Nucleic Acids Research, 1995, Vol. 23, No. 14 ***2791–2798***

'Well-determined' regions in RNA secondary structure prediction: analysis of small subunit ribosomal RNA

Michael Zuker and Ann B. Jacobson[1],*

Institute for Biomedical Computing, Washington University, St Louis, MO 63110, USA and [1]Department of Microbiology, State University of New York, Stony Brook, NY 11794-5222, USA

Received January 6, 1995; Revised and Accepted May 10, 1995

ABSTRACT

Recent structural analyses of genomic RNAs from RNA coliphages suggest that both well-determined base paired helices and well-determined structural domains that are identified by 'energy dot plot' analysis using the RNA folding package *mfold*, are likely to be predicted correctly. To test these observations with another group of large RNAs, we have analyzed 15 ribosomal RNAs. Published secondary structure models that were derived by comparative sequence analysis were used to evaluate the predicted structures. Both the optimal predicted fold and the predicted 'energy dot plot' of each sequence were examined. Each prediction was obtained from a single computer run on an entire ribosomal RNA sequence. All predicted base pairs in optimal foldings were examined for agreement with proven base pairs in the comparative models. Our analyses show that the overall correspondence between the predicted and comparative models varies for different RNAs and ranges from a low of 27% to a high of 70%, with a mean value of 49%. The correspondence improves to a mean value of 81% when the analysis is limited to well-determined helices. In addition to well-determined helices, large well-determined structural domains can be observed in 'energy dot plots' of some 16S ribosomal RNAs. The predicted domains correspond closely with structural domains that are found by the comparative method in the same RNAs. Our analyses also show that measuring the agreement between predicted and comparative secondary structure models underestimates the reliability of structural prediction by *mfold*.

INTRODUCTION

Two main methods are currently employed to predict RNA secondary structure; comparative sequence analysis (1–3) and free energy minimization (4–10).The former method proceeds from the assumption that structure is much more highly conserved than sequence during evolution. Base pairs are inferred by finding positions in aligned sequences that co-vary so as to conserve base pairing potential. In contrast, free energy minimization requires only a single sequence and proceeds automatically without the labor-intensive steps of iterative alignment and base pair detection that comparative sequence analysis requires.The major problem with energy minimization has been the lack of reliability of the predictions.

Coupling the suboptimal folding algorithm of Zuker (9,11) with the energy rules of Turner and colleagues (12–14) has led to impressive improvements in RNA secondary structure prediction. However, despite the dramatic improvement in the overall quality of prediction, variability is observed in the reliability of prediction among different sequences, as well as within different regions of the same sequence. Thus a user confronted with the structural prediction for a new RNA sequence would like to be able to assess the reliability of a particular prediction. The analyses presented here show that helices and structural domains that are well determined are predicted more reliably than helices and structural domains that are poorly determined. The identification of well-determined structures is generally done by viewing 'energy dot plots' by eye. However, we describe a first step towards quantifying our subjective notion of well-determined base pairs and provide a small computer program for this purpose.

MATERIALS AND METHODS

We have analyzed 15 small subunit rRNAs from bacteria (14) and chloroplasts (1). The names and accession numbers are given in Table 1. Published structures obtained by comparative sequence analysis were used to evaluate the predicted foldings (15–17). Only base pairs that have been proven by co-variance were used in the analysis.

RNA folding predictions use programs in version 2.2 (December 1992) of the *mfold* package (9,11,18,19). New software has been written to create and analyze energy dot plots. The energy dot plots used in these analyses show base pairs in structures within 12 kcal from minimum folding energies. For this manuscript, a new type of dot plot, called an overlaid energy dot plot, was created. The overlaid dot plots afford a view of how close the optimal folding(s) is (are) to the comparative model and

* To whom correspondence should be addressed

whether or not base pairs in the comparative model are predicted to be well-determined. The program that creates the overlaid dot plot is available from the first author upon request.

Table 1. Ribosomal sequences with GenBank accession nos

Sequence	Accession no.
Anacystis nidulans 16S	X00512
Bacillus subtilis 16S	K00637, M10606
Desulfovibrio desulfuricans 16S	M34113
Escherichia coli 16S	J0 1695
Heliobacterium chlorum 16S	M11212
Haloferax volcanii 16S	K00421
Mycoplasma gallisepticum 16S	M22441
Methanospirillum hungatei 16S	M60880
Thermococcus celer 16S	M21529
Thermoproteus tenax 16S	M35966
Thermomicrobium roseum 16S	M34115
Thermoplasma acidophilum 16S	M32297, M20822
Thermus thermophilus 16S	X07998
Thermotoga maritima 16S	M21744
Zea mays chloroplast 16S	Z00028

The rRNA sequences were obtained from the RDP database maintained at the University of Illinois (15).

Although the dot plots can be analyzed visually to determine what optimal base pairs are 'well-determined' in the sense of having few competitors in suboptimal foldings, it is useful to have a quantitative measure for this concept and we have devised a measure which we call *H*-num. It derives from an earlier function called *P*-num (9,11). The *P*-num function was introduced to describe how 'well-determined' individual nucleotides are in terms of their predicted status in a secondary structure. In a molecule of n nucleotides and for a given energy dot plot, *P*-num(i) is the total number of dots in the ith row and column of the dot plot. In an unfiltered dot plot this is the same as the total number of base pairs that can be formed with the ith nucleotide in all possible structures within the prescribed degree of suboptimality. Proceeding from here, we can define the *H*-num function as follows. For a base pair between ribonucleotides i and j, let *H*-num(i,j) = *P*-num(i) + *P*-num(j) − 1. This is the total number of dots in the ith and jth rows and columns of the dot plot. *H*-num(i,j) is at least 1 if i:j is a valid base pair within the prescribed free energy increment. Thus *H*-num(i,j) counts the total number of base pairs in all secondary structures within the prescribed free energy increment that contain nucleotides i or j. The *H*-num value for a helix is defined as the average *H*-num value for the base pairs in that helix. A helix with a low *H*-num value is said to be 'well-determined'. The definition of low is subjective. For these studies helices with a value ≤60 are considered well-determined. The cut-off of 60 was obtained by visually inspecting the ribosomal dot plots and choosing the maximum *H*-num value for which all helices appear to be well-determined by eye. A more rigorous approach will be developed in future studies (see Discussion). The current program for *H*-num analysis is available from the first author upon request.

RESULTS

Well-determined helices are well predicted

The number of well-determined helices that are found in any given folding varies for different RNAs. The variation among 15 ribosomal sequences is shown in Table 2. Both the number and the percent of well-determined helices are shown. The values range from a low of two (2%) well-determined helices for *Anacystis nidulans* 16S rRNA to 75 (85%) well-determined helices for 16S rRNA from *Thermococcus celer*, with a mean value of 27 (27%). The predicted free energy (ΔG) for the optimal folding of each of the RNAs is shown in the last column of the table. Five of the rRNAs shown in the table are found in thermophilic bacteria. These RNAs have a large number of well-determined helices and their RNA secondary structures are predicted to be unusually stable.

Table 2. The percent of well-determined helices found in the predicted foldings of small subunit ribosomal RNAs

Sequence	Total	Well-determined	Percent	ΔG (kcal/mol)
Anacystis nidulans 16S	105	2	2	–417.2
Bacillus subtilis 16S	150	8	5	–455.1
Desulfovibrio desulfuricans 16S	112	8	7	–437.6
Escherichia coli 16S	111	7	6	–434.6
Heliobacterium chlorum 16S	106	7	7	–453.7
Haloferax volcanii 16S	98	53	54	–515.1
Mycoplasma gallisepticum 16S	148	7	5	–349.4
Methanospirillum hungatei 16S	101	47	47	–458.6
Thermus thermophilus 16S	112	38	34	–585.0
Thermococcus celer 16S	88	75	85	–656.2
Thermoproteus tenax 16S	98	59	60	–704.8
Thermomicrobium roseum 16S	116	19	16	–566.7
Thermoplasma acidophilum 16S	94	44	47	–482.3
Thermotoga maritima 16S	112	24	21	–621.4
Zea mays chloroplast 16S	109	4	4	–432.2

The last column of the table shows the predicted ΔG for the optimal folding of each rRNA sequence.

Table 3. The agreement between base pairs found in optimal foldings and those in comparative models of small subunit ribosomal RNAs

Sequence	Total base pairs			Well-determined base pairs		
	Predicted	Comparative	Agreement (%)	Total	Agreement (%)	Improvement (Δ%)
Anacystis nidulans 16S	482	412	244 (51)	10	6 (60)	9
Bacillus subtilis 16S	682	435	285 (42)	61	59 (97)	55
Desulfovibrio desulfuricans 16S	518	426	212 (41)	44	38 (86)	45
Escherichia coli 16S	512	440	239 (47)	53	52 (98)	51
Heliobacterium chlorum 16S	496	422	210 (42)	46	38 (83)	40
Haloferax volcanii 16S	492	417	342 (69)	314	252 (80)	11
Mycoplasma gallisepticum 16S	670	415	246 (37)	40	35 (87)	51
Methanospirillum hungatei 16S	492	414	299 (61)	242	182 (75)	14
Thermus thermophilus 16S	532	431	249 (47)	197	163 (83)	36
Thermococcus celer 16S	496	427	310 (63)	431	302 (70)	8
Thermoproteus tenax 16S	524	439	296 (57)	345	253 (73)	17
Thermomicrobium roseum 16S	542	427	234 (43)	90	65 (72)	29
Thermoplasma acidophilum 16S	486	422	276 (57)	262	205 (78)	21
Thermotoga maritima 16S	527	443	300 (57)	130	100 (77)	20
Zea mays chloroplast 16S	504	407	136 (27)	24	23 (96)	69

The agreement between all predicted base pairs amd base pairs in the comparative model is juxtaposed with the agreement between well-determined predicted base pairs and base pairs in the comparative model. Well-determined base pairs are defined as base pairs in well-determined helices, i.e. helices with h-num values ≤60.

The agreement between the optimal predicted folding of each 16S rRNA and the comparative model of its secondary structure is given in the first three columns in Table 3. Base pairs rather than entire helices were used for the comparison, because predicted helices often have more base pairs than comparable helices in the comparative models. The observed agreement between the predicted optimal foldings and the comparative models ranges from 27% for 16S rRNA from *Zea mays* to 70% for 16S rRNA from *Haloferax volcanii*, with a mean agreement of 49%.

Two measures can be used to determine the agreement between the predicted and comparative models. One gives the fraction of base pairs or helices in the comparative model that are found by the algorithm (18,20). The second measure gives the fraction of base pairs predicted by the algorithm that are found in the comparative model. The latter measure gives a lower estimate of the reliability of *mfold* prediction, because of the number of unpaired nucleotides in the comparative models. It was used for the current analysis, because we are interested in evaluating the reliability of structural prediction for RNAs of unknown secondary structure.

The improvement in prediction that is achieved by restricting the analysis to well-determined helices (helices with an *H*-num value ≤60) is shown in the last three columns of Table 3. The base pairs shown in column 4 are those found in the well-determined helices of Table 2, column 2. Agreement with the comparative models is shown in column 5. The values range from 60% for *A.nidulans* to 97% for *Bacillus subtilis*. On average, 81% of well-determined helices agree with matching helices in the comparative models. The average improvement (column 6) is 32%. Overall secondary structure prediction is more reliable for RNAs with many well-determined helices (columns 3 and 4). As a result, these RNAs show less relative improvement (column 6).

A more detailed view of the improvement in prediction that is achieved when the analysis of RNA secondary structure is confined to well-determined helices is afforded in Figure 1. Optimal helices were sorted by increasing *H*-num value, from 'best-determined' to 'poorest-determined'. For each *H*-num value, the percent of predicted base pairs in agreement with the comparative model was plotted. Examination of the plots reveals that the number of predicted helices that are in agreement with the comparative model always decreases with increasing *H*-num and shows, once more, that well-determined helices are predicted more reliably than poorly-determined helices.

The *H*-num plots that are shown were selected to illustrate the variation in well-determined and/or well-predicted structure in different 16S sequences. In Figure 1A (*T.celer*), the majority of predicted helices are clustered in the left half of the plot. They are relatively well-determined and are relatively well-predicted. In contrast, most helices are 'poorly-determined' and poorly predicted in Figure 1B (*A.nidulans*) and cluster in the right half of the plot. The plot shown in Figure 1C (*Escherichia coli*) has intermediate properties.

Use of the comparative models lead us to underestimate of the reliability of structural prediction by *mfold*

The comparative models used for testing the reliability of the predicted optimal foldings derive from rigorous analysis of co-variance and provide the best structural models currently available with which to test the quality of *mfold* prediction (3). Nonetheless, a strict comparison of the two types of structural models leads us to underestimate slightly the reliability of RNA secondary structure prediction by *mfold*, because of inherent differences in the predicted and comparative models. *mfold* predicts the potential structure of RNAs in solution in the absence of protein. Secondary structures deduced from comparative analysis are structures conserved in evolution because of their

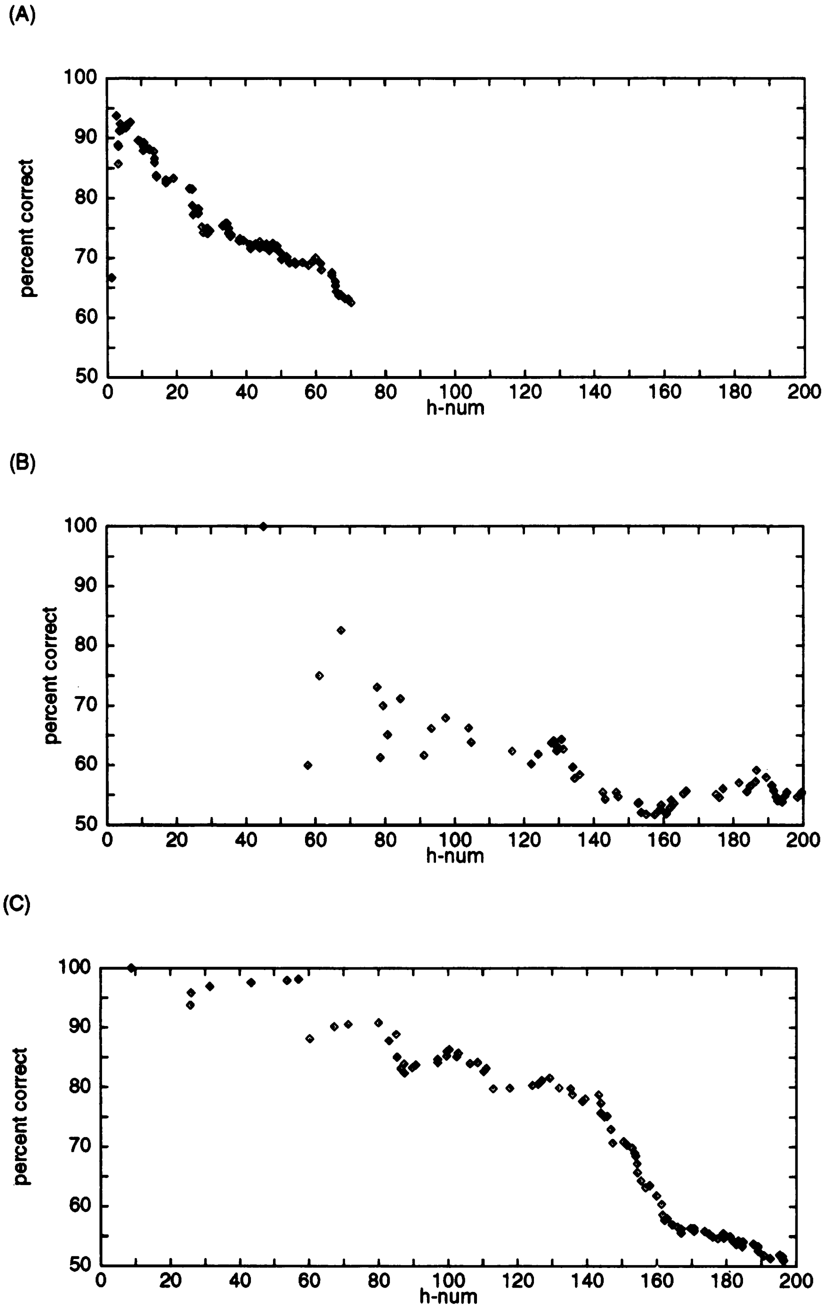

Figure 1. *H*-num plots for three 16S rRNAs. At each *H*-num, the percent of optimal base pairs that are in agreement with base pairs in the helices of the comparative model are plotted. The plots are cumulative, with each point representing all helices up to that point. (**A**) *T.celer* 16S rRNA, (**B**) *A.nidulans* 16S rRNA, (**C**) *E.coli* 16S rRNA.

functional significance in living organisms. Thus the models could differ if predicted structures do not have functional significance and are, therefore, not conserved in related RNAs or if conserved structures in the comparative models are not stable in solution without bound proteins. The models may also differ if the nucleotide sequence of a helix is strictly conserved and its existence cannot be proved by comparative analysis.

Some differences that are found among well-determined helices and those in the comparative models are shown in Figure 2 and Table 4. The structures shown are from the optimal prediction of *H.volcanii* 16S rRNA; in all cases predicted helices contain more base pairs (shown boxed) than are found in the comparative model. Most of the differences that have been observed fall into two classes. (i) One or more additional base pairs are present at the beginning or the end of the helix. For example, in the helix U548–G568, two of the eight predicted base pairs are missing. The *H*-num value for this helix is 2.9 and the agreement with the comparative model for this helix is 75%. Thermodynamic studies with small model compounds support the formation of additional base pairs in regions of this type (14). Approximately 68% of all well-determined helices in the *H.volcanii* 16S prediction that differ from helices in the

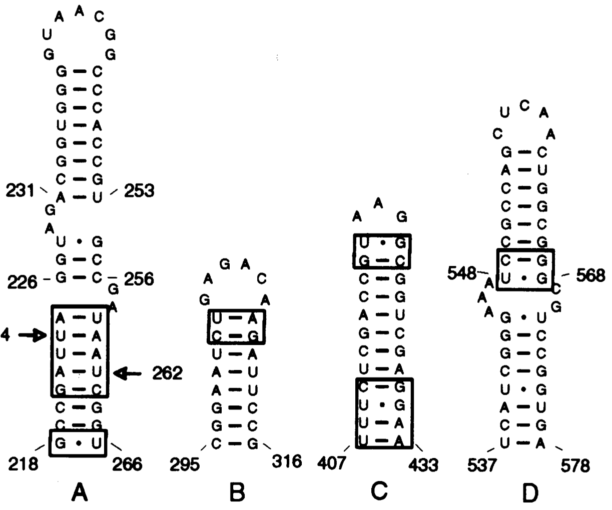

Figure 2. Comparison of the predicted and comparative secondary structure models for four well-determined hairpins in the optimal predicted folding of *H.volcanii* 16S rRNA. Boxed base pairs correspond to single-stranded regions in the comparative model. The arrows at U^{224} and U^{262} indicate that these nucleotides form a U:U pair in the comparative model.

comparative model are of this type. (ii) Longer base paired segments or entire helices are missing from the comparative model. In the lower region of hairpin C, for example, four base pairs are missing. The base pairs are conserved among ribosomal RNA but exhibit no co-variations, thus they are not proved by comparative analysis and are scored as differences in our analyses.

Table 4. Details of the agreement between predicted and comparative secondary structure models for several well-determined hairpins in *Halerofax volcanii* 16S rRNA

Helix	Helix length	Agreement	*H*-num
548–568	8	6	2.9
407–433	12	6	4.8
537–578	8	8	5.1
295–316	8	6	6.6
231–253	8	8	10.1
218–266	8	3	11.5
226–256	3	3	13.0

Two-dimensional models for these hairpins are shown in Figure 2. The nucleotide numbers given correspond to the first and last nucleotides of each helix.

A rare difference is illustrated in the lower helix of hairpin A. The helix is formed by seven canonical base pairs and one G:U pair. The comparative model for this region is primarily single-stranded; it has two canonical base pairs and, in addition, a non-canonical U:U pair (indicated by arrows in the figure). While comparative studies provide compelling evidence for many non-canonical base pairs in ribosomal RNAs (3), they are not included in *mfold*. In the current example, allowing U:U pairs would not alter the predicted structure for the helix, since the five missing canonical pairs are likely to form in solution.

If the differences between the comparative and predicted models like those illustrated in Figure 2 are ignored, 95% of well-determined helices in the optimal folding of *H.volcanii* 16S rRNA would be reliably predicted. This is up from the level of 80.3% given in Table 3. A similar computation can be made for all of the ribosomal RNAs analyzed in the current study. Thus measuring the agreement between the predicted and the comparative models may underestimate the magnitude of reliable prediction by *mfold*.

The prediction of large structural domains

Large well-determined structural domains are identified by visual inspection of an 'energy dot plot'. These plots show suboptimal output from the RNA folding algorithm (9). A dot in row *i* and column *j* of the plot represents a base pair between the *i*th and *j*th ribonucleotides in a sequence. The plot shows the superposition of all optimal and close to optimal foldings in a single figure. These plots often contain a mixture of clear regions and cluttered regions. Clear regions define well-determined structural domains in the optimal folding. Nucleotides within these regions do not interact with other regions of the molecule in suboptimal foldings. Cluttered regions indicate portions of the molecule that have the potential to form numerous alternative structures. Predicted structures in the optimal folding that are located in cluttered regions of the plots are considered to be poorly determined.

Three 'overlaid energy dot plots' are shown in Figure 3. Like standard 'energy dot plots', they show base pairs in the predicted optimal foldings (lower left triangle), as well as all possible base pairs in all possible foldings within 12 kcal of the optimal predicted foldings (upper right triangle). In addition, base pairs from the comparative models are overlaid on each plot as larger red and blue dots.

Some qualitative features and differences are immediately apparent among the energy dot plots. The plot of *E.coli* 16S rRNA is the most cluttered and seems to have the greatest overall density of dots. The plot of *H.volcanii* 16S rRNA has one large clear rectangular area. The plot of *T.celer* 16S rRNA is unusually sparse, with large clear areas over much of the plot.

Two structural domains are present in the 'energy dot plot' of *H.volcanii* 16S rRNA (Fig. 3B). The large clear rectangle with corners at (1,494), (1,1474), (1474,494) and (494,494) shows that nucleotides in the region 1–494 do not interact with nucleotides from the remainder of the molecule. The structure formed within nucleotides 1–494 is well-determined; secondary structures within the domain formed by nucleotides 494–1474 are poorly-determined. The entire 'energy dot plot' for *T.celer* 16S rRNA (Fig. 3C) is unusually well-determined. Four structural domains, extending from nucleotide 1 to 498, 499 to 868, 869 to 1352 and 1353 to 1486 can be identified.

Of the 15 ribosomal sequences that have been analyzed in this study, five have well-determined structural domains. The nucleotide positions of these domains have been compared with the position of comparable domains in the respective comparative models (Table 5). In three of them (*H.volcanii, T.celer* and *Thermoproteus tenax*) the domains in the predicted and comparative models coincide. For the remainder, the nucleotide positions of the structural domains correspond less well to the comparative models, although the magnitude of the differences that are observed is small.

The position of red and blue dots within each dot plot (Fig. 3) provides additional information about the agreement between the predicted and comparative models In general, the preponderance of correctly predicted helices (blue dots) are located along the diagonal of the plot, where short local hairpins are situated. Red dots, indicating regions of the phylogenetic model that are missing from predicted folding, are more often long-range and are usually located in regions of the plot that are poorly-determined.

The reliability of structural prediction within well-determined structural domains is better than in poorly determined regions. However, the level of improvement observed is variable and generally depends on the number of well-determined helices within the structure. Structures within domain III of the *T.celer* rRNA provide an exception to this generalization. Although 96% of the helices in domain III are well-determined, only 58% are in agreement with structures in the comparative model. Consistent with this, we note that structural prediction for domain III agrees poorly with the comparative models for domain III in all ribosomal sequences that we have studied. The reason for these discrepancies, as well as their significance both for the comparative and predicted models, remains to be explored. It should also be noted that we do not find consistent well-determined domains or local hairpins throughout the 15 ribosomal sequences that have been examined.

DISCUSSION

The results presented here show that the computer program *mfold* predicts RNA secondary structures more reliably for helices that are 'well-determined' than for helices that are 'poorly-determined'. In addition, the overall prediction of RNA secondary structure is more reliable for RNAs with many well-determined helices. The results also show that well-determined structural domains correspond to structural domains in the comparative models. Structures within these domains are generally predicted more reliably than structures in other regions of the RNA.

As shown in Table 3, only 47% of the predicted helices in *E.coli* 16S rRNA are in agreement with the comparative model. *Escherichia coli* 16S rRNA has often been used as a standard with which to determine the reliability of prediction for large RNAs, because it was the first large RNA whose secondary structure was well established. The *E.coli* 16s rRNA dot plot (Fig. 3A) is dense and contains few well-determined helices. In view of the correlation that has been observed between the overall quality of structural prediction and the appearance of the 'energy dot plots', it is not surprising that the overall prediction for this RNA is relatively poor compared with other rRNAs (Table 3). These results suggest that the reliability of prediction for *E.coli* RNA need not indicate the reliability of prediction for other large RNAs of unknown structure.

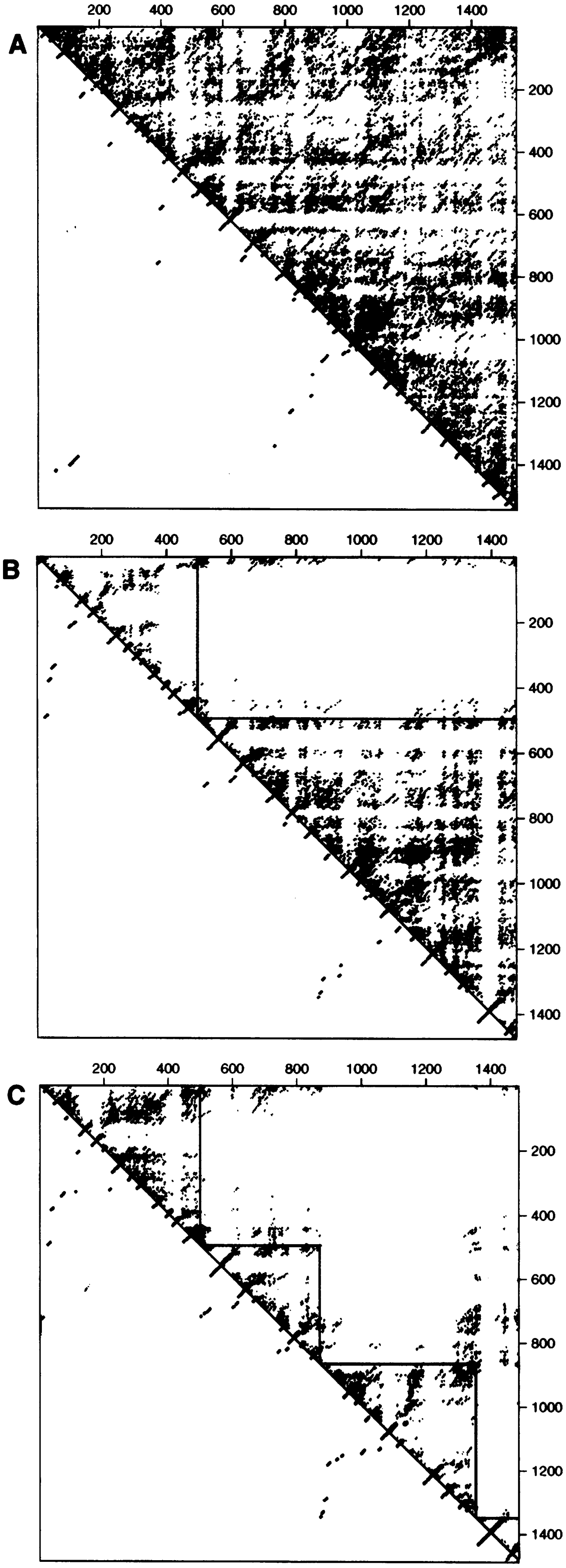

Figure 3. 'Overlaid energy dot plots' of three rRNAs. Predicted optimal base pairs are plotted as black dots and are shown both in the lower left and upper right triangles of the plot. Suboptimal base pairs are shown in grey in the upper right triangle of the figure, with darker to lighter dots corresponding to suboptimal foldings within 0–4, 4–8 and 8–12 kcal/mol respectively from the computed minimum energies. The base pairs of the comparative model are plotted as large red or blue dots in the upper right triangle. Blue dots represent base pairs found both in the comparative model and in an optimal folding. Red dots represent base pairs in the comparative model that are missing from the predicted optimal foldings. In addition, the plots have been annotated with horizontal and vertical black lines at domain boundaries.The energy dot plots were filtered. All possible optimal base pairs are shown, but only suboptimal helices of ≥3 bp are plotted. This eliminates isolated and doublet base pairs in suboptimal structures. (**A**) *E.coli* 16S rRNA, (**B**) *H.volcanii*, (**C**) *T.celer.*

Table 5. Predicted structural domains versus comparative models

Sequence				
Anacystis nidulans 16S	None			
Bacillus subtilis 16S	None			
Desulfovibrio desulfuricans 16S	None			
Escherichia coli 16S	None			
Heliobacterium chlorum 16S	None			
Haloferax volcanii 16S	1–494	495–1474		
	1–494	495–1474		
Mycoplasma gallisepticum 16S	1–584	585–939	940–1323	1324–1519
	1–554	555–915	916–1371	1372–1519
Methanospirillum hungatei 16S	1–487	488–882	883–1466	
	1–478	488–856	857–1466	
Thermococcus celer 16S	1–498	499–853	854–1352	1353–1486
	1–498	499–853	854–1352	1353–1486
Thermoproteus tenax 16S	1–515	516–1370	1371–1503	
	1–515	516–1370	1371–1503	
Thermomicrobium roseum 16S	None			
Thermoplasma acidophilum 16S	None			
Thermus thermophilus 16S	None			
Thermotoga maritima 16S	None			
Zea mays chloroplast 16S	None			

For each rRNA, the nucleotide positions of predicted structural domains are shown in the first row and nucleotide positions of structural domains taken from the corresponding comparative model are shown in the second row.

The studies with ribosomal RNAs support our published analysis by energy dot plot of the sequence of coliphage Qβ RNA (21). In the coliphage study we reported the presence of five well-determined structural domains that were consistent in position with unusually stable structural features that are visualized by electron microscopy in Qβ RNA. In the coliphage studies we also reported the presence of numerous local well-determined hairpins that were consistent in structure with our studies in solution by chemical modification (22). Experimental studies with native and mutant coliphage RNAs suggest that some of the potential structural heterogeneity that is visualized in energy dot plots correlates with helix stability; that weak structures are more likely to vary in conformation (21; unpublished studies). Energy dot plots of the ribosomal RNAs have many poorly-determined regions. The relative stability of individual structural features within the comparative secondary structure models for ribosomal RNAs have not been measured. It remains to be seen whether any structures in rRNAs that are poorly-determined by energy dot plot analysis are less stable than well-determined structures.

Le and colleagues (23–25) identify their equivalent of 'well-determined' regions by finding segments of an RNA molecule where the folding energy is much less than expected at random. These methods cannot determine separate folding domains in large molecules, nor can they separate 'well-determined' regions from poorly determined ones. Our point of view is different, because it is clear to us that random RNA will have structure and that parts may indeed turn out to be 'well-determined' by our dot plot analysis. We would expect such structures to form reliably in solution if the random RNA could be synthesized. In the folding of 50 randomly generated RNA sequences of size 1500 with equal expected A:C:G:U content (M. Zuker, unpublished results) we found many 'well-determined' regions.

Our current folding model allows only the six canonical and G:U pairings. Although these account for the vast majority of base pairs derived by comparative sequence analysis, there is clear evidence for others, especially A:G, A:C, A:A, C:C, G:G and U:U (3,26,27). Although these could be incorporated into the current folding program, there is still a lack of sufficient measurements to quantify the effects of all the different base pair stackings that would become possible. In addition, results for selected pairs of mismatches within helices containing otherwise canonical base pairs indicate that simple nearest neighbor rules are insufficient to explain the resulting stabilities (27). Ultimately, we would expect better results with a folding model that allowed these non-canonical base pairs.

The comparative results (Fig. 3), as well as unpublished studies, show that long-range base pairs are harder to predict reliably than short-range ones. One reason for this may be that *mfold* fails to predict many multi-branched loops correctly, in part because of the simplistic assumptions used in computing their energies. There is both comparative (1,26) and experimental (28) evidence for co-axial stacking of adjacent helices in multi-branched loops. The experimental evidence is compatible with a co-axial stacking interaction roughly equivalent to nearest neighbor stacking within helices. The reordering of groups of optimal and suboptimal foldings predicted by *mfold* based on re-evaluation of the energy of multi-branched loops to include possible co-axial stacking and the use of the Jacobson–Stockmayer theory (29) improves prediction of RNA structure (28). We expect that the incorporation of more realistic energy rules in multi-branch loop energy computation would improve RNA secondary structure prediction. What is not clear at this time is whether the inclusion of non-canonical base pairs and better energy rules for multi-branched loops in *mfold* will improve the prediction of 'well-defined' domains.

In addition to improving the energy rules for long-range structure prediction, studies are underway to analyze a larger set of 16S and 23S rRNAs. Although the studies described here support our early report regarding the significance of well-determined structural domains in genomic RNA from the RNA coliphage Qβ, additional analyses of a larger set of both 16S and 23S rRNA are needed in order to quantify the reliability of prediction for large domains. We note that large domains consist of both local and long-range interactions. In view of the uncertainties in long-range prediction, we anticipate that the reliability of domain prediction will correlate with the number of well-determined local hairpins within the domain. Although the predicted energy dot plots of many RNAs lack well-determined domains, they are found in the predicted plots from a variety of organisms, including genomic animal virus RNAs and eukaryotic mRNAs (unpublished studies). The broad distribution of these structures in RNAs from a variety of organisms encourages us to undertake more extensive analyses of their properties.

ACKNOWLEDGEMENTS

MZ thanks the Canadian Institute for Advanced Research for fellowship support. ABJ was supported by grants DMB9010123 and MCB9316501 from the National Science Foundation and grant GM48702 from the National Institutes of Health. We thank R.Gutell for providing secondary structure models.

REFERENCES

1 Woese,C.R., Gutell,R., Gupta,R. and Noller,H.F. (1983) *Microbiol. Rev.*, **47**, 621–669.
2 Winker,S., Overbeek,R., Woese,C.R., Olsen,G.J. and Pfluger,N. (1990) *Comput. Appl. Biosci.*, **6**, 365–371.
3 Gutell,R.R., Larsen,N. and Woese,C.R. (1994) *Microbiol. Rev.*, **58**, 10–26.
4 Zuker,M. and Stiegler,P. (1981) *Nucleic Acids Res.*, **9**, 133–148.
5 Sankoff,D., Kruskal,J.B., Mainville,S. and Cedergren,R.J. (1984) In Sankoff,D. and Kruskal,J.B. (eds), *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison.* Addison-Wesley, Reading, UK, pp. 93–120.
6 Zuker,M. and Sankoff,D. (1984) *Bull. Math. Biol.*, **46**, 591–621.
7 Williams,A.L. and Tinoco,I.,Jr (1986) *Nucleic Acids Res.*, **14**, 299–315.
8 Zuker,M. (1986) *Lectures Math. Life Sci.*, **1**, 87–124.
9 Zuker,M. (1989) *Science*, **244**, 48–52.
10 McCaskill,J.S. (1990) *Biopolymers*, **29**, 1105–1119.
11 Jaeger,J.A., Turner,D.H. and Zuker,M. (1990) *Methods Enzymol.*, **183**, 281–306.
12 Freier,S.M., Kierzek,R., Jaeger,J.A., Sugimoto,N., Caruthers,M.H., Neilson,T. and Turner,D.H. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 9373–9377.
13 Turner,D.H., Sugimoto,N., Jaeger,J.A., Longfellow,C.E., Freier,S.M. and Kierzek,R. (1987) *Cold Spring Harbor Symp. Quant. Biol.*, **52**, 123–133.
14 Turner,D.H., Sugimoto,N. and Freier,S.M. (1988) *Annu. Rev. Biophys. Chem.*, **17**, 167–192.
15 Larsen,N., Olsen,G.J., Maidak,B.L., McCaughey,M.J., Overbeek,R., Macke,T.J., Marsh,T.L. and Woese,C.R. (1993) *Nucleic Acids Res.*, **21** (suppl.), 3021–3023.
16 Gutell,R.R. (1993) *Nucleic Acids Res.*, **21**, 3051–3054.
17 Gutell,R.R., Gray,M.W. and SchnareM.N. (1993) *Nucleic Acids Res.*, **21**, 3055–3074.
18 Zuker,M., Jaeger,J.A. and Turner,D.H. (1991) *Nucleic Acids Res.*, **19**, 2707–2714.
19 Zuker,M. (1994) *Methods Mol. Biol.*, **25**, 267–294.
20 Jaeger,J.A., Turner,D.H. and Zuker,M. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 7706–7710.
21 Jacobson,A.B. and Zuker,M. (1993) *J. Mol. Biol.*, **233**, 261–269.
22 Skripkin,E.A. and Jacobson,A.B. (1993) *J. Mol. Biol.*, **233**, 245–260.
23 Le,S.-Y., Chen,J.-H., Currey,K.M. and Maizel,J.V.,Jr (1988) *Comput. Appl. Biosci.*, **4**, 153–159.
24 Le,S.-Y. and Maizel,J.V.,Jr (1989) *J. Theor. Biol.*, **138**, 495–510.
25 Le,S.-Y., Chen,J.-H., and Maizel,J.V.,Jr (1990) In Sarma,R.H. and Sarma,M.H. (eds), *Structure and Methods: Human Genome Initiative and DNA Recombination.* Adenine Press, New York, NY, Vol. 1, pp. 127–136.
26 Gutell,R.R. (1993) *Curr. Opin. Struct. Biol.*, **3**, 313–322.
27 SantaLucia,J.,Jr, Kierzek,R. and Turner,D.H. (1991) *Biochemistry*, **30**, 8242–8251.
28 Walter,A.E., Turner,D.H., Kim,J., Lyttle,M.H., Muller,P., Mathews,D.H. and Zuker,M. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 9218–9222.
29 Jacobson,H. and Stockmayer,W.H. (1950) *J. Chem. Phys.*, **18**, 1600–1606.