J. Math. Biol. (2008) 56:93–105
DOI 10.1007/s00285-007-0129-z

Mathematical Biology

Boltzmann ensemble features of RNA secondary structures: a comparative analysis of biological RNA sequences and random shuffles

Chi Yu Chan · Ye Ding

Received: 1 January 2007 / Revised: 31 August 2007 / Published online: 2 October 2007

Abstract Ensemble-based approaches to RNA secondary structure prediction have become increasingly appreciated in recent years. Here, we utilize sampling and clustering of the Boltzmann ensemble of RNA secondary structures to investigate whether biological sequences exhibit ensemble features that are distinct from their random shuffles. Representative messenger RNAs (mRNAs), structural RNAs, and precursor microRNAs (miRNAs) are analyzed for nine ensemble features. These include structure clustering features, the energy gap between the minimum free energy (MFE) and the ensemble, the numbers of high-frequency base pairs in the ensemble and in clusters, the average base-pair distance between the MFE structure and the ensemble, and between-cluster and within-cluster sums of squares. For each of the features, we observe a lack of significant distinction between mRNAs and their random shuffles. For five features, significant differences are found between structural RNAs and random counterparts. For seven features including the five for structural RNAs, much greater differences are observed between precursor miRNAs and random shuffles. These findings reveal differences in the Boltzmann structure ensemble among different types of functional RNAs. In addition, for two ensemble features, we observe distinctive, non-overlapping distributions for precursor miRNAs and random shuffles. A distributional separation can be particularly useful for the prediction of miRNA genes.

1 Introduction

The secondary structures of RNA molecules are important for their functions in gene regulation. For structural RNAs, e.g., transfer RNAs (tRNAs) and ribosomal RNAs

C. Y. Chan · Y. Ding (✉)
Wadsworth Center, New York State Department of Health, Center for Medical Science,
150 New Scotland Avenue, Albany, NY 12208, USA
e-mail: yding@wadsworth.org

(rRNAs), secondary structures have been well elucidated by comparative sequence analysis [7,25]. The importance of microRNAs (miRNAs) [2,3] in gene-regulation has become increasingly appreciated. The stem-loop structure for the precursor of a miRNA is involved in cellular processing of the precursor miRNA into mature miRNA [16,28]. A messenger RNA (mRNA), on the other hand, is unlikely to adopt a single and stable conformation, but rather exists in a population of structures [4,9]. Several studies have been conducted to assess the thermodynamic stability of secondary structures by evaluating the minimum free energy (MFE) values of biological sequences against those of randomized sequences. Seffens and Digby reported that mRNAs on average have significantly lower folding free energies than random RNAs of the same mononucleotide frequencies [24]. Workman and Krogh challenged this finding based on the argument that preserving the dinucleotide distribution is critical in the random shuffling process due to the dominant dinucleotide base-pair stacking energies for RNA secondary structure [26]. Their study found no evidence of lower folding free energies in mRNAs than randomly shuffled sequences with preserved dinucleotide frequencies. Bonnet and colleagues studied a set of non-coding RNAs and observed lower folding free energies in biological sequences than in random sequences for the majority of precursor miRNAs but not for their sets of tRNAs and rRNAs [5]. Clote and colleagues performed another analysis on a set of structural RNAs and also a set of mRNAs [10]. They concluded that structural RNAs have lower folding energies than random shuffles of the same dinucleotide frequencies, while the mRNAs possess folding energies that are comparable to those of their randomized counterparts.

Although several attempts have been made to examine the differences between biological and random sequences, the examination of the stability of secondary structure as indicated by the MFE has been the most common means for comparison. In recent years, ensemble-based approaches to RNA secondary structure prediction have become increasingly appreciated [11]. Based on a dynamic programming algorithm for calculating moments of the Boltzmann free energy distribution, Miklos and colleagues considered three ensemble characteristics: the Boltzmann probability of the MFE structure, the free energy distance between the MFE structure and the expected free energy value of the remaining free energy distribution, and the variance of the free energies of the Boltzmann distribution [20]. For a set of precursor miRNAs, a set of tRNAs and a set of 5S rRNA, they observed significant differences between biological sequences and random shuffles for all three characteristics. Here, we consider nine complementary ensemble features based on sampling and clustering of Boltzmann ensemble of RNA secondary structures that have been shown to have advantages over the MFE approach [13–15], and use these features to further examine the differences between biological sequences and random sequences. A majority of these features were included in our previous comparison between mRNAs and structural RNAs [14]. The results of our analyses support previous findings. Of particular interest, for two ensemble features, we observed distinctive, non-overlapping distributions for precursor miRNAs and random shuffles. A distributional separation can be particularly useful for the prediction of miRNA genes.

2 Materials and methods

2.1 RNA sequences

We included three separate sets of biological RNA sequences in this study, with a total of 118 sequences of lengths ranging from 64 to 983 nt. We started the sequence selection process by extracting those from our previous study on clustering of mRNA secondary structures [14], from our previous analysis on the performance of centroid structures [13], and from a comparative study by Bonnet and colleagues on differences in folding free energies between biological and random sequences [5]. The use of these sequences facilitates comparison of results among studies. There are a total of 227 sequences in the three extracted datasets, including 100 full-length human mRNAs randomly selected from the NCBI Reference Sequence (RefSeq) database [23], 81 structural RNAs of diverse types that were drawn by stratified random sampling from various online databases, and 46 experimentally validated precursor miRNAs for *C. elegans*, *D. melanogaster* and *H. sapiens*. (For a complete listing of all mRNAs, structural RNAs and precursor miRNAs, please see [5,13,14], respectively). Due to the need to generate a sufficiently large number of random shuffles for each biological sequence and time-consuming calculation of partition functions for structure sampling, we set a limit of 1,000 nt for sequence length to make the computations manageable under time and resource constraints. For structural RNAs, we included ten tRNAs, ten 5S rRNAs, ten RNase P RNAs, ten SRP RNAs, ten tmRNAs, eight group I introns, one group II intron, and one 23S rRNA. The length limit removed all of ten 16S rRNAs, nine of ten 23S rRNAs, one group I intron, and one group II intron from the original set of 81 structural RNAs. The final set of 118 RNA sequences for this study includes 12 mRNAs, 60 structural RNAs, and 46 precursor miRNAs.

2.2 Dinucleotide shuffling and test of significance

Because RNA secondary structure depends largely on base-pair stacking interactions, shuffling by preserving dinucleotide frequencies has been considered more appropriate than mononucleotide shuffling [26]. A dinucleotide shuffling algorithm was reported by Altschul and Erickson [1]. A program implementing this algorithm has been developed by the Clote lab [10] and is used here. Specifically, we generated 100 random shuffles for each of the 118 biological sequences. Ensemble statistics were computed for every random sequence and then an average value was calculated for each ensemble statistic for the 100 random shuffles. At the end, for each biological sequence, we have a list of its ensemble statistics for nine ensemble features described below, and also a corresponding list of averaged statistics for the random shuffles. We repeated this computation for all of the 118 RNA sequences.

Next, for each ensemble statistic, we would like to assess whether there exists a significant difference between the biological sequences of the same type and the random shuffles. Here, the type of RNAs is mRNA, or structural RNA, or precursor miRNA. Since there is a one-to-one correspondence between a biological sequence and its random shuffles, we performed paired Student's t tests for making two-group comparisons. Mean values of the two groups and the corresponding P value are reported for each ensemble statistic in the Results section. We did not combine the three types of RNAs for comparison, because they are functionally different and might have different features. Thus, we report findings for each type, and then compare and contrast the results among different RNA types.

2.3 Sampling and clustering of RNA secondary structures

Partition functions and samples from the Boltzmann-weighted ensemble of RNA secondary structures were computed with our Sfold software [12]. The forward step of the Sfold algorithm computes equilibrium partition functions for all substrings of an RNA sequence based on the Turner thermodynamic parameters [19,27]. The traceback step then applies a recursive sampling algorithm that draws a new base pair or unpaired base(s) given a partially formed structure, based on the conditional probabilities computed with the partition functions. In this study, a statistically representative sample of 1,000 structures was generated for each biological or random sequence. It has been shown that a sample size of 1,000 is sufficient to guarantee statistical reproducibility in typical sampling statistics, even for long sequences with enormous numbers of possible structures [14,15]. Statistics of the Boltzmann ensemble were then estimated using this sample of structures. Details of the partition function calculation and the sampling algorithm were reported in [15].

RNA clustering procedures were performed on each structure sample, and clustering statistics were computed. Our clustering approach has been described comprehensively in [8,13,14]. In short, the Sfold clustering module employs the top-down divisive hierarchical clustering method, Diana [18], in partitioning the sampled ensemble into structural clusters. Base-pair distance, as defined in [13,14], is used as the basis for evaluating dissimilarity between structures. The optimal number of clusters is determined by the CH index [6], which examines the ratio of between-cluster sums of squares to within-cluster sums of squares. Once the number of clusters is determined, we next determine to which cluster the MFE structure belongs, based on the radii of clusters and distances between the MFE structure and the clusters. The MFE structure is predicted by version 3.1 of the mfold software [30,31] for the same set of Turner thermodynamic parameters [19,27] that have been implemented in Sfold. Furthermore, we have introduced the notion of a centroid for any given set of structures. The centroid is the structure in the entire structure ensemble space that has the shortest total base-pair distance to the set of structures [13]. Ensemble centroid, the structure with the shortest total base-pair distance to the structures in the sampled ensemble, and cluster centroid, the structure with the shortest total base-pair distance to all structures in a cluster, can be easily computed as representatives of the ensemble and the cluster [13].

2.4 Ensemble features

A total of nine ensemble and clustering features are considered in our analyses. Statistics for these features are calculated with sampled ensemble of 1,000 structures. These features and calculations are described in detail below:

- *Number of clusters*. This is simply the optimal number of structural clusters determined by the CH index [6] for the set of 1,000 sampled structures. The use of CH index in our clustering procedures was discussed in [14].
- *Size of the largest cluster*. The size of a cluster is computed by the number of structures classified into that cluster divided by 1,000, the total number of sampled structures. It is a sample estimate of the probability of the cluster in the Boltzmann ensemble. The size of the largest cluster is computed here.
- *Size of the cluster containing the MFE structure*. The cluster to which the MFE structure belongs is determined by the method as described in [14]. The size of that cluster is reported here.
- *Energy gap between the MFE and the ensemble*. The average free energy of the sampled ensemble of 1,000 structures is first calculated. The difference between this average free energy and the MFE of this sequence is then computed. This energy gap is normalized by the length of sequence, which has been used before as the normalization factor for free energies of RNA secondary structures [22]. This sample-based statistic is essentially different from the "deviation" measure based on an exact algorithm [20]. The deviation was calculated from the free energy distribution with the MFE excluded while the MFE is not specifically removed in our average energy calculation. Thus, results from the two measures cannot be directly compared, particularly for short sequences for which a sample contains a substantial number of structures with MFE.
- *Number of high-frequency base pairs in the ensemble*. Base pairs with a sample frequency >0.5 are defined as high-frequency base pairs. We simply count the number of base pairs that appear in more than 500 structures in the sampled ensemble of 1,000 structures. Since the number of base pairs in a structure grows roughly linearly with sequence length, sequence length is used here as the normalization factor to allow comparisons among sequences of different lengths.
- *Average number of high-frequency base pairs per cluster*. This number is based on the application of the same base-pair frequency threshold of >0.5 to the clusters. For every cluster in the structure sample for a sequence, we first find the number of base pairs that appear in more than half of the structures in that cluster. We sum up the numbers for all clusters, divide the total by the number of clusters and then divide it again by the normalization factor of sequence length to obtain the normalized average number of high-frequency base pairs per cluster for the sequence.
- *Average base-pair distance between the MFE structure and the ensemble*. Base-pair distance between two structures is defined as the sum of the number of base pairs present only in one structure and the number of base pairs present only in the other structure. The mathematical definition of base-pair distance was described in [14]. We compute the base-pair distance between the MFE structure and each of the structures in the sampled ensemble, then sum up the base-pair distances over the structure sample and divide the sum by 1,000, and then finally normalize the value by sequence length.
- *Between-cluster sum of squares (BSS)*. BSS is a measure of closeness among clusters of structures for a sequence [14]. It is defined as $\text{BSS} = \sum_{1\leq i\leq k} n_i D(\text{EC}, \text{CC}^i)$, where k is the optimal number of clusters, n_i is the number of structures in the ith

cluster, $D(\cdot, \cdot)$ is the base-pair distance between two structures, EC is the ensemble centroid and CC^i is the centroid of the ith cluster. Here, ensemble centroid is defined as the structure in the entire structure ensemble space that has the shortest total base-pair distance to the 1,000 sampled structures [13]. Likewise, cluster centroid is defined as the structure in the entire structure ensemble that has the shortest total base-pair distance to all structures in that cluster [13]. It was shown that the structure formed by the high-frequency base pairs in the sampled ensemble is the ensemble centroid [13], and the structure formed by the high-frequency base pairs in a cluster is the cluster centroid [13]. The BSS value is normalized by the length of sequence for comparison purpose. Regression results from our previous analysis showed that normalization by sequence length for these sums of squares is appropriate [14].

- *Within-cluster sum of squares (WSS).* WSS is a measure of compactness of clusters of structures for a sequence [14]. It is defined as $\mathrm{WSS} = \sum_{1\le i\le k}\sum_{1\le j\le n_i} D(\mathrm{CC}^i, \mathrm{I}^{ij})$, where k is the optimal number of clusters, n_i is the number of structures in the ith cluster, $D(\cdot, \cdot)$ is the base-pair distance between two structures, CC^i is the centroid of the ith cluster and I^{ij} is the jth structure of the ith cluster. The WSS value is also normalized by the length of sequence for comparison purpose.

3 Results

For the nine ensemble features, the results of the comparison with random shuffles differ for the three types of RNAs. For mRNAs, we did not observe significant difference between biological sequences and random sequences for any of the features (Table 1). This shows that, on average, mRNAs and their shuffled sequences are not

Table 1 Comparison of ensemble and clustering features between biological and random sequences for the 12 mRNAs

Ensemble and clustering features	Mean and standard deviation		P value from paired t test
	Biological sequences	Random shuffles	
Number of clusters	2.92 ± 1.31	3.04 ± 0.17	0.7435
Size of the largest cluster	0.7606 ± 0.1697	0.7265 ± 0.0127	0.4927
Size of the cluster containing the MFE structure	0.5142 ± 0.3239	0.5588 ± 0.0395	0.6498
Energy gap between the MFE and the ensemble[a]	0.0291 ± 0.0049	0.0276 ± 0.0029	0.1095
Number of high-frequency base pairs in the ensemble[a]	0.2102 ± 0.0416	0.2137 ± 0.0116	0.7716
Average number of high-frequency base pairs per cluster[a]	0.2204 ± 0.0231	0.2262 ± 0.0116	0.3550
Average base-pair distance between the MFE structure and the ensemble[a]	0.2521 ± 0.0832	0.2448 ± 0.0141	0.7640
Between-cluster sum of squares[a]	50.62 ± 34.50	56.59 ± 4.62	0.5512
Within-cluster sum of squares[a]	137.54 ± 26.45	134.79 ± 8.29	0.6747

[a] Sequence length has been applied as the normalization factor

Table 2 Comparison of ensemble and clustering features between biological and random sequences for the 60 structural RNAs

Ensemble and clustering features	Mean and standard deviation		*P* value from paired *t* test
	Biological sequences	Random shuffles	
Number of clusters	3.25 ± 1.92	3.35 ± 0.36	0.6704
Size of the largest cluster	0.7265 ± 0.1598	0.7134 ± 0.0170	0.5193
Size of the cluster containing the MFE structure	0.6449 ± 0.2425	0.5982 ± 0.0386	0.1356
Energy gap between the MFE and the ensemble[a]	0.0260 ± 0.0083	0.0277 ± 0.0046	0.0620
Number of high-frequency base pairs in the ensemble[a]	0.2583 ± 0.0595	0.2243 ± 0.0202	3.559E-06
Average number of high-frequency base pairs per cluster[a]	0.2694 ± 0.0446	0.2401 ± 0.0200	1.243E-07
Average base-pair distance between the MFE structure and the ensemble[a]	0.1634 ± 0.0982	0.2192 ± 0.0327	3.301E-05
Between-cluster sum of squares[a]	56.58 ± 45.07	74.12 ± 10.91	0.0028
Within-cluster sum of squares[a]	92.01 ± 50.00	113.91 ± 22.16	0.0001

[a] Sequence length has been applied as the normalization factor

Table 3 Comparison of ensemble and clustering features between biological and random sequences for the 46 precursor miRNAs

Ensemble and clustering features	Mean and standard deviation		*P* value from paired *t* test
	Biological sequences	Random shuffles	
Number of clusters	5.61 ± 5.11	3.54 ± 0.32	0.0096
Size of the largest cluster	0.6708 ± 0.2046	0.7065 ± 0.0172	0.2447
Size of the cluster containing the MFE structure	0.6443 ± 0.2371	0.6460 ± 0.0261	0.9595
Energy gap between the MFE and the ensemble[a]	0.0161 ± 0.0084	0.0317 ± 0.0030	2.052E-17
Number of high-frequency base pairs in the ensemble[a]	0.3484 ± 0.0369	0.2163 ± 0.0181	5.429E-28
Average number of high-frequency base pairs per cluster[a]	0.3445 ± 0.0334	0.2335 ± 0.0162	3.236E-27
Average base-pair distance between the MFE structure and the ensemble [a]	0.0547 ± 0.0276	0.1850 ± 0.0150	8.969E-32
Between-cluster sum of squares[a]	29.27 ± 17.50	86.55 ± 6.96	2.133E-25
Within-cluster sum of squares[a]	32.10 ± 18.69	93.43 ± 9.05	8.488E-26

[a] Sequence length has been applied as the normalization factor

distinguishable. Structural RNAs, on the other hand, display significant differences from their randomized sequences for five of the nine features (Table 2). When compared to random sequences, structural RNAs have, on average, a larger number of high-frequency base pairs in the ensemble and in clusters, a smaller average base-pair distance between the MFE structure and the ensemble, a smaller between-cluster and within-cluster sums of squares. The precursor miRNAs show greater levels of distinction from their random shuffles (Table 3). Seven of the nine features are significantly different between the biological sequences and random sequences. These include the five significant features for the structural RNAs, the number of clusters and the energy gap between the MFE and the sampled ensemble. For the five features significant for both the structural RNAs and the precursor miRNAs, the P values from paired t tests for the precursor miRNAs are drastically lower than those for the structural RNAs. In the tables, we observe substantial differences in the standard errors for some of the features. A key assumption for the parametric t test is equal variances. We thus also performed the corresponding nonparametric Wilcoxon matched pairs signed-rank test that is not based on the equal variance assumption. The nonparametric test confirms the significance or insignificance by the t test for all features and all three RNA types, with the only exception for the number of clusters for the precursor miRNAs.

The distributions for two of the nine ensemble features are particularly intriguing. For each type of RNA, a histogram plot for the number of high-frequency base pairs is presented in Fig. 1. For mRNAs (Fig. 1a), a large degree of overlapping between the two distributions for the biological sequences and random sequences is observed. Their mean values, 0.2102 for mRNAs and 0.2137 for random shuffles, are comparable as shown in the plot, with a poor P value of 0.7716 from the t test for testing the difference in the means. For structural RNAs (Fig. 1b), the two mean values, 0.2583 for biological sequences and 0.2243 for random shuffles, are clearly separable from the plot, which is consistent with a low P value of 3.559E-06 for testing the difference in the means. The two distributions, however, still share an extensive overlapped region. For precursor miRNAs, Fig. 1c exhibits a clear separation between the two distributions for biological sequences and random sequences, with a significant gap between the two means (0.3484 for biological, 0.2163 for random, and P value of 5.429E-28 for the difference in the means). Intriguingly, the means for random shuffles for all three types of RNAs are quite comparable. The observed differences are mainly due to the increasing trend in the number of high-frequency base pairs from mRNAs to structural RNAs and to precursor miRNAs, implying more stable and conserved structures for precursor miRNAs and structural RNAs than for the mRNAs.

The other intriguing feature is the average base-pair distance between the MFE structure and the sampled ensemble. Figure 2 shows the histogram plot of the average base-pair distance for the mRNAs (Fig. 2a), the histogram for the structural RNAs (Fig. 2b), and the histogram for the precursor miRNAs (Fig. 2c). Similarly, we observe that the separation between distributions of the biological and random sequences increases from mRNAs to structural RNAs and to precursor miRNAs. For precursor miRNAs, in particular, the two distributions do not overlap at all. The average base-pair distance between the MFE structure and the ensemble for biological sequences is substantially shorter than the distance for random shuffles. This indicates that the similarity between the MFE structure and the sampled structures for precursor miRNAs is greater than that for their random shuffles.

4 Discussion

In this work, we have considered nine ensemble features for investigating the differences between biological sequences and random shuffles. We found that mRNAs are similar to their random shuffles, structural RNAs are different from random shuffles,

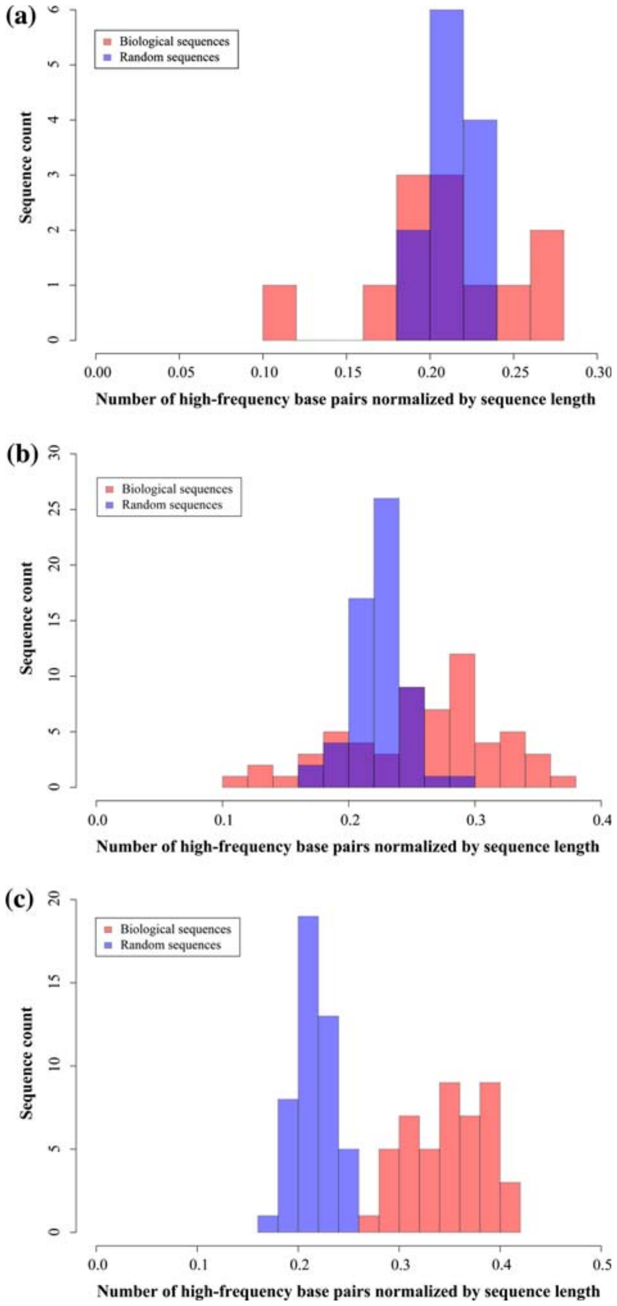

Fig. 1 Distribution of the number of high-frequency base pairs in the ensemble normalized by sequence length for biological sequences (*in red*), and the distribution for random sequences (*in blue*). The overlap of the distributions is in purple. The histograms are shown for the 12 mRNAs (**a**), the 60 structural RNAs (**b**), and the 46 precursor miRNAs (**c**)

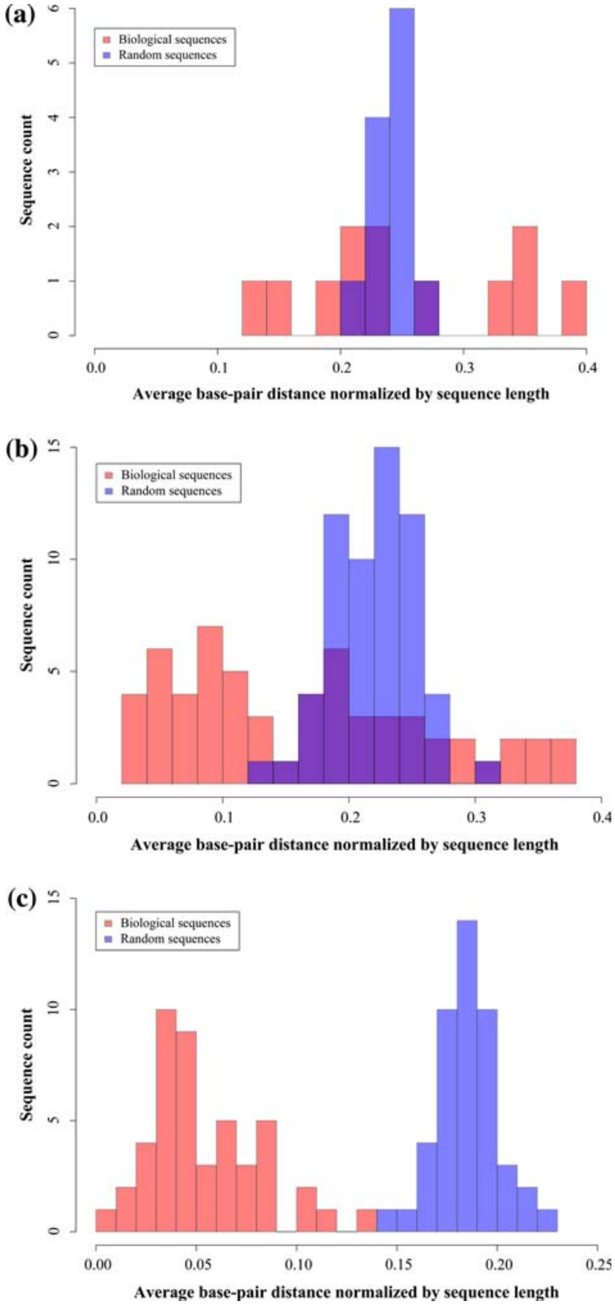

Fig. 2 Distribution of the average base-pair distance between the MFE structure and the ensemble normalized by sequence length for biological sequences (*in red*), and the distribution for random sequences (*in blue*). The overlap of the distributions is in purple. The histograms are shown for the 12 mRNAs (**a**), the 60 structural RNAs (**b**), and the 46 precursor miRNAs (**c**)

and there are much greater differences for precursor miRNAs. These findings are consistent with the previous reports [5,10,20,26], thus further confirming conclusions from these studies.

For randomization of RNA sequences, although dinucleotide shuffling has been argued to be more appropriate than mononucleotide shuffling, it does not guarantee the preservation of base-pair stacks in the predicted structure of the shuffled sequence. Thus alternative randomization methods that can preserve certain structural features warrant further investigation.

Our methods of investigation are different from methods used in the previous studies. The calculation of the ensemble statistics takes advantage of our structure sampling and clustering algorithms that have been shown to improve predictions for structural RNAs [13] and to effectively represent the likely population of mRNA structures [14], by overcoming the limited representation by the optimal folding (i.e., the MFE structure) or a heuristic set of suboptimal foldings [29]. Our previous studies [14,15] have shown that a sample size of 1,000 structures is sufficient to guarantee statistical reproducibility in typical sampling statistics. Larger structure samples will yield improved precisions in sampling statistics, increased power for the statistical tests and improved statistical significance. However, a drawback is the increased computational costs for structure clustering. In particular, memory requirement for clustering will become an issue.

Base-pair distance as the measure of dissimilarity between structures provides adequate discriminatory power for the purpose of clustering, and allows the transformation of the nonlinear problem of identifying the centroid structure into a simple linear problem [13]. Alternative distance metrics [17,21] may better address insertions and deletions. However, for comparison of structures generated for the same sequence, evolutionary consideration is not relevant here. Furthermore, these alternative metrics will introduce additional complexity to structure clustering, and the identification of the centroid structure is an open problem.

Due to the heterogeneous nature of our sequence set, sequence length has been chosen as the normalization factor for many of the ensemble features in order to facilitate comparisons among the diverse types of RNAs. This normalization scheme has proved to work well for energy-based models. However, it should be noted that the percentage of base pairs in structures determined by comparative analysis is not highly linear in sequence length. Nevertheless, this normalization scheme is a good approximation in our context, because our comparisons are concerned with predicted structures. For several ensemble features, e.g., the between/within-cluster sums of squares and the number of high-frequency base pairs, our previous study on comparing structural RNAs and mRNAs [14] has shown that the use of sequence length for normalization is well justified.

For precursor miRNAs, a significant finding is the clear distributional separation for the number of high-frequency base pairs in the ensemble, and for the average base-pair distance between the MFE structure and the ensemble. Experimental cloning for the discovery of new miRNA genes is both time-consuming and expensive, and the success rate largely depends on the precision of technology for detecting small RNAs. Computational methods can complement experimental techniques in miRNA gene identification. Our finding here can be particularly useful in this endeavor.

Acknowledgments The Computational Molecular Biology and Statistics Core at the Wadsworth Center is acknowledged for providing computing resources for this work. This work was supported in part by National Science Foundation grants DMS-0200970 and DBI-0650991, and National Institutes of Health grant R01 GM068726 to Y.D.

References

1. Altschul, S.F., Erickson, B.W.: Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. Mol. Biol. Evol. **2**, 526–538 (1985)
2. Ambros, V.: The functions of animal microRNAs. Nature **431**, 350–355 (2004)
3. Bartel, D.P.: MicroRNAs: genomics, biogenesis, mechanism, and function. Cell **116**, 281–297 (2004)
4. Betts, L., Spremulli, L.L.: Analysis of the role of the Shine–Dalgarno sequence and mRNA secondary structure on the efficiency of translational initiation in the Euglena gracilis chloroplast atpH mRNA. J. Biol. Chem. **269**, 26456–26463 (1994)
5. Bonnet, E., Wuyts, J., Rouze, P., Van de Peer, Y.: Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics **20**, 2911–2917 (2004)
6. Calinski, R.B., Harabasz, J.: A dendrite method for cluster analysis. Commun. Stat. **3**, 1–27 (1974)
7. Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Muller, K.M., Pande, N., Shang, Z., Yu, N., Gutell, R.R.: The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinf. **3**, 2 (2002)
8. Chan, C.Y., Lawrence, C.E., Ding, Y.: Structure clustering features on the Sfold Web server. Bioinformatics **21**, 3926–3928 (2005)
9. Christoffersen, R.E., McSwiggen, J.A., Konings, D.: Application of computational technologies to ribozyme biotechnology products. J. Mol. Struct. (Theochem) **311**, 273–284 (1994)
10. Clote, P., Ferre, F., Kranakis, E., Krizanc, D.: Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. RNA **11**, 578–591 (2005)
11. Ding, Y.: Statistical and Bayesian approaches to RNA secondary structure prediction. RNA **12**, 323–331 (2006)
12. Ding, Y., Chan, C.Y., Lawrence, C.E.: Sfold web server for statistical folding and rational design of nucleic acids. Nucleic Acids Res. **32**, W135–141 (2004)
13. Ding, Y., Chan, C.Y., Lawrence, C.E.: RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. RNA **11**, 1157–1166 (2005)
14. Ding, Y., Chan, C.Y., Lawrence, C.E.: Clustering of RNA secondary structures with application to messenger RNAs. J. Mol. Biol. **359**, 554–571 (2006)
15. Ding, Y., Lawrence, C.E.: A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. **31**, 7280–7301 (2003)
16. Han, J., Lee, Y., Yeom, K.H., Nam, J.W., Heo, I., Rhee, J.K., Sohn, S.Y., Cho, Y., Zhang, B.T., Kim, V.N.: Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell **125**, 887–901 (2006)
17. Jiang, T., Lin, G., Ma, B., Zhang, K.: A general edit distance between RNA structures. J. Comput. Biol. **9**, 371–388 (2002)
18. Kaufman, L., Rousseeuw, P.J.: Finding groups in data : an introduction to cluster analysis. Wiley series in probability and mathematical statistics. Applied probability and statistics, vol. xiv, 342 p. Wiley, New York (1990)
19. Mathews, D.H., Sabina, J., Zuker, M., Turner, D.H.: Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. **288**, 911–940 (1999)
20. Miklos, I., Meyer, I.M., Nagy, B.: Moments of the Boltzmann distribution for RNA secondary structures. Bull. Math. Biol. **67**, 1031–1047 (2005)
21. Moulton, V., Zuker, M., Steel, M., Pointon, R., Penny, D.: Metrics on RNA secondary structures. J. Comput. Biol. **7**, 277–292 (2000)
22. Pervouchine, D.D., Graber, J.H., Kasif, S.: On the normalization of RNA equilibrium free energy to the length of the sequence. Nucleic Acids Res. **31**, e49 (2003)
23. Pruitt, K.D., Tatusova, T., Maglott, D.R.: NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. **33**, D501–504 (2005)

24. Seffens, W., Digby, D.: mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucleic Acids Res. **27**, 1578–1584 (1999)
25. Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., Steinberg, S.: Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res. **26**, 148–153 (1998)
26. Workman, C., Krogh, A.: No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. Nucleic Acids Res. **27**, 4816–4822 (1999)
27. Xia, T., SantaLucia, J. Jr., Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C., Turner, D.H.: Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry **37**, 14719–14735 (1998)
28. Zeng, Y., Yi, R., Cullen, B.R.: Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. Embo. J. **24**, 138–148 (2005)
29. Zuker, M.: On finding all suboptimal foldings of an RNA molecule. Science **244**, 48–52 (1989)
30. Zuker, M.: Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. **31**, 3406–3415 (2003)
31. Zuker, M., Stiegler, P.: Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. **9**, 133–148 (1981)