Downloaded from rnajournal.cshlp.org on August 9, 2022 - Published by Cold Spring Harbor Laboratory Press

Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency

PETER CLOTE,[1] FABRIZIO FERRÉ,[1] EVANGELOS KRANAKIS,[2] and DANNY KRIZANC[3]
[1]Department of Biology, Boston College, Chestnut Hill, Massachusetts 02467, USA
[2]School of Computer Science, Carleton University, Ottawa, Ontario, K1S 5B6, Canada
[3]Department of Mathematics and Computer Science, Wesleyan University, Middletown, Connecticut 06459, USA

ABSTRACT

We present results of computer experiments that indicate that several RNAs for which the native state (minimum free energy secondary structure) is functionally important (type III hammerhead ribozymes, signal recognition particle RNAs, U2 small nucleolar spliceosomal RNAs, certain riboswitches, etc.) all have lower folding energy than random RNAs of the same length and dinucleotide frequency. Additionally, we find that whole mRNA as well as 5′-UTR, 3′-UTR, and cds regions of mRNA have folding energies comparable to that of random RNA, although there may be a statistically insignificant trace signal in 3′-UTR and cds regions. Various authors have used nucleotide (approximate) pattern matching and the computation of minimum free energy as filters to detect potential RNAs in ESTs and genomes. We introduce a new concept of the asymptotic *Z*-score and describe a fast, whole-genome scanning algorithm to compute asymptotic minimum free energy *Z*-scores of moving-window contents. Asymptotic *Z*-score computations offer another filter, to be used along with nucleotide pattern matching and minimum free energy computations, to detect potential functional RNAs in ESTs and genomic regions.

Keywords: tRNA; folding energy; RNA secondary structure; structural RNA; asymptotic *Z*-score

INTRODUCTION

In Le et al. (1990b), it was shown that RNA stem–loop structures situated 3′ to frameshift sites of retroviral gag–pol and pro–pol regions of several viruses (human immunodeficiency virus HIV-1, Rous sarcoma virus RSV, etc.) are thermodynamically stable and recognizable among positions 300 nt upstream and downstream of the frameshift site. Using Zuker's algorithm[4] (Zuker and Stiegler 1981; Mathews et al. 2000; Zuker 2003) to compute the minimum free energy (mfe) secondary structure for RNA, Le et al. (1990a) showed that certain RNAs have lower folding energy (i.e., minimum free energy of predicted secondary structure) than random RNA of the same mononucleotide (or compositional) frequency. This was measured by performing permutations (i.e., mononucleotide shuffles) of nucleotide positions, subsequently computing the *Z*-score[5] of the minimum free energy (mfe) of real versus random RNA—see Materials and Methods for details.

In Seffens and Digby (1999), it was shown that the folding energy of mRNA is lower than that of random RNA of the same mononucleotide frequency, as measured by the *Z*-score of the mfe secondary structure of mRNA versus mononucleotide shuffles of mRNA. In Rivas and Eddy (2000), a moving-window, whole-genome scanning algorithm was developed to compute *Z*-scores of windows of a genome with respect to mononucleotide shuffles of the window contents. By constructing artificial data with samples of real RNA (RNase-P RNA, T5 tRNA, soy bean SSU, etc.) planted in the center of a background sequence of random RNA of the same compositional frequency, Rivas and Eddy (2000; see their Figs. 4–11) found that the planted RNA had a low *Z*-score, as expected; however, other regions of the artificial data displayed low *Z*-scores as well, and by considering *p*-values for an assumed extreme value distribution, Rivas and Eddy subsequently argued that determining *Z*-scores of genomic window contents is statistically not

Reprint requests to: Peter Clote, Department of Biology, Higgins 416, Boston College, Chestnut Hill, MA 02467, USA; e-mail: clote@bc.edu; fax: (617) 552-2011.

Article and publication are at http://www.rnajournal.org/cgi/doi/10.1261/rna.7220505.

[4]Zuker's algorithm was first implemented in Zuker's mfold, subsequently in Hofacker et al.'s Vienna RNA Package RNAfold, and most recently in Mathews and Turner's RNAstructure.

[5]The *Z*-score of x (with respect to a histogram or probability distribution) is the number of standard deviation units to the left or right of the mean for the position where x lies, that is, $(x - \mu)/\sigma$.

 Copyright © 2005 RNA Society.

Downloaded from rnajournal.cshlp.org on August 9, 2022 - Published by Cold Spring Harbor Laboratory Press

reliable enough to allow one to construct an RNA gene finder on this basis.[6]

In Workman and Krogh (1999), it was noted that Zuker's algorithm (Zuker and Stiegler 1981) computes secondary structure minimum free energy (mfe) by adding contributions of negative (stabilizing) energy terms for stacked base pairs and positive (destabilizing) energy terms for hairpin loops, bulges, internal loops, and multiloops. In Zuker's algorithm, experimentally determined stacked base pair energies and loop energies for various lengths of hairpin, bulge, and internal loop are used, as determined by D. Turner's lab (see Mathews et al. 1999). The energy term contributed by a base pair depends on the base pair (if any) upon which it is stacked; for instance, Turner's current rules (Xia et al. 1998) at 37°C assign stacking free energy of −2.24 kcal/mol to

$$5' - AC - 3'$$
$$3' - UG - 5'$$

of −3.26 kcal/mol to

$$5' - CC - 3'$$
$$3' - GG - 5'$$

and of −2.08 kcal/mol to

$$5' - AG - 3'$$
$$3' - UC - 5'.$$

For this reason, Workman and Krogh (1999) argued that random RNA must be generated with the same dinucleotide frequency, for any valid conclusions to be drawn. Their experiments using mfold indicated that, in contrast to the results of Seffens and Digby (1999) mentioned above, mRNA does not have any statistically significant lower mfe than random RNA of the same dinucleotide frequency. This is consistent with the idea that mRNA exists in an ensemble of low-energy states, lacking any functional structure. Workman and Krogh additionally considered a small sample of five rRNAs and five tRNAs; for the latter, they stated, "Surprisingly, the tRNAs do not show a very clear difference between the native sequence and dinucleotide shuffled, and one of the native sequences even has a higher energy than the average of the shuffled ones" (Workman and Krogh 1999).

In this paper, we use Zuker's algorithm as implemented in version 1.5 of Vienna RNA Package RNAfold (http://www.tbi.univie.ac.at/~ivo/RNA/) to compute minimum free energy for RNA sequences, and analyze the following RNA classes: tRNA, hammerhead type III ribozymes, SECIS[7] elements, U1 and U2 small nuclear RNA (snRNA) components of the spliceosome, signal recognition particle RNA (srpRNA), seven classes of riboswitches (namely, Purine, Lysine, Cobalamin, THI element, S-box leader, RFN element, ykoK element), 5S ribosomal RNA, entire mRNA, as well as the 3′-UTR (untranslated region), 5′-UTR, and coding sequence (cds) of mRNA. Structural RNAs were chosen using information from the Rfam database (Griffiths-Jones et al. 2003) and the SCOR (Structural Classification of RNA) database (Klosterman et al. 2002). While Workman and Krogh (1999) use a heuristic to perform dinucleotide shuffle, their heuristic is not guaranteed to correctly sample random RNAs having a given number of dinucleotides, and thus we have implemented the provably correct procedure of Altschul and Erickson (1985). We provide both Python source code as well as a Web server for our implementation of the Altschul-Erikson algorithm[8] (see http://clavius.bc.edu/~clotelab/). The work of the present paper validates the conclusion of Workman and Krogh (1999) concerning mRNA. Concerning their conclusion about tRNA, by using the database of 530 tRNAs (Sprinzl et al. 1998), where we generated 1000 random RNAs for each tRNA considered,[9] we show that *Z*-scores for tRNA are low (~−1.5), although not as low as certain other classes of structural RNA (~−4), and that there is a statistically significant, although moderate signal in the *Z*-scores of tRNA from Sprinzl's database, with *p*-values of ~0.12. See the related work of Bonnet et al. (2004), who investigate *Z*-scores and *p*-values[10] of minimum free energy for precursor microRNAs.

Additionally, in this paper, we introduce the novel concept of the "asymptotic *Z*-score," and by proving an asymptotic limit for the mean and standard deviation of minimum free energy per nucleotide for random RNA, we indicate how to perform certain precomputations that entail an enormous speed-up when computing asymptotic *Z*-scores for whole-genome, sliding-window scanning algorithms. This method provides a filter, which may be used along with (approximate) pattern matching, minimum free energy computations, and other filters, when attempting to determine putative functional RNA genes in ESTs and genomic data.

Various researchers have used a combination of filters to

[6]Figures 12 and 13 of Rivas and Eddy (2000) are similar to some of the graphs presented in this paper; however, unlike our work, Rivas and Eddy (2000) use mononucleotide shuffles to produce random sequences. As previously observed in Workman and Krogh (1999) when computing *Z*-scores for minimum free energies of RNA, it is important to generate random sequences that preserve dinucleotide frequency of the given RNA. Our work presents a careful analysis of a large class of RNAs using the dinucleotide shuffling Algorithm 4.

[7]SECIS abbreviates "selenocysteine insertion sequence," a small (30–45 nt) portion of the 3′-UTR that forms a stem–loop structure necessary for the UGA stop codon to be retranslated to allow selenocysteine incorporation.

[8]After completion of this paper, we learned of the more general Web server *Shufflet* (Coward 1999).

[9]The work of Workman and Krogh (1999) focuses on mRNA, and only at the end of their article do they consider a small collection of five tRNAs, where 100 random RNAs are generated per tRNA.

[10]Bonnet et al. (2004) compute *p*-values of minimum free energy not not *p*-values of *Z*-scores as done in this paper.

Downloaded from rnajournal.cshlp.org on August 9, 2022 - Published by Cold Spring Harbor Laboratory Press

TABLE 1. *Z*-score statistics for structural RNA compared to random RNA of the same *expected* dinucleotide frequency using Algorithm 3

RNA type	Number of sequences	Mean	Stdev	Max	Min
tRNA	530	−1.348202	0.611164	0.269411	−3.124041
Hammerhead III	114	−2.053881	0.664340	−0.001203	−3.387384
SECIS	5	−3.800337	0.883944	−2.832499	−5.237905
srpRNA	94	−2.037159	1.030724	0.010698	−4.961649
U1	53	−1.083326	0.547852	0.012102	−2.508698
U2	62	−2.243978	0.599099	0.920614	−3.479369
mRNA whole length	41	0.090522	0.783253	1.667423	−1.711233
mRNA 3′-UTR	41	0.152680	0.646208	0.870732	−2.132468
mRNA 5′-UTR	41	0.183972	0.628083	0.893692	−1.940810
mRNA cds	41	−0.209889	0.681839	1.268412	−2.218905

determine potential RNAs of interest. Kryukov et al. (1999) developed the program SECISearch, which uses PATSCAN (Dsouza et al. 1997) to filter for approximate matching nucleotide sequences for SECIS elements (e.g., there is a required AA dinucleotide in an internal loop region of the secondary structure of the SECIS element, as well as certain other nucleotide constraints). Subsequently, SECISearch uses Vienna RNA Package RNAfold to compute free energies related to the SECIS secondary structure. Lescure et al. (1999) developed a filter using the tool RNAMOT (Gautheret et al. 1990; Laferriere et al. 1994) to find approximate pattern matches in human ESTs for known SECIS stem–loop structure with certain nucleotide constraints. After experimentally validating the SECIS elements found in Lescure et al. (1999), the secondary structure of valid SECIS elements was found by chemical probing in Fagegaltier et al. (2000).

In Lim et al. (2003), vertebrate microRNA (miRNA) genes were found by devising a computational procedure, MiRscan, to identify potential miRNA genes. MicroRNAs (Harborth et al. 2003; Tuschl 2003) are 21-nt RNA sequences that form a known stem–loop secondary structure, are (approximately) the reverse complement of a portion of transcribed mRNA, and prevent the translation of protein product. MiRscan (Lim et al. 2003) involves a moving-window scan of 110-nt regions of the genome, and by using the Vienna RNA Package (C. Burge, pers. comm.), determines stem–loop structures, then assigns a log-likelihood score to each window to determine how well its attributes resemble those of certain experimentally verified miRNAs of *Caenorhabditis elegans* and *Caenorhabditis briggsae* homologs.

Klein et al. (2002) scanned for GC-rich regions in the AT-rich genomes of *Methanococcus jannaschii* and *Pyrococcus furiosus* to determine noncoding RNA genes. Recently, Hofacker et al. (2004) developed a fast, whole-genome version of RNAfold, which determines the minimum free energy structure of RNA from whole genomes, in which base-paired indices i, j are required to be of at most a user-specified distance (e.g., 100 nt). See the additional relevant work of Eddy (2001, 2002), Macke et al. (2001), and Washietl and Hofacker (2004).

Although Rivas and Eddy (2000) argued that genome scanning computations of *Z*-scores, in which randomized window contents preserve mononucleotide frequency (Algorithm 2), are not statistically significant enough to be used as a base for a general ncRNA gene finder, it is nevertheless possible that *Z*-score computations, in which randomized window contents preserve dinucleotide frequency (Algorithms 3 or 4), may be used as one of several filters to determine RNA of interest. Such *Z*-score computations, especially for large window size, are enormously time-consuming. Owing to a precomputation phase, asymptotic *Z*-scores, introduced in this paper, may provide a computationally efficient filter to identify certain RNA. In all of our computational experiments, asymptotic *Z*-scores, when compared to (classical) *Z*-scores, have substantially higher signal-to-noise ratio,[11] although at present we have no understanding of why this is so.

RESULTS

As described in detail in Materials and Methods, we performed experiments on tRNA, SECIS elements, hammerhead type III ribozymes, and other structural RNAs, as well as whole mRNA and the cds, 5′-UTR and 3′-UTR regions of mRNA. For each RNA sequence s from a given class (e.g., tRNA), we compute the minimum free energy of s, as well as that of a large number of random RNA having the same *expected* (Algorithm 3) or the same *exact* (Algorithm 4) dinucleotide frequency as that of s. From these data, we compute the *Z*-score (number of standard deviation units to the right or left of the mean) for each RNA sequence, and produce histograms summarized in Tables 1 and 2 and related figures.

[11]Average *Z*-scores have value 0, while average asymptotic *Z*-scores are >0, making a greater contrast with negative scores of functional RNA in computational experiments.

Downloaded from rnajournal.cshlp.org on August 9, 2022 - Published by Cold Spring Harbor Laboratory Press

TABLE 2. *Z*-score and *p*-value statistics for structural RNA compared to random RNA of the *same* dinucleotide frequency using Algorithm 4

RNA type	Number of sequences	Mean	Stdev	Max	Min	*p*-value
tRNA	530	−1.591106	0.889903	0.732033	−4.034804	0.123123
Hammerhead III	114	−3.188341	0.870615	−1.202616	−5.34491	0.007526
SECIS	5	−4.736209	1.122621	−3.48201	−6.944927	0.0
srpRNA	94	−3.564441	2.139954	−0.099144	−9.254801	0.045528
U1	53	−1.750205	0.930827	0.156993	−4.041211	0.101509
U2	62	−4.224552	1.215934	−1.83139	−7.068373	0.002468
mRNA whole length	41	−0.180843	1.619402	2.90517	−4.207065	0.478049
mRNA 3′-UTR	41	−0.111613	1.021312	1.483879	−3.198117	0.526512
mRNA 5′-UTR	41	0.17506	1.092026	1.862059	−2.97943	0.459195
mRNA cds	41	−0.132962	1.646607	3.284421	−3.739057	0.514634
S-box leader (SAM)	70	−2.391071	1.039610	0.163915	−4.81384	0.047614
RFN element (FMN)	48	−1.430811	0.919783	0.225564	−3.460798	0.150170
THI element (TPP)	141	−0.819171	0.996411	1.825061	−3.646955	0.279965
Purine ribos.	37	−1.570689	1.305818	0.453687	−4.412738	0.171417
Lysine ribos.	48	−2.156423	1.575269	0.791328	−8.927446	0.104809
Cobalamin ribos.	82	−1.388310	1.378731	1.643258	−6.089820	0.205617
ykoK ribos.	40	−2.406571	1.391506	−0.095575	−6.506148	0.073231
5S rRNA	100	−1.537079	1.128463	0.59728	−5.788211	0.154

Tables 1 and 2 give details on the number of sequences, mean, standard deviation, maximum and minimum *Z*-scores[12] for each investigated class of RNA. For Table 1, we computed *Z*-scores with respect to random RNA of the same *expected* dinucleotide frequency, using Algorithm 3, while in Table 2 we computed *Z*-scores with respect to random RNA of the same (*exact*) dinucleotide frequency using the provably correct Altschul-Erikson Algorithm 4. Since we correct an assertion of Workman and Krogh (1999) concerning tRNA, we implemented their method of computing *p*-values and list in Table 2 the *p*-values for all investigated classes of RNA.

All classes of structurally important RNA, which we investigate (with the exception of the TPP riboswitch–THI element), show a significantly lower folding energy than random RNAs of the same dinucleotide frequency, using both Algorithms 3 and 4. In contrast, for entire mRNA, as well as in 5′-UTR, 3′-UTR, and cds of mRNA, the folding energy is approximately that of random RNA of the same (both expected and exact) dinucleotide frequency. Figures 1 and 2 present histograms of *Z*-score data for all RNA classes, where *Z*-scores were computed with respect to random RNA of the same expected dinucleotide frequency as generated by Algorithm 3. Figures 3 and 5 present similar histograms, differing only in that *Z*-scores were computed with respect to random RNA as computed by Algorithm 3 in the former and by Algorithm 4 in the latter. Figure 4 presents histograms of *Z*-score data for the seven classes of riboswitches that are found in the current Rfam release, using Algorithm 4. As an additional test of our assertion that structural RNA[13] has lower folding energy than random RNA of the same dinucleotide frequency (as generated by Algorithm 4), Figure 6 graphs *p*-scores against *Z*-scores for nonstructural RNA, while Figures 7 and 8 graph *p*-scores against *Z*-scores for structural RNAs and for the seven classes of riboswitches in the current release of Rfam, respectively. Note that Figure 6 is similar to Figure 2 of Workman and Krogh (1999), although we additionally compute separate *Z*-scores for 5′-UTR, 3′-UTR, and cds regions of mRNA as well as whole mRNA, and we use the Altschul-Erikson algorithm to generate random RNA. Figures 7 and 8 furnish additional evidence that tRNA and other structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. A Web server and Python source code for our implementation of this algorithm is available at the previously given Clote Lab Web site. We are currently computing *Z*-scores and *p*-values for all of Rfam. When completed, results will be summarized on this Web site.

In the Results section (explained in more detail in Materials and Methods), we introduce the new concept of "asymptotic *Z*-score" and state a new theorem, whose proof is given in the Appendix. This theorem postulates that for every complete set of dinucleotide frequencies $\vec{q}_{xy}$, there exist values $\mu(\vec{q}_{xy})$ (asymptotic mean minimum free energy per nucleotide) and $\sigma(\vec{q}_{xy})$ (asymptotic standard deviation of minimum free energy per nucleotide), with the following properties. If $x_0, x_1, x_2, \ldots$ is a sequence of random variables generated by a first-order Markov process from the dinucleotide frequencies $\vec{q}_{xy}$, then the limits

[12] *Z*-score is often used as a statistical measure of deviation from the mean in units of standard deviation. See Materials and Methods for formal definition.

[13] By structural RNA, we mean naturally occurring classes of RNA whose functionality depends on the native state, where we identify the native state with the minimum free energy secondary structure if the structure is not experimentally determined.

Downloaded from rnajournal.cshlp.org on August 9, 2022 - Published by Cold Spring Harbor Laboratory Press

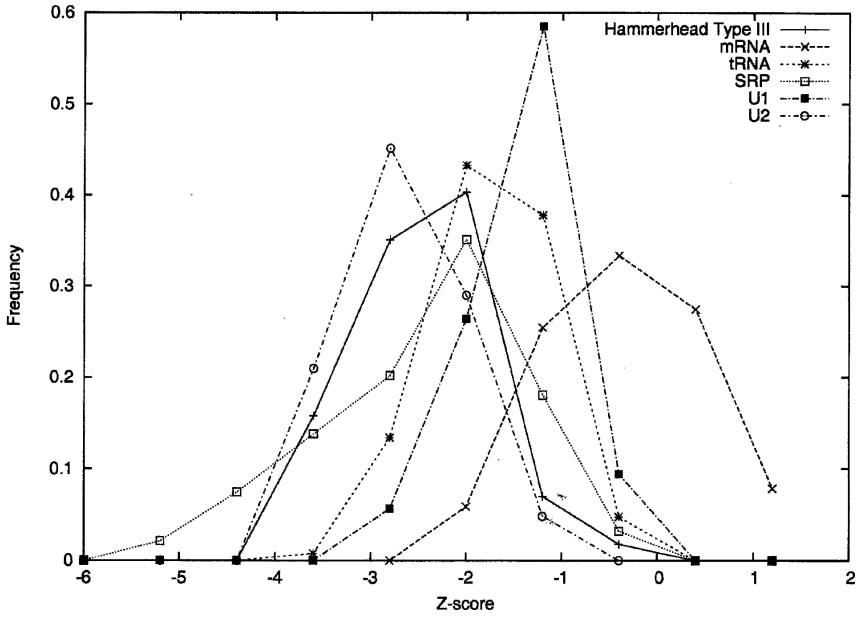

FIGURE 1. Histograms of *Z*-scores of minimum free energy (mfe) of RNA classes versus 1000 random RNAs of the same expected dinucleotide frequency using Algorithm 3. The curves, in *left* to *right* order, correspond to signal recognition particle (srp) RNA, U2 small nucleolar particle, Hammerhead type III ribozyme, 530 tRNAs from Sprinzl's database, U1 small nucleolar particle, and the 41 whole length mRNAs considered in Workman and Krogh (1999). Structurally important RNAs have *Z*-score curves shifted toward negative values with respect to the curve for mRNA.

$$\lim_{n\to\infty} \frac{E[\mathrm{mfe}(x_0, \ldots, x_n)]}{n} = \mu(\vec{q}_{xy})$$

and

$$\lim_{n\to\infty} \sqrt{\frac{E[(\mathrm{mfe}(x_0, \ldots, x_n))^2] - E[\mathrm{mfe}(x_0, \ldots, x_n)]}{n^2}} = \sigma(\vec{q}_{xy})$$

both exist and depend only on $\vec{q}_{xy}$.

We can now pre-compute a table of values $\mu(\vec{q}_{xy})$ and $\sigma(\vec{q}_{xy})$ for all complete sets $\vec{q}_{xy}$ of dinucleotide frequencies, where dinucleotide frequencies are specified up to (say) two decimal places. Given RNA nucleotide sequence $a_1, \ldots, a_n$, compute the dinucleotide frequencies $\vec{q}_{xy}$ of $a_1, \ldots, a_n$. The asymptotic minimum free energy *Z*-score, defined by

$$\frac{\mathrm{mfe}(a_1, \ldots, a_n)/n - \mu(\vec{q}_{xy})}{\sigma(\vec{q}_{xy})},$$

can be computed by one application of Zuker's algorithm with input $a_1, \ldots, a_n$, together with table look-up of the pre-computed (approximations) of $\mu(\vec{q}_{xy})$, $\sigma(\vec{q}_{xy})$. Figure 9 displays both *Z*-scores and asymptotic *Z*-scores for all windows of size 32 in the artificial genome constructed by planting RNA SECIS element fruA in the middle of random RNA of the same expected mononucleotide frequency. In this figure, *Z*-scores were computed using the Altschul-Erikson dinucleotide shuffle, Algorithm 4, and asymptotic *Z*-scores were computed by Algorithm 7. Note that although we are unsure why this is the case, there is a greatly improved signal-to-noise ratio in using asymptotic *Z*-scores compared to *Z*-scores.

DISCUSSION

In Seffens and Digby (1999), it was observed that mRNA has lower folding energy than random RNA of the same mononucleotide frequency, which latter is obtained by permuting nucleotide positions. Later, Workman and Krogh (1999) made an important observation that preserving dinucleotide frequency is critical, because of the nature of base-stacking free energies, and that mRNA cannot be distinguished from random RNA of the same dinucleotide frequency with respect to folding energy. Workman and Krogh additionally asserted that it appeared, according to their limited data set of five tRNAs, that the same was true of tRNA.

Our computation of both *Z*-scores and *p*-scores on the much larger data set of 530 tRNAs from the tRNA database of M. Sprinzl, K.S. Vassilenko, J. Emmerich, and F. Bauer, at URL http://www.staff.uni-bayreuth.de/~btc914/search/, indicates that tRNAs from Sprinzl's database have lower *Z*-scores than random RNA of the same dinucleotide frequency, although the *p*-value is only around 0.12. More generally, by considering tRNAs, type III hammerhead ribozymes, SECIS sequences, srpRNAs, snRNAs, and so on, we show that many important classes of structural RNA have lower folding energy than random RNA of the same dinucleotide frequency. Our careful tabulation of *Z*-scores may prove useful in future work involving a moving-window, genome-scanning algorithm, where one might attempt to detect particular structural RNA by looking at regions whose *Z*-score is close to that listed in Table 2.

It is known that tRNA has certain modified nucleotides; for example, aspartyl tRNA from *Saccharomyces cerevisiae* with PDB identity number 1ASY includes two dihydrouridines, three pseudouridines, one 5-methylcytidine, and one 1-methylguanosine. For this paper, we replaced all modified nucleotides
as annotated in Sprinzl's database by unmodified nucleotides (e.g., dihydrouridine is replaced by uridine) and subsequently applied RNAfold to the resulting tRNA sequences. It seems likely that computed energies of tRNA might differ from their experimentally determined energies, and that such a discrepancy would similarly influence pre-

Downloaded from rnajournal.cshlp.org on August 9, 2022 - Published by Cold Spring Harbor Laboratory Press

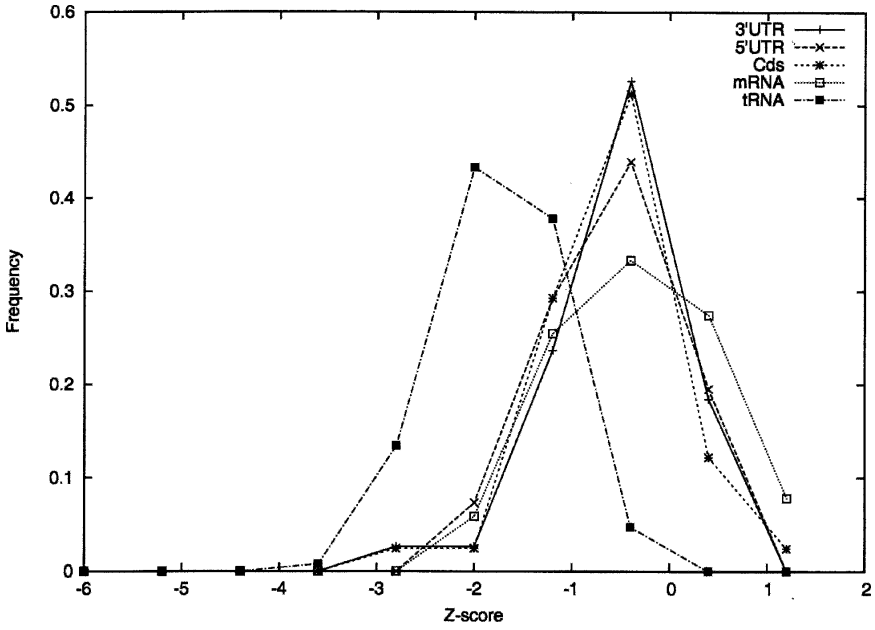

FIGURE 2. Histograms of *Z*-scores of minimum free energy (mfe) of RNA classes versus 1000 random RNAs of the same expected dinucleotide frequency using Algorithm 3. The curves, in *left* to *right* order, correspond to 530 tRNAs from Sprinzl's database, and to coding sequence (cds), 3′-untranslated region (UTR), 5′-UTR, and whole length mRNA of the 41 mRNAs considered in Workman and Krogh (1999). Different regions of the mRNAs show similar curves, centered around the 0.

dicted energies of randomizations of tRNA. This might explain the relatively high *Z*-scores and *p*-values of tRNA, when compared to other structural RNA classes.

While Workman and Krogh (1999) had considered whole mRNA, we additionally considered 5′-UTR, 3′-UTR, and cds of the same mRNA analyzed in those investigated by Workman and Krogh. Tables 1 and 2 provide evidence that these mRNA subclasses do not have lower folding energy than random RNA of the same dinucleotide frequency, although it should be noted that Table 2 shows negative *Z*-scores of -0.111613 (respectively, −0.132962) for 3′-UTR (respectively, cds) of mRNA, suggesting a slightly discernable signal in both the 3′-UTR and cds of mRNA (for a recent review, see Wilkie et al. 2003). A possible explanation for the statistically insignificant signal in the 3′-UTR, which contains regulatory elements, is that these structural, regulatory elements are short and dispersed in the UTR, which in many cases may be very long.

Moreover, we present evidence that riboswitches (metabolites binding domains that are found within certain messenger RNAs) have lower folding energy than random RNA of the same dinucleotide frequency similarly to structural RNAs, with the only exception of the THI element (TPP riboswitch). The TPP riboswitch is found in the 5′ region of mRNAs of genes involved in thiamine biosynthesis and transport (Miranda-Rios et al. 2001), and is able to bind thiamine and its pyrophosphate derivatives (Winkler et al. 2002), resulting in the reduction of translation. The interaction with thiamine is thought to be dependent on the secondary structure assumed by this riboswitch; therefore, the *Z*-score close to zero of this class of riboswitches is unexpected, and we do not have any valid argument to justify this observation.

Figures 1–5 present superposed histograms of *Z*-scores for the RNAs analyzed. The general trend is a shift toward negative values in the curves associated with structural RNAs; *Z*-score curves obtained using both Algorithms 3 and 4 are quite similar, although the small discrepancy between algorithms in the case of 3′-UTR regions of mRNA suggests that one should prefer the use of Algorithm 4, if possible.

The work of Seffens and Digby (1999) and of Workman and Krogh (1999) together provide strong evidence that the mononucleotide shuffle, Algorithm 2, and the 0-order Markov chain, Algorithm 1, should never be used when computing *Z*-scores. The slight discrepancy between Tables 1 and 2 for 3′-UTR regions of mRNA suggests that Algorithm 4 should be used if possible over Algorithm 3, when computing *Z*-scores.

Additionally, based on new mathematical results concerning asymptotic comportment of random RNA (see the Appendix), we define the concept of "asymptotic *Z*-score" (see Definition 6 in Materials and Methods) and show how to radically reduce the computation time for moving-window, whole-genome algorithms that compute *Z*-scores of window contents. Rather than computing *Z*-scores *on the fly* for each window's randomized contents, we use table look-up for precomputed asymptotic *Z*-scores and call Zuker's algorithm only once, rather than tens or hundreds of times, per window. Our approach, combined with the $O(NL^2)$ genome-scanning version[14] of Vienna RNA Package RNAfold (Hofacker et al. 2004), permits $O(NL^2)$ genome-scanning asymptotic *Z*-score computations of whole genomes.[15]

Asymptotic *Z*-scores are computed with respect to large random RNA sequences (in the current paper, we used sequences of length 1000 nt) of the same expected dinucleo-

[14]For a genome of length N, successive applications of Zuker's algorithm to window contents of size L require time $O(NL^3)$. By re-using partial computations from previous window contents, Hofacker et al. (2004) describe an improvement to $O(NL^2)$.

[15]In this paper, we present a proof of concept. In work in progress, we are computing dinucleotide frequencies, within two decimal places, of viral and bacterial genomes and are computing tables necessary for a general application of our method, to be reported elsewhere.

Downloaded from rnajournal.cshlp.org on August 9, 2022 - Published by Cold Spring Harbor Laboratory Press

tide frequency as that of window contents using Algorithm 3, unlike computations of *Z*-scores in Le et al. (1990a), Seffens and Digby (1999), and Rivas and Eddy (2000), which used random RNA sequences of the same size as that of the moving window, generated by Algorithm 2. Although we have no explanation at the present, in all cases we have observed a greater signal-to-noise ratio in using asymptotic *Z*-scores to detect RNA genes (data not shown). This is, indeed, the case for Figure 9, which plots *Z*-scores and asymptotic *Z*-scores for 32-nt windows of artificial data obtained by planting SECIS element fruA CCUCGAGGGGAACCCGAAAGGGAC CCGAGAGG in the middle of random RNA of compositional frequency $A = 0.28125$, $C = 0.28125$, $G = 0.40625$, and $U = 0.03125$ (i.e., of same compositional frequency as that of fruA). Our preliminary work on the asymptotic *Z*-score raises the hope of effectively using this approach along with other heuristic filters to detect RNA of interest.

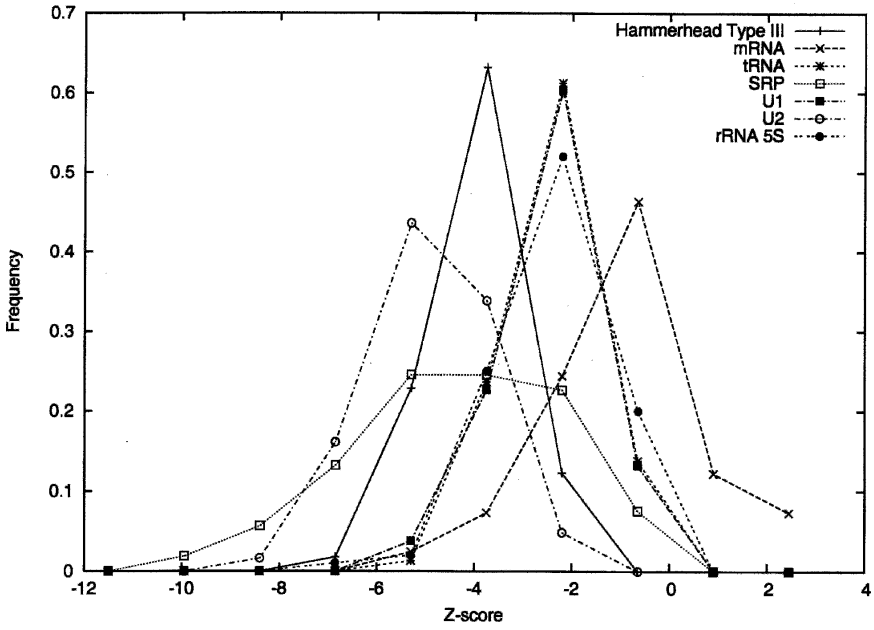

FIGURE 3. Histograms of *Z*-scores of minimum free energy (mfe) of RNA classes versus 1000 random RNAs of the same expected dinucleotide frequency using Algorithm 4. The curves, in *left* to *right* order, correspond to U2 small nucleolar particle, signal recognition particle (srp) RNA, Hammerhead type III ribozyme, U1 small nucleolar particle, tRNAs from Sprinzl's database, 100 5S rRNAs sampled from the Rfam seed alignment, and the 41 whole length mRNAs considered in Workman and Krogh (1999). As in Figure 1, structurally important RNAs have *Z*-score curves shifted toward negative values with respect to the curve of mRNA.

MATERIALS AND METHODS

For expository reasons, in this section, we describe the computer experiments we performed for tRNA. Additional experiments on mRNA, SECIS elements, hammerhead type III ribozymes, and so on were set up identically. Unless otherwise stated, we generated 1000 random RNAs per (real) RNA sequence, for each experiment. Using the mono- and dinucleotide frequencies for tRNA from Table 1, we generated random RNAs for each of the 530 tRNAs in the database of Sprinzl et al. (1998) according to two methods, which we respectively dub First-order Markov (Algorithm 3) and Dinucleotide Shuffle (Algorithm 4), and computed the mfe using RNAfold. The method First-order Markov generates random RNAs as a first-order Markov chain, and was considered in Workman and Krogh (1999), although it is unclear whether they generated the first nucleotide using sampling (as we do), or using the uniform probability of A, C, G, and U.

Algorithm 1 (sampling from 0-order Markov chain)

INPUT: *An RNA sequence* $a = a_1, \ldots, a_n$

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 $F_1(a)$ of $a = a_1, \ldots, a_n$; thus, $F_1(a)[A] = q_A$, $F_1(a)[C] = q_C$, $F_1(a)[G] = q_G$, $F_1(a)[U] = q_U$.
2. for i = 1 to n
   x = random in (0,1)
   if $x < q_A$ return 'A'
   else if $x < q_A + q_C$ return 'C'
   else if $x < q_A + q_C + q_G$ return 'G'
   else return 'U'

In their computation of *Z*-scores, Rivas and Eddy (2000) considered the following mononucleotide shuffle.

Algorithm 2 (Mononucleotide 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. generate a random permutation $\sigma \in S_n$
   for i = 1 to n
   $x_i = a_{\sigma(i)}$

Recall that Seffens and Digby (1999) observed negative *Z*-scores having large absolute value, when computing *Z*-scores of mRNA using Algorithm 2, while Workman and Krogh (1999) computed *Z*-scores approximately equal to 0 when computing *Z*-scores of mRNA using Algorithm 3.

Downloaded from rnajournal.cshlp.org on August 9, 2022 - Published by Cold Spring Harbor Laboratory Press

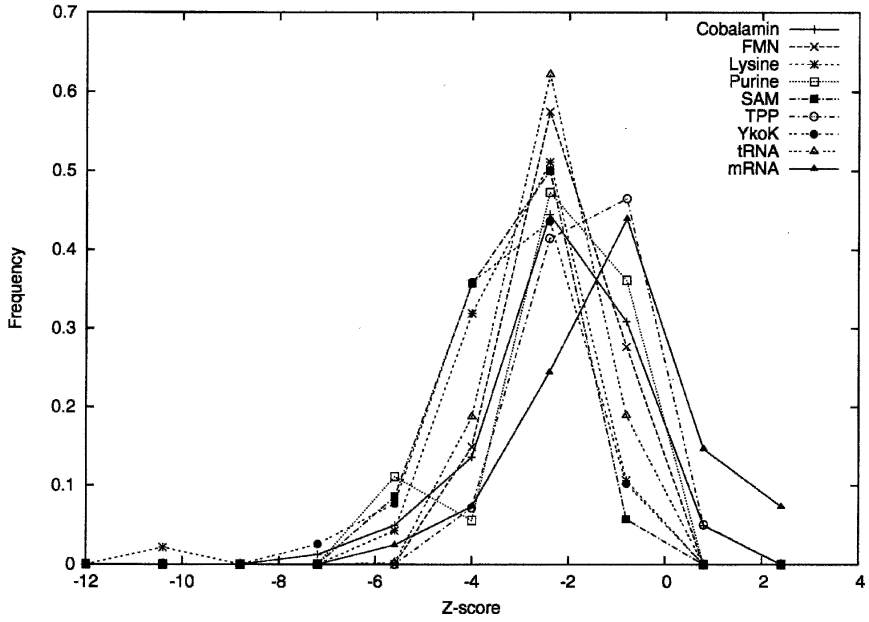

FIGURE 4. Histograms of *Z*-scores of minimum free energy (mfe) of RNA riboswitch classes versus 1000 random RNAs of the same expected dinucleotide frequency using Algorithm 4. The curves, in *left* to *right* order, correspond to lysine riboswitch, ykoK element, cobalamin riboswitch, S-box leader (SAM riboswitch), RFN element (FMN riboswitch), tRNAs from Sprinzl's database, purine riboswitch, THI element (TPP riboswitch), and whole length mRNAs considered in Workman and Krogh (1999).

Algorithm 3 (sampling from first-order Markov chain)

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 $Pr[Y \mid X]$, where $Pr[Y \mid X]$ equals the dinucleotide frequency that nucleotide Y follows X divided by the mononucleotide frequency of nucleotide X.

Algorithm 4 (Dinucleotide Shuffle)

Altschul and Erickson (1985).

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 $x \rightarrow y$ 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 $x \rightarrow y$ is an edge in E. If there is a vertex of G that 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 that had been removed.
5. For $i = 1$ to $n - 1$, generate x_{i+1} by taking the next available nucleotide such that $x_i \rightarrow x_{i+1}$ 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 Altschul and Erickson (1985) for details of Algorithm 4 and its extensions.

Before describing our experiments, we need to recall that the *Z*-score of a number x with respect to a sequence $s_1, \ldots, s_N$ of numbers is defined by $(x - \mu)/\sigma$, where μ, respectively σ, is the average respective standard deviation of $s_1, \ldots, s_N$. In (Workman and Krogh 1999), *p*-values associated with *Z*-scores are computed as the ratio N/D, where the numerator N is the number of *Z*-scores of random RNAs that exceed the *Z*-score of a fixed mRNA, and D is the number of *Z*-scores considered (see Workman and Krogh 1999 for details and an explicit graph of *Z*-scores vs. *p*-values for mRNA). Following the method of Workman and Krogh, we compute *p*-values and plot *Z*-scores and associated *p*-values for all classes of RNA investigated, where random RNA sequences were obtained by the Altschul-Erikson method.

We now describe our experiments. Lengths in Sprinzl's collection (Sprinzl et al. 1998) of 530 tRNAs range from 54 to 95. For each tRNA, we generated 1000 random RNAs of the same expected dinucleotide frequency (using Algorithm 3) and 1000 random RNAs of the same dinucleotide frequency (using Algorithm 4). For each tRNA, we computed the *Z*-score of its minimum free energy (mfe) using version 1.5 of Vienna RNA Package RNAfold with respect to the mfe of the corresponding 1000 random RNAs, separately using Algorithm 3 and Algorithm 4 to generate the random sequences. We followed the same procedure for each class of RNA we investigated: 530 tRNAs from Sprinzl's database, five SECIS elements from A. Böck of Ludwig-Maximilians-Universität München (pers. comm.), 114 hammerhead type III ribozymes, 53 U1 and 62 U2 small nucleolar spliceosomal RNAs, 94 signal recognition particle RNAs (srpRNAs), seven classes of riboswitches (namely, 70 S-box leaders, 48 RFN elements, 141 THI elements, 37 purine riboswitches, 48 lysine riboswitches, 82 cobalamin riboswitches, 40 ykoK elements), and 100 5S rRNAs. The hammerhead ribozymes, U1, U2, srpRNAs, and riboswitches sequences were taken from their respective Rfam seed alignment (Griffiths-Jones et al. 2003). The 100 5S rRNAs were sampled randomly from the very large Rfam seed alignment. Moreover, we considered the

Downloaded from rnajournal.cshlp.org on August 9, 2022 - Published by Cold Spring Harbor Laboratory Press

same mRNAs previously considered by Seffens and Digby (1999) and Workman and Krogh (1999); here, owing to the sequence length of mRNAs, we generated only 10 random RNAs per mRNA. Seffens and Digby considered 51 mRNAs.

Workman and Krogh considered a subset of 46 mRNAs, previously investigated in Seffens and Digby (1999), and explained their reasons for not including five spurious mRNAs considered by Seffens and Digby. We were not able to find five of these mRNAs in the latest GenBank release (namely, HUMIFNAB, HUMIFNAC, HUMIFNAH, SOYCHPI, XELSRBP); therefore, we included in the analysis 41 mRNAs, for which we considered the whole-length mRNA, and separately the untranslated regions (3′-UTR and 5′-UTR) and the coding sequence (cds) alone.

We now describe a new concept of "asymptotic *Z*-score," motivated by a new theorem concerning an asymptotic limit result for the mean and standard deviation of minimum free energy per nucleotide for random RNA. This result, formalized in Theorem 5, is proved in detail in the Appendix.

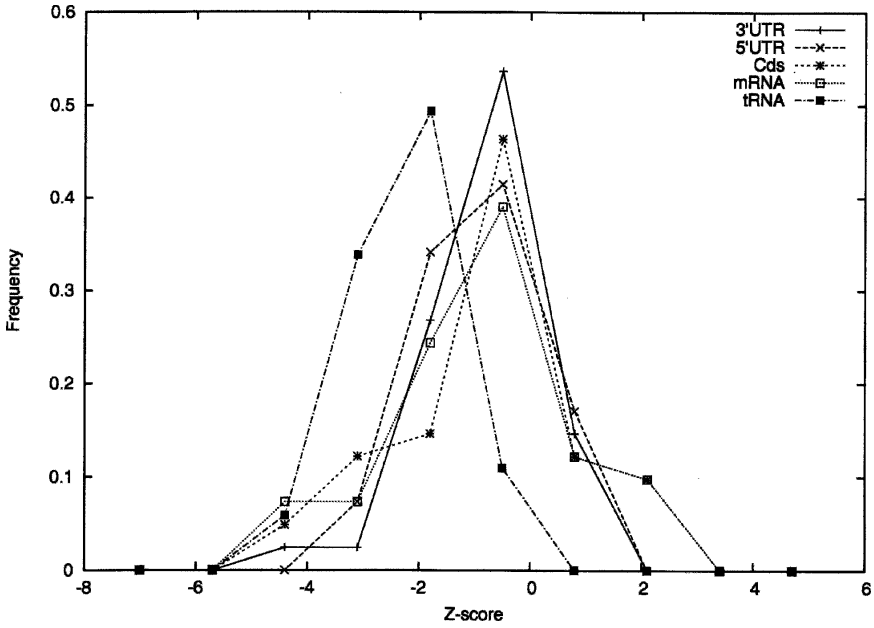

FIGURE 5. Histograms of *Z*-scores of minimum free energy (mfe) of RNA classes versus 1000 random RNAs of the same expected dinucleotide frequency using Algorithm 4. The curves, in *left* to *right* order, correspond to 530 tRNAs from Sprinzl's database, whole length mRNAs considered in Workman and Krogh (1999), coding sequences (cds), 3′-untranslated region (UTR), and 5′-UTR.

Let $F_2 = \{q_{xy} : x, y \in \{A, C, G, U\}\}$ be any complete set of dinucleotide frequencies; that is, $0 \le q_{xy} \le 1$ for all $x, y \in \{A, C, G, U\}$ and $\sum_{x,y} q_{xy} = 1$, where the sum is taken over all $x, y \in \{A, C, G, U\}$. Define $F_1 = \{q_x : x \in \{A, C, G, U\}\}$ to be the corresponding set of mononucleotide frequencies; that is, $q_x = \sum_u q_{ux}$, where the sum ranges over $u \in \{A, C, G, U\}$. We may at times say that the mononucleotide distribution F_1 is induced by the complete dinucleotide distribution F_2; moreover, we may use the notation $\vec{q}_{xy}$ to abbreviate F_2, and $\vec{q}_x$ to abbreviate F_1.

Theorem 5

Let $\vec{q}_{xy}$ be a complete set of dinucleotide frequencies, let $\vec{q}_x$ be the induced set of mononucleotide frequencies, and let X denote the infinite sequence of random variables $x_0, x_1, x_2, \ldots$ such that x_0 has the distribution $\vec{q}_x$, and for all i, x_{i+1} has the distribution given by the conditional probabilities

$$Pr[x_{i+1} = x] = \frac{q_{u,x}}{Pr[x_i = u]}.$$

For all $0 \le s \le t$, define random variables $X_{s,t} = \text{mfe}(x_s, \ldots, x_{t-1})$, where mfe denotes minimum free energy as measured by Zuker's algorithm. Then the limits

$$\lim_{n\to\infty} \frac{E[\text{mfe}(x_0, \ldots, x_n)]}{n} = \frac{E[X_{0,n}]}{n} = \mu(\vec{q}_{xy})$$

and

$$\lim_{n\to\infty} \sqrt{\frac{E[X_{0,n}^2] - (E[X_{0,n}])^2}{n^2}} = \sigma(\vec{q}_{xy})$$

both exist and depend only on $\vec{q}_{xy}$.

Although the proof gives no information on the rate of convergence, convergence appears to be fast (data not shown), and hence

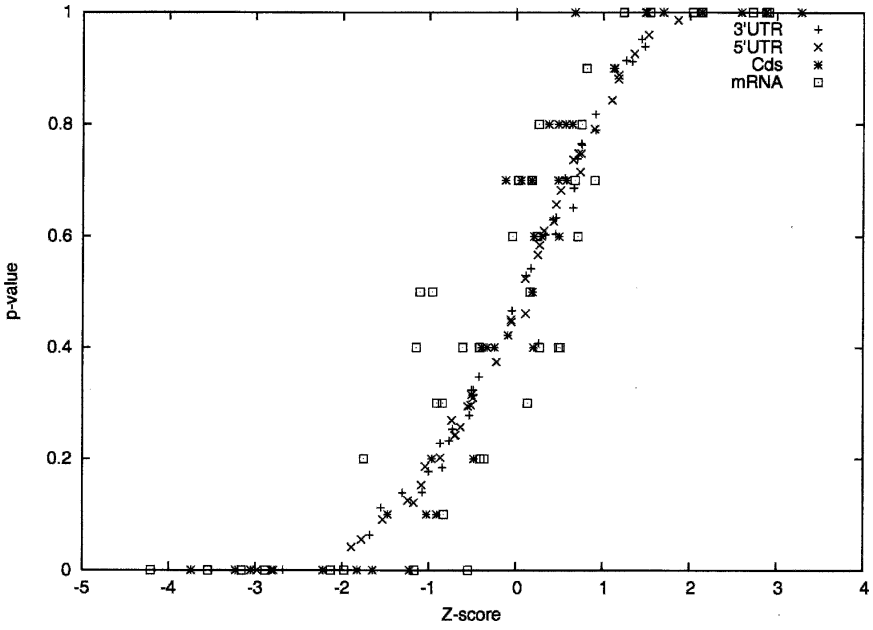

FIGURE 6. *Z*-score and *p*-value correlation for nonstructural RNAs.

Downloaded from rnajournal.cshlp.org on August 9, 2022 - Published by Cold Spring Harbor Laboratory Press

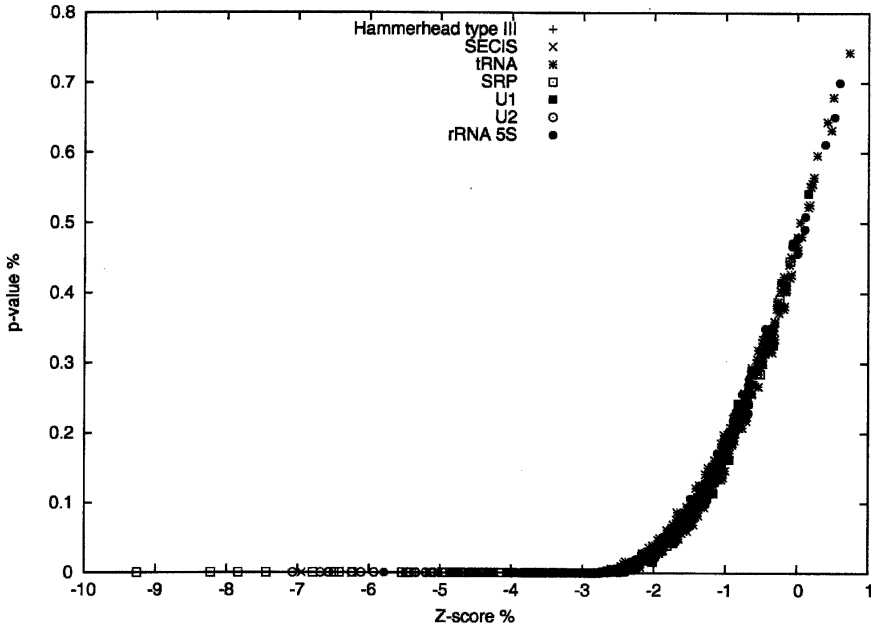

FIGURE 7. *Z*-score and *p*-value correlation for structural RNAs.

we can compute an approximation for the asymptotic mean, denoted by $\mu(\vec{q}_{xy})$, [respectively standard deviation, denoted by $\sigma(\vec{q}_{xy})$] per nucleotide of the minimum free energy of random RNA generated by a first-order Markov chain from dinucleotide frequencies $\vec{q}_{xy}$.

1. Compute minimum free energies for m random RNAs, each of length n nucleotides, as generated by Algorithm 3. In Figure 9, we used $m = 50$ and $n = 1000$.
2. Compute the mean and (sample) standard deviation for this collection, and divide both values by n so as to normalize these values with respect to sequence length.

Since m, n must be fixed for this computation, we denote the approximate mean by $\mu(\vec{q}_{xy}, m, n)$, and the approximate standard deviation by $\sigma(\vec{q}_{xy}, m, n)$. Thus, if $s_1, \ldots, s_m$ is a collection of m random RNA sequences, each s_1 has length n and is generated by Algorithm 3 from dinucleotide frequencies $\vec{q}_{xy}$, then

$$\mu(\vec{q}_{xy}, m, n) = \frac{\sum_{k=1}^{m} \mathrm{mfe}(s_i)/m}{n}$$

$$\sigma(\vec{q}_{xy}, m, n) = \frac{\sqrt{\dfrac{\sum_{k=1}^{m} \mathrm{mfe}^2(s_i)}{m-1} - \left(\dfrac{\sum_{k=1}^{m} \mathrm{mfe}(s_i)}{m}\right)^2 \cdot \dfrac{m}{m-1}}}{n}.$$

We now define as follows the "asymptotic, normalized mfe *Z*-score," with respect to random RNA of dinucleotide frequencies q_{xy}: Given RNA sequence s of length n_0 (generally n_0 is much less than n), compute the dinucleotide frequencies q_{xy} of s, and define

$$Z^2_{m,n}(s) = \frac{\mathrm{mfe}(s)/n_0 - \mu(\vec{q}_{xy}, m, n)}{\sigma(\vec{q}_{xy}, m, n)}.$$

Notice that when $n_0 = n$, we obtain the usual definition of *Z*-score, where randomization is performed with Algorithm 3.

As noted above, one should respect dinucleotide frequencies when performing *Z*-score computations. Taking this into account, we now define the "asymptotic, normalized mfe *Z*-score," with respect to random RNA of dinucleotide frequency q_{xy}, as follows.

Definition 6

Given RNA sequence s *of length* n_0 *(generally* $n_0 \ll n$*), compute the dinucleotide frequencies* q_{xy} *of* s. *Define*

$$Z^2_{m,n}(s) = \frac{\mathrm{mfe}(s)/n_0 - \mu(q_{xy}, m, n)}{\sigma(q_{xy}, m, n)}$$

This concludes the description of asymptotic *Z*-scores. Figure 9 illustrates the approach on small artificial data involving the SECIS element fruA. In future work, we plan to make available pre-computed tables of $\mu(q_{xy}, m, n)$, $\sigma(q_{xy}, m, n)$ for $n = 1000$, $m = 50$ over a range of dinucleotide frequencies found in windows of viral and bacterial genomes. Although not yet available, we can now describe an algorithm to efficiently compute asymptotic *Z*-scores in a moving-window scanning algorithm on a whole genome.

Algorithm 7

INPUT: *An entire genome* $g_1, \ldots, g_N$, *and window size* n_0

OUTPUT: *Values* (i, z_i), *where* $1 \le i \le N - n_0 + 1$ *is the starting position for the* i*-th window, and* z_i *is the asymptotic Z-score of the (reverse complement) of the* i*-th window*

for i = 1 to $N - n_0 + 1$
  s = reverse complement of $g_i, \ldots, g_{i+n_0-1}$
    compute mfe (s)
    compute dinucleotide frequencies q_{xy} of s
    for $x,y \in \{A, C, G, U\}$
  $q_{xy} = int(100 * q_{xy})/100$
    find $\mu(q_{xy}, m, n)$, $\sigma(q_{xy}, m, n)$ by table look–up
return $z_i = \dfrac{\mathrm{mfe}(s)/n_0 - \mu(q_{xy}, m, n)}{\sigma(q_{xy}, m, n)}$

Note that the instruction $q_{xy} = \mathrm{int}(100 * q_{xy})/100$ truncates each dinucleotide frequency q_{xy} to two decimal places. By using arrays with indirect addressing, table look-up does not require linear or logarithmic time, but rather unit time. Since Zuker's algorithm is applied only once, for each window, the run time of Algorithm 7 is $O(Nn_0^3)$. By using the genome-scan version of RNAfold (see Hofacker et al. 2004), we can reduce the run time of Algorithm 7 to $O(Nn_0^2)$.

ACKNOWLEDGMENTS

We thank the anonymous referees and Alice Tommasi di Vignano (Harvard Medical School) for helpful suggestions concerning this

Downloaded from rnajournal.cshlp.org on August 9, 2022 - Published by Cold Spring Harbor Laboratory Press

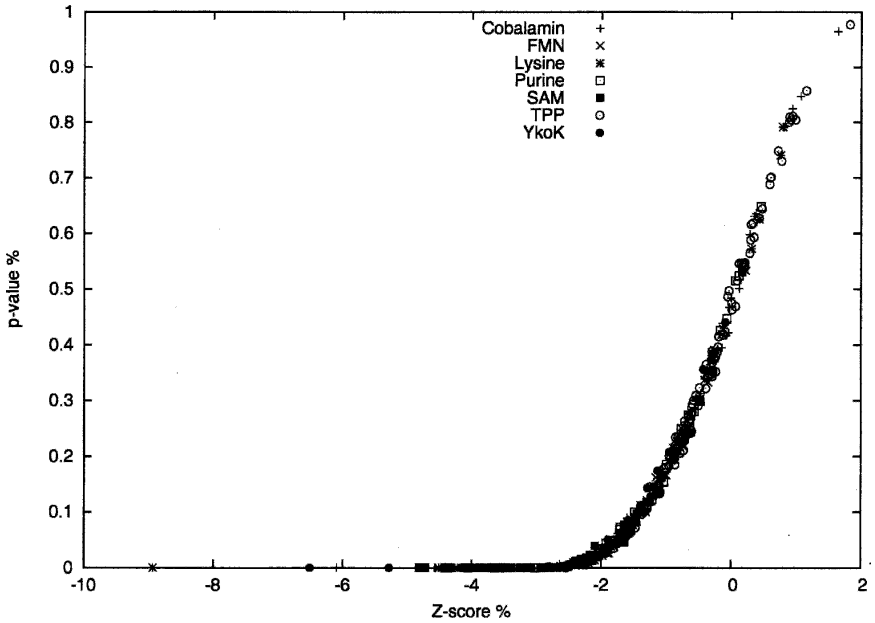

FIGURE 8. *Z*-score and *p*-value correlation for riboswitches.

Kingman's theorem has applications ranging from Ulam's problem concerning the asymptotic expected length of the longest increasing sequence [i.e., $1 \leq i_1 < i_2 < \ldots < i_k \leq n$ such that $\sigma(i_1) < \sigma(i_2 < \ldots < \sigma(i_k)]$ in a random permutation $\sigma \in S_n$ (Kingman 1973), to problems concerning restriction enzyme coverage (Waterman 1995). While Kingman's theorem proves the existence of an asymptotic limit λ, it can be a very difficult open problem to determine the precise value of λ for concrete cases.

Let $\vec{q}_{xy}$ denote any *complete* set $\{q_{xy} : x, y \in \{A, C, G, U\}\}$ of dinucleotide frequencies; that is, $0 \leq q_{xy} \leq 1$ for all $x, y \in \{A, C, G, U\}$ and $\sum_{x,\ y} q_{xy} = 1$, where the sum is taken over all $x, y \in \{A, C, G, U\}$. Define $\vec{q}_x$ to denote the set $\{q_x : x \in \{A, C, G, U\}\}$ of *induced* mononucleotide frequencies; that is, $q_x = \sum_u q_{ux}$, where the sum ranges over $u \in \{A, C, G, U\}$. We say that the mononucleotide distribution q_{x^-} is *induced* from the complete dinucleotide distribution $\vec{q}_{xy}$.

work. Research was supported in part by NSERC (Natural Sciences and Engineering Research Council of Canada) and MITACS (Mathematics of Information Technology and Complex Systems) grants.

Received October 28, 2004; accepted February 12, 2005.

APPENDIX

In this section, we state and prove Theorem 5, which provides the mathematical justification for our algorithm to compute (approximate) asymptotic *Z*-scores. The following theorem, due to Kingman (1973), provides the existence of a limit for certain types of subadditive stochastic processes.

Theorem 8 (Kingman 1973)

Let $X_{s,\ t}$, for nonnegative integers $0 \leq s \leq t$, denote a family of doubly indexed random variables that satisfy the following.

1. $X_{s,\ t} \leq X_{s,\ r} + X_{r,\ t}$ for all $s < r < t$.
2. The joint distribution of $X_{s,\ t}$ is the same as that of $X_{s+1,\ t+1}$ for all $0 \leq s \leq t$.
3. There exists $K < 0$ such that the expectation $E[X_{0,\ n}] = \mu_n$ exists and satisfies $\mu_n \geq K \cdot n$, for all natural numbers n.

Then there exists λ, for which $\lim_{n \to \infty} E[X_{0,\ n}]/n = \lambda$.

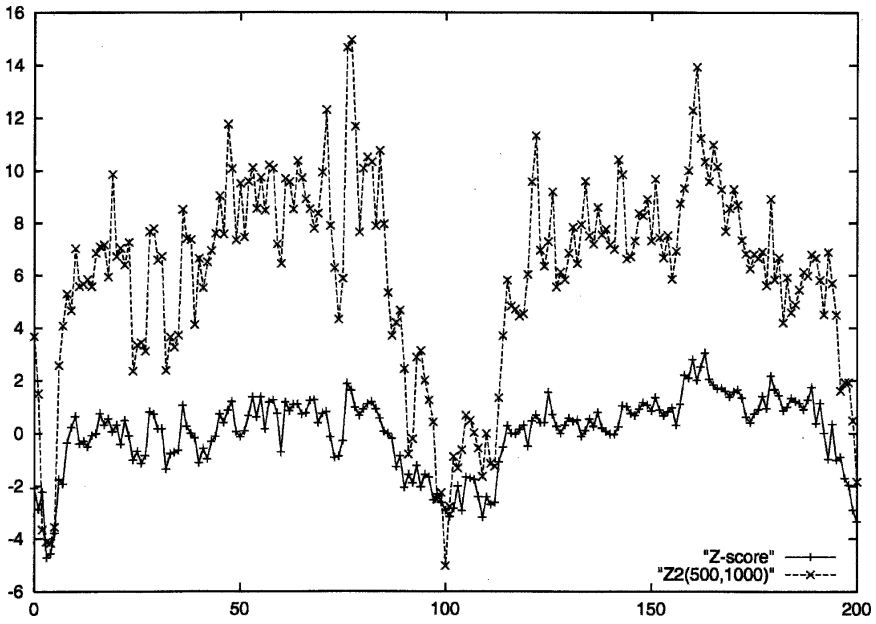

FIGURE 9. A plot of *Z*-scores and asymptotic *Z*-scores for 32-nt windows of artificial data obtained by planting SECIS element fruA CCUCGAGGGGAACCCGAAAGGGACCC GAGAGG in the middle of random RNA of compositional frequency $A = 0.28125$, $C = 0.28125$, $G = 0.40625$, and $U = 0.03125$ (i.e., of the same compositional frequency as that of fruA). For each size 32 window, *Z*-scores were computed with respect to 25 random RNAs of length 32, obtained by applying Algorithm 4 to the current window contents; thus, each randomization of current window contents had the same dinucleotide frequency as that of the corresponding current window contents. Asymptotic *Z*-scores were computed by table look-up of pre-computed means and standard deviations of 50 random RNAs, each of length 1000, having the same expected dinucleotide frequency as that of current window contents (only within two decimal places), as computed by Algorithm 3. We computed and stored all dinucleotide frequencies (only up to two decimal places), and pre-computed *Z*-scores with respect to much larger (1000 nt vs. 32 nt) random RNA. Justification for this approach follows from an asymptotic limit stated in the text.

Downloaded from rnajournal.cshlp.org on August 9, 2022 - Published by Cold Spring Harbor Laboratory Press

Theorem 9

Let $\vec{q}_{xy}$ be a complete set of dinucleotide frequencies, let $\vec{q}_x$ be the induced set of mononucleotide frequencies, and let X denote the infinite sequence of random variables x_0, x_1, x_2, … such that x_0 has the distribution $\vec{q}_x$, and for all i, x_{i+1} has the distribution given by the conditional probabilities

$$Pr[x_{i+1} = x] = \frac{q_{u,x}}{Pr[x_i = u]}.$$

For all $0 \le s \le t$, define random variables $X_{s,t} = \text{mfe}(x_s, \ldots, x_{t-1})$, where mfe denotes minimum free energy as measured by Zuker's algorithm. Then the limits

$$\lim_{n\to\infty} \frac{E[\text{mfe}(x_0, \ldots, x_n)]}{n} = \frac{E[X_{0,n}]}{n} = \mu(\vec{q}_{xy})$$

and

$$\lim_{n\to\infty} \sqrt{\frac{E[X_{0,n}^2] - (E[X_{0,n}])^2}{n^2}} = \sigma(\vec{q}_{xy})$$

both exist and depend only on $\vec{q}_{xy}$.

PROOF: To prove the existence of the first limit stated in Theorem 9, we claim that the collection of doubly indexed random variables $X_{s,t}$ satisfies the three conditions of Kingman's subadditive ergodicity Theorem 8.

By analysis of the pseudocode of Zuker's algorithm, it is clear that minimum free energy of RNA is subadditive, and hence condition (1) holds. Indeed, in the Turner energy model (Mathews et al. 1999), stacking free energies and loop energies are additive, hence the minimum free energy of the concatenation $x_s, \ldots, x_{t-1}$ of subsequence $x_s, \ldots, x_{u-1}$ and subsequence $x_u, \ldots, x_{t-1}$ satisfies $\text{mfe}(x_s, \ldots, x_{t-1}) \le x_s, \ldots, x_{u-1} + \text{mfe}(x_u, \ldots, x_{t-1})$. Here is a concrete example:

$$\text{mfe}(ACGUACGUACGU) = -1.20$$
$$\text{mfe}(CAGUCCAUUUGGG) = -0.90$$
$$\text{mfe}(ACGUACGUACGUCAGUCCAUUUGGG) = -2.20$$

To show that condition (2) holds, we first claim that for all nonnegative integers s, $Pr[x_s = x] = Pr[x_0 = x] = q_x$, for any given $x \in \{A, C, G, U\}$. This is done by induction on s. When $s = 0$, this is by definition of x_0. Assume that $Pr[x_s = x] = Pr[x_0 = x] = q_x$, and consider x_{s+1}. Then

$$\begin{aligned} Pr[x_{s+1} = x] &= \sum_u Pr[x_s = u] \cdot Pr[x_{s+1} = x | x_s = u] \\ &= \sum_u Pr[x_s = u] \cdot \frac{Pr[x_s = u, x_{s+1} = x]}{Pr[x_s = u]} \\ &= \sum_u Pr[x_s = u, x_{s+1} = x] = q_x \end{aligned}$$

where the last equality follows from the definition of induced mononucleotide frequency q_x. It thus follows by induction that $Pr[x_s = u] = q_u$, for all natural numbers s and all $u \in \{A, C, G, U\}$. Since the sequence $x_0, x_1, x_2, \ldots$ of random variables follows a first-order Markov condition, clearly $Pr[x_{s+1} = yx_s = x] = Pr[x_{s'+1} = yx_{s'} = x]$ holds for all natural numbers s, s', and thus by induction on n, we have

$$Pr[x_s = a_0, \ldots, x_{s+n} = a_n] = Pr[x_{s'} = a_0, \ldots, x_{s'+n} = a_n]$$

and hence the doubly indexed random variable $X_{s,t}$ has the same joint distribution as that of $X_{s+1,t+1}$, for all natural numbers $0 \le s \le t$. Thus, condition (2) of Kingman's theorem is satisfied.

We now turn to establish condition (3) of Kingman's theorem. For fixed n, $E[X_{0,n}] = \mu_n$ must exist, since the sample space $\Omega = \{A, C, G, U\}$ is finite, all probability distributions for fixed n are finite, and we consider only finitely many random variables $x_0, \ldots, x_n$. Let K_0 be the minimum value, −3.42 kcal/mol, over all base stacking free energies from Turner's current rules (Xia et al. 1998)—for example, see "Stacking enthalpies in kcal/mol" from M. Zuker's Web site, http://www.bioinfo.rpi.edu/~zukerm/rna/energy/. Note that base stacking free energies are all negative; hence, we are choosing that base stacking free energy whose absolute value is largest. Except for the (negative) base stacking free energies, all other energies (hairpin, bulge, internal loop, multiloop) are positive. The nearest neighbor energy model with Turner's experimentally measured energies (Mathews et al. 1999) is additive, and there are at most $n/2$ base pairs in an RNA sequence of length $n + 1$ (going from 0 to n); hence, $K_0 \cdot n/2 \le \mu_n$ for all n. It follows that (3) holds, and hence the existence of limit

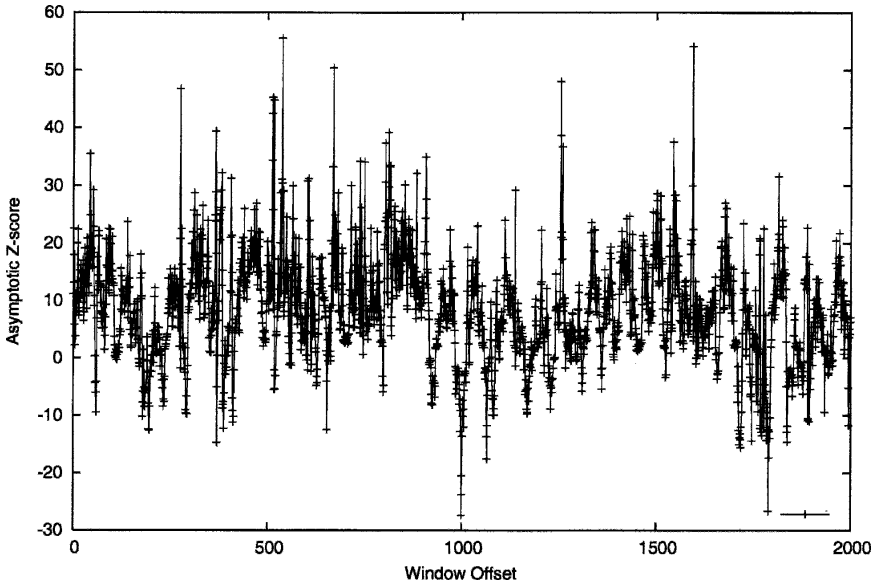

FIGURE 10. A plot of asymptotic Z-scores for 32 nt. Windows of artificial data obtained by planting SECIS element fruA at position 1000 in random RNA of compositional frequency $A = 0.28125$, $C = 0.28125$, $G = 0.40625$, and $U = 0.03125$ (i.e., of the same compositional frequency as that of fruA). Asymptotic Z-scores were computed by table look-up of pre-computed means and standard deviations of only 10 random RNAs, each of length 1000, having the same expected dinucleotide frequency as that of current window contents (only within two decimal places), as computed by Algorithm 3. We computed and stored all dinucleotide frequencies (only up to two decimal places), and pre-computed Z-scores with respect to much larger (1000 nt vs. 32 nt) random RNA.

Downloaded from rnajournal.cshlp.org on August 9, 2022 - Published by Cold Spring Harbor Laboratory Press

$$\lim_{n\to\infty} \frac{E[\mathrm{mfe}(x_0, \ldots, x_n)]}{n} = \mu(\vec{q}_{xy})$$

depending only on $\vec{q}_{xy}$ follows by application of Kingman's theorem.

To prove the existence of the second limit stated in Theorem 9, let $K = 3.42 = -K_0$, and define random variables $Z_{s,\,t} = K(t-s) + X_{s,\,t}$, and

$$Y_{s,t} = \frac{Z_{s,t}^2}{t-s} = \frac{(K(t-s) + \mathrm{mfe}(x_s, \ldots, x_{t-1}))^2}{t-s}$$

for all $0 \le s \le t$. We will show that the collection $Y_{s,\,t}$, for all $0 \le s \le t$, satisfies conditions (1), (2), and (3) of Kingman's ergodicity theorem. To prove the subadditivity condition (1), that is, that $Y_{s,\,t} \le Y_{s,\,r} + Y_{r,\,t}$ for all $0 \le s \le r \le t$, fix $0 \le s \le r \le t$, and temporarily let

$$\begin{aligned} A &= Z_{s,t} = K(t-s) + X_{s,t} \\ B &= Z_{s,r} = K(r-s) + X_{s,r} \\ C &= Z_{r,t} = K(t-r) + X_{r,t} \\ m &= r-s \\ n &= t-r \\ m+n &= t-s. \end{aligned}$$

Now

$$\begin{aligned} 0 &\le (nB - mC)^2 \\ 0 &\le n^2B^2 + m^2C^2 - 2mnBC \\ 2mnBC &\le n^2B^2 + m^2C^2 \\ mnB^2 + mnC^2 + 2mnBC &\le n(m+n)B^2 + m(m+n)C^2 \\ \frac{mnB^2 + mnC^2 + 2mnBC}{mn(m+n)} &\le \frac{n(m+n)B^2 + m(m+n)C^2}{mn(m+n)}. \\ \frac{B^2 + C^2 + 2BC}{m+n} &\le \frac{B^2}{m} + \frac{C^2}{n} \\ \frac{(B+C)^2}{m+n} &\le \frac{B^2}{m} + \frac{C^2}{n} \end{aligned}$$

Replacing B, C, m, n by the values they denote, we have shown that

$$\frac{(Z_{s,r} + Z_{r,t})^2}{t-s} \le \frac{Z_{s,r}^2}{r-s} + \frac{Z_{r,t}^2}{t-r}.$$

Since we have already established that $X_{s,\,t} \le X_{s,\,r} + X_{r,\,t}$, it follows that $K(t-s) + X_{s,\,t} \le K(r-s) + X_{s,\,r} + K(t-r) + X_{r,\,t}$; hence $Z_{s,\,t} \le Z_{s,\,r} + Z_{r,\,t}$. Since $Z_{s,\,t} \ge 0$, $Z_{s,\,r} \ge 0$, $Z_{r,\,t} \ge 0$, it follows that $Z_{s,\,t}^2 \ge (Z_{s,\,r} + Z_{r,\,t})^2$.[16] Thus

$$\frac{Z_{s,t}^2}{t-s} \le \frac{Z_{s,r}^2}{r-s} + \frac{Z_{r,t}^2}{t-r}$$

and hence $Y_{s,\,t} \le Y_{s,\,r} + Y_{r,\,t}$. This establishes subadditivity condition (1).

The proof that the joint distribution of $Y_{s,\,t}$ is the same as that of $Y_{s+1,t+1}$ for all $0 \le s \le t$ is as in our treatment of $X_{s,t}$ and $X_{s+1,t+1}$. This establishes condition (2) of Kingman's theorem.

Finally, since

$$Y_{s,t} = \frac{Z_{s,t}^2}{t-s} \ge 0,$$

condition (3) of Kingman's theorem holds, thus by application of Kingman's theorem, it follows that the limit

$$\lim_{n\to\infty} \frac{E[Y_{0,n}]}{n} = \zeta(\vec{q}_{xy})$$

exists and depends only on complete dinucleotide frequencies $\vec{q}_{xy}$. Note that

$$\begin{aligned} \lim_{n\to\infty} \frac{E[Y_{0,n}]}{n} &= \zeta(\vec{q}_{xy}) \\ &= \lim_{n\to\infty} \frac{E[(Kn + X_{0,n})^2/n]}{n} \\ &= \lim_{n\to\infty} \frac{E[K^2 n]}{n} + \frac{2KE[X_{0,n}]}{n} + \frac{E[X_{0,n}^2]}{n^2} \\ &= K^2 + 2K\,\mu(\vec{q}_{xy}) + \lim_{n\to\infty} \frac{E[X_{0,n}^2]}{n^2}. \end{aligned}$$

Define $\lambda(\vec{q}_{xy}) = \zeta(q_{xy}) - K^2 - 2K\mu(\vec{q}_{xy})$. It follows that

$$\lim_{n\to\infty} \frac{E[X_{0,n}^2]}{n^2} = \lambda(\vec{q}_{xy}).$$

Now the variance of $X_{0,\,n}$ satisfies $\mathrm{Var}[X_{0,\,n}] = E[X_{0,\,n}^2] - (E[X_{0,\,n}])^2$, thus dividing by n^2 and taking square roots of both sides of the equality, we have

$$\begin{aligned} \sigma(\vec{q}_{xy}) &= \lim_{n\to\infty} \sqrt{\frac{E[X_{0,n}^2] - (E[X_{0,n}])^2}{n^2}} \\ &= \sqrt{\lim \frac{E[Y_{0,n}]}{n} - \left(\lim \frac{E[X_{0,n}]}{n}\right)^2} \\ &= \sqrt{\lambda(\vec{q}_{xy}) - \mu^2(\vec{q}_{xy})}. \end{aligned}$$

This completes the proof of Theorem 9.

REFERENCES

Altschul, S.F. and Erickson, B.W. 1985. Significance of nucleotide sequence alignments: A method for random sequence permutation that preserves dinucleotide and codon usage. *Mol. Biol. Evol.* **2:** 526–538.

Bonnet, E., Wuyts, J., Rouze, P., and Van de Peer, Y. 2004. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* **20:** 2911–2917.

Dsouza, M., Larsen, N., and Overbeek, R. 1997. Searching for patterns in genomic data. *Trends Genet.* **13:** 497–498.

Eddy, S.R. 2001. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* **2:** 919–929.

———. 2002. Computational genomics of noncoding RNA genes. *Cell* **109:** 137–140.

Fagegaltier, D., Lescure, A., Walczak, R., Carbon, P., and Krol, A. 2000. Structural analysis of new local features in SECIS RNA hairpins. *Nucleic Acids Res.* **28:** 2679–2689.

Gautheret, D., Major, F., and Cedergren, R. 1990. Pattern searching/alignment with RNA primary and secondary structures: An effective descriptor for tRNA. *Comput. Appl. Biosci.* **6:** 325–331.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy,

[16] In order to obtain this last inequality, we needed $Z_{s,\,t} \ge 0$. This is the reason for working with $Z_{s,\,t}$, rather than $X_{s,\,t}$.

Downloaded from rnajournal.cshlp.org on August 9, 2022 - Published by Cold Spring Harbor Laboratory Press

S.R. 2003. Rfam: An RNA family database. *Nucleic Acids Res.* **31:** 439–441.

Harborth, J., Elbashir, S.M., Vandenburgh, K., Manninga, H., Scaringe, S.A., Weber, K., and Tuschl, T. 2003. Sequence, chemical, and structural variation of small interfering RNAs and short hairpin RNAs and the effect on mammalian gene silencing. *Antisense Nucleic Acid Drug Dev.* **13:** 83–105.

Hofacker, I.L., Priwitzer, B., and Stadler, P.F. 2004. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics* **20:** 186–190.

Kingman, J.F.C. 1973. Subadditive ergodic theory. *Ann. Probability* **1:** 893–909.

Klein, R.J., Misulovin, Z., and Eddy, S.R. 2002. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc. Natl. Acad. Sci.* **99:** 7542–7547.

Klosterman, P.S., Tamura, M., Holbrook, S.R., and Brenner, S.E. 2002. SCOR: A structural classification of RNA database. *Nucleic Acids Res.* **30:** 392–394.

Kryukov, G.V., Kryukov, V.M., and Gladyshev, V.N. 1999. New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements. *J. Biol. Chem.* **274:** 33888–33897.

Laferriere, A., Gautheret, D., and Cedergren, R. 1994. An RNA pattern matching program with enhanced performance and portability. *Comput. Appl. Biosci.* **10:** 211–212.

Le, S.Y., Chen, J.H., and Maizel, J.V.J. 1990a. Efficient searches for unusual folding regions in RNA sequences. In *Structure and methods: Human Genome Initiative and DNA recombination* (eds. R.H. Sarma and M.H. Sarma), pp. 127–136. Adenine Press, Schenectady, NY.

Le, S.Y., Malim, M.H., Cullen, B.R., and Maizel, J.V. 1990b. A highly conserved RNA folding region coincident with the Rev response element of primate immunodeficiency viruses. *Nucleic Acids Res.* **18:** 1613–1623.

Lescure, A., Gautheret, D., Carbon, P., and Krol, A. 1999. Novel selenoproteins identified in silico and in vivo by using a conserved RNA structural motif. *J. Biol. Chem.* **274:** 38147–38154.

Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. 2003. Vertebrate microRNA genes. *Science* **299:** 1540.

Macke, T.J., Ecker, D.J., Gutell, R.R., Gautheret, D., Case, D.A., and Sampath, R. 2001. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.* **29:** 4724–4735.

Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288:** 911–940.

Mathews, D.H., Turner, D.H., and Zuker, M. 2000. Secondary structure prediction. In *Current protocols in nucleic acid chemistry* (eds. S. Beaucage et al.), pp. 11.2.1–11.2.10. Wiley, New York.

Miranda-Rios, J., Navarro, M., and Soberon, M. 2001. A conserved RNA structure (thi box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proc. Natl. Acad. Sci.* **98:** 9736–9741.

Rivas, E. and Eddy, S.R. 2000. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* **16:** 583–605.

Seffens, W. and Digby, D. 1999. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.* **27:** 1578–1584.

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., and Steinberg, S. 1998. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* **26:** 148–153.

Tuschl, T. 2003. Functional genomics: RNA sets the standard. *Nature* **421:** 220–221.

Washietl, S. and Hofacker, I.L. 2004. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.* **342:** 19–30.

Waterman, M.S. 1995. *Introduction to computational biology: Maps, sequences and genomes.* Chapman and Hall, London, New York.

Wilkie, G.S., Dickson, K.S., and Gray, N.K. 2003. Regulation of mRNA translation by 5′- and 3′-UTR-binding factors. *Trends Biochem. Sci.* **28:** 182–188.

Winkler, W., Nahvi, A., and Breaker, R.R. 2002. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419:** 952–956.

Workman, C. and Krogh, A. 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.* **27:** 4816–4822.

Xia, T., SantaLucia Jr., J., Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C., and Turner, D.H. 1998. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37:** 14719–14735.

Zuker, M. 2003. Mfold Web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31:** 3406–3415.

Zuker, M. and Stiegler, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9:** 133–148.

Downloaded from rnajournal.cshlp.org on August 9, 2022 - Published by Cold Spring Harbor Laboratory Press

Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency

PETER CLOTE, FABRIZIO FERRÉ, EVANGELOS KRANAKIS, et al.

RNA 2005 11: 578-591

References This article cites 32 articles, 5 of which can be accessed free at:
http://rnajournal.cshlp.org/content/11/5/578.full.html#ref-list-1

License

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

To subscribe to *RNA* go to:
http://rnajournal.cshlp.org/subscriptions

Copyright 2005 by RNA Society