4816–4822 *Nucleic Acids Research, 1999, Vol. 27, No. 24*

No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution

Christopher Workman and Anders Krogh*

Center for Biological Sequence Analysis, Technical University of Denmark, Building 208, 2800 Lyngby, Denmark

Received July 12, 1999; Revised and Accepted October 22, 1999

ABSTRACT

This work investigates whether mRNA has a lower estimated folding free energy than random sequences. The free energy estimates are calculated by the mfold program for prediction of RNA secondary structures. For a set of 46 mRNAs it is shown that the predicted free energy is not significantly different from random sequences with the same dinucleotide distribution. For random sequences with the same mononucleotide distribution it has previously been shown that the native mRNA sequences have a lower predicted free energy, which indicates a more stable structure than random sequences. However, dinucleotide content is important when assessing the significance of predicted free energy as the physical stability of RNA secondary structure is known to depend on dinucleotide base stacking energies. Even known RNA secondary structures, like tRNAs, can be shown to have predicted free energies indistinguishable from randomized sequences. This suggests that the predicted free energy is not always a good determinant for RNA folding.

INTRODUCTION

The secondary structure of single-stranded RNA is known to implicate tertiary structure and function. Localized structures in mRNA have been shown to play important functional roles in translational regulation of some genes (1,2). However, it is unclear whether more global structures are formed by mRNA (3). In a recent paper (4) the folding free energy of mRNA from various organisms was predicted by the mfold program (5,6) and compared to that of random sequences with the same nucleotide distribution. The paper concludes that the native sequences on average have a significantly lower predicted free energy than the random sequences, and thus suggests that mRNA is likely to form secondary structure and that this biases the selection of codons.

The methods typically used for predicting RNA structure attempt to minimize the free energy of the molecule by maximizing the number of favorable base pairing interactions (7,8). The main contribution to the stability of RNA secondary structure is the free energy associated with base pairing. This free energy can be well approximated by stacking energies that depend not only on a single base pair, but on two neighboring base pairs (9,10). For instance, a C–G base pair is more favorable than a G–C base pair when stacked on top of an A–U base pair. It is well known that the dinucleotide distribution of DNA sequences is quite different from what would be expected from the nucleotide distribution alone (see for example 11). For this reason the dinucleotide frequency bias in the transcribed RNA should be important for the predicted free energy.

In Seffens and Digby (4) the predicted free energy of folding a native sequence was compared to several types of random sequences, such as random shuffling of the native sequence and randomly shuffled coding regions, but none of them preserved the dinucleotide distribution. The codon-shuffled sequences also tested in the paper are the most conservative in terms of dinucleotide statistics, and they turned out to have a lower average predicted free energy than the randomly shuffled sequences. In this paper we perform a similar analysis, but using random sequences with the same dinucleotide distribution as the native sequence. We find no evidence that (on average) mRNAs have lower predicted free energies than the random sequences.

The method is also tested on two well-known RNA structures: tRNA and the 18S rRNA from the ribosome small subunit. This analysis suggests that the method is not always sensitive enough to discriminate between random sequences and RNA with a known secondary structure. This may also indicate that the fold prediction method is not sensitive enough to detect small localized structures in long mRNA sequences (300–1200 nt).

MATERIALS AND METHODS

Data

An attempt was made to extract all the 51 mRNA sequences used in Seffens and Digby (4) from GenBank release 109.0. However, some were not found (HUMHPBS, HUMIFNAF and PHVLBA) and some had significantly different lengths (ECOALKA and ECODAPA) than that reported in Seffens and Digby (4) and were excluded. This set is listed in Tables 1 and 2. A set of five tRNA sequences was selected from the tRNA Sequence Database (12) and five ribosomal RNA sequences from Van de Peer *et al*. (13). These sets are listed in Table 3.

*To whom correspondence should be addressed. Tel: +45 4525 2477; Fax: +45 4593 1585; Email: krogh@cbs.dtu.dk

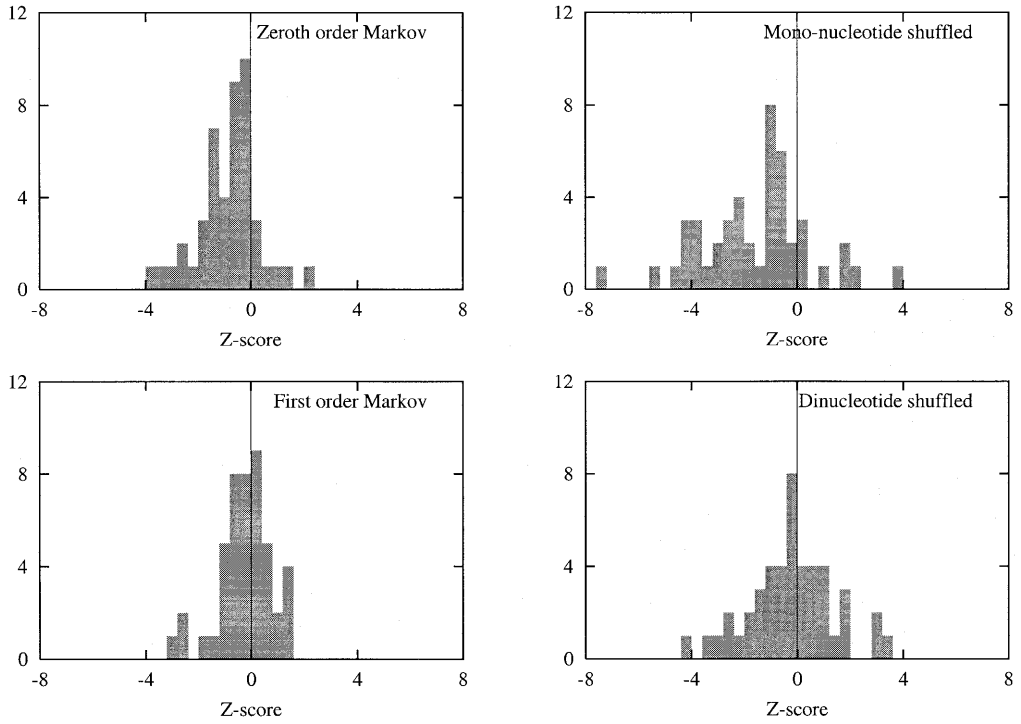

Figure 1. Histograms of the *Z*-scores for the four types of random sequences. The top row are for the zero order randomizations, the zero order Markov (left) and shuffled sequences (right), while the bottom row are for the first order randomizations, first order Markov (left) and dinucleotide shuffled (right).

Comparison between wild-type and random sequences

The newest version of the mfold program was obtained from Michael Zuker (mfold v.3.0). The mfold program minimizes a free energy function, which sums contributions from stacking, loop lengths, etc. It actually estimates the difference between the free energy of the unfolded state and the folded state. For any given RNA sequence length, the lower the energy estimate the more stable the predicted fold. The minimization is done by a dynamic programming method that always finds the secondary structure with the minimum free energy under a simplified secondary structure model (7,14). Although the various contributions to the free energy are obtained from experiments (9,10), a simplified model of RNA structure that disregards pseudoknots and other tertiary structures usually does not give a 100% correctly predicted secondary structure and occasionally the prediction is completely wrong.

For each native mRNA the minimum free energy prediction was found using mfold for energy calculations at 37°C. Then 10 random sequences of the same length as the native were generated and the minimum free energy prediction was found for each. The average and standard deviation were calculated and from these values the 'segment score' was found for each of the native sequences (4). The segment score is the number of standard deviations by which the predicted free energy of the native sequence is lower than the average of the random sequences. This is also called the *Z*-score, which will be the term used from here on. If the *Z*-score is positive then the native sequence has a higher minimum predicted free energy than the average of the random sequences and therefore is thought to have less secondary structure than random sequence. *Z*-scores were calculated for all the native sequences with the four different types of random sequences described below.

Random sequences

Making random sequences with exactly the same number of each nucleotide as the native sequence is trivial, one simply makes a random permutation of the nucleotides. It is less trivial to make a random sequence with exactly the same number of each dinucleotide as the native. We have made two different types of random sequences based on the dinucleotide distribution of the native sequence. Similarly, we made two types of random sequences based on the mononucleotide distribution to be used for comparison. The four types are detailed below.

Zero order Markov. The mononucleotide frequencies, P(*b*), for the native RNA sequence were calculated and used to generate a random sequence in which bases were simply chosen at random from P(*b*) until the length of the native sequence was reached (zero order Markov process).

Mononucleotide shuffled. The mononucleotide counts for the native RNA sequence were calculated and the rand-seq program, written by Gerald Hertz (unpublished), was used to generate a shuffled version. Rand-seq is given the sequence length and nucleotide counts and draws at random, weighted by the nucleotide proportions, until all counts are depleted.

First order Markov. From the native sequence the conditional probability P(*a*|*b*) of nucleotide *a* given *b* is found from the frequencies of the 16 possible pairs *a*,*b*. A random sequence is generated by first choosing a random nucleotide x_1, and then generate a sequence by choosing each nucleotide $x_i + 1$ from

Table 1. Comparison of the *Z*-scores and *P* values from each of the four random sequence models

	shuffled mono-nuc.		zero order		shuffled di-nuc.		first order	
gene name	Z-score	p-value	Z-score	p-value	Z-score	p-value	Z-score	p-value
DROCSKB	-0.71	0.252	-0.04	0.478	0.03	0.506	0.22	0.619
DROMETO	-0.68	0.247	-0.26	0.401	1.12	0.846	-0.05	0.493
DROSIST	1.97	0.969	0.39	0.642	1.70	0.927	0.78	0.769
DROTU4A	3.72	0.999	1.37	0.898	1.88	0.933	1.39	0.894
DROUBXDR	-7.43	0.000	-1.71	0.072	-4.30	0.002	-2.53	0.005
DROVMP	1.83	0.953	0.72	0.737	3.05	0.992	1.56	0.923
ECOADD	-0.63	0.264	-0.74	0.220	-1.33	0.114	-0.86	0.215
ECOCMA	-3.76	0.011	-3.44	0.005	-2.46	0.030	-2.59	0.018
HUMALR	-4.08	0.001	-1.31	0.123	-1.66	0.090	0.28	0.611
HUMCAL	-3.10	0.013	-1.87	0.049	-0.01	0.473	0.52	0.654
HUMCALCI	-2.51	0.021	-1.59	0.076	-0.81	0.223	-0.06	0.470
HUMGRP5E	-2.39	0.035	-3.75	0.004	0.07	0.536	0.12	0.539
HUMGST	0.17	0.547	0.94	0.799	0.82	0.773	0.02	0.515
HUMHEMBP	-4.25	0.001	-2.84	0.009	-3.14	0.017	-1.08	0.156
HUMHIS4	-2.89	0.015	-1.27	0.145	-1.38	0.113	-0.82	0.215
HUMIFNAB	-3.72	0.011	-2.28	0.026	-2.53	0.017	-1.40	0.118
HUMIFNAC	-4.08	0.001	-0.84	0.229	-0.67	0.255	-0.58	0.287
HUMIFNAH	-4.65	0.002	-2.71	0.004	-2.21	0.027	-0.64	0.266
HUML12A	-2.09	0.039	-0.49	0.350	-0.53	0.278	-0.57	0.304
HUMOGC	-1.20	0.132	-0.36	0.373	1.17	0.858	0.26	0.602
HUMP1BX	-0.36	0.386	-0.31	0.386	0.42	0.661	0.24	0.601
MMU03711	0.30	0.600	-0.29	0.367	0.43	0.651	0.45	0.648
MUSCASK	-5.57	0.001	-1.19	0.140	-3.39	0.006	-1.98	0.060
MUSCRYGD	-1.32	0.126	-1.70	0.077	-0.12	0.444	-0.88	0.217
MUSCTNCA	2.22	0.974	2.12	0.980	3.54	0.997	1.52	0.909
MUSGBPA	-0.72	0.265	0.38	0.638	0.94	0.807	1.01	0.839
MUSGLOBZ	-0.13	0.459	-0.43	0.338	-0.22	0.422	0.17	0.551
MUSHIS3A	-0.88	0.207	-0.98	0.196	-0.91	0.197	-0.54	0.322
MUSLACPI	0.24	0.605	0.28	0.611	1.37	0.912	1.15	0.850
MUSMK2P	-1.74	0.069	-2.56	0.033	-0.52	0.329	-0.80	0.224
MUSNGF7S	-1.10	0.168	-0.48	0.334	2.93	0.993	0.61	0.727
BNANAP	-1.89	0.045	-0.70	0.275	-0.38	0.371	-0.52	0.349
PEAABN1M	-0.97	0.169	-0.05	0.449	-0.22	0.375	-0.13	0.427
PHVCHM	-3.80	0.001	-0.30	0.378	-1.85	0.073	-2.93	0.010
SOYCHPI	-2.03	0.048	-1.56	0.084	-0.95	0.203	-0.28	0.364
SOYHSP176	-0.69	0.264	-0.60	0.266	-0.01	0.469	-0.21	0.378
TAHI02	0.99	0.811	-0.07	0.444	0.59	0.694	0.09	0.541
TOMRBCSD	-1.03	0.178	-0.36	0.374	-0.33	0.361	-0.15	0.426
XELGSCHB	-2.48	0.028	-0.52	0.312	0.51	0.666	0.09	0.546
XELHISH1	-1.08	0.156	-0.68	0.275	1.80	0.930	1.36	0.891
XELIGFIA	-2.42	0.033	-1.50	0.104	-1.00	0.192	-0.14	0.429
XELLBL	-0.85	0.249	-0.63	0.289	0.11	0.540	-0.01	0.469
XELPCNA	-2.10	0.044	-1.48	0.081	-1.22	0.131	-0.70	0.277
XELPYLA	-0.56	0.301	-0.25	0.394	0.16	0.574	0.60	0.738
XELRIGA	-3.52	0.006	-0.98	0.180	-0.76	0.253	-0.75	0.238
XELSRBP	-1.05	0.186	-1.35	0.124	-0.02	0.482	-0.56	0.294
mean	-1.59		-0.83		-0.22		-0.20	

The shuffled mono-nuc. and zero order Markov columns correspond to the zero order random models. The shuffled di-nuc. and first order Markov columns show the first order randomization statistics.

the probability $P(x_i + 1|x_i)$ (first order Markov process). The process is stopped when the sequence has exactly the same length as the native.

Dinucleotide shuffled. Dinucleotide shuffling is performed in the following way. At each iteration a random trinucleotide is chosen (e.g. ATT). Then all the non-overlapping trinucleotides that begin and end with the same bases (e.g. AAT, ACT, AGT and ATT) are shuffled at random. This is done *N* times, where *N* was chosen to be 10 times the length of the native sequence.

The random sequences generated by one of the Markov processes will be 'truly' random, meaning that the only relation to the native sequence is the mononucleotide (zero order) or dinucleotide (first order) distribution. However, the exact number of each nucleotide or dinucleotide will fluctuate around the numbers in the native sequence.

Table 2. Comparison of the native folding energies with the averages from each of the four random sequence models

gene name	length	native	shuffled mono-nuc.	zero order Markov	shuffled di-nuc.	first order Markov
DROCSKB	946	-227.7	-222.8	-227.0	-228.0	-231.6
DROMETO	301	-78.2	-73.6	-76.1	-83.2	-77.7
DROSIST	782	-245.8	-257.2	-252.3	-258.0	-260.8
DROTU4A	625	-166.2	-187.1	-186.3	-176.4	-183.3
DROUBXDR	663	-179.5	-152.9	-152.8	-156.2	-153.1
DROVMP	434	-120.0	-134.1	-129.3	-136.7	-139.1
ECOADD	1039	-369.3	-360.5	-360.2	-360.4	-359.6
ECOCMA	901	-264.6	-240.3	-232.7	-240.4	-230.7
HUMALR	1132	-438.3	-409.9	-407.7	-427.6	-442.9
HUMCAL	791	-291.9	-276.5	-272.6	-291.8	-302.1
HUMCALCI	681	-262.3	-242.8	-245.9	-258.0	-261.5
HUMGRP5E	797	-271.4	-255.3	-244.3	-271.9	-273.2
HUMGST	909	-204.0	-205.3	-209.8	-208.2	-204.2
HUMHEMBP	822	-343.0	-305.7	-299.2	-321.9	-324.7
HUMHIS4	390	-163.6	-151.9	-145.8	-156.2	-155.0
HUMIFNAB	1041	-298.8	-262.3	-261.6	-277.8	-277.8
HUMIFNAC	963	-263.6	-240.1	-250.7	-257.9	-250.9
HUMIFNAH	985	-270.4	-232.3	-235.1	-253.7	-259.5
HUML12A	612	-200.1	-187.1	-192.2	-197.0	-193.7
HUMOGC	891	-275.8	-267.5	-270.8	-281.7	-281.3
HUMP1BX	480	-196.4	-193.3	-192.5	-198.8	-198.7
MMU03711	618	-186.9	-188.0	-183.7	-190.1	-192.3
MUSCASK	785	-189.0	-158.8	-163.2	-168.8	-163.2
MUSCRYGD	599	-219.2	-208.8	-205.6	-218.1	-209.4
MUSCTNCA	703	-230.8	-245.3	-255.8	-250.9	-253.9
MUSGBPA	478	-158.6	-152.3	-162.4	-165.0	-171.2
MUSGLOBZ	556	-174.5	-173.5	-169.9	-173.4	-178.0
MUSHIS3A	595	-225.7	-216.8	-210.5	-217.8	-214.8
MUSLACPI	844	-218.3	-219.8	-222.8	-230.2	-233.3
MUSMK2P	728	-278.3	-264.3	-252.5	-272.3	-268.0
MUSNGF7S	830	-265.7	-258.2	-256.2	-277.2	-273.6
BNANAP	718	-201.2	-191.3	-190.7	-198.8	-193.2
PEAABN1M	609	-152.7	-146.7	-152.1	-151.5	-150.6
PHVCHM	1132	-415.8	-391.8	-410.5	-401.5	-385.7
SOYCIIPI	425	-124.9	-112.5	-112.2	-118.5	-121.0
SOYHSP176	718	-197.5	-194.0	-188.3	-197.4	-194.6
TAHI02	626	-235.9	-244.8	-235.0	-239.3	-238.2
TOMRBCSD	778	-205.1	-199.9	-201.3	-202.7	-202.1
XELGSCHB	1069	-329.9	-315.0	-318.7	-334.3	-332.4
XELHISH1	1180	-326.1	-319.5	-316.8	-342.6	-342.3
XELIGFIA	941	-252.8	-232.1	-231.6	-244.9	-250.1
XELLBL	796	-192.2	-184.5	-186.1	-193.1	-192.1
XELPCNA	1018	-312.3	-296.4	-295.7	-301.1	-301.7
XELPYLA	411	-100.5	-96.5	-98.6	-101.3	-105.8
XELRIGA	499	-160.2	-147.1	-149.1	-154.2	-150.2
XELSRBP	892	-243.1	-235.9	-231.0	-243.0	-238.0

The shuffled mono-nuc. and zero order Markov columns correspond to the zero order random model folding energy averages. The shuffled di-nuc. and first order Markov columns correspond to the folding energy averages of the first order randomizations.

The shuffling methods preserve the exact nucleotide or dinucleotide composition. The mononucleotide shuffling produces a truly random sequence (by the definition above). A dinucleotide shuffled sequence will have exactly the same number of each dinucleotide, but may be 'less random' due to fewer possible dinucleotide-preserving permutations. In fact, one can think of extreme examples where the sequence is not changed at all (e.g. of the form AAAATTTT), but in the real examples we have looked at, the shuffled sequence has no resemblance to the native, and we believe them to be randomized very effectively. These sequences always start and end with the same nucleotides as the native sequence, but for sequences that are all longer than 300 bases, we consider that to be of no importance. A careful and detailed treatment of dinucleotide shuffling is given in Altschul and Erickson (15), though any dinucleotide-preserving randomization method will be limited by sequences with few dinucleotide permutations, as discussed above. The two first types of randomization (zero order Markov and mononucleotide shuffled) will be referred to as zero order randomizations and the last two as first order randomizations.

Statistical significance

For a single mRNA one would like to know whether the *Z*-score is significantly different from that of a random sequence. The distribution of *Z*-scores for random sequences turns out to be fairly well approximated by a normal distribution with mean 0 and standard deviation 1, although this has no theoretical justification (one would rather expect an extreme value distribution because of the minimization over all possible folds). Under this approximation a *Z*-score of less than –2.33 is significant at 1%.

To account for the deviations from normal we have directly estimated significance levels by order statistics. To determine significance of the *Z*-score, we must make comparison with the distribution of *Z*-scores for random sequences with the same lengths and nucleotide statistics as the native sequences. For each of the 46 native sequences a set of 101 random sequences are generated (by one of the four methods) and their free energies estimated from the predicted fold. The following bootstrap type procedure (16) was repeated 2000 times. For each of the 46 groups, a random sequence is selected (the test sequence) and a random subset of 10 sequences is selected from the remaining 100 sequences [there are $\binom{100}{10}$ or $\approx 10^{13}$ ways to calculate a mean and a variance in this manner]. The *Z*-score is calculated for the random test sequence from the mean and variance of the other 10 random sequences. The average *Z*-score over the 46 test sequences is also found. For a given native sequence, the fraction of random sequences with a *Z*-score lower than that of the native gives a very good approximation to the probability that a score lower than this value occurred at random. We call this the *P* value for the sequence.

In the limit of large sample size the *Z*-scores for the random sequences would have mean 0 and a standard deviation of 1. Therefore the average *Z*-score over 46 sequences would be normally distributed with average 0 and a standard deviation of $1/\sqrt{46}$ to a very good approximation. For a sample of just 10 sequences the standard deviation would be slightly larger than 1, but if we disregard that effect, an average *Z*-score of less than $-2.33/\sqrt{46} = -0.344$ would be significant at the 1% level. Although this is a good guideline, we have again used the order statistics described above. The fraction of the 2000 average random *Z*-score values that are lower than the average *Z*-score for the native sequences is the *P* value for the average.

RESULTS AND DISCUSSION

The *Z*-scores and *P* values for the mRNAs are shown in Table 1 for all four types of random sequences. The predicted free energies used for these calculations are shown in Table 2. The energies reported in Seffens and Digby (4) are all significantly larger than the ones we obtain, which could be due to different parameter settings or differences between the mfold versions. The *Z*-scores for the two zero order randomizations agree reasonably well with the findings of Seffens and Digby (4), who obtained an average *Z*-score of –1.23 for randomly shuffled sequences, for which we obtain –1.59 (and –0.83 for the Markov random). This difference may be because of the mRNAs missing in our set, the different version of the mfold program or because of fluctuations due to differences in the random sequences. The important point is that the trend is the same: on average the energies are lower for the native sequences than for the zero order random sequences. The distributions of *Z*-scores can be seen in Figure 1. The average *Z*-scores are clearly significant for the zero order shufflings with *P* values much less than 0.01 (which corresponds to a *Z*-score of approximately –0.35 as discussed in Materials and Methods).

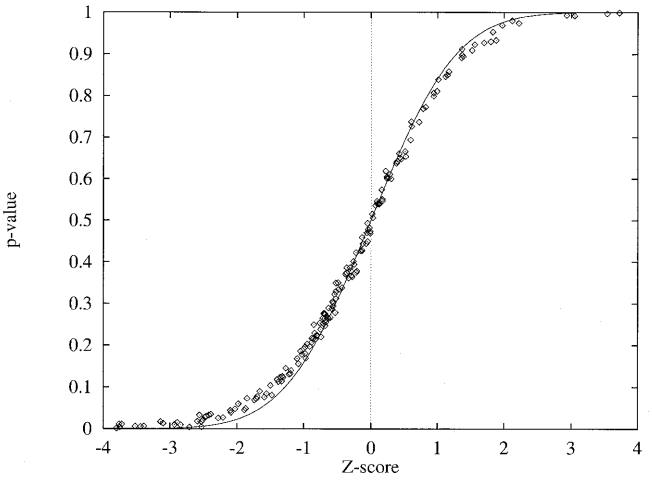

Figure 2. Correlation between *Z*-score and *P* value for all mRNAs and all four types of randomization. The different types of randomization are indistinguishable. The curve for a normal distribution of mean 0 and a standard deviation of 1 is also displayed.

The large difference between the zero order Markov sequences and the shuffled sequences is due to larger fluctuations in the energies for the Markov sequences than the others. This is primarily due to fluctuations in GC content that occur in Markov random sequences whereas sequence shufflings maintain constant GC:AT ratios. When the variance is larger the *Z*-scores attenuate. The average folding energies for the two types of random sequences are essentially the same. Table 1 also shows the *Z*-scores for the first order randomizations. Here the average is much smaller: –0.22 for the dinucleotide shuffled sequences and –0.20 for the first order Markov random sequences. Although there is still a negative trend in the score, it is quite clear from the score histograms in Figure 1 that it is insignificant. The histograms look almost symmetrical around 0, whereas a clear skew to the negative side is observed for the zero order randomizations. The order statistics for random sequences described in Materials and Methods gives a probability of 0.11 for a mean *Z*-score value of less than –0.22 for the dinucleotide shuffled sequences and a probability of 0.13 for a mean *Z*-score value of less than –0.20 for the first order Markov random sequences.

If we consider the sequences individually, we see that many individual *Z*-scores calculated with zero order randomizations appear significant while the large majority calculated with first order randomizations do not. If we use 0.01 as a significance threshold, we find 10 significant sequences from either the mononucleotide shuffled or zero order Markov *P* values. Only three sequences appear significant for either shuffled dinucleotide or first order Markov random and this supports the poor significance values we find for the average *Z*-scores. We also find three sequences with significantly higher predicted free

energies than expected from random (*P* values >0.99) in either of the first order randomization methods and only one sequence *P* value above this for the zero order models. The ratios of significantly low to significantly high sequence *P* values are 10:1 for the zero order methods and 1:1 for the first order methods. In Figure 2 the *P* values are plotted against the *Z*-scores for all types of randomizations. Notice that the distribution is close to normal, but the standard deviation is larger than 1, as expected (see Materials and Methods).

The process was repeated for five tRNAs and five 18S rRNAs for the monomer shuffled and dimer shuffled random models. The results are shown in Table 3. 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. Estimating *P* values from Figure 2 suggests that two of the zero order *Z*-scores are potentially significant while none of the dinucleotide shuffled *Z*-scores appear to be significant. For the rRNA there is quite a significant difference for all rRNA sequences. On average the predicted free energy of the native sequence is >8 SD from the random sequences, and they all have predicted free energies lower than the average of the random sequences. For these molecules there is only little difference in the results between the zero and first order randomizations.

Table 3. Comparison of the native folding energies with the averages from the shuffled mononucleotide and dinucleotide sequence models for selected tRNA and rRNA sequences

gene name	length	native energy	shuffled mono-nuc mean energy	shuffled mono-nuc Z-score	shuffled di-nuc mean energy	shuffled di-nuc Z-score
DE6281	72	-20.4	-21.60	0.32	-19.96	-0.16
DC8101	72	-16.4	-18.64	0.62	-19.04	1.01
DE7741	72	-26.2	-16.89	-6.75	-23.64	-0.95
DL9991	82	-37.8	-30.70	-2.27	-32.74	-1.71
DW6740	75	-23.2	-20.96	-0.76	-22.51	-0.17
mean				-1.77		-0.40
SPRRNASS	1842	-647.3	-566.39	-10.16	-569.37	-6.37
MMRNA18	1869	-776.6	-711.81	-5.40	-722.94	-4.54
DRORGAB	1995	-639.5	-579.35	-4.89	-581.21	-4.17
HSRRN18S	1869	-775.2	-714.39	-7.83	-720.18	-5.30
THARGAA	1471	-650.1	-545.28	-9.99	-546.37	-19.29
mean				-7.65		-7.93

The tRNAs are very short sequences (~70 bases) while the rRNAs are long (1500–2000 bases). Both are known to have global secondary structures. We suggest that extended structures

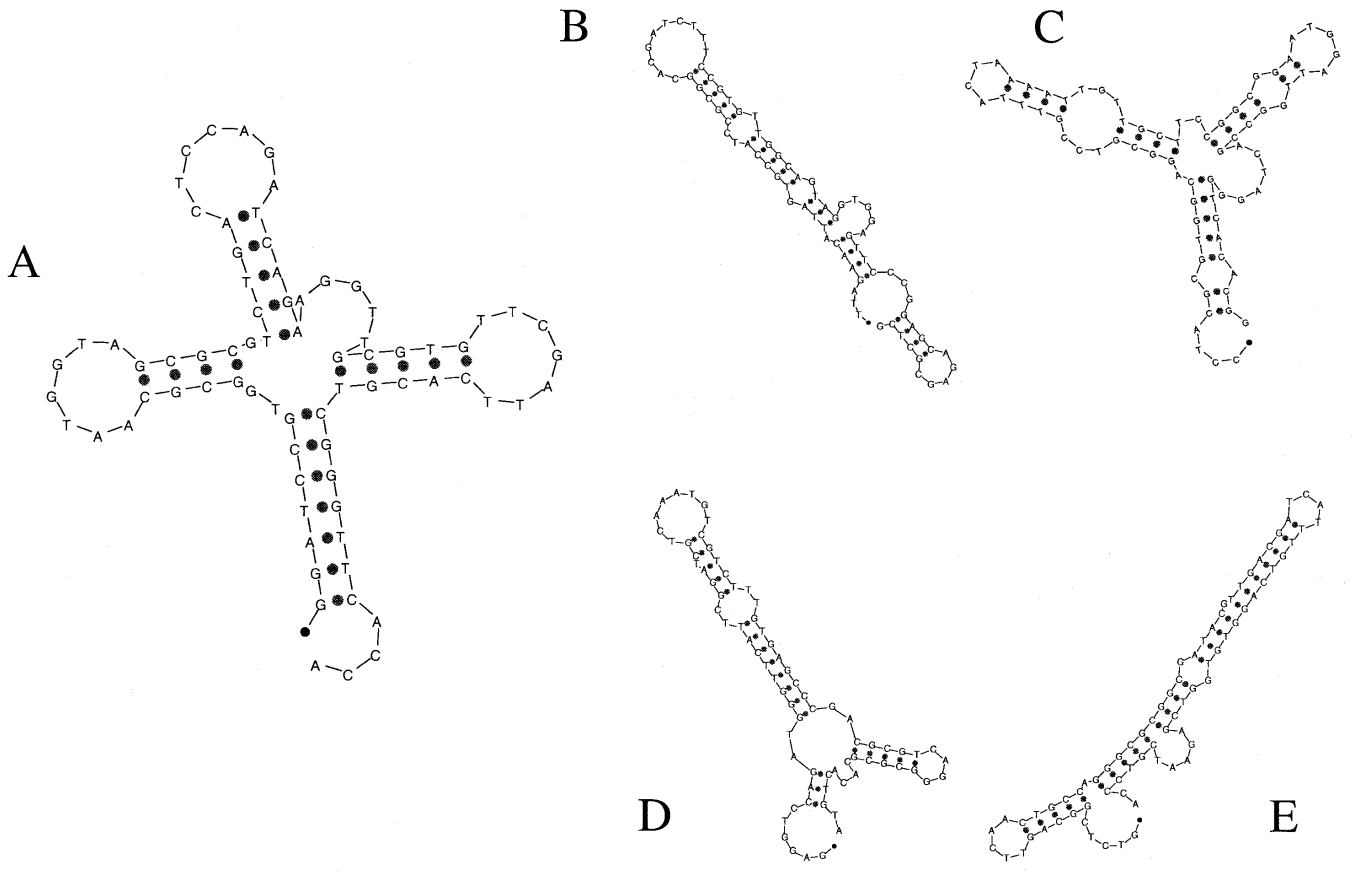

Figure 3. The fold predicted by mfold for an *Arabidopsis* tRNA (**A**) compared to secondary structure predictions for randomized versions of the same molecule. Structures (**B**) and (**C**) are examples from mononucleotide shuffled randomization while (**D**) and (**E**) show dinucleotide shuffled randomizations. Drawings were created with plt22ps by D. Stewart and M. Zuker.

of long RNA sequences need to be significantly more energetically favorable than those of short sequences for kinetic reasons. If there exist many suboptimal foldings with free energies close to the native, then they will compete with the native fold and the folding process will be extremely slow (17,18). Another explanation for the difference between tRNA and rRNA behavior is that tertiary structure, which is not taken into account in the predicted free energy calculation, contributes more significantly to the predicted free energy of tRNAs. Furthermore, since it is known that rRNAs have extended global secondary structures and both zero and first order randomization methods are capable of disrupting this global structure, one would expect the same trend in mRNA sequences should they have global secondary structures.

It is interesting that apparently 'well-folded' structures can be obtained from short random sequences. In Figure 3 a few structures predicted from shuffled tRNA sequences are shown.

CONCLUSIONS

The comparison of the predicted free energy of mRNA and random sequences with the same dinucleotide distribution shows no significant difference between the two for the 46 mRNAs studied here. This suggests that mRNA, in general, does not form more stable extended structures than random sequences. Considering the inability to distinguish short tRNAs (with well-known secondary structures) from randomized tRNAs, the method is probably not sensitive enough to determine whether mRNAs form localized structures. For example, stable hairpin loops are known to be important in translational control of some genes but are not likely to be detected in the predicted folding energy of an otherwise unfavorable global secondary structure. It was shown that the dinucleotide distribution is important for determining the significance of secondary structure, which is not surprising since stacking energies are crucial for the stability of RNA structure. It is unlikely that the dinucleotide distribution of mRNA is influenced by a need to form secondary structure, because the dinucleotide distribution is generally very similar in other types of DNA from the same organism (i.e. non-coding DNA) and varies greatly between coding regions of different organisms.

ACKNOWLEDGEMENTS

We would like to thank the referees for valuable suggestions and for bringing Altschul and Erickson (15) to our attention. This work was supported by a grant from the Danish National Research Foundation.

REFERENCES

1. de Smit,M.H. (1998) In *RNA Structure and Function*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 495–540.
2. Pantopoulos,K., Johansson,H.E. and Hentze,M.W. (1994) *Prog. Nucleic Acid Res. Mol. Biol.*, **48**, 181–238.
3. Doktycz,M.J., Larimer,F.W., Pastrnak,M. and Stevens,A. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 14614–14621.
4. Seffens,W. and Digby,D. (1999) *Nucleic Acids Res.*, **27**, 1578–1584.
5. Zuker,M. (1989) *Methods Enzymol.*, **180**, 262–288.
6. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) *J. Mol. Biol.*, **288**, 911–940.
7. Zuker,M. and Stiegler,P. (1981) *Nucleic Acids Res.*, **9**, 133–148.
8. Schuster,P., Fontana,W., Stadler,P.F. and Hofacker,I.L. (1994) *Proc. R. Soc. Lond. Ser. B Biol. Sci.*, **255**, 279–284.
9. Freier,S.M., Kierzek,R., Jaeger,J.A., Sugimoto,N., Caruthers,M.H., Neilson,T. and Turner,D.H. (1986) *Proc. Natl Acad. Sci. USA*, **83**, 9373–9377.
10. Turner,D.H., Sugimoto,N., Jaeger,J.A., Longfellow,C.E., Freier,S.M. and Kierzek,R. (1987) *Cold Spring Harbor Symp. Quant. Biol.*, **52**, 123–133.
11. Karlin,S. and Mrázek,J. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 10227–10232.
12. Sprinzl,M., Horn,C., Brown,M., Ioudovitch,A. and Steinberg,S. (1998) *Nucleic Acids Res.*, **26**, 148–153.
13. Van de Peer,Y., Robbrecht,E., de Hoog,S., Caers,A., De Rijk,P. and De Wachter,R. (1999) *Nucleic Acids Res.*, **27**, 179–183.
14. Nussinov,R., Pieczenik,G., Griggs,J.R. and Kleitman,D.J. (1978) *SIAM J. Appl. Math.*, **35**, 68–82.
15. Altschul,S.F. and Erickson,B.W. (1985) *Mol. Biol. Evol.*, **2**, 526–538.
16. Efron,B. and Tibshirani,R.J. (1993) *An Introduction to the Bootstrap*. Chapman and Hall, London, UK.
17. Uhlenbeck,O.C. (1995) *RNA*, **1**, 4–6.
18. Treiber,D.K and Williamson,J.R. (1999) *Curr. Opin. Struct. Biol.*, **9**, 339–345.