

# Combinatorial approaches to RNA folding

## Part I: Basics

Matthew Macauley

Department of Mathematical Sciences  
Clemson University  
<http://www.math.clemson.edu/~macaule/>

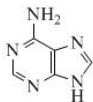
Math 4500, Fall 2016

## What is RNA?

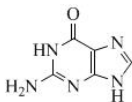
There are three major **macromolecules** that are essential to all forms of life:

- **RNA** (*Ribonucleic acid*)
  - **DNA** (*Deoxyribonucleic acid*)
  - **Proteins**
- } nucleic acids
- } biochemical compounds

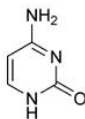
**Nucleic acids** are biological molecules built from strings of **nucleotides**.



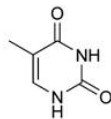
adenine



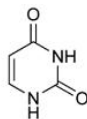
guanine



cytosine



thymine



uracil

A and G are *purines*. C, T, and U are *pyrimidines*.

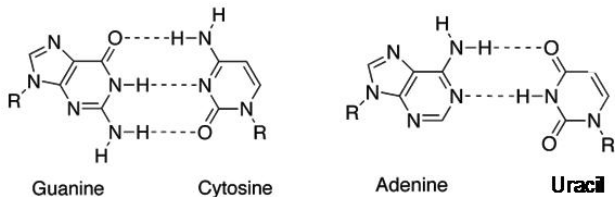
**DNA** strands consist of A, C, G, and T.

**RNA** strands consist of A, C, G, and U.

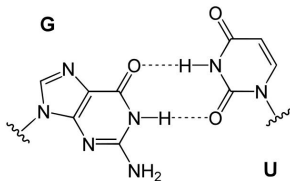
## What is RNA?

Combinatorially, an RNA strand is a length- $n$  sequence (of **bases**, or **nucleotides**), over the alphabet  $\{A, C, G, U\}$ .

Bases can **bond**: A with U, and C with G. (*Watson-Crick base pairs*.)



Additionally, U can bond with G. (Called a *wobble-pair*).



## Nucleic acid strands

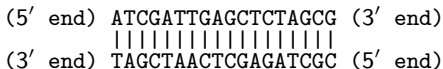
Other bonds are either chemically impossible (GT, AC), or thermodynamically unstable (purine–purine, pyrimidine–pyrimidine) and thus very rare.

Nucleotides are strung together along a sugar-phosphate backbone, called a **strand**.

Strands of nucleic acid have directionality: a **5' end** “five prime end” and a **3' end** “three prime end.”

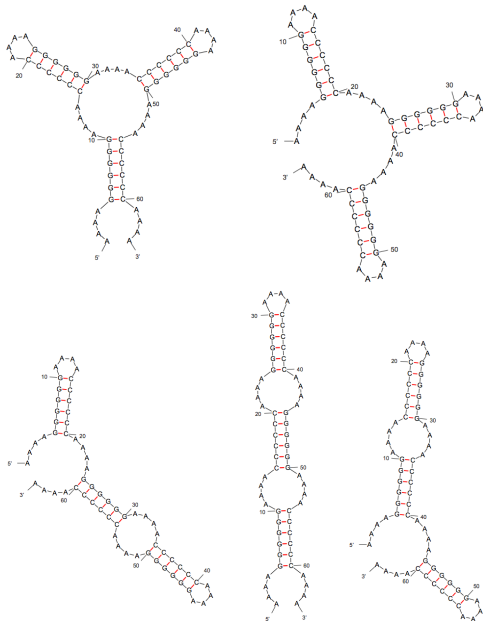
Single strands of DNA or RNA are written in the 5'-to-3' direction.

*DNA consists of two strands* that bond together, in opposite directions. One strand thus determines the other stand. For example:



*RNA consists of a single strand*. It can fold and bond to itself. It is much less structurally constrained than DNA! (And more mathematically interesting!)

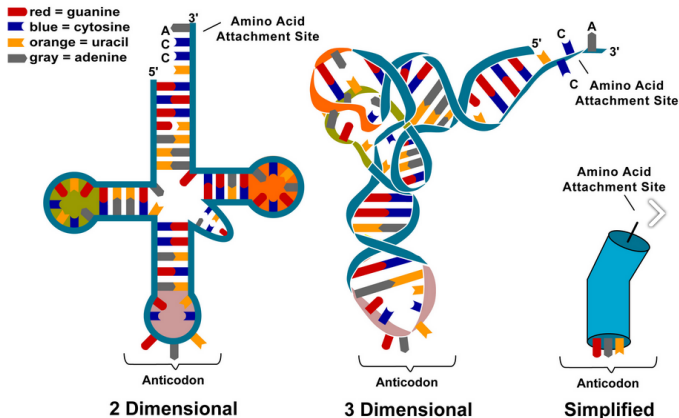
# How does RNA fold? [image from C. Heitsch; Georgia Tech]



# RNA folding

The physical structure of a folded RNA strand can be described on several levels.

- *Primary structure*: The raw sequence of nucleotides.
- *Secondary structure*: The bonding between nucleotides on a single strand.
- *Tertiary structure*: Embedding (e.g., twisting, knotting, etc.) of the strand in 3-dimensional space.



Dept. Biol. Penn State ©2002

## Central questions about RNA folding

### Questions

1. Given an RNA strand, can we predict how it will fold?
2. How does the structure that an RNA strand (or protein) folds into affect its function? ("*structure-to-function problem*")

Question 2 above is more purely biological.

In contrast, Question 1 can be attacked by mathematicians, computer scientists, engineers, without too much biology knowledge.

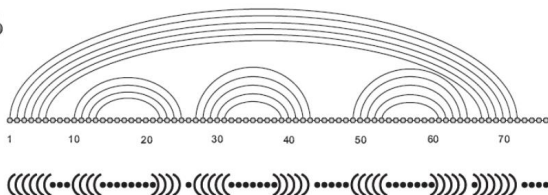
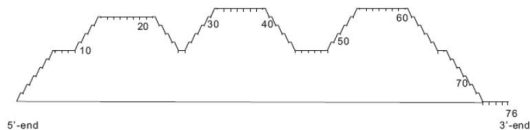
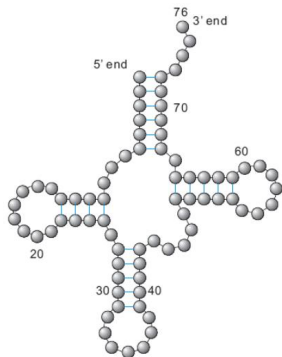
Before we proceed, we will need to establish a combinatorial framework for describing RNA strands.

# Combinatorial models of RNA

To each **base**, we associate a **vertex**. We use an **edge** to denote a **bond**.

The **arc diagram** of an RNA folding consists of **vertices**  $V = [n] = \{1, \dots, n\}$  and a collection of **edges**, or **arcs**,  $E = \{(i, j) \mid i < j\} \subsetneq V \times V$ .

There are several natural combinatorial models we can associate with RNA strands:

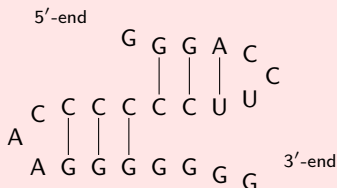




## Secondary structures

### Exercise

Consider the following fold of the RNA sequence GGGACCUUCCCCCAAGGGGGGG:



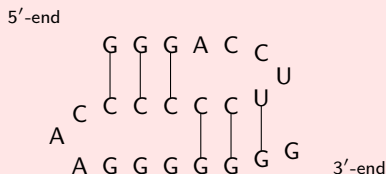
- (i) Draw the corresponding **arc diagram**.
- (ii) Write out this secondary structure in **point-bracket notation**.
- (iii) Draw the corresponding **Motzkin path**.

You should notice that your arc diagram has no crossings.

Formally, two arcs  $(i_1, j_1)$  and  $(i_2, j_2)$  (with  $i_1 < i_2$ ) are *crossing* if  $i_1 < i_2 < j_1 < j_2$ . An arc diagram is **non-crossing** if it has no crossing arcs. Such an RNA structure is (unfortunately) called a **secondary structure**.

## Exercise

Consider the following fold of the same RNA sequence:



- (i) Draw the corresponding **arc diagram**.
- (ii) Write out this secondary structure in **point-bracket notation**.
- (iii) Draw the corresponding **Motzkin path**.

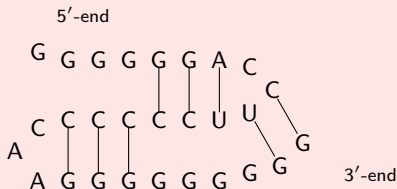
*Which of these go wrong, now that there are crossing arcs?*

An RNA structure is a **pseudoknot** if its arc diagram has crossings.

An arc diagram is  **$k$ -noncrossing** if there is no set of  $k$  mutually crossing arcs.

## Exercise

Consider the following fold of the same RNA sequence:



- (i) Draw the corresponding **arc diagram**. What is the smallest  $k$  for which this is  $k$ -noncrossing .
- (ii) What if the first G bonds with the C “directly below” it (vertex 17). Does this change the  $k$  from the previous part?
- (iii) Draw a picture of a folded RNA strand (like the one above) that is 4-noncrossing but not 3-noncrossing.

## Parameters

The **length** of an arc  $(i, j)$  is  $|i - j|$ . An arc of length  $k$  is called a  **$k$ -arc**.

A **stack** (or **stem** or **helix**) is a sequence of nested arcs:

$$(i, j), (i + 1, j - 1), \dots, (i + (\sigma - 1), j - (\sigma - 1)),$$

and a maximal such  $\sigma$  is its *size*.

For thermodynamical reasons, there are several key features of interest to us:

- The **minimum loop size** (i.e., arc-length),  $\lambda$ .
- The **minimum stack size**,  $\sigma$ .

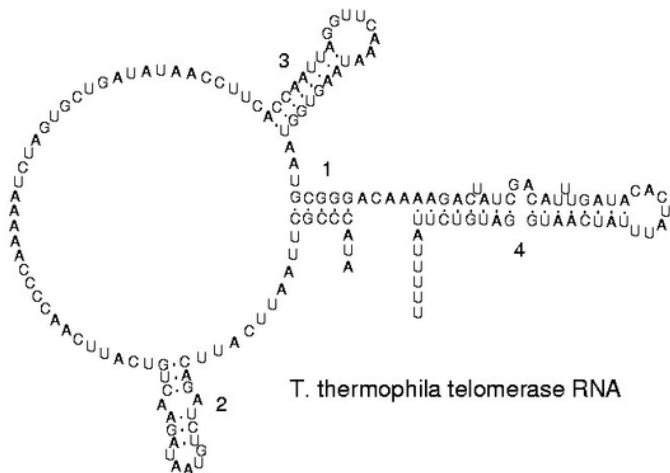
It is common to assume that  $\sigma = 2$  and  $\lambda = 3$  or  $\lambda = 4$ .

## Mathematical questions

- How can we enumerate the number of structures with certain parameters? This may require *asymptotic analysis*.
- How can we *uniformly generate* an RNA structure?
- What is the distribution of certain motifs (e.g., base-pairs, hairpin loops, etc.) in these structures?
- What is the *topology* of one of these structures?

## Loop decomposition

Every secondary structure can be described by its **loops**, which come in different types.



T. thermophila telomerase RNA

## Loop decomposition

Given a basepair  $(i, j)$  with  $i < v < j$ , say that  $v$  is **accessible** from  $(i, j)$  if there is no basepair  $(i', j')$  such that  $i < i' < v < j' < j$ .

Loosely speaking,  $v$  is accessible from  $(i, j)$  if it can “look up and see the arc  $(i, j)$ .”

A basepair  $(v, w)$  is accessible from  $(i, j)$  if both  $v$  and  $w$  are accessible.

The  **$k$ -loop** closed by  $(i, j)$  is the set of  $(k - 1)$  basepairs and the isolated bases that are accessible from  $(i, j)$ .

We do NOT include either  $i$  or  $j$  in the  $k$ -loop closed by  $(i, j)$ .

The **size** of a loop is the number of **isolated bases** in it

### Loop types

0. The vertices not accessible from any arcs form the unique 0-loop, or **null loop**  $L_0$ .
1. A 1-loop is called a **hairpin loop**
2. There are three types of 2-loops: **bulge loops**, **interior loops**, and **stacked pairs**.
3. A  $k$ -loop for  $k \geq 3$  is called a **multiloop**.

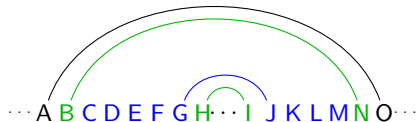
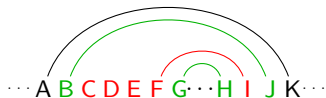
# Loop decomposition

## 2-loops

Suppose  $(i', j')$  is the unique accessible base pair from  $(i, j)$ . Then the resulting 2-loop is:

- 2a. a **stacked pair** if  $i - i' = j' - j = 1$ ;
- 2b. a **bulge loop** if exactly one of  $i - i'$  and  $j' - j$  is  $> 1$ ;
- 2c. an **interior loop** if both  $i - i'$  and  $j' - j$  are  $> 1$ ;

Two 2-loops: a bulge loop (left) and an interior loop (right). Each secondary structure also contains two 2-loops that are stacked pairs.



## Loop decomposition with pseudoknotting

Things get a little more complicated when the diagram contains a pseudoknot, but there is still a well-defined decomposition. (We won't go into details.)

