doi:10.1016/j.jmb.2006.01.056

J. Mol. Biol. (2006) **359**, 554–571

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Clustering of RNA Secondary Structures with Application to Messenger RNAs

Ye Ding[1]*, Chi Yu Chan[1] and Charles E. Lawrence[1,2]*

[1] *Wadsworth Center, New York State Department of Health Center for Medical Science 150 New Scotland Avenue Albany, NY 12208, USA*

[2] *Center for Computational Molecular Biology, and Division of Applied Mathematics, Brown University, 182 George Street Providence, RI 02912, USA*

**Corresponding author*

There is growing evidence of translational gene regulation at the mRNA level, and of the important roles of RNA secondary structure in these regulatory processes. Because mRNAs likely exist in a population of structures, the popular free energy minimization approach may not be well suited to prediction of mRNA structures in studies of post-transcriptional regulation. Here, we describe an alternative procedure for the characterization of mRNA structures, in which structures sampled from the Boltzmann-weighted ensemble of RNA secondary structures are clustered. Based on a random sample of full-length human mRNAs, we find that the minimum free energy (MFE) structure often poorly represents the Boltzmann ensemble, that the ensemble often contains multiple structural clusters, and that the centroids of a small number of structural clusters more effectively characterize the ensemble. We show that cluster-level characteristics and statistics are statistically reproducible. In a comparison between mRNAs and structural RNAs, similarity is observed for the number of clusters and the energy gap between the MFE structure and the sampled ensemble. However, for structural RNAs, there are more high-frequency base-pairs in both the Boltzmann ensemble and the clusters, and the clusters are more compact. The clustering features have been incorporated into the Sfold software package for nucleic acid folding and design.

Keywords: RNA secondary structure; Boltzmann ensemble; clustering

Introduction

A growing number of gene regulation processes are known to be mediated by mRNA secondary structure. In prokaryotes, the efficiency of translation initiation is modulated by the structure of the ribosome binding site.[1] In eukaryotes, translation can be inhibited by increasing secondary structure in either the $5'$ untranslated region (UTR) or the coding region,[2,3] or can be regulated by structural features as mediators for the binding of proteins to repress or activate initiation of translation.[4] In addition, gene expression can be post-transcriptionally regulated through interaction between the mRNA and a small nucleic acid molecule of complete or partial complementarity. Examples include translational inhibition by antisense oligonucleotides and target cleavage by ribozymes in both eukaryotes and prokaryotes,[5,6] and more recently gene silencing by RNA interference (RNAi) mediated by short interfering RNAs (siRNA) in eukaryotic organisms,[7,8] and translational repression by microRNAs (miRNAs) in animals and plants.[9,10] Furthermore, in prokaryotes, transcriptional and translational regulation by riboswitches in the $5'$ UTRs has emerged as an important *cis*-regulatory mechanism,[11] in addition to non-metabolite-mediated transcriptional attenuation by alternative structures of leader transcripts[12,13] and regulation of mRNA stability and decay by stem–loop structures in the $3'$ UTRs.[14–16] In eukaryotes, riboswitches may control pre-mRNA splicing.[17–19] RNA secondary structure features are also involved in RNA editing.[20,21] The recently discovered regulatory mechanisms of RNAs have generated much excitement in the scientific community.

The strength of the base-pairing interaction between an mRNA and a small nucleic acid molecule is modulated by the secondary structure of the target mRNA.[22–29] Riboswitches regulate

Abbreviations used: UTR, untranslated region; MFE, minimum free energy; CDS, coding sequence; MDS, multidimensional scaling.

E-mail addresses of the corresponding authors: yding@wadsworth.org; lawrence@dam.brown.edu translation through conformational change in the 5′ UTR of the mRNA upon binding by a metabolite.[30] For the *cIII* gene of bacteriophage λ, two alternative structures were elucidated experimentally for the short mRNA of 132 nt and were proposed to regulate *cIII* expression.[31] One structure favors translation through accessible initiation codon and the Shine–Dalgarno sequence, and the other inhibits translation with poor accessibility. Although the structure of mRNA is important for numerous gene regulatory mechanisms, experimental determination of mRNA structure is generally difficult, and computational approaches have so far proved to be less than optimal. One impediment to the prediction of mRNA structures may stem from the likelihood that mRNAs exist in a population of structures,[32] as evidenced by multiple mRNA conformations in an equilibrium mixture.[33] Accordingly, computational secondary structure predictions based on free energy minimization may be poorly suited to this task.

In a departure from the long-established and productive paradigm of predicting RNA secondary structure *via* free energy minimization, we recently described an alternative for the characterization of the Boltzmann-weighted ensemble of RNA secondary structures.[34] In this approach, a statistically representative sample from the Boltzmann-weighted ensemble of RNA secondary structures is drawn. Such samples can faithfully and reproducibly characterize structure ensembles of enormous sizes. For example, it is striking but not surprising that samples of 1000 structures reproducibly represent Boltzmann ensemble of over 10^{300} RNA secondary structures.[34]

Here, we describe a procedure to capture the main features of the Boltzmann-weighted ensemble of structures by a small number of representative structures. This procedure has three steps: drawing of a statistical sample of RNA secondary structures,[34] clustering of the sampled structures, and determination of the centroids of these clusters.[35] The clustering step has been outlined briefly.[35] Here, we focus on describing the clustering method and cluster visual representations in complete details. The procedure is applied to 100 full-length human mRNAs that are randomly selected. The minimum free energy (MFE) structures for these 100 mRNAs have Boltzmann probabilities ranging from 1.07×10^{-5} for the shortest mRNA of 425 nt to 2.54×10^{-141} for the longest mRNA of 8458 nt. For each of these 100 mRNAs, 1000 secondary structures are sampled from the Boltzmann ensemble, and the MFE structure is determined. The sampled structures for each mRNA are then clustered, and the centroid of the ensemble and the centroids of individual clusters of structures are identified. For only 29% of the mRNA sequences, the MFE structure is in a dominant cluster. The ensemble centroid represents the ensemble substantially better than does the MFE structure, and the improvement is the largest when the MFE structure is not in the dominant cluster. For a given cluster, the cluster centroid represents the cluster better than does either the ensemble centroid or the MFE structure. In the case of no single dominant cluster, the cluster centroids represent the ensemble better than does the ensemble centroid. The clusters and centroids effectively delineate statistical characteristics in the Boltzmann ensemble. Cluster-level characteristics and statistics are shown to be also statistically reproducible. In a comparison between mRNAs and structural RNAs, similarities and differences in clustering features are identified. The clustering features and tools have been incorporated into the Sfold software package for nucleic acid folding and design.

Results

Clustering results of mRNAs

As an example to illustrate the clustering procedure and distinct features among the clusters, we begin this section with a description of the clustering results for the mRNA of *Homo sapiens* dystrophin (muscular dystrophy, Duchenne and Becker types (DMD)) for transcript variant Dp40 (1634 nt; GenBank accession no. NM_004019, and sequence ID 081). This description is followed by findings from the clustering results for all of the 100 mRNAs. The last two subsections give a comparison between the representativeness of these centroids and the representativeness of the MFE structure.

An example of clustering

For the mRNA of *H. sapiens* dystrophin, three distinct clusters are indicated by the CH index (see Methods) (Figure 1). Since our procedure

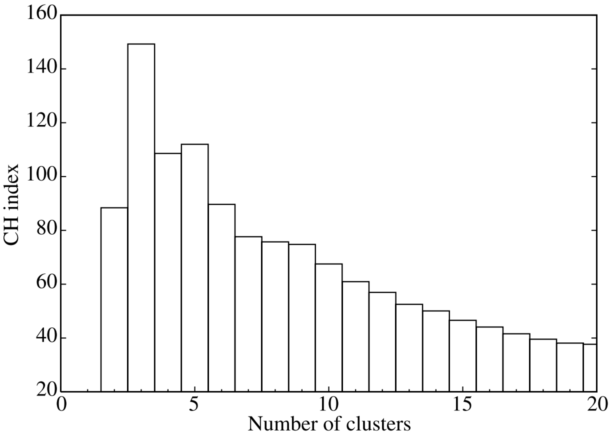

Figure 1. CH index profile for numbers of clusters ranging from 2 to 20 for the mRNA sequence (1634 nt, GenBank accession no. NM_004019, and ID 081) of transcript variant Dp40 for *H. sapiens* dystrophin (muscular dystrophy, Duchenne and Becker types (DMD)). The optimal number of clusters for this sequence is three, the highest value on the profile.

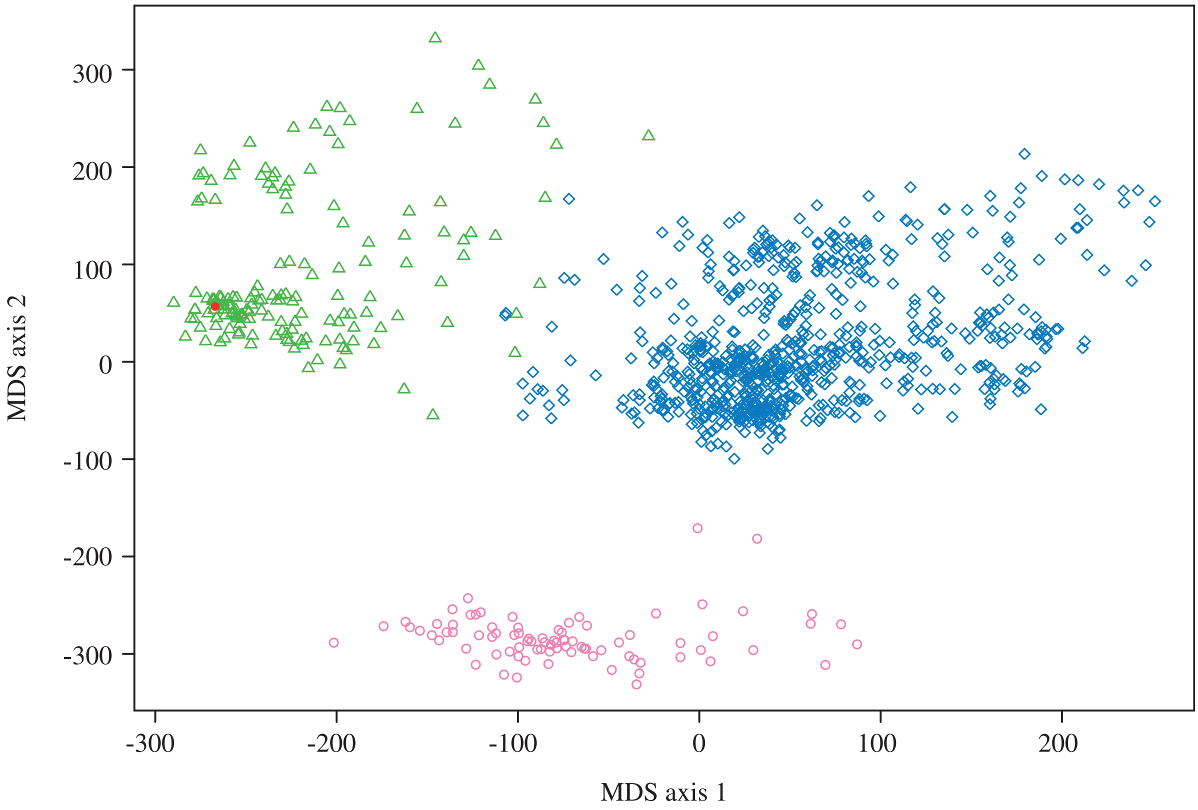

Figure 2. Two-dimensional representation of multi-dimensional scaling of all of the 1000 sampled structures and the MFE structure for the mRNA sequence of transcript variant Dp40 for *H. sapiens* dystrophin. Blue diamonds, green triangles and pink circles represent structures in the largest cluster, the second-largest cluster, and the smallest cluster, respectively. The MFE structure is in the second-largest cluster, and is represented by the red filled circle with coordinates $(X, Y)=(-266.9, 56.6)$.

samples structures in accordance with their Boltzmann-weighted probabilities, the probability of a cluster can be estimated by the frequency of the cluster in the sample. The probabilities of the three clusters are 0.754, 0.161, and 0.085, respectively. The MFE structure is in the second largest cluster. The three clusters are well separated in the two-dimensional representation of the multi-dimensional scaling (Figure 2).

We use the centroid of a cluster as a representative of the cluster.[35] Substantial differences among the clusters can be appreciated from the two-dimensional histograms of the clusters (Figure 3(a)–(c)) and from the circle diagrams of the centroids of the clusters (Figure 3(d)–(f)). The circle diagrams were produced by the Sir_graph program developed by Stewart & Zuker.[36] For example, in the first centroid structure, a local helical structure involves bases in the nucleotides 377 to 437 region and bases in the nucleotides 492 to 566 region (Figure 3(d)), and another local helical structure involves bases in the nucleotides 1156 to 1206 region and bases in the nucleotides 1376 to 1445 region. In the centroid of the second cluster, however, the above local structures are replaced by long-range interactions involving bases in the nucleotides 485 to 568 region and bases in the nucleotides 1156 to 1320 region (Figure 3(e)). Transition from one of these local structures to the other would require breakage and reformation of 113 base-pairs. The centroid of the third cluster requires another reconfiguration of the first local helical structure in the centroid of the first cluster, as bases in nucleotides 385 to 418 region are involved in long-range interactions with bases in nucleotides 1059 to 1086 region (Figure 3(f)), a region that is involved in mid-range interactions for the centroids of the first and the second cluster (Figure 3(d) and (e)). It is noted that pseudoknots may be partly responsible for the competing local structures. Our sampling algorithm does not permit pseudoknots; however, competing local structures in the sample may provide clues for pseudoknots.

Clustering results for all 100 human mRNAs

For the 100 human mRNAs, application of the CH index indicates that most (92 of 100) can be well characterized by five or fewer clusters (Figure 4(a)). Although the number of conformational states grows exponentially with sequence length, the number of clusters does not increase, on average, with the length of the sequence (Figure 4(b) and (c)). We found little evidence of a correlation between the cluster number and the sequence length (correlation coefficient$=$0.0692, p-value$=$0.4939). For the probability of the largest cluster, the cumulative probability distribution for the 100

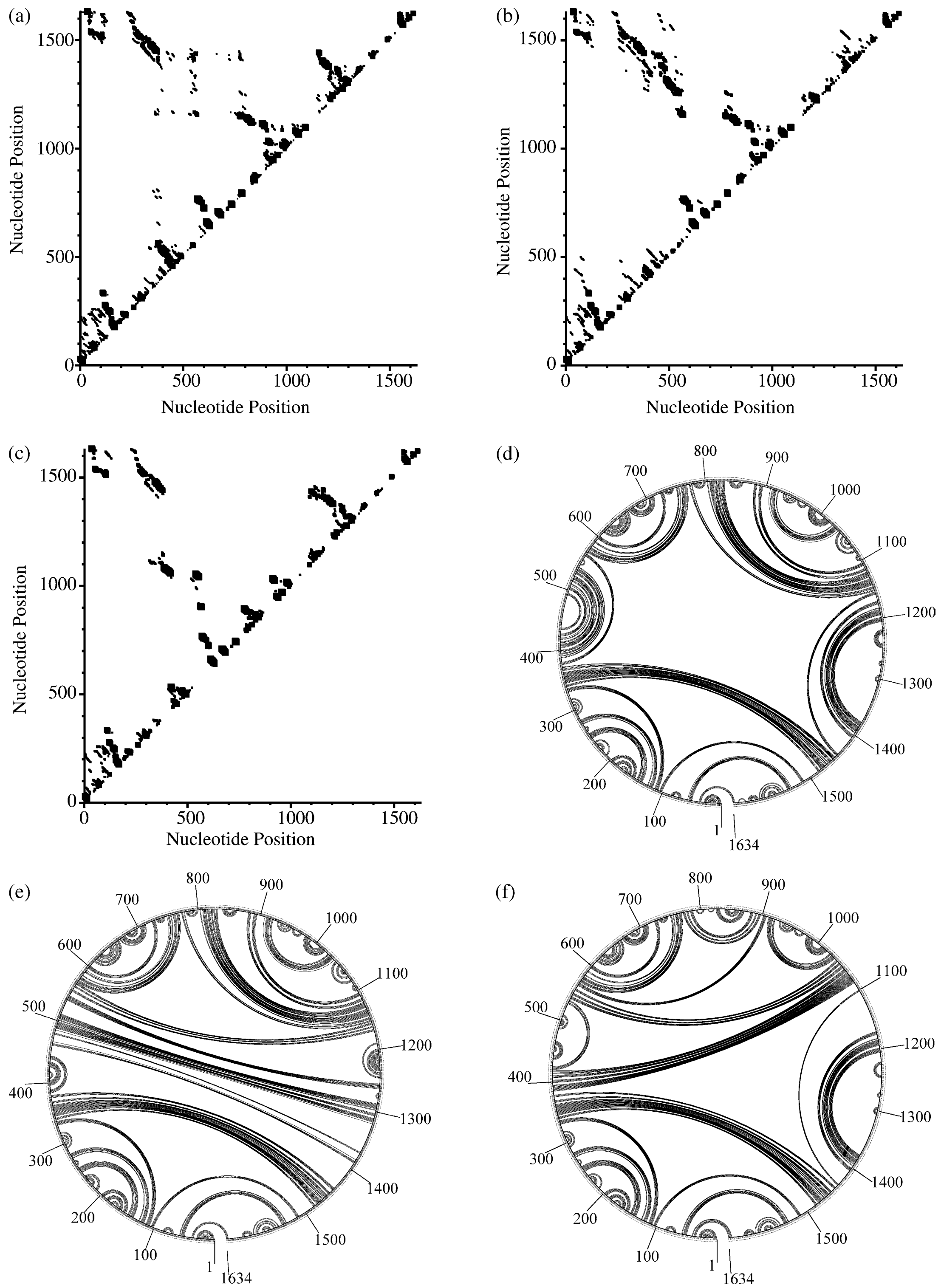

Figure 3. Two-dimensional histograms of individual clusters and circle diagrams of the cluster centroids for three distinct clusters (arranged in descending order of cluster size), for the mRNA sequence of transcript variant Dp40 for *H. sapiens* dystrophin. In the histogram for a cluster, the area of a square at position (i, j) is proportional to the frequency of base-pair $i \cdot j$ in the cluster. In a circle diagram, bases are positioned along a circle in the clockwise orientation, and a base-pair is shown by an arc that connects the two bases.

mRNAs is shown in Figure 5. For these sequences, the minimum for the probability of the largest cluster is 0.377. The size of the largest cluster is approximately uniformly distributed between 0.5 and 0.95 (Figure 5).

If we define a dominant cluster to be one with a probability of greater than 0.7, then 43% of the messages do not have a dominant cluster. By considering whether the MFE structure is in the largest cluster and whether the largest cluster is

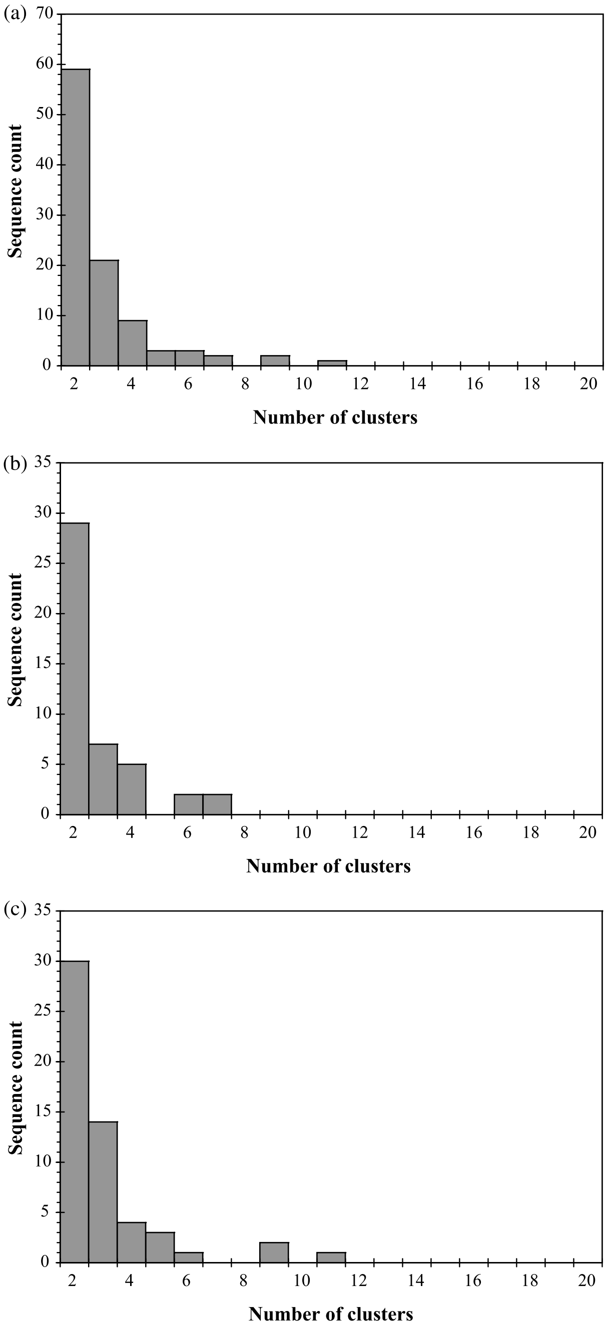

Figure 4. Distribution of the number of clusters for (a) the entire dataset of 100 human mRNA sequences, for (b) 45 sequences ≤2500 nt in length, and for (c) 55 sequences >2500 nt in length.

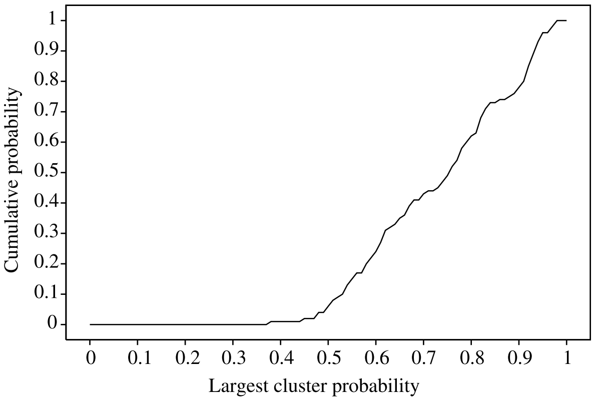

Figure 5. Cumulative distribution function for the size of the largest cluster for the 100 mRNA sequences. For each sequence, the size (i.e. the probability) of a cluster is estimated by the frequency of the cluster in the sample.

dominant, we find that these messages fall into one of four groups. Each of these groups is exemplified by two messages reported in Table 1. For the two messages from group 1, the MFE structure is in a dominant cluster. For group 2, the MFE structure is in the largest, but not the dominant cluster. For group 3 and group 4, the MFE structure is in a cluster of secondary probability. Sequence 046 (accession number NM_005656) presents an extreme case for which the MFE structure is not similar to any structure in the sample. The clustering results for all 100 mRNAs are available in Supplementary Data (Table S1). For 45 of the 100 messages, the MFE structure is clustered into the largest cluster (groups 1 and 2 in Table 2), but for only 29 of these 45 messages does the largest cluster dominate the ensemble (group 1 in Table 2). In addition, for 57 of the 100 messages, the MFE structure is contained in a cluster with a probability under 0.5. These findings suggest that the MFE structure does not generally represent well the Boltzmann-weighted ensemble of structures, and motivate our search for more representative structures.

Ensemble centroid represents Boltzmann-weighted ensemble substantially better than does the MFE structure

For each sequence, the ensemble centroid and cluster centroids are computed using the sampled 1000 structures. In the light of the population view of mRNA structures, we are interested in assessing how well a single structure represents the Boltzmann-weighted ensemble of secondary structures. For this purpose, we consider the average base-pair distance between a predicted structure and the sampled ensemble. As indicated in Figure 6, the ensemble centroid is at least 30% closer to the sampled ensemble than is the MFE structure for 66 messages, with an average distance improvement of 36.9% for all the 100 messages. To obtain summary statistics for a group of sequences, we first normalize the average distance by sequence length (see the legend to Table 2), and then average the normalized distances for the group. As indicated in Table 2, the ensemble centroid is about 27% closer to the sampled ensemble than is the MFE structure, when the MFE is present in the largest cluster (groups 1 and 2), and it is 40% to 50% closer when the MFE is absent in the largest cluster (groups 3 and 4).

Table 1. Clustering results for human mRNAs

Group[a]	Sequence ID[b]	GenBank accession no.	Sequence length (nt)	Number of clusters	Cluster probabilities (*cluster of MFE structure)
1	007	NM_004381	2622	4	0.813* 0.125 0.049 0.013
1	011	NM_002231	1623	2	0.980* 0.020
2	078	NM_016815	1019	2	0.642* 0.358
2	054	NM_130474	5640	2	0.590* 0.410
3	004	NM_021785	2338	3	0.919 0.062 0.019*
3	081	NM_004019	1634	3	0.754 0.161* 0.085
4	046	NM_005656	3226	3	0.518 0.482 0.000*
4	059	NM_033172	2711	3	0.592 0.277* 0.131

[a] Group 1, MFE structure is present in the largest cluster with a cluster probability >0.70. Group 2, MFE structure is present in the largest cluster with a cluster probability ≤0.70. Group 3, MFE structure is absent from the largest cluster with a cluster probability >0.70. Group 4, MFE structure is absent from the largest cluster with a cluster probability ≤0.70.

[b] Sequence ID is the identification number of the sequence in the sample of 100 mRNA sequences in this study.

Improved representation of ensemble and clusters by cluster centroids

Analogous to the ensemble centroid as a single representative of the Boltzmann-weighted ensemble, the centroid of a cluster is a representative of the cluster. To compare the degrees of representation of a cluster by its centroid, the ensemble centroid, and the MFE structure, we compute normalized distances to the cluster. We also compute the normalized distance from all cluster centroids to the sample (see footnote b of Table 2), for an ensemble-level comparison with the ensemble centroid and the MFE structure.

For the mRNA of *H. sapiens* γ-glutamyltransferase-like activity 1 (GenBank accession no. NM_004121, sequence ID 041 in Supplementary Data, Table S1), there are two major clusters. Cluster 1 has a probability of 0.554, and cluster 2 has a probability of 0.446. The MFE structure is in the smaller cluster. The distributions of the normalized distances to the structure sample for both the MFE structure and the ensemble centroid are bimodal (Figure 7(a)). In such cases, not surprisingly, neither the ensemble centroid nor the MFE structure represents the populations of structures well. However, centroids specific to the clusters represent the individual clusters and the ensemble substantially better than does any other single structure (Figure 7(b) and (c); and the last column in Table 2).

These results indicate that when there is no single dominant cluster, distance distributions tend to be multimodal. In such cases, no single structure is likely to be a good representative of the entire ensemble, and cluster-specific centroids are more appropriate representatives for individual clusters as well as for the Boltzmann ensemble.

Statistical reproducibility of clustering results

We have previously shown that ensemble-level sampling statistics are statistically reproducible.[34] Here, we examine reproducibility at the cluster level. For the mRNA sequence of transcript variant Dp40 for *H. sapiens* dystrophin, we generated two new independent samples, each with a sample size of 1000 structures. The two samples do not have a

Table 2. Improvement (in distance to the structure sample) by the ensemble centroid over the MFE structure, and by the cluster centroids over the ensemble centroid

Group	Number of sequences	Average normalized distance from the MFE structure to the sample[a]	Average normalized distance from the ensemble centroid to the sample[a]	Average normalized distance from cluster centroids to the sample[b]	Average improvement by the ensemble centroid over the MFE structure (%)[c]	Average improvement by cluster centroids over the ensemble centroid (%)[d]
1	29	25.07±7.60	17.61±4.05	15.95±3.40	27.30±11.68	8.87±6.01
2	16	24.90±5.98	17.84±2.86	14.74±2.38	26.67±9.96	17.05±7.25
3	28	33.84±6.73	16.28±2.43	15.10±2.17	50.44±10.03	7.08±5.18
4	27	31.77±6.46	18.84±3.63	15.51±2.95	39.21±11.65	17.30±7.34
Total	100	29.30±7.81	17.61±3.44	15.40±2.81	36.89±14.63	11.95±7.82

[a] Normalized distance from the representative structure to the sample=(average base-pair distance between the representative structure and the sample of 1000 structures)/(length of sequence)×100. The mean and standard deviation are calculated for the normalized distances of all sequences in the group.

[b] Normalized distance from cluster centroids to the sample= (average base-pair distance between a sampled structure and its closest cluster centroid for each of the 1000 structures in the sample)/(length of sequence)×100. The mean and standard deviation are calculated for the normalized distances of all sequences in the group.

[c] Improvement by ensemble centroid over MFE structure=[1−(normalized distance from the ensemble centroid to the sample)/(normalized distance from the MFE structure to the sample)]×100%. The mean and standard deviation are calculated for the improvement values of all sequences in the group.

[d] Improvement by cluster centroids over ensemble centroid=[1−(normalized distance from the cluster centroids to the sample)/(normalized distance from the ensemble centroid to the sample)]×100%. The mean and standard deviation are calculated for the improvement values of all sequences in the group.

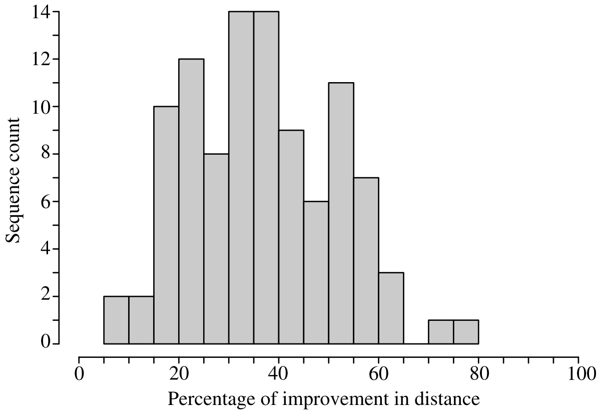

Figure 6. Distribution of the percentage improvement in average base-pair distance between the ensemble centroid and the sample over the average distance between the MFE structure and the sample.

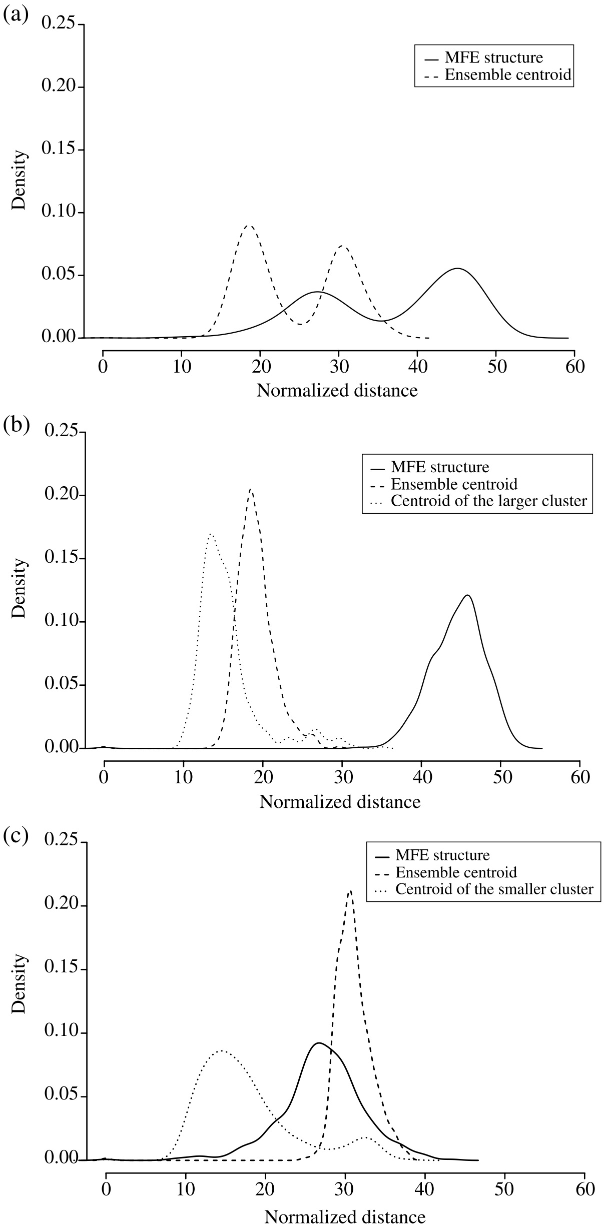

Figure 7. Overlaid distributions (smoothed histograms) of (a) the normalized distances to the entire structure sample for the mRNA of *H. sapiens* γ-glutamyltransferase-like activity 1 (Genbank accession no. NM_004121, 2414 nt, and ID 041), for the MFE structure and the ensemble centroid; (b) the normalized distances to the larger cluster for the MFE structure, the ensemble centroid and the centroid of this cluster; and (c) the normalized distances to the smaller cluster for the MFE structure, the ensemble centroid, and the centroid of this cluster.

single structure in common. The statistical reproducibility at the ensemble level is again illustrated by both the two-dimensional histograms for base-pair frequencies (Figure 8(a) and (b)) and the ensemble centroids (Figure 8(c) and (d)) for these two samples.

Each of these two non-overlapping samples contains three clusters. The reproducibility of these clusters can be appreciated from both the cluster-level two-dimensional histograms (Figure 9(a)–(f)), and the centroids of the clusters (Figure 10(a)–(f)). Furthermore, the cluster-level reproducibility is observed on a combined multi-dimensional scaling (MDS) plot for the two samples (Figure 11). These show that a cluster of significant size is represented by structures that can be completely different from one sample to another, and yet cluster-level characteristics and statistics are reproducible. We note that, because MDS is a method for displaying high-dimensional objects through reduction of dimensionality, objects that do not overlap in the original dimension may appear to overlap in the two-dimensional display.

To further assess the generality of cluster-level reproducibility, we selected nine other sequences of various lengths from the set of 100 human mRNAs, for a total of ten sequences. For each of the nine sequences, a second independent sample was generated and the clustering procedure was performed on this sample (see Supplementary Data, Table S2 for the list and the clustering results of the ten sequences). For six of the ten sequences, the clustering results from the first sample were confirmed by the second sample, in terms of the number of clusters, the two-dimensional histograms and the cluster centroids. For three of the remaining four sequences, the second sample differs from the first sample by only one or two small clusters of probability <0.04. In other words, for these three sequences, the clustering results are reproducible for all major clusters. A large difference in the number of clusters between two samples was observed for only one sequence (*H. sapiens* adducin 2 (β) transcript variant β-3a mRNA, 3014 nt in length). For this sequence, the first sample contains 11 clusters while the second sample has only two clusters. By a careful examination of the clustering output, we found that the combination of six of the 11 clusters in the first sample corresponds to cluster 1 in the second sample (Figure 12(a) and (b)), and the combination of the other five clusters in

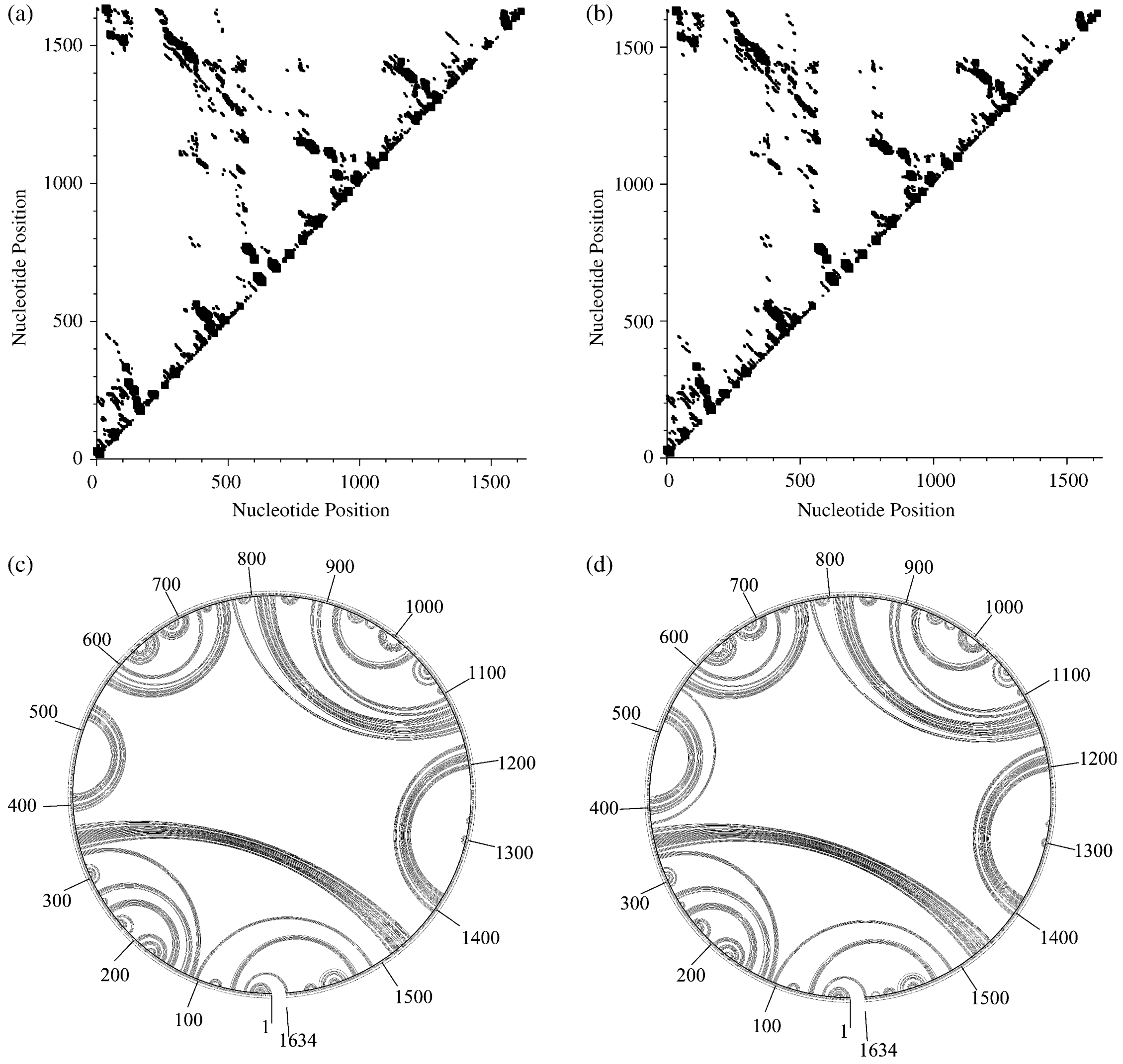

Figure 8. Two-dimensional histograms of (a) sample 1 and (b) sample 2, and circle diagrams of the ensemble centroids for (c) sample 1 and (d) sample 2, for the mRNA sequence of transcript variant Dp40 for *H. sapiens* dystrophin (GenBank accession no. NM_004019, 1634 nt, and ID 081). The two structure samples were generated independently with a sample size of 1000 structures. The two samples do not have a single structure in common.

the first sample corresponds to cluster 2 in the second sample (Figure 12(c) and (d)). This shows that the clustering output is not completely insensitive to sampling variations. When the CH index curve has multiple peaks with comparable heights, different numbers of clusters may be indicated by the CH index for different samples, simply as a result of sampling variations. However, even in such rare cases, reproducibility may still be observed with a proper combination of small clusters.

Comparison between mRNAs and structural RNAs

Because mRNAs and structural RNAs are functionally different classes of molecules, differences in folding features may be expected. Here, we investigate this problem by taking advantage of the capability to cluster RNA secondary structures. We compare the clustering results for the group of 100 human mRNAs and the clustering results for the group of 81 structural RNAs of diverse types that were used in our previous analysis on the performance of centroid structures.[35] For tRNAs, RNase P RNAs, transfer messenger RNAs, signal recognition particle (SRP) RNAs, small subunit (16 S or 16 S-like) rRNAs, large subunit (23 S or 23 S-like) rRNAs, and 5 S rRNAs, ten sequences were randomly selected for each RNA type. The group also includes nine group I introns without undetermined nucleotides and two group II introns that are available from online databases (for complete information on these sequences, see Ding *et al.*[35]).

Similar numbers of clusters

We first examine the numbers of clusters and the sizes of clusters for the two groups of RNAs. The average number of clusters for the mRNAs is 2.93,

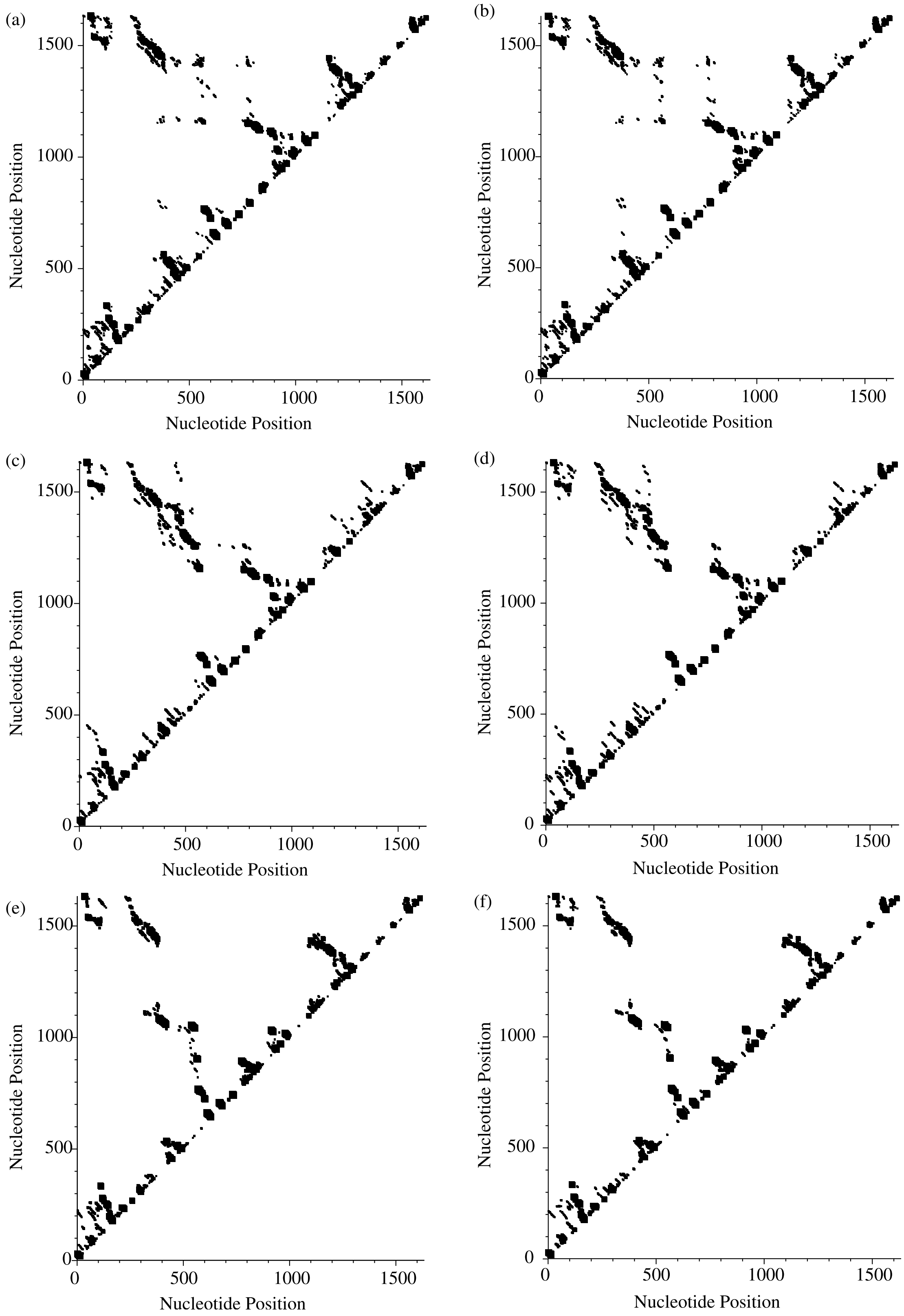

Figure 9. Cluster-level two-dimensional histograms of (a) the largest cluster from sample 1 and (b) the largest cluster from sample 2, (c) the second largest cluster from sample 1 and (d) the second largest cluster from sample 2, and (e) the smallest cluster from sample 1 and (f) the smallest cluster from sample 2, for the mRNA sequence of transcript variant Dp40 for *H. sapiens* dystrophin. The cluster sizes (i.e. estimated probabilities by sample frequencies) are (0.724, 0.169, 0.107) for sample 1, and (0.732, 0.168, 0.100) for sample 2.

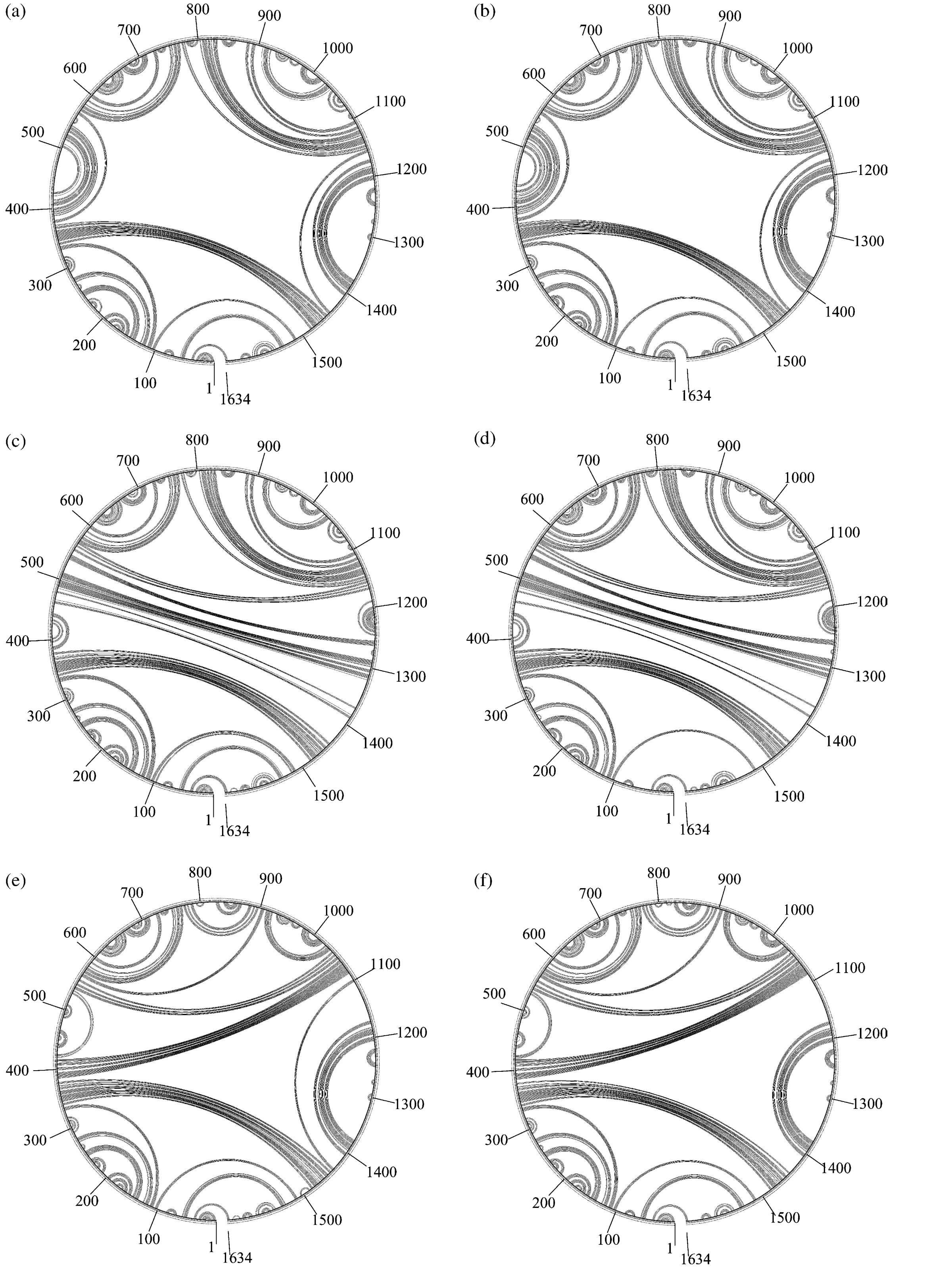

Figure 10. Circle diagrams of the cluster centroids for (a) the largest cluster from sample 1 and (b) the largest cluster from sample 2, (c) the second largest cluster from sample 1 and (d) the second largest cluster from sample 2, and (e) the smallest cluster from sample 1 and (f) the smallest cluster from (f) sample 2, for the mRNA sequence of transcript variant Dp40 for *H. sapiens* dystrophin.

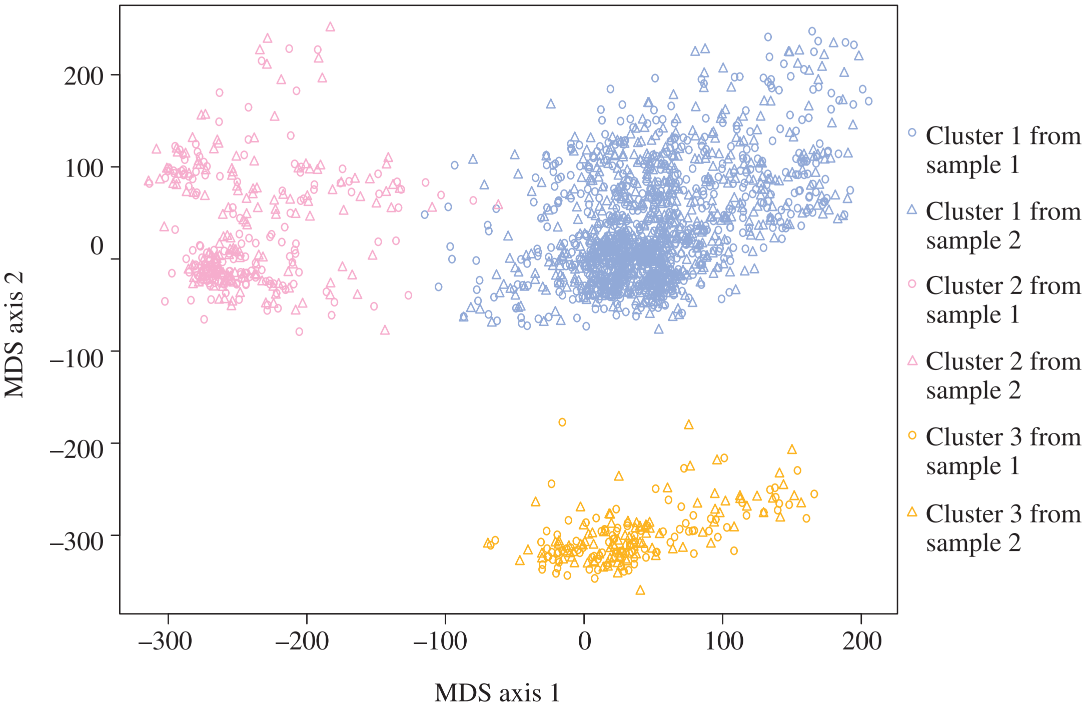

Figure 11. Two-dimensional representation of multi-dimensional scaling of the combined set of 2000 structures from sample 1 and sample 2 for the mRNA sequence of transcript variant Dp40 for *H. sapiens* dystrophin. Structures from sample 1 are plotted as circles whereas structures from sample 2 are plotted as triangles. The light blue, pink, and yellow colors represent structures from the largest cluster (cluster 1), the second largest cluster (cluster 2), and the smallest cluster (cluster 3), respectively.

whereas the average for the structural RNAs is 3.23. The histograms for the numbers of clusters show very similar distributional patterns for the two groups, as shown in Figure 4(a) for the mRNAs and Figure 13 for the structural RNAs. For the *t*-test used for testing the difference between the averages of cluster numbers of the two groups, the *p*-value is 0.2421, indicating no significant difference.

The average cluster size (probability) for all 293 clusters for the group of mRNAs is 0.3413, while the average cluster size for all 262 clusters for the structural RNA group is 0.3092. For the largest cluster for each sequence, the average size is 0.7338 for the mRNAs and 0.7143 for the structural RNAs. Results from the *t*-tests did not show a significant difference between the groups of mRNAs and structural RNAs (*p*-values are 0.2283 for all clusters, and 0.4160 for the largest cluster).

Substantial performance improvement by ensemble centroid over the MFE structure

For each of the structural RNAs, we have previously reported performance of the ensemble centroid using the structure determined by comparative sequence analysis as the standard.[35] However, such a standard is generally not available for an mRNA. To make a comparison based on a measure that is applicable to both RNA groups, we consider the average base-pair distance between a predicted structure and the sampled ensemble. Here, a predicted structure is either the MFE structure or the ensemble centroid. Performance is measured by which structure, the MFE structure or the ensemble centroid, best represents the ensemble. For the structural RNAs, the distribution of the percentage improvement in average base-pair distance between the ensemble centroid and the sample over the average distance between the MFE structure and the sample is shown in Figure 14. The distribution for the mRNAs is shown in Figure 6. The average percentage improvement of the ensemble centroid over the MFE structure is 36.89% for the mRNAs, and 22.52% for the structural RNAs. Thus, for both RNA groups, the ensemble centroid is substantially better than the MFE structure as a single representation of the Boltzmann-weighted ensemble. Furthermore, the degree of average improvement for the mRNAs is significantly higher than that for the structural RNAs (*p*-value of the *t*-test is 6.42×10^{-10}).

Both Boltzmann ensembles and clusters of structural RNAs contain more base-pairs with high frequencies

Because structural RNAs are expected to have more stable structures than do mRNAs, we compared the numbers of predicted base-pairs for the two groups, at both the ensemble level and the cluster level. Here, those base-pairs with a frequency >0.5 are considered to be high-frequency base-pairs. Because the ensemble centroid or a cluster centroid is formed by all base-pairs with

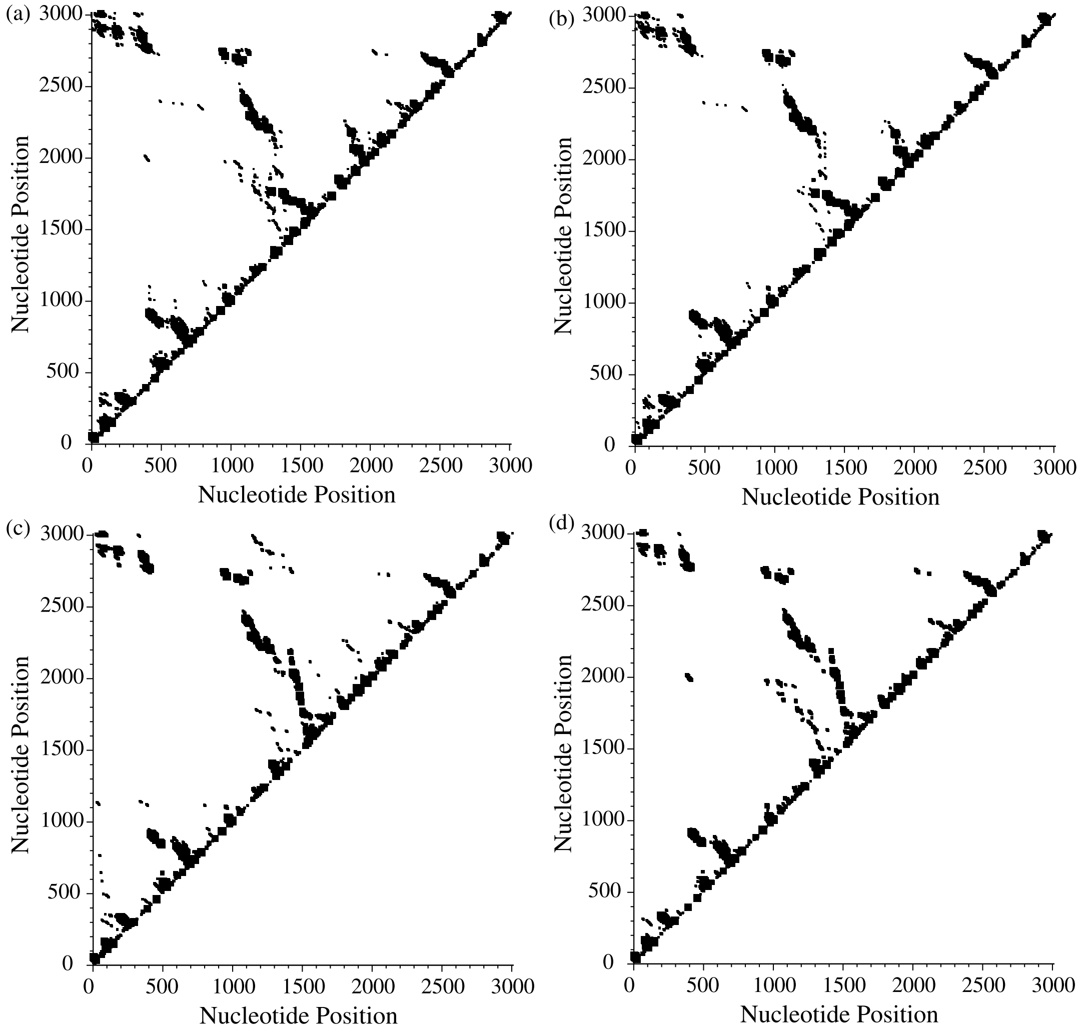

Figure 12. Cluster-level two-dimensional histograms for (a) the combined set of structures from clusters 1, 2, 3, 4, 5 and 8 from sample 1, for (b) structures in cluster 1 from sample 2, for (c) the combined set of structures from clusters 6, 7, 9, 10 and 11 from sample 1, and for (d) structures in cluster 2 from sample 2, for the mRNA of *H. sapiens* adducin 2 (β) transcript variant β-3a (Genbank accession no. NM_017483, 3014 nt, and ID 097). The cluster sizes are (0.535, 0.184, 0.117, 0.044, 0.042, 0.042, 0.013, 0.011, 0.006, 0.004, 0.002) for sample 1, and (0.935, 0.065) for sample 2. The sizes of the two combined sets for sample 1 are (0.933, 0.067).

frequency >0.5 in the sampled ensemble or the cluster, the numbers of base-pairs in the ensemble centroid and the cluster centroids can be used in the comparison of the numbers of high-frequency base-pairs. To account for differences in sequence length, the sequence length was used to normalize these base-pair numbers, as the number of base-pairs in a structure appears to grow approximately linearly with the sequence length. At the ensemble level, the group of structural RNAs was found to have, on average, 16.3% more high-frequency base-pairs per nucleotide than the group of mRNAs. The difference was statistically significant (p-value of the t-test is 5.59×10^{-7}). For cluster-level comparison, we first summed the numbers of base-pairs in all cluster centroids for a sequence. This number was divided by the total number of clusters to obtain an average number of high-frequency base-pairs per cluster centroid for the sequence, and was normalized by the sequence length. The average of

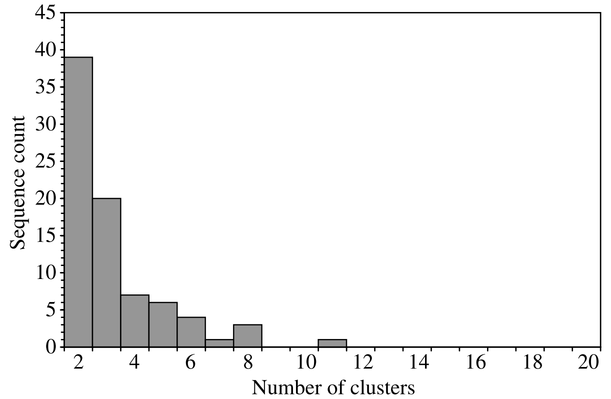

Figure 13. Distribution of the number of clusters for the set of 81 structural RNAs.

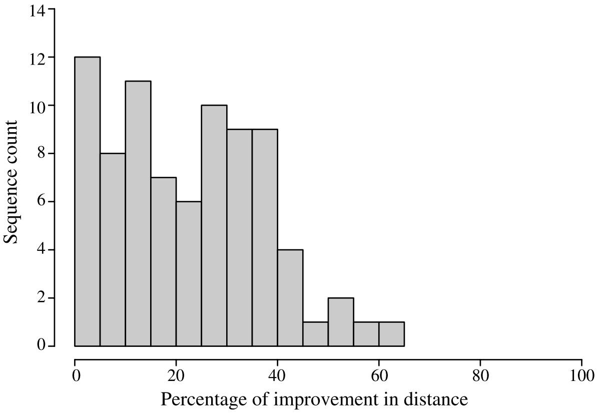

Figure 14. Distribution of the percentage improvement in average base-pair distance between the ensemble centroid and the sample over the average distance between the MFE structure and the sample for the set of 81 structural RNAs.

these normalized numbers for all sequences in a group is then computed. On average, the group of structural RNAs was found to have 14.9% more high-frequency base-pairs per nucleotide in cluster centroids than the group of mRNAs, with a highly significant p-value of 6.42×10^{-10}. Furthermore, we folded the $5'$ UTR, the $3'$ UTR and the coding sequence (CDS) of each mRNA separately, and performed calculations as described above at both the ensemble level and the cluster level for each of the three regions. The numbers are plotted in Figure 15, which shows that the observation of a significantly higher number of high-frequency base-pairs for the structural RNAs also holds for each of the three mRNA regions, with p-values of 2.29×10^{-5}($5'$ UTR), 8.62×10^{-9} ($3'$ UTR) and 1.46×10^{-6} (CDS) for the ensemble-level comparisons, and p-values of 2.92×10^{-4} ($5'$ UTR), 3.68×10^{-13} ($3'$ UTR) and 2.72×10^{-8} (CDS) for the cluster-level comparisons.

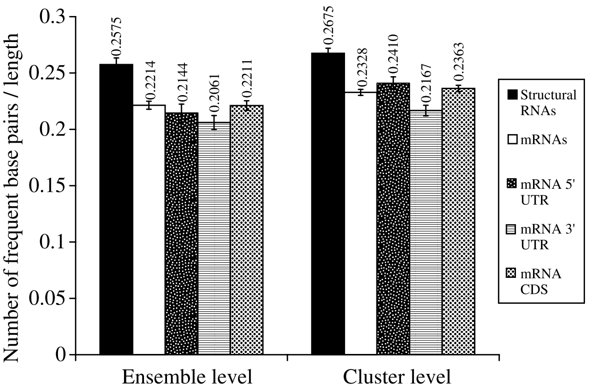

Figure 15. Bar plots of the (length-adjusted) average number of high-frequency (frequency >0.5) base-pairs at the ensemble level (the left group of bars) and the cluster level (the right group of bars), for the set of 81 structural RNAs, the set of 100 full-length human mRNAs, the set of the $5'$ UTRs of the human mRNAs, the set of the $3'$ UTRs of the human mRNAs, and the set of the coding sequences of the human mRNAs. Each full-length mRNA and the three regions are folded separately, and each of the resulting structure samples is clustered individually. The standard error bar and the average value are shown for each bar in the plot.

An alternative method for addressing the effect of sequence length and for detecting the difference between the two RNA groups is a linear regression analysis on the combined set of sequences. Here, *FBP*, the number of frequent base-pairs, is the dependent variable. The two independent variables (predictors) are *seqlen*, the sequence length, and *ind*, an indicator variable. For an mRNA sequence, $ind=0$; for a structural RNA, $ind=1$. The linear regression model is $\mathrm{FBP}=f_0+(f_1\times seqlen)+(f_2\times \mathrm{ind})$, where f_0, f_1, and f_2 are regression coefficients. For the ensemble-level comparison, both the sequence length and the indicator variable have significant p-values of 6.0×10^{-125} and 0.0019, respectively. The R^2 value of the regression is 0.9708. For a separate regression analysis for the cluster-level comparison, the p-values are also significant (2.0×10^{-136} for *seqlen* and 0.0059 for *ind*), and the R^2 value is 0.9787. We also performed regressions with $seqlen^2$ as an additional predictor. The p-value for $seqlen^2$ is 0.9945 for the ensemble-level analysis, and 0.4833 for the cluster-level analysis, indicating that normalization by sequence length is appropriate in this context.

Comparable energy gap between the MFE structure and the sampled ensemble

For structural RNAs, the energy gap between the ground state (MFE) and the rest of the Boltzmann ensemble may be greater than that of mRNAs. To examine this, we computed the energy difference between the MFE structure and the average free energy of all sampled structures. This difference was normalized by the sequence length, which has been used as the normalization factor for free energies of RNA structures.[37] The average normalized energy difference for the structural RNAs is 0.0259 kcal/mol whereas the average for the mRNAs is 0.0272 kcal/mol. No statistically significant difference was found for the two groups of RNAs (p-value of 0.1501).

Clusters are more compact for structural RNAs

Here, we examine the separation between clusters, and the compactness of clusters. These can be measured by the between-cluster sum of squares (BSS) and the within-cluster sum of squares (WSS), introduced for the calculation of the CH index. To adjust for the effect of sequence length, we normalize these sums of squares by the sequence length. The normalization here is based on the following observations: these sums of squares are computed by the base-pair distance between two structures (see Methods); the number of base-pairs in a structure grows roughly linearly with the sequence length. The average normalized BSS is

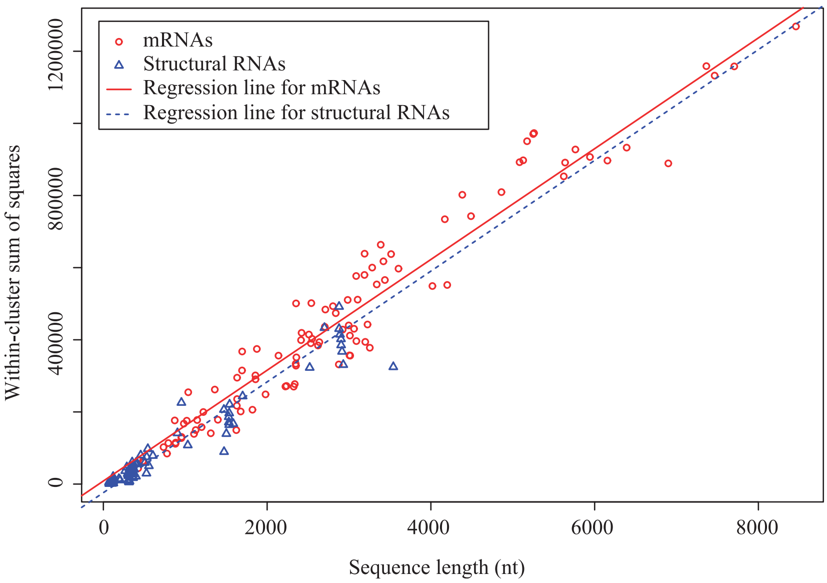

Figure 16. Scatter plot of the within-cluster sum of squares (WSS) *versus* sequence length for the combined set of 81 structural RNAs and 100 human mRNAs. Structural RNAs are represented as blue triangles, while mRNAs are represented as red circles. The regression lines determined by parameters from the linear regression model are drawn separately for the structural RNAs (blue dotted line) and the mRNAs (red continuous line). The slope for each of the two parallel regression lines is 153.55, whereas the vertical difference in WSS between the two lines for any fixed sequence length is 32,340.11. Since none of the structural RNAs is >4000 nt, we also performed a separate regression analysis on 81 structural RNA and 78 mRNA sequences that have lengths ≤ 4000 nt. This analysis returned highly similar findings (thus the results are not shown here).

51.88 for the structural RNAs, and 49.34 for the mRNAs. The average normalized WSS is 100.40 for the structural RNAs, and 153.99 for the mRNAs. For the sample size of 1000 structures, 51.88/1000 = 0.052 (49.34/1000 = 0.049) is the average distance per base between a cluster centroid and the ensemble centroid for structural RNAs (mRNAs); and 100.40/1000 = 0.100 (153.99/1000 = 0.154) is the average distance per base between a structure and the centroid of its cluster for structural RNAs (mRNAs). Results from the *t*-tests showed no significant evidence for the difference in the normalized BSS (*p*-value of 0.6592), but the normalized WSS for structural RNAs is significantly lower than that of the mRNAs (*p*-value of 2.67×10^{-15}). The latter finding suggests that clusters in the Boltzmann ensembles for structural RNAs are more compact than those for mRNAs.

To further assess the appropriateness of normalization by sequence length and the difference in the compactness of the clusters between the two RNA groups, we performed a linear regression analysis using combined data for both RNA groups. In this analysis, WSS is the dependent variable to be predicted. The sequence length, *seqlen*, is the first independent variable (predictor). An indicator variable *ind* is the second predictor. For an mRNA, *ind* = 0; for a structural RNA, *ind* = 1. The linear regression model is $\mathrm{WSS}=w_0+(w_1\times seqlen)+(w_2\times ind)$, where w_0, w_1, and w_2 are regression coefficients. From the regression analysis, $w_1=153.55$, and $w_2=-32340.11$ ($w_0=8457.70$). Both the sequence length and the indicator variable have significant *p*-values of 2.58×10^{-108} and 3.39×10^{-3}, respectively. The R^2 value of the regression is 0.9605, indicting that this linear model explains over 96% of the variations in the data. The scatter plot of the WSS *versus* sequence length is shown in Figure 16, with the regression lines drawn for both RNA groups. The regression results confirm that normalization by sequence length is appropriate, and that clusters for structural RNAs are significantly more compact than those for mRNAs.

Computational costs and software availability

The main memory requirement for the clustering procedure is the storage of the distance matrix. The computation of the centroid is a linear operation. The CPU times and memory requirements for our version of the partition function calculation, for sampling of 1000 structures, and for clustering and centroid calculation are given in Supplementary Data, Table S3 for several sequences of various lengths. Clustering features, including centroids, are available through the module Srna of the Sfold software for folding and design of nucleic acids. Sfold is available through Web servers†. Sample output for a folded sequence is also available‡.

Discussion

For over two decades, algorithms for computing the MFE structure have dominated computational approaches to prediction of RNA secondary structure. Application of this paradigm implicitly assumes that an RNA has a single, stable structure. On the other hand, it is unlikely that mRNAs exist in the unfolded state. Between these two extremes is an ensemble of an enormous number of possible structures. Here, we have explored this vast intermediate structural space through an automated procedure for identifying and representing structural clusters in the Boltzmann-weighted ensemble.

† http://sfold.wadsworth.org and http://www.bioinfo.rpi.edu/applications/sfold
‡ http://sfold.wadsworth.org/demo

The application was illustrated for a random sample of full-length human mRNAs. Our results indicate that the MFE structure often is not a good representative of the ensemble. The MFE structure is in a dominant cluster for only 29% of the sequences. We found that that the centroid structure is substantially closer to the members of the ensemble, particularly when the MFE structure is not in the largest cluster. In such cases, the centroid structures provide a substantially improved summary of this structure space. When there is no dominant structure, the ensemble of structures often shows a multimodal distribution. In these cases, no single structure provides an adequate summary of the Boltzmann-weighted space of structures, whereas centroids of a small number of clusters do.

Determination of the optimal number of clusters in clustering analysis is a difficult problem. Thus, although we have used a well-rated procedure for this purpose, the predicted number of clusters may not always be accurate. Nevertheless, the observation of competing helices and the findings of two or more modes in the distributions of distances indicate that multiple clusters of secondary structures do occur frequently. Furthermore, the result that multiple cluster centroids frequently represent the space of secondary structures substantially better than does the ensemble centroid further indicates the value of the cluster-specific centroids, even if there remains some uncertainty in the exact number of clusters. All of our results are dependent on the energy rules and parameters compiled for RNA secondary structures over the years. In fact, since the structure sampling algorithm generates a representative sample of structures from the Boltzmann ensemble, the implications of the findings presented are more a reflection of the energy rules and parameters than they are of the sampling algorithm.

Envisioning structure space from the perspective of a population of states in the Boltzmann-weighted ensemble represents a departure that is more difficult to characterize than is a single "best" structure. However, our finding that human mRNA structures populate a small number of clusters makes it possible to characterize the major features of this space with a small number of centroid structures that capture the central tendencies of these clusters. These centroids thus provide an efficient framework for representing the realities embodied in these mRNAs, according to the established secondary structure energy rules and parameters.

The observation of statistical reproducibility at the cluster-level is expected, because the structure sampling algorithm guarantees a statistical representation of the Boltzmann ensemble such that a major cluster is always represented in a sample of sufficient size. For two non-overlapping samples, there is no single common structure in the two representations of the cluster, yet the cluster characteristics and statistics are comparable for the two non-overlapping representations. In the comparison between mRNAs and structural RNAs, we observed similarity for the distribution and the average number of clusters, the energy gap between the MFE structure and the sampled ensemble, and the between-cluster sum of squares. Some of these similarities may be due to the incompleteness of the free energy model; others may be attributable to possibly conserved secondary structure elements in the coding regions of eukaryotic mRNAs[38] and structures in the UTRs. Significant differences were observed for the number of high-frequency base-pairs in the sampled ensemble and the clusters, and the compactness of the clusters. These differences are not surprising, as structures for structural RNAs are expected to be more stable than those of mRNAs. Clustering may also be useful for examining the differences between biological RNA sequences and random sequences.[39–43]

Methods

Base-pair distance

For an RNA sequence of n nucleotides, a secondary structure I can be expressed by an upper triangular matrix of base-pairing indicators $\{I_{ij}\}$, $1 \leq i < j \leq n$. $I_{ij} = 1$ if the ith base is paired with the jth base, or $I_{ij} = 0$ otherwise. The requirement of at least three unpaired intervening bases between any base-pair implies $I_{ij} = 0$ for $j = i+1$, $i+2$ and $i+3$, $1 \leq i$, $i+3 \leq n$. The indicators are not independent of each other, because they are subject to constraints. The assumption of no pseudoknots implies $I_{ij} I_{i'j'} = 0$ for $i' < i < j' < j$. Also, when base-triples are prohibited, $\sum_{1 \leq i \leq n} I_{ij} \leq 1$, and $\sum_{1 \leq j \leq n} I_{ij} \leq 1$. While the base-pair indicators are binary and under constraints, they are also coordinates in a Euclidean space of dimension $(n-1)n/2$. For two structures $I_1 = \{I_{ij}^1\}$ and $I_2 = \{I_{ij}^2\}$, we consider the following metric D_1 and the squared Euclidean distance D_2:$D_1(I_1, I_2) = \sum_{1 \leq i < j \leq n} |I_{ij}^1 - I_{ij}^2|$ and $D_2(I_1, I_2) = \sum_{1 \leq i < j \leq n} (I_{ij}^1 - I_{ij}^2)^2$. Both metrics are equal to the number of different base-pairs in I_1 and I_2. In other words, both of the metrics are the well-known base-pair distance $D(I_1, I_2)$. This interpretation does not apply to the square-root of D_2, i.e. the Euclidean distance. The discriminatory power of this distance is adequate in our context, because we are interested in comparing structures sampled for the same RNA sequence. In other circumstances, e.g. when structures of homologous sequences are under comparison, alternative metrics[44] may be more appriopriate to account for insertions and deletions. The base-pair distance also facilitates the identification of centroid structures.[35]

Sample and MFE structure

For an RNA sequence of even several hundred nucleotides, the MFE structure is highly unlikely to be observed in a statistical sample, because the Boltzmann probability of the MFE structure is very small for a sequence of moderate length. For example, for three human mRNAs of 632 nt, 1187 nt, and 2158 nt (ribosomal protein L3 (RPL3), GenBank accession no. NM_000976; *N*-acetylglucosamine kinase (NAGK), GenBank accession no. NM_017567.1; apolipoprotein L 1 (APOL1), transcript variant 3, GenBank accession no. NM_145344), the

Boltzmann probabilities of the MFE structures are 3.9336×10^{-10}, 3.5507×10^{-18}, and 5.1345×10^{-36}, respectively. Accordingly, for long RNA sequences, there is little or no overlap between two independent samples of moderate size; nevertheless, Boltzmann-weighted sampling statistics are reproducible.[34] Furthermore, regardless of sequence length, a cluster with an appreciable probability of occurrence is expected to be represented in a sample of 1000 structures. Larger samples would reveal additional clusters that are insignificant at a significance level of 0.001. Thus, for an RNA sequence, we first cluster 1000 structures sampled from the Boltzmann ensemble, and then determine the cluster for the MFE structure.

Clustering procedure

We first construct a 1000×1000 matrix of the pair-wise distances for the 1000 sampled structures. This distance matrix is used as the input for clustering. Hierarchical clustering is selected for our application, for its simplicity and its speed of computation.[45] Hierarchical clustering methods can be further divided into two subclasses: the agglomerative approach, and the divisive methods such as Diana.[45] It has been suggested that the top-down divisive method is likely to produce more sensible output if the focus is on identifying a few clusters.[46] Furthermore, among several other common clustering procedures, the Diana method has been found to be the most effective in achieving good separation in other settings.[47] Accordingly, in this study, we employed the Diana method as implemented in the R statistical package†.

Determination of number of clusters

The Diana method generates a tree structure depicting the separations of structures at every individual step. However, it does not address the problem of the optimal number of clusters. In an evaluation of 30 procedures for determining the optimal number of clusters, the Calinski & Harabasz index (CH index)[48] was the best performer.[49] The CH index is defined as $CH(k)=[B(k)/(k-1)]/[W(k)/(n_{\text{total}}-k)]$, where k is the number of clusters, n_{total} is the total number of objects to be clustered (here $n_{\text{total}}=1000$), $B(k)$ is the between-cluster sum of squares, and $W(k)$ is the within-cluster sum of squares. Thus, the CH index is analogous to the F-statistic in univariate analysis of variance.[48] The goal is to maximize $CH(k)$ over the number of clusters $k\geq2$ ($CH(k)$ is undefined for $k=1$).

As described above, a secondary structure as expressed by an upper triangular matrix is an object in a high-dimensional Euclidean space. For the calculation of the sum of squares, we can use either the average of the corresponding Euclidean coordinates for all structures in a cluster, or the centroid of the cluster.[48] We have recently introduced the notion of centroid for a set of structures.[35] For the centroid-based calculation, $B(k)=\sum_{1\leq i\leq k} n_i D(\text{EC},\text{CC}^i)$, and $W(k)=\sum_{1\leq i\leq k}\sum_{1\leq j\leq n_i} D(\text{CC}^i,\text{I}^{ij})$, where n_i is the number of structures in cluster i, $D(\text{EC}, \text{CC}^i)$ is the base-pair distance between the ensemble centroid (EC) and the centroid of cluster i (CC^i), and $\text{D}(\text{CC}^i, \text{I}^{ij})$ is the base-pair distance between the centroid of cluster i and the jth structure in this cluster. The distance calculation using averages of Euclidean coordinates is a quadratic operation, whereas the distance calculation using the centroid is a linear operation. We did not observe appreciable differences in the clustering results by the two methods. Therefore, we use centroids in the implementation of the CH index.

For all human mRNAs analyzed in this study, we calculate the CH index for k ranging from 2 to 20, because there is usually a gradual decrease in the index for k larger than 10, such that the upper bound of 20 is sufficient for finding the maxima on the CH index profile. The cluster number with the highest CH index value is the optimal number of clusters. This number is then used to determine the structural clusters, through identification of the corresponding divisive level for the clustering tree produced by Diana.

After clustering the sampled structures, we compute the MFE structure with mfold 3.1[36] for the same set of Turner thermodynamic parameters[50,51] that are currently implemented by our sampling algorithm.[34] To identify the cluster to which the MFE structure belongs, we first identify the cluster whose centroid has the shortest base-pair distance to the MFE structure. If this distance is less than or equal to the longest base-pair distance between a structure in the cluster and the cluster centroid, the MFE structure belongs to this cluster; otherwise, the MFE structure does not belong to any cluster in the sample, i.e. it is in a new cluster by itself.

Visual representation of clusters

Multidimensional scaling (MDS) is a method for a visual representation of the patterns of proximities (i.e. similarities or distances) among a set of objects.[52] It can be useful for displaying clusters of high-dimensional data in two-dimensional space, when the clusters are well-separated. The input for MDS is a distance matrix. For RNA, the same distance matrix for clustering is used for MDS. MDS is available in the R package. In addition, a cluster can be individually represented by a two-dimensional histogram[34] for displaying the frequencies of base-pairs in the cluster.

Centroid structures as ensemble and cluster representatives

Recently, we introduced the notion of centroid structure as a representative of the central tendency for any given set of structures.[35] For a set of structures, the centroid is defined as the structure in the entire structure ensemble that has the shortest total base-pair distance to all structures in the set. The centroid is referred to as the ensemble centroid when the structure set is the sampled ensemble, and it is a cluster centroid when the structure set is a cluster in the sample. We have described a procedure for centroid identification and have shown that these centroid structures are useful for improved prediction of RNA secondary structure.[35] Here, we focus on the utility of these centroids as the representatives of the Boltzmann ensemble of secondary structures.

Selection of human mRNA sequences

To select representative mRNA sequences with reliable annotation, we consider the Reference Sequence (RefSeq) database from the National Center for Biotechnology Information (NCBI)‡. The non-redundant collection of

† http://www.r-project.org/

‡ http://www.ncbi.nlm.nih.gov/RefSeq

human mRNA sequences was used as the basis for selection. Filters were used to identify records that have been reviewed by NCBI staff and collaborators, and to extract those sequences that are marked as complete on both the 5′ and the 3′ ends, i.e. full-length mRNA sequences including the 5′ UTR, the coding region and the 3′ UTR. As of 22 March 2004, only 1290 sequences satisfied the filtering criteria. Among these, 1249 sequences have length less than or equal to 9000 nt, a length that can be efficiently managed by our computer system for a large number of CPU-intensive and memory-intensive RNA folding jobs. From the 1249 mRNA sequences, 100 were drawn randomly. They range from 425 nt to 8458 nt in length, with an average of 2927 nt. These sampled mRNA sequences were given sequence identification numbers from 001 to 100 according to the order that they were drawn. A list of these 100 sequences can be found in the Supplementary Data (Table S1).

Acknowledgements

The Computational Molecular Biology and Statistics Core at the Wadsworth Center is acknowledged for providing computing resources for this work. The authors thank David Mathews for suggestions including a comparison between mRNAs and structural RNAs, and Kristin Bennett for suggesting multi-dimensional scaling for representation of RNA structure clusters. This work was supported, in part, by National Science Foundation grant DMS-0200970 and National Institutes of Health grant GM068726 (to Y.D.), and by National Institutes of Health grant HG01257 (to C.E.L.).

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.01.056

References

1. Schlax, P. J. & Worhunsky, D. J. (2003). Translational repression mechanisms in prokaryotes. *Mol. Microbiol.* **48**, 1157–1169.
2. Pelletier, J. & Sonenberg, N. (1987). The involvement of mRNA secondary structure in protein synthesis. *Biochem. Cell Biol.* **65**, 576–581.
3. Liebhaber, S. A., Cash, F. & Eshleman, S. S. (1992). Translation inhibition by an mRNA coding region secondary structure is determined by its proximity to the AUG initiation codon. *J. Mol. Biol.* **226**, 609–621.
4. Crucs, S., Chatterjee, S. & Gavis, E. R. (2000). Overlapping but distinct RNA elements control repression and activation of nanos translation. *Mol. Cell*, **5**, 457–467.
5. Zamecnik, P. C. & Stephenson, M. L. (1978). Inhibition of Rous sarcoma virus replication and cell transformation by a specific oligodeoxynucleotide. *Proc. Natl Acad. Sci. USA*, **75**, 280–284.
6. Scherer, L. J. & Rossi, J. J. (2003). Approaches for the sequence-specific knockdown of mRNA. *Nature Biotechnol.* **21**, 1457–1465.
7. Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E. & Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, **391**, 806–811.
8. Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K. & Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, **411**, 494–498.
9. Lee, R. C., Feinbaum, R. L. & Ambros, V. (1993). The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.
10. Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B. & Bartel, D. P. (2002). MicroRNAs in plants. *Genes Dev.* **16**, 1616–1626.
11. Nudler, E. & Mironov, A. S. (2004). The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.* **29**, 11–17.
12. Landick, R., Turnbough, C. L., Jr & Yanofsky (1996). Transcriptional attenuation. In Escherichia coli *and* Salmonella: *Cellular and Molecular Biology* (Neidhardt, F. C., Curtiss, R., III & Ingraham, J. L., eds), pp. 1263–1286, American Society for Microbiology, Washington, DC.
13. Merino, E. & Yanofsky, C. (2005). Transcription attenuation: a highly conserved regulatory strategy used by bacteria. *Trends Genet.* **21**, 260–264.
14. Higgins, C. F., Causton, H. C., Dance, G. S. C. & Mudd, E. A. (1993). The role of the 3′end in mRNA stability and decay. In *Control of Messenger RNA Stability* (Belasco, J. & Brawerman, G., eds), pp. 13–30, Academic Press, San Diego, CA.
15. Grzybowska, E. A., Wilczynska, A. & Siedlecki, J. A. (2001). Regulatory functions of 3′UTRs. *Biochem. Biophys. Res. Commun.* **288**, 291–295.
16. Mignone, F., Gissi, C., Liuni, S. & Pesole, G. (2002). Untranslated regions of mRNAs. *Genome Biol.* **3**. REVIEWS0004.
17. Sudarsan, N., Barrick, J. E. & Breaker, R. R. (2003). Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA*, **9**, 644–647.
18. Kubodera, T., Watanabe, M., Yoshiuchi, K., Yamashita, N., Nishimura, A., Nakai, S. *et al.* (2003). Thiamine-regulated gene expression of *Aspergillus oryzae* thiA requires splicing of the intron containing a riboswitch-like domain in the 5′-UTR. *FEBS Letters*, **555**, 516–520.
19. Kim, D. S., Gusti, V., Pillai, S. G. & Gaur, R. K. (2005). An artificial riboswitch for controlling pre-mRNA splicing. *RNA*, **11**, 1667–1677.
20. Bhalla, T., Rosenthal, J. J., Holmgren, M. & Reenan, R. (2004). Control of human potassium channel inactivation by editing of a small mRNA hairpin. *Nature Struct. Mol. Biol.* **11**, 950–956.
21. Reenan, R. (2005). Molecular determinants and guided evolution of species-specific RNA editing. *Nature*, **434**, 409–413.
22. Vickers, T. A., Wyatt, J. R. & Freier, S. M. (2000). Effects of RNA secondary structure on cellular antisense activity. *Nucl. Acids Res.* **28**, 1340–1347.
23. Zhao, J. J. & Lemke, G. (1998). Rules for ribozymes. *Mol. Cell Neurosci.* **11**, 92–97.
24. Heale, B. S., Soifer, H. S., Bowers, C. & Rossi, J. J. (2005). siRNA target site secondary structure predictions using local stable substructures. *Nucl. Acids Res.* **33**, e30.

25. Luo, K. Q. & Chang, D. C. (2004). The gene-silencing efficiency of siRNA is strongly dependent on the local structure of mRNA at the targeted region. *Biochem. Biophys. Res. Commun.* **318**, 303–310.
26. Kretschmer-Kazemi Far, R. & Sczakiel, G. (2003). The activity of siRNA in mammalian cells is related to structural target accessibility: a comparison with antisense oligonucleotides. *Nucl. Acids Res.* **31**, 4417–4424.
27. Vickers, T. A., Koo, S., Bennett, C. F., Crooke, S. T., Dean, N. M. & Baker, B. (2003). Efficient reduction of target RNAs by small interfering RNA and RNase H-dependent antisense agents. A comparative analysis. *J. Biol. Chem.* **278**, 7108–7118.
28. Bohula, E. A., Salisbury, A. J., Sohail, M., Playford, M. P., Riedemann, J., Southern, E. M. *et al*. (2003). The efficacy of small interfering RNAs targeted to the type 1 Insulin-like growth factor receptor (IGF1R) is influenced by secondary structure in the IGF1R transcript. *J. Biol. Chem.* **278**, 15991–15997.
29. Lee, N. S., Dohjima, T., Bauer, G., Li, H., Li, M. J., Ehsani, A. *et al*. (2002). Expression of small interfering RNAs targeted against HIV-1 rev transcripts in human cells. *Nature Biotechnol.* **20**, 500–505.
30. Winkler, W. C., Nahvi, A., Sudarsan, N., Barrick, J. E. & Breaker, R. R. (2003). An mRNA structure that controls gene expression by binding *S*-adenosylmethionine. *Nature Struct. Biol.* **10**, 701–707.
31. Altuvia, S., Kornitzer, D., Teff, D. & Oppemheim, A. B. (1989). Alternative mRNA structures of the cIII gene of bacteriophage λ determine the rate of its translation initiation. *J. Mol. Biol.* **210**, 265–280.
32. Christoffersen, R. E., McSwiggen, J. A. & Konings, D. (1994). Application of computational technologies to ribozyme biotechnology products. *J. Mol. Struct. (Theochem)*, **311**, 273–284.
33. Betts, L. & Spremulli, L. L. (1994). 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.
34. Ding, Y. & Lawrence, C. E. (2003). A statistical sampling algorithm for RNA secondary structure prediction. *Nucl. Acids Res.* **31**, 7280–7301.
35. Ding, Y., Chan, C. Y. & Lawrence, C. E. (2005). RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, **11**, 1157–1166.
36. Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucl. Acids Res.* **31**, 3406–3415.
37. Pervouchine, D. D., Graber, J. H. & Kasif, S. (2003). On the normalization of RNA equilibrium free energy to the length of the sequence. *Nucl. Acids Res.* **31**, e49.
38. Meyer, I. M. & Miklos, I. (2005). Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. *Nucl. Acids Res.* **33**, 6338–6348.
39. Seffens, W. & Digby, D. (1999). mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucl. Acids Res.* **27**, 1578–1584.
40. Workman, C. & Krogh, A. (1999). No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucl. Acids Res.* **27**, 4816–4822.
41. Miklos, I., Meyer, I. M. & Nagy, B. (2005). Moments of the Boltzmann distribution for RNA secondary structures. *Bull. Math. Biol.* **67**, 1031–1047.
42. Clote, P., Ferre, F., Kranakis, E. & Krizanc, D. (2005). Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, **11**, 578–591.
43. Rivas, E. & Eddy, S. R. (2000). Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–605.
44. Moulton, V., Zuker, M., Steel, M., Pointon, R. & Penny, D. (2000). Metrics on RNA secondary structures. *J. Comput. Biol.* **7**, 277–292.
45. Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.
46. Chipman, H., Hastie, T. J. & Tibshirani, R. (2003). Clustering microarray data. In *Statistical Analysis of Gene Expression Microarray Data* (Speed, T., ed.), pp. 159–200, Chapman & Hall, New York.
47. Datta, S. & Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, **19**, 459–466.
48. Calinski, R. B. & Harabasz, J. (1974). A dendrite method for cluster analysis. *Commun. Stat.* **3**, 1–27.
49. Milligan, G. W. & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159–179.
50. Xia, T., SantaLucia, J., Jr, Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X. *et al*. (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, **37**, 14719–14735.
51. Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911–940.
52. Kruskal, J. B. & Wish, M. (1977). *Multidimensional Scaling*, Sage Publications, Beverly Hills, CA.

Edited by D. E. Draper

(Received 16 December 2005; accepted 12 January 2006)
Available online 2 February 2006