RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble

YE DING,[1] CHI YU CHAN,[1] and CHARLES E. LAWRENCE[1,2]

[1]Bioinformatics Center, Wadsworth Center, New York State Department of Health, Albany, New York 12208, USA
[2]Center for Computational Molecular Biology and Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912, USA

ABSTRACT

Prediction of RNA secondary structure by free energy minimization has been the standard for over two decades. Here we describe a novel method that forsakes this paradigm for predictions based on Boltzmann-weighted structure ensemble. We introduce the notion of a centroid structure as a representative for a set of structures and describe a procedure for its identification. In comparison with the minimum free energy (MFE) structure using diverse types of structural RNAs, the centroid of the ensemble makes 30.0% fewer prediction errors as measured by the positive predictive value (PPV) with marginally improved sensitivity. The Boltzmann ensemble can be separated into a small number (3.2 on average) of clusters. Among the centroids of these clusters, the "best cluster centroid" as determined by comparison to the known structure simultaneously improves PPV by 46.5% and sensitivity by 21.7%. For 58% of the studied sequences for which the MFE structure is outside the cluster containing the best centroid, the improvements by the best centroid are 62.5% for PPV and 31.4% for sensitivity. These results suggest that the energy well containing the MFE structure under the current incomplete energy model is often different from the one for the unavailable complete model that presumably contains the unique native structure. Centroids are available on the Sfold server at http://sfold.wadsworth.org.

Keywords: secondary structure prediction; centroid; Boltzmann ensemble

INTRODUCTION

RNA molecules are key elements in some of the cell's most fundamental processes, including catalysis, RNA splicing, and regulation of transcription and translation. To a large degree, the function of a structural RNA molecule is determined by its structure. Computational methods for modeling RNA secondary structure have proven to be valuable in many cases in which crystal structures are not available.

Free energy minimization is a long-established paradigm in computational structural biology that is based on the assumption that, at equilibrium, the solution to the underlying molecular folding problem is unique, and that the molecule folds into the lowest energy state. Applications of this paradigm include RNA folding (Zuker 1989), protein folding (Anfinsen 1973; Abagyan 1993), and transmembrane helix packing (Pappu et al. 1999). The prediction of RNA secondary structure has been widely applied, with good success. Efficient algorithms for computing the minimum free energy (MFE) structure and a set of suboptimal structures (Zuker and Stiegler 1981; Mathews et al. 1999, 2004) are based on free energy parameters that are estimated or extrapolated from chemical melting experiments (Xia et al. 1998; Mathews et al. 1999, 2004). An alternative approach computes all suboptimal foldings within an energy increment above the MFE (Wuchty et al. 1999). The exponential growth in the number of these foldings motivated recent development of the RNAshapes method for the efficient representation of the near optimal set (Giegerich et al. 2004). In a drastic departure from the MFE perspective, efforts have been made to characterize the ensemble of structures (McCaskill 1990; Bonhoeffer et al. 1993). Recently, we have presented an algorithm that draws samples from the ensemble of secondary structures in proportion to their Boltzmann weights (Ding and Lawrence 2003). In other words, our algorithm guarantees the generation of a statistically representative sample of the Boltzmann-

Reprint request to: Ye Ding, Bioinformatics Center, Wadsworth Center, New York State Department of Health, 150 New Scotland Avenue, Albany, NY 12208, USA; e-mail: yding@wadsworth.org; fax: (518) 402-4623; or Charles E. Lawrence, Center for Computational Molecular Biology and Division of Applied Mathematics, Brown University, 182 George Street, Providence, RI 02912, USA; e-mail: lawrence@dam.brown.edu; fax: (401) 863-1355.

Article and publication are at http://www.rnajournal.org/cgi/doi/10.1261/rna.2500605.

weighted ensemble of structures, and thus enables the calculation of sampling statistics for structural features (Ding and Lawrence 2001, 2003). In this report, we examine the utility of such samples in RNA secondary structure prediction.

In applications of the sampling algorithm, we found that there often exist distinct clusters in the Boltzmann ensemble (Ding and Lawrence 2003), with each cluster containing similar structures.

To aid in the characterization of the sampled structural space, we introduce the centroid structure as an efficient means to characterize the central tendency for a set of structures and present a procedure for its identification. We examine the predictive value of the centroid of the entire sampled ensemble, the centroid of the largest cluster, and the cluster centroid that is closest to the structure determined by comparative sequence analysis.

RESULTS

Clustering results

From online RNA databases, 81 RNA sequences from nine structural RNA classes were selected (see Materials and Methods section for sequence selection process). For each sequence, 1000 structures sampled by our algorithm (Ding and Lawrence 2003) are clustered, and then the cluster to which the MFE structure belongs is determined (see Materials and Methods section for clustering procedure). Although a structural RNA may have unique structure in solution, we have found that there exists a small number (3.2 on average) of distinct clusters of similar structures in the Boltzmann ensemble. Although the number of conformational states grows exponentially with sequence length, we found no evidence that the number of clusters increases with the length of the sequence (correlation coefficient = −0.1180, P-value = 0.294). Since the algorithm samples structures in accordance with their Boltzmann-weighted probabilities, the probability of a cluster is estimated by the frequency of structures in that cluster, i.e., the number of structures in the cluster divided by the sample size. In the case of multiple occurrences of the same structure, each occurrence is counted in the calculation. The MFE structure is present in the largest cluster for 55 of the 81 RNAs (68%). For 36 of these 55 sequences, the largest cluster dominates the structure space, with a probability of 0.7 or higher. Thus, the MFE structure is in a dominant cluster for only 44% (36/81) of the RNAs. The clustering results in Table 1 for 12 sequences exemplify possible scenarios for the cluster of the MFE structure. The cluster of the MFE structure can be the largest cluster with either a dominant probability or a moderate probability. The MFE cluster can also be in a cluster that is secondary in size, or in some cases a cluster of only small or negligible probability. The 23S rRNA sequence for *Chlamydomonas reinhardtii* (accession number X15727) presents an extreme case for which the MFE structure is not similar to any structure in the sampled ensemble. These findings suggest that the MFE structure does not always represent well the Boltzmann-weighted ensemble, thus motivating our search for more reliable representatives.

Centroid structures as representatives

As an alternative to the MFE structure, we propose the centroid structure. The centroid for a given set of struc-

TABLE 1. Clusters for sampled structures and MFE structure

RNA type	Organism	GenBank accession no.	Length (nt)	Number of clusters	Cluster probabilities[a]		
SSU (16S) rRNA	*Bordetella bronchiseptica*	U04948	1532	2	0.930*	0.070	
tRNA	*Crossostoma lacustre*	M91245	70	2	0.906*	0.094	
Group I intron	*Acanthamoeba griffini*	U02540	556	2	0.950*	0.050	
LSU (23S) rRNA	*Thermus thermophilus*	X12612	2915	3	0.339*	0.335	0.326
Group I intron	*Acanthamoeba griffini*	S81337	526	5	0.464*	0.400	0.052
					0.048	0.036	
Group II intron	*Saccharomyces cerevisiae*	AJ011856	2520	4	0.557*	0.432	
					0.007	0.004	
5S rRNA	*Agrobacterium tumefaciens*	X02627	120	2	0.591		
					0.409*		
tmRNA	*Dehalococcoides ethenogenes* strain 195	GSP[b]	352	2	0.578		
					0.422*		
RNase P	*Tarsius syrichta*	L08801	286	3	0.552		
					0.446*	0.002	
LSU (23S) rRNA	*Chlamydomonas reinhardtii*	X15727	2902	3	0.907	0.093	0.000*
RNase P	*Dermocarpa* sp.	X97396	359	2	0.803	0.197*	
RNase P	*Leptospirillum ferrooxidans*	AF296042	327	2	0.804	0.196*	

[a]Asterisk indicates cluster of MFE structure.
[b]http://tigrblast.tigr.org/ufmg/index.cgi?database=d_ethenogenes|seq.

tures is the structure in the entire structure ensemble that has the minimum total base-pair distance to the structures in the set. Thus, the centroid structure can be considered as the single structure that best represents the central tendency of the set. A centroid is refered to as the *ensemble centroid* when the set is the entire collection of structures sampled from the ensemble. A centroid of a cluster of similar structures is referred to as a *cluster centroid*. The mathematical definition of centroid structure and the derivation for its identification are presented in the Materials and Methods section. Ever since the emergence of mfold (multiple folds) it has been a common practice to report a number of suboptimal folds for predicitive purposes (Zuker 1989, 2003). Both the optimal fold and the best from a long list of suboptimal folds are of interest for performance evaluation (Mathews et al. 1999). Here we employ a similar evaluation strategy and report on the predictive performance of the ensemble centroid and that of the best of a short list of cluster centroids, i.e., the best centroid. In other words, when a reference structure is available as the standard, the best cluster centroid is defined as the cluster centroid that has the shortest base-pair distance to this known structure. Of course, just as with the best suboptimal, the identity of this best centroid cannot be determined when a reference structure is not available. In keeping with accepted practice in this field, we employed structures determined from comparative sequence analysis as the standard for comparison and for the identification of the best centroid.

Performance measures

We consider three measures for making performance comparisons between the MFE structure and centroids: base-pair distance, sensitivity, and PPV. More specifically, we compute the base-pair distance between the MFE structure and the structure determined by comparative sequence analysis and between the ensemble centroid or a cluster centroid and the structure determined by comparative sequence analysis. The sensitivity for a predicted structure is the percentage of base pairs in the structure determined by comparative sequence analysis that are also present in the predicted structure. The PPV is the percentage of base pairs in the predicted structure that are in the structure determined by comparative sequence analysis. These two complementary measures have become the standards for measuring predictive accuracy (Mathews et al. 1999; Dowell and Eddy 2004; Mathews 2004). The sensitivity focuses on predicting base pairs in the structure determined by comparative sequence analysis without regard to false positive base pair predictions, while the PPV focuses on accuracy of the predicted base pairs without regard to false negative base pairs. A perfect prediction is achieved if both the sensitivity and the PPV are 100%, in which case the two structures being compared are identical and have a distance of zero base pairs.

Centroids are closer to the structure determined by comparative sequence analysis than is the MFE structure

The ensemble centroid, the centroid of the largest cluster, and the best centroid are closer in base-pair distance to the structure determined by comparative sequence analysis than is the MFE structure for 66 (81.5%), 60 (74.1%), and 74 (91.4%) sequences, respectively. Furthermore, these centroids are either closer to the structure determined by comparative sequence analysis than is the MFE structure or are as close to the structure determined by comparative sequence analysis as is the MFE structure for 73 (90.1%), 71 (87.7%), and 80 (98.8%) of the 81 sequences, respectively.

For each sequence, the percentage of distance improvement by a centroid over the MFE structure is calculated by $[1 - D(\mathrm{C}, \mathrm{P})/D(\mathrm{M}, \mathrm{P})] \times 100\%$, where $D(\mathrm{C}, \mathrm{P})$ is the base-pair distance between the centroid and the structure determined by comparative sequence analysis and $D(\mathrm{M}, \mathrm{P})$ is the base-pair distance between the MFE structure and the structure determined by comparative sequence analysis. For each RNA type, the averaged percentage of improvement is calculated; these values are presented in Table 2. For the best cluster centroid, the average improvement is $> 19\%$ for every RNA type. For the ensemble centroid and the centroid of the largest cluster, substantial improvements are obtained, except for the SRP RNAs.

Centroids yield comparable or improved sensitivities

For each RNA type, the averaged sensitivity by the MFE structure and the average percentage of improvement in sensitivity by each of the three centroids are presented in Table 3. For the MFE structure, the ensemble centroid, and the centroid of the largest cluster, the results are comparable with marginal overall improvements by the centroids. Furthermore, the ensemble centroid and the largest cluster centroid show equal or improved sensitivity for $> 60\%$ of the sequences. For the best cluster centroid, there is an average improvement of 21.74% for all RNA types, and negative improvement is only observed for group II introns.

Centroid predictions yield fewer errors

For each RNA type, the averaged PPV by the MFE structure and the average percentage of improvement in PPV by each of the three centroids are presented in Table 4. For both the ensemble centroid and the best centroid, there is an improvement over the MFE structure, with an overall average of 30.0% and an overall average of 46.5%, respectively. For the best centroid, in particular, the PPV is either the same or improved for 79 of the 81 sequences (97.5%). For the centroid of the largest cluster, there is an improvement for seven of the nine RNA types, with an overall average improvement of 17.6%.

TABLE 2. Distance improvement by centroids over MFE structure

RNA type	Number of sequences	Average percentage of improvement in base-pair distance, with respect to MFE structure[a]		
		Ensemble centroid	Largest cluster centroid	Best cluster centroid
SSU (16S) rRNA	10	12.88 ± 4.82	9.11 ± 6.15	24.03 ± 13.28
LSU (23S) rRNA	10	19.66 ± 13.37	16.78 ± 13.57	21.52 ± 10.53
5S rRNA	10	13.18 ± 19.68	17.75 ± 28.59	27.90 ± 33.08
Group I intron	9	15.98 ± 22.09	6.10 ± 23.46	23.02 ± 7.19
Group II intron	2	21.73 ± 0.01	21.11 ± 1.40	23.03 ± 1.31
RNase P	10	8.72 ± 15.35	2.41 ± 14.06	21.46 ± 17.40
SRP RNA	10	–9.48 ± 46.06	–14.64 ± 64.23	19.55 ± 16.60
tmRNA	10	19.10 ± 19.65	12.47 ± 16.56	25.20 ± 15.72
tRNA	10	21.11 ± 17.87	11.77 ± 15.78	56.32 ± 36.35
Total	81	12.83 ± 23.37	8.07 ± 28.56	27.32 ± 22.97

[a]Distance improvement by a centroid with respect to the MFE structure = [1 – (base-pair distance between structure determined by comparative sequence analysis and centroid)/(base-pair distance between structure determined by comparative sequence analysis and MFE structure)] × 100%.

MFE predictions break down when the MFE structure is in the wrong cluster

The degree of improvement by the best centroid largely depends on the location of the MFE structure. For 34 sequences (42.0%) for which the MFE structure is in the cluster of the best centroid, the base-pair distance and the PPV are substantially improved, and the sensitivity is improved appreciably; for the other 47 sequences (58.0%) for which the MFE structure is outside the cluster of the best centroid, the improvements by the best centroid are 36.5% for base-pair distance, 62.5% for PPV, and 31.4% for sensitivity (Table 5). The latter case is illustrated by the energy landscape of the sampled ensemble and representative structures for *Agrobacterium tumefaciens* 5S rRNA (Fig. 1).

Large standard deviations due to a wide range of improvements

The unusually large standard deviations in Tables 2–5 are due to a wide range of improvements, as illustrated by Figure 2 for the best centroid. For base-pair distance, only one sequence has a negative improvement of −1.1%, and the improvement is as high as 100% (Fig. 2A). In terms of sensitivity, there are small to moderate negative improvements for 20 of the 81 sequences (24.7%), with an average of −9.6%, and the positive improvement is as high as 245.5% (Fig. 2B). For PPV, the improvement is as high as 313.6%, with only two sequences having negative improvements of −8.3% and −25.3% (Fig. 2C).

TABLE 3. Sensitivity of MFE structure and sensitivity improvement by centroids over MFE structure

RNA type	Number of sequences	Sensitivity[a] of MFE structure	Average percentage of improvement in sensitivity with respect to MFE structure[b]		
			Ensemble centroid	Largest cluster centroid	Best cluster centroid
SSU (16S) rRNA	10	49.80 ± 14.62	–4.14 ± 9.31	–2.75 ± 8.18	9.10 ± 16.51
LSU (23S) rRNA	10	35.35 ± 13.26	0.75 ± 15.41	0.42 ± 15.33	5.46 ± 12.43
5S rRNA	10	55.93 ± 24.52	2.41 ± 37.38	15.15 ± 61.20	41.81 ± 84.82
Group I intron	9	45.48 ± 19.97	6.46 ± 21.72	4.60 ± 24.19	29.06 ± 55.63
Group II intron	2	44.48 ± 6.74	0.54 ± 9.46	0.33 ± 8.08	–2.09 ± 4.66
RNase P	10	48.47 ± 18.52	–5.60 ± 13.95	–13.37 ± 33.04	4.48 ± 20.35
SRP RNA	10	76.20 ± 13.20	–1.93 ± 9.48	–0.76 ± 14.97	4.00 ± 5.73
tmRNA	10	36.16 ± 19.06	31.50 ± 81.06	24.06 ± 78.30	42.64 ± 85.71
tRNA	10	64.16 ± 17.55	–0.25 ± 3.39	9.76 ± 31.90	42.83 ± 38.90
Total	81	51.34 ± 21.29	3.54 ± 33.73	4.53 ± 39.81	21.74 ± 50.24

[a]Sensitivity = (number of base pairs in common between structure determined by comparative sequence analysis and predicted structure)/(number of base pairs in structure determined by comparative sequence analysis) × 100%.
[b]Sensitivity improvement by a centroid with respect to the MFE structure = [(sensitivity of centroid)/(sensitivity of MFE structure) – 1] × 100%.

TABLE 4. Positive predictive values (PPV) for MFE structure and PPV improvement by centroids over MFE structure

RNA type	Number of sequences	PPV[a] of MFE structure	Average percentage of improvement in PPV with respect to MFE structure[b]		
			Ensemble centroid	Largest cluster centroid	Best cluster centroid
SSU (16S) rRNA	10	48.10 ± 14.37	14.08 ± 10.67	9.54 ± 11.17	28.31 ± 22.97
LSU (23S) rRNA	10	33.68 ± 13.80	43.86 ± 52.10	36.84 ± 47.51	46.37 ± 43.28
5S rRNA	10	59.78 ± 25.76	26.52 ± 52.76	21.60 ± 57.60	51.42 ± 88.62
Group I intron	9	37.95 ± 20.81	36.37 ± 44.36	20.04 ± 32.99	56.50 ± 57.24
Group II intron	2	29.31 ± 26.14	33.33 ± 4.37	31.55 ± 1.26	32.84 ± 3.09
RNase P	10	42.89 ± 16.49	14.16 ± 18.84	–3.86 ± 37.00	27.99 ± 27.41
SRP RNA	10	77.59 ± 15.83	1.69 ± 13.23	–0.48 ± 15.64	9.63 ± 12.99
tmRNA	10	30.28 ± 19.83	76.78 ± 116.97	39.57 ± 91.33	95.55 ± 122.81
tRNA	10	55.15 ± 20.34	26.49 ± 36.50	15.29 ± 30.62	60.07 ± 44.22
Total	81	47.84 ± 23.28	30.00 ± 55.19	17.64 ± 46.87	46.51 ± 64.02

[a]Positive predictive value (PPV) = (number of common base pairs between structure determined by comparative sequence analysis and predicted structure)/(number of base pairs in predicted structure) × 100%.
[b]PPV improvement by a centroid with respect to the MFE structure = [(PPV of centroid)/(PPV of MFE structure) − 1] × 100%.

Computational costs and 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 partition function calculation for sampling 1000 structures and for clustering and centroid calculation are given in Table 6 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 at http://sfold.wadsworth.org and http://www.bioinfo.rpi.edu/applications/sfold. Sample output for a folded sequence is available at http://sfold.wadsworth.org/demo.

DISCUSSION

Our main finding is that ensemble centroids yield more specific predictions with average improvements of 30.0% in PPV and 3.5% in sensitivity. More strikingly, the best of a small number of cluster centroids improves the PPV by 46.5% while simultaneously increasing the sensitivity by > 20%. Perhaps our most provocative finding is that the MFE structure falls outside the cluster containing the best centroid for over half of the studied sequences. In such cases, > 31% more base pairs are correctly identified with > 62% fewer predictive errors (Table 5).

The consistent finding of improved PPV suggests that the MFE structure may tend to overpredict. It has been argued that the structure determined by comparative sequence analysis is a minimal model for RNA secondary structure, because only base pairs for which comparative evidence exists are included in the structure model (Larsen and Zwieb 1991). This raises the possibility that overprediction by MFE structure is in part due to underrepresentation of base pairs in the structure determined by comparative sequence analysis. However, recent comparisons of structure determined by comparative sequence analysis with crystal structures indicate that covariation analysis for 16S and 23S rRNAs identifies nearly all base pairings (Cannone et al. 2002). For 16S and 23S rRNAs, the improvements by the centroids are substantial (Tables 2–4), and thus cannot be attributed to potential underrepresentation of base pairs in the structure determined by comparative sequence analysis. For other types of RNAs, comprehensive data for comparing crystal structures with structure determined by comparative sequence analysis are needed for making a more general assessment.

To a large degree, the ensemble centroid is reflective of the high-frequency base pairs in the structure sample. Because the base-pair frequencies are sampling estimates of the base-pair probabilities computed by partition functions (McCaskill 1990), the finding of improved PPV by the ensemble centroid is consistent with the recent report that base pairs in MFE structure that have high probabilities have a significantly higher PPV than that of base pairs with lower probabilities (Mathews 2004).

TABLE 5. Improvement by the best centroid with respect to the location of the MFE structure

Performance measure	Improvement by best centroid over MFE structure (%)	
	MFE structure in the cluster of the best centroid (34 sequences)	MFE structure outside the cluster of the best centroid (47 sequences)
Base-pair distance	14.62 ± 14.28	36.51 ± 23.79
Sensitivity	7.03 ± 43.20	31.44 ± 52.87
Positive predictive value (PPV)	23.77 ± 52.49	62.52 ± 66.87

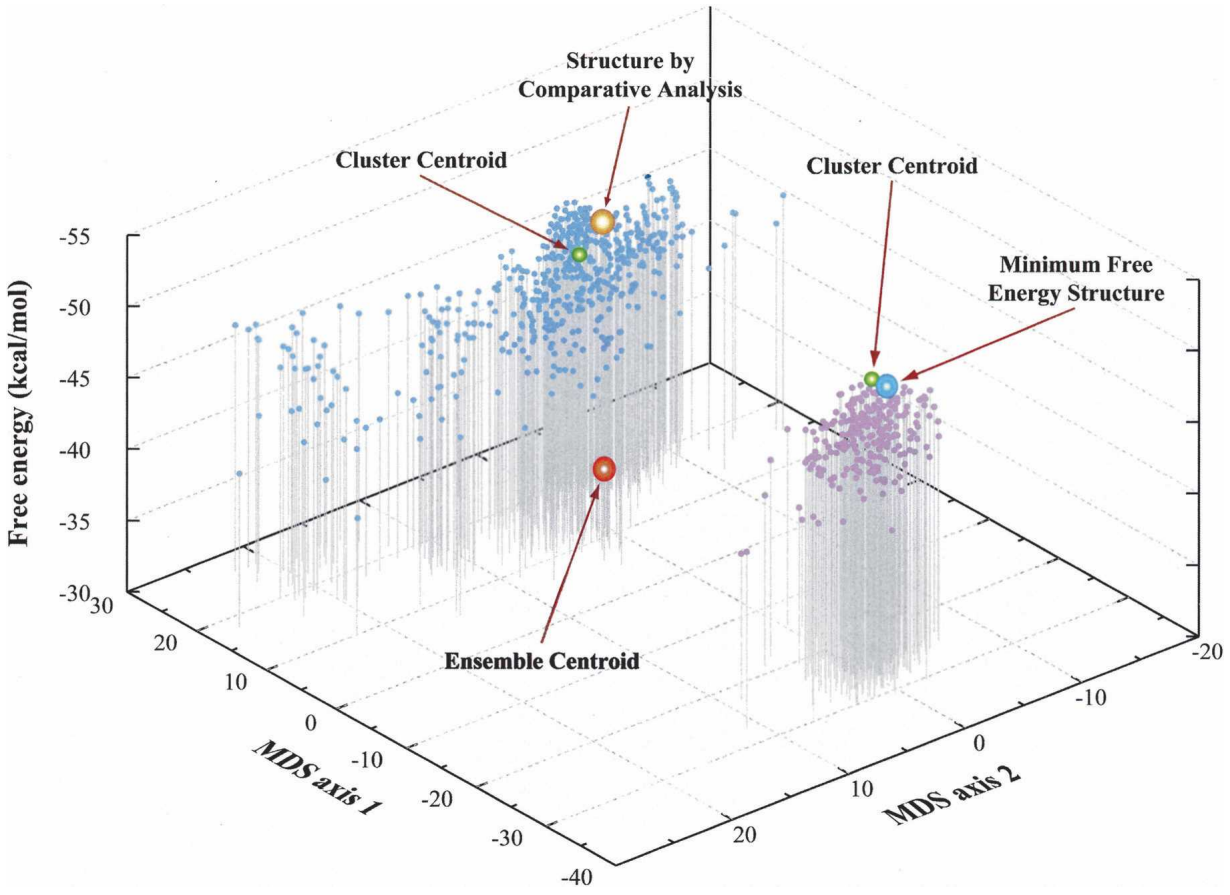

FIGURE 1. The energy landscape of the sampled ensemble and representative structures for *Agrobacterium tumefaciens* 5S rRNA (GenBank accession number X02627) of 120 nt. The structure determined by comparative sequence analysis is in the larger (blue color) cluster with a probability of 0.591 and the MFE structure is in the smaller cluster (purple color) with a probability of 0.409. The coordinates for a structure is (axis 1, axis 2, energy), where the horizontal axes are from multidimensional scaling (MDS; Kruskal and Wish 1977) for presenting high-dimensional objects in typically two dimensions, and the vertical axis is the free energy of a secondary structure. The base-pair distances between structures (see Materials and Methods section) are used for MDS. The coordinates are (21.50, −5.73, −46.80) for the structure determined by comparative sequence analysis, (−27.92, −0.45, −50.50) for the MFE structure, (6.55, 3.15, −36.40) for the ensemble centroid, (20.14, −2.88, −45.80) for the larger cluster centroid, and (−25.95, −0.34, −50.50) for the smaller cluster centroid.

All of the ensemble centroids in our analysis are based on samples of structures. However, we could also use the base-pair probabilities calculated from partition functions (McCaskill 1990) for this purpose. Because sample base-pair frequencies used for centroid calculation approach the base-pair probabilities as the sample size increases, our sample-based-centroid will approach the partition-function-based centroid. However, because base-pair probabilities give only the marginal probabilities of individual base pairs, the identification of clusters of similar structures based on base-pair probabilities alone is at best difficult. In contrast, because sampled structures are realizations from the joint high-dimensional distribution of all base pairs (Ding and Lawrence 2003), clustering is greatly facilitated. Accordingly, a statistical sample enables the decomposition of the two-dimensional histogram of base pairs into subhistograms of distinct structural clusters (Ding and Lawrence 2003).

Although the best centroids are the best predictors, these centroids cannot be defined when a reference structure is unavailable. However, it is an appealing feature that the best centroid predictions are based on only three to four clusters, on average. The small number of cluster centroid predictions can facilitate further structural determination by allowing the incorporation of other types of information, e.g., partial structure information from enzymatic or chemical probing. In order that our comparison be as direct and clear-cut as possible, all predicted structures in this analysis are based on the same set of energy rules (Xia et al. 1998; Mathews et al. 1999). We have not compared these approaches using recently revised energy rules (Mathews et al. 2004). Comparisons incorporating constraints (e.g., for forcing modified bases in tRNAs to be unpaired or for the incorporation of other partial structure information) and coaxial stacking also await further study. However, we currently see no reason why the advantages of these sample-based predictions should not extend to other

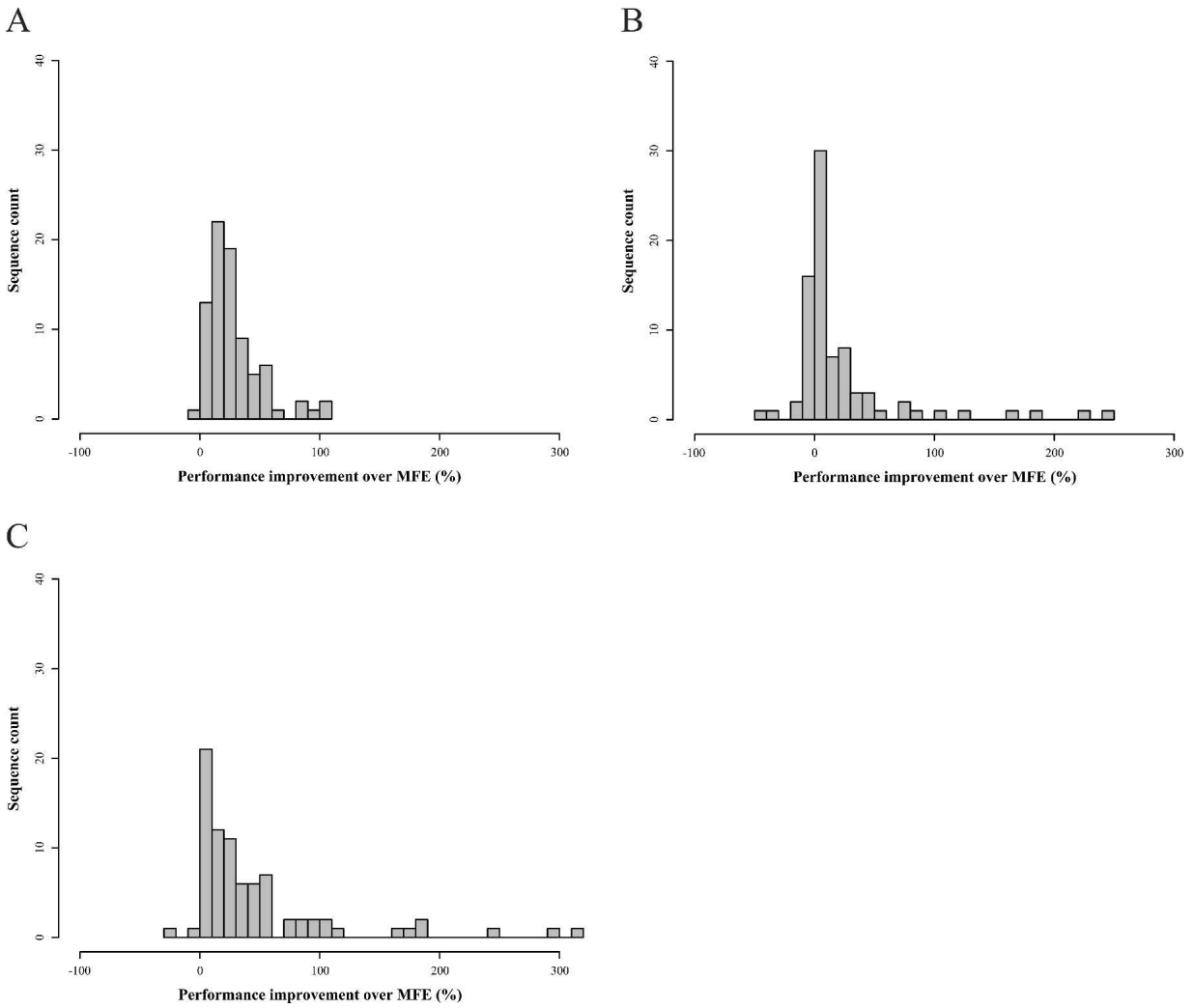

FIGURE 2. Distribution of the improvement percentage by the best centroid for base pair distance (*A*), sensitivity (*B*), and PPV (*C*). The best centroid is the cluster centroid with the shortest base-pair distance to the structure determined by comparative sequence analysis.

cases. We also expect that the use of experimental constraints may improve the predictions, as demonstrated for predictions based on free energy minimization (Mathews et al. 1999, 2004).

We have examined the constrained MFE structure, using base pairs in the ensemble centroid as the constraints. In comparison with the ensemble centroid, we found that on average

TABLE 6. Computational costs of clustering and centroid identification in comparison to those for computing partition functions (PFs) and sampling 1000 structures

		CPU time[a] (seconds)			Memory usage (MB)	
RNA sequence[b] (GenBank accession no.)	Length (nt)	PFs	Sampling	Clustering and centroid computation	PFs + Sampling	Clustering and centroid computation
tRNA (M91245)	70	0.06	1.74	9.92	2.65	65.51
5S rRNA (X02627)	120	0.33	3.84	9.73	2.89	65.82
RNase P RNA (X97396)	359	6.54	11.26	11.81	5.62	67.12
Group I intron (U02540)	556	25.57	22.00	12.66	10.00	68.82
Small subunit rRNA (U04948)	1532	527.06	160.09	21.31	57.34	77.04
Large subunit rRNA (X12612)	2915	4082.19	430.21	37.22	198.75	95.79

[a]Benchmarked on an AMD Opteron 1.8-GHz processor under the Linux operating system.
[b]The name of the organism of the sequence can be found in Table 1.

the sensitivity is improved by 2.22% for the constrained MFE structure as a result of more predicted base pairs; however, this small improvement has a cost of 7.05% average decrease in PPV. It remains an open question whether the combination of centroid prediction with other approaches can further improve structure prediction. The base pair frequencies for the entire structure sample as well as for individual clusters might be used as weights by the maximal weighted matching or the iterated loop matching method (Tabaska et al. 1998; Ruan et al. 2004) for calculating representative structures with pseudoknots.

As alternatives to examining clusters and centroids for sampled structures, one might consider clustering the 1000 structures with the lowest energies computed by RNAsubopt of the Vienna RNA package (Hofacker 2003) or the abstract shape representation of foldings within an energy increment from the MFE (Giegerich et al. 2004). These two methods focus on the lowest end of the free energy density of states, whereas structure sampling allows characterization of Boltzmann-weighted density of states (Ding and Lawrence 2003). Thus, the energy landscape is examined from two different perspectives. For short sequences such as tRNAs, there is generally a good correspondence between our centroids and the abstract shapes or centroids for 1000 best structures (data not shown), as the low energy structures are well represented in a sample. However, the degree of correspondence and the overlap in the energy coverage diminishes as sequence length increases, because apparently the Boltzmann-weighted density of states becomes increasingly dictated by structures at an energy distance from the MFE, and these structures far outnumber those with energies near the MFE. For example, for the rabbit β-globin mRNA of 589 nt (GenBank accession no. V00879), the 1000 best structures represent a small free energy range of 1.3 kcal/mol for default parameter settings of RNAsubopt, while a statistical sample presents representative structures from a much wider energy interval of 39.80 kcal/mol. In addition, a statistical sample can reveal "entropic clusters" (Ding and Lawrence 2003). For an entropic cluster, each member has a probability too small to command individual attention, but collectively the cluster has an appreciable probability because of the large number of cluster members. In computer RNA folding applications, it is a common practice among users to examine structures from mfold. Because the structure sample from mfold is heuristic rather than statistically or low-energy representative, the method presented here and the RNAshapes approach present two improved and complementary alternatives. The two methods are also complementary because RNAshapes provides an alternative method for the identification of structure clusters with cluster members having a common shape. It will be interesting to apply the RNAshapes algorithm to a sampled ensemble. The larger number of shape representatives than the number of our clusters suggests that our clustering procedure reports major clusters whereas the abstract shape approach may reveal more subtle structural dissimilarities.

As pointed out by Abagyan (1993), two major components are needed to solve macromolecular folding problems. First, all essential terms of the free energy of a trial conformation must be calculated with sufficient accuracy, and, second, a procedure is needed to find the minimum of this energy function. For RNAs, the global minimum can be found for an incomplete energy model, i.e., only the secondary structure model. Thus, it may not be a surprise that, for a majority of analyzed sequences (47 of 81 sequences), a centroid of an alternative cluster (probably representing an alternative energy well) is about 37% closer to the structure determined by comparative sequence analysis than the MFE structure computed with the incomplete energy function (Table 5). As argued by Abagyan, since only a small number of such alternative structural classes are needed to find this best alternative, an approach that employs a post-analysis filter function maybe a productive path for selecting among clusters for improved structure prediction. Since for protein structural models neither of Abagyan's two components is attainable, our findings argue that a more comprehensive examination of the energy landscape of the approximate models for structures of proteins and other macromolecules may also be worthy of further investigation. Even in the case of the complete model and energy function, the Boltzmann ensemble view is also important, e.g., for the investigation of metastable states, particularly RNA conformational switches (Barrick et al. 2004; Voss et al. 2004).

MATERIALS AND METHODS

RNA sequences

From publicly available databases (Larsen and Zwieb 1991; Sprinzl et al. 1998; Brown 1999; Cannone et al. 2002; Alm Rosenblad et al. 2003; Zwieb et al. 2003), we took samples of sequences for diverse types of structural RNAs with secondary structures determined by comparative sequence analysis. For tRNAs, RNase P RNAs, tmRNAs, signal recognition particle (SRP) RNAs, small subunit (16S or 16S-like) rRNAs, large subunit (23S or 23S-like) rRNAs, and 5S rRNAs, 10 sequences were randomly selected for each RNA type. In addition, nine group I introns without undetermined nucleotides and two group II introns that are available in the databases were also included in our analysis. The list of the 81 sequences is available from the authors upon request and will also be posted on the Sfold Web server.

Clustering procedure

For comparing two secondary structures, we use the base-pair distance. The discriminatory power of this distance is adequate in our context, because we are interested in comparing multiple

structures sampled for a given RNA sequence. In other circumstances, e.g., when structures of homologous sequences are compared, alternative metrics (Moulton et al. 2000) may be more appropriate to account for insertions and deletions. For hierarchical clustering of structures, we employ *Diana* (Kaufman and Rousseeuw 1990), a top-down divisive method that has performed well in other contexts (Datta and Datta 2003). For determining the number of clusters, we use the CH index (Calinski and Harabasz 1974) that was assessed as the best in a comprehensive study (Milligan and Cooper 1985). For a given divisive level on the clustering tree from *Diana*, the CH index is calculated 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, $B(k)$ is the between-cluster sums of squares, and $W(k)$ is the within-cluster sums of squares. The number of clusters at which the CH index is maximized is the optimal number of clusters. This number is then used to determine the structural clusters by identifying the corresponding divisive level for the hierarchical clustering tree produced by *Diana*. As described below, a secondary structure is an object in a high-dimensional Euclidean space that can be expressed by an upper triangular matrix. The cluster means required for the calculation of the sum of squares can be computed in two ways: averaging the corresponding Euclidean coordinates for all structures in a cluster or using the centroid of a cluster as its mean and computing the base-pair distance between the centroid and any structure in the cluster (see below for centroid calculation). The former is computationally intensive for long sequences whereas the later is a linear operation. We did not observe appreciable differences in the clustering results by the two methods. Therefore, we decided to use cluster centroids in the implementation of the CH index.

For every RNA sequence, we first cluster 1000 statistically sampled structures. We compute the MFE structure with mfold 3.1 (Zuker 2003) for the same set of Turner thermodynamic parameters (Xia et al. 1998; Mathews et al. 1999) that are currently implemented by our sampling algorithm (Ding and Lawrence 2003). To decide which cluster the MFE structure belongs to, 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. The structure sample size of 1000 has been shown to be sufficiently large to guarantee statistical reproducibility in typical sampling statistics including base-pair frequencies, even when two independent samples do not share a single structure (Ding and Lawrence 2003). As reported below, base pair frequencies are all that are needed for centroid identification. In addition, regardless of sequence length, a cluster with an appreciable probability 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.

Matrix representation of RNA secondary structure

For an RNA sequence of n nucleotides, a secondary structure I can be expressed by an upper triangular matrix of indicators $\{I_{ij}\}$, $1 \le i < j \le n$, where I_{ij} indicates base-pairing status between base i and base j. Specifically, $I_{ij} = 1$ 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 $i+3$, $1 \le i$, $i+3 \le n$. The indicators are not independent of one another, 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, $\Sigma_{1\le i\le n} I_{ij} \le 1$, and $\Sigma_{1\le j\le n} I_{ij} \le 1$.

Base-pair distance

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 $I_2 = \{I_{ij}^2\}$, consider the following metric D_1 and the squared Euclidean distance D_2: $D_1(I_1, I_2) = \sum_{1\le i<j\le n} |I_{ij}^1 - I_{ij}^2|$ and $D_2(I_1, I_2) = \sum_{1\le i<j\le n} (I_{ij}^1 - I_{ij}^2)^2$. In general, both metrics have sufficient discriminatory power for the purpose of clustering. In our context, 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 in fact 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.

Definition and derivation of centroid

For a set of m secondary structures $I_1, I_2, \ldots, I_m$, with $I_k = \{I_{ij}^k\}$, $1 \le k \le m$, the centroid for the set is defined as the structure in the *entire ensemble* of secondary structures that has the shortest total base-pair distance to the structures in the set. To compute the centroid structure, we need to find the secondary structure $I = \{I_{ij}\}$ that minimizes the following sum of distances under the constraints discussed above:

$$\begin{aligned}
&\Sigma_{1\le k\le m}\Sigma_i\Sigma_j \left(I_{ij}^k - I_{ij}\right)^2 \\
&= \Sigma_{1\le k\le m}\Sigma_i\Sigma_j \left[\left(I_{ij}^k\right)^2 - 2I_{ij}I_{ij}^k + \left(I_{ij}\right)^2\right] \\
&= \Sigma_{1\le k\le m}\Sigma_i\Sigma_j \left(I_{ij}^k\right)^2 - 2\,\Sigma_i\Sigma_j\left(\Sigma_{1\le k\le m} I_{ij}^k\right) I_{ij} + m\,\Sigma_i\Sigma_j I_{ij} \\
&= C_s + \Sigma_i\Sigma_j\left(m - 2\,C_{ij}\right) I_{ij} \qquad (1)
\end{aligned}$$

where $C_s = \Sigma_{1\le k\le m}\Sigma_i\Sigma_j (I_{ij}^k)^2$ is a constant for the given structure set, and $C_{ij} = \Sigma_{1\le k\le m} I_{ij}^k$ is the total number of occurrences of base pair $i \cdot j$ in the structure set. Because $I_{ij}^2 \equiv I_{ij}$, the nonlinear programming problem is in fact a linear programming problem with nonlinear constraints. For a base pair with a frequency under 50%, it cannot be in the centroid because $(m - 2C_{ij}) > 0$, and thus I_{ij} must be 0 for the centroid. A base pair with a frequency of 50% does not influence the double sum in (1), because $(m - 2C_{ij}) = 0$. For a base pair with a frequency $> 50\%$, because $(m - 2C_{ij}) < 0$, the inclusion of this base pair (i.e., $I_{ij} = 1$) decreases the double sum in (1). Furthermore, any two base pairs with frequencies $> 50\%$ do not form a pseudoknot, because no base pairs in the structure set are involved in pseudoknots. Thus, the consensus structure formed by all base pairs with a frequency $> 50\%$ is a centroid. We note that for base pairs with a frequency of 50%, inclusion of any compatible combination into the $> 50\%$ consensus does define another centroid. However, the $> 50\%$ consensus structure is always the unique centroid with the smallest number of base pairs and is the one we use for analysis.

The centroid is referred to as the ensemble centroid when the structure set is the statistical sample generated by our sampling algorithm, typically with $m = 1000$. In this case, C_{ij} is the observed

count for base pair $i \cdot j$ in the sample. For a cluster of similar structures identified from the statistical sample, the centroid is referred to as a cluster centroid. In this case, C_{ij} is the observed count for base pair $i \cdot j$ in the cluster.

ACKNOWLEDGMENTS

The Computational Molecular Biology and Statistics Core at the Wadsworth Center is acknowledged for providing computing resources for this work. This work was supported in part by National Science Foundation 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. We are grateful to the suggestions and observations of the anonymous referees that led to drastic improvements in the run times of clustering and centroid identification as well as the presentation of the article.

Received March 7, 2005; accepted May 9, 2005.

REFERENCES

Abagyan, R.A. 1993. Towards protein folding by global energy optimization. *FEBS Lett.* **325:** 17–22.

Alm Rosenblad, M., Gorodkin, J., Knudsen, B., Zwieb, C., and Samuelsson, T. 2003. SRPDB (Signal Recognition Particle Database). *Nucleic Acids Res.* **31:** 363–364.

Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. *Science* **181:** 223–230.

Barrick, J.E., Corbino, K.A., Winkler, W.C., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I., et al. 2004. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc. Natl. Acad. Sci.* **101:** 6421–6426.

Bonhoeffer, S., McCaskill, J.S., Stadler, P.F., and Schuster, P. 1993. RNA multi-structure landscapes. A study based on temperature dependent partition functions. *Eur. Biophys. J.* **22:** 13–24.

Brown, J.W. 1999. The ribonuclease P database. *Nucleic Acids Res.* **27:** 314.

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

Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Muller, K.M., et al. 2002. The comparative RNA Web (CRW) site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinform.* **3:** 2.

Datta, S. and Datta, S. 2003. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* **19:** 459–466.

Ding, Y. and Lawrence, C.E. 2001. Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. *Nucleic Acids Res.* **29:** 1034–1046.

———. 2003. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.* **31:** 7280–7301.

Dowell, R.D. and Eddy, S.R. 2004. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinform.* **5:** 71.

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

Hofacker, I.L. 2003. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31:** 3429–3431.

Kaufman, L. and Rousseeuw, P.J. 1990. *Finding groups in data: An introduction to cluster analysis.* John Wiley & Sons, New York.

Kruskal, J.B. and Wish, M. 1977. *Multidimensional scaling.* Sage Publications, Beverly Hills, CA.

Larsen, N. and Zwieb, C. 1991. SRP-RNA sequence alignment and secondary structure. *Nucleic Acids Res.* **19:** 209–215.

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

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

Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M., and Turner, D.H. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci.* **101:** 7287–7292.

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

Milligan, G.W. and Cooper, M.C. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50:** 159–179.

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

Pappu, R.V., Marshall, G.R., and Ponder, J.W. 1999. A potential smoothing algorithm accurately predicts transmembrane helix packing. *Nat. Struct. Biol.* **6:** 50–55.

Ruan, J., Stormo, G.D., and Zhang, W. 2004. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics* **20:** 58–66.

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

Tabaska, J.E., Cary, R.B., Gabow, H.N., and Stormo, G.D. 1998. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics* **14:** 691–699.

Voss, B., Meyer, C., and Giegerich, R. 2004. Evaluating the predictability of conformational switching in RNA. *Bioinformatics* **20:** 1573–1582.

Wuchty, S., Fontana, W., Hofacker, I.L., and Schuster, P. 1999. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* **49:** 145–165.

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

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

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

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

Zwieb, C., Gorodkin, J., Knudsen, B., Burks, J., and Wower, J. 2003. tmRDB (tmRNA database). *Nucleic Acids Res.* **31:** 446–447.