RNA
Secondary Structure
RNA

Ribonucleic Acid (RNA) is the workhorse chemical in the body. It is responsible for making the proteins and other biochemicals that the body needs to function.

Unlike its relative DNA which holds the genetic information of the body in its double stranded structure, RNA is a single stranded compound of nucleic acids held together on a backbone of polysaccharides. Another basic difference is the replacement of Thiamine (T) in DNA by Uracil (U) in RNA.

Being single stranded, these long RNA molecules can bend over and attach to themselves creating what is called a secondary structure.
Secondary Structure
Secondary Structure

- Single strand
- Double strand
- Single-nucleotide bulge
- Three-nucleotide bulge
- Hairpin loop
- Symmetric internal loop
- Asymmetric internal loop
- Two-stem junction or coaxial stack
- Three-stem junction
- Four-stem junction
Secondary Structure

A Crude Model

To model the folding of RNA molecules, we represent the molecule by points on the real line (each point a nucleic acid) and join points by edges if they bond in the fold.
A Crude Model

A graph of this type could represent a folding if and only if the edges do not intersect. This can be phrased algebraically as follows:

Let the vertices be consecutively ordered on the real line. Let $E$ be the set of edges corresponding to bonds and let each edge be represented by ordered pairs $(a,b)$ with $a < b$. The set $E$ is said to be *noncrossing* if for $(a,b)$ and $(c,d) \in E$, $a \leq c \leq b$ if and only if $a \leq d \leq b$.

This type of graph, with a noncrossing edge set, is a special case of a more general structure.
We will let $[n] := \{1,2,...,n\}$. A partition of $[n]$ is a set of non-empty disjoint subsets of $[n]$ whose union is $[n]$. The sets in a partition are called blocks.

We can represent a partition by a hypergraph whose vertices are the elements of $[n]$ and whose edges are the blocks of the partition.

A partition is called **noncrossing** if whenever four elements satisfying $1 \leq a < b < c < d \leq n$ are such that $a$ and $c$ are in the same block and $b$ and $d$ are in the same block, then the blocks coincide.
Noncrossing Partitions

The noncrossing partitions that correspond to secondary structure of RNA are those where the hypergraph is a graph (i.e., all the blocks have size two).

In the biological literature the secondary RNA structure is referred to as a *binary macromolecule*.

In the mathematical model we permit bonds between sites that are near each other (even adjacent). The data however indicate that there must be at least 4 or 5 sites between sites that are bonded. In other words, we don't find hairpin turns in nature.
Dyck Words

A **Dyck word** of length 2n is a string of n X's and n Y's with the property that each initial segment has at least as many X's as Y's.

**Example**: The Dyck words of length 6 are:

XXXYYY  XYXXYY  XYXYXY  XYYXYY  XXYXYY

If we interpret the X's as opening parentheses (i.e. left ( ) and the Y's as closing parentheses (right ) ), then a Dyck word is a *properly formed expression* in terms of parentheses.

((( )))    ( ) ( )  ( )( )  ( ) ( ) ( )  ( ) ( ) ( )
Dyck Paths

Dyck words have a geometric interpretation as well. Consider the integer lattice points and paths from (0,0) to (2n,0) where the only allowed steps are Up = (1,1) and Down = (1,-1). The paths of this type which never drop below the x-axis correspond to Dyck words and are called *Dyck paths*.
Lattice (Monotonic, Catalan) Paths

Another way to visualize Dyck words as paths uses the two dimensional integer lattice as well. This time we want paths from (0,0) to (n,n) with allowable steps being East = (1,0) and North = (0,1). Paths which do not rise above the line y=x correspond to Dyck words and are called Lattice or Monotone or Catalan paths.
A rooted binary tree is **full** if every vertex either has 2 or 0 children. Consider the number of rooted binary trees which have $n+1$ leaves (pendant vertices, vertices with no children.) For $n = 3$ we have:
Associating Binary Operations

Consider \( n+1 \) elements and a binary operation on these elements. If the elements are written in a fixed order, in how many ways can we introduce parentheses to give a proper association for a binary operation. For example, with \( n = 3 \):

\[
((ab)c)d \quad (a(bc))d \quad a((bc)d) \quad a(b(cd)) \quad (ab)(cd)
\]

By labeling the leaves of a full binary tree with \( n+1 \) leaves we can see the relationship between these two things.
Consider the number of ways to divide a convex polygon with $n+2$ sides into triangular regions by inserting diagonals that do not intersect in the interior.

For $n = 4$
A Surprising Equivalence

The number of triangulations of a convex polygon of \( n+2 \) sides is the same as the number of full binary trees with \( n+1 \) leaves.

Label the sides, except for one which we will call the base, consecutively with the symbols \( a_1, a_2, ..., a_{n+1} \). The internal chords of a triangulation will be labeled as the product of the other two sides (in lexicographic order). The triangulation can then be identified with the final label that is given to the base.

The association between such labels and full binary trees has already been shown.
A Surprising Equivalence

\[(a_1 a_2)(a_3 (a_4 a_5)) (a_6 a_7)\]
Catalan Numbers

Let $C_n$ denote the number of Dyck words of length $2n$. By letting $X = +1$ and $Y = -1$, we can view a Dyck word as a sequence $a_1, a_2, \ldots, a_{2n}$ of $n$ +1's and $n$ -1's such that the partial sums $a_1 + a_2 + \ldots + a_k \geq 0$ for all $1 \leq k \leq 2n$.

Let $U_n$ denote the number of sequences of $n$ +1's and $n$ -1's which are not Dyck words. Then we have:

$$C_n + U_n = \binom{2n}{n},$$

since we can get all sequences by choosing $n$ positions out of the $2n$ to make -1's.
Catalan Numbers

We will now count the $U_n$. For a non-Dyck word, there is a smallest $k$ for which the partial sum is negative. For this $k$ we must have $a_1 + a_2 + \ldots + a_{k-1} = 0$ and $a_k = -1$. Thus, $k-1$ must be even (i.e., $k$ is odd) since we need equally many $+1$'s and $-1$'s to get $0$.

Now change the sign of the first $k$ terms, and get a sequence consisting of $(n+1) +1$'s and $(n-1) -1$'s.

For example,

Non-Dyck word: $+ + - - - + + - + -$ with $n = 5$

$k = 5$

new sequence $- - + + + + + - + -$ having 6 $+$'s and 4 $-$'s (notice ending is unchanged)
Catalan Numbers

Furthermore, any sequence consisting of \((n+1)\) +1's and \((n-1)\) -1's can be used to get a non-Dyck word. In such a sequence, there must be a first time where the number of +1's exceeds the number of -1's. By changing the sign of all the terms up to and including this point, we get a sequence of \(n\) +1's and \(n\) -1's which is not a Dyck word.

Example:

- - + - + + + - + +

has 6 +'s and 4 -'s \((n = 5)\)

7th term is first place where there are more +'s than -"s.

++ - + - - - - + +

is a non-Dyck word

So, non-Dyck words with \(n\) +1's and \(n\) -1's are equivalent to sequences of \((n+1)\) +1's and \((n-1)\) -1's.
Thus, we can count $U_n$ to get:

$$U_n = \binom{2n}{n+1}$$

$$C_n = \binom{2n}{n} - \binom{2n}{n+1}$$

$$C_n = \frac{(2n)!}{n! n!} - \frac{(2n)!}{(n+1)!(n-1)!} = \frac{(2n)!}{n!(n-1)!} \left( 1 - \frac{1}{n} \right) = \frac{(2n)!}{n!(n-1)!} \left( \frac{1}{n(n+1)} \right)$$

$$C_n = \frac{(2n)!}{n! n!} \left( \frac{1}{n+1} \right) = \frac{1}{n+1} \binom{2n}{n}.$$

The numbers $C_n$ are called the **Catalan numbers**.
Catalan Recursion

Consider the Dyck paths of length 2n. There are \( C_n \) of these since they correspond bijectively to Dyck words of length 2n. Such a path may hit the x-axis before it reaches the end \((2n,0)\), but must do so at a point with even first coordinate, say \((2k,0)\), \(1 \leq k \leq n-1\). When this occurs for the first time, we have a special type of Dyck path from \((0,0)\) to \((2k,0)\) followed by any Dyck path from \((2k,0)\) to \((2n,0)\). The number of the latter is \( C_{n-k} \). The special Dyck path at the beginning starts with an Up (as all Dyck paths must), ends with a Down (ditto) and never drops below the line \( y = 1 \). Removing the first and last steps, you are left with a Dyck path from \((1,1)\) to \((2k-1,1)\) and there are \( C_{k-1} \) of these if \( k > 1 \). Thus, \( C_{k-1} C_{n-k} \) Dyck paths meet the x-axis for the first time at \((2k,0)\) for \(2 \leq k \leq n-1\).
Catalan Recursion

We haven't counted the Dyck paths that don't meet the x-axis except at the start and finish, but these are just like the special Dyck paths that we have counted, and there are $C_{n-1}$ of them. We also haven't counted those with $k = 1$, but again there are $C_{n-1}$ of them. So we have,

$$C_n = 2C_{n-1} + \sum_{k=2}^{n-1} C_{k-1}C_{n-k}.$$ 

Let $r = k-1$. Define $C_0 = 1$ and we get:

$$C_n = C_0 C_{n-1} + \sum_{r=1}^{n-2} C_r C_{n-1-r} + C_{n-1}C_0 = \sum_{r=0}^{n-1} C_r C_{n-1-r}.$$
Catalan Sequence

The first 26 Catalan numbers (starting with n = 0) are:

1, 1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786, 208012, 742900, 2674440, 9694845, 35357670, 129644790, 477638700, 1767263190, 6564120420, 24466267020, 91482563640, 343059613650, 1289904147324, 4861946401452, …
Triangulations

Let $t_n$ be the number of triangulations of a convex polygon of $n+2$ sides. We will determine the relationship between $t_n$ and $t_{n+1}$.

Given a polygon $P$ with $n+2$ sides, pick a side as the base. Triangulate $P$ and select any of its $2n+1$ edges ($n+2$ sides + $n-1$ internal diagonals) and orient it. There will be $2(2n+1)t_n$ ways to do this. Now, let $Q$ be a polygon with $n+3$ sides, and pick a side as the base. Triangulate and mark one of the sides other than the base. There will be $(n+2)t_{n+1}$ ways to do this.

We will show that these are the same numbers by setting up a bijection between the objects.
Triangulations

Taking a triangulated P with directed edge, we split the directed edge, separate the parts and form a new triangle by adding a new edge which we mark. This gives a marked triangulation of Q.

On the other hand we can collapse the triangle of Q which has the marked side to go in reverse.
Triangulations

So we have:

\[(4n+2)t_n = (n+2)t_{n+1}\]

or by using \(m = n + 1\) we have:

\[t_m = \frac{4m-2}{m+1} t_{m-1}.\]

We prove that \(t_m = C_m\) by induction. \(t_1 = 1 = C_1\) is easily seen. Assume the statement true for \(m-1\). Then

\[t_m = \frac{4m-2}{m+1} \frac{1}{m} \left( \frac{2(m-1)}{m-1} \right) = \frac{4m-2}{m+1} \frac{1}{m} \frac{(2m-2)!}{(m-1)! (m-1)!} \frac{m}{m} = \frac{1}{m+1} \frac{(2m)!}{m! m!} = \frac{1}{m+1} \binom{2m}{m}.\]
The Catalan Family

We now have 6 counting problems all of which have the Catalan numbers as solutions. Actually, there are many more counting problems with these answers.

In his text Enumerative Combinatorics (vol. 2), Richard Stanley lists 66 counting problems of this type. In a more recent addendum, he has increased this number to well over a hundred. [See supplements section for these two lists]

It has become somewhat of a game to find bijections between the objects being counted that are in this Catalan Family of problems. We shall look at a few more members of this family.
A Little History

The Catalan numbers first arose in Leonhard Euler's work on counting triangulations of convex polygons. He obtained the general formula for $C_n$ after a “quite laborious” induction.

The sequence is named after Eugene Charles Catalan who, in 1838, worked on the problem of placing parentheses in an expression of a binary operation.

The surprising equivalence between these problems was given by H.G. Forder in 1961.
Non-crossing Partitions

Consider any Catalan (monotonic) path from \((0,0)\) to \((n,n)\). Number all the East edges consecutively starting with 1. Now interpret the E edges as left parentheses ‘(‘ and the N edges as right parentheses ‘)’. Since the path corresponds to a Dyck word, every right parentheses is matched with a unique left parentheses and hence an E edge (which is numbered). Label the N edge of this right parentheses with the number of the associated E edge. Finally, consider the partition of \([n]\) whose parts are the numbers on the vertical sections of the path.

This partition is a non-crossing partition.
Non-crossing Partitions

For example:

1 2 3 4 5 6 |
3 4 |
7 8

1 2 5 6 | 3 4 | 7 8
Non-crossing Partitions

We can also reverse the procedure and construct a Catalan path from a non-crossing partition. We will build the path in stages, first constructing the N segments and then connecting these with E segments. For each part of the partition, build a sequence of N steps, one for each element of the partition part and labeled from top to bottom by the elements in increasing order. Order these segments from left to right in the order of their biggest (bottom-most) elements. Now, starting at (0,0), put in and consecutively label E-steps until you have labeled an E-step with bottom element of the first N-segment. From the top of this N-segment the path continues with E-steps (continuing the labeling from the last E-step) until it reaches the next N-segment. Repeat until done.
Non-crossing Partitions

For example: Consider the non-crossing partition

1 7 8 | 2 | 3 4 6 | 5

1 2 3 4 5 6 7 8
Non-crossing Chords

We have shown that the number of non-crossing partitions of \([n]\) is \(C_n\). We did this by showing the equivalence of these partitions with the monotonic paths. But a monotonic path is equivalent to a Dyck word of length 2n, and if we view the Dyck word in terms of parentheses, the pairing of parentheses gives a non-crossing binary partition of \([2n]\).

Another way to look at a non-crossing binary partition of \([2n]\), is to take the number line from 0 to 2n and bend it into a circle. The paired elements of the partition now determine a chord of the circle and none of these chords cross. So, \(C_n\) counts the number of ways that 2n points on a circle can be joined by n chords which do not cross each other.
Binary Macromolecules

We are now in a position to count the number of binary macromolecules that can be formed from an RNA strand of length $n$.

The difference between this problem and counting the non-crossing binary partitions is that there are sites on the RNA strand which do not get paired (the bulges and turns for instance), but in a binary partition everything is paired.

We shall count the binary macromolecules by starting with smaller binary partitions and adding singletons in all possible ways.
Distributions

Suppose that a binary macromolecule of length \( n \) has \( 2r \) paired sites and \( n-2r \) singleton sites. We can think of the \( n \) positions as being lined up and then picking \( 2r \) of these positions to be paired up. Once this choice is made, there will be \( C_r \) ways to form a binary non-crossing partition of the \( 2r \) positions.

Thus, for a given \( r \) there will be

\[
\binom{n}{2r} C_r
\]

binary macromolecules of length \( n \) with \( 2r \) paired sites.
We can now count the number of binary macromolecules that can be formed by a single string RNA molecule of length \( n \) by varying \( r \).

\[
\sum_{r=0}^{\left\lfloor \frac{n}{2} \right\rfloor} \binom{n}{2r} C_r,
\]

with \( C_r \) being the \( r^{\text{th}} \) Catalan number.
Example

With \( n = 5 \) we should obtain:

\[
\binom{5}{0}C_0 + \binom{5}{2}C_1 + \binom{5}{4}C_2 = 1 + 10 + 10 = 21
\]

binary macromolecules of length 5.