Emerging structural themes in large RNA molecules
Nicholas J Reiter, Clarence W Chan and Alfonso Mondragón

Extensive networks of tertiary interactions give rise to unique, highly organized domain architectures that characterize the three-dimensional structure of large RNA molecules. Formed by stacked layers of a near-planar arrangement of contiguous coaxial helices, large RNA molecules are relatively flat in overall shape. The functional core of these molecules is stabilized by a diverse set of tertiary interaction motifs that often bring together distant regions of conserved nucleotides. Although homologous RNAs from different organisms can be structurally diverse, they adopt a structurally conserved functional core that includes preassembled active and/or substrate binding sites. These findings broaden our understanding of RNA folding and tertiary structure stabilization, illustrating how large, complex RNAs assemble into unique structures to perform recognition and catalysis.

Address
Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208, USA

Corresponding author: Mondragón, Alfonso
(a-mondragon@northwestern.edu)

Introduction
RNAs assume diverse roles in numerous cellular processes, including serving as molecules to store or transfer genetic information, acting as catalysts (ribozymes) to accelerate specific chemical reactions, and functioning as gene-regulatory elements (riboswitches) to direct gene expression. Comparable to protein molecules, many RNAs adopt complex three-dimensional folds in order to function. Recent structures of several large RNA molecules in complex with their natural substrates or ligand cofactors have provided important insights into the mechanisms of RNA folding, RNA-based molecular recognition, and RNA-based enzymatic catalysis.

Early insights into the principles underlying RNA tertiary architecture were derived from the three-dimensional structures of tRNA²̂ [1,2]; naturally occurring small nucleolytic ribozymes, such as the hammerhead [3,4], hepatitis delta virus [5], and hairpin ribozymes [6]; and domains of significantly more complex molecules, such as the group I P4–P6 domain [7] and the RNase P specificity domain [8,9]. Collectively, these structures revealed an architecture dominated by groups of single and coaxial helical stems that are interconnected by a now familiar set of secondary and tertiary structural motifs serving to stabilize their overall fold, such as coaxial helical stacking, tetraloop–tetraloop receptor interactions, pseudoknots, ribose zippers, and dinucleotide platforms (reviewed in [10]). However, subsequent studies have shown that these motifs also act as essential components in the formation of larger, multidomain RNA-based molecules, which prompts the questions: Are there architectural themes that are unique to larger RNAs? Is there a recurrence of structural features that enable these RNAs to recognize their natural targets? Specifically, do large catalytic and regulatory RNAs, like most protein enzymes, contain preassembled active sites and utilize shape complementarity to perform molecular recognition?

In this review, structural studies of large RNA molecules (>100 nucleotides) in complex with their natural targets or ligand cofactors will be examined with a focus on ribonuclease (RNase) P [11] and self-splicing (group I and group II) introns [12,13]. The architectural themes emerging from these molecules are further illustrated by features present in various riboswitch structures [14,15,16,17] (reviewed in [18–20]). We confine our analysis to large RNA molecules that are stable without or with minimal assistance from protein molecules, and hence very large ribonucleoprotein (RNP) complexes, such as the spliceosome, telomerase, and the ribosome, are beyond the scope of this review.

Parallelisms in the tertiary folds
Consistent with observations that small nucleolytic RNAs and domains of large RNAs assume a relatively planar tertiary conformation (reviewed in [10,22]), the recent crystal structures of a bacterial RNase P holozyme and various full-length group I and II introns in complex with their respective molecular targets demonstrate that even large RNAs possess a relatively flat overall fold (Figure 1). The dominant tertiary motif that helps to preserve their relatively planar conformation and to form their core scaffold appears to be coaxial helical stacking. Several stems — typically separated by internal loops and noncanonical structures — pack tightly in a coplanar fashion to form large quasi-two-dimensional
Architecture of RNA molecules. (a) Diagrams of various RNA molecules (>100 nucleotides) illustrating some of their general properties. Two perpendicular views of each structure are shown. Large RNAs are formed by stacked stems that are stabilized by a variety of tertiary interactions. The *Thermoanaerobacter tengcongensis* glmS riboswitch ribozyme structure illustrates the packing of two long stems that form a single, extended layer. The *Thermotoga maritima* lysine riboswitch illustrates an elongated molecule formed by three packed stems, forming a two layer molecule. The two major stems interact with each other through a kissing loop (red/orange). The *Bacillus subtilis* M-box riboswitch also forms parallel stacked stems. More complex molecules have a more globular, but still quasi-planar structure. In some cases, the addition of protein increases their functionality. In the case of the phage Twort group I ribozyme, binding to the homodimeric *N. crassa* CYT-18 protein promotes the splicing of many introns. Note that the protein binds on the surface of the RNA. The *Oceanobacillus iheyensis* group II intron structure represents an remarkable example of an RNA molecule with a complex fold. This large RNA forms a tight, compact structure around the active site region (red). Similar to proteins, large RNA molecules can consist of functional domains. The structure of the RNase P holoenzyme is an excellent example of a ribozyme composed of two functional domains, a catalytic (blue) and a specificity (yellow) domain. In this case, the protein also binds on the surface and does not alter significantly the RNA structure. (b) Diagrams of the structure of the *Thermus thermophilus* 30S ribosome subunit with and without the protein subunits. The RNA moiety (left) is also quasi-planar, despite its large size. In the presence of the proteins (right) the overall shape remains, with the protein molecules mostly found on the surface. The structures (on a clockwise fashion starting from the top left corner) correspond to PDBs: 2Z75, 3DIL, 2Q8Z, 2RKJ, 3IGI, and 3OK7 [11**,13**,14,16,17**,43**,48]. The structures on each panel are drawn at the same scale, the bar on the right corresponds to 100 Å.

Layers that can stack to produce a final structure that is generally only 1–2 helical layers deep in thickness, conferring their characteristic shape. Within these RNA molecules, distinct domains, such as the P4–P6 domain in group I introns [7] and the specificity and catalytic domains in bacterial RNase P [8,9,23], exhibit a quasi-planar, stem-rich fold. This type of domain architecture is prevalent in nearly all large RNA molecules characterized to date, including the RNA of the 30S subunit of the bacterial ribosome. Notable exceptions to this pattern include the large subunits of the bacterial and eukaryotic ribosomes, which contain multiple embedded proteins, forming isotropically arranged multidomain RNP complexes [24,25].
Conservation of the functional core

Comparative sequence analysis of related RNA molecules from different organisms has shown that it is possible to distinguish conserved structural elements despite significant divergence in the corresponding sequences [26–29]. In many instances, covariant base-pair changes in helical stems are easily accommodated at the structural level, even though it has been noted in many RNA molecules that strongly conserved sequences are associated with structurally and functionally important regions. In fact, until recently it was unclear whether RNA molecules of identical function but with significant sequence differences would resemble each other structurally. Initial structural analysis of the specificity domains of two different RNase P molecules [8], followed by comparisons of various different group I introns [30] and of the full-length RNA component of RNase P from two bacteria [23], have shown that the core of these molecules is structurally conserved despite variable peripheral elements. These analyses demonstrate that different combinations of peripheral elements can stabilize a particular conserved core scaffold (Figure 2) and suggest, as previously proposed [8,30], that the evolutionary pressure to preserve the core RNA fold is strong.

Although the generality of this observation is unknown due to the paucity of determined structures of homologous RNA molecules showing significant sequence-level differences, it is possible to anticipate that the core of homologous molecules will likely be conserved even when the sequence variability is high. For example, determining whether different group II introns resemble each other at the catalytic core and share many of the essential structural features will be an interesting test of this hypothesis, particularly because their sequence variability is even higher [31] than that observed in both group I introns and RNase P.

Preassembled functional cores

The structures of RNase P and of the group I and II introns all suggest predominantly preassembled and organized active sites that require metals for catalysis [11**,12,13**,32–34] (Figures 2 and 3). Whereas metal-assisted catalysis is not a universal feature of ribozymes, as there are both small [6] and large [35] ribozymes that operate without direct involvement of metals, preassembled active sites appear to be a common feature. A preassembled active site does not preclude local conformational changes or movements of important functional groups, but instead implies that the main components are already in place and that no large-scale conformational rearrangements are needed to bring them together for activity. This appears to hold true for both small and large RNA molecules with the caveat that defining the correct boundaries of small functional RNAs existing within the context of a larger RNA molecule is critical [3].

The active site environment is typically located within loop regions or at the junction of coaxial helical stacks. These elements are stabilized by RNA tertiary motifs, metal ions, and sometimes accessory protein subunits. Conserved noncanonical base-pairs and/or extrahelical nucleobases are also common features at the active site. Whereas many small endonucleolytic ribozymes utilize the 2'-hydroxyl of an internal ribose for catalysis [36], both RNase P and group I introns use an external metal-activated nucleophile. These are not the only possibilities as group II introns utilize both the 2'-hydroxyl of an internal adenosine and at least two metal ions to catalyze the splicing reaction. Current data suggest that any significant conformational changes associated with catalytic activity are confined to local regions within the active site [37**]. To provide a clearer understanding of their catalytic mechanism and to address active site plasticity in these large ribozymes, additional structures at atomic-resolution and at different states of their catalytic cycle are needed.

Molecular recognition

Comparative analysis of large RNA structures highlights the common theme of shape complementarity and specific interactions in target recognition (Figure 4). Whereas self-cleaving RNA molecules recognize their target cleavage site in cis to perform single-turnover reactions and primarily use specific base-pair interactions to facilitate recognition [36], multiple-turnover ribozymes, like RNase P and the ribosome, recognize their substrate in trans. In the case of bacterial RNase P, substrate recognition primarily occurs via specific base-stacking interactions between bases in highly conserved regions in the specificity domain and the tRNA, in addition to the formation of an A-minor motif [11**]. Base-pairing involving the 3'-CCA tRNA sequence also contributes to recognition, but this is not a universal feature as some mature tRNAs acquire the 3'-CCA residues through a modification following 3' and 5' end-processing [38]. Structural evidence shows that group I introns also utilize extensive tertiary motifs to create precisely shaped binding scaffolds [12]. Not surprisingly, base-pairing and base stacking are widespread in RNA recognition. Although riboswitches recognize small metabolites, they also utilize both specific base-contacts and precise shape complementarity to distinguish exquisitely between closely related molecules. For instance, the lysine and purine riboswitches show at least a 5000-fold level of discrimination between lysine and ornithine and a 10 000-fold level of discrimination between adenine and guanine, respectively [39]. One of the best examples of a selective and preassembled binding site is the large glmS ribozyme-riboswitch [16,40], where the RNA structure does not change upon binding its target glucosamine-6-phosphate molecule. Instead, ligand binding serves to function as the catalytic cofactor in the reaction. Taken together, these large ribozyme and riboswitch structures...
Figure 3

Proposed active sites based on RNase P, group I, and group II intron structures. Whereas these catalytic RNAs require at least two metal ions, putatively magnesium (pink spheres), not all ribozymes contain metal cofactors within their active sites. In these large RNA molecules the overall active site structure appears to be largely preassembled, although correct substrate positioning likely enhances the rigidity of the active site and binding of catalytic metal ions. (a) Structure of the T. maritima RNase P holoenzyme in complex with tRNA [11**] (PDB 3OKB) shows components that comprise the active site scaffold. (b) The Azoarcus group I intron (ribo-I1G) with bound exons [42] (PDB 1ZZN). The group I active site (light blue) contains two A-rich regions and structurally conserved base triples that help to position an external nucleotide (I1G, orange) and exons (red) during cleavage and ligation. The nucleophile and scissile bond for the ligation step are shown in yellow. (c) Structure of the O. iheyensis group II intron shows three metal ions in the region where exon ligation occurs [33,46] (PDB 3EOG). Yellow dashed lines indicate metal-ligand interactions that are ≤3.0 Å for the RNase P and group II structures, and within ≤2.2 Å for the Azoarcus group I structure.

support the notion that RNA conforms to the same principles that are common to all macromolecular recognition processes: shape complementarity and specific atomic interactions.

The role of adenosines

Of the four nucleobases, adenine appears to be preferred in functional cores and junctions. In RNase P, a constellation of conserved adenosines surrounds the active site scaffold and helps to maintain the integrity of the active site environment. Interestingly, adenosine is the nucleotide that is universally conserved with greatest frequency across all RNase P/MRP family members [41]. Other large RNA molecules, such as the group I introns and the glycine riboswitch [12,15**] also contain critical adenosines within loops and junctions. Conserved A-rich regions (J4/5 and J8/7) in group I introns are important for substrate recognition and splice-site selection, serving as a platform during exon ligation, helping to orient catalytic metals, and promoting the formation of stabilizing A-minor and ribose zipper tertiary interactions [30,42]. The glycine riboswitch, which utilizes two RNA domains to bind glycine cooperatively, contains both a cluster of purine nucleotides around its ligand-binding interface and a highly conserved adenosine that extrudes from the binding pocket to intercalate with an A-rich junction, facilitating interdomain tertiary contacts and cooperative ligand binding [15**]. Finally, approximately 43% of all 23S and 16S rRNA loop and interhelical bridging regions are composed of adenosines, corresponding to a sequence preference 18–28% greater than that of C, G, or U nucleotides [10]. Whether this purine base is indeed preferred at loop or interhelical regions in the assembly of higher order, multidomain RNA enzymes remains to be confirmed, and will require the structural elucidation of additional large RNA and RNP complexes.

The role of proteins

Proteins greatly extend the functional versatility of RNA molecules by stabilizing a specific three-dimensional fold,
Recognition of molecular targets by large RNA molecules is based on shape complementarity and specific interactions. (a) *T. tengcongensis* glmS ribozyme-riboswitch [16] (PDB 2Z75) in complex with its ligand cofactor, glucosamine-6-phosphate (red), reveals the importance of specific stabilizing interactions that orient and position the substrate within the cofactor binding pocket. (b) *T. maritima* RNase P holozyme recognition of tRNA\(^{3\prime}\)-CCA sequence (red) [11\(^{**}\)] (PDB 30K8) is primarily achieved by precise positioning of two base-stacks (marked by an *) and an A-minor motif (●) in the specificity domain with the unique, tertiary fold of the tRNA 5\'TC-loop and D-loop (pink panel). The complex is further stabilized by base-pairing interactions between the RNase P catalytic domain and the tRNA 5\'-CCA sequence (blue panel). (c) *Azoarcus* group I intron [42] (PDB 1ZZN) with both exons (red) positioned by base-pairing and base-stacking interactions in a catalytic state before the exon ligation step (pink and blue panels). Three base-stacks (*, green panel) stabilize an important junction (dark purple) in close proximity to the active site. The 5\'-terminal guanosine (1G5, *) nucleophile is both positioned by and embedded within a conserved cluster of stacked base triples (orange panel). In all panels, purple and white spheres represent magnesium atoms and water molecules, respectively.

by making an unstructured RNA region more accessible for its biological function, or by guiding molecular recognition. Group I and II introns and RNase P all contain important accessory protein subunits that assist in the folding and activation of the RNA catalyst. For example, the *Neurospora crassa* CYT-18 protein, a tyrosyl-tRNA synthetase, has adapted to function in splicing by binding and stabilizing some phage group I introns at a similar location as the peripheral group I intron PSabc domain [43**]. Thus it appears that CYT-18 functionally replaces parts of the peripheral fold in some group I introns, consistent with the theme that proteins help to extend the functionality of a ribozyme and were likely later additions in evolutionary time. Similarly, DEAD-box proteins CYT-19 and/or Mss116 help to facilitate RNA folding and specific nucleic acid rearrangements *in vivo* in the group II splicing reaction [44,45]. In addition, it is likely that other maturase proteins form part of the *in vivo* group II RNP enzyme by binding and stabilizing core RNA domains and by guiding the flexible domain VI within the active site [46]. In contrast, the protein of bacterial RNase P is not known to affect significantly the folding or the final conformation of the RNA, but rather, docks on to a conserved RNA surface of the ribozyme and serves to position the precleaved tRNA 5\'-leader region [11**].

The RNase P and group I and II intron protein subunits are not embedded in the RNA structures, but are instead positioned at the surface or periphery, helping to promote a stable and functional RNP complex *in vivo*. This is not the case in more complex ribonucleoprotein complexes. For instance, essential RNAs in the spliceosome are surrounded and stabilized by multiple proteins [47]. Thus, the role of proteins spans a broad spectrum, with bacterial RNase P and group I and II introns whose protein components are not essential *in vitro* near one end, and larger assemblies with multiple essential protein components, like the ribosome and the spliceosome, near the other.

**Conclusions and outlook**

The advent of several large RNA crystal structures, both alone and in complex with their substrates and associated proteins, has permitted analysis of the general properties important for RNA architecture and the fundamental relations between structure and function. Large RNAs are composed mainly of packed stacked stems that are linked together by a variety of tertiary interactions, often forming extensive layers that aggregate into anisotropic molecules. Although there is variability in the structure of homologous RNAs, which is mostly confined to peripheral elements, functional RNA molecules have well-defined structures with preassembled active sites or binding regions. As in all biological molecules, recognition involves specific atomic interactions and shape complementarity, and in the case of RNA, base-pairing and base-stacking play a central role.

To date, most of the known structures correspond to the first functional RNA molecules identified. As more structures become available it is likely that some of these observations will need to be revisited or revised. Additional structures of more noncoding RNA molecules spanning a wider range of functions are needed, along with more examples of homologous RNA structures. This
is an important task for the future if we want to achieve a complete understanding of RNA function and structure. Indeed, many surprises await as we discover even more about RNA and its numerous roles in biology.

Acknowledgements
Work was supported by a National Institutes of Health grant (GM84843) to AM, a NRSA postdoctoral fellowship to NJR, and an NIH Molecular Biophysics Training grant (T32GM008382) fellowship to CWC.

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:
• of special interest
•• of outstanding interest

11. Reiter NJ, Osterman A, Torres-Larios A, Swinger K, Pan T, Mondragon A: Structure of a bacterial ribonuclease P holoenzyme in complex with tRNA. Nature 2010, 468:784-789. This study illustrates how RNase P is preassembled to engage in shape selective recognition of its tRNA target. The two structures support earlier models of substrate recognition and catalysis, identify the location of the active site, and show how the protein component increases RNase P functionality.
13. Toor N, Keating KS, Taylor SD, Pyle AM: Crystal structure of a self-spliced group II intron. Science 2008, 320:77-82. One of the most elaborate RNA-only structures determined to date. This work reveals how multiple RNA helices come together to stabilize the catalytic core of this ribozyme. This study provides a crucial starting point for future structural investigations of group II introns and the spliceosome.

Structures of free and ligand-bound states of one domain of the glycine riboswitch show how a unique RNA junction and metal ions create a highly selectively binding pocket for the smallest amino acid. Structural analysis and footprinting studies provide insight into how the entire, two domain riboswitch forms specific interdomain interactions, thereby facilitating cooperative ligand binding.

An insightful review describing how an RNA molecule can recognize a small molecule and act as a genetic switch.


