^{-\frac{1}{\mu}} - 1} \leq x_n \leq n - \frac{\alpha_{k,\sigma}}{1 - \alpha_{k,\sigma}}$$

the conditional probability of having an unique irreducible substructure s of size x_n is strictly monotone in x_n and for $x_n = n - \frac{\alpha_{k,\sigma}}{1-\alpha_{k,\sigma}}$ given by

$$(5.3) \qquad \lim_{n\to\infty} \mathbb{P}\left(|s| = n - \frac{\alpha_{k,\sigma}}{1 - \alpha_{k,\sigma}} \mid s \text{ is unique}\right) = (1 - \alpha_{k,\sigma})\,\alpha_{k,\sigma}^{\frac{\alpha_{k,\sigma}}{1-\alpha_{k,\sigma}}}.$$

Lemma 2 and the above observations show that a unique largest irreducible component is of size almost n. The few, remaining unpaired nucleotides are size $O(1)$, see Figure 7. The random structure distributions of length 85 derived in Lemma 2 are given in Figure 15 together with the distributions for 2- and 3-(nc), 3-(ca) RNA mfe structures of length $n = 85$. As for random structures, we observe that in this case the dominant singularities are $\alpha_{2,3} = 0.6053$ and $\alpha_{3,3} = 0.4914$ and the maximal probabilities as specified in eq. (5.3) are given by

$$(1 - \alpha_{2,3})\,\alpha_{2,3}^{\frac{\alpha_{2,3}}{1-\alpha_{2,3}}} \approx 0.18$$
$$(1 - \alpha_{3,3})\,\alpha_{3,3}^{\frac{\alpha_{3,3}}{1-\alpha_{3,3}}} \approx 0.25,$$

respectively. The distributions of 2- and 3-(nc), 3-(ca) mfe structures exhibit similar features as those of random structures. Since the data on mfe structures are obtained by sampling random sequences having a unique, irreducible substructure, they represent a refinement of the data given in Figure 14. Next we consider the case of two irreducible substructures. As in the case of a unique irreducible substructure, here, the larger of the two irreducibles contains almost all nucleotides. The proofs, however, become substantially more involved.

Lemma 3. *Suppose we are given an RNA structure S, that contains exactly the two irreducible substructures, s_1 and s_2, where $x_n = |s_1| \geq |s_2|$. Then*

$$(5.4) \qquad \mathbb{P}(|s_1| = x_n \mid S \text{ contains } s_1, s_2) \sim \begin{cases} o(1) & \text{for } (n - x_n) \to \infty \\ 2(1 - \alpha_{k,\sigma})^3 \cdot c_a \cdot (\alpha_{k,\sigma})^a & \text{for } (n - x_n) \to a < \infty, \end{cases}$$

where $\mu = (k - 1)^2 + (k - 1)/2$, $c_a > 0$ and $a \geq 1$. In particular, in the limit of long sequences, s_1 has a.s. a size of at least $(n - O(1))$.
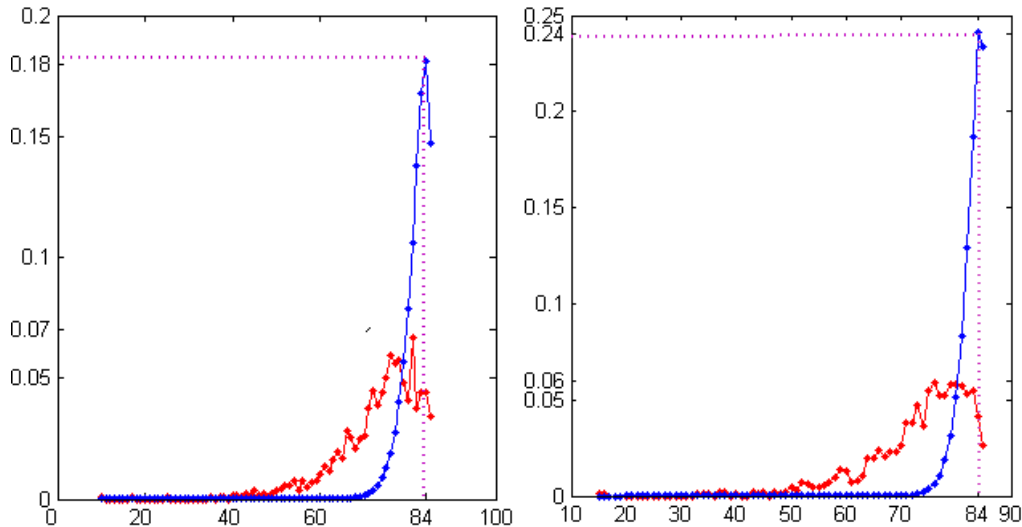
FIGURE 15. The distribution of the size of the unique irreducible component for mfe
and random structures of length 85: The lhs displays 2-(nc), 3-(ca) mfe (red) and random
structures (blue), the rhs shows 3-(nc), 3-(ca) mfe (red) and random structures (blue).

According to Lemma 3, we have for any $(n - x_n) \to a < \infty$

$$(5.5) \qquad \lim_{n \to \infty} \mathbb{P}(|s_1| = x_n \mid S \text{ contains } s_1, s_2) = 2(1 - \alpha_{k,\sigma})^3 \cdot c_a \cdot (\alpha_{k,\sigma})^a.$$

It follows from the proof of Lemma 3 in Section 7 that $c_a = \sum_{i=1}^{a} \binom{a-i+2}{2} R_{k,\sigma}(i)$. Let $\hat{a}$ be
a positive constant for which $c_{\hat{a}} \cdot (\alpha_{k,\sigma})^{\hat{a}}$ maximal. Consequently the probability $\mathbb{P}(|s_1| = x_n \mid$
$S \text{ contains } s_1, s_2)$ is maximal at $x_n = n - \hat{a}$, implying that the size of the largest irreducible
component is in the limit of long sequences typically $n - \hat{a}$, with probability $(1 - \alpha_{k,\sigma})^3 c_{\hat{a}} \cdot (\alpha_{k,\sigma})^{\hat{a}}$.
Note that for 3-(nc), 3-(ca) random structures of length n, we have $\alpha_{3,3} = 0.4914$. Maximizing
the term $c_{\hat{a}} \cdot (\alpha_{k,\sigma})^{\hat{a}}$ yields $\hat{a} = 14$ and accordingly the size of the largest component is likely to be
$n - 14$. Figure 16 confirms that already for $n = 85$, the size of the largest irreducible component is
typically 70. Figure 16 shows first that the probability $\mathbb{P}(|s_1| = x_n \mid S \text{ contains } s_1, s_2)$ is sharply
concentrated at $x_n = n - \hat{a}$, as implied by Lemma 3. Second, as n increases, the distribution of
component sizes shifts into a limit distribution which is sharply concentrated at $n - \hat{a}$. Furthermore
we remark that, by construction, $\hat{a}$ is independent of n, see Figure 17. For $n = 75$ and $n = 85$, the
size of the largest irreducible component is localized at $x_{75} = 60$ and $x_{85} = 70$.
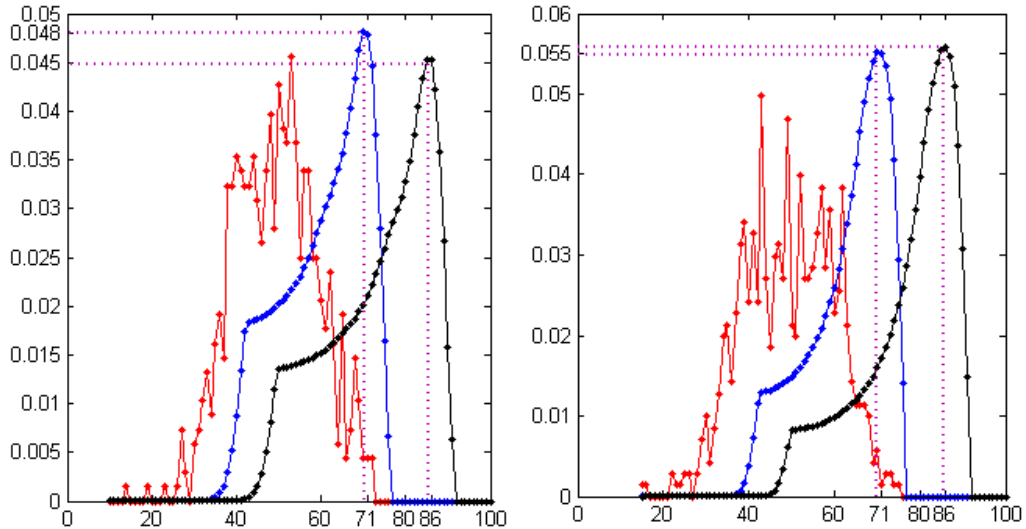
FIGURE 16. The distribution of the sizes of the giant for mfe and random structures having exactly 2 irreducible components. The lhs: 2-(nc), 3-(ca) mfe structures (red) and random structures for $n = 85$ and 100 (blue/black). The rhs: 3-(nc), 3-(ca) mfe structures (red) and random structures for $n = 85$ and 100 (blue/black).

6. DISCUSSION

Let us integrate our results and put them into context. Employing the Motzkin-path (Figure 1) interpretation, it is straightforward to construct random RNA secondary structures. However, random RNA pseudoknot structures are a different matter. Their inherent cross-serial dependencies (Figure 2) prohibit constructive recurrence relations and despite their D-finiteness [5], at present time, there exists no computer algorithm that can construct a random RNA pseudoknot structure in polynomial time with uniform probability. Consequently, any data on limit distributions of random RNA pseudoknot structures are nontrivial and virtually impossible to obtain computationally. In this paper we have shown that random RNA structures decompose into a small number of irreducible substructures, see Figures 2, 6 and 7. We established in Section 5 that one of these irreducibles is in fact a "giant", i.e. it contains almost all nucleotides. Key structural parameters, like the maximum number of mutually crossing bonds, as well as the minimum stack size do not fundamentally change this picture, see Figure 8 and Figure 11. In Section 3 we discussed the distribution of random structures and mfe structures. We actually used the limit of long
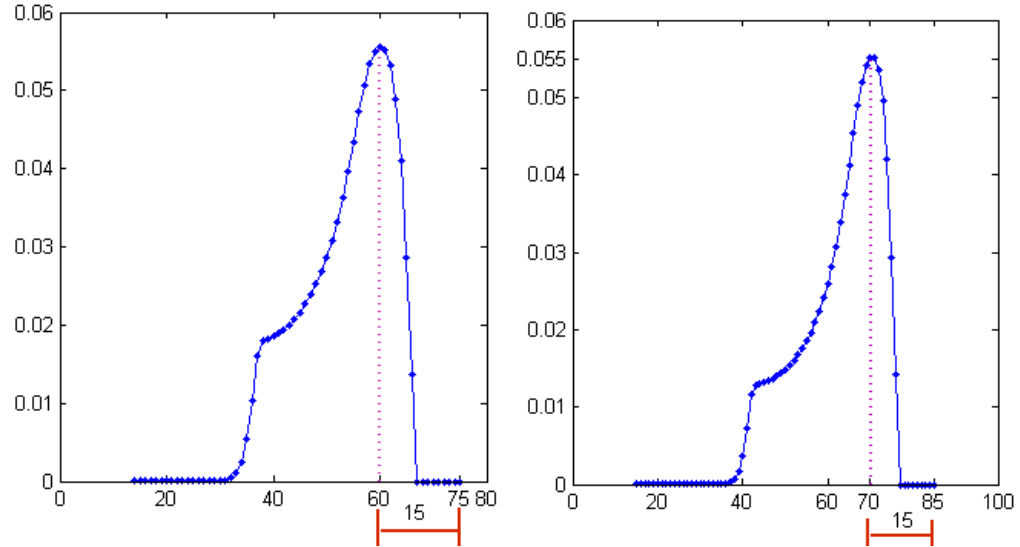
FIGURE 17. $(n - x_n)$ is independent n: we display the distribution of the sizes of the giants conditional on the existence of two irreducibles. The lhs shows 2-(nc), 3-(ca) random structures for $n = 75$ and the rhs for $n = 85$, respectively. In both distributions we highlight the distance (red) between the typical size of the giant and the end of the sequence.

sequences in order to prove a shift-result, allowing for the reduction of k-(nc) σ-(ca) structures over n to k-(nc), $(\sigma - j)$-(ca) structures over $n - f(j)$. Figure 10 illustrates that the original and the correspondingly shifted distributions of mfe structures virtually "coincide". Mfe and random structures exhibit significant differences. Most striking maybe is the vast preference of noncrossing RNA pseudoknot structures over their crossing counterparts. While the percentage of 63% of folded noncrossing configurations for $n = 75$ does not seem to be particularly remarkable, Theorem 1 of Section 2 shows, that the above percentage is equivalent to a factor of $300\,000\,:\,1$, relative to random sampling. This is certainly a consequence of the currently implemented pseudoknot-loop energy parameters. At this point it is pure speculation whether or not different energy parameters or significantly longer sequence length will alter this picture. In [4] the reader can find more data on the fraction of noncrossing configurations of 3-(nc) mfe structures. In any case, the overrepresentation of noncrossing configurations implies, that the distributions of 3-(nc) mfe structures are in fact more similar to those of random RNA secondary structures.

Having established that, even in the limit of long sequences, only a few irreducibles exist, the next question is to determine their respective sizes. In Section 4 and Section 5 we achieve this by studying returns, that is the endpoints of irreducible substructures and the distribution of their sizes in Lemma 2 and Lemma 3. For random structures we observe, that the largest irreducible component is a giant as it contains almost all nucleotides. For mfe structures we observe a systematic shift towards smaller sizes of the giant, however a giant irreducible substructure typically also exists in mfe structures. Aside from these structural results we present in Section 5, eq. (5.3), a simple formula for identifying the typical size of an unique giant and confirm in Figure 15 its applicability to mfe structures. Along these lines we furthermore localize the typical size of the giant in case of two irreducible substructures. In addition we make its dependence on k, σ and n explicit.

7. Proofs

Proof of Proposition 1.

Proof. D-finiteness [13] of $\mathbf{R}_{k,\sigma}(z)$ guarantees the existence of analytic continuation in some simply connected domain containing zero around the dominant singularity $\alpha_{k,\sigma}$. Therefore the singular expansion of $\mathbf{R}_{k,\sigma}(z)$ at its dominant singularity $\alpha_{k,\sigma}$ exists and in case of $k \equiv 1 \mod 2$ we have, setting $\mu = (k-1)^2 + \frac{k-1}{2}$,

$$(7.1) \qquad \mathbf{R}_{k,\sigma}(z) = \tau_k - c_k \left(1 - \frac{z}{\alpha_{k,\sigma}}\right)^{\mu-1} \ln\left(1 - \frac{z}{\alpha_{k,\sigma}}\right)(1 + o(1)), \qquad \text{where } c_k > 0.$$

In case of $k \equiv 0 \mod 2$, the singular expansion is given by

$$(7.2) \qquad \mathbf{R}_{k,\sigma}(z) = \tau_k - c_k \left(1 - \frac{z}{\alpha_{k,\sigma}}\right)^{\mu-1}(1 + o(1)), \qquad \text{where } c_k > 0.$$

In the following we restrict ourselves to the analysis of the case $k \equiv 1 \mod 2$. The arguments for $k \equiv 0 \mod 2$ are completely analogous. From the singular expansion of $\mathbf{R}_{k,\sigma}(z)$ we derive

$$R_{k,\sigma}(n) \sim n^{-\mu} \left(\frac{1}{\alpha_{k,\sigma}}\right)^n$$

and, apparently, $R_{k,\sigma+1}(n) < R_{k,\sigma}(n)$. For sufficiently large n, we have

$$
\frac{R_{k,\sigma+1}(n+x_n^{(k)})}{R_{k,\sigma}(n)} = \frac{(n+x_n^{(k)})^{-\mu}\left(\frac{1}{\alpha_{k,\sigma+1}}\right)^{n+x_n^{(k)}}}{n^{-\mu}\left(\frac{1}{\alpha_{k,\sigma}}\right)^{n}}
$$

$$
= \left(1+\frac{x_n^{(k)}}{n}\right)^{-\mu} \cdot \frac{(\alpha_{k,\sigma})^{n}}{(\alpha_{k,\sigma+1})^{n+x_n^{(k)}}}.
$$

Suppose $\lim_{n\to\infty}\left(\frac{R_{k,\sigma+1}(n+x_n^{(k)})}{R_{k,\sigma}(n)}\right) = c > 0$, then

$$
(7.3) \qquad x_n^{(k)} = \left\lfloor n\left(\frac{\ln(\alpha_{k,\sigma})}{\ln(\alpha_{k,\sigma+1})}-1\right)\right\rfloor.
$$

Accordingly

$$
(7.4) \qquad \frac{R_{k,\sigma+1}(n+x_n^{(k)})}{R_{k,\sigma}(n)} \sim \left(\frac{\ln(\alpha_{k,\sigma})}{\ln(\alpha_{k,\sigma+1})}\right)^{-\mu} < 1
$$

and eq. (3.1) follows. Using the singular expansion of $\mathbf{R}_{k,\sigma}(z)$ given in eq. (7.1) we derive

$$
\beta_{k,\sigma,j}(n) = [z^n]\mathbf{R}_{k,\sigma}(z)^j\left(\frac{1}{1-z}\right)^{j+1}
$$

$$
= [z^n]\left[\tau_k - c_k\left(1-\frac{z}{\alpha_{k,\sigma}}\right)^{\mu-1}\ln\left(1-\frac{z}{\alpha_{k,\sigma}}\right)(1+o(1))\right]^j\left(\frac{1}{1-z}\right)^{j+1}
$$

$$
\sim c_{k,\sigma}^{(1)}\cdot n^{-\mu}\alpha_{k,\sigma}^{-n} \qquad \text{for some constant } c_{k,\sigma}^{(1)} > 0
$$

and consequently

$$
\beta_{k,\sigma+1,j}(n+y_{n,k}) \sim c_{k,\sigma+1}^{(1)}\cdot(n+y_{n,k})^{-\mu}\alpha_{k,\sigma+1}^{-n-y_{n,k}} \qquad \text{for some constant } c_{k,\sigma+1}^{(1)} > 0.
$$

We observe that $\beta_{k,\sigma,j}(n) \approx \beta_{k,\sigma+1,j}(n+y_{n,k})$ holds only if $y_{n,k} = \lfloor c_k\cdot n\rfloor$ for some constant $c_k > 0$ with the property

$$
(7.5) \qquad \left(\frac{\alpha_{k,\sigma}}{(\alpha_{k,\sigma+1})^{1+c_k}}\right)^n = \frac{c_{k,\sigma}^{(1)}}{c_{k,\sigma+1}^{(1)}}(1+c_k)^\mu.
$$

The solution c_k of eq. (7.5) is asymptotically

$$
c_k \approx \frac{\ln(\alpha_{k,\sigma})}{\ln(\alpha_{k,\sigma+1})},
$$

whence $y_{n,k} = x_n^{(k)}$ and eq. (3.2) is established.   □

Proof of Proposition 2.

Proof. We begin by splitting a given structure s at j and i:

$$a_i = \sum_{j \leq i} T_{k,\sigma}(j) \cdot R_{k,\sigma}(i-j) \cdot T_{k,\sigma}(n-i) \quad = \quad T_{k,\sigma}(n-i) \cdot \sum_{j \leq i} T_{k,\sigma}(j) \cdot R_{k,\sigma}(i-j)$$

and $\sum_{j \leq i} T_{k,\sigma}(j) \cdot R_{k,\sigma}(i-j) = [z^i] \left(\mathbf{T}_{k,\sigma}(z) \mathbf{R}_{k,\sigma}(z) \right)$. Furthermore, we derive

$$\begin{aligned} \sum_{i=1}^{n} a_i &= \sum_{i=1}^{n} T_{k,\sigma}(n-i) \cdot [z^i] \left(\mathbf{T}_{k,\sigma}(z) \mathbf{R}_{k,\sigma}(z) \right) \\ &= \sum_{i=1}^{n} [z^{n-i}] \mathbf{T}_{k,\sigma}(z) \cdot [z^i] \left(\mathbf{T}_{k,\sigma}(z) \mathbf{R}_{k,\sigma}(z) \right) \\ &= [z^n] \mathbf{T}_{k,\sigma}^2(z) \mathbf{R}_{k,\sigma}(z). \end{aligned}$$

Consequently, the probability of a return at position i is given by

$$(7.6) \qquad \mathbb{P}[i \in \chi(s)] \quad = \quad \frac{T_{k,\sigma}(n-i) \cdot [z^i] \left(\mathbf{T}_{k,\sigma}(z) \mathbf{R}_{k,\sigma}(z) \right)}{[z^n] \mathbf{T}_{k,\sigma}^2(z) \mathbf{R}_{k,\sigma}(z)}.$$

Using $\mathbf{R}_{k,\sigma}(z) = 1 - z - \frac{1}{\mathbf{T}_{k,\sigma}(z)}$ we rewrite the probability $\mathbb{P}[i \in \chi(s)]$ as

$$(7.7) \qquad \mathbb{P}[i \in \chi(s)] \quad = \quad \frac{T_{k,\sigma}(n-i) \cdot [z^i] \left((1-z) \mathbf{T}_{k,\sigma}(z) - 1 \right)}{[z^n] (1-z) \mathbf{T}_{k,\sigma}^2(z) - \mathbf{T}_{k,\sigma}(z)}.$$

We next use the singular expansion of $\mathbf{T}_{k,\sigma}(z)$ at the dominant singularity $\alpha_{k,\sigma}$

$$(7.8) \qquad \mathbf{T}_{k,\sigma}(z) = \begin{cases} O((1 - \frac{z}{\alpha_{k,\sigma}})^{\mu-1} \ln(1 - \frac{z}{\alpha_{k,\sigma}})) & \text{for } k \text{ odd as } z \to \alpha_{k,\sigma} \\ O((1 - \frac{z}{\alpha_{k,\sigma}})^{\mu-1}) & \text{for } k \text{ even as } z \to \alpha_{k,\sigma}, \end{cases}$$

where $\mu = (k-1)^2 + (k-1)/2$. We restrict ourselves proving the case of $k \equiv 1 \mod 2$, the case $k \equiv 0 \mod 2$ follows analogously. Since D-finite power series form an algebra, $\mathbf{Q}_{k,\sigma}(z) = (1-z) \mathbf{T}_{k,\sigma}^2(z) - \mathbf{T}_{k,\sigma}(z)$ is D-finite and analytic continuation and singular expansion exist. The latter is given by

$$\begin{aligned} \mathbf{Q}_{k,\sigma}(z) &= (1-z)O\left(\left[(1 - \frac{z}{\alpha_k})^{\mu-1} \ln(1 - \frac{z}{\alpha_k}) \right]^2 \right) - O\left((1 - \frac{z}{\alpha_k})^{\mu-1} \ln(1 - \frac{z}{\alpha_k}) \right), \quad z \to \alpha_{k,\sigma} \\ &= O\left((1 - \frac{z}{\alpha_k})^{\mu-1} \ln(1 - \frac{z}{\alpha_k}) \right), \quad z \to \alpha_{k,\sigma}. \end{aligned}$$

Therefore, extracting the coefficients of the singular expansion, we obtain

$$(7.9) \qquad\qquad [z^n]\mathbf{Q}_{k,\sigma}(z) \sim n^{-\mu}\alpha_{k,\sigma}^{-n}.$$

Using $[z^i]\mathbf{T}_{k,\sigma}(z) = T(i) \sim i^{-\mu}\alpha_{k,\sigma}^{-i}$ as $i \to \infty$ we arrive at

$$
\begin{aligned}
\mathbb{P}[i \in \chi(s)] \quad &\sim \quad \frac{T(n-i)\left[[z^i]\mathbf{T}_{k,\sigma}(z) - [z^{i-1}]\mathbf{T}_{k,\sigma}(z)\right]}{[z^n]\mathbf{Q}_{k,\sigma}(z)} \\
&\sim \quad \frac{(n-i)^{-\mu}\alpha_{k,\sigma}^{-n+i}[i^{-\mu}\alpha_{k,\sigma}^{-i} - (i-1)^{-\mu}\alpha_{k,\sigma}^{-i+1}]}{n^{-\mu}\alpha_{k,\sigma}^{-n}} \\
&= \quad \left(1 - \frac{i}{n}\right)^{-\mu}\left[i^{-\mu} - (i-1)^{-\mu}\alpha_{k,\sigma}\right].
\end{aligned}
$$

From this we immediately conclude in case of $i \to \infty$ and $(n-i) \to \infty$

$$
\begin{aligned}
\frac{\mathbb{P}[i+1 \in \chi(s)]}{\mathbb{P}[i \in \chi(s)]} \quad &\sim \quad \frac{\left(1 - \frac{i+1}{n}\right)^{-\mu}[(i+1)^{-\mu} - i^{-\mu}\alpha_{k,\sigma}]}{\left(1 - \frac{i}{n}\right)^{-\mu}[i^{-\mu} - (i-1)^{-\mu}\alpha_{k,\sigma}]} \\
&\sim \quad \left(\frac{n-i}{n-i-1}\right)^{\mu} \qquad i \to \infty, (n-i) \to \infty.
\end{aligned}
$$

$\square$

Proof of Lemma 2.

Proof. Using the singular expansion of $\mathbf{R}_{k,\sigma}(z)$, eq. (7.1) and eq. (7.2) we obtain

$$(7.10) \qquad \delta_{n,1} = [z^n]\left(\frac{1}{1-z}\right)^2 \mathbf{R}_{k,\sigma}(z) \sim c \cdot \left(\frac{1}{1-\alpha_{k,\sigma}}\right)^2 n^{-\mu}\alpha_{k,\sigma}^{-n}, \qquad \text{for some } c > 0.$$

We proceed by computing

$$(7.11) \qquad \delta_{n,1,x_n} = [z^{x_n}]\mathbf{R}_{k,\sigma}(z) \cdot [z^{n-x_n}]\left(\frac{1}{1-z}\right)^2 \sim (n - x_n + 1)c \cdot x_n^{-\mu}\alpha_{k,\sigma}^{-x_n}$$

and combining eq. (7.10) and eq. (7.11), we arrive at

$$(7.12) \qquad \frac{\delta_{n,1,x_n}}{\delta_{n,1}} \sim \frac{(n - x_n + 1)c \cdot x_n^{-\mu}\alpha_{k,\sigma}^{-x_n}}{c \cdot \left(\frac{1}{1-\alpha_{k,\sigma}}\right)^2 n^{-\mu}\alpha_{k,\sigma}^{-n}}.$$

The critical term here in eq. (7.12) is readily identified to be $\alpha_{k,\sigma}^{n-x_n}$ and consequently

$$(7.13) \qquad\qquad \lim_{n \to \infty} \frac{\delta_{n,1,x_n}}{\delta_{n,1}} > 0 \quad\Longrightarrow\quad x_n = n - O(1),$$

whence the lemma. $\qquad\qquad\square$

Proof of Lemma 3.

Proof. We distinguish the cases $x_n < \frac{n}{2}$ and $x_n \geq \frac{n}{2}$. In case of $x_n < \frac{n}{2}$ we have

(7.14) $$\delta_{n,2} = [z^n] \left(\frac{1}{1-z}\right)^3 \mathbf{R}_{k,\sigma}^2(z) \sim c \cdot \left(\frac{1}{1-\alpha_{k,\sigma}}\right)^3 n^{-\mu} \alpha_{k,\sigma}^{-n}, \quad \text{for some } c > 0.$$

Therefore the number of structures containing s_1 of size x_n is given by

$$
\begin{aligned}
\delta_{n,2,x_n} &= \sum_{\max(s,t)=x_n} [z^s]\mathbf{R}_{k,\sigma}(z) \cdot [z^t]\mathbf{R}_{k,\sigma}(z) \cdot [z^{n-s-t}] \left(\frac{1}{1-z}\right)^3 \\
&= 2[z^{x_n}]\mathbf{R}_{k,\sigma}(z) \cdot \sum_{t=1}^{x_n} [z^t]\mathbf{R}_{k,\sigma}(z) \cdot [z^{n-x_n-t}] \left(\frac{1}{1-z}\right)^3
\end{aligned}
$$

The term $[z^{n-x_n-t}] \left(\frac{1}{1-z}\right)^3$ represents the number of compositions of the integer $u = n - x_n - t$ into at most 3 distinct parts, denoted by $P(u,3)$. Assuming the first part to be i, ranging from 1 to u, the number of ways of dividing $(u - i)$ into at most 2 parts is $(u - i + 1)$, whence

$$[z^u] \left(\frac{1}{1-z}\right)^3 = \sum_{i=0}^{u} P(u-i,2) = \sum_{i=0}^{u}(u-i+1) = \binom{u+2}{2}.$$

Consequently, we can rewrite $\delta_{n,2,x_n}$ as

(7.15) $$\delta_{n,2,x_n} = 2[z^{x_n}]\mathbf{R}_{k,\sigma}(z) \cdot \sum_{t=1}^{x_n} \binom{n-x_n+2}{2} \cdot [z^t]\mathbf{R}_{k,\sigma}(z).$$

Claim. Suppose x_n and $n - x_n$ tend to infinity, as n tends to infinity. Then there exists some constant $\kappa > 0$ such that

(7.16) $$\sum_{i=1}^{x_n} \binom{n-x_n-i+2}{2} [z^i]\mathbf{R}_{k,\sigma}(z) = \kappa \cdot \binom{n-2x_n+2}{2} [z^{x_n}]\mathbf{R}_{k,\sigma}(z).$$

According to the Claim

$$
\begin{aligned}
\delta_{n,2,x_n} &= 2[z^{x_n}]\mathbf{R}_{k,\sigma}(z) \cdot \sum_{t=1}^{x_n} \binom{n - x_n - t + 2}{2} \cdot [z^t]\mathbf{R}_{k,\sigma}(z) \\
&\sim 2[z^{x_n}]\mathbf{R}_{k,\sigma}(z) \cdot \kappa \binom{n - 2x_n + 2}{2}[z^{x_n}]\mathbf{R}_{k,\sigma}(z) \\
&\sim 2\kappa c^2 \cdot x_n^{-2\mu}\alpha_{k,\sigma}^{-2x_n}\binom{n - 2x_n + 2}{2},
\end{aligned}
$$

whence the probability of containing the largest component of size $x_n < \frac{n}{2}$ is given by

(7.17) $$ \frac{\delta_{n,2,x_n}}{\delta_{n,2}} \sim \frac{2\kappa c^2 \cdot x_n^{-2\mu}\alpha_{k,\sigma}^{-2x_n}\binom{n-2x_n+2}{2}}{c \cdot \left(\frac{1}{1-\alpha_{k,\sigma}}\right)^3 n^{-\mu}\alpha_{k,\sigma}^{-n}} = o(1). $$

In case of $x_n \geq \frac{n}{2}$ we derive,

$$
\begin{aligned}
\delta_{n,2,x_n} &= \sum_{\max(s,t)=x_n} [z^s]\mathbf{R}_{k,\sigma}(z) \cdot [z^t]\mathbf{R}_{k,\sigma}(z) \cdot [z^{n-s-t}]\left(\frac{1}{1-z}\right)^3 \\
&= 2[z^{x_n}]\mathbf{R}_{k,\sigma}(z) \cdot \sum_{t=1}^{n-x_n} [z^t]\mathbf{R}_{k,\sigma}(z) \cdot [z^{n-x_n-t}]\left(\frac{1}{1-z}\right)^3 \\
&= 2[z^{x_n}]\mathbf{R}_{k,\sigma}(z) \cdot [z^{n-x_n}]\mathbf{R}_{k,\sigma}(z)\left(\frac{1}{1-z}\right)^3.
\end{aligned}
$$

Suppose $(n - x_n) \to \infty$, then the singular expansion implies

$$ [z^{n-x_n}]\mathbf{R}_{k,\sigma}(z)\left(\frac{1}{1-z}\right)^3 = \kappa \cdot (n - x_n)^{-\mu}(\alpha_{k,\sigma})^{-n+x_n}. $$

Accordingly, we derive the following asymptotic expression for $\delta_{n,2,x_n}$

$$
\begin{aligned}
\delta_{n,2,x_n} &= 2[z^{x_n}]\mathbf{R}_{k,\sigma}(z) \cdot [z^{n-x_n}]\mathbf{R}_{k,\sigma}(z)\left(\frac{1}{1-z}\right)^3 \\
&\sim 2c\kappa \cdot x_n^{-\mu}\alpha_{k,\sigma}^{-x_n}(n - x_n)^{-\mu}\alpha_{k,\sigma}^{-(n-x_n)},
\end{aligned}
$$

Therefore we arrive at

(7.18) $$ \frac{\delta_{n,2,x_n}}{\delta_{n,2}} \sim \frac{2c\kappa \cdot x_n^{-\mu}\alpha_{k,\sigma}^{-x_n}(n - x_n)^{-\mu}\alpha_{k,\sigma}^{-(n-x_n)}}{c \cdot \left(\frac{1}{1-\alpha_{k,\sigma}}\right)^3 n^{-\mu}\alpha_{k,\sigma}^{-n}} = 2\kappa(1 - \alpha_{k,\sigma})^3 \left[\frac{x_n(n - x_n)}{n}\right]^{-\mu}. $$

Note that $(n - x_n) \to \infty$ implies, that $\lim_{n\to\infty} \frac{x_n}{n} = \nu$ for $\frac{1}{2} \leq \nu < 1$. Consequently

(7.19) $$ \frac{\delta_{n,2,x_n}}{\delta_{n,2}} \sim 2\kappa(1 - \alpha_{k,\sigma})^3 \left[\frac{x_n(n - x_n)}{n}\right]^{-\mu} = 2\kappa(1 - \alpha_{k,\sigma})^3 [\nu \cdot n(1 - \nu)]^{-\mu} = o(1), $$

from which we immediately conclude

$$(7.20) \qquad \lim_{n\to\infty} \frac{\delta_{n,2,x_n}}{\delta_{n,2}} = 0 \quad \text{for } x_n \geq \frac{n}{2} \text{ and } (n - x_n) \to \infty.$$

In case of $(n - x_n) \to a < \infty$, we set

$$(7.21) \qquad c_a = \sum_{i=1}^{a} \binom{a - i + 2}{2} R_{k,\sigma}(i) = [z^a]\mathbf{R}_{k,\sigma}(z) \left(\frac{1}{1-z}\right)^3.$$

Accordingly

$$\delta_{n,2,x_n} = 2[z^{x_n}]\mathbf{R}_{k,\sigma}(z) \cdot [z^{n-x_n}]\mathbf{R}_{k,\sigma}(z) \left(\frac{1}{1-z}\right)^3$$

$$\sim 2c_a \cdot c \cdot x_n^{-\mu} (\alpha_{k,\sigma})^{-x_n}.$$

We accordingly arrive at

$$(7.22) \qquad \frac{\delta_{n,2,x_n}}{\delta_{n,2}} \sim \frac{2c_a \cdot c \cdot x_n^{-\mu} (\alpha_{k,\sigma})^{-x_n}}{c \cdot \left(\frac{1}{1-\alpha_{k,\sigma}}\right)^3 n^{-\mu}\alpha_{k,\sigma}^{-n}} = 2c_a \cdot (1 - \alpha_{k,\sigma})^3 \cdot (\alpha_{k,\sigma})^a,$$

We observe that $x_n < \frac{n}{2}$ implies $(n - x_n) \to \infty$, therefore we conclude in the case $(n - x_n) \to \infty$,

$$\lim_{n\to\infty} \frac{\delta_{n,2,x_n}}{\delta_{n,2}} = 0.$$

While $(n - x_n) \to a < \infty$ implies $x_n \geq \frac{n}{2}$, we conclude that in the case $(n - x_n) \to a < \infty$,

$$\lim_{n\to\infty} \frac{\delta_{n,2,x_n}}{\delta_{n,2}} = 2c_a \cdot (1 - \alpha_{k,\sigma})^3 \cdot (\alpha_{k,\sigma})^a.$$

completing the proof of the lemma. $\qquad \square$

Proof of the Claim.

Proof. Set

$$(7.23) \qquad A_i = \binom{n - x_n - i + 2}{2}[z^i]\mathbf{R}_{k,\sigma}(z) = \binom{n - x_n - i + 2}{2} R_{k,\sigma}(i)$$

We first show that A_{x_n} is the maximal term A_i for $1 \le i \le x_n$. In view of the fact that $x_n < \frac{n}{2}$,

$$
\begin{aligned}
\lim_{n\to\infty} \frac{A_{i+1}}{A_i} &= \lim_{n\to\infty} \frac{n - x_n - i}{n - x_n - i + 2} \cdot \frac{R_{k,\sigma}(i+1)}{R_{k,\sigma}(i)} \\
&= \lim_{n\to\infty} \frac{1}{1 + \frac{2}{n - x_n - i}} \frac{R_{k,\sigma}(i+1)}{R_{k,\sigma}(i)} \\
&> \lim_{n\to\infty} \frac{1}{1 + \frac{2}{\frac{n}{2} - i}} \frac{R_{k,\sigma}(i+1)}{R_{k,\sigma}(i)} \\
&= \frac{R_{k,\sigma}(i+1)}{R_{k,\sigma}(i)} > 1.
\end{aligned}
$$

Therefore A_{x_n} is maximal. We shall show next

(7.24) $$\forall \, 0 < \alpha < 1; \qquad \sum_{i < x_n - n^\alpha} A_i = o(1) \cdot A_{x_n}.$$

In view of $\sum_{i < x_n - n^\alpha} A_i < (x_n - n^\alpha) A_{x_n - n^\alpha}$, we obtain

$$
\frac{\sum_{i < x_n - n^\alpha} A_i}{A_{x_n}} < \frac{(x_n - n^\alpha) A_{x_n - n^\alpha}}{A_{x_n}} = \frac{(x_n - n^\alpha)\binom{n - 2x_n + n^\alpha + 2}{2} R_{k,\sigma}(x_n - n^\alpha)}{\binom{n - 2x_n + 2}{2} R_{k,\sigma}(x_n)}
$$

Using $R_{k,\sigma}(n) \sim c_1 n^{-\mu} \alpha_{k,\sigma}^{-n}$, for some $c_1 > 0$, we arrive at

$$
\begin{aligned}
\frac{\sum_{i < x_n - n^\alpha} A_i}{A_{x_n}} &< \frac{(x_n - n^\alpha)\binom{n - 2x_n + n^\alpha + 2}{2} R_{k,\sigma}(x_n - n^\alpha)}{\binom{n - 2x_n + 2}{2} R_{k,\sigma}(x_n)} \\
&\sim \left(1 - \frac{n^\alpha}{x_n}\right)^{-\mu} (x_n - n^\alpha)\left[1 + \frac{n^\alpha}{n - 2x_n + 2}\right]\left[1 + \frac{n^\alpha}{n - 2x_n + 1}\right] \alpha_{k,\sigma}^{n^\alpha} = o(1),
\end{aligned}
$$

whence eq. (7.24). Next we claim that the terms close to x_n contribute at most $O(A_{x_n})$, i.e.

(7.25) $$\sum_{j=0}^{n^\alpha} A_{x_n - j} = O(A_{x_n}).$$

To prove this, we compute for $1 \le j \le n^\alpha$

$$
\begin{aligned}
\frac{A_{x_n - j}}{A_{x_n}} &\sim \left[1 + \frac{j}{n - 2x_n + 2}\right] \cdot \left[1 + \frac{j}{n - 2x_n + 1}\right] \cdot \left[1 - \frac{j}{x_n}\right]^{-\mu} \alpha_{k,\sigma}^j \\
&\le \left[1 + \frac{j}{n - 2x_n + 2}\right] \cdot \left[1 + \frac{j}{n - 2x_n + 1}\right] \alpha_{k,\sigma}^j \\
&< \left(1 + \frac{j}{2}\right)(1 + j) \alpha_{k,\sigma}^j.
\end{aligned}
$$

Taking the sum over all j we obtain

$$\sum_{j=0}^{n^\alpha} \frac{A_{x_n-j}}{A_{x_n}} < \sum_{j=0}^{n^\alpha} \left(1 + \frac{j}{2}\right)(1+j)\alpha_{k,\sigma}^j$$

$$= \frac{\alpha_{k,\sigma}^{n^\alpha+1}\left[(n^\alpha+1)^2(\alpha_{k,\sigma}-1)^2(n^\alpha+1)(\alpha_{k,\sigma}-2)^2 - n^\alpha + 1\right] - 2}{2(\alpha_{k,\sigma}-1)^3},$$

whence

$$\lim_{n\to\infty}\sum_{j=0}^{n^\alpha}\frac{A_{x_n-j}}{A_{x_n}} < \frac{1}{(1-\alpha_{k,\sigma})^3}.$$

Therefore, we obtain $\sum_{j=0}^{n^\alpha} A_{x_n-j} < \frac{1}{(1-\alpha_{k,\sigma})^3}A_{x_n}$ and we arrive at

$$(7.26) \qquad \sum_{i=0}^{x_n} A_i = \sum_{i<x_n-n^\alpha} A_i + \sum_{i\geq x_n-n^\alpha} A_i = o(A_{x_n}) + O(A_{x_n}) = \kappa \cdot A_{x_n},$$

for some constant $\kappa > 0$, proving the Claim. $\qquad\qquad\qquad\qquad\qquad\qquad\square$

REFERENCES

[1] N.T. Cameron and L. Shapiro, *Random walks, trees and extensions of Riordan group techniques* Annual Joint Mathematics Meetings, Baltimore, MD, U.S. (2003).

[2] J.A. Howell, T.F. Smith and M.S. Waterman, *Computation of generating functions for biological molecules* SIAM J. Appl. Math. **39**(1980), 119–133.

[3] Fenix W.D. Huang, Wade W.J. Peng and C. M. Reidys, *Folding 3-Noncrossing RNA Pseudoknot Structures* (2008), submitted.

[4] Fenix W.D. Huang, Linda Y.M. Li and C. M. Reidys, *Sequence-structure relations of pseudoknot RNA* BMC, (2008), to appear.

[5] E.Y. Jin, J. Qin and C.M. Reidys, *Combinatorics of RNA structures with pseudoknots* Bull. Math. Biol. **70**(1)(2008), 45–67.

[6] E.Y. Jin and C.M. Reidys, *Combinatorial Design of Pseudoknot RNA* Adv. Appl. Math. (2008), doi: 10.1016/j. aam. 2008. 06. 003.

[7] E.Y. Jin and C.M. Reidys, *On the decomposition of k-noncrossing RNA structures*, submit, 2009.

[8] D.A.M. Konings and R.R. Gutell, *A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs* RNA. **1**(1995), 559–574.

[9] A. Loria and T. Pan, *Domain structure of the ribozyme from eubacterial ribonuclease P* RNA. **2**(1996), 551–563.

[10] G. Ma and C.M. Reidys, *Canonical RNA Pseudoknot Structures* J. Comp. Biol. **15**(10)(2008), 1257–1273.

[11] R.C. Penner and M.S. Waterman, *Spaces of RNA secondary structures* Adv. Math. **101**(1993), 31–49.

[12] D.B. Searls, *Formal language theory and biological macromolecules* Series in Discr. Math. and Theor. Comp. Sci. **47**(1999), 117–140.

[13] R.P. Stanley, *Differentiably finite power series* Europ. J. Combinatorics. **1**(1980), 175–188.

[14] C. Tuerk, S. MacDougal and L. Gold, *RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase.* Proc. Natl. Acad. Sci. USA **89**(1992), 6988–6992.

[15] Vienna RNA Package–RNA Secondary Structure Prediction and Comparison. www.tbi.univie.ac.at/RNA/.

[16] M.S. Waterman, *Combinatorics of RNA hairpins and cloverleafs* Stud. Appl. Math. **60**(1979), 91–96.

[17] M.S. Waterman, *Secondary structure of single - stranded nucleic acids* Adv. Math.I (suppl.) **1**(1978), 167–212.

[18] M.S. Waterman and W.R. Schmitt, *Linear trees and RNA secondary structure* Discr. Appl. Math. **51** (1994), 317–323.

[19] E. Westhof and L. Jaeger, *RNA pseudoknots* Current Opinion Struct. Biol. **2** (1992), 327–333.

⋆ CENTER FOR COMBINATORICS, LPMC-TJKLC, † COLLEGE OF LIFE SCIENCES, NANKAI UNIVERSITY, TIANJIN 300071, P.R. CHINA, PHONE: *86-22-2350-6800, FAX: *86-22-2350-9272

E-mail address: reidys@nankai.edu.cn