

Introduction to RNA

The biophysics of RNA is as complex and multifaceted as the molecules themselves. To introduce RNAs and the biophysical concepts that describe them, this introduction will use a generic RNA structure as a prototypical example of study. This apparent reduction of complexity is possible because RNA molecules are effectively modular: in their folded forms, they contain structural motifs that can be autonomous or part of the tertiary structure. There are several canonical secondary structural motifs that constitute RNA building blocks; the primary sequence of these motifs can vary, but the structural elements are constant. These motifs appear in discussions of RNA molecules, so it is important to recognize them.

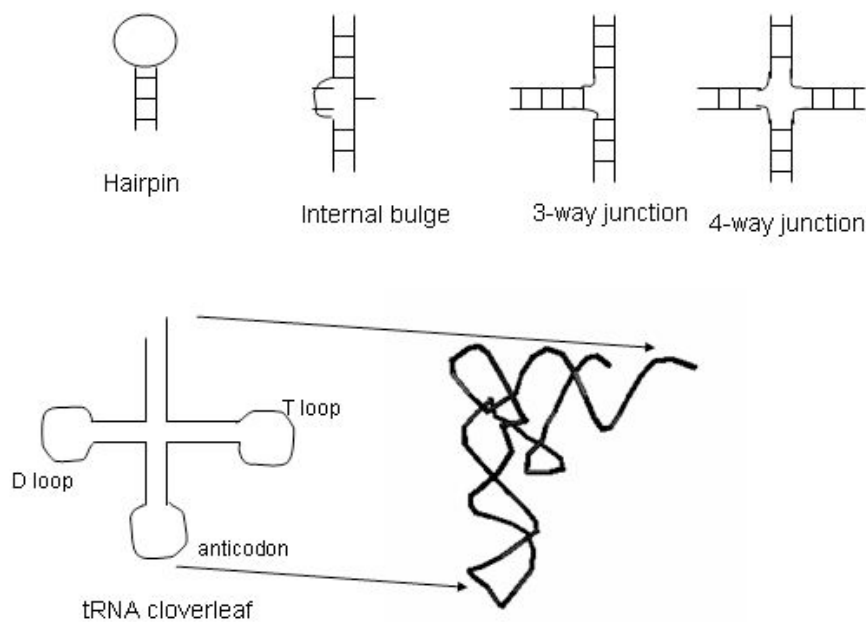


Figure 1. RNA secondary structure motifs. Hairpin, 3-way and 4-way junctions; internal bulge. tRNA cloverleaf secondary structure and tertiary L-fold.

Here, we will use RNA hairpins as examples of a prototypical RNA structural motif, with the acknowledgement that within this broad class of structural elements, there are subclasses that confer unique properties on an RNA molecule. Hairpins are found in virtually

all RNAs: the most common example is probably the 76 nucleotide tRNA's with their three hairpins (the D-stem, T-stem and anticodon stemloops). While it is generally true that an RNA hairpin can be removed from a larger RNA for detailed characterization, a tRNA illustrates an exception: its D-loop and T-loop nucleotides interact (a tertiary interaction) to give the tRNA its L-shape. However, its anticodon stemloop is independent of the rest of the structure, and has been studied on its own as an isolated component. It is this modular property of RNAs that in general makes the study of their component pieces both tractable and biologically relevant, and allows us to use a hairpin to introduce general biophysical concepts.

An RNA hairpin. A hairpin contains a stem and a loop. The stem can be a perfect duplex with Watson-Crick base pairs, or contain noncanonical base pairs such as G:U, A:G, U:C, and others; it might also contain an internal bulge or mismatch. RNA duplexes are A-form (DNA duplexes are typically B-form in solution), and in the context of a long RNA strand, the duplex regions tend to be short (< 6 base pairs). An A-form duplex is thermodynamically more stable than a B-form duplex, thus these short stems maintain their integrity within a longer RNA molecule.

The loop size varies with the RNA, of course, but the smallest loop is restricted by steric constraints to be three nucleotides (a triloop). Large loops, greater than 15 nucleotides, are rarely encountered as independent entities, simply because possible base pairing combinations in a large loop will give it a structure or provide a site for tertiary interactions. The sarcin-ricin loop in 23S and 28S rRNA is an example of a 17 nucleotide loop that has a noncanonical structure (Seggerson & Moore, 1998), and is a regulatory site within the ribosome. Unlike stems, (most) loops do not offer a thermodynamic advantage to the RNA despite their ubiquity as a structural element; instead, they function as binding sites for small ligands, proteins, ions, and other RNAs.

RNA stability. To estimate the thermodynamic stability of a hairpin, several computational methods make use of the large database of RNA thermodynamic parameters. Two of the most popular platforms for predicting the secondary structures of RNA are *mfold*

[mfold.bioinfo.rpi.edu/cgi-bin/rna-form1.cgi] (Zuker, 2003; Mathews et al., 1999) and Vienna RNA (www.tbi.univie.ac.at/~ivo/RNA/). Even before we examine the biophysics of thermodynamics parameters of RNA, we can run this software as a black box to begin to think about RNA folding.

For this in-class **Exercise 1**, we use the RNA sequence 5'GCCAGCUACGCAAUGGCUGGC3' and the *mfold* platform.

FIRST: using paper and pen, predict the secondary structure of this sequence.

THEN: Type in the RNA sequence and submit it directly to the server; for this short sequence, the results will be returned within minutes. Follow the directions in *mfold* which include a query for the number of alternative folds to be presented ('percent suboptimality'). The default value is 5, but prudence dictates at least 25 should be entered (Why?).

Interpretation of these predictions requires knowledge of the origins of the thermodynamic parameters upon which they are based. These calculated free energies come from experimental data of short duplex RNA melting in 1 M NaCl with an ad hoc addition of a penalty for adding a loop of N-nucleotides (Mathews et al., 2004). While the duplex data are very robust, the loop penalties are not; a simple calculation of sequence complexity will show why this data disparity arises and that it is unlikely to be remedied.

RNA sequences can often adopt several structures. These "alternative conformations" reflect the ability of an RNA strand to fold into 'non-native' structures as a consequence of nearest nucleotide neighbors, temperature, salt conditions, specific ion binding, transcription rates, and concentration. By defining a native fold as the active conformation of the RNA (using whatever assay is appropriate), then non-native folds could be off-pathway dead-ends, or intermediates in the folding process. The implicit assumption in these calculations is that the lowest free energy structure is the functional structure, but there are many circumstances in which that assumption is dubious.

Exercise 2. Make a mutation in the RNA, predict the structure yourself, then re-run the prediction. What mutation(s) result in alternative structures, and how would you explain their effects?

The structures that are predicted by mfold, or by your imagination, are secondary structures. Their three dimensional structure is known in part, since the stem is certainly an A-form duplex, but what is the structure of the loop? Here there is no guide to assist you – the structures of loops cannot be predicted. We can make some guesses, however, based on known structures and some fundamental properties of nucleic acids in solution. What properties of ribonucleotides are of use for these guesses? This loop contains A,C and G nucleotides; purine bases will stack on each other, as will cytosine; uridine will not since it is polar and hydrophilic. None of the bases in our loop will ‘want’ to be extruded into solution. Assuming that the loop bases stack onto the loop-closing base pair at the stem/loop junction, then the A-form duplex structure will be maintained until there is an inevitable need to make a turn. Where will this turn take place? And what portion of the nucleotide will accommodate this turn? The ribose pucker? The phosphate backbone? Will there be only one structure of the loop, or will there be several that interconvert (that is, their folding free energies will be similar and the energy barriers low)? Will the loop structure be rigid or flexible? If the loop structure is flexible, what is the timescale of motions? This apparently simple hairpin is in reality a complex structure!

Exercise 3: Compare the structures of three loops: the UUCG tetraloop (1F7Y), the hepatitis B virus encapsidation signal (2IXY), and the conserved stemloop from E coli SRP (28SR). Are these loop structures static or flexible? Is there any common feature shared by all? How are the structures dependent on the sequence?

The predicted secondary structures have folding free energies assigned to them, based on experimental data. These Gibbs free energies (ΔG°) predict the stability of the structure at 37° C, but of course there are no data to account for any structures in the loop that would either stabilize or destabilize the hairpin. To test the prediction requires experiments, typically a thermal unfolding experiment (a ‘melting’ experiment), measuring UV absorption at 260 and 280 nm as a function of temperature. These measurements are designed to determine the contribution of the RNA loop to the energetics of unfolding, and thus obtain the ‘loop penalty’

(e.g. Mathews et al., 1999), but also to examine the transition to determine whether it is two-state. If bases/ribose in a loop make no hydrogen bonding interactions, then loop structure will be controlled by base stacking interactions, which will propagate from the loop-closing base pair. Base stacking is not sensitive to salt concentration (and it is not a hydrophobic interaction), but it does decrease with increasing temperature, so the absorbance change from loop unstacking will give a sloping baseline at the start of the melt. If there is inter-nucleotide hydrogen bonding within the loop, then there could be two separable unfolding transitions, with the expectation that loop structure melts out at lower temperature.

RNA unfolding experiments are an important means to probe the structure of both simple RNAs, such as the hairpin, but also complex structures such as the Tetrahymena Group I intron. Like their protein folding analogues, these experiments measure unimolecular transitions: Native ↔ Unfolding transitions can be measured as a function of temperature, urea concentration, and in the presence and absence of divalent ions, but the results should not be dependent on the RNA concentration. That is, the experimental concentration of the RNA should not affect the parameters of the melting curve; if there is a concentration dependence, then higher order complexes have formed (dimers are the typical off-pathway complex formed from hairpins). The melting curves obtained from a temperature / absorbance experiment provides the melting temperature $T_{m,i}$ (melting temperature of transition i) and ΔH_i (van't Hoff enthalpy) (Xia et al., 1998). The melting temperature is defined as the midpoint of the transition, and it depends on the solution conditions. Free energies of unfolding are typically extrapolated to 25° C (37° C) using the equation $\Delta G^\circ = \Delta H^\circ(1 - 298(310)/T_M)$.

Nature uses RNA “thermometers” to regulate gene expression in prokaryotes.

RNA and Ions. RNA has a complicated relationship with ions. As a polyanion, it has a large net negative charge from its phosphates, which if unshielded would repel each other. An RNA in water without neutralizing counterions would indeed appear as an extended strand. An

RNA that needs to fold, such as a riboswitch, ribosomal RNA, tRNA, or ribozymes, must rely on counterions to pack locally high densities of negatively charged phosphates. In a cell, K^+ ions are the dominant monovalent ions, typically estimated at 0.15 M. Other monovalent ions, particularly NH_4^+ and Na^+ , are of interest for RNA folding due to their different ionic radii and numbers of bound waters. For historical reasons, data for thermodynamic characterizations were collected in NaCl, but it can be important to realize that these are not the common ions in vivo.

While monovalent ions make the dominant contribution to facilitation of RNA folding, it is the divalent ions, particularly Mg^{2+} , that are critical to the formation of many unique RNA structures. For our RNA hairpin, the addition of 1 mM $MgCl_2$ to 100 mM KCl will raise the ionic strength slightly, but because the divalent ion is able to associate preferentially with the RNA phosphates (Grilley et al., 2007), it will raise the RNA melting temperature. The complication comes from the possibility that there is a specific interaction of the loop structure with Mg^{2+} . The concept of specific site binding of divalent ions to an RNA began with the crystal structures of tRNA (for example, yeast tRNA^{PHE} from Westhof & Sundaralingam, 1986). In those structures, both Mg^{2+} ions and spermine molecules (a tetra-amine with a +4 charge) were found at specific loci in the folded tRNA, and together with tRNA folding experiments with and without Mg^{2+} ions, led to the idea that RNA folding was facilitated by tightly bound divalent ions. While there is a wealth of experiments that demonstrate that RNA does indeed use Mg^{2+} ions for folding, the idea that there are specific binding sites in a given RNA must be very carefully applied.

Exercise 4: Review the tRNA x-ray crystal structures (1TRA) and examine the placement of Mg^{2+} ions. What interactions with the RNA are described? What role do the crystallographic waters play?

RNA Dynamics. While A-form duplexes are more rigid than B-form duplexes, much of an RNA strand is found in loops, single strands, internal bulges, and junctions. These regions could be

tightly structured (perhaps facilitated by divalent ions), or they might be flexible to anticipate their function (such as a riboswitch binding to its ligand). The motions of an RNA molecule therefore are often intrinsic to their function, and are part of the characterization of any RNA.

What properties of an RNA contribute to its motions, and how do solution conditions affect its dynamics? In the example of the hairpin, its loop could very well have complex motions as the bases stack and unstack at some rate or the ribose puckers fluctuate. Base stacking is an enthalpy-driven process that depends on the particular nucleobase and temperature. It is not a hydrophobic interaction. Stacked nucleotides are not static; the simplest example of their transient association is seen in NMR experiments that observe the imino protons of a base pair. The terminal base pair of a duplex at room temperature and physiological salt (0.1 M KCl or NaCl) always has a weak imino proton intensity, indicating that it is in exchange with solvent water. Since exchange demands that the base pair opens, these data show that stacking and hydrogen bonding allow frequent excursions of both bases away from their A-form duplex structure. While base pairs within a longer duplex have reduced dynamic motions, they do break and reform with a frequency dependent on sequence, temperature, and salt.

The origins and timescales of RNA motions in hairpins have been monitored by spectroscopic methods and molecular dynamics simulations, and this research is still ongoing.

RNA:RNA interactions. Beyond the obvious formation of RNA duplexes, RNAs interact with each other in tertiary interactions that are very difficult to predict and anticipate. Prediction of RNA tertiary interactions remains one of the major challenges for understanding RNA folding and ultimately design of novel RNAs.

RNA:RNA complexes are critically important in biology. A most striking example is that of RNA interference [RNAi] and microRNAs [miRNA]. In these complexes of short RNAs that base pair to complementary sites on mRNA, perfect complementarity leads to mRNA degradation in RNAi, while pairing of miRNAs to mRNA (typically in the 3'UTR) leads to

Exercise 5. Select one of the above examples (or find your own) and report on its biological role in regulation.

translation inhibition. Other examples abound, including the pairing of U4:U6 snRNAs in the spliceosome, tRNA anticodon:mRNA codon pairing in the ribosome; guide RNAs pairing to mRNA targets for RNA editing; snoRNA recognition of sites for modification; E. coli RhyB sRNA; etc etc. These interactions are often transient within the context of the biological process they mediate, and they rely on structure and thermodynamics (and proteins) to modulate their lifetimes.

Not all RNA:RNA interactions rely on extensive base pairing. Some use noncanonical interactions between bases and riboses to act as staples for large RNAs, and specific examples are found in many different RNAs. One such tertiary interaction is the GAAA:receptor complex, others are the U-turn and kink-turn; and pseudoknots are common. These interactions are typically difficult to predict based on sequence alone, but they constitute building blocks of larger RNA structures.

The GAAA:receptor complex is a model system to study the thermodynamics and kinetics of tertiary interactions by ensemble and single molecule methods.

RNA:Ligand interactions. RNA molecules specifically recognize small molecules as part of their regulatory functions in cells, and these properties can be exploited to develop artificial probes of cellular processes as well as potentially to develop new therapeutics. The malleability of RNA and its ability to adapt itself to complex surfaces led to SELEX experiments; in vitro selection experiments using random sequence pools of transcribed RNAs that were incubated with target molecules (proteins or small molecules) to identify RNAs that would specifically bind. These experiments produced a number of RNA “aptamers” selected to bind to a his-tag, to biotin, BSA, cocaine, thrombin, yeast prion proteins, oligosaccharides, theophylline, etc etc.

Nature had already put this property of RNA to use, of course. Long before there was SELEX, there were riboswitches in bacteria, archea, fungi, and plants. Discovered by Breaker in 2002, these RNA elements bind to the small molecule ligand that is the product of the genes they control: s-adenyl methionine (SAM), thiamine pyrophosphate (TPP), guanine, adenine,

lysine, queosine, flavin mononucleotide (FMN), and more to come. Riboswitches are typically located in the 5' untranslated region (UTR) of a gene, and are more complex than the in vitro aptamers, since they not only need to bind the ligand, but then to transmit the signal to downstream RNA to terminate transcription or block translation. The regulatory mechanisms of riboswitches are not clear, but are presumed to be kinetically controlled by rates of transcription and binding of ligand.

Ribosomal RNA also binds to small molecules. Many antibiotics target the ribosome, where one target is the peptidyl transferase center to inhibit the enzyme. These interactions are seen in crystal structures of ribosomes with bound carbomycin A, spiramycin, tylosin, sparsomycin, puromycin, chloramphenicol that reveal intricate networks of hydrogen bonds and packing of the antibiotic with the RNA bases, riboses, and phosphates. The difficulties inherent in de novo design of drugs that bind to bacterial ribosomes are easily appreciated from these co-crystals.

Exercise 6: The HIV trans-activating region (TAR) RNA is the target of the HIV tat protein for transcriptional regulation of the genome. The entire HIV protein can be replaced by an argininamide molecule that mimics its binding. Examine the structure of this R-TAR complex (1AJU), and identify the sequence-specific interactions.

Riboswitches will be beautiful model systems to study the properties of specific recognition of molecules by RNA.

RNA Folding. RNA is transcribed as a single strand, and it must adopt secondary and tertiary structures in order to become biologically active. A hairpin is the simplest of structures because it is a local structure: it is easy to imagine it forming as transcription progresses. Now imagine how a 2400 nucleotide ribosomal RNA must fold, with all its domains and inter-connections and with all its proteins bound. This story of folding and assembly is still on-going. The 414 nucleotide Tetrahymena Group I intron is another folding story, and one that has been studied by many methods and in many conditions.

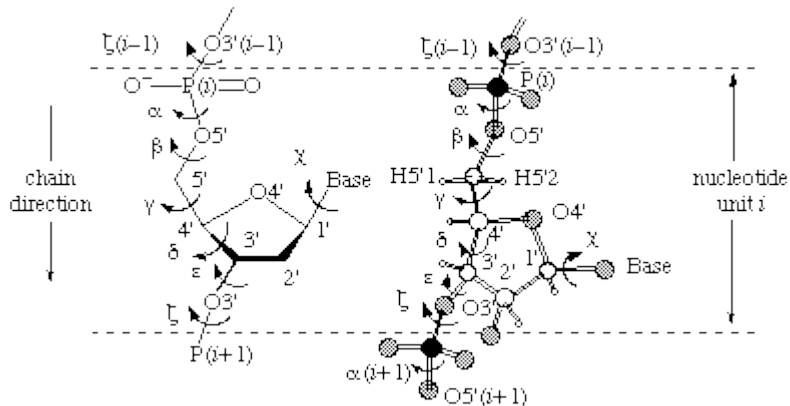
In order to fold into a compact tertiary structure, any RNA must overcome several problems. The first is simply electrostatic – how to shield its negatively charged phosphates in order to collapse the structure. Here, ions and proteins are often both required. A second problem is avoiding local minima – folded structures that are not active can be energetically favorable with high barriers to escape. This is a thermodynamic problem that has been characterized by transcribing the RNA, purifying it under denaturing conditions, then refolding it in experimentally controlled conditions. These studies have identified folding pathways and intermediates, and established rates of interconversion between structures. From these experiments, we gain an appreciation of the folding landscape, the driving forces for folding, and the equilibrium distribution of structures.

The next generation of experiments will look at co-transcriptional folding. This is the kinetic problem: How does the folding of an RNA depend on the rate of synthesis and interactions with RNA polymerase and associated proteins? This challenge will likely be met by single molecule experiments that regulate polymerase at the same time as the nascent RNA strand is emerging and monitored. This is the future.

RNA enzymes. In addition to its regulatory roles in cells, RNA molecules are also enzymes. The discovery of the Group I intron made it famous and earned Tom Cech a Nobel Prize, but other RNA enzymes (ribozymes) include the ribosome, the spliceosome, the hammerhead, the hairpin, Group II introns, RNase P, and others to be discovered. The active sites of many ribozymes are designed to do transesterification reactions, which require precise alignment of attacking and leaving groups, and in some cases divalent ions also participate. The chemistry of a ribozyme can depend on local pKa effects that protonate a base, on backbone flexibility to position the phosphate oxygens to act as nucleophiles, and on ribose puckering to present the 2' hydroxyl for attack.

Introductory RNA concepts to explore.

1. Different components of nucleotide can confer structural flexibility and/or constraints on any sequence. In contrast to only two degrees of freedom [torsion (ϕ, ψ) angles] in a peptide backbone, a nucleotide has seven. There are five backbone dihedral angles (ϕ, ψ, χ, γ and δ), a sugar pucker dihedral (ϵ) and the glycosidic dihedral angle (ζ).



Which of these dihedral angles are constrained in a) an RNA duplex? b) a single stranded RNA? c) an RNA loop?

2. Thermodynamic measurements of RNA loops are constrained by the sheer number of sequences to be considered. For examples, how many sequences are possible in a tetraloop, assuming that only the four canonical bases are present? Since there are 4 sites and 4 possible bases per site, then there are $(4 \text{ bases at site 1}) \times (4 \text{ bases at site 2}) \times (4 \text{ bases at site 3}) \times (4 \text{ bases at site 4}) = 256$ possible tetraloops! Generalize this calculation as a combinatorial problem: there are only four bases, but the number of sites (*n*) can be any number to give 4^n as the general formula. If necessary, convince yourself that this is the correct algorithm, keeping in mind that the sequence *XY* is not equivalent to *YX* (use *n*=2 to save time and effort).

3. Structural elements such as a hairpin can be removed from a larger RNA and studied as autonomous units. What is the analogous smallest structural element that can be removed from a protein?