Chapter 10-RNA Secondary Structure - The Hidden Architects of Cellular Function

Reza Rezazadegan
Shiraz University
www.dreamintelligent.com

Learning Outcomes

After completing this chapter, you should be able to:

- Explain how RNA secondary structure forms from base-pairing interactions and describe common structural motifs, including pseudoknots.
- Describe major functional classes of RNA—cellular and viral—and relate their functions to their structures.
- Understand the thermodynamic principles behind RNA folding, including minimum free energy, motif-based energy models, and ensemble descriptions.
- Outline how computational methods predict RNA structure, including dynamic programming, pseudoknot-capable algorithms, integration of chemical probing data, and modern deep-learning approaches.
- Describe how RNA sequence space can be viewed as an (n)-dimensional Hamming cube and how secondary structures can be represented as dot-bracket strings, trees, or arc diagrams.
- Explain the mapping from sequence space to structural space, the concept of RNA neutral networks, and how neutrality contributes to robustness and evolvability.
- Summarize the key forces and interactions involved in RNA tertiary folding and the main limitations of current prediction methods.

RNA Structure, Folding, and Functional Roles

RNA molecules are central to gene expression, regulation, catalysis, and viral replication. Their ability to fold into specific structures underlies nearly all these functions. While proteins rely on a broad set of chemical interactions to reach complex 3D shapes, RNA structure is built primarily from base pairing, stacking interactions, and modular motifs. Despite this relative simplicity, RNA structural space is vast and supports remarkable biological versatility.

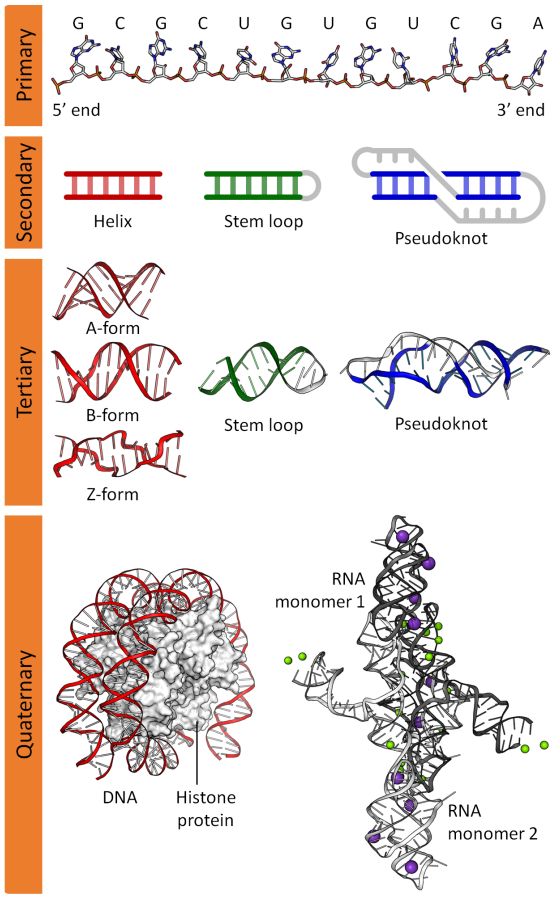
This chapter provides a comprehensive overview of RNA functions, thermodynamics of folding, secondary and tertiary structural motifs, modern structure-prediction strategies (including deep learning and chemical probing), pseudoknot-capable algorithms, and evolutionary concepts such as neutral networks and the mapping from sequence space to structural space. Definitions of primary and tertiary structure are only briefly recalled, as they were covered in the protein chapter.

1. Overview of RNA Structure and Function

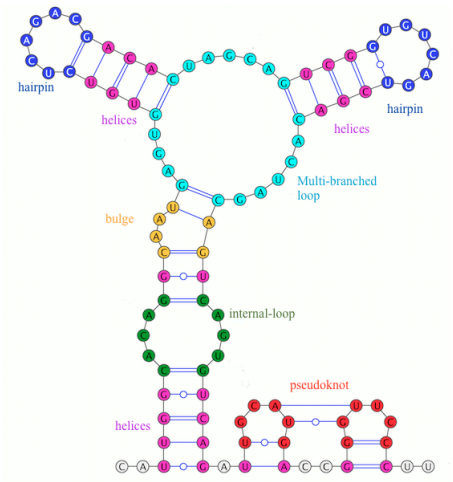
1.1 Primary to Tertiary Structure

RNA primary structure is a linear sequence of nucleotides (A, U, G, C). Secondary structure refers to the set of intramolecular base pairs that form motifs such as hairpins, internal loops, bulges, and junctions. Tertiary structure arises from longer-range interactions—coaxial stacking, pseudoknots, ribose zippers, and metal-ion mediated contacts—that bring distant elements into spatial proximity.

Although chemically simpler than proteins, RNA can form highly elaborate architectures: ribozymes, riboswitches, tRNA, the ribosome, viral genomes, structured UTRs, and long noncoding RNAs all achieve function through their folded shapes.



Similar to proteins, RNA has a hierarchical structure: from primary to tertiary and sometimes quaternary



Different motifs in RNA secondary structure

1.2 Biological Roles of Structured RNA

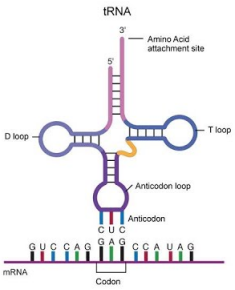
Messenger RNA (mRNA)

Beyond encoding proteins, mRNAs contain structured elements that regulate translation, localization, and stability. For example:

- hairpins near the start codon can modulate initiation
- structured 5′ or 3′ UTR elements bind proteins or small RNAs
- long-range interactions influence circularization and translation control

Transfer RNA (tRNA)

An archetype of RNA architecture: a cloverleaf secondary structure folds into an L-shaped 3D form stabilized by tertiary interactions. tRNAs demonstrate the principle that stable secondary structure is a scaffold for precise tertiary folding.

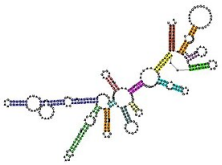


Ribosomal RNA (rRNA)

Forms the catalytic core of the ribosome. rRNA structure is a mixture of:

- highly conserved helices
- complex junctions
- long-range tertiary interactions
- metal-ion stabilisation

rRNA folding shows how RNA can achieve large, hierarchical assemblies.



Regulatory RNAs

Small RNAs serve as the "scouts" of the cellular world, utilizing their compact, simple structures to navigate complex molecular landscapes. By acting as high-precision guides, classes such as **microRNAs (miRNAs)**, **small interfering RNAs (siRNAs)**, and **Piwi-interacting RNAs (piRNAs)** enable cells to identify and regulate specific genetic targets.

The efficiency of these molecules lies in their ability to bind with specific protein complexes—like the RNA-induced silencing complex (RISC)—and lead them directly to a matching messenger RNA (mRNA) or DNA sequence. Ultimately, these small RNAs function as essential regulators that ensure proteins reach their correct destination to silence genes or modify cellular activity.

Catalytic RNAs (Ribozymes)

Ribozymes shatter the old "DNA makes RNA makes Protein" dogma by proving that RNA can act as both a genetic blueprint and a functional engine. Unlike proteins, which use a diverse set of 20 amino acids, ribozymes are limited to just four bases. To compensate, they rely on **noncanonical base pairing** (such as G-U wobbles) to fold into intricate 3D shapes.

These folds often create a specialized "pocket" for **metal ions** (like Mg^{2+}). These ions are crucial because they neutralize the negatively charged RNA backbone and directly participate in the chemistry of the reaction, such as stabilizing the transition state during **self-cleavage** or the **peptide bond formation** catalyzed by the ribosome—which is, at its heart, a massive ribozyme.

Riboswitches

a riboswitch is a regulatory segment of a messenger RNA molecule that binds a small molecule, resulting in a change in production of the proteins encoded by the mRNA. Thus, an mRNA that contains a riboswitch is directly involved in regulating its own activity, in response to the concentrations of its effector molecule.

Found primarily in the 5' untranslated regions (UTRs) of mRNA, they consist of two functional parts: an **aptamer domain** (the sensor) and an **expression platform** (the switch).

When a specific metabolite—such as a vitamin, amino acid, or metal ion—reaches a certain concentration, it binds directly to the aptamer. This binding event triggers a dramatic **conformational change** in the expression platform. This "switch" can physically hide a ribosome-binding site to *stop translation or create a termination signal to halt transcription*, allowing the cell to respond to its environment in real-time without needing a middleman protein.

Long Noncoding RNAs (lncRNAs)

While ribozymes and riboswitches are often defined by rigid, global folds, **lncRNAs** (transcripts longer than 200 nucleotides) are more like "molecular Swiss Army knives." They are frequently composed of **modular domains**, where only specific "islands" of the RNA are highly structured, while the rest remains flexible or disordered.

These structured local regions serve as precise docking stations. A single lncRNA can act as a **scaffold**, bringing multiple proteins together to form a complex, or as a **decoy**, soaking up microRNAs like a sponge. Even if the overall sequence of an lncRNA evolves rapidly and appears poorly conserved, these small, functional structural motifs are often maintained by evolution because they are essential for "guiding" epigenetic modifiers to specific spots on the genome.

1.3 Viral RNA Structure

Many viruses use RNA genomes or RNA intermediates. Viral RNAs rely heavily on structure:

- **Structured UTRs** regulate replication and interaction with host factors
- **Replication signals** often rely on long-range pairing and pseudoknots

Viral RNAs frequently adopt multiple conformations (structural switching), enabling compact genomes to encode diverse regulatory functions.

---

2. Secondary Structure Motifs

2.1 Differences between Protein and RNA secondary structure

- Instead of 20 amino acids we have only 4 nucleotides: A, U, C and G.
- Bonds between bases are given by the Watson-Crick (A-U, C-G) and the non-canonical (G-U) base pairs.
- Nucleotides are not flat and thus we do not have the equivalent of the torsion angles ϕ, ψ for RNA.
- Unlike proteins, the energy of an RNA fold is additive with respect to the structure motifs. (Next section.)

2.2 RNA Secondary Structure Motifs

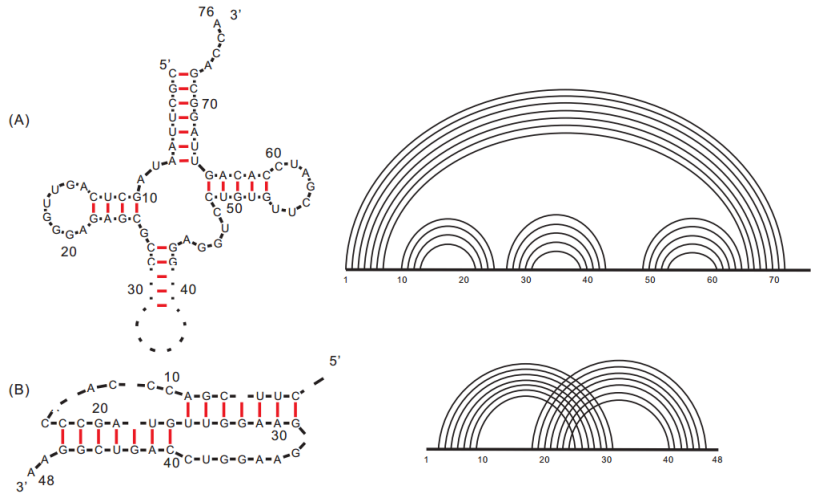
RNA motifs differ significantly from proteins.

- **RNA Base Pairing**: Primarily involves **canonical Watson-Crick base pairing** (A-U and G-C).
- **Non-canonical Pairing**: **G-U base pairing** is also common, although it is less stable than Watson-Crick pairs and usually occurs within double-strand helices surrounded by canonical pairs.

- **Double Helix (Stem)**: Perfectly base-paired regions.
- **Hairpin Loop (Stem-loop)**: The most common motif; a stem closed by a loop of unpaired residues.
- **Bulge Loop**: Unpaired residues on **one side** of a helix, introducing kinks.
- **Interior Loop**: Unpaired residues on **both sides** of a helix.
- **Multibranch Loop (Junction)**: A central loop connecting three or more helices; vital for complex $3\mathrm{D}$ folding.
- **Pseuoknots:** *Base pairs that cross in sequence order.* Pseudoknots are crucial for ribosomal frameshifting, telomerase RNA, viral replication elements, and many ribozymes.

2.3. Arc Diagrams for Representing RNA secondary structure

RNA secondary structures can be represented by arc diagrams by flattening the backbone.



Two RNA structures with and without crossings and their corresponding arc diagrams. Source: Huang, Reidys, and Rezazadegan, Fatgraph models of RNA structure

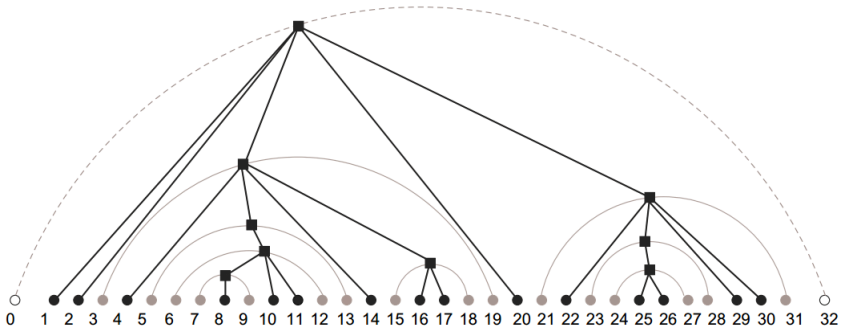
2.4 Dot-Bracket and Tree representations

They exist for non-crossing structures only.

Dot-bracket representation (example)

```
(((..((....))..)))
```

Tree representation



The tree corresponding to an arc diagram. The hypothetical over-arc is called *rainbow arc*. Source: Chen, Reidys, Waterman, "RNA Secondary Structures with Given Motif Specification: Combinatorics and Algorithms"

3. Thermodynamics of RNA Folding

RNA folding is guided by base-pairing and stacking interactions. The free energy of a structure is commonly approximated as a sum of local motif contributions.

RNA folding is a classic energy minimization problem:

- **Minimum Free Energy (**MFE**)**: The classical assumption is that the native structure is the most thermodynamically stable conformation, possessing the lowest total free energy (ΔG_{total}). This energy is calculated by summing the contributions (ΔG) of individual motifs (Turner rules).
- **Free-Energy Ensemble**: In reality, RNA *samples an* ***ensemble*** *of thermodynamically accessible structures*. The probability of any structure S is given by the Boltzmann distribution. This approach, used by tools like **RNAfold**, provides a more realistic view, offering base-pairing probabilities and suboptimal structures.

3.1 Minimum Free Energy (MFE) Principles

Most classical algorithms assume that an RNA adopts the **minimum free-energy structure**, defined by:

$$\Delta G_{\text{total}} = \sum_i \Delta G(\text{motif}_i)$$

Energy parameters $\Delta G(\text{motif}_i)$, called Turner rules, are derived from extensive thermodynamic measurements and updated periodically.

In this framework, predicting the secondary structure of an RNA sequence involves finding the structure with minimum free energy (MFE) among all the structures that are compatible with the sequence. Since for each sequence there are a lot of compatible structures, this task can be performed using Dynamic Programming (DP) *for non-crossing structures (without pseudoknots)*. For general structures with pseudoknots there is not a recursion that can be used to obtain a DP algorithm.

3.2 Free-Energy Landscapes and Ensembles

RNA rarely exists in a single conformation. Instead, it samples a **Boltzmann ensemble**, with probability:

$$P(S) = \frac{\exp(-\Delta G(S)/RT)}{Z}$$

where T is the temperature, R is the Boltzmann constant and

$$Z = \sum_S \exp(-\Delta G(S)/RT)$$

is the **partition function**. Tools such as **ViennaRNA** compute:

- base-pairing probabilities
- centroid structures
- MEA (maximum expected accuracy) structures
- Boltzmann samples

This provides a more realistic view than a single MFE structure.

4. Modern Computational Prediction of RNA Structure

4.1 Dynamic Programming for Pseudoknot-Free Structures

Classical algorithms such as ViennaRNA compute:

- MFE structures
- partition functions
- suboptimal structures

They assume non-crossing base pairs.

4.2 Pseudoknot Prediction

Because *pseudoknots break the noncrossing assumption*, they require special methods:

- heuristic (HotKnots)
- integer programming (IPknot)
- stochastic search (Knotty)
- coarse-grained folding

These methods are slower and often produce approximate solutions, but continue to improve.

4.3 Chemical-Probing–Guided Methods

Experimental constraints (SHAPE, DMS, MaP-seq) provide reactivities that correlate with flexibility.

Input includes:

- high reactivity → likely unpaired

- low reactivity → likely paired or structured

Programs (RNAstructure, ViennaRNA, ShapeKnots) integrate reactivities to produce more accurate models.

4.4 Deep-Learning Approaches

Recent neural-network models can outperform MFE alone, especially when paired with base-pair probability inference.

Examples of modern architectures:

- end-to-end folding networks
- 2D contact-matrix prediction
- transformer-based RNA embeddings
- convolutional architectures leveraging sequence locality

Examples include UFold, SPOT-RNA, and E2Efold. These methods learn pairing patterns directly from known structures and can model tertiary-informed constraints implicitly.

---

5. RNA Structure in Evolution: Neutral Networks

RNA is the classic biological example of high-dimensional genotype–phenotype maps.

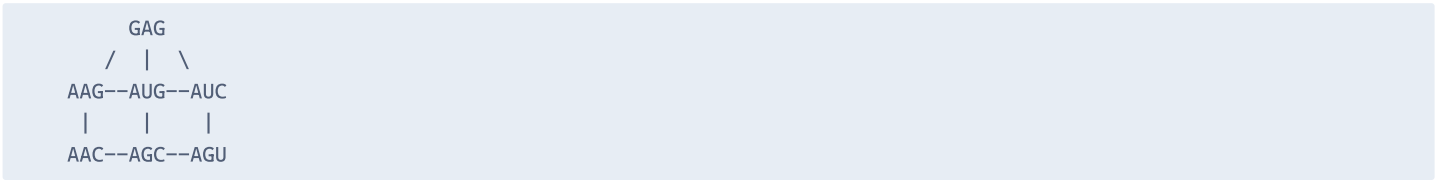
5.1 Sequence Space as a Hamming Cube

For an RNA of length (n), the space of all possible sequences is:

- dimension: (n)
- alphabet: {A, U, G, C}
- each sequence is a node
- edges connect sequences that differ by exactly one mutation

This forms a **4-ary Hamming cube**: $\{A, U, G, C\}^n$. It is a graph whose nodes are the RNA sequences of length n and two sequences are connected by an edge if they differ by only one mutation.

Small schematic (n=3):



The true shape is an (n)-dimensional hypercube.

5.2 Structural Space

Secondary structures can be represented as:

- dot-bracket strings
- trees
- arc diagrams

Each structure corresponds to many sequences that fold into it.

5.3 The Genotype–Phenotype Map: Sequence → Structure

If $\mathcal{S}$ is the set of secondary structures then the **folding map**:

$$F : \{A, U, G, C\}^n \longrightarrow \mathcal{S}$$

is many-to-one. Vast regions of sequence space map to the same structure. In other words $F^{-1}(S)$ for a secondary structure can be very large. It is a subgraph of the Hamming cube of dimension n and is called the **neutral network** of the structure S.

Consequences:

- many sequences share one structure.
- structures occupy “volumes” in sequence space.
- sequences for a given structure form connected **neutral networks**.

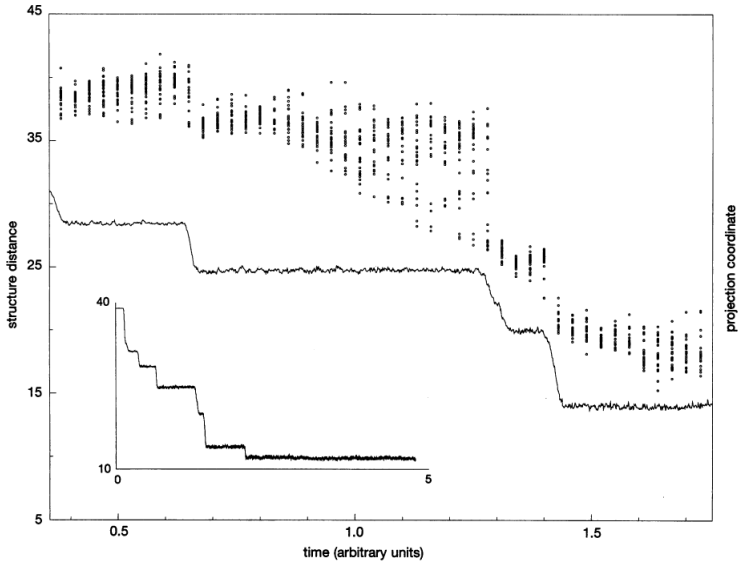
These networks are extensive and often percolate through sequence space, *allowing populations to mutate without losing function.*

5.4 Neutrality, Robustness, and Evolvability

RNA neutral networks explain key evolutionary properties:

High mutational robustness

Many mutations do not change the structure because nearby sequences in the hypercube often map to the same fold.



The structure-distance of a mutating population of RNA sequences to a target sequence over time. The plateaus correspond to neutral networks. Source: Huynen, et al. Smoothness within ruggedness: The role of neutrality
in adaptation

Evolvability through neutrality

Populations can drift on a neutral network and encounter “portals” leading to new structural phenotypes.

Landscape topology

Instead of isolated fitness peaks, RNA structure space contains:

- plateaus (neutral regions)
- ridges (shared motifs)
- basins (stable structures)

This explains how RNA molecules explore sequence space efficiently.

5.2.2. Related Research: Fold Switching

Recent research has focused on **RNA fold switching**, where a sequence can switch between two distinct, stable functional structures in response to an environmental trigger (e.g., temperature or ligand binding). This phenomenon leverages the properties of neutral networks where two different functional structures are encoded by the same sequence and represent two deep, distinct basins on the folding energy landscape. This structural plasticity is crucial for **riboswitches** and many **viral regulatory elements**.

---

6. RNA Tertiary Structure and its Prediction

Tertiary folding involves:

- long-range base pairing
- coaxial stacking
- A-minor interactions
- ribose zippers
- metal-ion binding
- pseudoknot-dependent architecture

Predicting tertiary structure computationally remains difficult. Modern approaches include:

- fragment assembly
- coarse-grained molecular dynamics
- cryo-EM guided modeling
- deep-learning embeddings coupled to 3D reconstruction
- integrated chemical-probing + modeling workflows

Accuracy is improving rapidly, especially where high-quality experimental constraints are available.

---

7. Summary

RNA structure, though built from a small set of chemical interactions, is remarkably versatile. Classical thermodynamic models describe the folding landscape well enough for small RNAs, while probabilistic ensembles and chemical probing have significantly improved accuracy. Deep-learning approaches now complement thermodynamic methods, providing new capabilities in base-pair prediction and tertiary structure inference.

From an evolutionary standpoint, RNA provides a rich example of genotype–phenotype relationships. The immense sequence hypercube collapses through folding into a smaller space of structures, with large neutral networks enabling robustness and evolvability. This interplay between biophysics, computation, and evolution explains why RNA remains both a central biological molecule and a powerful model system for theoretical biology.