Available online at www.sciencedirect.com

ScienceDirect

Discrete Applied Mathematics 155 (2007) 759–787

DISCRETE APPLIED MATHEMATICS

www.elsevier.com/locate/dam

Asymptotic expected number of base pairs in optimal secondary structure for random RNA using the Nussinov–Jacobson energy model

Peter Clote[a, b, *, 1], Evangelos Kranakis[c, 2], Danny Krizanc[d], Ladislav Stacho[e, 3]

[a] *Department of Biology (courtesy), Higgins Hall 355, Boston College, Chestnut Hill, MA 02467, USA*
[b] *Department of Computer Science, Higgins Hall 355, Boston College, Chestnut Hill, MA 02467, USA*
[c] *School of Computer Science, Carleton University, Ottawa, Ont., Canada K1S 5B6*
[d] *Department of Mathematics and Computer Science, Wesleyan University, Middletown, CT 06459, USA*
[e] *Department of Mathematics, Simon Fraser University, Burnaby, BC, Canada V5A 1S6*

Received 17 July 2004; received in revised form 22 April 2005; accepted 22 April 2005
Available online 12 October 2006

Abstract

Motivated by computer experiments, we study asymptotics of the expected maximum number of base pairs in secondary structures for random RNA sequences of length n. After proving a general limit result, we provide estimates of the limit for the binary alphabet $\{G, C\}$ with thresholds $k \geqslant 0$. We prove a general theorem entailing the existence of an asymptotic limit for the mean and standard deviation of free energy per nucleotide, as computed by `mfold`, for random RNA of any fixed compositional frequency; higher order moment limits are additionally shown to exist.

Keywords: Random RNA; Secondary structure; Nussinov–Jacobson algorithm; Zuker algorithm; Asymptotic Z-score

1. Introduction

It is well known that there is a compositional bias in nucleotide usage of various classes of RNA, depending on function (see, for instance [18]). For example, the mononucleotide (or compositional) frequency of 530 tRNAs from Sprinzl's tRNA database [26,27] is given by $q_A = 0.239922$, $q_C = 0.253383$, $q_G = 0.275618$, $q_U = 0.231076$, while that from a collection of 155 16S ribosomal RNAs [1] is $q_A = 0.2642$, $q_C = 0.2101$, $q_G = 0.3178$, $q_U = 0.2079$. This suggests the following motivating question: *To what extent might compositional frequency of a class of RNAs determine or constrain the stability of optimal secondary structures* (*hence the function*) *for members of that class*?

* Corresponding author.
E-mail addresses: clote@bc.edu (P. Clote), kranakis@scs.carleton.ca (E. Kranakis), dkrizanc@mail.wesleyan.edu (D. Krizanc), lstacho@sfu.ca (L. Stacho).

[1] Research partially supported by NSF DBI-0543506.
[2] Research supported in part by the NSERC (Natural Sciences and Engineering Research Council of Canada) and MITACS (Mathematics of Information Technology and Complex Systems) grants.
[3] Research supported in part by the NSERC (Natural Sciences and Engineering Research Council of Canada).

0166-218X/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2005.04.022

Table 1
Table of number *BP* of base pairs, ratio of base pairs to sequence length, etc. For random binary sequences of length n generated by Algorithm 1 to have expected mononucleotide frequencies: $q_C = q_G = 0.5$

n	*BP*	StDev	*BP*/n	Error	Max	Min
10	3.1800	0.7795	0.3180	0.0779	4	1
100	44.1200	1.9610	0.4412	0.0196	46	35
200	90.2900	2.0312	0.4515	0.0102	92	82
300	136.8600	1.9850	0.4562	0.0066	140	127
400	183.0200	1.9848	0.4576	0.0050	186	175
500	229.6200	1.9939	0.4592	0.0040	233	220
600	276.2100	2.3845	0.4603	0.0040	280	266
700	322.8600	2.1449	0.4612	0.0031	326	311
800	369.3800	2.0188	0.4617	0.0025	372	358
900	416.0800	1.9114	0.4623	0.0021	420	411
1000	462.5500	2.1325	0.4626	0.0021	466	457

Our implementation of the Nussinov–Jacobson algorithm was used with threshold 1, and sequence length up to 1000. Average values were taken over 100 iterations, where error means Stdev/n; points are indicated along with error bars. The asymptotic limit appears to be at least 0.4626.

In this paper, which vastly extends the preliminary report of [9], we study asymptotic properties of random RNA generated by a 0th order Markov chain from fixed *mononucleotide* or *compositional* frequencies of nucleotides A, C, G, U; in the appendix, we consider random RNA generated by kth order Markov chains. Our investigation is different from the work of either Hofacker et al. [17] or of Nebel [22,23]. These authors consider a *stickiness parameter*,[4] which gives the probability that any two positions can base pair. In [17], Hofacker et al. extend the technique of Stein and Waterman [29] to compute asymptotic limits of the expected number of base pairs divided by sequence length, the number of secondary structures of a given order, etc. They do this by deriving appropriate recurrence relations and proceed by application of Bender's Theorem (see [29]), a very powerful tool for solving asymptotic limits when generating functions satisfy a particular functional relation. In [22], Nebel computes precise rth order moments of asymptotic numbers of secondary structures by using sophisticated extensions of the generating function technique of [30]. For example, Theorem 10 of [22] states that "the average number of unpaired bases in a secondary structure of size n is asymptotically $\frac{n}{\sqrt{5}} + \frac{3}{10} + \frac{1}{\sqrt{5}} + \mathrm{O}(n^{-1})$". This, however, concerns the expected number of unpaired bases among *all* secondary structures, even those which are not optimal, where additionally any bases may pair (i.e. not just Watson–Crick or GU wobble pairs). While the results of Hofacker et al. and of Nebel are both interesting and deep, they do not concern the questions addressed in this paper. In particular, the asymptotic limits we establish concern the expected *maximum* number of base pairs (and higher order moments) of random RNA of a given compositional frequency (or of a given dinucleotide or more generally k-tuple frequency). This is *not* the same mathematical model as the Bernoulli model with a given stickiness parameter. In particular, the asymptotic value $P_n/nS_n = 0.2051$ of expected number of base pairs from the model of Hofacker et al. (cf. [17, Table 3]) is quite different from the asymptotic value of approximately 0.46 suggested by our computer experiments summarized in the Table 1 and graph from Fig. 1. For this latter comparison, for both the stickiness model of Hofacker et al. and our model, compositional frequency is $p_G = 0.5 = p_C$ and $p_A = 0 = p_U$, and the threshold (i.e. minimum number of unpaired bases in hairpin loops) is 1. While the models of Hofacker et al. and of Nebel concern the collection of all secondary structures compatible with a given RNA sequence, we consider only *optimal* secondary structures having a maximum number of base pairs for a given RNA sequence.

In this paper, we consider different possible values $t \geqslant 0$ for a minimum *threshold* on the number of unpaired bases between any two paired bases (i.e. hairpin loops are required to have at least t unpaired bases in the loop region). We prove a general limit theorem, which states that there is an asymptotic limit for the ratio of the expected maximum number of base pairs in random RNA divided by sequence length; moreover, this limit depends only on the compositional frequency used to generate the random RNA. In this regard, our simulations suggest that this limit is a minimum when

[4] Stickiness parameter $p = 2(p_A p_U + p_C p_G + p_G p_U)$ if Watson–Crick and GU wobble pairs are allowed, while $p = 2(p_A p_U + p_C p_G)$ if Watson–Crick but no GU wobble pairs are allowed. Here, p_A, p_C, p_G, p_U denote the compositional frequency of a class of RNA.

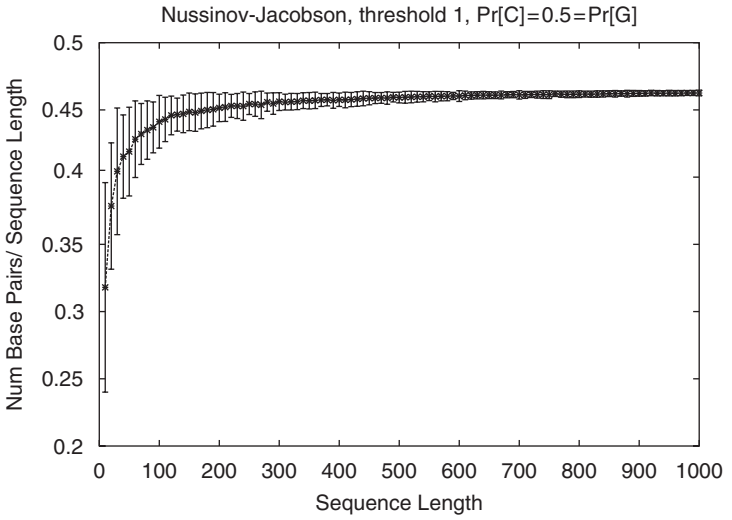

Fig. 1. Graph of the average number of base pairs in random RNA divided by length of RNA sequence. Values graphed come from columns 1 and 4 of Table 1.

the compositional frequency is 0.25 for each base A,C,G,U.[5] In the Appendix, we extend the asymptotic limit result in three directions: (i) we consider the more realistic Turner [33] energy model using Zuker's algorithm [20,34], as well as the Nussinov–Jacobson [24] energy model; (ii) we consider random RNA as generated from a kth order Markov chain, for arbitrary but fixed $k \geqslant 1$ (the asymptotic limit proved in the main part of the text concerns 0th order Markov chain); (iii) we consider not only the mean minimum free energy (mfe) per nucleotide of random RNA, but the standard deviation of the mfe per nucleotide as well as higher order moments. This extension is placed in an appendix, since the focus of the current paper is combinatorial; i.e. to prove exact values or lower and upper bounds for the asymptotic limit of the expected maximum number of base pairs of random RNA as a function of compositional frequency. A companion paper [7] to this article focuses on the Turner energy model, dinucleotide frequencies, random RNA generated by a first-order Markov chain, Z-scores, p-values and asymptotic Z-scores to quantify the extent to which (structural) RNA has lower folding energy[6] than random RNA of the same dinucleotide frequency.

Obtaining provable, exact values for asymptotic limits of expected maximum number of base pairs for random RNA of different compositional frequencies seems currently to be an intractable problem, so to shed light on this problem, we study the expected maximum number of base pairs for random strings in $\{0, 1\}^*$ having a minimum number k of hairpin loops, each having a threshold of size t. Here, in analogy to RNA, we allow base pairings between distinct symbols (0 with 1, but not 0 with 0 or 1 with 1)—alternatively expressed, we consider RNA strings containing only A, U or only C, G. For this binary alphabet problem, the asymptotic ratio of the expected maximum number of base pairs over n is compared to $D(p)$, the "dual" of the well-known constant $L(p)$, the latter defined as the asymptotic ratio of the expected length of the longest common subsequence (LCS) of two random strings of length n divided by n, where bits are generated randomly and independently, 1 with probability p and 0 with probability $1 - p$.

2. Computer experiments

Figures and tables from our computer experiments are found at the end of the paper. Some additional data, as well as a short, self-contained proof of Lemma 7, is available in the web supplement found at `http://bioinformatics.bc.edu/clotelab/`. Throughout the paper, we consider an RNA sequence s to be a word over the finite alphabet

[5] If only Watson–Crick base pairing is allowed, then clearly the maximum number of base pairs for RNA sequence s is bounded above by $\min(|s|_A, |s|_U) + \min(|s|_C, |s|_G)$, where $|s|_x$ denotes the number of occurrences of x in sequence s. To avoid obvious trivialities of this form, Conjecture 5 requires that mononucleotide frequencies $q_A = q_U$ and $q_C = q_G$.

[6] The folding energy of an RNA sequence s is the minimum free energy of s, as computed by Zuker's algorithm [34] using the Turner energy model [33]—i.e. using Zuker's `mfold` or `RNAfold` from the Vienna RNA Package.

$\{A, C, G, U\}$; i.e. $s \in \{A, C, G, U\}^*$. Given RNA sequences s, t, we write the concatenation of s with t by $s \cdot t$, or sometimes even st.

Recall that a *secondary structure* for an RNA sequence $a = a_1 \cdots a_n \in \{A, C, G, U\}^n$ is an expression $s = s_1 \cdots s_n$ involving dot, left and right parenthesis, which is well-balanced, such that nucleotides corresponding to matching parentheses are either Watson–Crick complements or GU wobble pairs. We say that a secondary structure has *threshold* θ, if hairpin loops have at least θ unpaired bases.

Formally, define a secondary structure S on RNA sequence $a_1, \ldots, a_n$ to be a set of ordered pairs (i, j) corresponding to base pair positions, where $i < j$ and the following requirements are satisfied.

(1) *Watson–Crick or GU wobble pairs:* If (i, j) belongs to S, then pair (a_i, a_j) must be one of the following canonical base pairs: (A, U), (U, A), (G, C), (C, G), (G, U), (U, G).[7]
(2) *Nonexistence of pseudoknots:* If (i, j) and (k, ℓ) belong to S, then it is not the case that $i < k < j < \ell$.
(3) *No base triples:* If (i, j) and (i, k) belong to S, then $j = k$; if (i, j) and (k, j) belong to S, then $i = k$.
(4) *Threshold requirement:* If (i, j) belongs to S, then $j - i > \theta$.

A base pair (x, y) is *interior* to base pair (i, j) if $i < x < y < j$; one also says that (i, j) is *exterior* to (x, y).

In [24] Nussinov and Jacobson present a dynamic programming algorithm to compute the maximum number of base pairs in a secondary structure for a given RNA sequence. This $O(n^3)$ time algorithm is the basis for the more realistic Zuker algorithm [34], as implemented in `mfold` and in Vienna RNA package `RNAfold`. Since the current paper concerns a mathematical analysis of asymptotic properties of RNA, we adopt the simpler Nussinov–Jacobson algorithm.

We now describe four methods of generating random RNA sequences: *Markov* 0, *Markov*1, *Shuffle*, *Dishuffle*. The first method is known as the random word model, or more precisely a 0th order Markov chain.

Algorithm 1 (*Markov 0*). INPUT: *An RNA sequence* $a_1, \ldots, a_n$.
OUTPUT: *An RNA sequence* $x_1, \ldots, x_n$ *of the same* expected *mononucleotide frequency as* $a_1, \ldots, a_n$.

(1) Compute the mononucleotide frequency of $a_1, \ldots, a_n$.
(2) For $i = 1, \ldots, n$, generate x_i by sampling from mononucleotide frequency.

The next method generates a random sequence by taking a random walk on a first-order Markov chain, whose transitional probabilities are obtained from measured dinucleotide frequencies.

Algorithm 2 (*Markov1*). INPUT: *An RNA sequence* $a_1, \ldots, a_n$.
OUTPUT: *An RNA sequence* $x_1, \ldots, x_n$ *of the same* expected *dinucleotide frequency as* $a_1, \ldots, a_n$.

(1) Compute the mono- and dinucleotide frequency of $a_1, \ldots, a_n$.
(2) Generate x_1 by sampling from mononucleotide frequency.
(3) Generate remaining nucleotides $x_2, \ldots, x_n$ by sampling from the conditional probabilities $Pr[X|Y]$, where $Pr[X|Y]$ equals the dinucleotide frequency that nucleotide X follows Y divided by mononucleotide frequency of nucleotide Y.

The next method is a trivial shuffle, familiar to beginning students of computer science.

Algorithm 3 (*Shuffle*). INPUT: *An RNA sequence* $a_1, \ldots, a_n$.
OUTPUT: *An RNA sequence* $x_1, \ldots, x_n$ *of the same* exact *mononucleotide frequency as* $a_1, \ldots, a_n$.

(1) Choose a random permutation $\sigma \in S_n$.
(2) For $i = 1$ to n, set $x_i = a_{\sigma(i)}$.

[7] At times, we may disallow wobble pairs. Note that there is even an option in `RNAfold` of Vienna RNA Package [16,15] which disallows wobble pairs.

The last method is a clever dinucleotide shuffle process, due to Altschul and Erikson [2], which preserves the same exact dinucleotide count. (Web server and Python source code for this algorithm is available in the web supplement. See also [10] for a recent web server, which implements the Altschul–Erikson algorithm for k-tuple shuffles, for arbitrary but fixed k.)

Algorithm 4 (*Dishuffle*). INPUT: *An RNA sequence* $a_1, \ldots, a_n$.

OUTPUT: *An RNA sequence* $x_1, \ldots, x_n$ *of the same* exact *dinucleotide frequency as* $a_1, \ldots, a_n$, *where* $x_1 = a_1, x_n = a_n$; *moreover, the Altschul–Erikson algorithm even produces the same number of dinucleotides of each type AA, AC, AG, AU, CA, CC, etc.*

(1) For each nucleotide $x \in \{A, C, G, U\}$, create a list L_x of edges $x \to y$ such that the dinucleotide xy occurs in the input RNA.
(2) For each nucleotide $x \in \{A, C, G, U\}$ distinct from the last nucleotide x_n, randomly choose an edge from the list L_x. Let E be the set of chosen edges (note that E contains at most three elements).
(3) Let G be the graph, whose edge set is E and whose vertex set consists of those nucleotides x, y such that $x \to y$ is an edge in E. If there is a vertex of G which is not connected to the last nucleotide a_n, then return to (2).
(4) For each nucleotide $x \in \{A, C, G, U\}$, permute the edges in $L_x - E$. Append to the end of each L_x any edges from E which had been removed.
(5) For $i = 1$ to $n - 1$, generate x_{i+1} by taking the next available nucleotide such that $x_i \to x_{i+1}$ belongs to the list L_{x_i}.

The proof of correctness of the Altschul–Erikson dinucleotide shuffle algorithm depends on well-known criteria for the existence of an Euler tour in a directed graph. See [2] for details of Algorithm 4 and its extensions.

Now, given an RNA sequence s of length n, by the previous four methods, we can generate many random sequences t of the same length n, guaranteed to have the same expected or exact mono- or dinucleotide frequency as that of s, depending on choice of algorithm. While the theoretical contribution of this paper focuses on the *random word model* or 0th order Markov chain, we experimented with each of the four algorithms to generate random sequences.

2.1. miRNA versus random RNA

The results of this section suggest that functionally important RNA, such as precursor micro-RNA (miRNA) from the Rfam database [13], have more base pairs than that of random RNA of the same expected mononucleotide and/or dinucleotide frequency, as computed by our implementation of the Nussinov–Jacobson algorithm [5,24].[8] In computations described in this section we allow GU wobble pairs, in addition to Watson–Crick pairs, and alternately investigate the situation with threshold 0 and 3.

We computed the mono- and dinucleotide frequencies of 506 precursor miRNAs, with sequence data taken from Bonnet et al. [3] (the data of Bonnet et al. was extracted from Rfam), as well as the minimum, maximum, average and standard deviation of the precursor miRNA lengths. Table 2 and Fig. 2 indicate clearly that precursor miRNA has more base pairs than random RNA, when applying the Nussinov–Jacobson algorithm with threshold 3, where random RNA is generated by each of Algorithms 1–4. In contrast, Table 3 and Fig. 3 indicate that for the biologically irrelevant case of threshold 0, there is no such phenomenon. See the web supplement for additional experiments with transfer RNA, type III hammerhead ribozymes and riboswitches, all of which yield that real RNA has more base pairs than random RNA. Though a crude approximation to the real energy model, the Nussinov–Jacobson energy model does indicate, for threshold 3, that structural RNA has more base pairs than random RNA. Unlike the Turner energy model, the Nussinov–Jacobson energy model is simple enough to allow us to establish numerical limits and upper and lower

[8] As shown in [6], most structurally important RNA has lower *folding energy* than random RNA, where folding energy is measured using Zuker's algorithm, as implemented in `mfold` or `RNAfold`. Although the Nussinov–Jacobson energy model, in particular computing the maximum number of base pairs, is a crude approximation to the real energy model, other classes of RNA (tRNA, hammerhead ribozymes, riboswitches) illustrate consistently that when applying the Nussinov–Jacobson algorithm for threshold 3, real RNA has more base pairs than random RNA. For the biologically irrelevant case of threshold 0, this is no longer the case. See additional data, tables and figures in the web supplement of this paper.

Table 2
Descriptional statistics for the number of base pairs divided by sequence length for a collection of 506 precursor miRNAs (miRNA sequence data from [3]) and for random RNA, according to Algorithms 1–4

	Mean	StDev	Max	Min
miRNA	0.396588	0.017623	0.445783	0.353535
Markov0	0.365564	0.019093	0.424242	0.247312
Markov1	0.366969	0.020685	0.427083	0.154930
Shuffle	0.371138	0.012318	0.418919	0.319444
Dishuffle	0.374058	0.012256	0.430380	0.323529

For each miRNA, 100 random RNAs of the same size were generated, and the number of basepairs was computed, using our implementation of the Nussinov–Jacobson algorithm, where Watson–Crick and GU base pairs are allowed, with threshold set to 3. Table values concern the ratio of number of base pairs over sequence length. The theoretical analysis of the current paper principally concerns random RNA generated by Algorithm 1. For each method of generating random RNA, the mean number of base pairs is less than that of real RNA.

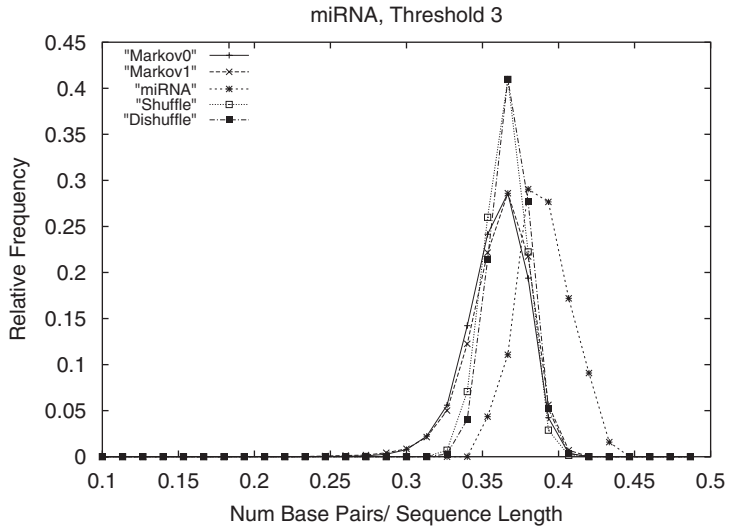

Fig. 2. This figure displays the ratio of number of base pairs over sequence length, for 506 precursor miRNAs (sequence data taken from [3]). Number of base pairs was computed using our implementation of the Nussinov–Jacobson algorithm, allowing Watson–Crick and GU wobble pairs with threshold of 3. The histogram of number of base pairs divided by sequence length for precursor miRNA lies to the right of the histograms produced by each of the four methods for generating random RNA—Algorithms 1–4. Histograms were obtained by generating, for each miRNA sequence, 100 random RNAs per real RNA, using each of the four methods discussed. Descriptional statistics for these graphs are given in Table 2.

Table 3
Descriptional statistics generated in an identical manner to those from Table 2, with the exception that threshold is set to 0

	Mean	StDev	Max	Min
miRNA	0.436927	0.017835	0.484536	0.368421
Markov0	0.424635	0.032262	0.494737	0.225806
Markov1	0.421845	0.034673	0.500000	0.140845
Shuffle	0.439826	0.017681	0.494118	0.363636
Dishuffle	0.437898	0.018040	0.494118	0.355263

Note the anomaly in this case of threshold 0, that random RNA obtained by both shuffling methods has a *larger* average number of base pairs divided by sequence length. Structural RNA has been under selective pressure to have lower folding energy than random RNA [6]. Although the Nussinov–Jacobson energy model is a crude approximation for the real energy model, in the case of threshold 3, random RNA appears to have fewer base pairs than real RNA. In the biologically irrelevant case of threshold 0, this is no longer the case.

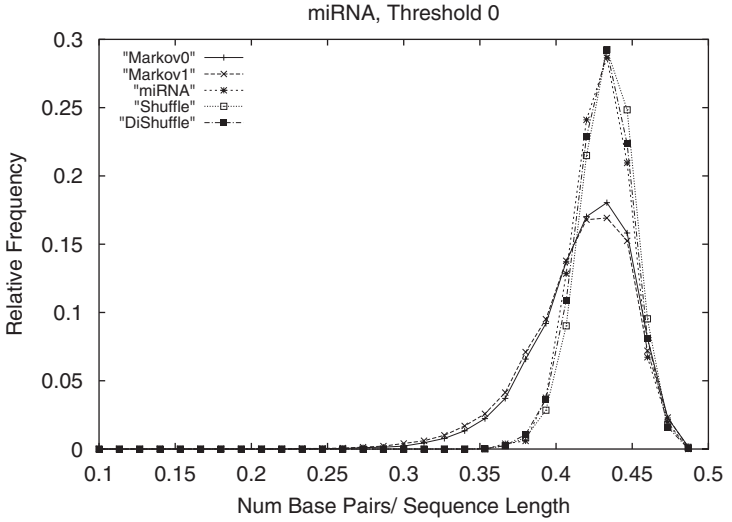

Fig. 3. This figure differs from Fig. 2 only in that threshold 0 was taken in the Nussinov–Jacobson algorithm, rather than threshold 3. Descriptional statistics are given in Table 3.

bounds for the maximum number of base pairs. In the companion paper [6], we compute Z-scores and p-values to study how the folding energy of real RNA compares with random RNA.

2.2. Simulations suggest an asymptotic limit

In Table 6, we give a table of the expected number BP of base pairs in random RNA with varying compositional frequency, no GU bonds, threshold 0, and string length 500, where the average was taken over 100 random sequences per fixed compositional frequency. While it is clear that a maximum number of base pairs is obtained for compositional frequency of 0.5 for each of A, U (or of C, G),[9] it is not obvious how to prove the following conjecture.

Conjecture 5. *Let $\theta \geqslant 0$ be a fixed threshold, and $n \geqslant \theta + 2$ arbitrary. Let q_A, q_C, q_G, q_U be compositional (mononucleotide) frequencies of A, C, G, U, satisfying $q_A = q_U$, $q_C = q_G$ and $q_A + q_U + q_C + q_G = 1$. Generate random RNA sequences of length n, obtained by appending nucleotides, where A is appended with probability q_A, C with probability q_C, G with probability q_G and U with probability q_U. Let $BP(n)$ be the expected maximum number of base pairs in such random sequences, where Watson–Crick but no GU pairs are allowed. Then $BP(n)$ achieves a minimum with the uniform distribution $q_A = q_U = q_C = q_G = 0.25$.*

Table 6 provides evidence for the likelihood of this conjecture, when $n = 500$ and threshold $\theta = 0$.

Some classes of RNA have compositional frequencies of approximately $q_A = q_U = q_C = q_G = 0.25$, so the above conjecture might suggest that such RNA is optimized for structural *instability*, which appears to contradict the data presented in Fig. 2. As previously mentioned, the case of threshold 0 is biologically irrelevant; moreover, for simplicity, we have disallowed GU wobble pairs, and in counting the maximum number base pairs, there is no distinction between GC and AU base pairs. These ignored factors are all biologically relevant. Table 4 illustrates the expected mfe, as computed by Version 1.4 of Vienna RNA Package `RNAfold`, in random RNA with varying compositional frequency, no GU bonds, threshold 3, and string length 500, where the average was taken over 100 random sequences per fixed compositional frequency. In this case, uniform compositional frequency $q_A = q_U = q_C = q_G = 0.25$ does not yield a maximum mfe value.

[9] The formal proof is left to the reader; however, the idea is that by replacing all A's by C's and U's by G's, the number BP cannot decrease.

Table 4
Table of expected minimum free energy (mfe) in random RNA with varying compositional frequency, no GU bonds, threshold 3, string length 500, average values over 100 iterations, using Vienna RNA Package `RNAfold`

q_A	q_C	q_G	q_U	Mean	StDev	Max	Min	Ratio
0.0000	0.5000	0.5000	0.0000	−475.5846	8.5438	−458	−495	−0.951169
0.0156	0.4844	0.4844	0.0156	−436.4168	12.5359	−402	−460	−0.872834
0.0312	0.4688	0.4688	0.0312	−408.6424	16.4497	−371	−441	−0.817285
0.0469	0.4531	0.4531	0.0469	−372.4908	14.5766	−333	−410	−0.744982
0.0625	0.4375	0.4375	0.0625	−348.0432	15.4277	−311	−377	−0.696086
0.0781	0.4219	0.4219	0.0781	−322.3006	13.6317	−294	−353	−0.644601
0.0938	0.4062	0.4062	0.0938	−296.3844	15.8813	−266	−339	−0.592769
0.1094	0.3906	0.3906	0.1094	−273.6242	14.7602	−230	−305	−0.547248
0.1250	0.3750	0.3750	0.1250	−251.7838	16.8178	−207	−285	−0.503568
0.1406	0.3594	0.3594	0.1406	−235.4784	14.4806	−198	−271	−0.470957
0.1562	0.3438	0.3438	0.1562	−214.0264	12.2795	−187	−245	−0.428053
0.1719	0.3281	0.3281	0.1719	−200.9744	12.8800	−171	−236	−0.401949
0.1875	0.3125	0.3125	0.1875	−185.1242	10.8777	−162	−205	−0.370248
0.2031	0.2969	0.2969	0.2031	−170.3640	13.3179	−146	−204	−0.340728
0.2188	0.2812	0.2812	0.2188	−156.9842	14.6035	−125	−196	−0.313968
0.2344	0.2656	0.2656	0.2344	−143.9242	11.6329	−123	−167	−0.287848
0.2500	0.2500	0.2500	0.2500	−132.1930	10.2426	−112	−157	−0.264386
0.2656	0.2344	0.2344	0.2656	−120.3692	7.8277	−103	−140	−0.240738
0.2812	0.2188	0.2188	0.2812	−109.9630	8.2964	−97	−127	−0.219926
0.2969	0.2031	0.2031	0.2969	−101.8248	9.4531	−76	−120	−0.203650
0.3125	0.1875	0.1875	0.3125	−93.6436	7.6687	−76	−112	−0.187287
0.3281	0.1719	0.1719	0.3281	−84.3984	7.3432	−72	−99	−0.168797
0.3438	0.1562	0.1562	0.3438	−78.6708	7.2501	−61	−93	−0.157342
0.3594	0.1406	0.1406	0.3594	−71.9054	7.5020	−54	−85	−0.143811
0.3750	0.1250	0.1250	0.3750	−65.8294	6.2562	−54	−85	−0.131659
0.3906	0.1094	0.1094	0.3906	−61.3780	5.3742	−50	−71	−0.122756
0.4062	0.0938	0.0938	0.4062	−59.4470	5.6267	−44	−76	−0.118894
0.4219	0.0781	0.0781	0.4219	−56.9226	4.0502	−44	−66	−0.113845
0.4375	0.0625	0.0625	0.4375	−56.4236	4.8504	−47	−68	−0.112847
0.4531	0.0469	0.0469	0.4531	−57.3960	4.4266	−50	−71	−0.114792
0.4688	0.0312	0.0312	0.4688	−62.3824	5.4518	−51	−75	−0.124765
0.4844	0.0156	0.0156	0.4844	−68.1114	4.8157	−57	−81	−0.136223
0.5000	0.0000	0.0000	0.5000	−79.6382	4.9443	−66	−88	−0.159276

Note that while Table 6 illustrates our conjecture that the uniform compositional frequency $p_A = p_C = p_G = p_U = 0.25$ yields the fewest base pairs as computed by the Nussinov–Jacobson algorithm, the situation is radically different when computing with Zuker's algorithm, as implemented in Vienna RNA Package. For the latter, $p_A = 0.4375 = p_U$, $p_C = 0.0625 = p_G$ appears to yield the highest (i.e. negative with smallest absolute value) minimum free energy. This is in part because AU bonds are weaker than GC bonds (for this experiment we have disallowed GU base pairs, to allow comparison of results between Tables 6 and 4.

Further simulations suggest an asymptotic limit phenomenon. For any fixed compositional frequency (see Fig. 4 and Table 5), for instance $q_A = q_C = q_G = q_U = 0.25$, we generated random RNA sequences of length n by Algorithm 1, computed the number BP of base pairs in the Nussinov–Jacobson optimal structure, and determined the ratio BP/n. See Figs. 1, 5 and 7, and Tables 6 and 7, which illustrate the dependence of this asymptotic limit on the compositional frequency, for fixed threshold. The remainder of the paper furnishes proof of this asymptotic limit phenomenon, as well as upper and lower bounds in the case of binary sequences.

3. Expected maximum number of base pairs in labeled secondary structures

We now prove the existence of an asymptotic limit, as suggested by the computer experiments from the previous section.

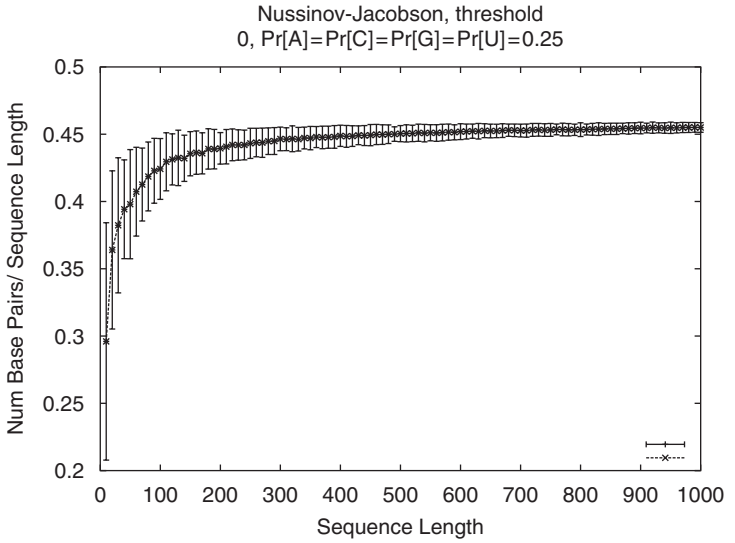

Fig. 4. Graph of the average number of base pairs in random RNA divided by length of RNA sequence. Values graphed come from columns 1 and 4 of Table 5.

Table 5
Table of number *BP* of base pairs, ratio of base pairs to sequence length, etc. for random RNA sequences of length n generated by Algorithm 1 to have expected mononucleotide frequencies: $q_A = q_C = q_G = q_U = 0.25$

n	*BP*	StDev	*BP*/n	Error	Max	Min
10	2.9600	0.8823	0.2960	0.0882	5	1
100	42.4200	2.2635	0.4242	0.0226	46	35
200	87.9100	2.3499	0.4395	0.0117	92	82
300	133.9500	2.6434	0.4465	0.0088	139	127
400	179.5300	3.1574	0.4488	0.0079	187	171
500	225.2100	2.9776	0.4504	0.0060	232	216
600	271.1300	3.5543	0.4519	0.0059	278	263
700	316.8800	3.3654	0.4527	0.0048	324	308
800	362.7400	4.0562	0.4534	0.0051	371	351
900	409.1800	3.8429	0.4546	0.0043	418	400
1000	455.0100	3.5539	0.4550	0.0036	463	443

Our implementation of the Nussinov–Jacobson algorithm was used with Watson–Crick base pairs (no GU base pairs), threshold 0, and sequence length up to 1000. Average values were taken over 100 iterations, where error means Stdev/n; points are indicated along with error bars.

Definition 6. A function f defined on the positive integers is said to be *superadditive* if for all integers s, s',

$$f(s) + f(s') \leqslant f(s + s'). \tag{1}$$

Similarly, a function f is said to be *subadditive* if

$$f(s + s') \leqslant f(s) + f(s'). \tag{2}$$

The following useful lemma is due to Fekete [11]; see also Steele [28] for extensions and additional information. For the sake of completeness, we include a short self-contained proof in the web supplement.

Lemma 7 (*Superadditivity Lemma, Fekete [11]*). *For any superadditive function f, the limit*

$$\lim_{n\to\infty} \frac{f(n)}{n} = \sup_{n\geqslant 1} \frac{f(n)}{n}$$

always exists.

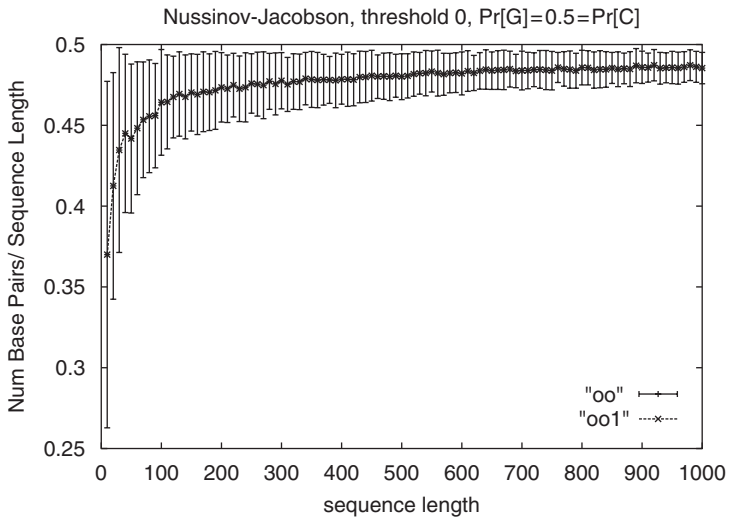

Fig. 5. Graph of data from Table 7; values from column 4 (expected number of base pairs of random RNA divided by sequence length n) are graphed as a function of those from column 1 (sequence length n).

Fix arbitrary compositional frequencies q_A, q_C, q_G, q_U with $q_A + q_C + q_G + q_U = 1$. For integer n let $E(n, q_A, q_C, q_G, q_U)$ denote the expected number of base pairs in an optimal secondary structure (i.e. having maximum number of base pairs) for random RNA of length n generated by sampling the compositional frequencies q_A, q_C, q_G, q_U (i.e. 0th order Markov chain). Since the compositional frequencies are fixed throughout, we write $E(n)$ instead of $E(n, q_A, q_C, q_G, q_U)$. Writing 4^n instead of $\{A, C, G, U\}^n$, let $N(s)$ be the number of base pairs in an optimal secondary structure on RNA sequence $s \in 4^n$ when applying the Nussinov–Jacobson algorithm (i.e. $N(s)$ is the maximum number of base pairs in a secondary structure on s). Of course $N(s)$ depends on fixed threshold θ, so we should really write $N(\theta, s)$, but the existence of a limit is independent of the value of θ.[10]

Lemma 8. *For fixed compositional frequencies q_A, q_C, q_G, q_U, $E(n)$ is superadditive.*

Proof.

$$
\begin{aligned}
E(n+m) &= \sum_{r \in 4^{n+m}} \Pr[r] \cdot N(r) \\
&= \sum_{s \in 4^n} \sum_{t \in 4^m} \Pr[s] \Pr[t] N(st) \\
&\geqslant \sum_{s \in 4^n} \sum_{t \in 4^m} \Pr[s] \Pr[t] (N(s) + N(t)) \\
&= \sum_{t \in 4^m} \Pr[t] \cdot \sum_{s \in 4^n} \Pr[s] N(s) + \sum_{s \in 4^n} \Pr[s] \cdot \sum_{t \in 4^m} \Pr[t] N(t) \\
&= \sum_{s \in 4^n} \Pr[s] N(s) + \sum_{t \in 4^m} \Pr[t] N(t) \\
&= E(n) + E(m).
\end{aligned}
$$

This concludes the proof of Lemma 8. □

[10] In `mfold`, θ is taken to be 3.

Table 6
Table of expected number *BP* of base pairs in random RNA with varying compositional frequency, no GU bonds, threshold 0, string length 500, average values were taken over 100 iterations

q_A	q_C	q_G	q_U	Mean	StDev	Max	Min	Ratio
0.0000	0.5000	0.5000	0.0000	240.0300	7.0999	250	220	0.480060
0.0156	0.4844	0.4844	0.0156	238.4600	6.6550	249	219	0.476920
0.0312	0.4688	0.4688	0.0312	237.6100	6.6210	247	222	0.475220
0.0469	0.4531	0.4531	0.0469	236.5900	6.4095	247	222	0.473180
0.0625	0.4375	0.4375	0.0625	236.0400	5.9547	246	220	0.472080
0.0781	0.4219	0.4219	0.0781	234.9800	5.5479	244	219	0.469960
0.0938	0.4062	0.4062	0.0938	234.2100	5.6909	243	216	0.468420
0.1094	0.3906	0.3906	0.1094	233.6500	5.6752	244	216	0.467300
0.1250	0.3750	0.3750	0.1250	232.9300	5.5177	242	217	0.465860
0.1406	0.3594	0.3594	0.1406	232.4200	5.6217	242	217	0.464840
0.1562	0.3438	0.3438	0.1562	232.0500	5.3859	241	217	0.464100
0.1719	0.3281	0.3281	0.1719	231.7900	5.2140	241	217	0.463580
0.1875	0.3125	0.3125	0.1875	231.4400	4.9302	241	220	0.462880
0.2031	0.2969	0.2969	0.2031	231.0700	4.9925	241	219	0.462140
0.2188	0.2812	0.2812	0.2188	231.0100	4.9163	242	217	0.462020
0.2344	0.2656	0.2656	0.2344	231.1300	5.0807	242	219	0.462260
0.2500	0.2500	0.2500	0.2500	231.0500	4.7167	241	221	0.462100
0.2656	0.2344	0.2344	0.2656	231.3900	4.4898	241	222	0.462780
0.2812	0.2188	0.2188	0.2812	231.3600	4.6444	241	222	0.462720
0.2969	0.2031	0.2031	0.2969	231.3500	4.7463	241	221	0.462700
0.3125	0.1875	0.1875	0.3125	231.5700	4.5503	240	222	0.463140
0.3281	0.1719	0.1719	0.3281	231.9200	4.7972	241	223	0.463840
0.3438	0.1562	0.1562	0.3438	232.2600	5.1529	241	217	0.464520
0.3594	0.1406	0.1406	0.3594	232.6500	5.3224	243	216	0.465300
0.3750	0.1250	0.1250	0.3750	233.0400	5.2305	246	218	0.466080
0.3906	0.1094	0.1094	0.3906	233.6700	5.4223	245	221	0.467340
0.4062	0.0938	0.0938	0.4062	234.5400	5.3989	245	220	0.469080
0.4219	0.0781	0.0781	0.4219	234.9500	5.4118	243	221	0.469900
0.4375	0.0625	0.0625	0.4375	236.0400	5.3646	245	223	0.472080
0.4531	0.0469	0.0469	0.4531	237.1000	5.7541	245	222	0.474200
0.4688	0.0312	0.0312	0.4688	237.6200	6.0527	247	221	0.475240
0.4844	0.0156	0.0156	0.4844	238.4600	6.5459	249	218	0.476920
0.5000	0.0000	0.0000	0.5000	240.0300	7.0999	250	220	0.480060

Table 7
Table of number *BP* of base pairs, ratio of base pairs to sequence length, etc. for random binary sequences of length n generated by Algorithm 1 to have expected mononucleotide frequencies: $q_G = q_C = 0.5$

n	*BP*	StDev	*BP*/n	Error	Max	Min
10	3.7000	1.0724	0.3700	0.1072	5	1
100	46.4200	3.2716	0.4642	0.0327	50	35
200	94.7000	4.3070	0.4735	0.0215	100	82
300	143.2900	5.2293	0.4776	0.0174	150	127
400	191.3900	6.1788	0.4785	0.0154	200	175
500	240.0300	7.0999	0.4801	0.0142	250	220
600	289.2600	7.9531	0.4821	0.0133	300	266
700	338.7600	8.3751	0.4839	0.0120	350	311
800	388.5900	8.5687	0.4857	0.0107	400	358
900	437.2800	9.0278	0.4859	0.0100	450	414
1000	485.3800	9.7445	0.4854	0.0097	500	462

Our implementation of the Nussinov–Jacobson algorithm was used with Watson–Crick base pairs (no GU base pairs), threshold 0, and sequence length up to 1000. Average values were taken over 100 iterations, where error means Stdev/n; points are indicated along with error bars. By Theorem 11, the asymptotic limit is 0.5.

Note that the previous lemma depends on two conditions.

(1) Random words are generated by a 0th order Markov process, which implies that $\Pr[st] = \Pr[s] \cdot \Pr[t]$, where st is the concatenation of sequence s followed by sequence t.
(2) $N(st) \geqslant N(s) + N(t)$. This is clear, since the union of a secondary structure for s and one for t yields a valid secondary structure for st, and so the maximum number of base pairs in a secondary structure for st is at least $N(s) + N(t)$.

The condition (2) is *not* always valid for the Turner energy rules. For instance, if $s = \texttt{CCCUUUGGG} = t$, then Vienna RNA package `RNAfold` yields

```
CCCUUUGGG
(((...)))
```
minimum free energy $= -0.90$ kcal/mol

```
CCCUUUGGGCCCUUUGGG
(((...(....)...)))
```
minimum free energy $= -3.30$ kcal/mol

where the mfe structure for s and for t each has 3 base pairs, but that for the concatenation st has only 4.

Theorem 9 (*Asymptotic expected maximum number of base pairs*). *For any compositional frequencies q_A, q_C, q_G, q_U, there exists a limit $L(q_A, q_C, q_G, q_U)$ such that the expected maximum number of base pairs in random RNA of given compositional frequency and of length n is asymptotically equal to $n \cdot L(q_A, q_C, q_G, q_U)$.*

Proof. By Lemma 8, $E(n)$ is superadditive, so by Lemma 7, the limit $\lim_{n\to\infty} E(n)/n$ exists. □

In the remainder of the paper, we provide rigorous upper and lower bounds for the asymptotic limit of the expected maximum number of base pairs for random RNA as a function of the compositional frequency. In the appendix, we prove an asymptotic limit for mean and standard deviation (as well as higher order moments) of mfe per nucleotide of random RNA of a given compositional frequency, where mfe is computed by Zuker's algorithm using the Turner energy model.

3.1. Motivation

We now turn to the question of computing the asymptotic limit, whose existence was just shown. Throughout the remainder of the paper, we will consider a binary alphabet 0, 1, instead of the usual RNA nucleotides A, C, G, U—this would correspond to the (unrealistic) case where an RNA sequence consisted only of C, G or only of A, U. It is hoped that our analysis of asymptotics for a binary alphabet may deliver techniques useful for the general problem.

For the purposes of combinatorial analysis a *secondary structure* is modeled as an outerplanar graph with vertices $1, 2, \ldots, n$ such that there is at most one edge between any two vertices.[11] A *base pair* (i, j), where $i < j$, is an edge connecting two positions of the RNA sequence $a_1, \ldots, a_n$. Base pair (x, y) is *interior* to base pair (i, j) if $i < x < y < j$; equivalently, (i, j) is said to be *exterior* to (x, y). A *labeled secondary structure*, denoted LSS, differs from a secondary structure in two respects: first, the bases are labeled by either 0 or 1, and second, a base is paired with another base only if their labels are *different*. Thus a LSS is an outerplanar graph with n vertices such that the valence of every vertex is at most one and vertices are labeled with either 0 or 1 in such a way that vertices i, j connected by an edge (i, j) must have different labels. Graph vertices will indistinguishably be called nodes and bases, and edges will be called base pairs.

[11] Technically, the graphs we consider are labeled, outerplanar graphs with valence 1. For display, we order the vertices $1, 2, \ldots, n$ along a horizontal line, and depict edges by arcs above this line. While an outerplanar graph has all edges depicted by arcs above the line, a planar graph could additionally have arcs below the line, corresponding for instance to a type-H pseudoknot.

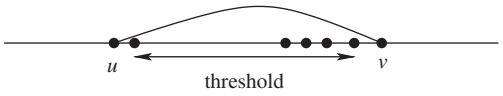

Fig. 6. Two bases u and v and a threshold of size t.

The structural components of RNA secondary structures (see [5,31]) are stacked base pairs, hairpin loops, bulges, interior loops, and multiloops. Such components, with the exception of hairpin loops, are not important for our analysis. In the sequel we are interested in secondary structures having at least k hairpin loops, each having at least threshold θ (see Fig. 6). For our purposes, a hairpin loop in a secondary structure for a given binary sequence is given by a base pair (i, j) having no interior base pairs. Such a base pair has threshold θ if $j - i = \theta + 1$; i.e. $j = i + \theta + 1$ and positions $i + 1, \ldots, i + \theta$ do not belong to any base pair.

Definition 10. Consider a LSS with a random distribution of $0 - 1$ labels. A *threshold position* is a collection of unpaired bases delimited by a base pair (see Fig. 6).

A 0-threshold is a base pair (i, j) with $j = i + 1$. In general, an LSS may have several hairpin loops, each having possibly different thresholds. The threshold of the LSS is defined to be the minimum threshold over all its hairpin loops.

3.2. Secondary structures with 0-thresholds

We can prove the following theorem that gives the asymptotic behavior of the expected maximum number of base pairs of a random labeled secondary structure.

Theorem 11 ($0 - 1$ *Algorithm*). *Let* $E_0(n, p)$ *be the expected maximum number of base pairs for a random word* $s \in \{0, 1\}^n$, *where* s *is generated by Algorithm* 1 *and probability of generating* 1 *is* p, *while that of* 0 *is* $1 - p$. *Then*

$$\lim_{n\to\infty} \frac{E_0(n, p)}{n} = \min\{p, 1 - p\}.$$

Moreover, the resulting max size base pairing has no threshold positions.

Proof. For any binary string s of length n which contains i many 0's, $E_0(s) = \min(i, n - i)$. To see this consider the following algorithm.

Algorithm	$0 - 1$ Algorithm
Input:	A string $s_1 s_2 \cdots s_n$ of bits of length n.
Output:	An optimal secondary structure.
	1. Repeat as long as two adjacent bases with different labels exist; 2. Basepair any two adjacent bases with different labels; 3. Remove the paired bases and go to step 1;

To analyze this algorithm we use the DeMoivre–Laplace theorem. Letting $q = 1 - p$, and recalling standard notation for the binomial probability distribution, where $b(i; n, p)$ denotes $\binom{n}{i} p^i q^{n-i}$, we have

$$E_0(n, p) = \sum_{i=0}^{n} \min(i, n - i) \binom{n}{i} p^i q^{n-i} = A + B,$$

where

$$A = \sum_{i=0}^{\lfloor n/2 \rfloor} i \binom{n}{i} p^i q^{n-i}$$

$$= np \sum_{i=1}^{\lfloor n/2 \rfloor} \binom{n-1}{i-1} p^{i-1} q^{(n-1)-(i-1)}$$

$$= np \sum_{j=0}^{\lfloor n/2 \rfloor - 1} \binom{n-1}{j} p^j q^{(n-1)-j}$$

$$= np \sum_{j=0}^{\lfloor n/2 \rfloor - 1} b(j; n-1, p),$$

$$B = \sum_{i=0}^{\lfloor (n-1)/2 \rfloor} i \binom{n}{i} q^i p^{n-i}$$

$$= nq \sum_{j=0}^{\lfloor (n-1)/2 \rfloor - 1} b(j; n-1, q).$$

Before proceeding, we define the notation $f \sim g$ to mean that $\lim_{n\to\infty} \frac{f}{g} = 1$.[12] Now, by the DeMoivre–Laplace theorem (a version of the central limit theorem—see Feller [12, p. 182]),

$$\sum_{j=0}^{\lfloor n/2 \rfloor - 1} b(j; n-1, p) \sim \Phi\left(\frac{n/2 - 1 - (n-1)p}{\sqrt{(n-1)pq}}\right) \sim \Phi\left(\frac{\sqrt{n-1}(\frac{1}{2} - p)}{\sqrt{p - p^2}}\right),$$

$$\sum_{j=0}^{\lfloor (n-1)/2 \rfloor - 1} b(j; n-1, q) \sim \Phi\left(\frac{(n-1)/2 - 1 - (n-1)q}{\sqrt{(n-1)pq}}\right) \sim \Phi\left(\frac{\sqrt{(n-1)}(p - \frac{1}{2})}{\sqrt{p - p^2}}\right),$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-t^2/2}\, \mathrm{d}t,$$

denotes the cumulative distribution function for the standard normal distribution with mean 0 and standard deviation 1. Thus

$$A = np \sum_{j=0}^{\lfloor n/2 \rfloor - 1} b(j; n-1, p) \sim np\Phi\left(\frac{\sqrt{n-1}(\frac{1}{2} - p)}{\sqrt{p - p^2}}\right),$$

$$B = nq \sum_{j=0}^{\lfloor (n-1)/2 \rfloor - 1} b(j; n-1, q) \sim nq\Phi\left(\frac{\sqrt{n-1}(p - \frac{1}{2})}{\sqrt{p - p^2}}\right).$$

Let x_n denote the expression $\frac{\sqrt{n-1}(\frac{1}{2}-p)}{\sqrt{p-p^2}}$. By properties of the normal distribution, the following is evident: for $p < \frac{1}{2}$, as n tends to infinity, $\Phi(x_n)$ tends to 1, while $\Phi(-x_n)$ tends to 0; for $p > \frac{1}{2}$, as n tends to infinity, $\Phi(x_n)$ tends to 0,

[12] In analysis and number theory (e.g. [14, p. 7]) and in some probability texts (e.g. [12]), the notation $\sim$ is used in this context. This should not be confused with the statistics notation $X \sim D$, which means that random variable X has probability distribution D.

while $\Phi(-x_n)$ tends to 1. Thus if $p < \frac{1}{2}$ we have

$$\lim_{n\to\infty} \frac{E_0(n, p)}{n} = p(1-0) + 0 = p,$$

while if $p > \frac{1}{2}$ we have

$$\lim_{n\to\infty} \frac{E_0(n, p)}{n} = p(0-1) + 1 = 1 - p.$$

This completes the proof of Theorem 11. □

4. Asymptotics of optimal secondary structures

In this section we consider asymptotics of optimal secondary structures with bases labeled with 0, 1. We will extend the asymptotic result of Theorem 11 to the case of secondary structures with a given threshold size.[13] In answer to a question of one of the referees:

Definition 12. Consider a sequence $s = s_1, s_2, \ldots, s_n$ of 0's and 1's. Given integers k, t we consider the combinatorial function $N_{k,t}(s)$ which is defined as the maximum number of base pairs of an optimal secondary structure over the string s with at least k threshold positions and each threshold position has at least t unpaired bases.

Let $s = s_1, s_2, \ldots, s_n$ be a sequence of independent and identically distributed $\{0, 1\}$-valued random variables, where 1s are generated with probability p and 0's with probability $1 - p$.

Assume that k is a *subadditive* integer valued function; i.e. k satisfies

$$k(m+n) \leqslant k(m) + k(n) \quad \text{for all integers } m, n. \tag{3}$$

Definition 13. Let $E_{k,t}(n, p)$ be the expected maximum number of base pairs of an optimal secondary structure over a random string s of length n with at least $k(n)$ threshold positions and each threshold position has size at least t unpaired bases. Formally we define

$$E_{k,t}(n, p) = \sum_{s\in\{0,1\}^n} N_{k(n),t}(s)\, p^{|s|_1}(1-p)^{|s|_0},$$

where $|s|_0$, $|s|_1$ is the number of 0s and 1s in s.

Lemma 14. *Assume that k is a subadditive integer valued function, and that $0 \leqslant p \leqslant 1$ is fixed. Then $E_{k,t}(n, p)$ is superadditive as a function of n; moreover, the limit*

$$E_{k,t}(p) := \lim_{n\to\infty} \frac{E_{k,t}(n, p)}{n}$$

exists.

Proof. Let s, s' be two strings of length m and n, respectively, with at least $k(m)$ and $k(n)$ thresholds each, respectively, and each threshold of size at least t. If we concatenate the two strings $s \in \{0, 1\}^m$ and $s' \in \{0, 1\}^n$ we form the string ss' which will have at least $k(m)+k(n)$ thresholds and each threshold of size at least t. It follows easily from Inequality (3) that

$$N_{k(m),t}(s) + N_{k(n),t}(s') \leqslant N_{k(m+n),t}(ss').$$

[13] From the annotation of base pairs in the 50 S large ribosomal unit (PDB code `1FFK`, NDB ID `RR0011`, we computed 75 hairpin loops, with an average of 5.4 unpaired bases per hairpin loop, using only *cis* (anti-parallel) Watson–Crick base pairs. This answers a question of one of the referees.

Finally, we can prove the superadditivity of $E_{k,t}(n, p)$. Indeed,

$$\begin{aligned}
E_{k,t}(m, p) + E_{k,t}(n, p) &= \sum_s N_{k(m),t}(s) p^{|s|_1} (1-p)^{|s|_0} \\
&\quad + \sum_{s'} N_{k(n),t}(s') p^{|s'|_1} (1-p)^{|s'|_0} \\
&= \sum_{ss'} N_{k(m),t}(s) p^{|ss'|_1} (1-p)^{|ss'|_0} \\
&\quad + \sum_{ss'} N_{k(n),t}(s') p^{|ss'|_1} (1-p)^{|ss'|_0} \\
&= \sum_{ss'} (N_{k(m),t}(s) + N_{k(n),t}(s')) p^{|ss'|_1} (1-p)^{|ss'|_0} \\
&\leqslant \sum_{ss'} N_{k(m+n),t}(ss') p^{|ss'|_1} (1-p)^{|ss'|_0} \\
&= E_{k,t}(m+n, p).
\end{aligned}$$

The existence of the limit is an immediate consequence of the superadditivity of $E_{k,t}(n, p)$. This completes the proof of Lemma 14. □

Our goal is to prove the following theorem.

Theorem 15. *Fix t, let $0 \leqslant p \leqslant 1$ and let k be any subadditive integer valued function which satisfies $\lim_{n\to\infty} k(n)/n = 0$. Then the limit $\lim_{n\to\infty} \frac{E_{k,t}(n,p)}{n}$ exists; moreover,*

$$\begin{aligned}
p(1-p) + p^2\left(1 - \frac{p}{p^2 - p + 1}\right) &\leqslant \lim_{n\to\infty} \frac{E_{k,t}(n, p)}{n} = \lim_{n\to\infty} \frac{E_{1,t}(n, p)}{n} \\
&\leqslant \min\{p, 1-p\}. \qquad (4)
\end{aligned}$$

The rest of this section is devoted to the proof of this theorem. The proof will follow a detour in which we will first consider the simpler problem of the longest dual-common subsequence of two random sequences (see Section 4.1) as well as an optimization result concerning the position of the threshold in an optimal secondary structure with a single threshold position (see Section 4.2).

4.1. Longest dual-common subsequences

Let $s, s' \in \{0,1\}^n$ be two strings $s = s_1 s_2 \cdots s_n$, $s' = s'_1 s'_2 \cdots s'_n$ of length n. A *common* subsequence of s and s' is determined by sequences $i_1 < i_2 < \cdots < i_k \leqslant n$ and $j_1 < j_2 < \cdots < j_k \leqslant n$ of indices such that $s_{i_r} = s'_{j_r}$, for all $r = 1, 2, \ldots, k$. The integer k is called the length of the common subsequence. Given $0 \leqslant p \leqslant \frac{1}{2}$, let the sequence of bits be generated randomly and independently, where 1s are generated with probability p and 0s with probability $1-p$. Let $C(n, p)$ be the expected length of the longest common subsequence of two random sequences s and s' and let $C(p) := \lim_{n\to\infty} C(p, n)/n$. The *longest common subsequence problem* goes back to [4,25] and concerns the computation of $C(\frac{1}{2})$. Related to this is proving that $C(\frac{1}{2}) < C(p)$, for $p \neq \frac{1}{2}$. Both of these are open problems.

Of interest to us is the dual problem which we now define. A *dual-common* subsequence of s and s' is determined by sequences $i_1 < i_2 < \cdots < i_k \leqslant n$ and $j_1 < j_2 < \cdots < j_k \leqslant n$ of indices such that $s_{i_r} \neq s'_{j_r}$, for all $r = 1, 2, \ldots, k$. The integer k is called the length of the dual-common subsequence. Let $D(n, p)$ be the expected length of the longest dual-common subsequence of two random sequences s and s' and let $D(p) := \lim_{n\to\infty} D(p, n)/n$. In this section we prove the following result.

Theorem 16. *For any* p, $D(p) > 2p(1-p) + p^2(1-p/(p^2-p+1))$.

Proof. Before proving the main theorem we digress in order to derive two useful results using Chernoff bounds. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent, identically distributed $\{0,1\}$-valued random variables such that 1s are generated with probability p and 0s with probability $1-p$. Let N be the random variable that counts the number of occurrences of the pattern 01 in $X_1, X_2, \ldots, X_n$. We have

Lemma 17.

$$\Pr\left[N \geqslant \left(1 - \sqrt{\frac{2\ln n}{np(1-p)}}\right)(n-1)p(1-p)\right] \geqslant 1 - \frac{2}{n}. \tag{5}$$

Proof. Consider the indicator random variables I_i, $i \geqslant 2$, where $I_i = 1$ if 01 ends in position i, and is 0 otherwise. Clearly, I_i, I_j are independent random variables if and only if $|i-j| \geqslant 2$. Define the random variables

$$N = \sum_{i=2}^{n} I_i, \quad N_0 = \sum_{i=1}^{\lfloor n/2 \rfloor} I_{2i}, \quad N_1 = \sum_{i=2}^{\lfloor (n+1)/2 \rfloor} I_{2i-1}. \tag{6}$$

Since $E[I_i] = p(1-p)$, it is clear that

$$\mu := E[N] = \sum_{i=2}^{n} E[I_i] = (n-1)p(1-p),$$

$$\mu_0 := E[N_0] = \sum_{i=1}^{\lfloor n/2 \rfloor} E[I_{2i}] = \lfloor n/2 \rfloor p(1-p),$$

$$\mu_1 := E[N_1] = \sum_{i=1}^{\lfloor (n+1)/2 \rfloor} E[I_{2i-1}] = \lfloor (n+1)/2 \rfloor p(1-p).$$

Using Chernoff bounds (see [21]) for $0 < \delta < 1$ we see that

$$\Pr[N_0 \geqslant (1-\delta)\mu_0] \geqslant 1 - \exp(-\mu_0\delta^2/2),$$

$$\Pr[N_1 \geqslant (1-\delta)\mu_1] \geqslant 1 - \exp(-\mu_1\delta^2/2).$$

Let A_0 and A_1 denote the events "$N_0 \geqslant (1-\delta)\mu_0$" and "$N_1 \geqslant (1-\delta)\mu_1$", respectively, and observe that

$$\begin{aligned}
\Pr[N \geqslant (1-\delta)\mu] &= \Pr[N_0 + N_1 \geqslant (1-\delta)\mu] \\
&\geqslant \Pr[N_0 \geqslant (1-\delta)\mu_0 \text{ and } N_1 \geqslant (1-\delta)\mu_1] \\
&= \Pr[A_0 \text{ and } A_1] \\
&= 1 - \Pr[\overline{A_0} \text{ or } \overline{A_1}] \\
&\geqslant 1 - \Pr[\overline{A_0}] - \Pr[\overline{A_1}] \\
&= \Pr[A_0] + \Pr[A_1] - 1 \\
&\geqslant 1 - \exp(-\mu_0\delta^2/2) - \exp(-\mu_1\delta^2/2).
\end{aligned}$$

Now if we choose $\delta = \sqrt{\frac{2\ln n}{np(1-p)}}$ then after a few elementary calculations we derive easily that

$$\Pr\left[N \geqslant \left(1 - \sqrt{\frac{2\ln n}{np(1-p)}}\right)(n-1)p(1-p)\right] \geqslant 1 - \frac{2}{n}.$$

This completes the proof of Lemma 17. □

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ be two sequences of independent, identically distributed $\{0, 1\}$-valued random variables such that 1s are generated with probability p and 0s with probability $1 - p$. Let $N(X)$ and $N(Y)$ be the random variables that count the number of occurrences of the pattern 01 and 10 in the sequences $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$, respectively. Finally, define the random variable $N(X, Y) := \min\{N(X), N(Y)\}$. We have:

Lemma 18.

$$E[N(X,Y)] \geqslant \left(1 - \sqrt{\frac{2\ln n}{np(1-p)}}\right)(n-1)p(1-p)\left(1 - \frac{2}{n}\right)^2. \tag{7}$$

Proof. Using Lemma 17 we derive that for $\delta = \sqrt{\frac{2\ln n}{np(1-p)}}$,

$$\begin{aligned}
&\Pr[N(X,Y) \geqslant (1-\delta)(n-1)p(1-p)] \\
&\quad = \Pr[\min\{N(X), N(Y)\} \geqslant (1-\delta)(n-1)p(1-p)] \\
&\quad = \Pr[N(X) \geqslant (1-\delta)(n-1)p(1-p)] \cdot \Pr[N(Y) \geqslant (1-\delta)(n-1)p(1-p)] \\
&\quad \geqslant \left(1 - \frac{2}{n}\right)^2.
\end{aligned}$$

Using this last result we can estimate the expected value of the random variable N, we have that

$$\begin{aligned}
E[N(X,Y)] &= \sum_{k=0}^{n} \Pr[N(X,Y) \geqslant k] \\
&\geqslant \sum_{0 \leqslant k \leqslant (1-\delta)(n-1)p(1-p)} \Pr[N(X,Y) \geqslant k] \\
&\geqslant \left(1 - \sqrt{\frac{2\ln n}{np(1-p)}}\right)(n-1)p(1-p)\left(1 - \frac{2}{n}\right)^2. \qquad \square
\end{aligned}$$

Now we can turn to proving the theorem. For this purpose, let us assume that $s = s_1 s_2 \cdots s_n$ and $s' = s'_1 s'_2 \cdots s'_n$ be two binary strings generated randomly and independently, where 1s are generated with probability p and 0s with probability $1 - p$.

Fix a nonnegative $r \leqslant n$ and consider all substrings 01 in $s_{r+1} s_{r+2} \cdots s_{n-r}$ and all substrings 10 in $s'_{r+1} s'_{r+2} \cdots s'_{n-r}$. Now pair the ith 01 substring with the ith 10 substring obtaining the ith *block*. Let B_r be the ordered set of blocks, from left to right say. Since s and s' have uniform distributions, we can use Lemma 18 to lower bound the expected size of B_r:

$$E[|B_r|] \geqslant \left(1 - \sqrt{\frac{2\ln(n-2r)}{(n-2r)p(1-p)}}\right)(n-2r-1)p(1-p)\left(1 - \frac{2}{n-2r}\right)^2. \tag{8}$$

Every block gives rise to two matchings (base pairs); see Fig. 7. Therefore the expected length of the longest dual-common subsequence between s and s' is at least $2E[|B_r|]$. Passing to the limit we observe that for any $r = o(n)$,

$$D(p) \geqslant \lim_{n\to\infty} \frac{2E[|B_r|]}{n} \geqslant 2p(1-p).$$

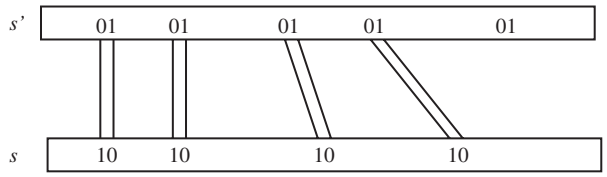

Fig. 7. Pairing patterns 01 and 10 from s and s', respectively, in order to form two matchings (or base pairs). As a result, we obtain four blocks.

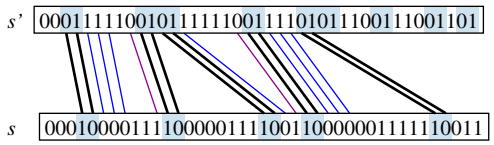

Fig. 8. There are 10 matchings arising from the five blocks. Further matchings can be obtained by considering 011^k and 0^k01 in s, and 100^k and 1^k10 in s', respectively.

Next we improve the lower bound by considering substrings in s of the form 011^k and 0^k01 for $1 \leqslant k \leqslant r \leqslant n$. First consider the set C_r of all substrings in s of the form 011^k where $1 \leqslant k \leqslant r$, the leading 01 is in a block $b \in B_r$, and 100^k is the substring of s' with leading 10 in the block b. Obviously, for each $011^k \in C_r$, we can add one matching (that has not been added yet) into the dual-common subsequence determined by B_r. In particular we can match the last 1 in $011^k \in s$ with the last 0 in the corresponding $100^k \in s'$; see Fig. 8.

Consider the indicator random variables I_i^k, $r+1 \leqslant i \leqslant n-r-2$, $1 \leqslant k \leqslant r$, where $I_i^k = 1$, if $011^k \in C_r$ and it starts in position i, and is 0 otherwise. Obviously,

$$|C_r| = \sum_{k=1}^{r} \sum_{i=r+1}^{n-r-2} I_i^k.$$

To estimate the probabilities $\Pr[I_i^k = 1]$, call an ordered pair (t, t') of two binary strings of equal length "good" if t has at most as many 01's as t' has 10's. If a pair (t, t') is "bad", i.e. t has more 01's as t' has 10's, then the pair (reverse(t'), reverse(t)) is "good". Thus, for each bad pair there is a unique good pair. Note that there are pairs that have equal number of 01 and 10, respectively, and these are good pairs. Thus, $\Pr[(t, t') \text{ is good}] > \frac{1}{2}$.

Let A be the event that (s, s') is good. Now,

$$\Pr[I_i^k = 1] \geqslant \Pr[I_i^k = 1 \text{ AND } A] = \Pr[A] \cdot \Pr[I_i^k = 1 \mid A] > \frac{p^{k+1}(1-p)^{k+1}}{2}.$$

The last inequality follows from conditioning on A since then each substring 01 of s is in a block and thus the first 10 in 100^k in s' is guaranteed by the event A.

Therefore,

$$\begin{aligned} E[|C_r|] &= \sum_{k=1}^{r} \sum_{i=r+1}^{n-r-2} E[I_i^k] > \sum_{k=1}^{r} (n-2r-2)\, p^{k+1}(1-p)^{k+1}/2 \\ &\geqslant \frac{(n-2r-2)}{2} \left[p^2 \frac{p^2-2p+1}{p^2-p+1} - \frac{p(1-p)(p-p^2)^r}{p^2-p+1} \right]. \end{aligned} \tag{9}$$

Second consider the set D_r of all substrings in s of the form 0^k01 where $1 \leqslant k \leqslant r$, the ending 01 is in a block $b \in B_r$, and 1^k10 is the substring of s' with ending 10 in the block b. Obviously, for each $0^k01 \in D_r$, we can add one matching (that has not been added yet) into the dual-common subsequence determined by $B_r \cup C_r$. (Notice that $C_r \bigwedge D_r \subseteq B_r$,

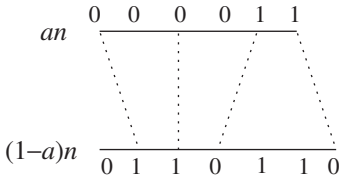

Fig. 9. An optimal matching between two arrays one of size an and the other of size $(1-a)n$.

where $\bigwedge$ returns substrings of s that appear in both C_r and D_r.) In particular we can match the first 0 in $0^k01 \in s$ with the first 1 in the corresponding $1^k10 \in s'$; see Fig. 8. One can show that

$$E[|D_r|] > \frac{(n-2r-2)}{2}\left[p^2\frac{p^2-2p+1}{p^2-p+1} - \frac{p(1-p)(p-p^2)^r}{p^2-p+1}\right]. \tag{10}$$

Therefore all together there will be at least $2|B_r| + |C_r| + |D_r|$ matchings (base pairs) between s and s'.

Using (8)–(10), and the linearity of expectation, the average number of matchings between two random strings will be

$$\begin{aligned} M &= 2E[|B_r|] + E[|C_r|] + E[|D_r|] \\ &> 2\left(1-\sqrt{\frac{2\ln(n-2r)}{(n-2r)p(1-p)}}\right)(n-2r-1)p(1-p)\left(1-\frac{2}{n-2r}\right)^2 \\ &\quad + (n-2r-2)\left[p^2\frac{p^2-2p+1}{p^2-p+1} - \frac{p(1-p)(p-p^2)^r}{p^2-p+1}\right]. \end{aligned}$$

Passing to the limit, we observe that for $r = \log(n)$,

$$D(p) > \lim_{n\to\infty}\frac{M}{n} \geqslant 2p(1-p) + p^2\left(1-\frac{p}{p^2-p+1}\right).$$

This completes the proof of Theorem 16. □

4.2. Dual-common subsequences and single thresholds

We would like to relate the dual-common subsequence problem and the expected maximum number of base pairs in secondary structures by showing that the number of base pairs is maximized when the threshold is at the centre of the secondary structure.

Before providing the details of the proof we explain several ideas on optimal secondary structures. Consider a sequence $s = s_1s_2\cdots s_n$ of 0s and 1s.

Definition 19. For a given rational number a, where $0 \leqslant a \leqslant 1$, an a-matching for s is a matching without crossings between the subsequences $s = s_1s_2\cdots s_{an}$ (depicted as the top row in Fig. 9) and $s = s_{an+1}s_{an+2}\cdots s_n$ (depicted as the bottom row in Fig. 9), where an edge between i (where $1 \leqslant i \leqslant an$) and j (where $an+1 \leqslant j \leqslant n$) may exist only if $s_i \neq s_j$.

Definition 20. An a-matching is called *optimal* if the number of its edges is maximum. Let $f_a(s)$ be the number of edges of an optimal a-matching for s.

Let $s = s_1s_2\cdots s_n$ be a sequence of independent and identically distributed $\{0, 1\}$-valued random variables, where 1s are generated with probability p and 0s with probability $1-p$, and $0 \leqslant p \leqslant \frac{1}{2}$. Let the expected size of an optimal

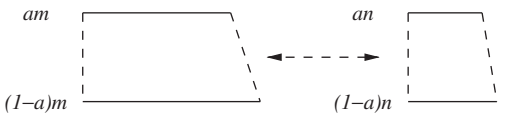

Fig. 10. a-matchings for two strings one of size m and another of size n.

a-matching of a random string s of length n be defined by

$$M_a(n, p) = \sum_{s\in\{0,1\}^n} f_a(s)p^{|s|_1}(1-p)^{|s|_0},$$

where $|s|_0$, $|s|_1$ is the number of 0s and 1s in s.

Lemma 21. *For each rational number* $0 \leqslant a \leqslant \frac{1}{2}$ *the limit*

$$\ell_a(p) := \lim_{n\to\infty} \frac{M_a(n, p)}{n}$$

exists. Moreover, $\ell_a(p)$ *is maximized for* $a = \frac{1}{2}$. *In particular,*

$$D(p) = 2 \cdot \ell_{1/2}(p).$$

Proof. The second identity $D(p) = 2 \cdot \ell_{1/2}(p)$ is an immediate consequence of the definition of the dual-common subsequence. So we concentrate on the rest of the lemma. The existence of the limit will follow from the superadditivity of the function $M_a(n)$. Indeed, we want to prove that for any m, n, $M_a(m) + M_a(n) \leqslant M_a(m+n)$. Let s and s' be two strings of length m and n, respectively, and consider the two a-matchings depicted in Fig. 10. By superimposing the two a-matchings a new a-matching is formed. The two arrays to the top form a new array of size $a(m+n)$ and the two arrays to the bottom a new array of size $(1-a)(m+n)$. Since the resulting a-matching includes all the edges of the two previous a-matchings it follows that $f_a(ss')$ is at least $f_a(s) + f_a(s')$, where ss' is the concatenation of s and s'. It follows that

$$\begin{aligned}
M_a(m, p) + M_a(n, p) &= \sum_{s\in\{0,1\}^m} f_a(s)p^{|s|_1}(1-p)^{|s|_0} \\
&\quad + \sum_{s'\in\{0,1\}^n} f_a(s')p^{|s'|_1}(1-p)^{|s'|_0} \\
&= \sum_{ss'} f_a(s)p^{|ss'|_1}(1-p)^{|ss'|_0} \\
&\quad + \sum_{ss'} f_a(s')p^{|ss'|_1}(1-p)^{|ss'|_0} \\
&= \sum_{ss'\in\{0,1\}^{m+n}} (f_a(s) + f_a(s'))p^{|ss'|_1}(1-p)^{|ss'|_0} \\
&\leqslant \sum_{ss'\in\{0,1\}^{m+n}} f_a(ss')p^{|ss'|_1}(1-p)^{|ss'|_0} \\
&= M_a(m+n, p).
\end{aligned}$$

The existence of the limit is now an immediate consequence of the superadditivity of the function $M_a(n, p)$ just proved.

Next we prove that ℓ_a is monotone in a. Indeed, assume that $a < b \leqslant \frac{1}{2}$. We will show that $\ell_a \leqslant \ell_b$. In the sequel we will provide a transformation $s \to s'$ that transforms a sequence s into a new sequence s' and an a-matching for s into

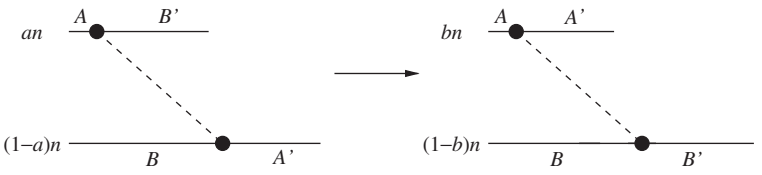

Fig. 11. Transforming a-matching of type $(a, 1-a)$ into a-matching of type $(b, 1-b)$.

a b-matching for s' (see Fig. 11). Consider a sequence s as depicted on the left side of Fig. 11. Since $b \leqslant \frac{1}{2}$ we observe that $1-a \geqslant \frac{1}{2} \geqslant b$. We are looking for a "cut" of the top and bottom rows of the leftmost sequence that forms pieces A, B' on the top and pieces B, A' on the bottom row, respectively, in such a way that $|A| = xn$, $|B'| = (a-x)n$ and $|B| = \frac{1-a}{a}xn$, $|A'| = (1-a)n - \frac{1-a}{a}xn$. We would like to "swap" the position of the pieces A', B' in such a way that the resulting top row (consisting of the pieces A, A') has length bn while the bottom row (consisting of the pieces B, B') has length $(1-b)n$ (see Fig. 11). The value of x that will achieve the desired cut is easy to determine by observing that thelength of A plus the length of A' must be equal to bn, i.e.

$$xn + (1-a)n - \frac{1-a}{a}xn = bn. \tag{11}$$

Solving Eq. (11) for x, we derive that

$$x = \frac{a(1-a-b)}{1-2a}.$$

Sequence s' is formed by attaching A' to A and B' to B.

Let $\bar{s}_1$ be formed from segments A and B and $\bar{s}_2$ be formed from segments B' and A', where $s = \bar{s}_1\bar{s}_2$. Let $\bar{s}'_2$ be formed by swapping A' and B'. Clearly, $s' = \bar{s}_1\bar{s}'_2$ and

$$f_a(\bar{s}_1) + f_a(\bar{s}_2) \leqslant f_b(\bar{s}_1\bar{s}'_2) = f_b(s'). \tag{12}$$

It follows from the definition of the expected value that

$$M_a\left(\frac{x}{a}n, p\right) = \sum_{\bar{s}_1} f_a(\bar{s}_1)p^{|\bar{s}_1|_1}(1-p)^{|\bar{s}_1|_0}$$

and

$$M_a\left(\frac{a-x}{a}n, p\right) = \sum_{\bar{s}_2} f_a(\bar{s}_2)p^{|\bar{s}_2|_1}(1-p)^{|\bar{s}_2|_0}.$$

Using these identities and inequality (12) we obtain

$$\begin{aligned}
M_a\left(\frac{x}{a}n, p\right) + M_a\left(\frac{a-x}{a}n, p\right) &= \sum_{\bar{s}_1} f_a(\bar{s}_1)p^{|\bar{s}_1|_1}(1-p)^{|\bar{s}_1|_0} \\
&\quad + \sum_{\bar{s}_2} f_a(\bar{s}_2)p^{|\bar{s}_2|_1}(1-p)^{|\bar{s}_2|_0} \\
&= \sum_{\bar{s}_1\bar{s}_2} (f_a(\bar{s}_1) + f_a(\bar{s}_2))p^{|\bar{s}_1\bar{s}_2|_1}(1-p)^{|\bar{s}_1\bar{s}_2|_0} \\
&\leqslant \sum_{\bar{s}_1\bar{s}'_2} f_b(\bar{s}_1\bar{s}'_2)p^{|\bar{s}_1\bar{s}'_2|_1}(1-p)^{|\bar{s}_1\bar{s}'_2|_0} \\
&= \sum_{s'} f_b(s')p^{|s'|_1}(1-p)^{|s'|_0} \\
&\leqslant M_b(n, p).
\end{aligned}$$

Dividing both sides of the resulting inequality by n this implies that

$$\begin{aligned}&\frac{M_a((x/a)n,\,p)}{n}+\frac{M_a(((a-x)/a)n,\,p)}{n}\\&=\frac{x}{a}\cdot\frac{M_a((x/a)n,\,p)}{(x/a)n}+\frac{a-x}{a}\cdot\frac{M_a(((a-x)/a)n,\,p)}{((a-x)/a)n}\\&\leqslant\frac{M_b(n,\,p)}{n}.\end{aligned}$$

Using the last inequality and passing to the limit as $n\to\infty$ we obtain that

$$\ell_a=\frac{x}{a}\cdot\ell_a+\frac{a-x}{a}\cdot\ell_a\leqslant\ell_b.$$

This completes the proof of Lemma 21. □

Proof of Theorem 15. The upper bound

$$\lim_{n\to\infty}\frac{E_{k,t}(n,\,p)}{n}\leqslant\lim_{n\to\infty}\frac{E_0(n,\,p)}{n}\leqslant\min\{p,\,1-p\}$$

is an immediate consequence of Theorem 11. It remains to prove that

$$\lim_{n\to\infty}\frac{E_{k,t}(n,\,p)}{n}\geqslant p(1-p)+p^2\left(1-\frac{p}{p^2-p+1}\right).$$

Divide the secondary structure into $k(n)$ pieces each of size $n/k(n)$. On each piece consider a string $s^{(i)}\in\{0,\,1\}^{n/k(n)}$, $i=1,\,2,\,\ldots,\,k(n)$. Observe that

$$N_{1,t}(s^{(1)})+N_{1,t}(s^{(2)})+\cdots+N_{1,t}(s^{(k(n))})\leqslant N_{k(n),t}(s^{(1)}s^{(2)}\cdots s^{(k(n))}),$$

where $s^{(1)}s^{(2)}\cdots s^{(k(n))}$ denotes string concatenation. It follows that

$$\begin{aligned}k(n)E_{1,t}(n/k(n),\,p)&=\sum_{i=1}^{k(n)}\sum_{s^{(i)}}N_{1,t}(s^{(i)})p^{|s^{(i)}|_1}(1-p)^{|s^{(i)}|_0}\\&=\sum_{s^{(1)}\cdots s^{(k(n))}}(N_{1,t}(s^{(1)})+\cdots+N_{1,t}(s^{(k(n))}))\\&\quad\times p^{|s^{(1)}\cdots s^{(k(n))}|_1}(1-p)^{|s^{(1)}\cdots s^{(k(n))}|_0}\\&\leqslant\sum_{s^{(1)}\cdots s^{(k(n))}}N_{k(n),t}(s^{(1)}\cdots s^{(k(n))})\\&\quad\times p^{|s^{(1)}\cdots s^{(k(n))}|_1}(1-p)^{|s^{(1)}\cdots s^{(k(n))}|_0}\\&=E_{k,t}(n,\,p).\end{aligned}$$

Dividing by n we obtain that

$$\frac{E_{1,t}(n/k(n),\,p)}{n/k(n)}\leqslant\frac{E_{k,t}(n,\,p)}{n}.$$

Clearly $\frac{E_{k,t}(n,p)}{n}\leqslant\frac{E_{1,t}(n,p)}{n}$, so in passing to the limit as $n\to\infty$, we have

$$\lim_{n\to\infty}\frac{E_{1,t}(n,\,p)}{n}=\lim_{n\to\infty}\frac{E_{k,t}(n,\,p)}{n},\tag{13}$$

for any subadditive function $k(n)$ satisfying $k(n)=o(n)$.

Consider now two random strings $s, s' \in \{0,1\}^n$, generated by appending independently generated random bits, 1 with probability p and 0 with probability $1-p$, and let s'' denote the string reversal of s'. For constant t, the expected maximum number of base pairs in a secondary structure on a string of the form $s0^t s'$, where the threshold spans the inserted subword 0^t, is clearly the expected maximum number of edges between s and s'', hence equal to $D(n, p)$.

Recall that by Theorem 16 we have $D(p) \geqslant 2p(1-p)+2p^2(1-\frac{p}{p^2-p+1})$, for any p, and that by Lemma 21 we have $D(p)=2\cdot\ell_{\frac{1}{2}}(p)$. It follows that $\ell_{\frac{1}{2}}(p) \geqslant p(1-p)+p^2(1-\frac{p}{p^2-p+1})$, for any p. It follows that $\frac{D(n,p)}{n} \leqslant \frac{E_{1,t}(2n+t,p)}{2n+t}\cdot\frac{2n+t}{n}$. By taking the limit as n tends to infinity and applying Theorem 16, we have

$$p(1-p)+p^2\left(1-\frac{p}{p^2-p+1}\right) \leqslant \frac{D(p)}{2} \leqslant \lim_{n\to\infty}\frac{E_{1,t}(n,p)}{n}.$$

This establishes the proof of Theorem 15. □

5. Conclusion

In this paper, we report results of various computer experiments concerned with random RNA; see the web supplement `http://clavius.bc.edu/~clotelab/` for some of the code used and data obtained. These results suggest an asymptotic limit phenomenon proved to exist in Theorem 9, for which we provide an exact numerical value in Theorem 11 for the case of binary sequences using threshold 0, and for which we give an upper and lower bound in Theorem 15 for the case of binary sequences using threshold t. As a tool, we investigate $D(p)$, the "dual" of the well-known constant $L(p)$; here, $n \cdot L(p)$ is asymptotically the expected length of the longest common subsequence (LCS) of two random sequences of length n, each generated by independently appending random bits, 1 with probability p and 0 with probability $1-p$. Our experiments suggest Conjecture 5, which asserts that under certain conditions the uniform distribution for nucleotides A, C, G, U yields a minimum expected number of base pairs in random RNA. One might wonder whether natural RNA tends roughly to have an equal mononucleotide frequency for each of A, C, G, U in order to maximize instability? Most assuredly not, as illustrated in Table 2 and Fig. 2, which suggest that real RNA has more base pairs than random RNA of the same mono- or dinucleotide frequency. In contrast to Table 6, Table 4 suggests that natural RNA neither has the maximum nor minimum free energy over all possible compositional frequencies.

Acknowledgments

We would like to thank the anonymous referees for their various helpful comments, leading to an improved presentation.

Appendix. Asymptotic limits for subadditive functions

In this section, we prove a far-reaching extension of Theorem 9. Specifically, we prove the existence of an asymptotic limit for the mean and standard deviation of the *minimum free energy* (mfe) per nucleotide, as computed either by the Nussinov–Jacobson algorithm [5,24] or by Zuker's algorithm `mfold` [34], for random RNA of any fixed compositional frequency; additionally, we prove the existence of limits for all higher order moments.

Definition 22. A real-valued function on the integers is subadditive (respectively, superadditive) if for all u, v, $f(u+v) \leqslant f(u)+f(v)$ (respectively, $f(u+v) \geqslant f(u)+f(v)$).

Lemma 23. *Consider a real-valued function f on the integers. If $f \geqslant 0$ (respectively, $f \leqslant 0$) and f is subadditive (respectively, f is superadditive) then so is the function $\frac{f^k(n)}{n^{k-1}}$, for all integers $k \geqslant 1$.*

Proof. If f is subadditive and $f \geqslant 0$ then $-f$ is superadditive and $-f \leqslant 0$. Therefore it is enough to prove the result when f is subadditive and $f \geqslant 0$. The proof is by induction on k. Clearly the result is true for $k = 1$. Assuming it is true

for k, we will show that the function $\frac{f^{k+1}(n)}{n^k}$ is subadditive. Indeed, from the induction hypothesis we have that for $1 \leqslant u, v$,

$$\begin{aligned}
\frac{f^{k+1}(u+v)}{(u+v)^k} &= \frac{f^k(u+v)}{(u+v)^{k-1}} \cdot \frac{f(u+v)}{u+v} \\
&\leqslant \left(\frac{f^k(u)}{(u+v)^{k-1}} + \frac{f^k(v)}{(u+v)^{k-1}} \right) \\
&\quad \cdot \frac{f(u)+f(v)}{u+v} \quad \text{(induction hypothesis)} \\
&\leqslant \left(\frac{f^k(u)}{u^{k-1}} + \frac{f^k(v)}{v^{k-1}} \right) \cdot \frac{f(u)+f(v)}{u+v} \\
&\quad \text{(as } u \leqslant (u+v),\ v \leqslant (u+v),\ f \geqslant 0) \\
&= \frac{u}{u+v} \frac{f^{k+1}(u)}{u^k} + \frac{v}{u+v} \frac{f^{k+1}(v)}{v^k} + \frac{u}{u+v} \frac{f^k(u) f(v)}{u^k} \\
&\quad + \frac{v}{u+v} \frac{f^k(v) f(u)}{v^k}.
\end{aligned}$$

Therefore it is enough to show that this last term is less than or equal to $\frac{f^{k+1}(u)}{u^k} + \frac{f^{k+1}(v)}{v^k}$. If we simplify we obtain that it is enough to prove that

$$\frac{u}{u+v} \frac{f^k(u) f(v)}{u^k} + \frac{v}{u+v} \frac{f^k(v) f(u)}{v^k} \leqslant \frac{v}{u+v} \frac{f^{k+1}(u)}{u^k} + \frac{u}{u+v} \frac{f^{k+1}(v)}{v^k}. \tag{14}$$

In turn, if we multiply out Inequality (14) by $(u+v)u^k v^k$ we obtain the equivalent Inequality (15).

$$uv^k f^k(u) f(v) + vu^k f^k(v) f(u) \leqslant vv^k f^{k+1}(u) + uu^k f^{k+1}(v). \tag{15}$$

After factorization, Inequality (15) becomes equivalent to Inequality (16).

$$\left(\frac{f^k(u)}{u^k} - \frac{f^k(v)}{v^k} \right) \cdot \left(\frac{f(u)}{u} - \frac{f(v)}{v} \right) \geqslant 0, \tag{16}$$

which is always true since $f \geqslant 0$. To see that Inequality (10) is always valid observe that if we put $a := \frac{f(u)}{u}$, $b := \frac{f(v)}{v}$ then the inequality becomes equivalent to

$$(a-b)^2 (a^{k-1} + a^{k-2}b + \cdots + ab^{k-2} + b^{k-1}) \geqslant 0,$$

which is always true since $a, b \geqslant 0$. This completes the proof of Lemma 23. □

As a corollary of this lemma we also obtain the following result.

Lemma 24. *Consider a real-valued function f on the integers. If for some constant $B \geqslant 0$ we have that $f(n) \geqslant -Bn$ (respectively, $f \leqslant Bn$) and f is subadditive (respectively, f is superadditive) then for all integers $k \geqslant 0$ the limit of $\frac{f^k(n)}{n^k}$ exists as $n \to \infty$.*

Proof. As before, without loss of generality we consider only the case when f is subadditive. The proof is by induction on k. Clearly, the result is trivial for $k = 0$. By induction hypothesis the limit of $\frac{f^i(n)}{n^i}$, exists as $n \to \infty$, for all integers $0 \leqslant i \leqslant k-1$.

Next consider the function $g(n) := f(n) + Bn$. Since f is subadditive so is g. It follows from Lemma 14 that for all integers $k \geqslant 1$, the function $\frac{g^k(n)}{n^{k-1}}$ is subadditive. By Fakete's Lemma the limit of $\frac{g^k(n)}{n^k}$ exists as $n \to \infty$. However,

$$\begin{aligned}\frac{g^k(n)}{n^k} &= \frac{(f(n)+Bn)^k}{n^k} \\ &= \sum_{i=0}^{k} \binom{k}{i} \frac{f^i(n)(Bn)^{k-i}}{n^k} \\ &= \sum_{i=0}^{k} \binom{k}{i} B^{k-i} \frac{f^i(n)}{n^i}.\end{aligned}$$

It follows that

$$\frac{g^k(n)}{n^k} = \frac{f^k(n)}{n^k} + \sum_{i=0}^{k-1} \binom{k}{i} B^{k-i} \frac{f^i(n)}{n^i}. \tag{17}$$

However, by induction hypothesis, the limits of all the terms in the sum occurring in Eq. (17) exist. Consequently, since the limit of $\frac{g^k(n)}{n^k}$ exists as $n \to \infty$ so does the limit of $\frac{f^k(n)}{n^k}$ as $n \to \infty$, and the proof of Lemma 24 is complete. □

We now apply Lemma 23 to the kth moment of a random variable X. In our case we have a subadditive random variable X measuring the minimum free energy of a random RNA structure. For such a random variable we know that there is a constant $B > 0$ such that inequality

$$X \geqslant -Bn \tag{18}$$

is valid. We can prove the following lemma.

Theorem 25. *Consider a subadditive random variable X satisfying Inequality* (18). *Then the limit of* $\frac{E[X^k]}{n^k}$ *exists as* $n \to \infty$.

Proof. We imitate the proof of Lemma 24. Consider the random variable $Y := X + Bn$. Since X is subadditive so is Y. It follows from Lemma 23 that for all integers $k \geqslant 1$, the function $\frac{Y^k}{n^{k-1}}$ is subadditive and therefore so is its expected value $\frac{E[Y^k]}{n^{k-1}}$. By Fakete's Lemma the limit of $\frac{E[Y^k]}{n^k}$ exists as $n \to \infty$. It is easy to see that

$$\frac{E[Y^k]}{n^k} = \frac{E[(X+Bn)^k]}{n^k} = \frac{1}{n^k} \sum_{i=0}^{k} \binom{k}{i} (nB)^{k-i} E[X^i]. \tag{19}$$

Now repeating the argument in the proof of Lemma 24, we can easily see that since the limit of $\frac{E[Y^k]}{n^k}$ exists as $n \to \infty$ so does the limit of $\frac{E[X^k]}{n^k}$ as $n \to \infty$. This completes the proof of Theorem 25. □

We can also apply this lemma to show that for the standard deviation of a subadditive random variable X satisfying Inequality (18) the limit of

$$\frac{\sqrt{Var(X)}}{n}$$

exists, as $n \to \infty$.

Theorem 26. *Let X be a subadditive random variable X satisfying Inequality* (18). *Then the limit of*

$$\frac{Var(X)}{n^2} = \frac{E[X^2] - E[X]^2}{n^2}$$

exists, as $n \to \infty$.

Proof. Indeed, by Fakete's lemma since X is subadditive the limit of $\frac{E[X]}{n}$ exists, as $n \to \infty$. As a consequence, also the limit of $\frac{E[X]^2}{n^2}$ exists, as $n \to \infty$. By Theorem 25, the limit of $\frac{E[X^2]}{n^2}$ must exist as $n \to \infty$. This completes the proof of Theorem 26. □

In Theorem 9, we established the existence of an asymptotic limit of the expected maximum number of base pairs in a secondary structure of random RNA, which is generated by Algorithm 1 to have a given expected mononucleotide frequency. Given the generality of the theorems we have just established, we can lift the asymptotic limit result of Theorem 9 in two directions: (i) to consider a more realistic energy model for secondary structure formation, (ii) to consider random RNA generated by Algorithm 2 (resp. by any kth order Markov process). The latter condition ensures that the random RNA which is generated has a given expected dinucleotide frequency (resp. k-tuple frequency). As earlier mentioned, Workman and Krogh [32] have pointed out the importance of conserving dinucleotide frequency when computing Z-scores for minimum free energy of random RNA, so this is a practical concern in applications.

To treat the Turner energy model [33,20], which is the current energy model used in Zuker's algorithm, as implemented in `mfold` and in Vienna RNA Package `RNAfold`, we here redefine (in a trivial manner) the Nussinov–Jacobson energy of RNA sequence $a_1, \ldots, a_n$ to be -1 times the maximum number of base pairs in any secondary structure on $a_1, \ldots, a_n$.[14] The properties for the energy function used in the proof of Theorem 9 (viewed from the standpoint of the new version of the Nussinov–Jacobson energy model) are: (i) subadditivity and (ii) the existence of a lower bound $-Bn$ for the minimum free energy of random RNA of length n, i.e. Inequality (18). Clearly the Nussinov–Jacobson energy model is subadditive and the Nussinov–Jacobson energy of an RNA sequence of length n is greater than or equal to $-n/2$ (a sequence of length n can have at most $n/2$ base pairs). The Turner energy model [33,20] is subadditive and the Turner energy of an RNA sequence of length n is greater than or equal to $-3.42 \cdot n/2$ (a sequence of length n has at most $n/2$ stacked base pairs, and the stacking free energy per base pair is at least -3.42).

Define random variable X by setting $X(n)$ to equal the mfe (according to either the Nussinov–Jacobson or the Turner energy model) of random RNA of length n, which is generated by Algorithm 1 and Algorithm 2 or by a kth order Markov process. Given random RNA sequences $a = a_1, \ldots, a_n$ and $b = b_1, \ldots, b_m$, clearly the mfe of a concatenated with b is at most the mfe of a plus the mfe of b; i.e. subadditivity $X(n+m) \leqslant X(n) + X(m)$. If we apply Theorems 25 and 26 to the random variable X, then we obtain the existence of an asymptotic limit for the expectation and standard deviation (as well as higher moments) *per nucleotide* of random RNA. Here, random RNA can be generated by Algorithms 1 and 2, or even by a kth order Markov process.

The ability to compute mean and standard deviation of the mfe per nucleotide, according to the Turner energy model, of random RNA generated by Algorithm 2 permits us to define the novel notion of *asymptotic Z-score*[8].[15] Let $\vec{q_{xy}} = \langle q_{xy} : x, y \in \{A, C, G, U\}\rangle$ be any *complete* set of *dinucleotide frequencies*; i.e. $0 \leqslant q_{xy} \leqslant 1$ for all $x, y \in \{A, C, G, U\}$ and $\sum_{x,y} q_{xy} = 1$, where the sum is taken over all $x, y \in \{A, C, G, U\}$. Let $\mu(\vec{q_{xy}})$ (resp. $\sigma(\vec{q_{xy}})$) denote the mean μ (resp. standard deviation σ) of mfe per nucleotide of random RNA, whose limit values μ, σ we have just proved to exist and to depend only on the given dinucleotide frequencies $\vec{q_{xy}}$. In practice, this can be approximated by generating according to Algorithm 2 many random RNAs of length n (for n sufficiently large), then computing the mean and standard deviation of the mfe of the random RNA, and finally dividing by n.

---

[14] In the main body of the text, we had defined Nussinov–Jacobson energy to be the maximum number of base pairs in any secondary structure on $a_1, \ldots, a_n$ (a positive number). The trivial change made here entails that energy is negative, and that the Nussinov–Jacobson energy is the minimum free energy according to the Nussinov–Jacobson energy model.

[15] Following Workman and Krogh [32], in Z-score computations involving mfe of RNA secondary structures, also called *folding energy*, it is important to generate random RNA so as to conserve given dinucleotide frequencies. This can be done by Algorithm 2, but not by Algorithm 1.

Definition 27. Given RNA sequence s of length n_0, compute the dinucleotide frequencies $q_{\vec{xy}}$ of s. Define

$$Z^2(s) = \frac{\text{mfe}(s)/n_0 - \mu(q_{\vec{xy}})}{\sigma(q_{\vec{xy}})}.$$

An alternative and more detailed proof, using Kingman's ergodocity theorem for subadditive stochastic processes [19], for the existence of an asymptotic limit for the mean and standard deviation of the mfe per nucleotide for random RNA, generated by Algorithm 2, is given in [8]. The proofs given in this appendix are new and extend both Theorem 9 and the limit theorem proved in [8]. See [8] for details concerning asymptotic Z-scores and applications to RNA.

References

[1] S. Akmaev, Collection of 155 16s rRNA's, personal communication.
[2] S.F. Altschul, B.W. Erikson, Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage, Molecular Biol. Evolution 2 (6) (1985) 526–538.
[3] E. Bonnet, J. Wuyts, P. Rouzé, Y. Vande Peer, Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences, Bioinformatics 20 (2004) 2911–2917.
[4] V. Chvátal, D. Sankoff, Longest common subsequence of two sequences, J. Appl. Probab. 12 (1975) 306–315.
[5] P. Clote, R. Backofen, Computational Molecular Biology: An Introduction, Wiley, New York, 2000, 279p.
[6] P. Clote, F. Ferrè, E. Kranakis, D. Krizanc, Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency, RNA, 2005, in press.
[7] P. Clote, F. Ferrè, E. Kranakis, D. Krizanc, Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency, RNA, 2005, in press.
[8] P. Clote, F. Ferré, E. Kranakis, D. Krizanc, Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency, submitted for publication.
[9] P. Clote, E. Kranakis, D. Krizanc, Asymptotics of random RNA, in: R. Spang, P. Béziat, M. Vingron (Eds.), Currents in Computational Molecular Biology 2003, IEEE, 2003, pp. 149–150, Poster paper.
[10] E. Coward, Shufflet: shuffling sequences while conserving the k-let counts, Bioinformatics 15 (2) (1999) 1058–1059.
[11] M. Fekete, Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzähligen Koeffizienten, Math. Z. 17 (1923) 228–249.
[12] W. Feller, An Introduction to Probability Theory and its Applications, third ed., vol. 1, Wiley, New York, 1968.
[13] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, S.R. Eddy, Rfam: an RNA family database, Nucleic Acids Res. 31 (1) (2003) 439–441.
[14] G.H. Hardy, E.M. Wright, An Introduction to the Theory of Numbers, fifth ed., Oxford University Press, Oxford, 1979.
[15] I. Hofacker, et al., Vienna RNA Package, ⟨http://www.tbi.univie.ac.at/∼ivo/RNA/⟩.
[16] I.L. Hofacker, W. Fontana, P.F. Stadler, L. Sebastian Bonhoeffer, Manfred Tacker, Peter Schuster, Fast folding and comparison of RNA secondary structures, Monatsh. Chem. 125 (1994) 167–188.
[17] I.L. Hofacker, P. Schuster, P.F. Stadler, Combinatorics of RNA secondary structures, Discrete Appl. Math. 88 (1998) 207–237.
[18] S. Karlin, J. Mrazek, A.M. Campbell, Codon usages in different gene classes of the *escherichia coli* genome, Molecular Microbiol. 6 (1998) 1341–1355.
[19] J.F.C. Kingman, Subadditive ergodic theory, Ann. Probab. 1 (6) (1973) 893–909.
[20] D.H. Matthews, J. Sabina, M. Zuker, D.H. Turner, Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure, J. Molecular Biol. 288 (1999) 911–940.
[21] R. Motwani, P. Raghavan, Randomized Algorithms, Cambridge University Press, Cambride, 1995.
[22] M.E. Nebel, Combinatorial properties of RNA secondary structure, J. Comput. Biol. 9 (3) (2002) 541–573.
[23] M.E. Nebel, Investigation of the Bernoulli-model of RNA secondary structures, Bull. Math. Biol. 66 (2004) 925–964.
[24] R. Nussinov, A.B. Jacobson, Fast algorithm for predicting the secondary structure of single stranded RNA, Proc. Nat. Acad. Sci., U.S.A. 77 (11) (1980) 6309–6313.
[25] M. Paterson, V. Dancik, Longest common subsequences, in: MFCS'94, Mathematical Foundations of Computer Science, Springer Lecture Notes in Computer Science, vol. 841, Springer, Berlin, 1994, pp. 127–142.
[26] M. Sprinzl, C. Horn, M. Brown, A. Ioudovitch, S. Steinberg, Compilation of tRNA sequences and sequences of tRNA genes, Nucleic Acids Res. 26 (1998) 148–153.
[27] M. Sprinzl, K.S. Vassilenko, J. Emmerich, F. Bauer, tRNA Database. ⟨http://www.staff.uni-bayreuth.de/∼btc914/search/index.html⟩.
[28] J.M. Steele, Probability Theory and Combinatorial Optimization, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 69, SIAM, 1997.
[29] P.R. Stein, M.S. Waterman, On some new sequences generalizing the Catalan and Motzkin numbers, Discrete Math. 26 (1978) 261–272.
[30] X.G. Viennot, M.V. de Chaumont, Enumeration of RNA's secondary structures by complexity, in: V. Capasso, E. Grosso, S.L. Paveri-Fontana (Eds.), Mathematics in Medicine and Biology, Springer Lecture Notes in Biomathematics, vol. 57, 1985, pp. 360–365.
[31] M.S. Waterman, Introduction to Computational Biology, Chapman & Hall/CRC Press, London, Boca Raton, 1995.

[32] C. Workman, A. Krogh, No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution, Nuclic Acids Res. 27 (1999) 4816–4822.
[33] T. Xia Jr., J. SantaLucia, M.E. Burkard, R. Kierzek, S.J. Schroeder, X. Jiao, C. Cox, D.H. Turner, Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs, Biochemistry 37 (1999) 14719–14735.
[34] M. Zuker, P. Stiegler, Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information, Nucleic Acids Res. 9 (1981) 133–148.