RNA (1995), *1*:559–574. Cambridge University Press. Printed in the USA.

A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs

DANIELLE A.M. KONINGS[1] and ROBIN R. GUTELL[1,2]

[1] Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Campus Box 347, Boulder, Colorado 80309-0347, USA
[2] Department of Chemistry and Biochemistry, University of Colorado, Campus Box 215, Boulder, Colorado 80309-0215, USA

ABSTRACT

To increase our understanding of the dynamics and complexities of the RNA folding process, and therewith to improve our ability to predict RNA secondary structure by computational means, we have examined the foldings of a large number of phylogenetically and structurally diverse 16S and 16S-like rRNAs and compared these results with their comparatively derived secondary structures. Our initial goals are to establish the range of prediction success for this class of rRNAs, and to begin comparing and contrasting the foldings of these RNAs. We focus here on structural features that are predicted with confidence as well as those that are poorly predicted. Whereas the large set of Archaeal and (eu)Bacterial 16S rRNAs all fold well (69% and 55% respectively), some as high as 80%, many Eucarya and mitochondrial 16S rRNAs are poorly predicted (~30%), with a few of these predicted as low as 10–20%. In general, base pairs interacting over a short distance and, in particular, those closing hairpin loops, are predicted significantly better than long-range base pairs and those closing multistem loops and bulges. The prediction success of hairpin loops varies, however, with their size and context. Analysis of some of the RNAs that do not fold well suggests that the composition of some hairpin loops (e.g., tetraloops) and the higher frequency of noncanonical pairs in their comparatively derived structures might contribute to these lower success rates. Eucarya and mitochondrial rRNAs reveal further novel tetraloop motifs, URRG/A and CRRG, that interchange with known stable tetraloop in the procaryotes.

Keywords: comparative sequence analysis; ribosomal RNA; RNA secondary structure; thermodynamic RNA folding

INTRODUCTION

Transforming a single RNA sequence into its biologically active three-dimensional structure is a major objective in molecular biology research. This quest involves a multitude of different disciplines, including thermodynamic and biophysical structural chemistry, and molecular, computational, and evolutionary biology. Basic to these efforts are several key issues that need to be understood in more detail. These are: (1) a more complete knowledge of the biologically significant RNA structure motifs and their thermodynamic stabilities; (2) relationships between sequence and structure; (3) the degree of influence thermodynamic and kinetic factors have on the folding of RNA molecules; (4) the contribution of ancillary factors (e.g., proteins, RNA, RNPs) contribute to the RNA folding pathways; and (5) to what extent RNA molecules undergo conformational rearrangements (e.g., folding intermediates) during their initial folding and in their functional states.

The potential complexity of RNA structures and their folding, and our lack of detailed information on a number of the issues raised here, would suggest that any attempt to fold an RNA sequence would produce a biologically meaningless answer. However, several RNA folding algorithms are reporting improvements in their ability to predict a secondary structure that has been solved by comparative methods (e.g., Abrahams et al., 1990; Zuker et al., 1991). The improvements observed from these algorithms are based on refinements in the folding algorithm and additional thermodynamic energy values for different RNA motifs (e.g., tetraloops;

Reprint requests to: Robin R. Gutell, Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Campus Box 347, Boulder, Colorado 80309-0347, USA; e-mail: robin.gutell@colorado.edu.

Jaeger et al., 1989; Woese et al., 1990b). These results suggest that additional changes in these folding algorithms and energy values will further improve our ability to predict comparatively derived RNA secondary structures. With the goal to enhance our understanding of how an RNA molecule folds and to improve these RNA folding algorithms, we have initiated a more detailed study to ascertain the strengths and weaknesses of the most widely utilized thermodynamic RNA folding program (Zuker, 1989). The information derived from these studies should suggest additional alterations in the RNA folding algorithm and encourage the experimental investigation of those structural motifs that are not well predicted. Although a related iterative process has already improved our ability to predict some RNA structures (e.g., stable tetraloop motifs), it is our strong belief that this approach can further be utilized to investigate and understand the thermodynamic, kinetic, and other factors that influence RNA folding. With these objectives in mind, we analyze here a larger collection of 16S and 16S-like rRNA folded structures.

Described here is our folding analysis of a large number of structurally and phylogenetically diverse 16S and 16S-like rRNAs. These comparatively derived secondary structure models are treated as the gold standard. We evaluate how well a thermodynamic-based RNA folding algorithm predicts these given secondary structures. The success of the foldings are determined for the entire rRNA molecule, for short- and long-range base pairs, and for base pairings in proximity with different structural features.

RESULTS

This study evaluates the performance of a thermodynamic-based folding algorithm. Our aims are to determine how well this algorithm folds a large and structurally diverse set of 16S rRNA sequences by comparing these folded secondary structures with the analogous set of comparatively derived secondary structure models (Gutell, 1994). Previously, smaller sets of rRNAs (*Escherichia coli*, (eu)Bacteria; *Haloferax volcanii*, Archaea; *Rattus norvegicus*, rat mitochondria; *Clamydomonas reinhardtii*, chloroplast) were compared in this manner, resulting in folding scores for helices that range from 57% to 65% in the absence of coaxial stacking contributions (Walter et al., 1994). Here, we analyze 56 16S rRNA sequences (Table 1), which include representatives from the five phylogenetic groups. Initially, we evaluate how well the entire 16S rRNA structure is predicted across this large collection of 16S rRNA structures. Subsequently, we evaluate how well specific classes of structural elements for these 16S rRNA structures are predicted. In comparing the data presented here and those reported in earlier folding

TABLE 1. Selection of 16S rRNA analyzed.

Archaea (8)	
Euryarchaeota	*Halobacterium marismortui* (HC10), *Haloferax volcanii*, *Methanobacterium formicicum*, *Methanococcus vannielli*, *Thermococcus celer*, *Thermoplasma acidophilum*
Crenarchaeota	*Sulfolobus solfataricus*, *Thermoproteus tenax*
(eu)Bacteria (15)	
Thermotoga	*Thermotoga maritima*
Deinococcus + relatives	*Thermus thermophilus*
Spirochetes + relatives	*Borrelia burgdorferi*
Chlamydia	*Chlamydia psittaci*
Cytopha.-Flexi.-Bacteroides	*Bacteroides fragilis*
Purple bacteria	*Agrobacterium tumefaciens*, *Pseudomonas testosteroni*, *Escherichia coli*, *Desulfovibrio desulfuricans*
Cyanobacteria	*Synechococcus* sp. 6301
Gram-positive	*Arthrobacter globiformis*, *Frankia* sp., *Bacillus subtilis*, *Mycoplasma hyopneumoniae*, *Mycoplasma gallisepticum*
Chloroplast (11)	
Protoctista	*Chlamydomonas reinhardtii*, *Chlorella vulgaris*, *Cryptomonas* sp., *Olisthodiscus luteus*, *Astasia longa*, *Euglena gracilis*, *Cyanidium caldarium*, *Palmaria palmata*
Plantae	*Nicotiana tabacum*, *Zea mays*, *Marchantia polymorpha*
Mitochondria (7)	
Protoctista	*Paramecium tetraulia*
Fungi	*Saccharomyces cerevisiae*, *Aspergillus nidulans*
Plantae	*Zea mays*
Animalia	*Bos taurus*, *Ascaris suum*, *Caenorhabditis elegans*
Eucarya (15)	
Archezoa	*Hexamita*, *Giardia muris*, *Giardia ardeae*, *Giardia intestinalis*, *Encephalitozoon cunuculi*, *Vairimorpha necatrix*
Protoctista	*Babesia bigemina*, *Gracilariopsis* sp., *Tritrichomonas foetus*
Fungi	*Saccharomyces cerevisiae*, *Cryptococcus neoformans*
Animalia	*Placopecten magellanicus*, *Xenopus laevis*, *Mus musculus*, *Homo sapiens*

studies (Jaeger et al., 1989; Zuker et al., 1991; Walter et al., 1994), one should keep in mind that Zuker and coworkers (1) utilized older comparatively derived structures that do not incorporate a number of recent minor refinements; and (2) performed foldings for the three known 16S rRNA domains separately, which improves the overall accuracy of prediction (in agreement with our own data, not shown).

Overall prediction of base pairs and helices

We investigate first how well the thermodynamic algorithm predicts the overall 16S rRNA structures for a structurally and phylogenetically diverse set of 16S and 16S-like rRNA secondary structures. To help ascertain an optimal method for calculating the folding success, predictions are calculated for two different structural units, base pairs and helices (see the Materials and methods for details). The folding prediction values for these two counting methods are similar in the averaged percentages within each of the five phylogenetic groups (Fig. 1). Thus, for the remainder of these studies, we utilize the base pair counting method.

Some of our results in Figure 1 are unexpected given the relatively good predictions in previous folding studies on a limited number of 16S rRNA structures (e.g., Zuker et al., 1991; Walter et al., 1994). Although our thermodynamic foldings on a larger number of Archaea and (eu)Bacteria are consistently well predicted, and similar to the values obtained previously for a smaller sampling (e.g., Zuker et al., 1991; Walter et al., 1994), our foldings of a phylogenetically diverse set of mitochondrial and Eucarya (nuclear) structures reveal a larger range of folding success and, on average, score significantly less than those for the Archaea and (eu)Bacteria. The predictabilities for chloroplasts have a broad range and, on average, are intermediate in success values between (eu)Bacteria and the mitochondria and Eucarya. The average and their low and high prediction values (calculated with the base pair counting method) for each of the five phylogenetic groups are: Archaea (69%, 55%, 81%); (eu)Bacteria (55%, 39%, 69%); chloroplast (48%, 32%, 71%); mitochondria (31%, 17%, 56%); and Eucarya (30%, 10%, 47%). The key observations for this analysis are: (1) For individual sequences, the Archaea *H. volcanii* has the best score at 81%, whereas the lowest score is for the Eucarya *Encephalitozoon cunuculi* at 18%. (2) Archaea 16S rRNAs are predicted, on average, over twice as well as the mitochondrial and Eucarya 16S-like rRNAs. (3) However, the highest individual scores within the chloroplast, mitochondria, and Eucarya are similar to scores in the two highest groups—Archaea and (eu)Bacteria. The ranges of folding success within the phylogenetic groups vary, with the greatest range in the chloroplast (39%) and mitochondria (39%), followed by Eucarya (37%), (eu)Bacteria (30%), and Archaea (26%).

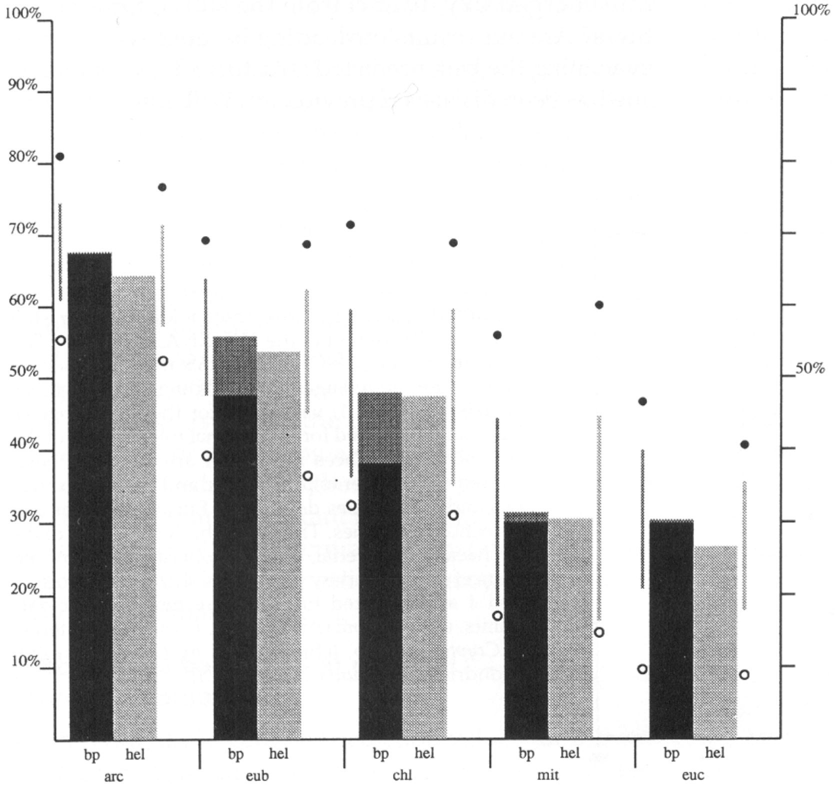

FIGURE 1. For the set of 16S rRNA sequences, secondary structure models proposed with a thermodynamic folding algorithm (Zuker, 1989) and with comparative sequence analysis (Gutell et al., 1992, 1994) are compared. We evaluate how well the folding algorithm predicts the comparatively derived secondary structure base pairs (bp) and helices (hel, as defined by Zuker et al., 1991). These two prediction values, shown with vertical bars, are determined for each individual sequence, and averaged within each phylogenetic group (see Table 1). The values shown are for the optimal folding. The thermodynamic folding for the bp prediction analysis is performed with and without the additional contribution for extra-stable tetraloops (i.e., −2.0 kcal, see Table 3). The upper gray portion of the bp bar denotes the folding success when the tetraloop thermodynamic bonus is included, whereas the lower black bar does not include this folding bonus. Standard deviations from the average prediction values are shown with a vertical line adjacent to their bar and pointing vertically from the average value of each group (i.e., half the length of the line represents the standard deviation). The values for the sequence with the highest and lowest predictions within each phylogenetic group are represented with closed (highest) and open (lowest) circles that are aligned to the standard deviation lines. arc, Archaea; eub, (eu)Bacteria; chl, chloroplasts; mit, mitochondria; euc, Eucarya nuclear.

TABLE 2. Length and base composition (GC%) of the five phylogenetic groups of studied 16S rRNA.

	Arc	Eub	Chl	Mit	Euc
Average length	1,480	1,530	1,496	1,302	1,630
Standard deviation of length	12	21	17	477	208
Minimal length	1,466	1,488	1,474	697	1,250
Maximal length	1,503	1,562	1,537	1,962	1,870
Average GC%	59	54	51	35	53
Standard deviation of GC%	5	6	5	10	9

GC content and length of 16S rRNAs

Is there a simple explanation for the large range of folding success and the phylogenetic group-specific results? There are a couple of obvious parameters, such as base composition and sequence length, that can have a direct influence on the folding potential. Because extreme biases in GC content (either very high AU or GC percentages) can greatly diminish accurate prediction (unpubl. results), we examine the GC percentages for the RNA structures under study to determine if there is a relationship between GC content and our folding predictabilities. The average base composition for the Eucarya group is very similar to that of (eu)Bacteria and Archaea (GC: ~55%), although the Eucarya have a larger variation in their base compositions (Table 2; Fig. 2). Mitochondria are, on average, AU rich (GC: 35%) and, like the Eucarya, the individual structures are more diverse in their GC percentages. These values by themselves do not resolve this issue. We further investigate the foldings for the 16S rRNA structures studied here by comparing their GC content with their percentage of base pairs predicted accurately (Fig. 2). The key observations are: (1) The GC contents and prediction values for the prokaryotes (Archaea and (eu)Bacteria) and chloroplasts are moderate; there are no extremely high or low GC percentages (40–67%), nor extremely low prediction values (32–81%). All of the best predictions for this entire study are in this broad class. (2) In contrast, the individual mitochondrial and Eucarya structures analyzed have larger, less constrained GC contents (22–75%) and prediction accuracies (10–56%). (3) There is no apparent relationship between GC contents and predictabilities, although the prediction values are less when the GC contents are at their highest and lowest.

Another factor to consider is the length of the sequence. It can be argued that significantly shorter sequences will be predicted more accurately than longer sequences with the thermodynamic folding algorithm, because shorter sequences have a lower number of competing potential helices than longer sequences. The length values of the studied rRNAs suggest that there is no obvious relationship between predictabilities and the lengths of the entire rRNAs studied here (data not shown). For example, structures with similar lengths (e.g., ~1,460 nt) show predictabilities that vary largely (10–80%). Alternatively, there are examples where the prediction values are similar (~55%), whereas the lengths vary (~600 nt). Furthermore, our shortest sequences (~700 nt) reveal very low accuracies of prediction (17–23%).

Sub-optimal foldings

Is the best or optimal thermodynamically folded structure energetically distinct from the sub-optimal structures? Are our results misleading because we are only evaluating the best predicted structure? This general issue has been discussed previously (Williams & Tinoco,

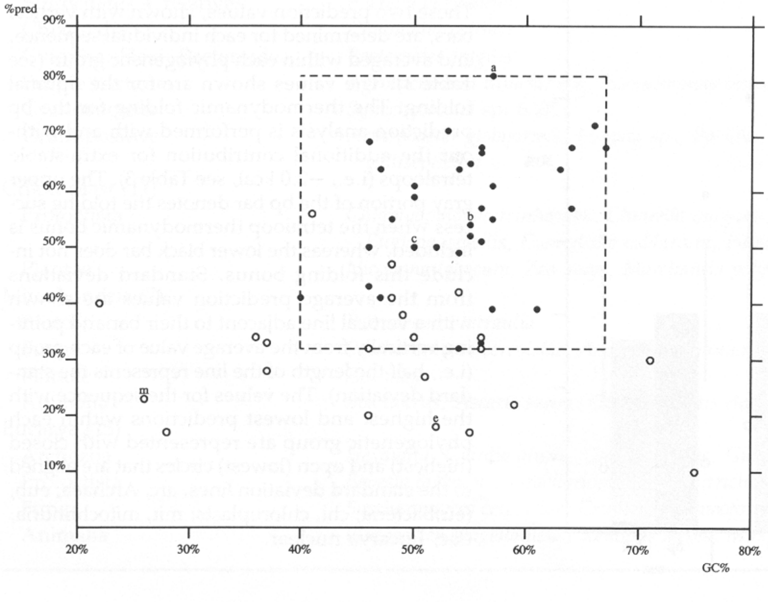

FIGURE 2. Thermodynamic-based folding is contrasted with the GC content of the 16S rRNA sequences. The percentages of (G+C) for the 16S rRNA sequences (x-axis) are plotted against the folding predictabilities (y-axis) (See Fig. 1; values are for the percentage of base pairs predicted for the optimal folding). All of the 16S rRNA sequences in Table 1 are analyzed here. Closed circles refer to prokaryotic and chloroplast values and open circles denote the Eucarya-nuclear and mitochondria values. The dotted box encloses all of the Archaea, (eu)Bacterial, and chloroplast values. For reference, the secondary structures shown in Figures 3 and 4 are indicated in the figure next to their data points. a, *H. volcanii* (Archaea); b, *E. coli* ((eu)Bacteria); c, *Cryptomonas* sp. (chloroplast); m, *C. elegans* (mitochondria); e, *E. cunuculi* (Eucarya).

1986; Jaeger et al., 1989; Konings & Hogeweg, 1989; Zuker et al., 1991) when evaluating the mapping of thermodynamic folding onto otherwise determined structures (e.g., comparative structures). Generally, many alternative thermodynamic foldings are energetically close to the optimal structure, which can be structurally very similar to or distinct from the optimal one. For our studies here, we also examine the prediction accuracy for the sub-optimal structure that is the closest to the comparatively derived structure (see the criteria in the Materials and methods). The optimal and best sub-optimal averaged values for the five structural groups are: Archaea, 69% versus 73%; (eu)Bacteria, 55% versus 64%; chloroplast, 48% versus 52%; mitochondrial, 31% versus 37%; and Eucarya, 30% versus 38%. On the whole, this additional analysis only increases the averaged prediction accuracy from 4 to 9%. Alternatively, this result reveals that the majority of the incorrectly folded part of these rRNAs are still not predicted correctly in one of the sub-optimal folds. However, it is still encouraging that some of the correct foldings that are missed in the optimal structure are present in the sub-optimal folds. In the general case where no comparatively derived reference structure is available, the question remains which of the many sub-optimal folds is the best and what additional factors will help us make better predictions. The availability of chemical and enzymatic modification results, in the rare case, can greatly aid in this discrimination between all potential sub-optimal foldings.

Recently, Walter et al. (1994) reported an extension of the original MFOLD program that incorporates a free-energy contribution for coaxial stacking and some other new structural features (e.g., a revised multi-branch loop function). After selecting a set of sub-optimal foldings with the existing MFOLD algorithm, this new version of the program recalculates their energies based on experimentally derived energy contributions for coaxial stacking. In this recalculation, a new optimal structure can be selected from the original pool of sub-optimal structures. Thus, regarding our analysis, the recalculation with coaxial-stacking contributions can, in principle, improve prediction accuracies up to the "best" suboptimal (i.e., an increase, on average, up to ~10%, see above). These authors show that this new aspect of the method indeed improves the accuracy of predicting a comparative structure by the optimal folding. For the general purpose of RNA folding, this is an important improvement because, in the absence of other structural information, one has no grounds to prefer any particular suboptimal folding over the optimal folding.

Prediction of base pairs at different structural contexts

Underlying these general prediction values for the 16S rRNA structures are discrete base pairs that are differentially predicted. A visual inspection of a few phylogenetically and structurally diverse secondary structures (*E. coli* [Fig. 3], *H. volcanii* [Fig. 4], Archaea; *Cryptomonas* sp., chloroplast; *Caenorhabditis elegans*, mitochondrion; *Encephalitozoon cunuculi*, Eucarya) reveals a few general patterns. For *E. coli*, where 54% of this 16S rRNA is folded correctly, the majority of the short-range helices (i.e., distance < 100 nt) are properly predicted, whereas the majority of the long-range interactions (i.e., distance > 100 nt) are not predicted. The same pattern is true for the chloroplast *Cryptomonas* sp., where 50% of the base pairs are folded correctly. The Archaea, *H. volcanii*, has the highest folding percentage at 81%. Here the percentage of correctly predicted short- and long-range base pairs are both higher. Only a few helices are not predicted; the majority of these are the long-range base pairings in the structural domain 3. The percentage of the rRNA properly predicted is very low in the mitochondrion, *C. elegans*, and the Eucarya, *E. cunuculi*, at 23% and 18%, respectively. For these two examples, the majority of the small number of helices properly predicted are short range. Only one long-range helix is predicted in this situation. These visually qualitative observations need to be followed up with a quantitative analysis of the entire set of structures folded (see Table 1), and evaluated with different structural parameters.

In our quest to understand why some 16S rRNA structures are predicted better than others, we begin distinguishing which structural elements are predicted better than others. The results from these enquiries will begin focusing attention on structural features that are not well predicted. For the studies here, we parse the entire 16S rRNA into a few of the more obvious structural elements that are differentially predicted: (1) short- versus long-range base pair interactions, measured in increments of 10 and 100 nt; (2) helical base pairings in proximity to four types of loops—hairpin, internal, bulge, and multistem; and (3) hairpin loop sizes.

Predicting base pairs separated with different length increments

Base pairs are separated by varying numbers of nucleotides. The distances separating the two interacting nucleotides are divided into increments of 100 nt up to the maximum base pair distance (see Fig. 5A and legend), and into increments of 10 nt for short-range interactions (see Fig. 5B and legend). The following conclusions can be drawn from the analysis in Figure 5A. (1) The majority of the base pairs in the comparative structure are separated by less than 100 nt (see Fig. 5A insert). (2) Short-range base pairs (distance < 100 nt) are predicted better than long-range interactions (distance > 100 nt) within each of the phylogenetic classes. (3) The best prediction values for the short-range base pairs is for the Archaea, followed by the (eu)Bacteria,

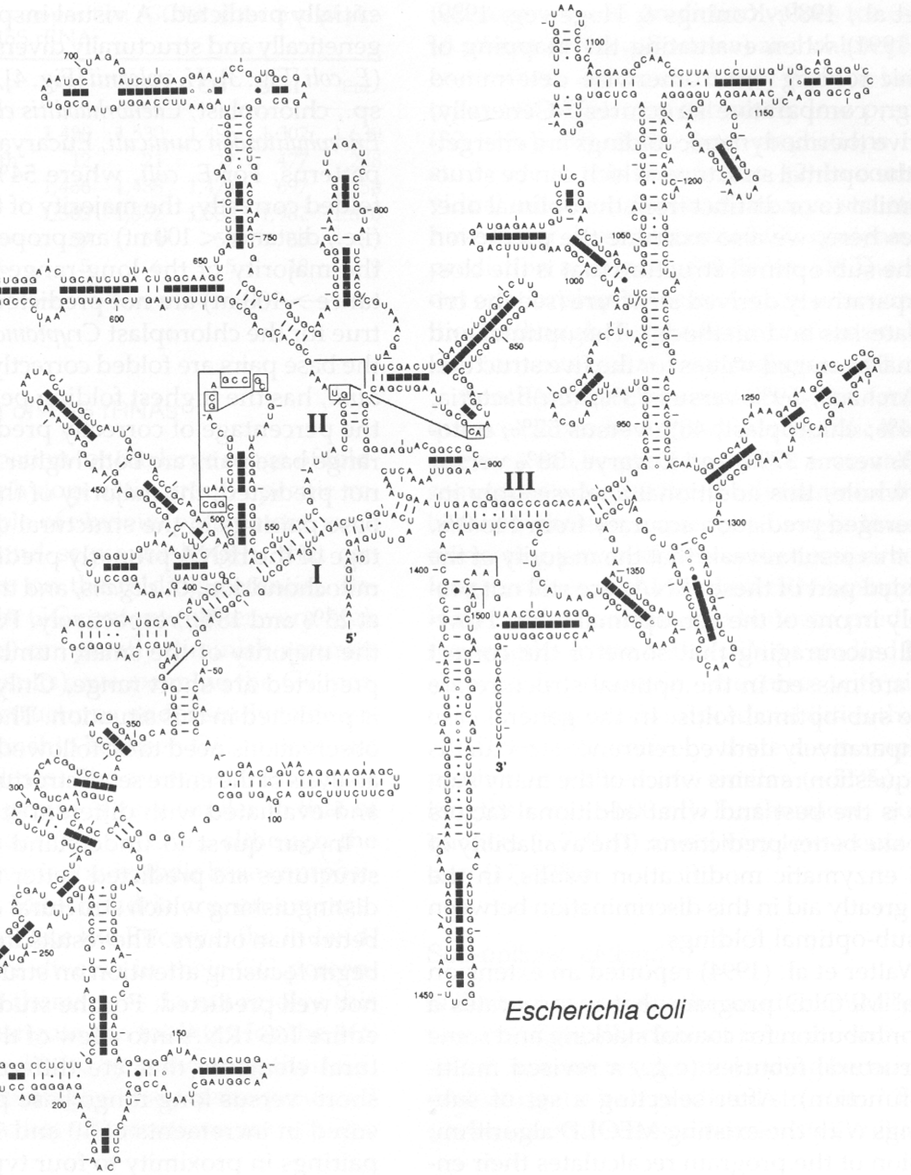

FIGURE 3. Secondary structure base pairs that are predicted as the optimal thermodynamic folding are indicated on the *E. coli* 16S rRNA comparative structure diagram with thick tick marks. This figure is shown as an example of the complexity of these rRNA higher-order structures and the types of structural elements that are well predicted. Tertiary interactions, including pseudoknots and noncanonical base pairs, are shown here, although the folding algorithm used in these studies does not predict these types of structures.

chloroplast, Eucarya, and mitochondria. (4) Overall, the accuracies of prediction for each of the length increments show the same ordering as described above: Archaea > (eu)Bacteria > chloroplast > mitochondria = Eucarya. (5) The predictability of the long-range base pairings (>200 nt) decreases with increasing distance. Very few base pairings separated by more than 300–400 nt are predicted in the Eucarya and mitochondria. The Archaea have the highest prediction values for these longer-range interactions.

Because more than 75% of the base pairs in the comparatively derived 16S rRNA structures are separated by less than 100 nt (see inset in Fig. 5A), the 3–100-nt length increment in Figure 5A has been expanded into length increments of 10 (Fig. 5B). The most significant conclusions from Figure 5B are: (1) The majority of the base pairs in the comparative structure are separated by less than 50 nt or so (see inset in Fig. 5B). (2) Within each phylogenetic group, the larger length increments are not predicted as well as the smaller base pair distances. (3) Within the Archaea group, the first eight length increments are predicted at approximately 80%, with a small decrease in prediction at lengths 90 and 100. The smallest length increments for the (eu)Bacteria and chloroplast groups have prediction values slightly less than those in the Archaea. Within the mi-

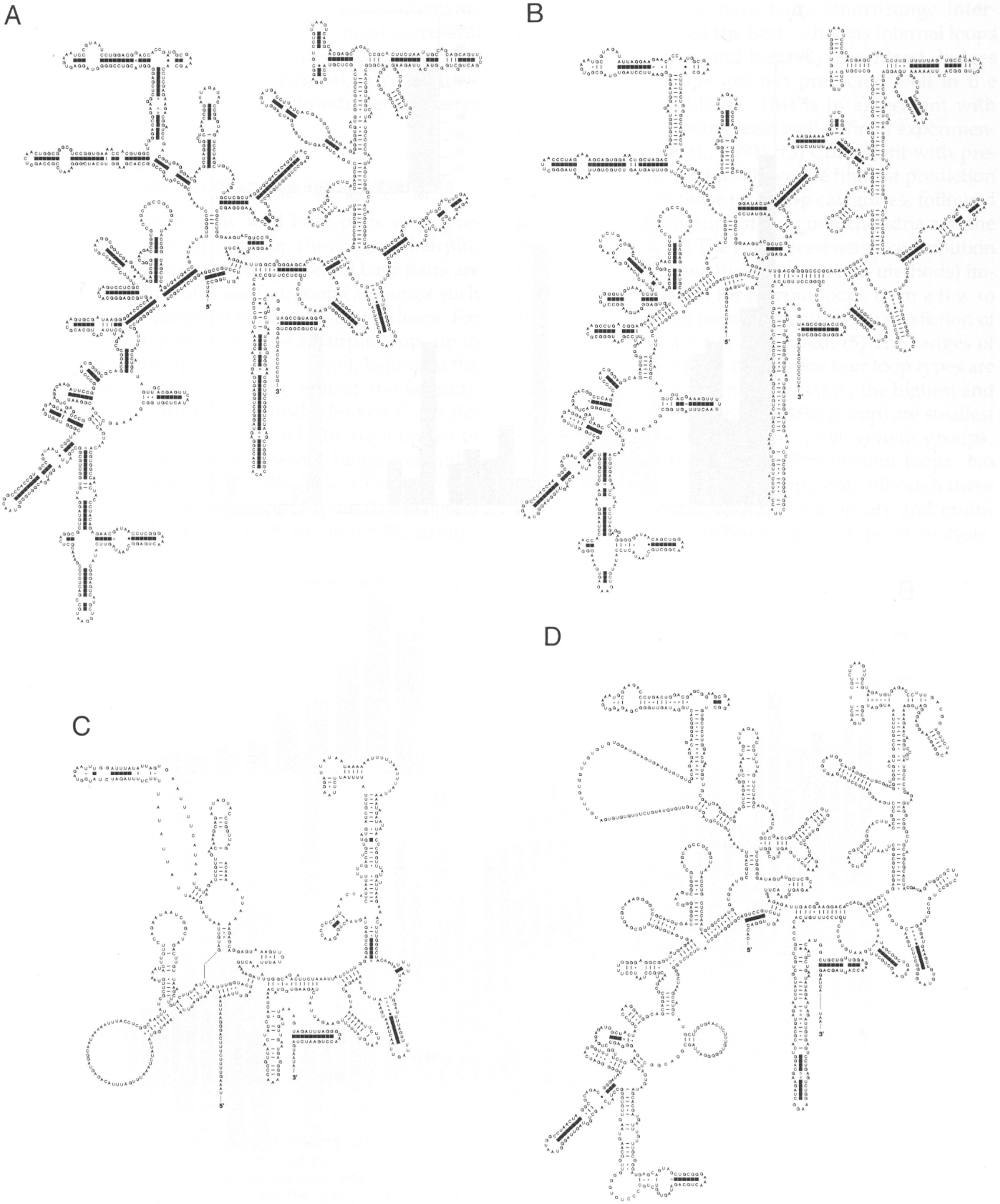

FIGURE 4. Four additional examples of the optimal thermodynamic folding superimposed on the comparatively derived 16S rRNA secondary structure diagram. Secondary base pairs predicted by the optimal thermodynamic folding are indicated with thick tick marks. Tertiary interactions identified with comparative methods are not shown on these diagrams, except for the pseudoknot interaction that closes domain I and II. **A:** *H. volcanii*, Archaea. **B:** *Cryptomonas* sp., chloroplast. **C:** *C. elegans*, mitochondria. **D:** *E. cunuculi*, Eucarya. A complete collection of these secondary structure diagrams is accessible via our WWW URL address: URL:http://pundit.colorado.edu:8080/root.html (see also Gutell, 1994).

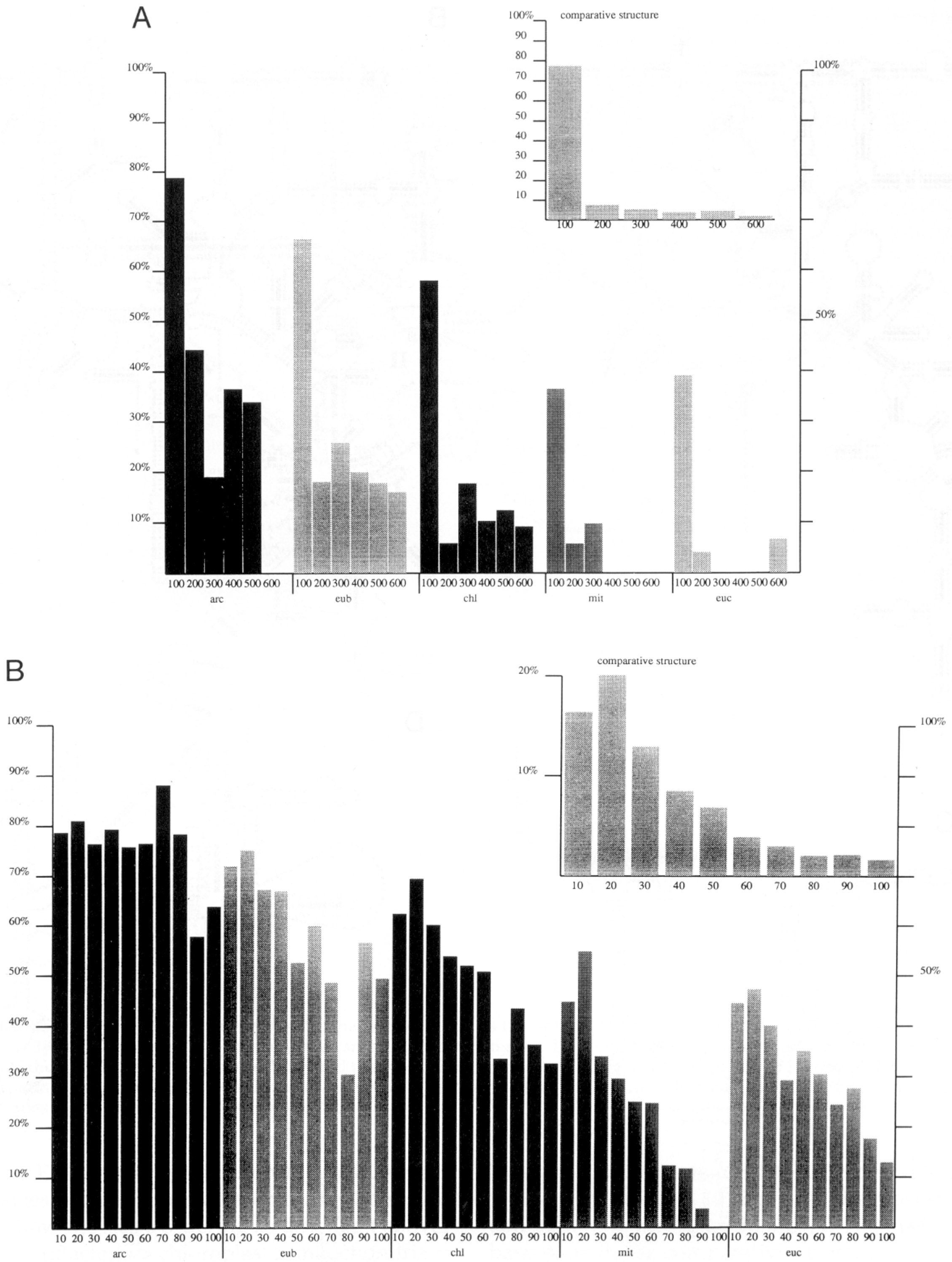

FIGURE 5. The thermodynamic prediction of the comparatively derived base pairs is evaluated for different sequence lengths spanning the two halves of each predicted base pair. This evaluation is performed for all of the analyzed 16S rRNA secondary structures (Table 1) and grouped into the five phylogenetic groups. **A:** These distances are grouped into increments of 100 nt, producing six categories [3–100 nt (100); 101–200 nt (200); 201–300 nt (300); 301–400 nt (400); 401–500 nt (500); and 501 nt and greater (600)]. The inset shows the contribution of each distance class in the (eu)Bacterial comparative structures (i.e., nearly 80% of the base pairs in a typical (eu)Bacterial structure are separated by less than 100 nt). This distribution is very similar for each of the phylogenetic groups, thus only the (eu)Bacteria distribution is shown. **B:** An expanded view for the short-range interactions spanning from 3 to 100 nt, in increments of 10 nt. This evaluation is performed for 10 increments (3–10 nt (10), 11–20 nt (20), ... 91–100 nt (100). The inset shows the average contribution of each distance class for the (eu)Bacterial comparative structures. Distributions for the other phylogenetic groups are very similar. See legend to Figure 1 for additional information on the presentation of data.

tochondria and Eucarya, the smallest size increments are predicted at about 50%. Thus, the most successful predictions occur for the shortest distances separating the two paired nucleotides. In parallel, Archaea have the highest scores, whereas mitochondria and Eucarya have the lowest.

Predicting base pairs in different loop environments

Next, we investigate how well base pairs were predicted based on the type of loop they close – hairpin, internal, bulge, or multistem. All helical base pairs are associated with one of these four loop categories such that a base pair is assigned to the loop that it closes. For example, all base pairs that close a hairpin loop, up to the first helix interruption (e.g., a bulge), belong to the "hairpin loop" category (see Fig. 6 inset and legend). As an example, in the comparatively derived *E. coli* 16S rRNA secondary structure model, the frequencies of base pairs that close hairpin, internal, bulge and multistem loops are, respectively, 35%, 26%, 19%, and 20%. The most significant results from this analysis (Fig. 6) are as follows. (1) Within each phylogenetic group, hairpin loop-closing base pairs (short-range interactions) are predicted the best, whereas internal loops are predicted the second best. (2) In contrast, bulges and multistem loops are not predicted well in the thermodynamic foldings. This is in agreement with their free energies being least well-defined experimentally (see Zuker et al., 1991). (3) Consistent with previous trends, the Archaea have the highest prediction values for each of these four loop categories, followed by (eu)Bacteria, chloroplast, and mitochondria and the nuclear Eucarya. (4) The extra free-energy contribution for the tetraloops (see the Materials and methods) improves the prediction for hairpin loops from a few to 10%. The analogous improvements in the prediction of other loop types are not calculated. (5) The ranges of prediction for base pairs closing the four loop types are quite variable. Overall, the ranges (i.e., the highest and lowest scores within a phylogenetic group) are smallest for the hairpin loops in the five phylogenetic groups, and greatest for the bulge and multistem loops. No simple explanation is readily apparent, although these results do help emphasize the complexity and multiple dimensionalities inherent in this type of analysis.

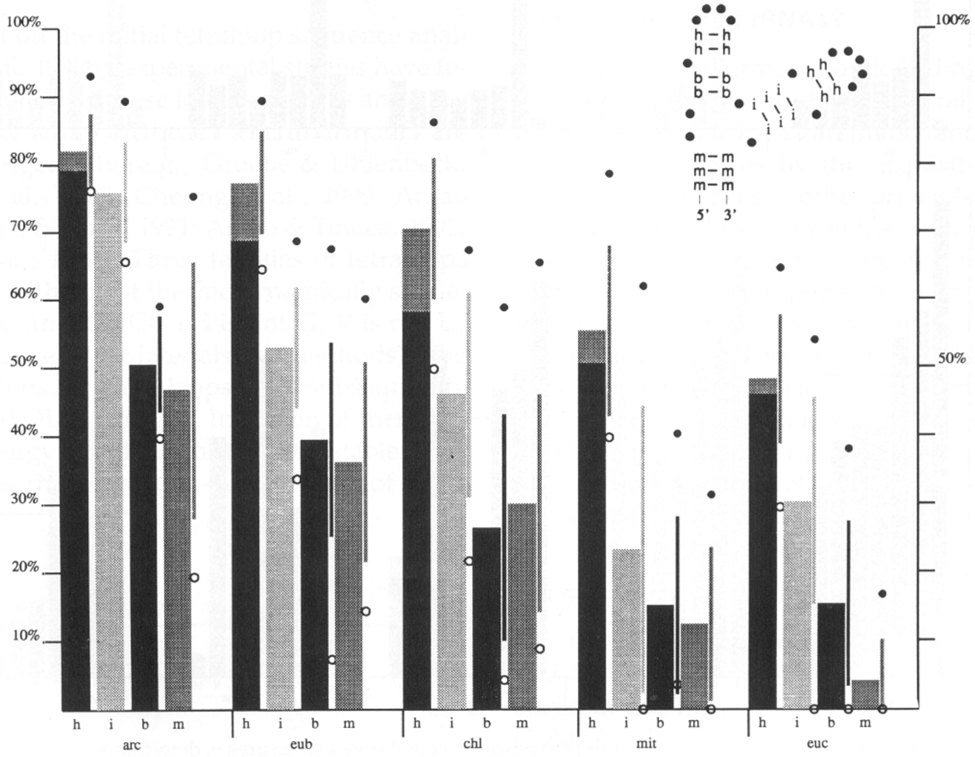

FIGURE 6. Thermodynamic prediction of comparatively derived base pairs is evaluated by the type of loop associated with them. All secondary structure base pairs occurring in uninterrupted helices are grouped into one of following loop classes: hairpin (h), internal (i), bulge (b), or multistem (m). For example, all base pairs that close a hairpin loop, up to the first new loop interruption (e.g., bulge), are classified under hairpin loop. For hairpin loops (h bar), the analysis is performed with (upper grey) and without (lower black) the additional contribution for extra stable tetraloops. See legend to Figure 1 for additional information on the presentation of data. The inset is a key explaining the association between base pairs and their loop.

Prediction of hairpin loops of different size

These studies have shown that base pairings that are in close proximity with hairpin loops are predicted with the highest accuracy. We now investigate hairpin loops of different sizes and compositions to help us ascertain similarities and differences in the prediction of base pairs in association with these hairpin loops. Because these hairpin structures are potential nucleation sites during the folding of an RNA molecule, they might well have an important function in the kinetic and thermodynamic folding process.

Many interesting questions can be put forward with regard to the folding of hairpin loops. For our initial studies, we address some of the more general issues: (1) the prediction of hairpin loops of different lengths with thermodynamic-based algorithms; and (2) the length distributions and sequence compositions of some hairpin loops in comparatively derived structures.

Predicting hairpin loops

Although the average number of hairpin loops per comparatively derived structure (calculated for each 1,000 nt) varies between 18 and 21 across the five phylogenetic groups (data not shown), the average percentage of loops predicted for each phylogenetic group are different (vertical bar in Fig. 7B). (A hairpin loop is considered predicted when the base pair closing that loop is predicted correctly.) Whereas 63% of the hairpin loops are predicted correctly for Archaea and (eu)Bacteria, only 40% and 34% are predicted for mitochondria and Eucarya, respectively. Again, the chloroplasts (54%) fall in between the procaryotes and the Eucarya and mitochondria. The prediction success for different size loops varies. The prediction values for the smaller loops within each phylogenetic group are above the average value for that group, whereas the larger loops tend to be lower than this group average (Fig. 7B).

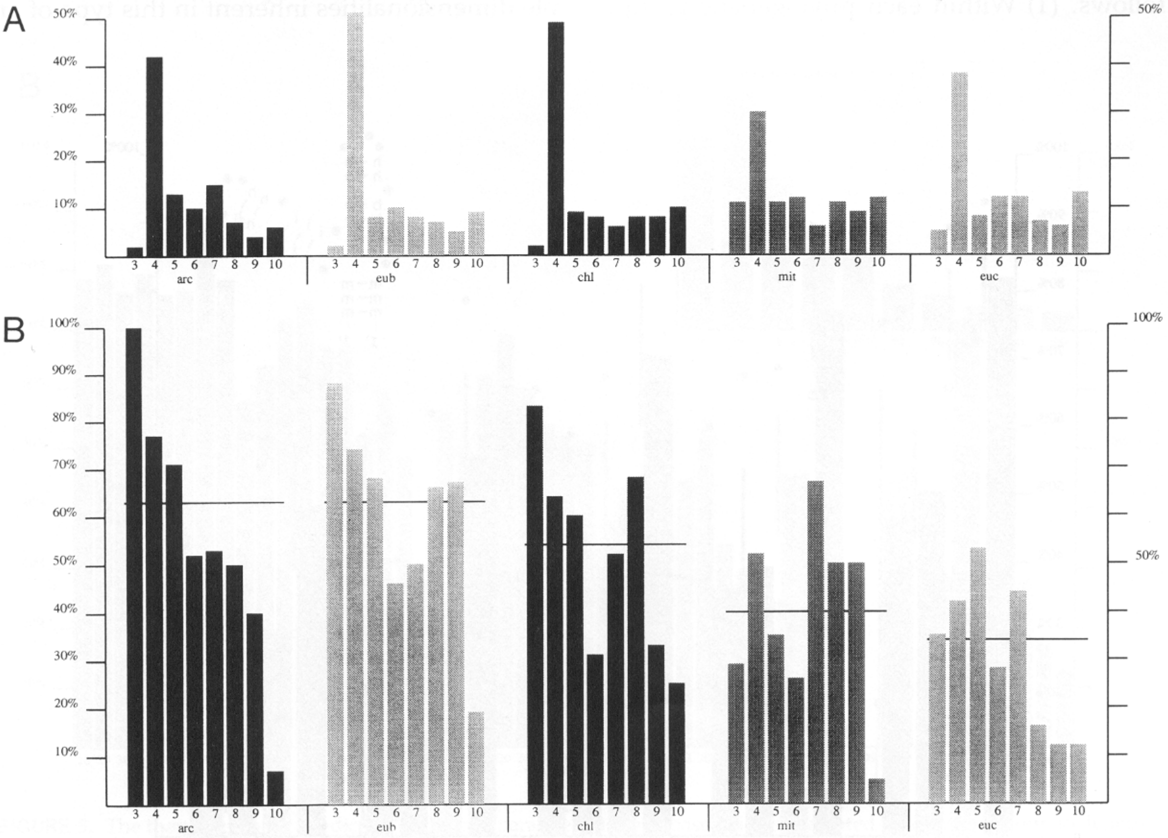

FIGURE 7. **A:** Distribution of hairpin loop sizes in the comparatively derived structures is determined. All of the hairpin loops in the set of 16S rRNA comparative structures are classified according to their length (x-axis) (loops with 10 nt and up are in class 10, the minimum size of a loop is 3). The average distribution of these classes is displayed in the y-axis. The average number of hairpin loops per 1,000 nt is 21 for the Archaea (arc), (eu)Bacteria (eub), and chloroplast (chl) phylogenetic groups, 18 for the mitochondria (mit), and 19 for the Eucarya (euc). **B:** Accurate prediction for hairpin loops are evaluated for all of the structures noted in Table 1. A hairpin loop is considered properly predicted when the closing base pair of that loop is associated correctly by the thermodynamic folding algorithm. This prediction, in percent (y-axis), is averaged for all structures within each of the five phylogenetic groups, and evaluated for eight different loop sizes (x-axis). The horizontal line in each phylogenetic group indicates the average prediction of hairpin loops. See legend to Figure 1 for more additional information on the presentation of data.

Thus, overall, the smaller-sized loops are predicted better than the larger loops.

Contribution of extra stable tetraloops

Tetraloops

Within comparatively derived secondary structures, the most frequent 16S rRNA hairpin loop size in the prokaryotes is four (Woese et al., 1990b). More detailed analysis on a larger number of 16S rRNA structures (Fig. 7A) reveals that the highest frequency of tetraloops (in comparison with all hairpin loops sizes) occurs in the (eu)Bacteria (51%) followed by chloroplasts (49%), Archaea (42%), Eucarya (38%), and the mitochondria (30%). Of the 256 possible sequences for these tetraloops, the majority of the prokaryotic 16S rRNA tetraloops fall into three sequence families: UWCG, GNRA, and CUUG (Woese et al., 1990b). A tetraloop analysis of the more phylogenetically diverse collection of 16S rRNA structures studied here reveals that the first two of these three tetraloop families make up a relatively high percentage of all tetraloops in (eu)Bacteria and Archaea (72–73%), whereas their contribution is lower in the Eucarya and mitochondria (44 and 56%, respectively), and intermediate in chloroplasts (65%) (Table 3).

Based in part on the initial tetraloop sequence analysis (Woese et al. 1990b), experimental studies have focused their attention on these loop sequences and have determined that some of these structural motifs are very stable energetically (e.g., Groebe & Uhlenbeck, 1988; Tuerk et al., 1988; Cheong et al., 1990; Antao et al., 1991; Heus & Pardi, 1991; Antao & Tinoco, 1992; SantaLucia et al., 1992). Three families of tetraloops were found to be the most thermodynamically stable: GHGA, GVAA, and UWCG (H is not G, V is not U, and W is A or U; see the Materials and methods). The free-energy values for these loops were subsequently added to the MFOLD program. Inclusion of these additional free-energy contributions for extra stable tetraloops improves the prediction of a variety of RNA molecules (Hughes et al., 1987; Jaeger et al., 1989; Zuker et al., 1991; see below).

TABLE 3. Distribution of known extra stable tetraloops.

	ghga% 4-lp[a]	gvaa% 4-lp[a]	uwcg% 4-lp[a]	Tot%[b]
Arc16s	12	47	13	72
Eub16s	21	34	18	73
Chl16s	16	39	10	65
Mit16s	7	49	0	56
Euc16s	13	29	2	44

[a] Percentage of the given tetraloop per all tetraloops in the comparative structures.
[b] Total frequency of the three discussed tetraloop classes among all tetraloops.

Prediction improvements with tetraloops

To investigate the contribution of the known stable tetraloops in the folding of a larger sampling of the rRNAs, the predictabilities of comparative structures with and without the extra free-energy contributions of these loops are compared (see the Materials and methods). As expected, the predictabilities of base pairs in general (see Fig. 1) and those closing hairpin loops (see Fig. 6) are higher when the extra stable energy values are included in the calculation for each of the five phylogenetic groups. The largest differences are observed in the (eu)Bacteria and chloroplasts, where the overall prediction of base pairs is improved by 8–10%. In contrast, the improvement is approximately only 1% for the other three groups (Fig. 1; compare Jaeger et al., 1989). These results correlate in part with the different frequencies of stable tetraloops: whereas the (eu)Bacteria and chloroplast rRNAs have the highest percentages of these stable tetraloops, these percentages for mitochondrial and Eucarya are minimal.

Distinct structural features in Eucaryotic and mitochondrial rRNAs?

The differential thermodynamic folding predictabilities of rRNAs of the five studied phyla raises the question to what extent Eucarya and mitochondrial rRNAs distinguish themselves by the acquisition of different structural properties or other principles that influence their proper folding. Within the context of the present work, a couple of rationales are explored here. In line with the suggested importance of hairpin loops and, in particular, tetraloops as a driving force for proper folding as discussed above, one possibility is that there are other stable hairpin loop configurations present in the Eucarya and mitochondrial rRNAs that are not currently known. Another possible explanation for their low predictions could relate to the increase in the number of noncanonical pairings.

Unusual tetraloop sequences

The three tetraloops, occurring at positions 297, 343, and 1516 (*E. coli* numbering), contain unique differences in the Eucarya and mitochondrial 16S rRNAs (Table 4). (These are three examples of this pattern. A more complete analysis of these tetraloops will be presented elsewhere.) GNRA is the predominant tetraloop sequence at positions 297 and 1516 in the Archaea, (eu)Bacteria, chloroplast, and mitochondria. In the Eucarya, the most frequent sequences at position 297 are URGG and CRGG – neither fall into the usual tetraloop families. At position 1516, the Eucarya sequence is

TABLE 4. Identity of tetraloops corresponding to *E. coli* positions 297, 343, and 1516.[a]

Positions/phyla	Pos-297	Pos-343	Pos-1516
Arc	GRRA 100%	UACG 99%	GGAA 100%
Eub	GAGA 99%	UACG 98%	GNRA 80%
Chl	GAGA 97%	UACG 97%	GGAA 100%
Mit	GARA 44%	CACG 54%	GGAA 97%
Euc	URGG 60%	UACG 5%	UGAA 96%
	UAGA 4%	UAAG 19%	
	CRGG 15%	CAAG 74%	

[a] For the identification of these tetraloop sequences, alignments with 77 Archaea, 1,560 (eu)Bacteria, 30 chloroplasts, 117 mitochondria, and 506 Eucarya sequences were used (see the Material and methods).

UGAA, another example of an unconventional tetraloop. The predominant sequences within the mitochondria and the Eucarya at position 343 are, respectively, CACG and (U/C)AAG, and both are different from the more common form present in the prokaryotes. The novel patterns at the three positions can be summarized as URRG/A and CRRG, with an exceptional CACG sequence at position 343 in mitochondria.

Noncanonical base pairings

Visual inspection of the secondary structure diagrams reveals a larger number of noncanonical base pairs in the mitochondrial and Eucarya 16S rRNA secondary structures (we define noncanonical as those base pairs other than G-C, A-U, and G-U). To quantitate this observation, the percentages of canonical and noncanonical base pairs that occur in the comparatively derived structures are analyzed. The mitochondria and Eucarya 16S rRNA do indeed have a higher percentage of noncanonical base pairs, having approximately twice as many as the prokaryotic 16S rRNAs. The ranking of noncanonical base pairs are: Eucarya (11.3%) > mitochondria (11.2%) > chloroplast (7.5%) > (eu)Bacteria (6.6%) > Archaea (5.9%). Thus, not only do the Eucarya and mitochondria have the lowest prediction values and the highest number of noncanonical base pairs, but the percentage of noncanonical pairings for each phylogenetic group is roughly inversely proportional to the overall prediction values for that group, the prediction scores for the five phylogenetic groups are: Archaea (69%) > (eu)Bacteria (55%) > chloroplast (48%) > mitochondria (31%) > Eucarya (30%). Many of the mitochondrial and Eucarya noncanonical pairings are distributed at internal locations in the helix, thus the average length of mitochondrial and Eucarya "canonical" helices are shorter than those helices in the three other phylogenetic groups (data not show). This result could also be one of the reasons why the Eucarya and mitochondrial secondary structures are not as well predicted.

SUMMARY AND DISCUSSION

The present study reports on the comparison of secondary structures derived by thermodynamic folding with the corresponding models derived by comparative sequence analysis. For this comparison, we have analyzed a large and phylogenetically diverse set of 16S and 16S-like rRNA sequences. Summarizing, our most significant findings are: (1) For a structurally diverse collection of 16S and 16S-like rRNAs, the range of folding success is quite broad, spanning over 70% (ranging from 10% for a Eucarya, up to 81% for a Archaea). (2) On average, for the five phylogenetic groups, the Archaea 16S rRNAs are predicted the best, followed by the (eu)Bacteria, chloroplast, mitochondria, and Eucarya. This trend is generally true for the different structural elements under investigation (e.g., the entire molecule, short-range and long-range base pairings, base pairs in proximity to different loops, and for hairpin loops of different sizes). (3) In general, base pairs interacting over a short distance (i.e., separated by less than 100 nt) and, in particular, those closing hairpin loops, are predicted significantly better than long-range base pairs and those closing multistem loops and bulges. In addition, smaller hairpin loops are usually predicted more accurately than larger hairpin loops. (4) The lower prediction success of mitochondrial and Eucarya 16S rRNA structures correlates with lower frequencies of comparatively derived hairpin loops of size four and, in particular, stable tetraloops, and with higher percentages of noncanonical pairings compared to these structural features of the other phylogenetic groups. (5) Eucarya and mitochondrial rRNAs reveal novel tetraloop motifs, URRG/A and CRRG, that occur at positions corresponding with known stable tetraloop in the procaryotes.

Our results, taken all together, show that, at least for some 16S rRNAs, the current thermodynamic-based folding algorithm is predicting the majority of the comparatively derived structure. At this point, we question if we can improve upon the predictions for 16S rRNAs structures (i.e., improvements for those structures that are already well predicted and for those structures that are not well predicted). Will we learn more about RNA folding in the process of trying to answer this question? In what follows, we address some questions related to this issue.

Predicting a base pair: How much of the 16S rRNA structure might we predict?

Although this is not a simple question to answer, we can begin to outline some of the factors that will affect its outcome. In several situations, the RNA folding algorithm will extend a helix beyond the boundaries of the comparative helix. Why is this? Almost all of the secondary structure base pairs in our current model have strong comparative support. In a few cases, these helices can then be extended with additional Watson–Crick base pairs (e.g., 16S rRNA base pairs, *E. coli* numbering: U516/A535, G517/U534; G567/U884, U827/A873, U828/A872; A1046/U1211; U1065/A1191, C1066/G1190, A1067/U1189), although there are no compensatory base changes at these putative pairings, and usually there is a small but significant number of noncoordinated changes resulting in mispairs. Thus, we do not include these in the comparative structure model, although the RNA folding algorithms might well include these pairings if it has already identified the adjacent comparative helix. In addition, several other comparative helices can be extended with regular consecutive antiparallel base pairings to the ends of an existing comparative helix. However, the current comparative secondary structure helix contains unusual base pairings in place of these apparently more stable and regular base pairings (e.g., the more stable 16S rRNA base pairs A243/U283, U244/A282, U245/G281 have been replaced with the noncanonical pair U245/U283; C47/G394 has been replaced with C47/G52; U437/G497 has been replaced with the putative triple (440 497)494, [the adjacent pair 438/497 is unusual—it interchanges from a U-A to G-G]; 570/880 has been replaced with 570/866 and 575/880; 576/765 has been replaced with putative triple 575(579/762) [R. Gutell, unpubl. triple]; 1055/1202 has been taken out, 1056/1201 has been replaced with 1056/1204, 1057/1200 has been replaced with 1057/1203). Thus, we would expect the current folding algorithm to identify the simpler versions of these helices and not predict these unusual base pairings. It should also be noted that the most current version of the 16S rRNA structure model (Gutell et al., 1994; R. Gutell, unpubl.) contains other examples of tertiary-like interactions. All of these unusual and conformationally interesting interactions are beyond the predictability of the current version of the RNA folding algorithm and our limited repertoire of unusual RNA structural motifs and their stability values. We now question how prevalent these unusual pairings are in our comparatively derived structures. Can they be the reason for less than perfect predictions? In one scenario, our predictions should be affected minimally, because these pairings amount to only a few percent of the total number of 16S rRNA base pairs. From this we extrapolate that only a small percentage of the base pairings in the 16S rRNA could be missed with the current thermodynamic-based folding algorithm. The majority of the 16S rRNA base pairs should be predictable. In another scenario, these unusual pairings occur in crucial positions of the RNA secondary structures. As a result, our inability to predict them correctly affects the folding far beyond their local context, i.e., in terms of much larger parts of the structure.

Will thermodynamic stability values for more structural elements improve the folding predictions?

Some of the recent improvements in RNA folding prediction are due to the inclusion of additional thermodynamic energy rules for tetraloops (Jaeger et al., 1989). More recently, thermodynamic data on coaxial stacking of helices have made more incremental improvements in the folding predictions (Walter et al., 1994), again suggesting that additional thermodynamic information will further improve our folding predictions. Where might we expect to find additional thermodynamic data for structural elements that will enhance our prediction scores? The analysis presented here begins to outline a few possibilities. Although a more detailed analysis is necessary to arrive at a comprehensive list of structural features that are and are not well predicted, at this time we can suggest a few structural motifs that should be studied for their thermodynamic stabilities. The analysis presented here reveals that some tetraloop sequences in the Eucarya and mitochondria are different from the conventional tetraloop sequences. Are these as stable as the conventional tetraloops? Results discussed here also reveal a larger number of noncanonical base pairs in the Eucarya and mitochondria 16S rRNAs. Are these base pairs destabilizing and randomly assorted in helices, or might they be placed in a specific context in the helix and comparable in stability to a Watson–Crick base pair? Alternatively, these noncanonical pairing juxtapositions and unconventional tetraloops might be recognition sites for proteins and are truly unstable. Furthermore, the frequencies of some structural features in comparatively derived structures are proportional to their thermodynamic stabilities (e.g., the frequencies of overall base pairs, D.A.M. Konings & R. Gutell, unpubl. data, and the frequencies of tetraloops). This fact suggests that further determination of frequencies for sequences associated with different structural motifs may reveal other frequent sequence/structure motifs that are energetically more stable.

Is the folding of rRNAs dependent on other factors and other folding principles?

It goes beyond the scope of the present study to discuss in any detail other factors that may contribute to the RNA folding process. Nevertheless, to maintain a proper prospective, we would like to touch upon a few of them here. First, currently not much is known about the kinetics of RNA folding, although two recent publications (Emerick & Woodson, 1994; Zarrinkar & Williamson, 1994) are beginning to shed some light on how kinetic principles could affect RNA structure. The group I intron from *Tetrahymena* LSU rRNA forms a series of metastable intermediates during its folding. Individual helices form quickly, followed by the slower assembly of these helices into the final structure (Zarrinkar & Williamson, 1994). Futher developments of RNA folding methods will have to evaluate and/or incorporate the consequences of these observed principles (compare e.g., Martinez, 1984; Abrahams et al., 1990). Second, in line with the strong evidence obtained for protein folding process, there is growing awareness that other molecules may facilitate the proper folding of RNA. These molecules include chaperones and small nucleolar RNAs (snoRNAs). Recent studies on the bacteriophage T4 group I intron (Coetzee et al., 1994) have shown that the ribosomal protein S12 facilitates splicing activity, suggesting its role as an RNA chaperone to prevent non-native structures from accumulating. The ribosomal protein S12, along with other ribosomal proteins, could also function as an rRNA chaperone, to facilitate in the rapid and accurate folding of the rRNAs during ribosome assembly (e.g., Nomura & Held, 1974). Within the past few years, studies are beginning to reveal that Eucaryotic snoRNAs, in association with proteins, are involved in ribosome biogenesis and transport (Filipowicz & Kiss, 1993; Fournier & Maxwell, 1993; Sollner-Webb, 1993). Yet another factor that may influence RNA folding is the presence of modified nucleotides in the RNA.

Although several factors are probably involved in rRNA folding, it is remarkable that some 16S rRNAs fold with such high prediction scores with only a thermodynamic consideration. Do these rRNAs truly fold in the absence of other molecules or factors (e.g., kinetics), or might other factors be involved such that there is redundancy or overdetermination? That some 16S rRNAs do not fold well could be due to a limited and possibly nonoptimal set of thermodynamic parameters, or these 16S rRNAs have acquired a larger protein dependency for proper folding or stabilization of their RNA structures (compare Hogeweg & Konings, 1985). Given our faith in the comparatively derived structures and the large number of phylogenetically diverse secondary structure models for 16S rRNAs, 23S rRNA (Gutell et al., 1993), group I introns (Damberger & Gutell, 1994), RNase P (Brown et al., 1994), etc., we are in a position to explore further some of the questions raised here. Ultimately, we seek to better understand the dynamics underlying the folding of RNA molecules.

MATERIALS AND METHODS

We establish the comparatively derived 16S and 16S-like rRNA structures as our gold standard. Our initial goal is to determine how well the thermodynamic-based folding algorithm (Zuker, 1989) is able to predict these 16S rRNA reference structures. 16S rRNA was chosen for this study for a number of reasons. First, we are very confident that these secondary structures are biologically correct due to the large number of compensatory base substitutions at the majority of all secondary structure base pairs (Gutell et al., 1994) and the large body of experimental data that is consistent with these structures (Hill et al., 1990; Nierhaus et al., 1993). Second, this collection of comparatively derived 16S rRNA structures is quite large and phylogenetically and structurally diverse (more than 3,500 complete or nearly complete 16S and 16S-like rRNA sequences are now available; Gutell, 1994; Maidak et al., 1994). Currently there are approximately 150 16S and 16S-like rRNA structure diagrams (Gutell, 1994) that span the three phylogenetic domains and the two Eucarya organelles. Third, the length of 16S rRNA (see Table 2) is particularly well suited for our studies on short-range (<100 nt) and long-range (>100 nt) secondary structure interactions. Table 2 shows that the Archaeal, (eu)Bacterial, and chloroplast sequences are all approximately of the same length, i.e., close to 1,500 nt. In contrast, the mitochondrial and Eucarya sequences have a large variation in their lengths, ranging from 697 to 1,962 nt and from 1,250 to 1,870 nt, respectively.

Nomenclature

We refer to five phylogenetic groups: the three phylogenetic domains (Archaea, (eu)Bacteria, and Eucarya) (Woese et al., 1990a) and the two Eucarya organelles (chloroplasts and mitochondria). The term "16S" rRNA is used to refer to both 16S and 16S-like rRNAs.

Comparative structures

A representative set of 16S rRNA sequences and their comparative structures for each of the five phylogenetic groups were selected from those presented in the recent compendium (Gutell, 1994). The selection used in the present study includes 8 Archaeal, 15 (eu)Bacterial, 11 chloroplast, 7 mitochondrial, and 15 Eucarya sequences (Table 1). We did not include sequences that have unidentified nucleotides due to difficulties in the analysis of their structures. The secondary structure diagrams are generated with the interactive RNA-specific graphics program XRNA (B. Weiser & H. Noller, University of California, Santa Cruz, unpubl.) on a SUN Sparc 2 workstation.

Thermodynamic foldings

Optimal and suboptimal foldings are calculated for the full-length 16S rRNA sequences through minimization of free energy by using the algorithm described by Zuker and coworkers (Jaeger et al., 1989, 1990; Zuker, 1989). We choose to fold the full-length sequences because our primary interest is not to report the best possible prediction, but to consider the case where we have no prior knowledge of the biological RNA secondary structure and therewith the borders of existing domains. For each sequence, 20 suboptimal foldings are generated with a certain structural difference among each other (window size of 10). The set of free-energy values given by Jaeger et al. (1989) was used in the analysis. These values include an extra stabilizing energy of −2.0 kcal for a selected set of tetraloops (GAAA, GCAA, GGAA, GAGA, GUGA, GCGA, UUCG, UACG). These tetraloops have been studied experimentally and shown to be more stable than the average tetraloop (e.g., Groebe & Uhlenbeck, 1988; Tuerk et al., 1988; Cheong et al., 1990; Antao et al., 1991; Heus & Pardi, 1991; Antao & Tinoco, 1992; SantaLucia et al., 1992). It should be noted that the tetraloop GUAA (GNRA) also occurs frequently in prokaryotes and chloroplasts (unpubl. data), although it was not incorporated as extra-stable tetraloop in the MFOLD program until very recently (Walter et al., 1994).

The foldings are also calculated (1) without including the extra free energies for the initial eight unusually stable tetraloops (see above), and (2) for the three separate domains for a small sampling of 16S rRNAs.

The comparison of thermodynamic foldings and comparative structures

The 56 comparatively inferred 16S rRNA structures (Table 1) are compared with their calculated thermodynamic structures. The following considerations are part of these comparisons. (1) Noncanonical base pairs (other than GU base pairs) in the comparatively derived structures are excluded from the comparison and treated as single-stranded because these pairings are not predictable by the thermodynamic folding algorithm. Pseudoknots and other tertiary interactions in the comparative structures are also excluded from the comparison (and treated as unpaired) for the same reason. (2) Small regions of some mitochondrial and Eucarya sequences have not been modeled in their entirety with comparative analysis due to the absence of strong comparative evidence for a secondary structure. The unstructured regions of these RNAs are not included in the analysis.

For the analysis performed here, we investigate how well the comparatively inferred base pairs are predicted by thermodynamic folding. The accuracy of this prediction is expressed by the percentage of the comparative structure that is calculated correctly. To discern the strengths and weaknesses of these thermodynamic foldings, three primary measures of folding success are utilized here.

The prediction of base pairs

These base pairs are initially evaluated with no context bias; a score is determined for the overall prediction success (Fig. 1). Base pairs are also classified according to the sequence distance between the two paired positions in question (Fig. 5), and to their proximity with four loops—hairpin, internal, bulge, and multistem (Fig. 6).

The prediction of helices

The prediction of a helix is calculated according to Zuker et al. (1991). A helix is defined by a double-stranded region of at least three base pairs that is possibly interrupted by internal loop or bulge of at most two bases. A helix is considered predicted when the number of correctly predicted base pairs is not more than two less the total number of base pairs in the comparative helix.

The prediction of hairpin loops

A hairpin loop is predicted correctly when the closing base pairs are identical. The properties addressed in these comparisons are averaged over each of the five phylogenetic and structural groups.

Identification of tetraloop sequences

For the identification of the tetraloop sequences at *E. coli* positions 297, 343, and 1516, the following alignments were used: (1) Archaea, 77 sequences; (eu)Bacteria, 1,560 sequences; chloroplasts, 30 sequences (Ribosomal Database Project, Maidak et al., 1994); (2) mitochondria, 117 sequences; Eucarya, 506 sequences (R.R. Gutell, unpubl. data).

ACKNOWLEDGMENTS

This work was supported by grants from the Colorado RNA Center and the NIH (GM48207) to R.R.G. We thank the W.H. Keck Foundation for support of RNA science on the Boulder Campus.

Received May 8, 1995; returned for revision June 27, 1995; revised manuscript received July 19, 1995

REFERENCES

Abrahams JP, van den Berg M, van Batenburg E, Pleij C. 1990. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Res 18*:3035–3044.

Antao VP, Lai SY, Tinoco I Jr. 1991. A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucleic Acids Res 19*:5901–5905.

Antao VP, Tinoco I Jr. 1992. Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucleic Acids Res 20*: 819–824.

Brown JW, Haas ES, Gilbert DG, Pace NR. 1994. The ribonuclease P database. *Nucleic Acids Res 22*:3660–3662.

Cheong C, Varrani G, Tinoco I Jr. 1990. Solution structure of an unusually stable RNA hairpin, 5′ GGAC(UUCG)GUCC. *Nature 346*: 680–682.

Coetzee T, Herschlag D, Belfort M. 1994. *Escherichia coli* proteins, including ribosomal protein S12, facilitate in vitro splicing of phage T4 introns by acting as RNA chaperones. *Genes & Dev 8*:1575–1588.

Damberger SH, Gutell RR. 1994. A comparative database of group I intron structures. *Nucleic Acids Res 22*:3508–3510.

Emerick VL, Woodson SA. 1994. Fingerprinting the folding of a group I precursor RNA. *Proc Natl Acad Sci USA 91*:9675–9679.

Filipowicz W, Kiss T. 1993. Structure and function of nucleolar snRNPs. *Mol Biol Reports 18*:149–156.

Fournier MJ, Maxwell ES. 1993. The nucleolar snRNAs: Catching up with the spliceosomal snRNAs. *Trends Biochem Sci 18*:131–135.

Groebe DR, Uhlenbeck OC. 1988. Characterization of RNA hairpin loop stability. *Nucleic Acids Res 16*:1172–1173.

Gutell RR. 1994. Collection of small subunit (16S- and 16S-like) ribosomal RNA structures: 1994. *Nucleic Acids Res 22*:3502–3507.

Gutell RR, Gray MW, Schnare MN. 1993. A compilation of large subunit (23S and 23S-like) ribosomal RNA structures: 1993. *Nucleic Acids Res 21*:3055–3074.

Gutell RR, Larsen N, Woese CR. 1994. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiol Rev 58*:10–26.

Gutell RR, Power S, Hertz G, Putz E, Stormo G. 1992. Constraints on the higher-order structure of RNA: Continued development and application of comparative sequence analysis methods. *Nucleic Acids Res 20*:5785–5795.

Heus HA, Pardi A. 1991. Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science 23*:191–194.

Hill WE, Dahlberg A, Garrett RA, Moore PB, Schlessinger D, Warner JR, eds. 1990. *The ribosome: Structure, function & evolution*. Washington D.C.: American Society for Microbiology.

Hogeweg P, Konings DAM. 1985. U1 snRNAs: The evolution of its primary and secondary structure. *J Mol Evol 21*:323–333.

Hughes JMX, Konings DAM, Cesareni G. 1987. The yeast homologue of U3 snRNA. *EMBO J 6*:2145–2155.

Jaeger JA, Turner DH, Zuker M. 1989. Improved predictions of secondary structures for RNA. *Proc Natl Acad Sci USA 86*:7706–7710.

Jaeger JA, Turner DH, Zuker M. 1990. Predicting optimal and suboptimal secondary structure for RNA. In: Doolittle RF, ed. *Molecular evolution: Computer analysis of protein and nucleic acid sequences. Methods Enzymol 183*:281–306.

Konings DAM, Hogeweg P. 1989. Pattern analysis of RNA secondary structure: Similarity and consensus of minimal-energy folding. *J Mol Biol 17*:4205–4216.

Maidak BL, Larsen N, McCaughey MJ, Overbeek R, Olsen GJ, Fogel K, Blandy J, Woese CR. 1994. The ribosomal database project. *Nucleic Acids Res 22*:3485–3487.

Martinez HM. 1984. An RNA folding rule. *Nucleic Acids Res 12*:323–334.

Nierhaus KH, Franceschi F, Subramanian AR, Erdmann VA, Wittmann-Liebold B, eds. 1993. *The translational apparatus: Structure, function, regulation, evolution*. New York: Plenum Press.

Nomura M, Held WA. 1974. Reconstitution of ribosomes: Studies of ribosome structure, function and assembly. In: Nomura M, Tissieres A, Lengyel P, eds. *Ribosomes*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press. pp 193–223.

SantaLucia J, Kierzek R, Turner DH. 1992. Context dependence of hydrogen bond free energy revealed by substitutions in an RNA hairpin. *Science 256*:217–219.

Sollner-Webb B. 1993. Novel intron-encoded small nucleolar RNAs. *Cell 75*:403–405.

Tuerk C, Gauss P, Thermes C, Groebe DR, Gayle M, Guild N, Stormo G, D'Aubenton-Carafa Y, Uhlenbeck OC, Tinoco I Jr, Brody EN, Gold L. 1988. CUUCGG hairpins: Extraordinarily stable RNA secondary structures associated with various biochemical processes. *Proc Natl Acad Sci USA 85*:1364–1368.

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

Williams AL, Tinoco I Jr. 1986. A dynamic programming algorithm for finding alternative RNA secondary structures. *Nucleic Acids Res 14*:299–315.

Woese CR, Kandler O, Wheelis ML. 1990a. Towards a natural system of organisms: Proposal for the domains archaea, bacteria and eucarya. *Proc Natl Acad Sci USA 87*:476–479.

Woese CR, Winker S, Gutell RR. 1990b. Architecture of ribosomal RNA: Constraints on the sequence of tetraloops. *Proc Natl Acad Sci USA 87*:8467–8471.

Zarrinkar PP, Williamson JR. 1994. Kinetic intermediates in RNA folding. *Science 265*:918–924.

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

Zuker M, Jaeger JA, Turner DH. 1991. A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. *Nucleic Acids Res 19*:2707–2714.