1384–1398 *Nucleic Acids Research, 2005, Vol. 33, No. 4*
doi:10.1093/nar/gki267

Modular RNA architecture revealed by computational analysis of existing pseudoknots and ribosomal RNAs

Samuela Pasquali, Hin Hark Gan[1] and Tamar Schlick[1,2,*]

Department of Physics, [1]Department of Chemistry and [2]Courant Institute of Mathematical Sciences,
New York University, 251 Mercer Street, New York, NY 10021, USA

Received December 1, 2004; Revised and Accepted February 5, 2005

ABSTRACT

Modular architecture is a hallmark of RNA structures, implying structural, and possibly functional, similarity among existing RNAs. To systematically delineate the existence of smaller topologies within larger structures, we develop and apply an efficient RNA secondary structure comparison algorithm using a newly developed two-dimensional RNA graphical representation. Our survey of similarity among 14 pseudoknots and subtopologies within ribosomal RNAs (rRNAs) uncovers eight pairs of structurally related pseudoknots with non-random sequence matches and reveals modular units in rRNAs. Significantly, three structurally related pseudoknot pairs have functional similarities not previously known: one pair involves the 3′ end of brome mosaic virus genomic RNA (PKB134) and the alternative hammerhead ribozyme pseudoknot (PKB173), both of which are replicase templates for viral RNA replication; the second pair involves structural elements for translation initiation and ribosome recruitment found in the viral internal ribosome entry site (PKB223) and the V4 domain of 18S rRNA (PKB205); the third pair involves 18S rRNA (PKB205) and viral tRNA-like pseudoknot (PKB134), which probably recruits ribosomes via structural mimicry and base complementarity. Additionally, we quantify the modularity of 16S and 23S rRNAs by showing that RNA motifs can be constructed from at least 210 building blocks. Interestingly, we find that the 5S rRNA and two tree modules within 16S and 23S rRNAs have similar topologies and tertiary shapes. These modules can be applied to design novel RNA motifs via build-up-like procedures for constructing sequences and folds.

INTRODUCTION

RNA secondary and tertiary structures are composed of modular substructures that often fold independently and in a hierarchical manner (1). This modularity of RNA secondary structure has been exploited in the design of novel functional molecules (2–6). Modular architecture also implies similarity of substructural motifs among existing RNAs, suggesting possible functional relationships. An interesting example noted by Mitchell *et al*. (7) is the occurrence of the box H/ACA motif of snoRNA in telomerase RNA, indicating a functional relationship between these RNAs; human telomerase RNA (hTR) H/ACA domain is essential for hTR accumulation, hTR 3′ end processing and telomerase activity. Indeed, similar functions are observed when the H/ACA snoRNA motif is substituted for the hTR H/ACA domain in human telomerase. As the repertoire of RNA structures increases through the discovery of natural (8,9) and synthetic or designed (3,10,11) RNAs, and as our interest in RNA structure and functions intensifies (12–14), automatic computer approaches are needed to detect structural similarity of RNAs within RNAs, as commonly done for establishing relatedness in protein families (15,16).

Most RNA structure comparison algorithms are designed for secondary structures because not many three-dimensional (3D) RNA structures are available for analysis (17–20). An exception is the recent PRIMOS program for comparing and identifying novel motifs in tertiary structures (21). Comparing RNAs based on secondary motifs/submotifs can yield insights about their relationships and topological properties because secondary structures belonging to the same functional group are generally conserved.

Current secondary-structure comparison algorithms have focused exclusively on tree structures owing to their relative simplicity for quantitative analysis. Tree structures refer to mathematical constructs for RNA secondary structures without pseudoknots (2,19,20). Various types of graphical tree representations have been used to develop RNA structure comparison and clustering algorithms, including ordered, labeled and

[*] To whom correspondence should be addressed: Tel: +1 212 998 3116; Fax: +1 212 995 4152; Email: schlick@nyu.edu

Table 1. Significant LALIGN aligned pseudoknot pairs

Pair	Aligned pair	$S(G)$	Sequence identity (%)	Aligned length (nt)	Score	<R>	Functional similarity
1	PKB134/PKB135	1.17	56	129	86	1.54	Known
2	PKB205/PKB233	0.47	63	32	50	1.47	Unknown
3	PKB178/PKB135	1.17	59	29	54	1.42	Unknown
4	PKB217/PKB174	0.47	62	47	57	1.24	Known (Figure 4)
5	PKB134/PKB173	1.17	80	19	48	1.23	Novel (Figure 5)
6	PKB134/PKB205	1.17	76	17	49	1.23	Novel (Figure 5)
7	PKB173/PKB233	0.47	63	35	48	1.20	Unknown
8	PKB173/PKB191	1.17	70	20	46	1.15	Unknown
9	PKB178/PKB173	1.17	71	24	38	1.14	Unknown
10	PKB205/PKB223	0.47	72	14	34	1.06	Novel (Figure 5)

Local alignment gap penalties: −15/−1. $S(G)$ values for simple and double pseudoknots are 0.47 and 1.17, respectively. <R> is R-value averaged over alignments with five randomized sequences.

unlabeled trees (17–20,22–24). The use of tree representations has made possible clustering of related tree structures (18,19), detection of point mutations with specific structural effects (18), search of recurring subtrees to deduce consensus structural motifs (20) and analysis of secondary structure statistics of large ensembles of random RNA sequences (24).

RNA pseudoknots, which cannot be represented as tree graphs, are common in nature, represented by many catalytic RNAs and small subunit (SSU) rRNA and their dominance in the RNA repertoire universe may increase with RNA size (6); see PseudoBase for a catalogue of known pseudoknots (25). In fact, our recent theoretical analysis of the abundance of RNA types (trees, pseudoknots and bridge topologies) predicts that the collection of possible motifs is dominated by pseudoknots (6).

In this study, we develop a structure-based RNA comparison algorithm using our 2D RNA dual graphical representation for both trees and pseudoknots (2,26,27), coupled with a graph similarity (isomorphism) search method and sequence alignment analysis. Because of available work in the field related to RNA trees, we focus here on uncovering structural and functional similarities of pseudoknots. Our graphical analysis of structure similarity emphasizes global topological similarity, a well-known fact for existing functional RNA families, such as tRNA, rRNA, RNase P and snoRNA. This automated approach is advantageous for comparing related RNAs with a low-sequence similarity. A disadvantage of graphical analysis is that both detailed base pair and structural information are necessarily ignored. For example, minor sequence and secondary structure changes in a viral frameshifting pseudoknot that preserve the RNA topology but disrupt the RNA function would be considered as a match (28,29). Such a subtle structural–functional relationship may not be amenable to automated analysis. A similar problem also arises in structure alignment of protein structures caused by, for example, mutations in functional sites (15,30). To reduce the error rates of false functional identification, such cases for proteins and RNAs are best screened by manual analysis after the detection of structural similarity. We adopt this two-step strategy for the identification of structural and functional similarity in RNA.

Our systematic comparison of structural similarities within 14 existing RNA pseudoknots—including viral frameshifting, tRNA-like, internal ribosome entry site (IRES), ribosomal RNA (rRNA) and tmRNA—unravels eight non-trivial matches, including three novel findings supported by sequence and functional data. Functional relationship can be deduced subsequent to finding structural matches, by analyzing sequence alignment and pseudoknot functional properties. One such structural and functional similarity example is an alternative hammerhead ribozyme pseudoknot of satellite cereal yellow dwarf virus (PKB173) embedded in the pseudoknot topology occurring in the 3′ end of brome mosaic virus genomic RNA (PKB134) (see Pair 5 in Table 1); both RNAs are templates for replicase in viral RNA replication. Another novel similarity example is between the viral IRES (PKB223) and the eukaryotic 18S rRNA (PKB205) (see Pair 10 in Table 1); both participate in translation initiation. Yet another related pair (Pair 6) involves 18S rRNA (PKB205) and viral tRNA-like pseudoknot (PKB134), which recruits host ribosomes for viral replication. We find that tRNA-like molecules possess an 8–11 nt region complementary to 18S rRNA (PKB205), suggesting that tRNA-like molecules may recruit ribosomes via structural mimicry (of tRNA shape) (31) and base complementarity. To the best of our knowledge, the structural and functional relationships identified in these three pseudoknot pairs have not been described previously.

In addition, we analyze the distribution of small tree modules or submotifs within the *Saccharomyces cerevisiae* 16S and 23S rRNAs using mathematically enumerated (distinct) tree graphs and our submotif search algorithm. We find that most small motifs (i.e. tree graphs with <6 vertices, or ∼100 nt) exist within rRNAs, whereas the larger submotifs (i.e. graphs with >9 vertices, or ∼160 nt) occur much less frequently. Specifically, our analysis suggests that rRNAs are constructed from at least 210 small distinct tree modules.

In summary, these quantitative analyses of similarity among pseudoknots and modular subunits of rRNAs may be applied to increasing repertoire of RNA structures and exploited for the identification and the modular design of novel RNA structures.

MATERIALS AND METHODS

We describe our computational methods for analyzing and comparing RNA secondary structures in the following subsections. Analyzing secondary rather than tertiary structures, as currently performed in many other studies, is appropriate because the former is conserved for functional RNA classes (e.g. tRNA, 5S rRNA and RNase P). Of course, analysis of tertiary folds is a separate and important problem.

Graphical representation of RNA secondary structures

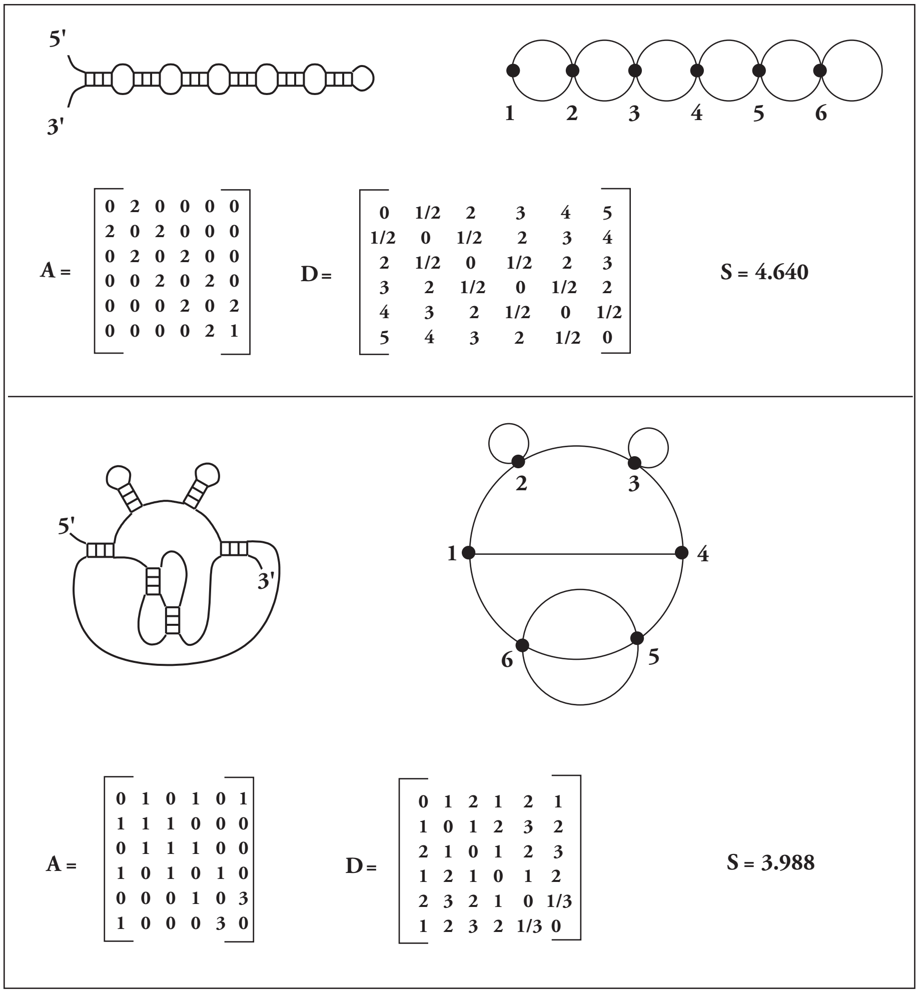
RNA secondary structures can be schematically represented as mathematical graphs to capture the essential topological connectivity of loops, bulges, junctions and stems (2,20). The use of simplified graphical representation allows RNA structures to be efficiently analyzed and compared. Recently, we developed a general class of RNA graphs called dual graphs capable of representing both RNA tree and pseudoknot structures (2) (compare dual and tree graphs in Figures 1 and 2, respectively). We use both tree and dual graphs in this work. The rules for mapping RNA secondary structures onto tree and dual graphs are given in (2) and our RNA-As-Graphs web resource (http://monod.biomath.nyu.edu/rna) (26,27).

Graph isomorphism search algorithm

The representation of RNAs as graphs provides a systematic framework to search for similar substructures in RNAs through the concept of graph isomorphism (32). Isomorphic graphs share the same pattern of connectivity among all vertices. The challenge is to identify, by an efficient computer algorithm, a graph as a subgraph of a larger graph or, in molecular terms, an existing RNA motif contained in a larger RNA. The computational complexity of identifying two structurally equivalent (i.e. isomorphic) graphs with V vertices is of the order of V factorial ($V!$) and known as the 'graph isomorphism problem' (32,33).

Our efficient method for testing graph isomorphism is based on graph topological numbers or invariants; a sketch of this idea was reported previously [(2), Appendix D]. We associate each graph or subgraph with one or more topological invariants, which are computed based on the patterns of connectivity among graph vertices. Thus, isomorphic graphs have the same topological invariants while dissimilar graphs have different topological invariants. The similarity or dissimilarity between graphs can be readily established by comparing their topological invariants (17). Our topological invariants allow the identification of equivalent graphs differing only in the placement of chain ends and/or hairpin loops, as explained below.

We define the topological invariants using a graph's connectivity as described by the adjacency matrix. Given a graph G with V vertices, the $V \times V$ adjacency matrix $A(G)$ of the graph specifies the connectivity between all pairs of vertices. For example, matrix element A_{ij} indicates the number of edges



$$A = \begin{bmatrix} 0 & 2 & 0 & 0 & 0 & 0 \\ 2 & 0 & 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 & 0 & 2 \\ 0 & 0 & 0 & 0 & 2 & 1 \end{bmatrix} \quad D = \begin{bmatrix} 0 & 1/2 & 2 & 3 & 4 & 5 \\ 1/2 & 0 & 1/2 & 2 & 3 & 4 \\ 2 & 1/2 & 0 & 1/2 & 2 & 3 \\ 3 & 2 & 1/2 & 0 & 1/2 & 2 \\ 4 & 3 & 2 & 1/2 & 0 & 1/2 \\ 5 & 4 & 3 & 2 & 1/2 & 0 \end{bmatrix} \quad S = 4.640$$

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 3 \\ 1 & 0 & 0 & 0 & 3 & 0 \end{bmatrix} \quad D = \begin{bmatrix} 0 & 1 & 2 & 1 & 2 & 1 \\ 1 & 0 & 1 & 2 & 3 & 2 \\ 2 & 1 & 0 & 1 & 2 & 3 \\ 1 & 2 & 1 & 0 & 1 & 2 \\ 2 & 3 & 2 & 1 & 0 & 1/3 \\ 1 & 2 & 3 & 2 & 1/3 & 0 \end{bmatrix} \quad S = 3.988$$

Figure 1. Two hypothetical RNA secondary structures and their dual graphical representations. The RNA graphs are quantitatively described using adjacency matrix A, distance matrix D (Equation 1) and topological invariant S (Equation 2). We use these measures to compare and search for similar substructures in existing RNAs.

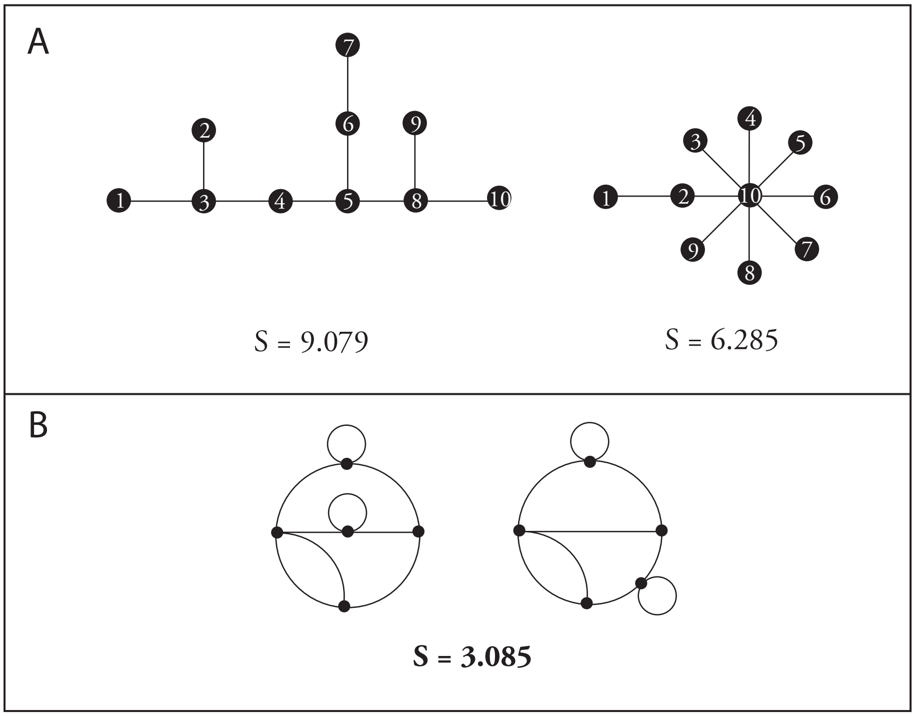


Figure 2. Tree and dual graphical representations and topological invariants of hypothetical RNA secondary structures. (**A**) RNA tree graphs and their topological invariants S^{tree}. (**B**) The only example we found (out of ∼100) of two distinct RNA-like pseudoknot graphs with the same topological invariant S; the pseudoknots only differ by the placement of a stem–loop. Although our structural comparison algorithm regards this structure pair as a positive match, in practice such cases are eliminated by manual screening.

connecting vertex i with vertex j. For RNA dual graphs, elements of the adjacency matrix have the following properties: A_{ij} is symmetric; the allowed number of edges (connections) between two vertices is 0, 1, 2 or 3; a vertex can have at most one self-loop (value of 1) representing a hairpin loop; and each vertex i is connected to four edges representing two incoming and two outgoing connecting strands, except where the chain ends occur.

Our graph invariant is defined based on a modified distance-like matrix $D(G)$ derived from the adjacency matrix $A(G)$. The elements of symmetric matrix (D_{ij}) are defined as follows:

$$\begin{aligned} D_{ij} &= A_{ij}^{-1} && \text{for } A_{ij} \neq 0 \\ D_{ij} &= d_{ij} && \text{for } A_{ij} = 0 \\ D_{ii} &= 0 && \text{for all } i, \end{aligned} \qquad \mathbf{1}$$

where d_{ij} is the minimum number of edges separating vertices i and j, i.e. the number of edges defining the shortest possible path between these two vertices. We define the graph invariant $S(G)$ as follows:

$$S(G)^2 = (V-1)^{-1} \sum_{i=1}^{V} \left(\sum_{j=1}^{V} D_{ij} \right)^2. \qquad \mathbf{2}$$

Approximately, $S(G)$ measures the average distance among the vertices in the graph G. For example, the component S_i, defined as $\sum_{j=1}^{V} D_{ij}$, measures how well-connected vertex i is to the other vertices in the graph; a peripheral vertex will produce a higher S_i value than a vertex located more in the center of the graph. Effectively, our scheme assigns adjacent vertices with smaller weights than non-adjacent vertices. Thus, a large S value corresponds to an extended structure with few branches and a small S value to a compact graph with high-order junctions or a complex pseudoknot fold. This property of S can also be used to rank and measure the topological distance between graphs. In general, such topological invariants are believed to have a better correspondence with the physicochemical properties of real molecules since many of the properties depend critically on the exposed surface and less critically on the buried elements (34). Figure 1 shows the S values for two distinct RNA topologies.

Our distance matrix $D(G)$ and topological invariant S can be extended to tree graphs, although their adjacency matrices have different properties from those for dual graphs. We call S^{tree} a tree topological invariant, which will be used for analyzing submotifs in 16S and 23S rRNAs. Figure 2A compares two topologically distinct 10-vertex tree graphs with different S^{tree} values.

Our topological invariants [$S(G)$ and S^{tree}] are degenerated with respect to the location of chain ends or hairpin loops, since these elements are not scored. Thus, topologically equivalent structures (i.e. same connectivity) embedded within a larger structure can be identified. All known topological invariants are imperfect since distinct structures can be incorrectly assigned as identical. That is, graphs with distinct topological invariants are non-isomorphic, but the converse is not true (not all graphs with same topological invariant are isomorphic). This is the well-known graph isomorphism problem (33). In our analysis of ∼100 connected graphs with our topological invariant S, we only encountered one pair of similar but non-isomorphic graphs (for $V = 5$) that resulted in the same invariant S (see Figure 2B). In a previous work, we have also used the Laplacian eigenvalue spectrum as a topological invariant yielding a comparable error rate of a few percent (26). Some false positive matches are reported because of this graph isomorphism problem. Since we manually examine all positive matches, the false positive matches are eliminated.

Searching for the motif of a 10-vertex graph in a 30-vertex graph requires ~20 min CPU time on an SGI 300 MHz MIPS R12000 IP27 processor. The search of a 15-vertex motif in a 40-vertex graph would imply 1000-fold increase in CPU time. In practice, our method can easily search (i.e. in minutes) for subgraphs of size ranging from 5 to 11 vertices within graphs of size up to ~30 vertices (corresponding to ~600 nt RNAs).

RESULTS AND DISCUSSION

RNA pseudoknots

We analyze the structural similarity of a set of 14 pseudoknots ranging from 41 to 161 nt. Figure 3 shows the secondary structures and graphical representations of the pseudoknots

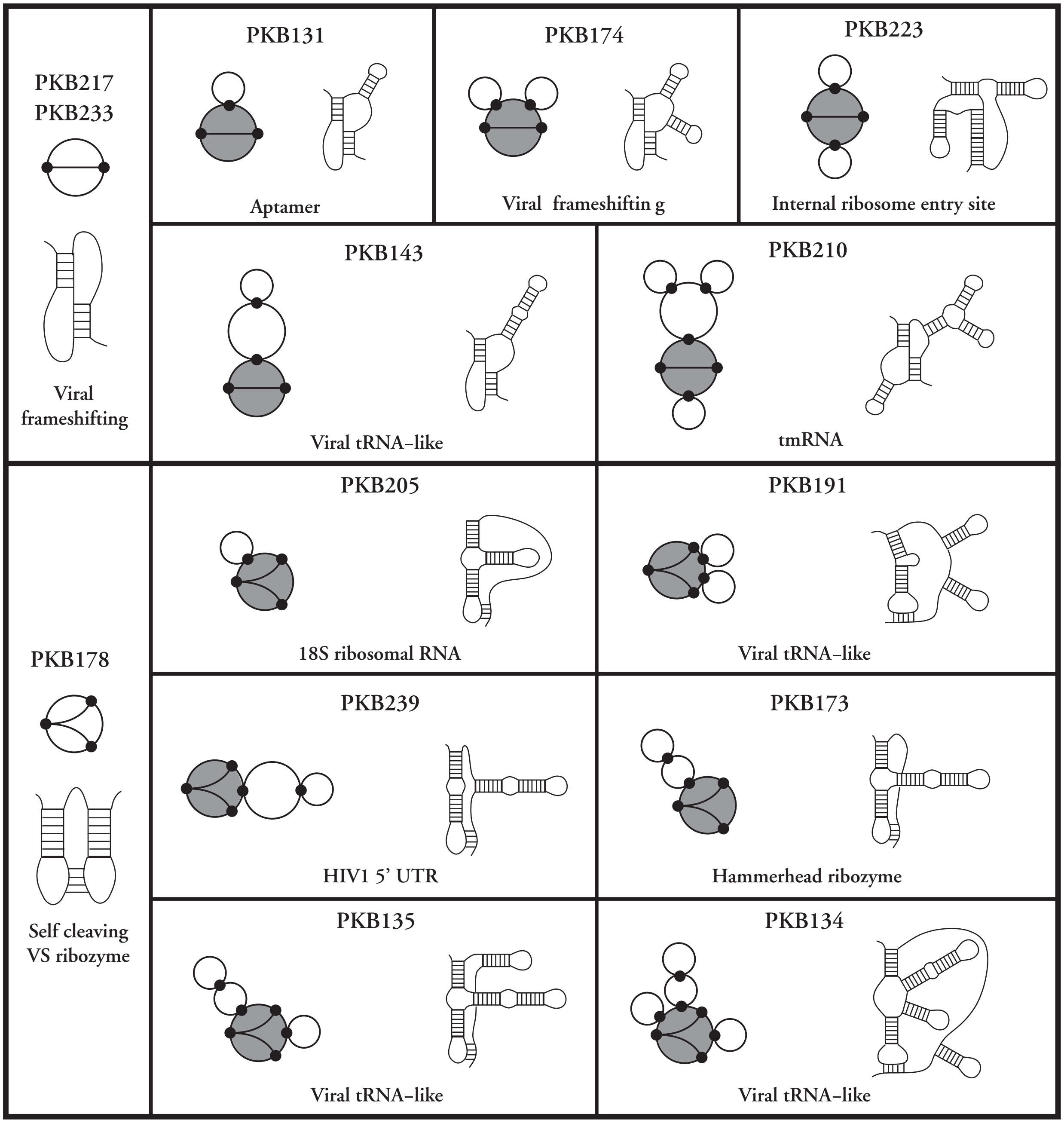


Figure 3. Topological matches generated by the simple (PKB217, PKB233) and double (PKB178) pseudoknots. The pseudoknots are shown as schematic secondary structures and dual graphs. There are 2 probe and 11 target pseudoknot topologies representing diverse functional RNA pseudoknots. For each probe pseudoknot structure, the corresponding matched submotifs/subgraphs in the larger pseudoknots are shaded.

selected from the PseudoBase (25) (http://wwwbio.leidenuniv.nl/~Batenburg/PKB.html). The RNA pseudoknots are viral frameshifting (PKB174, PKB217, PKB233), viral tRNA-like (PKB134, PKB135, PKB143, PKB191), IRES RNA (PKB223), 18S rRNA (PKB205), HIV-1 5′-untranslated region (5′-UTR) (PKB239), alternative pseudoknot of hammerhead ribozyme (PKB173), aptamer that binds human nerve growth factor (PKB131), tmRNA (PKB210) and self-cleaving *Neurospora* vs ribozyme (PKB178). (PKB233 and PKB217 appear the same but have different sequences, and this difference affects subsequent sequence alignment analysis.) These pseudoknot classes are not known to be functionally related. We choose more than one member for some classes because of exhibited topological differences.

The above pseudoknot structures were derived from mutational analysis (PKB173, PKB174, PKB191, PKB233, PKB239), enzymatic and chemical structure probing (PKB131, PKB174, PKB191, PKB239), phylogenetic or sequence comparison analysis (PKB131, PKB135, PKB143, PKB205, PKB217, PKB233, PKB234, PKB191, PKB210, PKB239) and 3D modeling (PKB134, PKB135, PKB210). The pseudoknots listed in more than one category were probed using two or more methods.

To use these structures for graphical analysis, we convert the base pairing information of each structure provided by PSEUDOBASE to a dual graphical representation using our dual graph rules [D1–D3 (2)]. We then use pseudoknot graphs to specify their corresponding adjacency matrices. These two steps are performed manually. For RNA tree structures, we use our RNA Matrix program to automatically convert a secondary structure into its corresponding adjacency matrix (26,35). The search for a small motif within a larger motif is performed using input adjacency matrices, and the comparison of matrices is performed via the topological invariant *S* (Equation 2).

Sequence and structural relationships of functionally related pseudoknots

We illustrate using several pseudoknot pairs the functional similarity identified by topological rather than sequence similarity. Figure 4 compares viral frameshifting pseudoknot structures. Viral RNA frameshifting is a process by which expression of overlapping open reading frames (ORFs) can be achieved (36). We compare three pairs from subfamilies Gag-pro, ORF1a/ORF1b and ORF2/ORF3 ribosomal frameshifting. Frameshifting signals involve two essential components: a ‘slippery’ (or sliding) sequence of nucleotides, where frameshifting takes place, and a stimulatory RNA secondary structure (usually a pseudoknot) located a few nucleotides downstream. Although members of the same frameshifting subfamilies have significant sequence similarity, global sequence identity between subfamilies is as low as 28%, similar to random sequences. Because all frameshifting pseudoknots in Figure 4A have the same topology and function, finding topological similarity between RNA structures has an advantage for functional annotation over sequence comparison.

Figure 4B shows that this finding also holds for a frameshifting pseudoknot pair (PKB217 and PKB174, Pair 4 in Table 1). The 72 nt PKB217 is the −1 frameshifting pseudoknot of lactate dehydrogenase-elevating virus (LDV) (37). Together with its sliding sequence UUUAAAC, it regulates the expression of ORF1a for a polyprotein containing proteases and ORF1b for an RNA-dependent RNA polymerase in LDV. The 127 nt PKB174 is the −1 frameshifting pseudoknot of Rous sarcoma virus (RSV) (36,38). This pseudoknot and sliding sequence AAAUUUA regulate the expression of Gag-Pro-Pol and Gag-Pol polyproteins. Although both the LDV and RSV frameshifting pseudoknots have a simple pseudoknot structure, the latter has two extra stem–loops. Despite their functional similarity, the global sequence identity is only 37% [computed using the program ALIGN (39) available at http://workbench.sdsc.edu], comparable with random sequences; their sliding sequences are also dissimilar. Using our local sequence alignment parameters, we find a large significant matched region in PKB217 (6833–6877), as shown in Figure 4B: AAACUGCUAGCCACCUCUGGUCUCGACC-GCUGUACUAGAGGUGG.

Although functional similarity between frameshifting pseudoknots is expected, this case illuminates how examining topological similarity aids in identifying pseudoknots with similar functions despite low overall sequence similarity.

Comparison of pseudoknot structures

Our two probe motifs—a simple pseudoknot (PKB217 or PKB233) and a double pseudoknot (PKB178)—are used to search larger RNAs. We generate two groups of topological matches, as shown in Figure 3, with the matched motifs shown darkened; we found 11 such matched pseudoknot pairs. In addition, we analyze sequence similarity of many of the 84 (14 × 13/2) possible pseudoknot pairs since topological similarities are apparent within each pseudoknot group generated by the probe motifs. Although not exhaustive, this additional analysis may help to identify pairs having similar substructures that are not detected by using fixed probes (simple and double pseudoknots). Table 1 lists 10 pseudoknot pairs (out of the 11 probe/structure matches and subset of the 84 structure–structure analysis) that have significant structural and sequence similarities. Of the total of 10 pairs listed (Pairs 1–10), two known functional pairs (1 and 4 in Table 1) are included for comparison with the eight other pairs that are not previously known to be functionally related.

We begin by analyzing the overall patterns of relationship among our 14 pseudoknots. Figure 3 organizes the matches into two groups: five matches for the simple pseudoknot (PKB217) and six for the double pseudoknot (PKB178). The matched motifs appear to be ‘modular constructions’ around the probe motif/module, with added stem–loops (three or fewer). To the best of our knowledge, these results reveal modular features and topological similarities of RNA pseudoknots that have not been described.

The biological significance of topological matches depends on the size and complexity of pseudoknots: a match for the double pseudoknot is probably more significant than for a simple pseudoknot. The six pseudoknots generated by the double pseudoknot probe PKB178 show striking topological as well as functional similarity (e.g. viral tRNA-like PKB134 and PKB135); we analyze their similarity with PKB178 and among themselves. Still, the matched pairs of RNAs are functionally diverse: besides PKB134 and PKB135, such examples

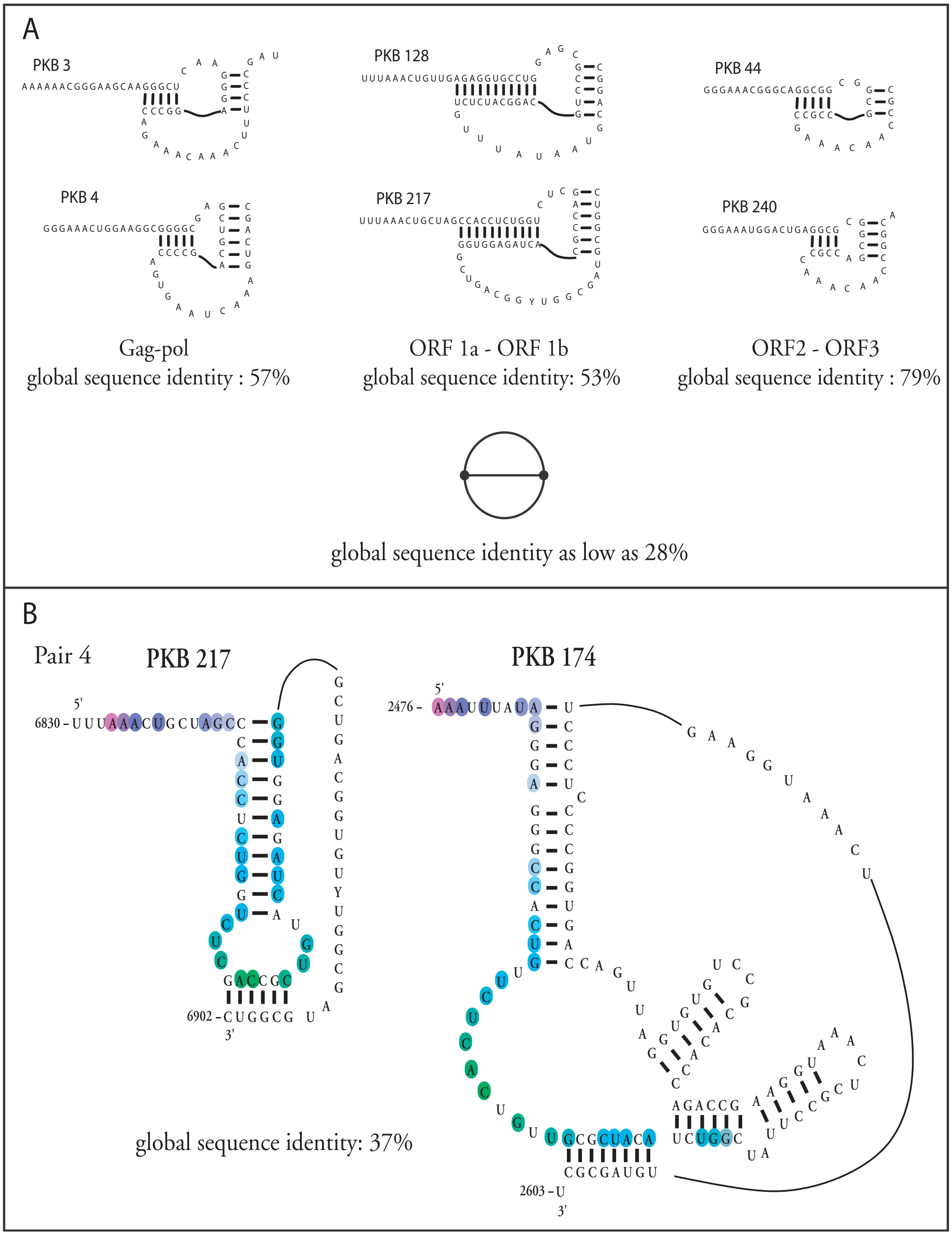


Figure 4. Sequence and structural comparisons of pseudoknots in the frameshifting functional families. (**A**) Two secondary structures and corresponding topologies for each of the three viral frameshifting subfamilies Gag-pro, ORF1a/ORF1b and ORF2/ORF3. Sequence identity within the same subfamily ranges between 50 and 80%; it is as low as 28% when comparing pseudoknots from different subfamilies. All frameshifting pseudoknots are functionally related and have the same topology (dual graph). (**B**) Secondary structures and their aligned nucleotides (color coded) for frameshifting pseudoknots of LDV (PKB217) and Rous sarcoma virus (PKB174). Global sequence alignments were performed using the program ALIGN with default (−16/−4) gap opening/gap extension penalties.

of diverse RNAs, include IRES (PKB223) RNA, 18S rRNA (PKB205, PKB234), HIV-1 5′-UTR (PKB239) and alternative pseudoknot of hammerhead ribozyme (PKB173). In this double pseudoknot group, six pairs (1, 3, 5, 6, 8 and 9 in Table 1) produce local sequence alignments [using LALIGN (40), http://workbench.sdsc.edu] that suggest significant results. In particular, the pair PKB134/PKB173 (Pair 5 in Table 1) indicates a significant match. The pairs PKB134/PKB205 (Pair 6) and PKB205/PKB223 (Pair 10 in Table 1) also display intriguing sequence and functional similarities.

These relationships suggest structural and functional similarities that are not previously known (Table 1).

To support the significance of these matches, we perform sequence alignments using Smith–Waterman parameters for gap opening and gap extension penalties (41). We choose the gap opening penalty value of -15 from a range of 0 to -19. To produce better alignments for large sequence segments (short alignments are less likely to be significant in our context), our gap extension penalty of -1 from a range of -1 to -8 allows larger extensions, consistent with the possibility of the insertion of long sequences and added 2D motifs. In Table 1, we also report the alignment score, which is the sum of nucleotide matches, mismatches and gap penalties.

Table 1 shows that the probe structure (PKB178) matches two other pseudoknots (PKB135, PKB173) with sequence identities ranging from 59 to 71% for 24–49 nt alignment lengths (compared with 56% sequence identity over 129 nt for the functionally related pair PKB134/PKB135, Pair 1). Significantly, the pairs PKB134/PKB173 (Pair 5), PKB134/PKB205 (Pair 6) and PKB205/PKB223 (Pair 10) have sequence identity range of 72–80%. The non-randomness of the 10 aligned pairs in Table 1 is also evident in the ratio *R*, which measures the level of significance above the random expectation, and has value 1 for random pairs:

R = (percent sequence identity)/(percent sequence identity when one of the aligned sequences is randomized).

We also define <*R*> as the *R*-value averaged over alignments with different randomized sequences. Table 1 shows that some pseudoknot pairs with no known functional relationship have <*R*> values ranging between 1.06 and 1.47, whereas the related pair PKB134/PKB135 has a value of 1.54; many matched pairs have <*R*> values <1. Thus, the topological similarities we identify led to matched pairs with non-random sequence alignment results. Non-random sequence matches may arise from topological and/or functional similarities, as illustrated below for three pseudoknot pairs (5, 6 and 10 in Table 1).

The rate of false positives generated by topological and <*R*> value screening can be easily estimated by considering all possible pairs from the set of six matched double pseudoknots in Figure 3. These pseudoknots form 15 distinct pairs, 6 of which have <*R*> values >1.14 [Table 1, with $S(G)$ = 1.17]. Table 1 shows that only four pairs have large <*R*> values (>1.2): Pair 1 (PKB134/PKB135) is related, Pair 5 (PKB134/PKB173) and Pair 6 (PKB134/PKB205) are most probably related, and Pair 3 (PKB178/PKB173) is of unknown relationship (in fact, both are self-cleaving ribozymes). Assuming that the Pair 3 is unrelated and therefore reflects a false positive, we have an error rate of 1/15, or 7%. If the inferred functionally related Pairs 5 and 6 are also considered as false positives, we then have a conservative error rate of 20%.

Functional relationship between PKB134 and PKB173 (Pair 5 in Table 1)

The 116 nt PKB134 pseudoknot occurs in the 3′ end of brome mosaic virus genomic RNA (31). The 73 nt PKB173 is a hammerhead ribozyme of satellite cereal yellow dwarf virus-RPV (satRPV) RNA capable of adopting an alternative pseudoknot conformation by forming an (L1–L2a) stem whose existence is established by experimental mutational analysis (42). Figure 3 shows that their topologies have the same 'core', but PKB134 has two extra stem–loops. The local sequence alignment of this pair using LALIGN (Figure 5) shows 19 bases aligned to 80% sequence identity and an <*R*> value of 1.23 (Table 1). There are two significant segments in PKB173 that are matched with those in PKB134: UACUGU-CUGACGACGUAUCC (nt 11–30) and UAGAAGGCUG-GUGCC (nt 39–53). These segments and their matched regions in PKB134 span the stem and unpaired elements of the secondary structures (see Figure 5). Significantly, both PKB134 and PKB173 pseudoknots serve as template for replicase, a protein that recognizes specific sequence regions to initiate replication of viral genomes (31,42). Moreover, a conserved sequence (CUGANGA) of PKB173 is in the aligned region (nt 11–30). Thus, our analysis suggests that PKB134 and PKB173 pseudoknots are structurally and functionally related.

Further analysis of pseudoknots PKB134 and PKB173 reveals specific domains interacting with the replicase. Experiments have shown that mutations in domains A, C, B1 and B2 [defined in (31)] of PKB134 impair viral RNA replication (43). Significantly, our two aligned sequence regions span domains A, B1, C and also D. Thus, the replicase recognizes several domains in the PKB134 pseudoknot structure. In PKB173, the pseudoknot forming five base pair helix (GCGCG) is implicated in replicase recognition or ligation (42). This pseudoknot also has conserved sequence regions ACAAA (61–65) or its complement (possibly the replication origin) and AGAAA (nucleotides attached to the 3′ end but not shown). However, these regions are not observed in PKB134.

Functional relationship between PKB205 and PKB223 (Pair 10 in Table 1)

The 63 nt PKB223 is the pseudoknot of IRES of capsid-producing cricket paralysis-like viruses (CrPV) for methionine-independent initiation of translation (44). Appearing immediately upstream from the capsid-coding sequence, it initiates translation without a canonical AUG (methionine) initiation codon. The 48 nt PKB205 is the pseudoknot substructure (V4 domain) of SSU of *Palmaria palmata* 18S rRNA, which was found using extensive comparative analysis (45,46). Figure 5 (Pair 10, lower panel) shows that PKB223 is a 4-vertex simple pseudoknot with two inserted stem–loops and PKB205 is a 4-vertex double pseudoknot with an inserted stem–loop. Figure 5 (middle panel) displays the local sequence alignment of PKB223 and PKB205, yielding a match of 14 nt with 72% sequence identity. This significant sequence similarity occurs in regions ACAAUAUCCAGGAA (6148–6161) and ACAUUAGCAUGGAA (765–778) of PKB223 and PKB205, respectively. Remarkably, we find that the 9 nt segment 5′-UUAGGUUAG-3′ (nt 6100–6108) of PKB223 is complementary to 3′-GGUACGAUU-5′ (nt 768–776) of PKB205 except at G6103. Our finding corroborates with the experimental identification that IRES elements in cellular mRNA contain a short 9 nt (nt 133–141, CCGGCGGGU) in the 5′-UTR of the homeodomain protein Gtx, which is complementary to nt 1132–1124 of 18S rRNA (47). Indeed, many such near complementary sequences have been found between the 3′ end of 18S rRNA and picornavirus RNAs (48).

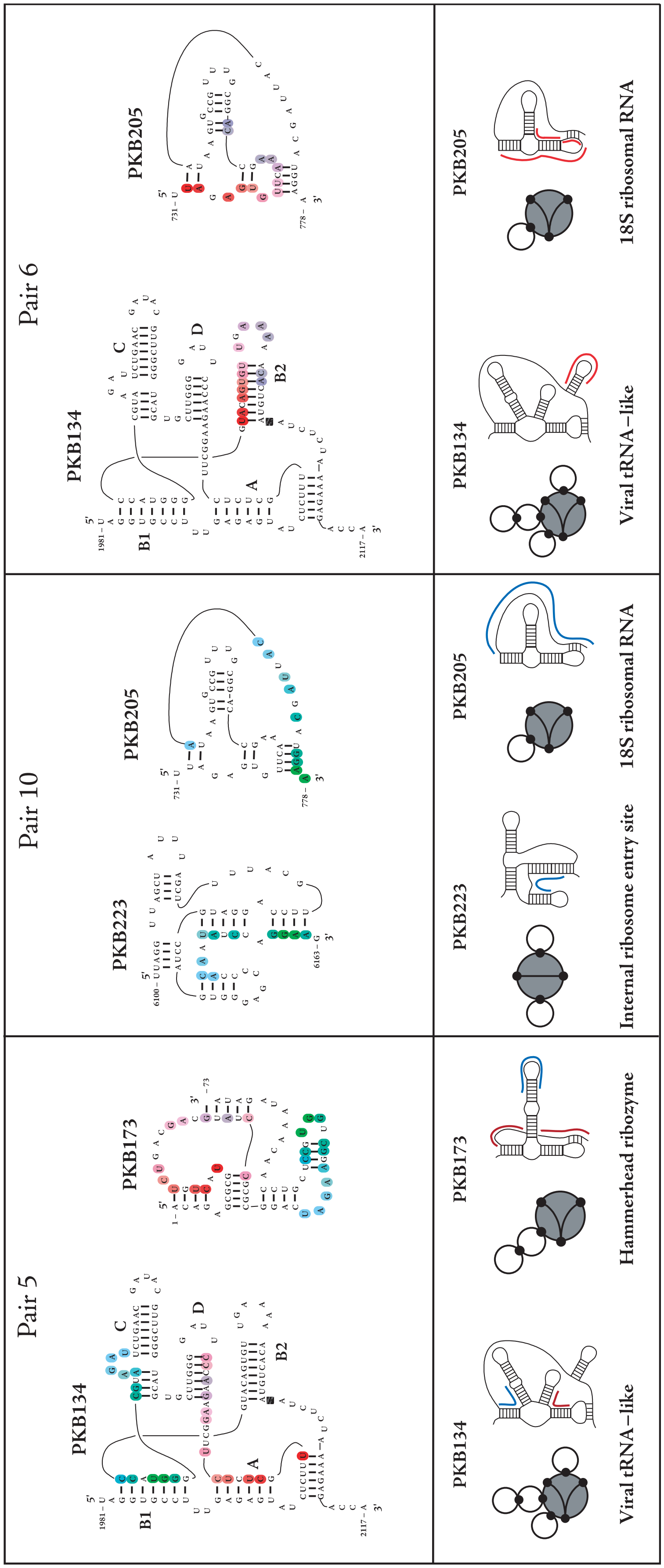


Figure 5. Three novel functionally related pseudoknot pairs inferred by topology matching and sequence alignment. Upper panels: Secondary structures and their aligned nucleotides (color coded) for pseudoknot pairs PKB134/PKB173 (Pair 5, Table 1; left panel), PKB205/PKB223 (Pair 10, Table 1; middle panel) and PKB134/PKB205 (Pair 6; right panel). PKB134 occurs in 3′ end of brome mosaic virus genomic RNA; PKB173 is a pseudoknot of satellite cereal yellow dwarft virus-RPV RNA; PKB223 is the pseudoknot of internal entry site (IRES) of caspid-producing CrPVs; and PKB205 is the pseudoknot substructure (V4 domain) of SSU of *P.Palmata* 18S rRNA. Lower panels: The aligned subgraphs or regions are shaded or marked.

Table 2. Short sequence regions in PKB205 complementary to tRNA-like molecules

PKB205 regions	Complementary sequences
3′-GGUACGAUU-5′ (768–776)	PKB223 5′-UUAGGUUAG-3′ (6100–6108), IRES
3′-GUGAGAUU-5′ (769–776)	PKB134 5′-CACUGUAAA-3′ (113–120), tRNA-like
3′-UACGAUUAC-5′ (767–775)	PKB191 5′-AUGCUCAUG-3′ (77–85), tRNA-like
3′-UGUGAGAUU-5′ (731–739)	PKB138 5′-ACACUUUAA-3′ (38–47), tRNA-like
3′-GUUUGCGGACC-5′ (746–756)	PKB17 5′-CAAAACCCUGG-3′ (19–29), tRNA-like

Intriguingly, for CrPVs, the complementary segments occur in pseudoknots with similar folds not involving the dangling 3′ end of 18S rRNA. Base pair complementary between IRES elements of mRNA and 18S rRNA is a mechanism for ribosome recruitment and translation. In viral IRES elements, the secondary and tertiary structures are believed to play critical roles in translation initiation (44).

Functional relationship between PKB205 and PKB134 (Pair 6 in Table 1)

The functional relationsip of the pair PKB205/PKB134 is similar to that for PKB205/PKB223 above. PKB134 belongs to a common viral tRNA-like functional class whose role is to recruit host ribosomes to initiate viral protein synthesis (49). It is known to interact with the ribosome via tRNA shape mimicry (31). Our analysis indicates that tRNA-like molecules also contain a region, which is nearly complementary to PKB205, a pseudoknot region of the 18S rRNA. Table 2 shows regions of PKB205 complementary to four tRNA-like molecules (PKB134, 191, 138, 17); the complementary region in PKB223 is shown for comparison. Significantly, each of the pseudoknots PKB223, PKB134 and PKB191 has a region complementary to the same 8–9 nt unpaired region in PKB205 (767–776). This single-stranded region is most probably available for base pairing with incoming tRNA-like molecules. In addition, two other regions in PKB205 are complementary to tRNA-like pseudoknots, PKB138 and PKB17; one region is a hairpin loop and another region is part of short stems. Thus, this analysis suggests that tRNA-like molecules may employ both structural mimicry and sequence complementarity to recruit SSU rRNA, an interaction mode similar to tRNA. PKB134 also has a region UACAGUGUUGAAAAACA (nt 2078–2094) closely matched to the region UAGAGUGUUCAAAGCCA (nt 732–748) in PKB205, with an average *R*-value of 1.23.

Tertiary structure similarity of topological matches in ribosomal RNAs

The preceding examination of three pseudoknot pairs shows that topological matches can suggest functionally related pseudoknots. It is also important to show that novel topological matches led to similar tertiary structures. We focus for this purpose on rRNA since several bacterial and archaeal ribosome structures have been solved in recent years. In fact, in our previous work, we have found several topological matches of the *S.cerevisiae* 5S rRNA motif within its larger rRNAs: two within 16S rRNA (domains II and III) and three within 23S rRNA (domains III, IV and VI) (2). From available X-ray structures for *Thermus thermophilus* 16S rRNA (PDB accession no. 1fjg) (50) and 23S rRNA from *Deinococcus radiodurans* (1nkw) and *Haloarcula marismortui* (1ffk) (51,52), only 5S rRNA motif's matches in domain II of 16S rRNA and domain III of 23S rRNA are relevant; the other three cases are not valid matches in these species.

Figure 6 displays the secondary diagrams, graphical representations and tertiary structures of 5S rRNA for bacterium *D.radiodurans* (1nkw) and archaea *H.marismortui* (1ffk), and their topologically matched modules within 16S and 23S rRNAs; the secondary diagrams are adapted from those in Gutell's comparative RNA website (http://www.rna.icmb.utexas.edu/) (53). The module in 23S is an identical match for the 6-vertex 5S graph except for a hairpin loop, which does not contribute to the topological invariant *S*. The 16S module is a 5-vertex graph because a 2 bp stem at the junction has a non-canonical base pair AG (dual graph rule D1 defines a stem as having at least two canonical and/or wobble base pairs); for *S.cerevisiae*, this stem consists of only canonical base pairs, making the 16S module an identical match for the 5S topology. As shown in Figure 6, the 16S and 23S modules have a similar L-shaped tertiary fold as the 5S structure. Tertiary structure similarity with 5S is more striking for the module of domain II (nt 653–753) of *T.thermophilus* 16S rRNA than for the module of domain III (nt 1033–1143) of 23S rRNA. Both the 5S structure and 16S module have a pronounced 3-stem junction, where the chain ends of the 16S module are located. Structural (backbone) overlap of the 5S structure and 16S module shows similar structural features, except for the orientations of the smaller stems at the junction. (We aligned the structures manually because the chain ends of 5S and 16S module occur at non-corresponding locations.) In contrast, the 23S module shows significant distortions on the shorter helical arm, with the hairpin loop folded back to the helix, as shown in the structural overlap of the 5S structure and 23S module. Structural similarities seen in these examples are probably due to geometric constraints, which are apparent in their secondary diagrams and graphical representations. Thus, 2D topological matches can lead to similar 3D structures. Below, we survey the distribution of modules in rRNAs.

Distribution of submotifs in 16S and 23S ribosomal RNAs

Characterizing the submotifs within the 16S and 23S rRNAs, the largest known RNA molecules, can help uncover the modular construction of RNA structures, as analysis of rRNA's tertiary modules has revealed. Furthermore, our previous analysis of rRNA's structural motifs indicated existence of characteristic features (35). For example, the distributions of paired/unpaired bases in stems, bulges/internal loops, hairpin loops and junctions follow specific functional forms.

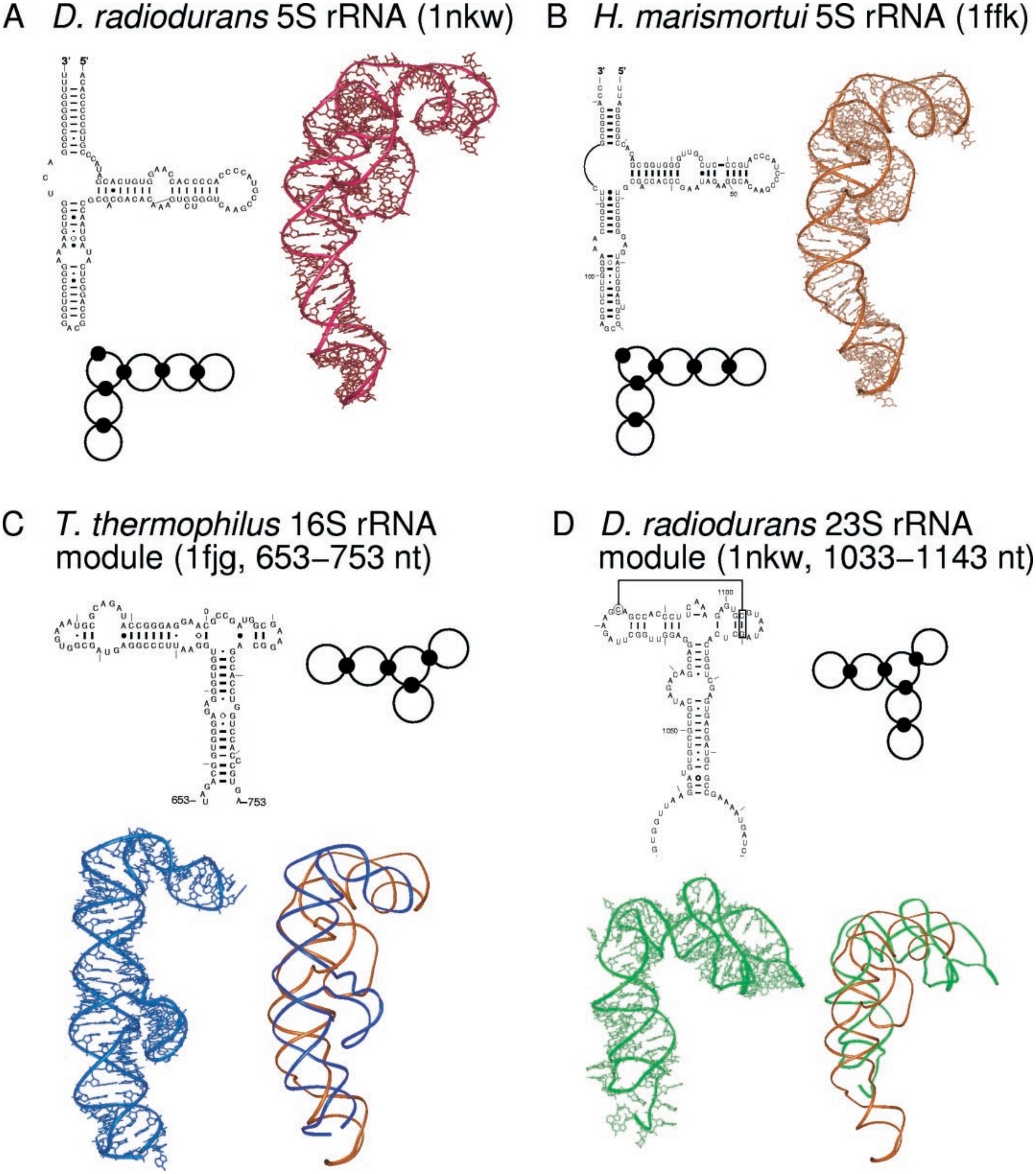

A D. radiodurans 5S rRNA (1nkw)
B H. marismortui 5S rRNA (1ffk)
C T. thermophilus 16S rRNA module (1fjg, 653–753 nt)
D D. radiodurans 23S rRNA module (1nkw, 1033–1143 nt)
3'
5'
50
100
653–
–753
1100
1050

Figure 6. Comparison of the graph topologies and 2D and 3D structures of 5S rRNAs (**A** and **B**) and modules in 16S and 23S rRNAs (**C** and **D**); (C and D) Two structural (backbone) overlaps of the 5S structure (1ffk; gold color) with 16S module (1fjg; blue) and 5S structure with 23S module (1nkw; green) are shown. The secondary diagrams are adapted from those in Gutell's comparative RNA website (http://www.rna.icmb.utexas.edu/) and 3D structures are from the Protein Data Bank (http://www.rcsb.org/pdb/). The 2D modules were identified computationally by our earlier work on finding ∼120 nt 5S rRNA motifs in 16S and 23S rRNAs of *S.cerevisiae*.(2) The equivalent tertiary structures and modules for 5S, 16S and 23S rRNAs from other organisms show that similar graph topologies can lead to similar secondary and tertiary structures. All structures have similar graph topologies. The *T.thermophilus* 16S module is a 5-vertex instead of 6-vertex graph because a 2 bp stem at the junction has a non-canonical base pair AG; for *S.cerevisiae*, this stem consists of only canonical base pairs, making the 16S module an identical match for the 5S topology.

Moreover, the percentages of nucleotides in stems, bulges/internal loops, hairpin loops and junctions do not vary for 16S and 23S rRNAs, even though the latter is twice the size of the former. Such motif characteristics help discriminate RNA-like from non-RNA-like molecules and guide the design of novel RNAs. Here, we analyze the distinct modular units—small tree submotifs—that makeup the rRNAs to advance such applications. Since all small tree graphs up to 10 vertices have been exhaustively enumerated (54,55), we can use our substructure search algorithm to identify the occurrences of all distinct tree submotifs in rRNAs.

Specifically, we use sets of tree graphs for $V = 4, 5, 6, 7, 8, 9$ and 10 comprising 2, 3, 6, 11, 23, 47 and 106 graphs, respectively, for a total of 198 tree motifs. The complete enumerated motifs can be found in the RAG web resource (http://monod.biomath.nyu.edu/rna/). Figure 7 shows, as an example, the 11 possible 7-vertex tree graphs and lists the frequency of occurrence of each tree graph in rRNAs. Remarkably, the same tree motifs are abundant in both 16S and 23S rRNAs. The most abundant (10 or 15 times) 7-vertex tree has two 3-stem junctions (topological invariant $S^{\text{tree}} = 6.160$); 4- and 5-stem junctions are the next most prevalent submotifs ($S^{\text{tree}} = 5.137$, 5.921 and 5.667). Interestingly, the unbranched tree ($S^{\text{tree}} = 7.219$) and highly branched trees ($S^{\text{tree}} = 4.509$) are rare.

Figure 8 plots the percentage of distinct tree graphs in each set V (fraction of motifs identified relative to possible motifs) found in rRNA. For small tree graphs, $V = 4$, 5 and 6, all possible topologies are present, but not for $V > 6$. In fact, the percentage of trees found declines with V. In 23S rRNA, ~90% of possible trees are found for $V = 7$, and only ~40% (representing ~47 tree topologies) are found for $V = 10$. Larger trees ($V > 10$) are expected to have much smaller percentages. This percentage decline is expected since small tree modules are less specific than large trees. In total, 111 submotifs out of 198 are identified. The observed trends in Figure 8 may be generally valid for rRNA structures of other organisms since they are highly conserved, although some statistical fluctuations are expected due to structural differences and RNA's finite size.

These data show that RNA structures are modular constructs of relatively few (tree) building blocks. Extrapolation of the curves in Figure 8 suggests that ~20% of $V = 11$ motifs (total of 235 trees) and ~10% of $V = 12$ motifs (total of 551 trees) probably exist in rRNAs. Thus, the total number of building blocks in rRNAs with $V < 13$ would be ~210. Although the percentages of trees with higher V values are low, they are more numerous and will contribute to the overall number of building blocks. We thus consider 210 as the minimum number of building blocks. The building block number may be exploited in the future design of novel functional RNA molecules using modular assembly from 210 distinct tree motifs. The tree motifs may be taken from substructures of existing RNAs, allowing simultaneous sequence and structure

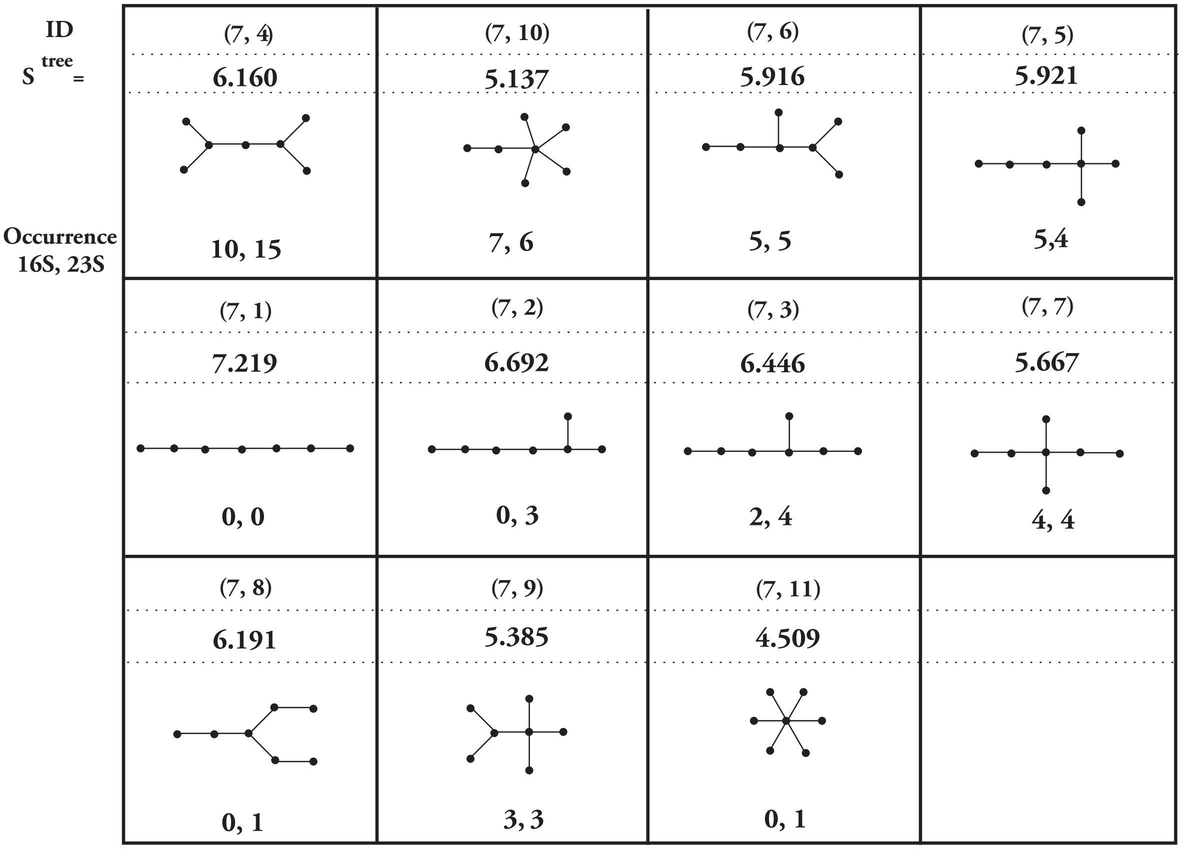


Figure 7. Occurrence frequency of 11 distinct 7-vertex tree topologies as submotifs of *S.cerevisiae* 16S and 23S rRNAs. The tree graphs are labeled using topological invariant S^{tree} (second top row) and ID number (i, j) (first row), which refers to motif identification number as used in RAG database (http://monod.biomath.nyu.edu/rna). The two numbers at the bottom of each graph refer to the number of times that topology is found embedded in 16S and in 23S rRNAs, respectively.

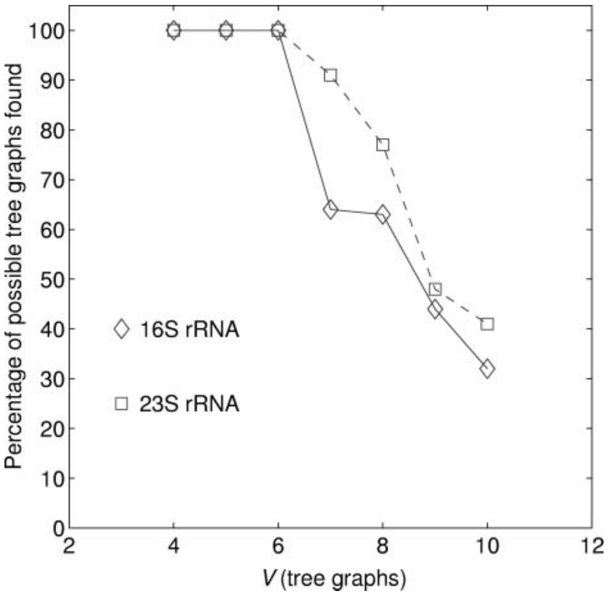


Figure 8. Distribution of all possible distinct tree modules with *V* vertices found in *S.cerevisiae* 16S and 23S rRNAs. In graph theory, each *V* is represented by a set of possible tree structures. For example, for *V* = 4, 5, 6, 7, 8, 9 and 10 there are 2, 3, 6, 11, 23, 47 and 106 possible trees, respectively. The distribution of trees quantifies the diversity of tree modules within rRNAs. This analysis suggests that rRNAs possess at least 210 tree modules.

design. Such a build-up approach appears feasible from our (2,6) and other works (3,5) and is akin to the 'buildup' technique for proteins pioneered by Vasquez and Scheraga (56).

The above estimate of building block number can be either a consequence of the limited number of available structures or the fact that RNA adopts only a subset of possible graphs. Certainly, future discoveries of novel RNA structures might indicate a larger repertoire of building blocks. Still, our surveys of natural RNAs (2,6) suggest that RNAs possess specific topological characteristics and favor a subset of the possible graph motifs. The missing motifs may not exist in nature for RNA owing to physical or functional constraints.

The building block number can also be defined in terms of irreducible graphs, i.e. trees that cannot be decomposed into smaller trees. However, this approach has a major drawback. Since all RNAs are composed of connected stems, their substructures can be reduced to two vertex graphs, which is not an informative description of complex RNA folds. Still, if we define 4-vertex trees as the smallest irreducible graphs, the number of irreducible graphs can be calculated. Based on our submotif data, we estimate that for *V* up to 10, there are about 42 irreducible tree modules in rRNAs.

SUMMARY AND CONCLUSION

Our automated comparison of RNA secondary structures based on graphical representation allows the examination of conserved 2D RNA topological features, which in turn suggest functional relationships. This approach has an advantage over sequence alignment because functionally related RNAs often lack sequence similarity, as shown for frameshifting pseudoknots with low global sequence identities. Of course, our structure alignment approach relies on available 2D RNA structures and thus will increase in scope as more RNA structures become available. Our eight identified pairs of potentially related pseudoknots (Table 1) reveal three pairs that are functionally related: the pseudoknots of 3′ end of brome mosaic virus genomic RNA (PKB134) and alternative hammerhead ribozyme pseudoknot (PKB173) are templates for replicase to initiate replication of viral genomes; the pseudoknots of internal entry site (IRES) of caspid-producing CrPVs (PKB223) and SSU of 18S rRNA (PKB205) are associated with translation initiation; and the pseudoknots of 18S rRNA (PKB205) and viral tRNA-like (PKB134) most probably interact via both structural mimicry and base complementarity. To the best of our knowledge, these functional relationships have not been described previously.

The developed RNA comparison algorithm also led us to quantify the modularity of 16S and 23S rRNAs using small enumerated tree motifs. Our results suggest that rRNAs are constructed from at least 210 distinct tree subtopologies (or modules) having up to ∼180 nt. Significantly, we find that 5S rRNA and two tree modules within 16S and 23S rRNAs with similar topologies have similar tertiary shapes, affirming the relationship between 2D graphical representation and 3D structure. The tree modules in rRNA may be used as building blocks for the design of novel RNA folds. A similar idea has been employed to design and improve functional RNA molecules (5,10). Design by assembly of modules exploits motifs present in natural RNAs unlike *de novo* design of sequences and structures (11). The issues of discovering novel topologies (6), building module library, and designing functional folds are likely to increase in importance with interest in novel RNAs.

Our algorithm has several limitations. Of course, its utility depends on availability of reliable RNA secondary folding algorithms (57,58). More fundamentally, although simplified graphical representations specify topological characteristics, detailed sequence and motif information is missing. An associated problem is that unusual RNA structure–function relationships due to deleterious mutations, as in frameshifting pseudoknots (28), and structural flexibility are not captured. More advanced graph constructs such as weighted graphs may, however, specify detailed motif features at the secondary and base pair levels. Another limitation of the present algorithm is the requirement to fix the probe structure or graph. As in protein structure alignment algorithms (15), future refinements of our algorithm should allow some flexibility in motif definitions as the similarity search proceeds. We invite RNA scientists to visit our RAG web resource at http://monod.biomath.nyu.edu/rna/ and suggest to us specific enhancements that will be useful in practice.

ACKNOWLEDGEMENTS

We thank Dinshaw Patel for comments that motivated this work. This research was supported by a Joint NSF/NIGMS Initiative in Mathematical Biology (DMS-0201160) and the Human Frontier Science Program (HFSP).

REFERENCES

1. Tinoco,I. and Bustamante,C. (1999) How RNA folds. *J. Mol. Biol.*, **293**, 271–281.

2. Gan,H.H., Pasquali,S. and Schlick,T. (2003) Exploring the repertoire of RNA secondary motifs using graph theory: implications for RNA design. *Nucleic Acids Res*., **31**, 2926–2943.
3. Soukup,G.A. and Breaker,R.R. (1999) Engineering precision RNA molecular switches. *Proc. Natl Acad. Sci. USA*, **96**, 3584–3589.
4. Breaker,R.R. (2002) Engineered allosteric ribozymes as biosensor components. *Curr. Opin. Biotechnol*., **13**, 31–39.
5. Ikawa,Y., Fukada,K., Watanabe,S., Shiraishi,H. and Inoue,T. (2002) Design, construction, and analysis of a novel class of self-folding RNA. *Structure*, **10**, 527–534.
6. Kim,N., Shiffeldrim,N., Gan,H.H. and Schlick,T. (2004) Candidates for novel RNA topologies. *J. Mol. Biol*., **341**, 1129–1144.
7. Mitchell,J.R., Cheng,J. and Collins,K. (1999) A box H/ACA small nucleolar RNA-like domain at the human telomerase RNA 3′ end. *Mol. Cell. Biol*., **19**, 567–576.
8. Doudna,J.A. and Cech,T.R. (2002) The chemical repertoire of natural ribozymes. *Nature*, **418**, 222–228.
9. Storz,G. (2002) An expanding universe of noncoding RNAs. *Science*, **296**, 1260–1263.
10. Breaker,R.R. and Joyce,G.F. (1994) Inventing and improving ribozyme function: rational design versus iterative selection methods. *Trends Biotechnol*., **12**, 268–275.
11. Andronescu,M., Fejes,A.P., Hutter,F., Hoos,H.H. and Condon,A. (2004) A new algorithm for RNA secondary structure design. *J. Mol. Biol*., **336**, 607–624.
12. Carter,R.J., Dubchak,I. and Holbrook,S.R. (2001) A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Res*., **29**, 3928–3938.
13. Doudna,J.A. (2000) Structural genomics of RNA. *Nature Struct. Biol*., **7**, 954–956.
14. Rivas,E., Klein,R.J., Jones,T.A. and Eddy,S.R. (2001) Computational identification of noncoding RNAs in *E.coli* by comparative genomics. *Curr. Biol*., **11**, 1369–1373.
15. Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–602.
16. Shindyalov,I.N. and Bourne,P.E. (2001) A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. *Nucleic Acids Res*., **29**, 228–229.
17. Benedetti,G. and Morosetti,S. (1996) A graph-topological approach to recognition of pattern and similarity in RNA secondary structures. *Biophys. Chem*., **59**, 179–184.
18. Margalit,H., Shapiro,B.A., Oppenheim,A.B. and Maizel,J.V.,Jr (1989) Detection of common motifs in RNA secondary structures. *Nucleic Acids Res*., **17**, 4829–4845.
19. Shapiro,B.A. and Zhang,K.Z. (1990) Comparing multiple RNA secondary structures using tree comparisons. *Comput. Appl. Biosci*., **6**, 309–318.
20. Le,S.Y., Nussinov,R. and Maizel,J.V. (1989) Tree graphs of RNA secondary structures and their comparisons. *Comput. Biomed. Res*., **22**, 461–473.
21. Duarte,C.M., Wadley,L.M. and Pyle,A.M. (2003) RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Res*., **31**, 4755–4761.
22. Chevalet,C. and Michot,B. (1992) An algorithm for comparing RNA secondary structures and searching for similar substructures. *Comput. Appl. Biosci*., **8**, 215–225.
23. Wang,L. and Zhao,J. (2003) Parametric alignment of ordered trees. *Bioinformatics*, **19**, 2237–2245.
24. Fontana,W., Konings,D.A.M., Stadler,P.F. and Schuster,P. (1993) Statistics of RNA secondary structures. *Biopolymers*, **33**, 1389–1404.
25. van Batenburg,F.H.D., Gultyaev,A.P., Pleij,C.W.A., Ng,J. and Oliehoek,J. (2000) PseudoBase: a database with RNA pseudoknots. *Nucleic Acids Res*., **28**, 201–204.
26. Gan,H.H., Fera,D., Zorn,J., Shiffeldrim,N., Tang,M., Laserson,U., Kim,N. and Schlick,T. (2004) RAG: RNA-As-Graphs database—concepts, analysis, and features. *Bioinformatics*, **20**, 1285–1291.
27. Fera,D., Kim,N., Shiffeldrim,N., Zorn,J., Laserson,U., Gan,H.H. and Schlick,T. (2004) RAG: RNA-As-Graphs web resource. *BMC Bioinformatics*, **5**, 88.
28. Chen,X.Y., Kang,H.S., Shen,L.X., Chamorro,M., Varmus,H.E. and Tinoco,I. (1996) A characteristic bent conformation of RNA pseudoknots promotes −1 frameshifting during translation of retroviral RNA. *J. Mol. Biol*., **260**, 479–483.
29. Kang,H.S., Hines,J.V. and Tinoco,I. (1996) Conformation of a non-frameshifting RNA pseudoknot from mouse mammary tumor virus. *J. Mol. Biol*., **259**, 135–147.
30. Gan,H.H., Perlow,R.A., Roy,S., Ko,J., Wu,M., Huang,J., Yan,S., Nicoletta,A., Vafai,J., Sun,D. *et al*. (2002) Analysis of protein sequence/structure similarity relationships. *Biophys. J*., **83**, 2781–2791.
31. Felden,B., Florentz,C., Giege,R. and Westhof,E. (1994) Solution structure of the 3′-end of brome mosaic-virus genomic RNAs—conformational mimicry with canonical transfer-RNAs. *J. Mol. Biol*., **235**, 508–531.
32. Gross,J.L. and Yellen,J. (1999) *Graph Theory and its Applications*. CRC Press, Boca Raton, FL.
33. Köbler,J., Schöning,U. and Torán,J. (1993) *The Graph Isomorphism Problem*. Birkhäuser, Boston, MA.
34. Randic,M. and Zupan,J. (2001) On interpretation of well-known topological indices. *J. Chem. Inform. Comput. Sci*., **41**, 550–560.
35. Zorn,J., Gan,H.H., Shiffeldrim,N. and Schlick,T. (2004) Structural motifs in ribosomal RNAs: implications for RNA design and genomics. *Biopolymers*, **73**, 340–347.
36. Jacks,T., Madhani,H.D., Masiarz,F.R. and Varmus,H.E. (1988) Signals for ribosomal frameshifting in the Rous-Sarcoma virus Gag–Pol region. *Cell*, **55**, 447–458.
37. Godeny,E.K., Chen,L., Kumar,S.N., Methven,S.L., Koonin,E.V. and Brinton,M.A. (1993) Complete genomic sequence and phylogenetic analysis of the lactate dehydrogenase-elevating virus (Ldv). *Virology*, **194**, 585–596.
38. Marczinke,B., Fisher,R., Vidakovic,M., Bloys,A.J. and Brierley,I. (1998) Secondary structure and mutational analysis of the ribosomal frameshift signal of Rous sarcoma virus. *J. Mol. Biol*., **284**, 205–225.
39. Myers,E.W. and Miller,W. (1988) Optimal alignments in linear space. *Comput. Appl. Biosci*., **4**, 11–17.
40. Huang,X.Q. and Miller,W. (1991) A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math*., **12**, 337–357.
41. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol*., **147**, 195–197.
42. Song,S.I., Silver,S.L., Aulik,M.A., Rasochova,L., Mohan,B.R. and Miller,W.A. (1999) Satellite cereal yellow dwarf virus-RPV (satRPV) RNA requires a DouXble hammerhead for self-cleavage and an alternative structure for replication. *J. Mol. Biol*., **293**, 781–793.
43. Dreher,T.W. and Hall,T.C. (1988) Mutational analysis of the transfer-RNA mimicry of brome mosaic-virus RNA—sequence and structural requirements for aminoacylation and 3′-adenylation. *J. Mol. Biol*., **201**, 41–55.
44. Kanamori,Y. and Nakashima,N. (2001) A tertiary structure model of the internal ribosome entry site (IRES) for methionine-independent initiation of translation. *RNA*, **7**, 266–274.
45. Wuyts,J., De Rijk,P., Van de Peer,Y., Pison,G., Rousseeuw,P. and De Wachter,R. (2000) Comparative analysis of more than 3000 sequences reveals the existence of two pseudoknots in area V4 of eukaryotic small subunit ribosomal RNA. *Nucleic Acids Res*., **28**, 4698–4708.
46. Van de Peer,Y., De Rijk,P., Wuyts,J., Winkelmans,T. and De Wachter,R. (2000) The European small subunit ribosomal RNA database. *Nucleic Acids Res*., **28**, 175–176.
47. Chappell,S.A., Edelman,G.M. and Mauro,V.P. (2000) A 9-nt segment of a cellular mRNA can function as an internal ribosome entry site (IRES) and when present in linked multiple copies greatly enhances IRES activity. *Proc. Natl Acad. Sci. USA*, **97**, 1536–1541.
48. Scheper,G.C., Voorma,H.O. and Thomas,A.A. (1994) Basepairing with 18S ribosomal RNA in internal initiation of translation. *FEBS Lett*., **352**, 271–275.
49. Barends,S., Rudinger-Thirion,J., Florentz,C., Giege,R., Pleij,C.W. and Kraal,B. (2004) tRNA-like structure regulates translation of Brome mosaic virus RNA. *J. Virol*., **78**, 4003–4010.
50. Carter,A.P., Clemons,W.M., Brodersen,D.E., Morgan-Warren,R.J., Wimberly,B.T. and Ramakrishnan,V. (2000) Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature*, **407**, 340–348.
51. Harms,J., Schluenzen,F., Zarivach,R., Bashan,A., Gat,S., Agmon,I., Bartels,H., Franceschi,F. and Yonath,A. (2001) High resolution structure of the large ribosomal subunit from a mesophilic Eubacterium. *Cell*, **107**, 679–688.
52. Ban,N., Nissen,P., Hansen,J., Moore,P.B. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 angstrom resolution. *Science*, **289**, 905–920.

53. Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D'Souza,L.M., Du,Y.S., Feng,B., Lin,N., Madabusi,L.V., Muller,K.M. *et al.* (2002) The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2.
54. Harary,F. and Prins,G. (1959) The number of homeomorphically irreducible trees and other species. *Acta. Math.*, **101**, 141–162.
55. Harary,F. (1969) *Graph Theory*. Addison-Wesley, Reading, Mass.
56. Vasquez,M. and Scheraga,H.A. (1985) Use of buildup and energy-minimization procedures to compute low-energy structures of the backbone of enkephalin. *Biopolymers*, **24**, 1437–1447.
57. Zuker,M., Mathews,D.H. and Turner,D.H. (1999) Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In Barciszewski,J. and Clark,B.F.C. (eds), *RNA Biochemistry and Biotechnology*. Kluwer Academic Publishers, Dordrecht, pp. 11–43.
58. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.