Comparing RNA secondary structures using a relaxed base-pair score

PHAEDRA AGIUS,[1] KRISTIN P. BENNETT,[2] and MICHAEL ZUKER[2]
[1]Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA
[2]Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York 12180, USA

ABSTRACT

The use of free energy-based algorithms to compute RNA secondary structures produces, in general, large numbers of foldings. Recent research has addressed the problem of grouping structures into a small number of clusters and computing a representative folding for each cluster. At the heart of this problem is the need to compute a quantity that measures the difference between pairs of foldings. We introduce a new concept, the relaxed base-pair (RBP) score, designed to give a more biologically realistic measure of the difference between structures than the base-pair (BP) metric, which simply counts the number of base pairs in one structure but not the other. The degree of relaxation is determined by a single relaxation parameter, *t*. When $t = 0$, (no relaxation) our method is the same as the BP metric. At the other extreme, a very large value of *t* will give a distance of 0 for identical structures and 1 for structures that differ. Scores can be recomputed with different values of *t*, at virtually no extra computation cost, to yield satisfactory results. Our results indicate that relaxed measures give more stable and more meaningful clusters than the BP metric. We also use the RBP score to compute representative foldings for each cluster.

Keywords: RNA secondary structure; consensus structure; minimum free energy structure; medioid structure; centroid structure; cluster; *k*-means clustering; spectral clustering; base-pair metric; relaxed base-pair score; representative structure

INTRODUCTION

Secondary structure prediction for single sequences uses efficient recursive algorithms to compute minimum free energy (mfe) foldings, including samples of foldings that are close to the mfe (Zuker 1989a, 2003; Zuker et al. 1999), or else they compute base-pair probabilities (McCaskill 1990; Hofacker et al. 1994; Markham and Zuker 2008),"stochastic samples" of foldings (Ding and Lawrence 2003; Markham and Zuker 2008), and other probabilistic quantities based on a Boltzmann statistical model. At the core of these methods are the so-called "nearest-neighbor" free energy parameters estimated and extrapolated from melting experiments performed on oligoribonucleotides (Walter et al. 1994; Serra and Turner 1995; Mathews et al. 1999, 2004).

The reliability of individual secondary structures predicted by energy minimization is low in general, especially for larger RNAs (Doshi et al. 2004). Methods have been developed to mitigate these uncertainties by annotating individual base pairs in predicted foldings using the crude "p-num" reliability index in mfold (Zuker 1989a; Zuker and Jacobson 1998), or by using base-pair probabilities (Zuker and Jacobson 1998; Mathews 2004; Markham and Zuker 2008). Such annotation highlights base pairs that are more likely to be correctly predicted than others.

The methods described above do not address the problem of computing a single or a small number of representative foldings from a large sample. Several years ago, clustering methods were brought to bear on this problem (Ding et al. 2006). A distance measure between pairs of secondary structures was required, and the base-pair (BP) metric was chosen. The BP distance between two secondary structures is the total number of base pairs that occur in one structure, but not in the other. Once a distance is defined between pairs of structures, it is straightforward to compute an array of distances between all pairs of structures in a sample. Standard clustering methods such as hierarchical clustering (Johnson 1967) or neighbor joining (Saitou and Nei 1987) may then be used to partition the foldings into clusters. Once clusters are defined, one can address the problem of computing a single "representative" folding for each cluster. When energy methods are used for folding prediction, the

Abbreviations: mfe, minimum free energy; BP, base pair (in the context of base-pair metric or base-pair score); RBP, relaxed base pair (score).

Reprint requests to: Michael Zuker, Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180, USA; e-mail: zukerm@rpi.edu; fax: (518) 276-4824.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.903510.

collection of base pairs with probabilities >50% defines the "consensus structure."

While the BP metric is simple to understand, easy to compute, and has been successfully used in past clustering work (Ding et al. 2005), it has limited discriminative power. Here, we show two examples of pairs of structures where the BP metric gives a score of 35 for both cases, even though the structures in the first example are clearly more similar to one another than are the structures in the second example.

Figure 1 contains the phylogenetically derived structure (Szymanski et al. 2002) of the 5S rRNA from *Methanothermobacter thermautotrophicus* on the left and a mfe structure on the right computed by UNAFold (Markham and Zuker 2008). Of the 41 base pairs in the phylogenetic folding, 23 are also found in the mfe folding (Fig. 1, black circles), while the remaining 18 (Fig. 1, red circles) are not found in the other structure. Similarly, of the 40 base pairs (bp) in the mfe folding, 17 (Fig. 1, green circles) do not occur in the phylogenetic structure. The BP score counts the total number of red (18) and green (17) base pairs, yielding a distance of 35, even though the two structures appear rather similar to one another. Figure 2 contains the phylogenetic structure of the 5S rRNA from *Haloarcula japonica* (Szymanski et al. 2002) on the left and a consensus structure for the same RNA as computed by UNAFold (Markham and Zuker 2008). Of the 38 base pairs in the phylogenetic structure, only 14 are also in the consensus folding, leaving 24 (Fig. 2, red circles) base pairs that are not. Of the 25 bp in the consensus folding, 11 (Fig. 2, green circles) are not in the phylogenetic structure, yielding a BP distance of 24 + 11 = 35, which is the same as the distance between the two structures in Figure 1. This seems inappropriate because the two *H. japonica* structures appear much more dissimilar than the two *M. thermautotrophicus* structures.

The necessity of a more discriminative method for comparing RNA secondary structures is what motivated the development of our comparative method—the relaxed base-pair (RBP) score. The RBP score generalizes the BP metric. It has a relaxation parameter, $t \geq 0$, that indicates the degree of relaxation. We denote the RBP score between two secondary structures, S and S', by $\rho_t(S, S')$. When $t = 0$, the RBP score is the same as the BP distance. When $t = \infty$ (i.e., very large), $\rho_t(S, S') = 0$ if $S = S'$, and is 1 otherwise. Useful values of t depend on the length of the RNA and on the average number of base pairs in foldings.

RBP ILLUSTRATION

The RBP score can be explained in words with the aid of illustrations. Details appear later (Materials and Methods). A secondary structure can be represented as a "structure dot plot." A base pair between the ith and jth ribonucleotides, r_i and r_j, respectively, where $i < j$, can be plotted as a "dot" or other symbol in row i and column j of a triangular matrix. Our convention is that the i (5′) index increases top to bottom, and the j (3′) index increases left to right. The distance between two base pairs is defined to be the maximum of the 5′ index difference and the 3′ index difference. For example, the distance between $G_{23} \bullet C_{94}$ and $U_{30} \bullet A_{86}$ is 8, because 30 − 23 = 7 and 94 − 86 = 8. Given two secondary structures, S_1 and S_2, the base pairs of S_1 can be plotted with red dots and those of S_2 with green dots. When a base pair is in both structures, a black dot is used. The BP metric counts the total number of red and green dots.

The RBP score, illustrated here with the relaxation parameter $t = 1$, is less strict. If the two structures are the same, all dots are black and the RBP score is set to 0. Otherwise, choose a red or green base pair that is the farthest away from a base pair of the opposite color (including, of course, black). Exclude it from further consideration. Then place a 1 × 1 square box centered about each base pair. With the exception of the

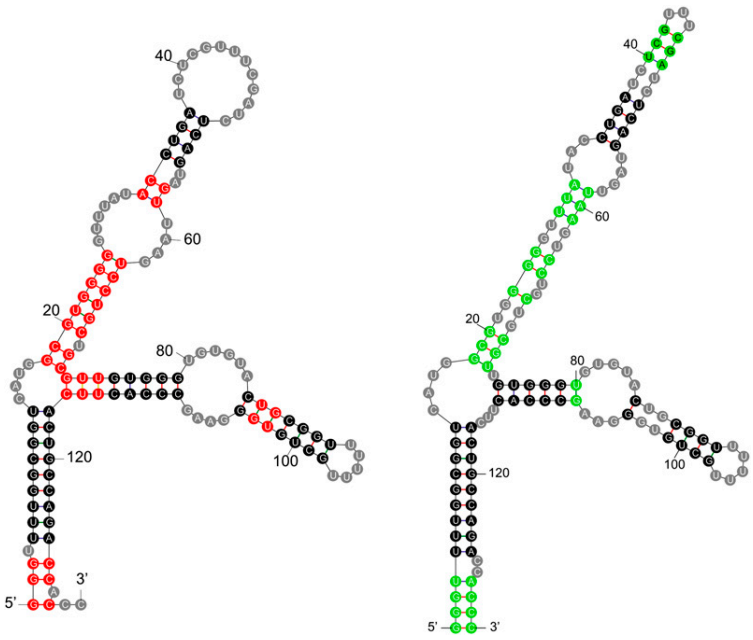

FIGURE 1. Two secondary structures for the 5S rRNA from the Archaeon, *Methanothermobacter thermautotrophicus* (strain δ-H; variant 2). (*Left*) The accepted phylogenetic structure (Szymanski et al. 2002). (*Right*) A minimum free energy (mfe) structure computed by UNAFold (Markham and Zuker 2008). Each base is drawn within a colored circle: (black circles) base pairs that are common to both structures, (red circles) base pairs in the phylogenetic structure that do not occur in the mfe structure, (green circles) base pairs in the mfe structure that are not in the phylogenetic structure, (gray circles) single-stranded bases.

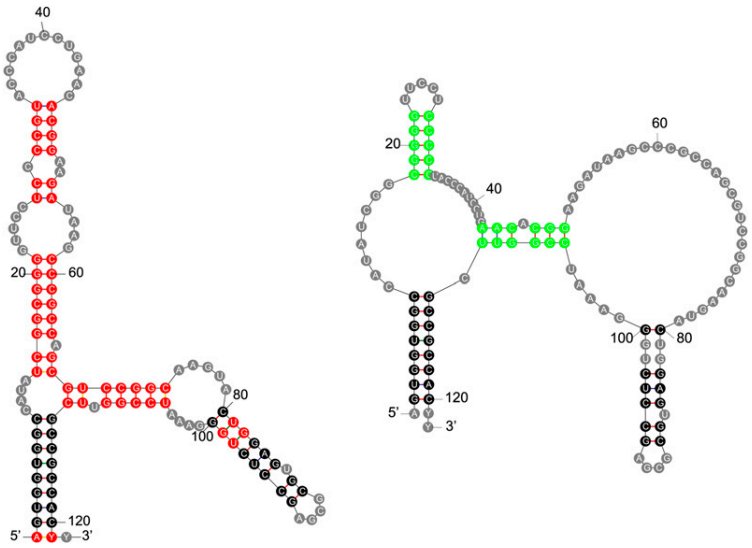

FIGURE 2. Two secondary structures for the 5S rRNA from the Archaeon, *Haloarcula japonica*. (*Left*) The accepted secondary structure computed by phylogenetic methods. (*Right*) A computed structure comprising base pairs that have probabilities >50%, computed by UNAFold (Markham and Zuker 2008). The coloring scheme is identical to that used in Figure 1, as is the total number of red (24) and green (11) circled base pairs. The BP distance between this pair of structures is 35, the same as the corresponding distance from Figure 1.

excluded base pair, if every red or green box overlaps with or touches a box of the other color, including black, along an edge or at a single vertex, then the RBP score is set to 1. If this is not possible, then another red or green base pair is excluded, and then each square box is expanded to be 2×2. This process continues until d base pairs have been excluded, and each $d \times d$ square box of the nonexcluded base pairs overlaps with or touches a box of the opposite color. At this point, the RBP score is set to d.

We illustrate the computation of the RBP score by comparing the same pairs of structures that appeared in Figures 1 and 2 using a dot plot format. In Figure 3, the squares are 3×3 and the RBP score is 3. The three green base pairs closest to the diagonal and near (40,40) have been excluded, although one of them touches a black square. However, this base pair did not touch a red or black base pair when the squares were 2×2, so excluding two base pairs was insufficient. These three excluded base pairs are drawn using a darker shade of green in Figure 1. At a square size of 3×3, all the nonexcluded red or green base pairs are close to a base pair of the opposite color (or black).

Figure 4 superimposes the secondary structure pair from Figure 2 in dot plot format. In this case, the squares are 18×18. For this very large square size, there are 10 red squares that do not touch or overlap with a green or black square, and six green squares that do not touch or overlap with a red or black square. That is, there are 16 excluded base pairs. However, if the square size is reduced to 17×17, the number of excluded base pairs rises to a number >17, yielding an RBP score of 18.

If a base pair is in one structure and not in another, the BP metric does not take into account how close it might be to a base pair in the other structure. The BP metric uses a rigid measure of closeness. Two base pairs are "close" when they are the same. The RBP score uses a relaxed degree of closeness, so that two base pairs are "close" when the distance between them is within some threshold. Moreover, the threshold of closeness is flexible. Initially, the threshold is 0. However, as base pairs are excluded, the threshold grows in proportion to the number of excluded base pairs. The reason for doing this is to achieve a balance between the degree of closeness and the number of excluded base pairs, which can be regarded as outliers.

When the t parameter is not 1, the boxes in the dot plot are $td \times td$ in size after d base pairs have been excluded. When $t = 0$, the "boxes" become single base pairs so that base pairs are excluded until the remaining ones are in both structures. This shows that the RBP score is equivalent to the BP metric when $t = 0$. The RBP score can only remain the same or decrease as t increases. If t is set to the length of the sequence (a large number), then no base pairs are excluded if the structures are identical, so the RBP score is 0. Otherwise, after one base pair is excluded, the boxes are all $t \times t$. They all intersect with one another, making the RBP score 1. That is, for t large enough, the RBP score reduces to a trivial case: 0 for identical structures, and 1 in all other cases.

COMPUTATIONAL RESULTS

Human mRNAs used for computations

Computations were performed on a selection of 13 human mRNA sequences of varying lengths (See Table 1). This collection overlaps the set used in a previous study (Ding et al. 2006). For each mRNA, 1000 foldings were computed by sampling from the Boltzmann distribution of all possible foldings. In addition, mfe structures were computed for each mRNA. All foldings were computed using the UNAFold package (Markham and Zuker 2008). Using different values of the t parameter, including $t = 0$ to reproduce the BP metric, RBP scores were computed for each of the $(1000 \times 999) \div 2$ pairs of structures for a given mRNA. Our results focus on the differences between the BP metric and RBP as seen for these mRNAs, using a variety of different computational tools.

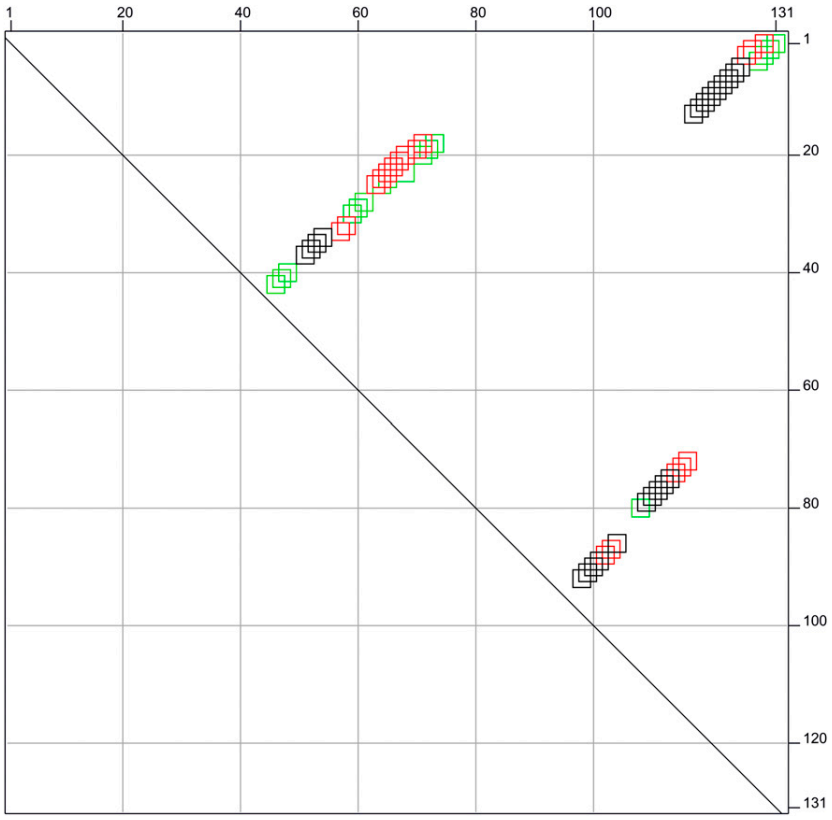

FIGURE 3. A structure dot plot that superimposes the phylogenetic structure of *Methanothermobacter thermautotrophicus 2* 5S rRNA (red and black squares) and a mfe structure for the same RNA (green and black squares). In this format, it is easy to visualize structural similarity. The squares are hollow to facilitate recognition of overlapping or touching squares. The squares are all 3 × 3.

Stability of clusters

Foldings were clustered using the spectral *k*-means clustering algorithm (Ng et al. 2001). The *k*-means clustering algorithm, the final step in the clustering process, is stochastic, meaning that cluster results differ when it is run several times on the same set of structures. Each clustering of 1000 structures was validated by taking random subsamples of size 100 and clustering them. The "Jaccard" score was used to measure the deviation of the subsample clustering to the way in which the same structures are grouped in the full 1000-structure clustering. This process was repeated a number of times to yield an average Jaccard score. Scores of 1 (maximum) indicate very stable clusterings and scores of 0 (minimum) indicate very unstable clusterings. Figure 5 shows line plots representing the clustering stability for each of the sequences. Red is used for the BP metric ($t = 0$) and blue for the RBP scores ($t > 0$). The RBP lines are almost always longer than the corresponding BP line, and are always longer for at least one value of $t > 0$. In addition, RBP scores often come close to the optimal value of 1. In some cases, the RBP scores appear "significantly" higher than the BP score. The subjective conclusion from these computations is that spectral *k*-means clustering works well with the RBP score and gives more stable clusterings when $t > 0$.

Computing representative foldings

The consensus secondary structure in Figure 2 contains, by definition, all base pairs whose probabilities in the Boltzmann distribution are >50%. It can be computed without using stochastic sampling. However, when structure samples are clustered, a consensus structure for each cluster is formed using the base pairs that occur in >50% of the foldings within the cluster. Such a structure is not contained in the sample, and it will vary somewhat each time the sampling process is repeated. When using the BP metric, the consensus structure is also a "centroid." For any pairwise scoring method, such as RBP, a centroid of a cluster of foldings is a folding such that the sum of all the scores between it and all members of the cluster is minimized. When $t = 0$, the centroid is unique and equal to the consensus structure (Ding and Lawrence 2003). This pleasing attribute is lost when $t > 0$. The consensus structure, while still of interest, is no longer a centroid. Furthermore, we are unaware of any algorithm that would compute a true centroid. It is also clear that a centroid is no longer unique. As an example, when t is very large, any structure in a cluster is a centroid for the cluster and any other structure is not.

For this reason, we compute "medioids" for clusters. A medioid is a member of a cluster that minimizes the sum of scores between it and all (other) members of the cluster. It is not unique in general.

Medioids and mfe structures

The mfe structure is by definition the most probable structure, but for the mRNAs we considered, the probability of any particular structure in the Boltzmann ensemble is so small that one would never expect to find the mfe structure in a stochastic sample of 1000 structures. However, it is valid to ask if the mfe structure belongs to a highly probable cluster of similar structures. Previous work using the BP metric (Ding et al. 2005) showed that the mfe structure could belong to highly probable or less probable clusters. For each of the 13 mRNAs we studied, the mfe structure was computed and assigned to the nearest cluster. The nearest cluster to any given structure is the cluster

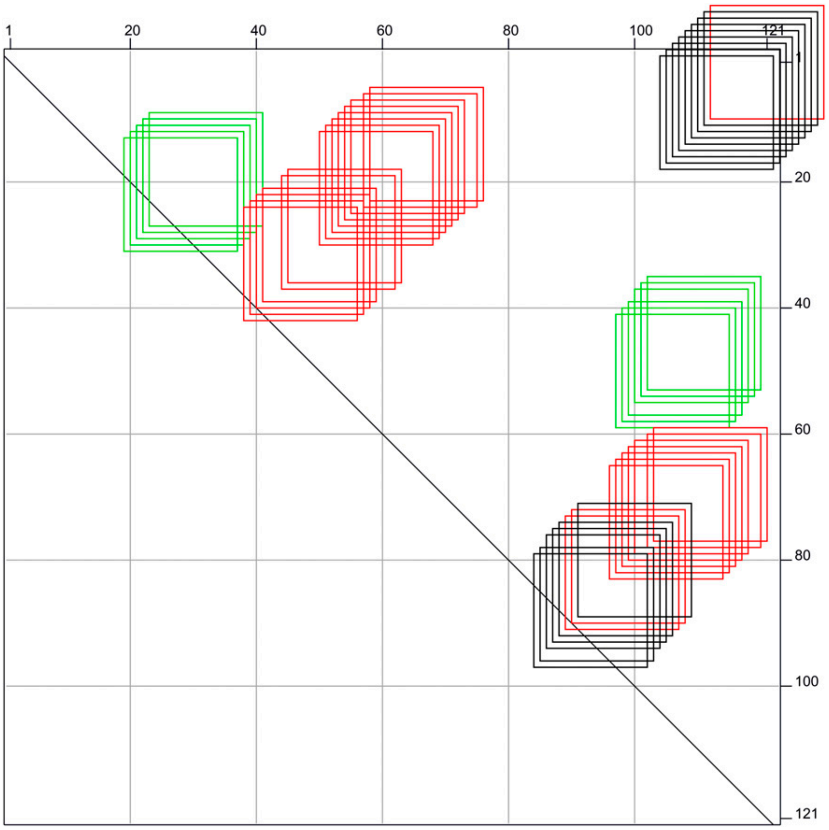

FIGURE 4. A structure dot plot that superimposes the phylogenetic structure of *Haloarcula japonica* (red and black squares) and a consensus structure for the same RNA (green and black squares).

whose medioid is closest to that structure, where structure closeness is the RBP score. The mfe structure S_{mfe} is assigned to a cluster only if $\rho_t(S_{\text{mfe}}, S_{\text{medioid}})$ does not exceed the maximal RBP score within the cluster, where S_{medioid} is the medioid structure. Otherwise, the mfe structure is considered to be an outlier sitting in a lone cluster. We define a dominant cluster to be one containing >70% of the structures, and a minor cluster to contain <10%. A large cluster is defined as nondominant, but containing at least 50% of the structures. Finally, a small cluster is nonminor and contains <50% of the structures.

The mfe structure is assigned to minor, dominant, and midsized clusters across the sequences. For some of the sequences, the assignment switches from one type of cluster to another as the *t* parameter changes. For NM_004381, the mfe structure is consistently assigned to a dominant cluster. The mfe structure is consistently assigned to a minor cluster for NM_002988. For NM_013961, the mfe is assigned to a minor cluster for six values of *t* and to a large cluster for the other six values of *t*.

Figure 6 shows bar plots for mRNAs NM_004381, NM_002988, and NM_013961. For NM_004381, the clustering is quite stable over the different values of *t*. For each value of *t*, a major and a small cluster are computed. For four values of *t*, an extra one or two very small minor clusters are computed. The major cluster varies somewhat in size and the mfe structure is consistently assigned to it. For NM_002988,

TABLE 1. Locus, gene name, and a brief description of the gene product for each of the 13 mRNA sequences used in our computations

Locus	Gene	Description	mRNA length	$\overline{M}_{\text{BP}}$	M_D
NM_130776	XAGE3	X antigen	493	138	1415
NM_002988	CCL18	C–C motif chemokine ligand 18	793	210	3879
NM_002475	MYL6B	Light chain myosin 6B	869	250	1999
NM_033142	CGB7	Chorionic gonadotropin β polypeptide 7	880	245	1599
NM_016815	GYPC	Glycophorin C	1019	294	3688
NM_013961	NRG1	Neuregulin 1	1199 (1703)	344	4222
NM_004019	DMD	Dystrophin	1634	446	7819
NM_002231	CD82	CD82 antigen	1715	524	6815
NM_000781	CYP11A1	Cytochrome P450	1821	533	7339
NM_021785	RAI2	Retinoic acid	2338	630	9606
NM_004381	ATF6B	Activating transcription factor 6β	2622	766	8686
NM_033172	B3GALT5	β 1,3-galactosyl-transferase	2711	808	12,528
NM_005656	TMPRSS2	Transmembrane serine protease	3226	948	15,330

The mRNAs are arranged in ascending order of length, an ordering that is preserved in figures that display all 13 loci. Sequence lengths vary from 493 nucleotides (nt) to 3226 nt. $\overline{M}_{\text{BP}}$ is the average number of base pairs in the $N(1000)$ foldings. If folding i has M_i base pairs, then $\overline{M}_{\text{BP}} = \sum_{i=1}^{N} M_i/N$. M_D is the number of distinct base pairs that occur in all N foldings. $M_{\text{D}} \leq N\overline{M}_{\text{BP}}$ because the first number does not count duplicate base pairs.

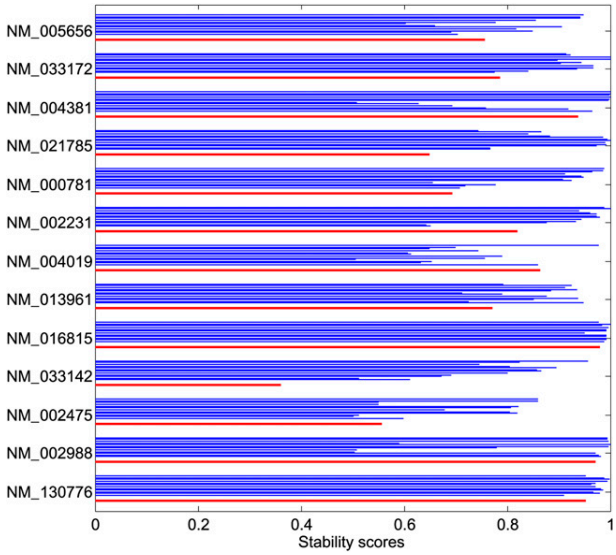

FIGURE 5. Clustering stability scores (horizontal axis) for the BP metric (t = 0, red) and RBP scores ($t > 0$, blue) for the 13 sequences listed (vertical axis, with sequences listed in order of length, shortest at the bottom). Stability scores approaching 1 indicate very stable clusters. The RBP relaxation parameters are t = 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, and 20, corresponding to consecutive blue lines, *bottom* to *top*, respectively, for each sequence. Higher t values appear to give more stable clusters.

the clustering itself changes dramatically for t = 0.1, 0.2, and 0.5. For the other values of t, a major and a minor cluster are computed, with some variation in size. The mfe structure is always assigned to the minor cluster. Examination of the underlying data reveals that the minor cluster that contains the mfe structure is stable over all values of t. It is the major cluster that is broken into multiple clusters when more than two clusters are computed. Both clustering and mfe structure assignment appear unstable for NM_013961. A detailed examination of which structures are assigned to various clusters reveals an interesting fact. The smallest minor cluster to which the mfe structure is assigned occurs when t = 0.02 and contains 48 structures. Without exception, each of these structures appear in an expanded minor cluster to which the mfe structure is assigned or else in a major cluster to which the mfe structure is assigned. In other words, there is a well-defined minor cluster that either expands into a larger minor cluster or is absorbed into a major cluster. The mfe structure is always assigned to the cluster containing these 48 structures. In other cases where mfe structure assignment appeared to be unstable, a careful analysis revealed a similar situation.

Two-dimensional cluster plots

Each structure in a sample, $\{S_1, S_2, \ldots, S_N\}$, corresponds to a vector in an N-dimensional space, where N = 1000. The jth coordinate of the ith vector is $\rho_t(S_i, S_j)$. These vectors were subjected to spectral k-means clustering (see Materials and Methods). The resulting clusters are groups of secondary structures corresponding to clusters of points in a 1000-dimensional space. We used principal component analysis (PCA) (Jolliffe 2002) to project these vectors into two dimensions in a way that minimizes the differences between the distances in the original high-dimensional space and those in the two-dimensional plots. We refer to the resulting plots as PCA plots.

Figure 7 contains pairs of PCA plots for six different mRNAs. The plots on the left are for the BP metric (t = 0), while those on the right are for t = 1, except for the last two RBP plots (j and l) where t = 0.1. We refer to the plots on the left as BP plots, and those on the right as RBP plots. The details of these plots are described by looking at them in pairs.

1. Plots A and B: The BP plot contains just two clusters, with the major one containing all but 38 of the sampled structures, while the RBP plot breaks up the sample into a dominant cluster, a small cluster, and a minor cluster. The dominant cluster in the BP plot looks like a union of two clusters (upper versus lower). We could have forced the creation of three clusters in this case, but this occurs automatically in the RBP plot. Moreover, the RBP plot shows a better separation of clusters.

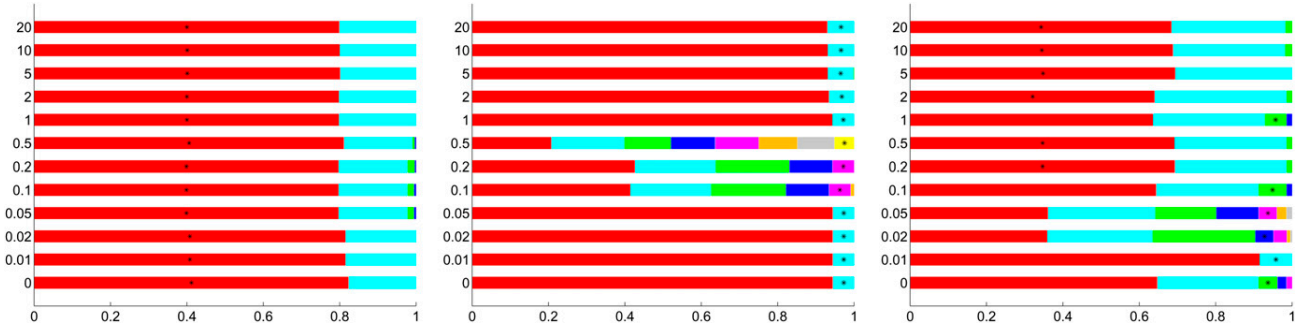

FIGURE 6. Bar plots for three of the 13 mRNAs: NM_004381 (*left*), NM_002988 (*middle*), and NM_013961 (*right*). Bar lengths indicate cluster sizes (horizontal axis), which sum to 1000. Different bars correspond to different values of the RBP relaxation parameter (t, vertical axis), which increase *bottom* to *top*. The colors are used only for visual effect. (*) The cluster, if any, to which the mfe structure is assigned.

2. Plots C and D: Both PCA plots have two clusters, and in both cases, the mfe structure is assigned to a dominant cluster. The RBP plot has compact clusters and avoids the intermingling of structures that appears in the BP plot. Note also that the mfe structure is very close to the medioid in the RBP plot, while this is not the case in the BP plot.
3. Plots E and F: The RBP plot contains a more compact dominant cluster than the BP plot. Two minor clusters from the BP plot (green and blue) are merged into a single cluster in the RBP plot (green). The smallest minor cluster in the BP plot (magenta) is retained in the RBP plot (blue). Both plots assign the mfe structure to equivalent minor clusters.
4. Plots G and H: The BP plot comprises two small clusters that are spread out and somewhat intermingled. The RBP plot contains a new minor cluster that is very well separated from the other two clusters. The large and small clusters (red and cyan, respectively) in the RBP plot are more compact than the two in the BP plot, and they are better separated. One can visualize drawing a straight line in the RBP plot that separates the two largest clusters. This cannot be done in the BP plot. The mfe structure, assigned to large clusters in both cases, appears closer to the cluster medioid in the RBP plot.
5. Plots I and J: The BP plot contains seven clusters, of which four are minor. The three large clusters and largest minor cluster seem well separated visually. The mfe structure is assigned to a minor cluster, but it nevertheless appears far from its own cluster medioid and very far from all the other cluster medioids. The RBP plot is for $t = 0.1$, which represents a very small degree of relaxation. There is a dramatic change. Only two clusters appear, well separated but with elongated "tails." The "tail" of the major cluster appears to point to the mfe structure, which is assigned to it. This dramatic change illustrates how the BP metric can give results that are unstable. The RBP plots yield two or three clusters for $t \geq 0.1$. Assignment of the mfe structure varies but a closer examination (data not shown) shows that cluster 5 of the BP plot is always completely absorbed into a cluster to which the mfe structure is assigned as t varies. As is the case with NM_013961 in Figure 6, there is a small cohort of structures that are always placed in the same cluster as t varies, and the mfe structure is always assigned to the cluster containing them, no matter what the size of the cluster.
6. Plots K and L: This is another case where changing t from 0 to 0.1 makes a dramatic difference. However, in this case, both plots contain five clusters. The point is that just a small degree of relaxation can result in greatly improved clustering. The RBP clusters, with the possible exception of the second largest, are more compact than those in the BP plot. Moreover, the RBP clusters are all well separated. The mfe structure assignment changes from minor to dominant, but in both plots, the mfe appears far from the cluster medioid. Although the number of RBP clusters increases to seven for $t = 1$ and then decreases to four for $t = 10$ (data not shown), mfe structure assignment is either to a dominant or a large cluster for $t \geq 0.05$. We consider this to be a stable assignment. The large cluster to which it is assigned in the RBP plot shown is almost dominant.

Because we used a different clustering algorithm than Ding et al. (2006), our cluster sizes and mfe structure assignments for the BP metric ($t = 0$), in general, will differ from what they reported. Table 2 summarizes the cluster sizes and mfe structure assignments given in Figure 7 and compares them with equivalent computations reported by Ding et al. (2006). We consider our ρ_t results to be "correct" in the sense of being as good as possible using our methods. Differences that stand out the most are contradictory assignments of the mfe structure for the first and last two sequences in Table 2. However, our computations clearly confirm once more that the mfe structure need not belong to a highly probable group of structures and that it can be isolated in structure space.

Medioid PCA plots

The use of the BP metric to cluster a sample of structures gives a single group of clusters and one medioid for each cluster. We saw that using the RBP score can result in the appearance of new clusters, and so we decided to compute medioids for all the different values of t used in this work and to display them all in a single PCA plot. Because these structures are medioids of clusters computed using many different values of t, we had to decide on some common scoring method to compare them. We chose the BP metric and were thus able to compute PCA plots. Figure 8 contains PCA plots of all the cluster medioids for four different mRNAs. In some cases, the medioids appear to aggregate into clusters of their own. For instance, the NM_130776 plot contains a cluster of medioids that includes a BP cluster medioid. In such a case, the BP metric was sufficient to identify a cluster, but the presence of RBP cluster medioids close to the BP medioid serves to reinforce the original BP cluster. On the other hand, isolated blue dots or small groups of blue dots far from any BP medioid (upper left of the NM_002231 plot) indicate clusters very different from what could be found using the BP distance alone.

Varying the relaxation parameter

At the beginning of this study, it was not at all clear what value or values of the relaxation parameter are appropriate for a particular RNA. Computations were performed for $t =$ 0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 50, 100, and so on. The intent was to span the range of "useful values." For

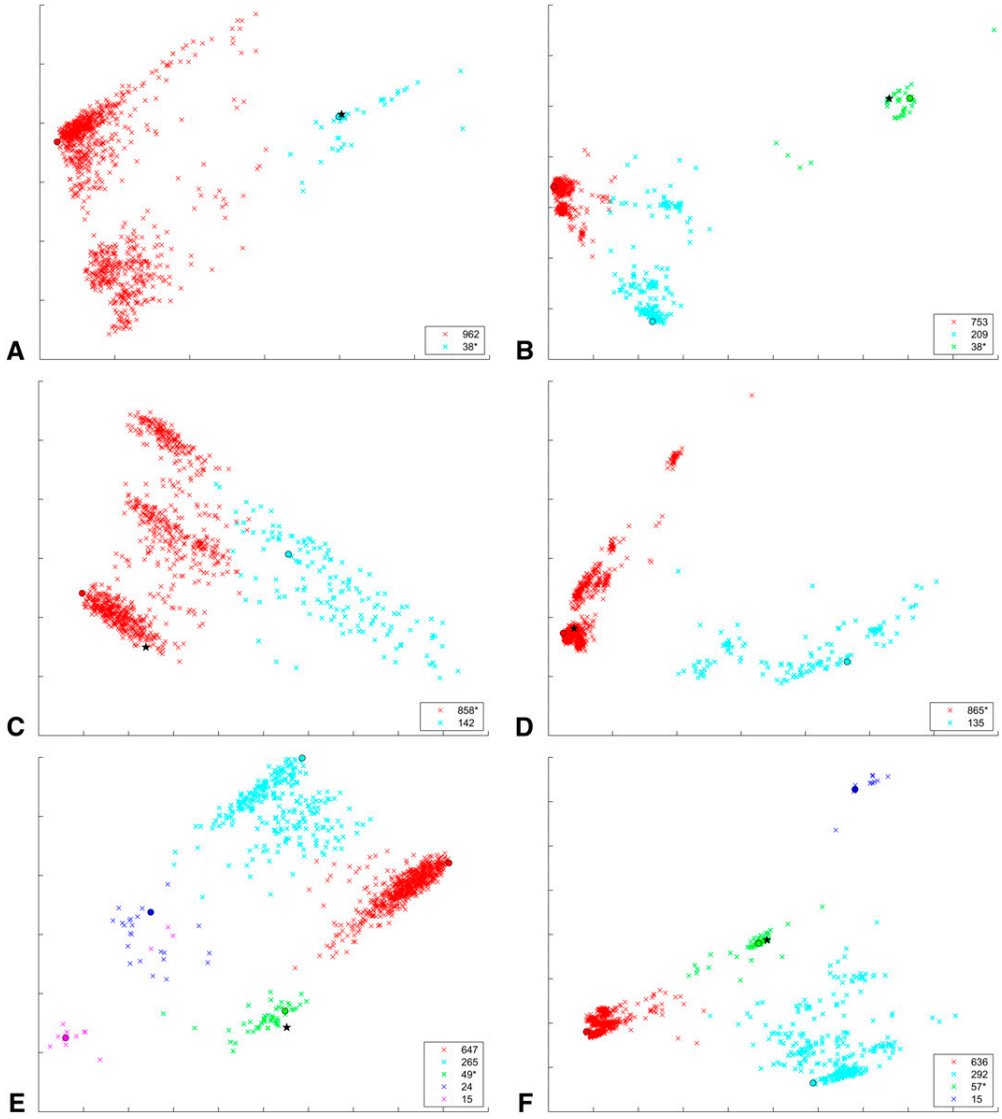

FIGURE 7. (continued on next page)

the sequences we analyzed, values >50 produced little further variation and gave trivial 0 or 1 scores for the smaller sequences.

Individual plots of ρ_t versus t may sometimes suggest appropriate values of t. Figure 9 shows the decrease of ρ_t with increasing t for the two pairs of 5S rRNAs depicted in Figures 1 and 2. Both scores are identical for the BP metric, $t = 0$, but the score for the *M. thermautotrophicus* pair initially decreases sharply with t, while the score for the *Haloarcula japonica* pair does not. For the first pair, values of $t < 0.2$, where the score is 10 and values of $t > 1$, where the score is 3, appear intuitively to be inappropriate. The score falls dramatically up to $t = 0.2$, while beyond 1 there is scarcely any further improvement. If we choose a subjective rule of not increasing t beyond the "point of diminishing returns," then $t = 1$ is an appropriate choice. For the second pair, there is a gradual decrease in the score as t increases. Visually, there appear to be no "appropriate values" for t that stand out. The subjective rule cannot be applied for all sequences.

In our quest to find appropriate values of t that could be used for all pairs of structures in a sample, we decided to investigate how RBP scores deviate from the BP metric as t increases. For this purpose, we employed the nonparametric Kendall τ statistic (see Materials and Methods). For a particular structure, S_i, the RBP scores create a ranking of all the structures in terms of increasing values of $\rho_t(S_i, S_j)$ for different js. The value of τ lies between 1 and -1, where 1 would occur if the rankings are identical and -1 indicates a total reversal. If two rankings are random and independent of one another, then the expected value of τ is 0. The Kendall τ statistic measuring the change in rankings

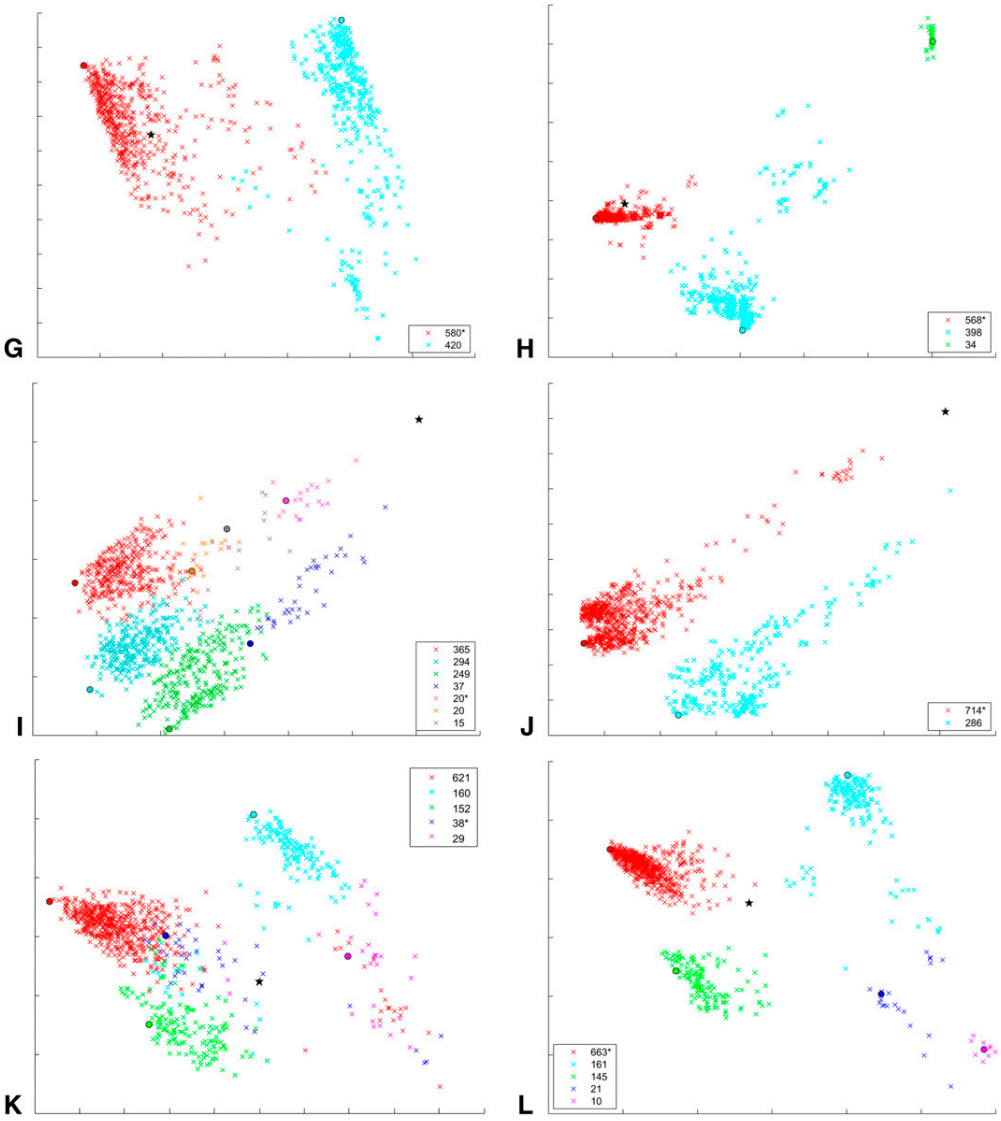

FIGURE 7. PCA plots for a variety of sequences and values of *t*. Colored crosses (×) indicate individual structures of the sample. The colors red, cyan, green, blue, magenta, orange, and gray correspond to the largest cluster, the next largest, and so on, respectively. Cluster sizes are shown in the boxed legends. (*) The mfe structure; (○ filled with the cluster color) cluster medioids. The number of clusters is computed automatically for each value of *t*. Each plot is labeled by LOCUS, *t* value: (*A*) NM_130776, 0; (*B*) NM_130776, 1; (*C*) NM_016815, 0; (*D*) NM_016815, 1; (*E*) NM_013961, 0; (*F*) NM_013961, 1; (*G*) NM_002231, 0; (*H*) NM_002231, 1; (*I*) NM_000781, 0; (*J*) NM_000781, 0.1; (*K*) NM_005656, 0; and (*L*) NM_005656, 0.1.

between ρ_0 and ρ_t is denoted by $\tau_t(i)$. Finally, τ_t was defined as the average value of $\tau_t(i)$ over all the structures. This statistic measures the overall change in rankings as t varies.

Figure 10 plots the average τ_t scores for the deviation from the BP metric as t increases from 0. They all decrease from 1 to 0 (or a slightly negative value). The decrease appears to be "somewhat linear" for middle values of t. We found no relation between sequence length and the rate of decrease.

For the ranking of 1000 items, a Kendall τ value of 0.07 or larger will occur by chance just 0.1% of the time if the rankings are totally uncorrelated (Sheskin 2007). What Figure 10 shows unambiguously is that for values of t that we would consider using in constructing clusters, say between 0.02 and 20, ρ_t is significantly different from the BP metric, but not so different as to be totally dissociated from it.

DISCUSSION

We began this work with two goals in mind. One was to group large numbers of foldings of an RNA sequence into a small number of well-separated clusters. The second was to compute an appropriate secondary structure to represent each cluster. The RBP score was designed to be an

TABLE 2. A comparison of cluster sizes and mfe structure assignment

Locus	BP	ρ_0	t	ρ_t
NM_130776	605* 392 2 1	962 38*	1	753 209 38*
NM_016815	642* 358	858* 142	1	865* 135
NM_013961	573 293 59* 43 16 14 2	647 265 49* 24 15	1	636 292 57* 15
NM_002231	980* 20	580* 420	1	568* 398 34
NM_000781	948 52 0*	365 294 249 37 20* 20 15	0.1	714* 286
NM_005656	518 482 0*	621 160 152 38* 29	0.1	663* 161 145 21 10

BP refers to published results (Ding et al. 2006) using the BP metric together with hierarchical clustering. Cluster sizes have been converted from probabilities. ρ_0 Refers to results computed in this work with the BP metric; t is the value of the relaxation parameter used in computing the results in the next column; ρ_t refers to results using the RBP score; and (*) refers to clusters in which the mfe structure is assigned (0* means unassigned).

improvement on the BP metric (Ding et al. 2005). We coupled this new scoring method with powerful clustering algorithms that have not, as far as we know, been used previously in clustering RNA secondary structures.

We were deliberate in not using hierarchical clustering (Johnson 1967; Manning et al. 2008) or the related neighbor-joining algorithm (Saitou and Nei 1987) to compute clusters of structures, even though such methods are simple to use and readily available (Felsenstein 1989). An underlying assumption of these algorithms is that the data are related by a tree that the algorithm (re)constructs as best it can. Such computations are inappropriate for our purposes. We are

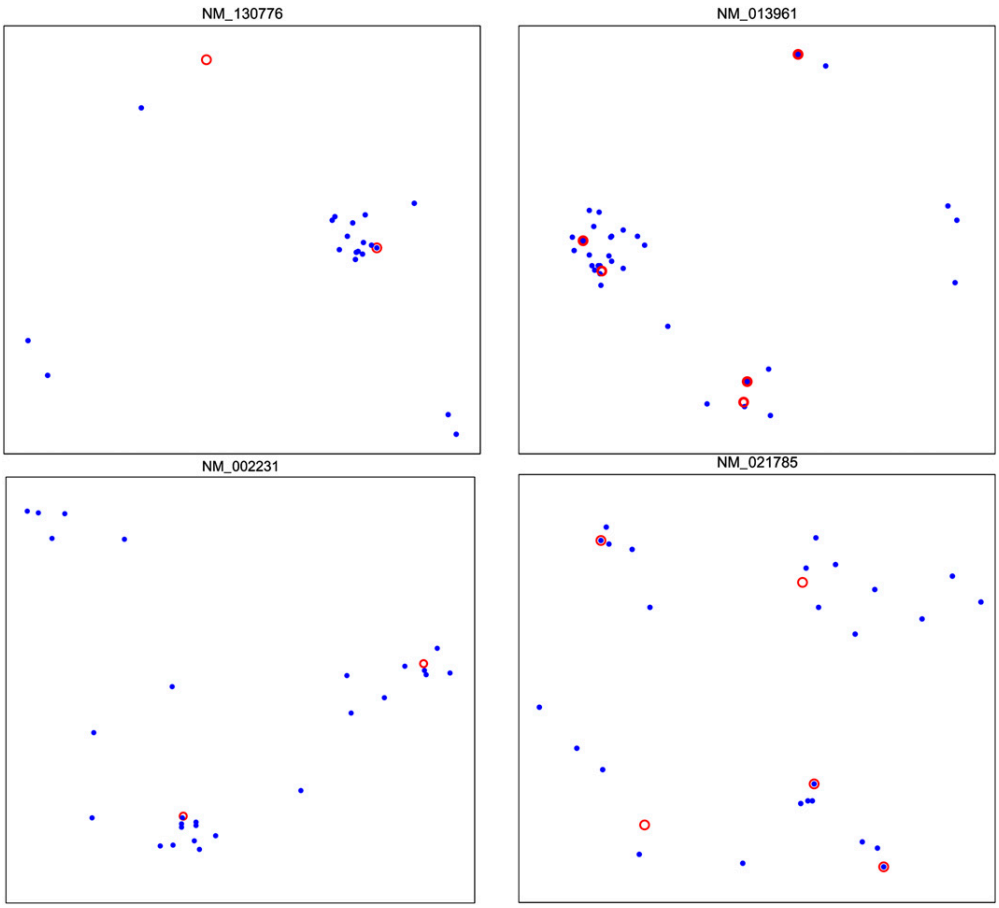

FIGURE 8. PCA plots of the cluster medioids derived by projecting medioid BP distances into two dimensions. The plots, clockwise from *top left*, are for NM_130776, NM_013961, NM_021785, and NM_002231. (Red circles) BP metric medioids, (blue dots) all the RBP medioids computed using various $t \neq 0$ values.

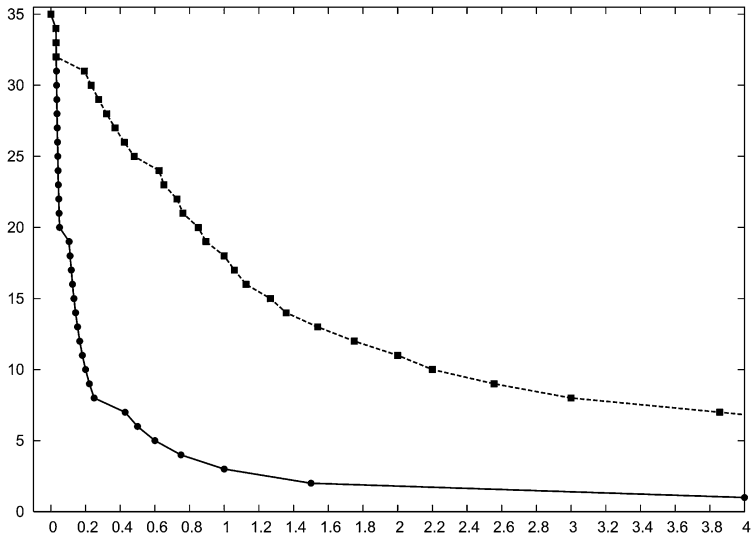

FIGURE 9. RBP scores (vertical axis) are plotted versus different relaxation parameters (horizontal axis) for two pairs of 5S rRNA secondary structures. (Solid line with circles) $\rho_t(S_{\text{phylo}}, S_{\text{mfe}})$ versus t, where S_{phylo} and S_{mfe} are the phylogenetic and mfe structures for *M. thermautotrophicus*, respectively. (Dashed line with boxes) $\rho_t(S_{\text{phylo}}, S_{\text{consen}})$ versus t, where S_{phylo} and S_{consen} are the phylogenetic and consensus structures for *H. japonica*, respectively. Values of t are automatically chosen for each plot so that every possible RBP score occurs within the range $0 \leq t \leq 4$. The RBP score reaches its minimum value of 1 for the *M. thermautotrophicus* pair when $t = 4$, so that 35 circles appear in the plot. The two plots coincide for the first four (identical) values of t.

not comparing homologous or orthologous RNA structures on different RNA sequences, for which tree construction is appropriate.

Other comparison methods

Numerous articles have been published over the years describing different approaches for comparing RNA secondary structures (Shapiro 1988; Shapiro and Zhang 1990; Moulton et al. 2000; Giegerich et al. 2004; Ding et al. 2005; Steffen et al. 2006; Liu et al. 2008). Viewing RNA secondary structures as trees and the use of tree edit algorithms to measure the difference between structure pairs was introduced ~20 yr ago (Shapiro and Zhang 1990). Tree edit algorithms tend to be slow, and in the past, computational complexity led to the use of heuristics that did not guarantee optimal solutions. In the context of RNA secondary structures, M base pairs would correspond to (at most) M nodes in a tree, and it is now known that such algorithms can run with a time and space complexity of $\mathcal{O}(M^3)$ (Demaine et al. 2007). While practical for comparing a limited number of structures, this sort of cost is prohibitive if the goal is to compute distances for all $N(N-1) \div 2$ pairs of structures in a sample of N structures. Other methods, including tree edit algorithms, are more suitable for comparing structures on different RNAs. In our opinion, the RNAshapes algorithm (Giegerich et al. 2004; Steffen et al. 2006) stands out as a useful tool for reducing huge numbers of possible structures to a smaller number of "shapes" together with actual secondary structures. At this time, it is not clear to us how to compare RNAshapes with our methods.

Computational and technical considerations

Unlike the BP metric, the RBP score is not in general a metric for $t > 0$. Although obviously symmetric, it sometimes fails the triangle inequality. For the purpose of deriving clusters and projecting into lower dimensions for display purposes, our computations were based on embedding RBP scores into high-dimensional Euclidean spaces, where the triangle inequality is satisfied.

The RNA structures in a sample will have different numbers of base pairs. Suppose that structure S_1 has M_1 base pairs and that structure S_2 has M_2 base pairs. The BP score between these two structures can be computed in $\mathcal{O}(M_1 + M_2)$ time by a simple scan of all the M_1 base pairs in the first structure, counting base pairs that do not occur in the second structure. A second scan is performed on the M_2 base pairs in the second structure. That is, computation

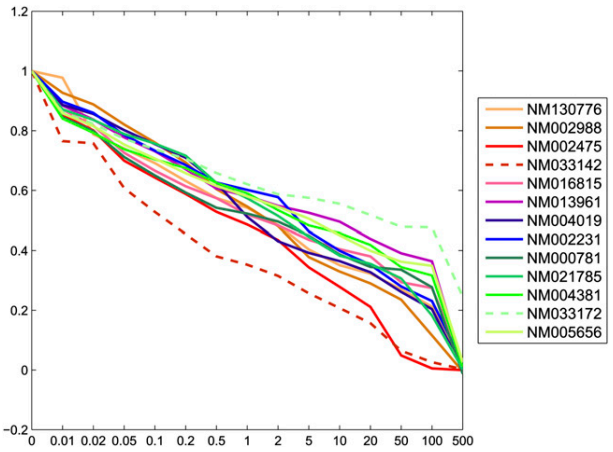

FIGURE 10. Average Kendall τ ranking scores (vertical axis) for the BP metric versus ρ_t plotted for different values of t (horizontal axis). The legend lists the sequences in order of length (shortest at the *bottom*). Dashed lines are used for the average τ_t scores for NM_033172 (pale green) and NM_033142 (red) because they exhibit, to some extent, the most extreme behavior. The scores were evaluated only for the values of t indicated along the horizontal axis, and straight lines were used to connect successive values of τ_t. No inference can be made on the behavior of τ_t between adjacent values of t.

time is linear in the total number of base pairs. This is no longer the case when computing ρ_t for $t > 0$. It appears at first that all M_1 base pairs in one structure must be compared with all M_2 base pairs of the other to compute the list of distances required to compute ρ_t. This yields a time complexity of $\mathcal{O}(M_1M_2)$. However, the base pairs of structure S_1 may be "sorted" in two dimensions by the construction of a 2-d tree using the K-d tree method (Bentley 1990). Computation time is $\mathcal{O}(M_1 \log M_1)$. For each base pair in S_2, the closest base pair in S_1 may be found in $\mathcal{O}(\log M_1)$ steps. Computing all the closest base pairs then requires $\mathcal{O}[(M_1 + M_2)\log M_1]$ steps, including the tree construction. Similarly a 2-d tree is computed for S_2, and nearest base pairs in S_2 are computed for base pairs in S_1. The entire process requires $\mathcal{O}[(M_1 + M_2)(\log M_1 + \log M_2)]$ steps. This is equivalent to $\mathcal{O}[(M_1 + M_2)\log (M_1 + M_2)]$ steps.

As stated above, there are $N(N-1) \div 2$ distinct pairs of structures in a sample of N structures, so that ρ_t must be computed for a huge number of structure pairs. In practice, this computation can become the bottleneck if done naively. Fortunately, the total number of distinct base pairs in a sample tends to be much lower than the number of structure pairs. The total number of distinct base pairs is observed to be a "small" multiple of the sequence length, where "small" ranges from 1.8 to 4.9 (3.7 ± 0.9) for the mRNAs analyzed in this work (Table 1). Thus, the entire computation could be performed by creating a 2-d tree for each structure, and searching each tree for the nearest base pair to each distinct base pair.

Choosing the relaxation parameter

In our opinion, the Kendall τ plots in Figure 10 serve to reassure us that values of t centered at 1 and varying between $1/c$ and c for $c = 20$ give useful results. One possibility for choosing t objectively would be to compute an average Kendall τ plot routinely before computing clusters. This plot would then be used to select a value of t that corresponds to a target value of τ, say somewhere between 0.5 and 0.7, thereby ensuring that the RBP scores do not deviate entirely from the BP metric, but that there is enough relaxation induced by the selected t to render RBP rankings sufficiently different from BP rankings. Although choosing a target value for τ is itself subjective, choosing t in this manner would be automatic and would not rely on human intervention.

Another solution to the problem of choosing t would be to eliminate choice by integrating the RBP score over all t. A second glance at Figure 9 suggests computing the "area under the curve" from $t = 0$ to the value of t where the RBP score becomes 1. For the *M. thermautotrophicus* pair of structures, $\rho_1 = 3$, while the area under the curve is 11.3. For the *H. japonica* pair, $\rho_1 = 18$, and the area under the curve is 107.9. The ratio of the ρ_1 scores is $3/18 \approx 0.17$ and the ratio of the two areas is $11.3/107.9 \approx 0.10$. They agree within a factor of 2.

Conclusions and future work

Our preliminary study of the novel RBP algorithm coupled with spectral k-means clustering has provided us with results that are compelling enough to instigate further investigation. With a properly chosen relaxation parameter, t, the RBP score yields a more biologically meaningful measure of the difference between two secondary structures on the same RNA. Computing clusters as described yields better defined, more compact, and better separated clusters than previously reported methods. In addition, when the mfe structure is assigned to a cluster, it is often placed close to the medioid structure when viewed in a PCA plot.

We intend to implement efficient computation of RBP scores as described above in order to incorporate clustering and medioid prediction into the UNAFold software package (Markham and Zuker 2008), into the existing mfold web server (Zuker 2003), and into future updated web servers. Consensus structures, although not centroids, are still important and will be computed for individual clusters by using base pairs that occur in over half the structures in a cluster.

We will pursue the question of computing appropriate values of t using the Kendall τ statistic and will also experiment with averaging over t. Our intuition is that computing with a single appropriate value of t will give sharper results than averaging over t. In any case, more computations need to be performed on different classes of RNA, including structural and noncoding RNA. Another approach will be to compute medioids for many values of t and cluster them. The relatively small number of medioids in Figure 8 were not clustered. This computational experiment may help to determine which clusters are genuine (stable over a range of relaxation parameters) and which are more akin to noise.

MATERIALS AND METHODS

Algorithms and statistical methods

The RBP algorithm

Denote a base pair between ribonucleotides i and j by $i{\cdot}j$. The distance between two base pairs $i{\cdot}j$ and $i'{\cdot}j'$ is defined by (Zuker 1989b)

$$\delta_{bb}(i \cdot j, i' \cdot j') = \max\{|i - i'|, |j - j'|\}.$$

The distance, δ_{bs}, between a base pair, $i{\cdot}j$, and a structure, S, is defined as the smallest distance between $i{\cdot}j$ and a base pair in S. That is,

$$\delta_{bs}(i \cdot j, S) = \min_{i' \cdot j' \in S} \{\delta_{bb}(i \cdot j, i' \cdot j')\}.$$

For two structures, S_1 and S_2, containing M_1 and M_2 base pairs, respectively, a total of $M_1 + M_2$ distances are computed: all M_1 distances between the base pairs of S_1 and the structure S_2 and all M_2 distances between the base pairs of S_2 and the structure S_1. These distances are by no means distinct. In particular, if S_1 and S_2 have M_3 base pairs in common, then 0 will occur M_3 times. These distances are sorted, with repeats, in decreasing order. We denote

these sorted distances by $\{\Delta_1, \Delta_2, \ldots, \Delta_{M_1 + M_2}\}$, where $\Delta_i \geq \Delta_j$ if $i < j$. For a relaxation parameter $t \geq 0$, we define

$$\rho_t(S_1, S_2) = \min \{m \in \mathbb{Z} \mid m \geq 0, \Delta_k \leq tm \text{ if } k > m\}.$$

This method of setting a flexible threshold of significance was inspired by the Prokhorov algorithm (Prokhorov 1956). Note that if $t = 0$, then tm is always 0, so that $\rho_0(S_1, S_2)$ is the number of nonzero distances, which is the BP score. If $t > \max \Delta_i$, then the condition is satisfied for $m = 1$, so the RBP score is 1 if the structures are different and 0 if they are identical. The sensitivity of the RBP algorithm can be manipulated using the t parameter. Changing t alters the threshold for significant differences. It is clear that for a fixed pair of structures, $\rho_t(S_1, S_2)$ is nonincreasing as t increases. Note that the computational cost for calculating the RBP scores for multiple values of t is minimal once the list of $M_1 + M_2$ distances is computed and sorted. RBP scores can easily be calculated for various t values by simply scanning the distances from largest to smallest.

Spectral k-means clustering

Clustering of secondary structures was performed using the spectral k-means clustering algorithm (Ng et al. 2001). This algorithm employs the k-means clustering algorithm (Lloyd 1982), which is stochastic. The exact algorithm is as follows:

1. Let $\{S_1, S_2, \ldots, S_N\}$ be a sample of N secondary structures on the same RNA molecule. Fix some relaxation parameter, t. For structure i, define a vector $\vec{s}_i = [\rho_t(S_i, S_1), \rho_t(S_i, S_2), \ldots, \rho_t(S_i, S_N)]$.
2. For a suitable $\sigma > 0$, which is chosen automatically by the algorithm, define an $N \times N$ affinity matrix, $A = (a_{ij})$, by

$$a_{ij} = \exp\left(-\frac{\left\|\vec{s}_i - \vec{s}_j\right\|^2}{2\sigma^2}\right) \text{ for } i \neq j \text{ and } = 0 \text{ for } i = j.$$

3. Construct a scaled matrix, $L = D^{-1/2} AD^{-1/2}$, where D is a diagonal matrix defined by $d_{ii} = \sum_{j=1}^{N} a_{ij}$.
4. If the N eigenvalues of L are ordered as $\lambda_1 \geq \lambda_2 \ldots$, then for any K between 1 and N, form an $N \times K$ matrix, X, where the K columns of X are orthonormal eigenvectors of L corresponding to the $\lambda_1, \lambda_2, \ldots, \lambda_K$, respectively.
5. Define the projected feature space, Y, to be the $N \times K$ matrix formed from X by normalizing the rows of X to unit length. Each secondary structure is now represented by a row in Y, which is a point in K-dimensional Euclidean space, $\mathbb{R}^K$.
6. Use k-means clustering on these N points in $\mathbb{R}^K$ to cluster them into K clusters.

In the above, only step (1) is specific to our application. Note that in the definition of a_{ij}, $\left\|\vec{s}_i - \vec{s}_j\right\|^2$ is used. This is $\sum_{h=1}^{N} \left[\rho_t(S_i, S_h) - \rho_t(S_j, S_h)\right]^2$, not $\rho_t(S_i, S_j)^2$. In general, the number of points to be clustered and the lengths of the vectors are not the same.

The Calinski and Harabasz (CH) index

The algorithm described above creates a specified number of clusters. It does not choose K, the number of clusters. A good clustering is one that forms closely knit clusters that are far apart from each other. The Calinski and Harabasz (CH) index (Calinski and Harabasz 1974) provides some measure of this. Ding et al. (2005) successfully used the CH index to choose K. Given N objects clustered into K groups, the CH index is defined to be

$$\mathrm{CH}(K) = \frac{B(K)}{K-1} \times \frac{N-K}{W(K)},$$

where $B(K)$ are the between-cluster sums of squares and $W(K)$ are the within-cluster sums of squares. The CH index does not have a global maximum and therefore a search space must be defined for $K \geq 2$ (the index is undefined for $K = 1$). We use the CH index to determine an optimal number of clusters within the range of 2–15.

The Jaccard similarity coefficient

The Jaccard index (Jaccard 1901) was used to measure the consistency of clusterings obtained by the k-means algorithm. Given a clustering $C(K)$ of 1000 structures into K groups, we measure the clustering stability by randomly subsampling h structures from the 1000 structures and clustering them separately into K groups to form the clustering $C'_h(K)$. This clustering is then compared to their clustering $C_h(K)$ in $C(K)$ using the Jaccard index as a measure of similarity. The Jaccard index is typically used to compare two clusterings and is defined to be the ratio $[N_{11}/(N_{11} + N_{10} + N_{01})]$, where N_{11} is the number of point pairs grouped together in both clusterings, and N_{10} and N_{01} are the number of point pairs grouped together in one clustering but not the other. Identical clusterings have a Jaccard index of 1. Therefore, using this as a measure of similarity $J[C'_h(K), C_h(K)]$ between the two clusterings for the subsampled h points, and repeating our subsampling n times, the stability score is defined to be

$$\text{Stability}[C(K)] = \frac{1}{n}\sum_{i=1}^{n} J\left[C_h^i(K), C(K)\right].$$

The resulting stability score is a number between 0 and 1, with stable clusterings having a score that approaches 1.

Kendall's τ statistic

Kendall's τ (Kendall 1975) is a nonparametric statistic that is used to compare two different rankings of the same objects. Given two rankings (orderings) of N objects, r_a and r_b, a pair of items in the rankings is defined to be concordant if both rankings agree on how the items are ordered. Otherwise, the pair is said to be discordant. The items represented in the rankings are compared pairwise and a count of the concordant pairs P and discordant pairs Q is taken. Kendall's τ is defined to be

$$\tau(r_a, r_b) = \frac{P-Q}{P+Q},$$

where $\tau \to 1$ as the rankings become identical, and $\tau = -1$ corresponds to a complete reversal of rankings. In the context of this work, there are N (1000) secondary structures, $(S_1, S_2, \ldots, S_N)$, in a sample. Any particular structure, S_i induces a ranking, $r_{i,t}$, on the $N - 1$ remaining structures determined by increasing values of $\rho_t(S_i, S_j)$. Let $\tau_t(S_i)$ be defined for the rankings $r_{i,0}$ and $r_{i,t}$. Finally, compute the average of all the $\tau_t(S_i)$,

$$\tau_t = \frac{1}{N}\sum_{i=1}^{N}\tau_{t,i}.$$

We call τ_t the rank score and use it to compare the BP metric to RBP scores as t varies. If two rankings of N objects are stochastically independent, then the associated statistic,

$$z = 3\tau\sqrt{\frac{N(N-1)}{2(2N+5)}},$$

converges to the normal distribution with mean 0 and variance 1 [N (0,1)] as $N \to \infty$ (Sheskin 2007). In practice, z is almost identical to $N(0,1)$ when $N > 40$. For $N > 40$, significance levels for Kendall's τ statistic can be computed using readily available tables for the normal distribution.

ACKNOWLEDGMENT

We thank the reviewers for their critical remarks and helpful suggestions.

Received December 24, 2009; accepted February 3, 2010.

REFERENCES

Bentley JL. 1990. K-d trees for semidynamic point sets. In *SCG '90: Proc. 6th Annual Symposium on Computational Geometry*, pp. 187–197. Association for Computing Machinery, New York.

Calinski RB, Harabasz J. 1974. A dendrite method for cluster analysis. *Comm Stat* **3:** 1–27.

Demaine ED, Mozes S, Rossman B, Weimann O. 2007. An optimal decomposition algorithm for tree edit distance. In *Automata, languages and programming,* Vol. 4596. Springer, Berlin.

Ding Y, Lawrence CE. 2003. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res* **31:** 7280–7301.

Ding Y, Chan CY, Lawrence CE. 2005. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA* **11:** 1157–1166.

Ding Y, Chan CY, Lawrence CE. 2006. Clustering of RNA secondary structures with application to messenger RNAs. *J Mol Biol* **359:** 554–571.

Doshi KJ, Cannone JJ, Cobaugh CW, Gutell RR. 2004. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics* **5:** 105. doi: 10.1186/1471-2105-5-105.

Felsenstein J. 1989. PHYLIP: Phylogeny inference package (version 3.2). *Cladistics* **5:** 164–166.

Giegerich R, Voss B, Rehmsmeier M. 2004. Abstract shapes of RNA. *Nucleic Acids Res* **32:** 4843–4851.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures (the Vienna RNA package). *Monatsh Chem* **125:** 167–188.

Jaccard P. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci s Nat* **37:** 547–579.

Johnson SC. 1967. Hierarchical clustering schemes. *Psychometrika* **2:** 241–254.

Jolliffe IT. 2002. Principal component analysis. In *Springer series in statistics*, 2nd ed. Springer, New York.

Kendall MG. 1975. *Rank correlation methods*. Griffin, London.

Liu Q, Olman V, Liu H, Ye X, Qiu S, Xu Y. 2008. RNACluster: An integrated tool for RNA secondary structure comparison and clustering. *J Comput Chem* **29:** 1517–1526.

Lloyd S. 1982. Least squares quantization in PCM. *IEEE Trans Inform Theor* **28:** 129–137.

Manning CD, Raghavan P, Schütze H. 2008. *Introduction to information retrieval*. Cambridge University Press, Cambridge, UK.

Markham NR, Zuker M. 2008. UNAFold: Software for nucleic acid folding and hybridization. In *Bioinformatics: Structure, functions and applications* (ed. JM Keith), Vol. 453, pp. 3–31. Humana Press, Totowa, NJ.

Mathews DH. 2004. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA* **10:** 1178–1190.

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

Mathews DH, Disney DH, Childs MD, Schroeder JL, Zuker M, Turner DH. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci* **101:** 7287–7292.

McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29:** 1105–1119.

Moulton V, Zuker M, Steel M, Pointon R, Penny D. 2000. Metrics on RNA secondary structures. *J Comput Biol* **7:** 277–292.

Ng AY, Jordan MI, Weiss Y. 2001. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems 14*, pp. 849–856. MIT Press, Cambridge, MA.

Prokhorov YV. 1956. Convergence of random processes and limit theorems in probability theory. *Teor Veroyatnost i Primenen* **1:** 157–214 (in Russian).

Saitou N, Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406–425.

Serra MJ, Turner DH. 1995. Predicting thermodynamic properties of RNA. *Methods Enzymol* **25:** 242–261.

Shapiro BA. 1988. An algorithm for comparing multiple RNA secondary structures. *Comput Appl Biosci* **4:** 381–393.

Shapiro BA, Zhang K. 1990. Comparing multiple RNA secondary structures using tree comparison. *Comput Appl Biosci* **6:** 309–318.

Sheskin DJ. 2007. *Handbook of parametric and nonparametric statistical procedures*, 3rd ed. CRC Press, Boca Raton, FL.

Steffen P, Voss B, Rehmsmeier M, Reeder J, Giegerich R. 2006. RNAshapes: An integrated RNA analysis package based on abstract shapes. *Bioinformatics* **22:** 500–503.

Szymanski M, Barciszewska MZ, Erdmann VA, Barciszewski J. 2002. 5S ribosomal RNA database. *Nucleic Acids Res* **30:** 176–178.

Walter AE, Turner DH, Kim J, Lyttle MH, Müller P, Mathews DH, Zuker M. 1994. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc Natl Acad Sci* **91:** 9218–9222.

Zuker M. 1989a. On finding all suboptimal foldings of an RNA molecule. *Science* **244:** 48–52.

Zuker M. 1989b. The use of dynamic programming algorithms in RNA secondary structure prediction. In *Mathematical methods for DNA sequences* (ed. MS Waterman), pp. 159–184. CRC Press, Boca Raton, FL.

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

Zuker M, Jacobson AB. 1998. Using reliability information to annotate RNA secondary structures. *RNA* **4:** 669–679.

Zuker M, Mathews DH, Turner DH. 1999. Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In *RNA biochemistry and biotechnology, NATO science partnership sub-series: 3: High technology, no. 70* (ed. J Barciszewski, BFC Clark), pp. 11–43. Kluwer Academic, Dordrecht, The Netherlands.

Comparing RNA secondary structures using a relaxed base-pair score

Phaedra Agius, Kristin P. Bennett and Michael Zuker

RNA 2010 16: 865-878 originally published online April 1, 2010
Access the most recent version at doi:10.1261/rna.903510