Julie Zorn[1]
Hin Hark Gan[1,3]
Nahum Shiffeldrim[1,3]
Tamar Schlick[1–3]

[1] Department of Chemistry,
New York University,
1021 Silver,
100 Washington Square East
New York,
NY 10003

[2] Courant Institute of
Mathematical Sciences,
New York University,
251 Mercer Street,
New York,
NY 10012

[3] Howard Hughes
Medical Institute

Received 29 April 2003;
accepted 15 May 2003

Structural Motifs in Ribosomal RNAs: Implications for RNA Design and Genomics

Abstract: *The various motifs of RNA molecules are closely related to their structural and functional properties. To better understand the nature and distributions of such structural motifs (i.e., paired and unpaired bases in stems, junctions, hairpin loops, bulges, and internal loops) and uncover characteristic features, we analyze the large 16S and 23S ribosomal RNAs of* Escherichia coli. *We find that the paired and unpaired bases in structural motifs have characteristic distribution shapes and ranges; for example, the frequency distribution of paired bases in stems declines linearly with the number of bases, whereas that for unpaired bases in junctions has a pronounced peak. Significantly, our survey reveals that the ratio of total (over the entire molecule) unpaired to paired bases (0.75) and the fraction of bases in stems (0.6), junctions (0.16), hairpin loops (0.12), and bulges/internal loops (0.12) are shared by 16S and 23S ribosomal RNAs, suggesting that natural RNAs may maintain certain proportions of bases in various motifs to ensure structural integrity. These findings may help in the design of novel RNAs and in the search (via constraints) for RNA-coding motifs in genomes, problems of intense current focus.* © 2003 Wiley Periodicals, Inc. Biopolymers 73: 340–347, 2004

Keywords: *ribosomal RNA; RNA structural motif; paired and unpaired bases; RNA design; RNA motif search*

Correspondence to: Tamar Schlick; email: schlick@nyu.edu

Contract grant sponsor: Joint NSF/NIGMS Initiative to Support Research Grants in the Area of Mathematical Biology
Contract grant number: DMS-0201160

Biopolymers, Vol. 73, 340–347 (2004)

INTRODUCTION

As well appreciated for proteins, the sequences of common structural motifs in natural RNAs, too, are nonrandom. Large ribosomal RNAs (rRNAs), for example, are known to have recurrent sequence motifs. Common motifs of 16S and 23S rRNAs include the UNCG and GNRA tetraloop motifs,[1,2] and others like bulge-G, bulge–helix–bulge, U-turn, biloop and triloop, and A-stack have also been identified.[1–5] These patterns are important because they participate in secondary and tertiary (e.g., loop–loop and loop–bulge) interactions that help stabilize compact structures of functional RNAs. Thus, analysis of various aspects of RNA motifs can help us understand the structural features that distinguish natural RNAs from non-RNA-like sequences. This understanding can be applied to the design of RNA-like structures and to searches for RNA-coding motifs in genomes.

Indeed, in the former application, the modular design of novel RNAs using existing RNA fragments implicitly utilizes such information to improve the functional properties of natural RNAs.[6,7] Such information may also be used to improve the in vitro selection technique, an experimental method for selecting functional RNAs from a large pool (10^{15}) of random sequences.[8,9] Instead of random sequences, using designed sequences with RNA-like motifs for these experiments may increase the probability of identifying novel functional RNAs.

The second challenging application—searching for RNA genes in genomes—often utilizes motif-search algorithms that require input about common motifs and motif lengths to narrow the search.[10–13] RNA genes are genomic sequences that lead to functional RNA molecules instead of proteins as end products. The recent RNAMotif scanning algorithm employed to identify known and novel RNA genes can greatly benefit from a greater understanding of RNA motifs.[12,13]

Here, we present new analyses on aspects of RNA motifs responsible for the stabilization of RNA secondary structure. Specifically, we examine the distributions of the number of paired bases in stems and unpaired bases in bulges/internal loops, hairpin loops, and junctions of large ribosomal RNAs in an effort to determine the character and range of structural motifs of functional RNAs (see Figure 1). From these distributions, we derive the ratio of the total unpaired to paired bases and fractions of the paired bases in stems and unpaired bases in bulges/internal loops, hairpin loops, and junctions. These distributions and ratios are not expected to be random because Nature has selected RNA molecules with motifs that are thermodynamically stable. We choose the large ribosomal RNAs for our analysis because they can yield statistically meaningful results, unlike small RNAs. We use *Escherichia coli* ribosomal RNAs in this study, but we have found results to be similar for other species, such as yeast, because of a high degree of structural conservation in ribosomes.

We find that 16S and 23S rRNAs have a similar range of paired or unpaired bases in stems, bulges, loops, and junctions, although the length of 23S is twice that of 16S. Moreover, for each motif, the frequency distribution of bases displays a different characteristic pattern. For example, the distributions for paired bases in stems and unpaired bases in hairpin loops decline linearly and exponentially, respectively, with the number of bases. In contrast, the frequency distribution of the number of unpaired bases in junctions shows a pronounced peak (at 6–10 bases); for bulges/internal loops, the distribution of unpaired bases exhibits a sharp drop beyond a certain number of bases (around 11). More significantly, our analysis reveals that the ratio of total unpaired to paired bases, and the allocation of paired bases in stems and unpaired bases in junctions, bulges/internal loops, and hairpin loops are roughly the same for 16S and 23S ribosomal RNAs. These constants suggest that natural RNAs may maintain these proportions or, equivalently, structural parameters, which likely reflect the global requirements for structural stability. Thus, delineated characteristic distributions, ranges, and parameters for RNA motifs may be used to aid RNA design efforts and the search for novel RNA motifs in genomes.

This article is organized as follows. The Methods section defines the RNA structural motifs and describes a program for analyzing motifs. The Results section presents the distributions, ranges, and parameters characterizing different RNA motifs. We conclude with a summary of the usage of such information in biological research on RNA structure and function.

METHODS

To survey structural motifs, we rigorously define RNA stem, bulge, hairpin and internal loops, and junction using the same definitions detailed recently in our approach for representing RNA motifs as graphs.[14] (Our RNAs-As-Graphs, or RAG, database is available on our group's website http://monod.biomath.nyu.edu.) Specifically, we define an RNA stem to consist of two or more complementary base pairs (GC and AU), with the GU wobble base pair considered a complementary base pair as well.[15] A nucleotide

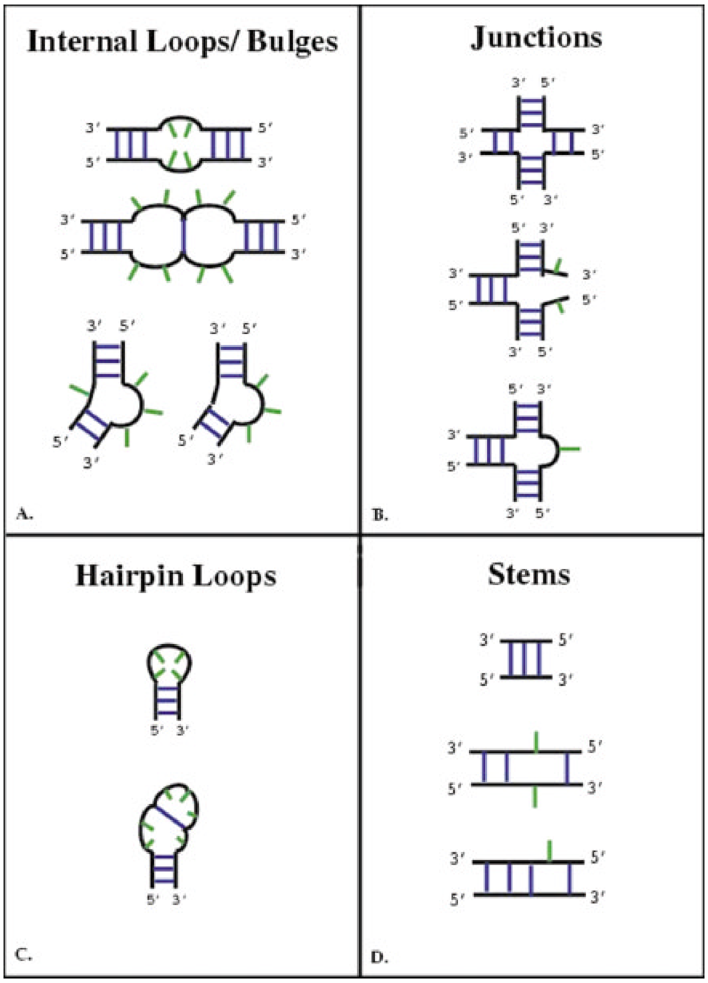

FIGURE 1 RNA structural motifs.

bulge, hairpin loop, or internal loop must have more than one unmatched nucleotide or noncomplementary base pair. A junction is a meeting point of three or more stems; the number of unpaired bases at a junction can be zero or more bases. Figure 1 shows various internal loops, bulges, junctions, hairpin loops, and stems statisfying these definitions.

Here, we analyze the *E. coli* 5S, 16S, and 23S ribosomal RNA sequences containing 120, 1542, and 2904 nt, respectively; the 5S is from *E. coli* strain *Pseudomonas putida* and the 16S and 23S are from *Actinobacillus actinomycetemcomitans.* The secondary structures of these RNAs, obtained from Gutell's website (http://www.rna.icmb.utexas.edu), are shown in Figure 2; they were reliably inferred by comparative RNA analysis using related experimental two and three-dimensional structures.[15,16] RNA structural motifs (i.e., stems, bulges/internal loops, junctions, and hairpin loops) can be evaluated using a "ct file" specifying which bases are paired or unpaired; the file is produced by Gutell's comparative analysis.[16] By using the ct file, we can determine the number of nucleotides associated with each of the RNA structural motifs using our analysis program that scans for existence of various structural motif types. The program determines for each segment of paired/unpaired bases whether it belongs to a stem, a bulge/internal loop, a hairpin loop, or a junction. Since our analysis is confined to secondary structures, we removed a few base pairs associated with tertiary structures in the ct files.

RESULTS

Characteristic Distributions of Paired/ Unpaired Bases in Structural Motifs

Figure 3 shows the characteristic frequency distributions of the number of paired bases in stems and unpaired bases in bulges/internal loops, hairpin loops, and junctions of the 16S and 23S rRNAs; the results for 5S are not shown because this small RNA has only a few structural motifs. For the 23S rRNA, the distribution of the occurrence frequency for paired bases in stems declines roughly linearly with the number of bases. The slope η of the linear

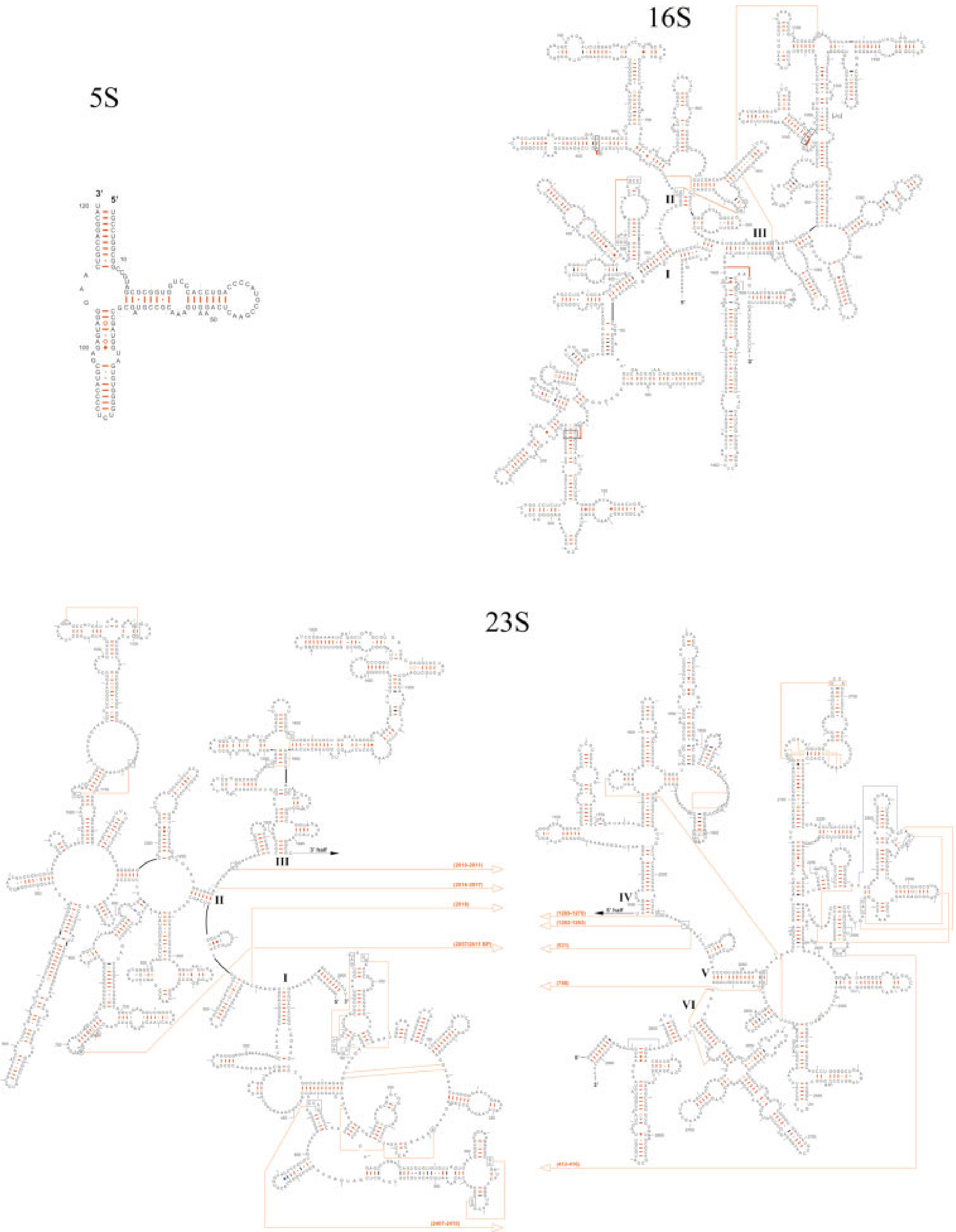

FIGURE 2 Secondary structures of the *E. coli* 5S, 16S, and 23S ribosomal RNAs.

fit is −1.18. Most stems in 23S rRNA have between 4 and 14 bases, or 2 and 7 base pairs. In contrast, the frequency distribution for unpaired bases in bulges/internal loops may be modeled, to a crude approximation, as a step-like function with an sharp decline at around 11 bases. The distribution for

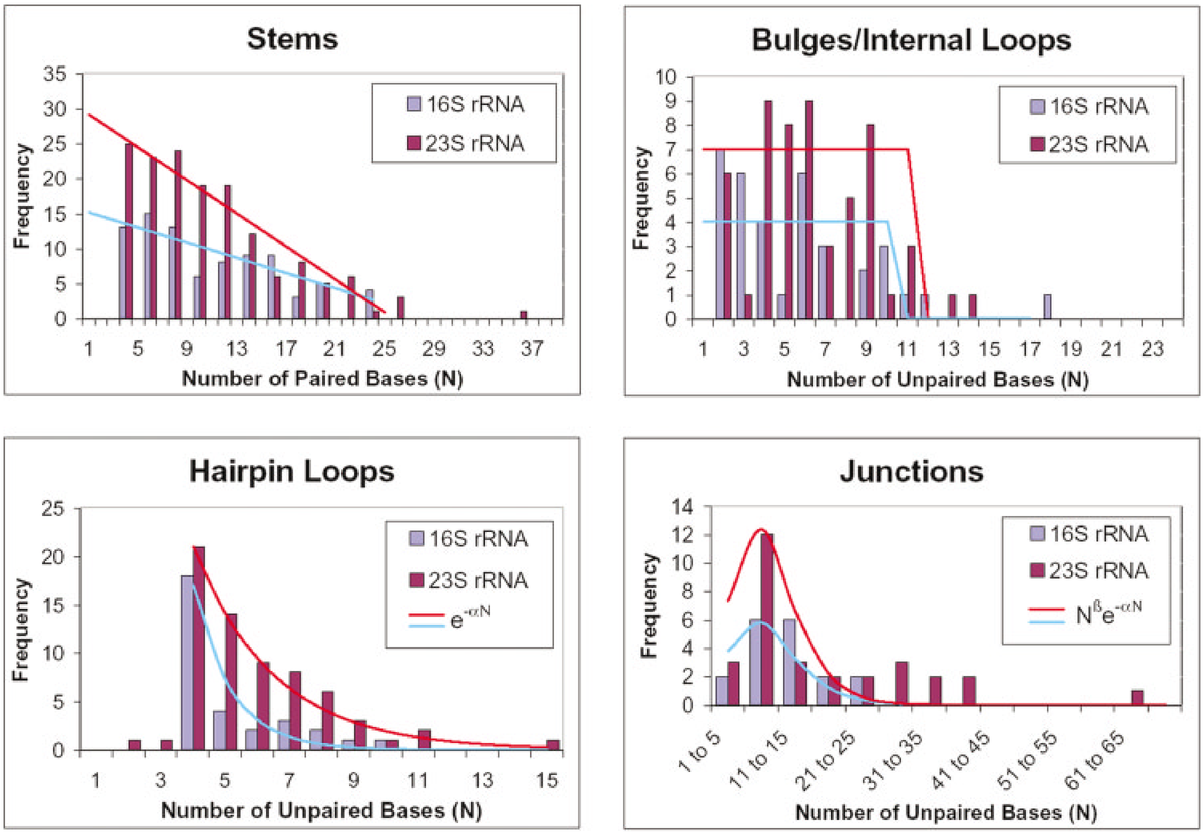

FIGURE 3 Histograms of the occurrence frequency of the number of paired and unpaired bases in the structural motifs of *E. coli* 16S and 23S ribosomal RNAs. Results for stems, bulges/internal loops, hairpin loops, and junctions are fitted using linear, step-like, exponential, and exponential-like functions, respectively.

unpaired bases in hairpin loops can be modeled as an exponentially decaying function, $\exp(-\alpha N)$, where α $(= 0.4)$ is a constant and N is the number of unpaired bases; it has about half the range, i.e., 4–10 bases, relative to that of the stems' distribution. For junctions, the frequency distribution of unpaired bases shows a prominent, narrow peak at 6–10 bases, indicating a strong bias in base number for this structural motif. (We have binned the distribution plot for junctions since there are fewer junctions than stems, loops, or bulges.) This distribution can be reasonably fitted to the function $N^{\beta} \exp(-\gamma N)$ where the constants $\beta = 4.0$ and $\gamma = 0.45$.

These trends for 23S rRNA are generally the same for the 16S rRNA, with minor differences in the distributions for stems and hairpin loops. We thus use the same fitting functions in Figure 3 for 16S results, but with different constants: $\alpha= 0.85$, $\beta= 3.5$, $\gamma = 0.4$, and $\eta = -0.54$. The stem size distribution for 16S has a secondary peak at 14 bases and most of 16S's hairpins are tetraloops whose common sequence motifs include the GNRA pattern.[1,2] We find that the ranges of the distributions of paired/unpaired bases (shown in Figure 3) for various motifs in both 16S and 23S rRNAs are nearly the same (see Table I), even though the latter is twice the size of the former. In particular, we find that the number of unpaired bases in junctions, bulges/internal loops and hairpin loops has about the same range (2–10), whereas the number of paired bases in stems has a broader range (4–24).

The above general conclusions about structural motifs of ribosomal RNAs of *E. coli* strains are also valid for other species like the yeast rRNAs (check the group's website http://monod.biomath.nyu.edu).

Invariant Features of Structural Motifs

Other motif features of interest are the proportions of paired and unpaired bases in structural motifs. Are these proportions variable depending on RNA type or are they fixed? We enumerate the number of bases in bulges (B), hairpin loops (H), junctions (J), entire RNA (L), helical mispairs (M), and stems (S), and summarize in Table II the fraction of paired bases in

Table I Ranges of the Number of Paired Bases in Stems and Unpaired Bases in Bulges/Internal Loops, Hairpin Loops, and Junctions of the Ribosomal RNAs of *E. coli*[a]

RNA	Stems	Bulges/Internal Loops	Hairpin Loops	Junctions
16S	4–24	2–10	4–7	6–15
23S	4–26	2–11	4–9	6–30

[a] These ranges are derived from the distributions of the number of paired and unpaired bases in RNA structural motifs in Figure 3 where the occurrence frequence is at least three. The distributions have the following shapes: linearly decreasing function for paired bases in stems; step-like function for unpaired bases in bulges/internal loops; exponentially decaying function for unpaired bases in hairpin loops; and exponential-like function for unpaired bases in junctions.

stems (S/L), fractions of unpaired bases in junctions (J/L) and in bulges/internal loops (B/L), and ratio of total unpaired to paired bases in the entire RNA. The results for the small 5S rRNA are shown for comparison. Despite their size difference, the 16S and 23S rRNAs have remarkably similar proportions of paired/unpaired bases in various structural motifs.

In particular, the ratio of unpaired to paired bases is 0.70 for 16S rRNA, and 0.79 for 23S rRNA, meaning the relative amount of unpaired bases is at most ~80% of paired bases for these RNAs. For the 5S rRNA, the relative amount of unpaired bases is even lower at 50% of paired bases. These results also indicate that the ratio depends on RNA size: smaller RNAs have a smaller proportion of unpaired bases. The ratio may saturate for larger RNAs.

How are the paired and unpaired bases distributed among stem, junction, and bulge motifs? As noted above, the ratio of unpaired to paired bases, (B+J+H+M)/S, is less than 0.8 for all rRNAs compared. As shown in Table II, for the large rRNAs, S/J ~ 0.58, J/L ~ 0.16, and B/L ~ H/L ~ 0.10 to 0.14; the sum S/L+J/L+B/L+H/L is more than 0.97, with the remaining small fraction of bases found in helical mispairs and 3′ and 5′ ends. These values again show that RNAs are stabilized by a large proportion of paired bases in stems. The slightly larger proportion of unpaired bases in junctions than in either bulges, internal loops or hairpin loops may suggest the need for greater flexibility in junctions where several stems meet; the J/L ratio also increases with RNA size. We emphasize that the B/L ratio is remarkably similar for all rRNAs. Overall, we find that about 60% of paired bases occur in stems, and 10–18% of unpaired bases occur in each of the following motifs, including junctions, bulges/internal loops, and hairpin loops, and less than 3% of the unpaired bases occur in helical mispairs and 3′ and 5′ ends.

RNA Structural Motifs and Euler Formula

Another approach for analyzing RNA secondary structures is to use graphical representations. In our recent work,[14] we introduce RNA graphical representations for quantitative analysis and enumeration of all possible RNA structures with the aim of estimating RNA's structural repertoire. One of our findings is that the numbers of different structural motifs in any given RNA are strictly related. Let us denote by N_{J}, N_{BH}, and N_{S} the number of junctions, bulges/internal loops/hairpin loops, and stems, respectively, in a given RNA structure. From the Euler formula in graph theory,[17] we have derived the following relation for any RNA[14]:

$$N_{\mathrm{J}} + N_{\mathrm{BH}} = N_{\mathrm{So}}. \qquad (1)$$

Table III summarizes the ratios $N_{\mathrm{J}}/N_{\mathrm{S}}$, $N_{\mathrm{BH}}/N_{\mathrm{S}}$, and $(N_{\mathrm{J}} + N_{\mathrm{BH}})/N_{\mathrm{S}}$ for all three rRNAs. We find that $N_{\mathrm{J}}/N_{\mathrm{S}} \sim 0.2$ and $N_{\mathrm{BH}}/N_{\mathrm{S}} \sim 0.8$, making $(N_{\mathrm{J}}+N_{\mathrm{BH}})/N_{\mathrm{S}} = 1$. Thus, the proportions of junctions and bulges/loops to stems are nearly constant for RNAs with very different sizes, although many other values are mathematically possible. In particular, the ratio $N_{\mathrm{J}}/N_{\mathrm{S}}$ is a measure of the degree of branching in RNA struc-

Table II Fractions of the Number of Unpaired to Paired Bases, Paired Bases in Stems (S/L), Unpaired Bases in Junctions (J/L), Unpaired Bases in Bulges/Internal Loops (B/L), and Unpaired Bases in Hairpin Loops (H/L) Found in *E. coli* Ribosomal RNAs[a]

RNA	(B + J + H + M[a])/S	S/L	J/L	B/L	H/L
5S	0.500	0.667	0.075	0.108	0.142
16S	0.702	0.598	0.137	0.129	0.104
23S	0.786	0.560	0.178	0.120	0.135

[a] M is the number of unpaired bases in helical mispairs and L the RNA sequence length.

Table III Number of RNA Structural Motifs (Number of Junctions (N_J) and Number of Bulges/Internal Loops/Hairpin Loops (N_{BH}), Relative to the Number of Stems

RNA	N_J/N_S	$N_{B\text{-}H}/N_S$	$(N_J + N_{BH})/N_S$
5S	0.167	0.833	1.00
16S	0.217	0.795	1.01
23S	0.199	0.808	1.01

tures. The observed N_J/N_S value suggests that rRNAs are moderately branched, as we have shown earlier.[14] Our results also imply that $N_J/N_S < N_{BH}/N_S$. Thus, examining RNAs from the perspective of relation (1) again suggests the existence of nonrandom relationships among RNA structural motifs.

Comparison with tRNA

It is instructive to compare the above results with those for the tRNA. We choose a 76-nt yeast tRNA-Phe whose crystal structure is known (see Nucleic Acids Database at http://ndbserver.rutgers.edu). The fractions of paired bases in stems and unpaired bases hairpin loops, and junctions are 0.55, 0.29, and 0.11, respectively; the tRNA has no bulges and internal loops. Thus, the proportions of paired bases in stems and unpaired bases in junctions are similar to those in rRNAs, but a larger fraction of bases in hairpin loop is found in tRNA. As for the ratios of structural motif numbers, we obtain $N_J/N_S = 0.25$ and $N_{BH}/N_S = 0.75$, which are comparable to those for rRNAs (Table III).

CONCLUSION

Our survey of ribosomal RNA motifs (and examination of tRNA for comparison), indicates that common features hold for large functional RNAs (Figure 3, Tables I–III). To establish the proportions and ranges of paired and unpaired base more firmly, future analysis should consider all existing RNA secondary structures. Since the fraction of paired bases (with respect to the total number) in an RNA ($\sim$60%) is related the overall energetics or stability of the secondary structure, most RNAs are likely to have similar values. It is also likely that the other related ratios in Table II may be generally valid for RNAs.

RNA's modularity and conformational flexibility have been exploited to synthetically design novel functional RNAs for biotechnology applications using in vitro selection experiments[8,9] and rational modular design where new RNAs are engineered by assembling existing RNA fragments.[6,7] Such design efforts can be made more productive by selecting RNAs with structural motifs resembling those for natural RNAs. For example, in vitro selection uses a large (random) sequence pool (of order 10^{15}) of which only a tiny fraction of sequences yields the selected functional properties. This is likely due to the poor folding characteristics of random sequences. To improve the selection, we suggest instead designing a smaller pool of RNA sequences possessing the structural motifs in proportions we found in Tables I–III. RNA-like motifs are likely to have a much greater probability of yielding functional RNAs than random sequences. By the same argument, the "parameters" we derived from rRNAs may be applied to other RNA design applications.

Another intriguing application of this work is in the search for novel RNA motifs and genes in genomes. Programs such as RNAMotif allow general search for RNA motifs, but the motif "descriptor" specifying stem, bulge, and loop sizes must have well-defined constraints to make the search effective.[12,18] Current uncertainty in defining such constraints limits the value of RNA motif search programs in terms of finding known functional RNA families in genomes. Since the RNA parameters in Tables I–III may approximate functional RNAs, they can be used to constrain the search for novel RNAs using motif search programs.

In sum, the information derived here on the global features of RNA structural motifs complements the accumulated knowledge about many sequence patterns and motifs found in RNAs.[1–5] Taken together, the information about RNA motifs can be exploited to advance the rapidly progressing fields of RNA design and genomics.

This work was supported by a Joint NSF/NIGMS Initiative to Support Research Grants in the Area of Mathematical Biology (DMS-0201160). TS is an investigator of the Howard Hughes Medical Institute.

REFERENCES

1. Burkard, M. E.; Turner, D. H.; Tinoco, I., Jr. In The RNA World; Gesteland, R. F., Cech, T. R., Atkins, J. F., Eds.; Cold Spring Harbor Laboratory Press: New York, 1999; Chap 10.
2. Moore, P. B. Annu Rev Biochem 1999, 68, 287–300.
3. Leontis, N. B.; Stombaugh, J.; Westhof, E. Biochemie 2002, 84, 961–973.
4. Gutell, R. R.; Cannone, J. J.; Konings, D.; Gautheret, D. J Mol Biol 2000, 300, 791–803.

5. Lee, J. C.; Cannone, J. J.; Gutell, R. R. J Mol Biol 2003, 325, 65–83.
6. Soukup, G. A.; Breaker, R. R. Proc Natl Acad Sci USA 1999, 96, 3584–3589.
7. Soukup, G. A.; Breaker, R. R. Curr Opin Struct Biol 2000, 10, 318–325.
8. Tuerk, C.; Gold, L. Science 1990, 249, 505–510.
9. Ellington, A. D.; Szostak, J. W. Nature 1990, 346, 818–822.
10. Lowe, T. M.; Eddy, S. E. Nucleic Acids Res 1997, 25, 955–964.
11. Rivas, E.; Klein, R. J.; Jones, T. A.; Eddy, S. E. Curr Biol 2001, 11, 1369–1373.
12. Macke, T. J.; David, J.; Ecker, D. J.; Gutell, R. R.; Gautheret, D., Case, D. A.; Sampath, R. Nucleic Acids Res 2001, 29, 4724–4735.
13. Dandekar, T.; Hentze, M. W. TIG 1995, 11, 45–50.
14. Gan, H. H.; Pasquali, S.; Schlick, T. Nucleic Acids Res 2003, 31, 2926–2943.
15. Cannone, J. J.; Subramanian, S.; Schnare, M. N.; Collett, J. R.; D'Souza, L. M.; Du, Y.; Feng, B.; Lin, N.; Madabusi, L.V.; Muller, K. M.; Pande, N.; Shang, Z.; Yu, N.; Gutell, R. R. BMC Bioinformatics 2002, 3, 2.
16. Gutell, R.; Lee, J. C.; Cannone, J. J. Curr Opin Struct Biol 2002, 12, 301–310.
17. Gross, J.; Yellen, J. Graph Theory and Applications; Addison-Wesley: Reading, MA, 1999.
18. Fogel, G. B.; Porto, V. W.; Weekes, D. G., Fogel, D. B.; Griffey, R. H.; McNeil, J. A., Lesnik, E.; Ecker, D. J.; Sampath R. Nucleic Acids Res 2002, 30, 5310–5317.

Reviewing Editor: Dr. David A. Case