On Some Properties of Catalan Numbers and Application in RNA Secondary Structure

Similien NDAGIJIMANA

College of Science and Technology

School of Science

Department of Applied Mathematics

Master of Science

Huye, 2016

On Some Properties of Catalan Numbers and Application in RNA Secondary Structure

By

Similien NDAGIJIMANA

PG 214003431

A dissertation submitted in Partial fulfillment of the
requirement for the degree of
Master of Science in Applied Mathematics-Statistical
Modelling and Actuarial Sciences

In the College of Science and Technology

Supervisor: Prof. Yishao Zhou

May, 2016

Declaration

This is to certify that this thesis is my own work, that due reference has been made in the text to all other material used, that it is less than 20,000 words in length, exclusive list of figures, tables and bibliographies, and that it has not been previously submitted for any comparable academic award.

Student's names : Similien NDAGIJIMANA
Date : May 2016

Signature:

Dedication

I dedicate first of all my thesis to my God who protected me in all of my life. I dedicate also this work my family, my friends and my colleges who have supported me throughout the process. A special feeling of gratitude to my loving parents, my brothers and Sisters and my fellow students for the encouragement. I dedicate this work and give special thanks to my best friend wife
Deborah MUKANDAMAGE.

Acknowledgment

First of all I address my sincerest gratitude to my supervisor, Prof. Yishao Zhou, who has supported me throughout my thesis with her patience and knowledge that allow me to achieve our objectives. I attribute the level of my Masters degree to her encouragement and effort and without her this thesis, too, would not have been completed or written. I address my gratitude to the university of Rwanda and the collaboration of Linköping University. I am also grateful to lecturers, in the Department of Applied Mathematics. I am extremely thankful and indebted to their for sharing expertise, and sincere and valuable guidance and encouragement extended to me. I also thank my parents for the unceasing encouragement, support and attention. I am also grateful to my beloved wife who supported me during my master's studies.

abstract

The purpose of this thesis is to understand the connections between the Catalan Numbers and Ribonucleic Acids (RNA) Secondary Structure. We show in this thesis the mathematical wealth of the Catalan numbers from several mathematical branches to one particular application in secondary structure of RNA used in protein folding. The different ways codes folding of RNA are represented. In particular, we show different representations of the Catalan numbers from the aspect of traditional combinatorial counting and generating functions as well as from classical function theory through orthogonal polynomial system where linear algebra plays role in computation.

Nomenclature

UR: University of Rwanda
LiU: Linköping University
RNA:Ribonucleic Acides

Key words

Catalan number, RNA secondary structure and code deciphering.

Contents

Chapter 1

Introduction

We start this thesis by some brief presentations on the background of the Calalan numbers and their properties and applications. We will also have a short discussion on Ribonucleic acid.

1.1. Motivation and background of Catalan number

From various sources like books and Wikipedia we see that in combinatorial mathematics, the Catalan numbers form a sequence of natural numbers that occur in various counting problems, often involving recursively-defined objects such as polygon triangulation, balanced parentheses, mountain ranges, diagonal avoiding paths and binary tree. They are named after the Belgian mathematician Eugène Charles Catalan. Using zero-order numbering, the n^{th} Catalan number is given directly in terms of binomial coefficients by

$$\begin{aligned} C_n &= \frac{1}{n+1}\binom{2n}{n} \\ &= \frac{(2n)!}{(n+1)!n!} \end{aligned} \tag{1.1}$$

It is not clear from this definition that the n^{th} Catalan number C_n $(n \geq 0)$ is a natural number [12]. Let us rewrite the formula defined by the relation (1.1)

$$\begin{aligned} C_n &= \frac{1}{n+1}\binom{2n}{n} = \frac{(2n)!}{(n+1)!n!} \\ &= (2n)!\left\{\frac{1}{(n+1)!n!}\right\} = (2n)!\left\{\frac{(n+1)-n}{(n+1)!n!}\right\} \\ &= (2n)!\left\{\frac{(n+1)}{(n+1)!n!} - \frac{n}{(n+1)!n!}\right\} = (2n)!\left\{\frac{1}{n!n!} - \frac{1}{(n+1)!(n-1)!}\right\} \\ &= \left\{\frac{(2n)!}{n!n!} - \frac{(2n)!}{(n+1)!(n-1)!}\right\} = \binom{2n}{n} - \binom{2n}{n+1}. \end{aligned}$$

Clearly, the last expression tells us that C_n is a natural number.

1.2. Commonly used properties of Catalan numbers

Let us describe some property of Catalan numbers from various paper and books such as [14]

i) $C_0 = 1$ and $C_{n+1} = \sum_{k=0}^{n} C_k C_{n-k}$ for $n \geq 0$

ii) $C_n = \frac{1}{n+1} \sum_{k=0}^{n} \binom{n}{k}^2$

iii) $C_0 = 1$ and $C_n = \frac{2(2n-1)}{n+1} C_{n-1}$

Proof. **i)** We use induction to prove that C_n satisfies the recursion. Clearly $C_0 = 1$ holds true. Now assume that for n, C_n satisfies the recursion for $n = l+1$, i.e., $C_{l+1} = \sum_{k=0}^{l} C_k C_{l-k}$. Now let $C(x)$ be the generating function of C_n, i.e. $C(x) = \sum_{n\geq 0} C_n x^n$. Then

$$C(x)^2 = \sum_{l\geq 0} \sum_{k=0}^{l} C_k C_{l-k} x^l \Leftrightarrow xC(x)^2 = \sum_{l\geq 0} C_{l+1} x^{l+1}$$

From this we see that the coefficient of x^{n+1} is

$$C_{l+2} = \sum_{k=0}^{l+1} C_k C_{l+1-k}$$

[3] as desired.

ii) By the Binomial Theorem

$$(1+x)^{2n} = \sum_{k=0}^{2n} \binom{2n}{k} x^k, \quad (1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k$$

So $(1+x)^{2n} = (1+x)^n (1+x)^n$ imply that

$$\sum_{k=0}^{2n} \binom{2n}{k} x^k = \sum_{i=0}^{2n} \sum_{k=0}^{i} \binom{n}{k} \binom{n}{n-k} x^i.$$

Comparing the coefficients of x^n yields

$$\binom{2n}{k} = \sum_{k=0}^{n} \binom{n}{k} \binom{n}{n-k} = \sum_{k=0}^{n} \binom{n}{k} \binom{n}{k}.$$

implying that

$$C_n = \frac{1}{n+1} \sum_{k=0}^{n} \binom{n}{k}^2$$

as given by [14]

iii) By using the following two identities

$$C_n = \frac{(2n)!}{(n+1)!n!} \tag{1.2}$$

$$C_{n-1} = \frac{(2(n-1))!}{(n-1)!n!} \tag{1.3}$$

we prove the property

$$C_n = \frac{2(2n-1)}{n+1} C_{n-1}$$

obtained by dividing the equation 1.2 to equation 1.3 and we get

$$\frac{C_n}{C_{n-1}} = \frac{2(n-1)}{n+1}.$$

Therefore by using cross product we get

$$C_n = \frac{2(n-1)}{n+1} C_{n-1}$$

□

Let us give an example that can illustrate the recursion

$$C_{n+1} = \sum_{k=0}^{n} C_k C_{n-k}$$

Example 1.2.1. By calculating C_5 using this recurrence relation above

$$C_4 = 14, C_3 = 5, C_2 = 2, C_1 = 1, C_0 = 1$$

Then by replacing the values above

$$C_5 = C_0C_4 + C_1C_3 + C_2C_2 + C_3C_1 + C_4C_0 = 42$$

Section 1.3 gives more examples that show clearly the property of Catalan numbers. These examples help us to understand the behaviour of Catalan numbers and how we can relate them to RNA secondary structure.

Definition 1.2.1. 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. Here X's are opening parentheses and Y's are closing parentheses. Then a *Dyck* word is a properly formed expression in terms of parentheses [5].

Theorem 1.2.1. The number of Dyck words of length $2n$ is

$$C_n = \frac{1}{n+1}\binom{2n}{n}.$$

Proof. Let c_n denote the number of Dyck words of length $2n$. By letting $X = +1$ and $Y = -1$ we can write a Dick words as a sequences $[d_1, d_2, d_3, ..., d_{2n}]$ of $n + 1$'s and $n - 1$'s such that the partial sums $d_1 + d_2 + d_3, ..., +d_k \geq 0$ for all $1 \leq k \leq 2n$. Let also U_n denote the number of sequences of $n + 1$'s and $n - 1$'s which are not *Dyck* words we can get all sequences by choosing n positions out of the $2n$ to make $-1's$. Then we have:

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

where

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

Thus

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

and therefore

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

□

Thus, the Catalan number C_n is the number of balanced parenthesis expressions of length $2n$ over alphabet of terminals as seen in [3].

1.3. Some applications of Catalan numbers

There is a set of examples which illustrate 66 different sequences of sets with the property that the n^{th} set of each collection has the same number C_n of objects, [18]. Here are some examples:

1 Let a binary operation on a set be given and let $x_1, x_2 \dots x_n$ be a word. Then the number of parenthesizing of this word is given by the Catalan number C_n.

2 Let A_n be a regular n polygon ($n \geq 3$). Then C_{n-2} is a number of possible triangulation of A_n.

3 Non-nesting matching on $2n$, i.e., ways of connecting $2n$ points in the plane lying on a horizontal line by n arcs, each arc connecting two of the points and lying above the points, such that no arc is contained entirely below another.

4 Ways of connecting $2n$ points in the plane lying on a horizontal line by n vertex-disjoint arcs, each arc connecting two of the points and lying above the points, such that the following condition holds: for every edge e let $n(e)$ be the number of edges e' that nest e , and let $c(e)$ be the number of edges e' that begin to the left of e and that cross e. Then $n(e) - c(e) = 0$ or 1.

5 Ways of linking any number of points in the plane lying on a horizontal line by non intersecting arcs lying above the points, such that the total number of arcs and isolated points is $n - 1$ and no isolated point lies below an arc.

6 Ways of connecting n points in the plane lying on a horizontal line by non-crossing arcs above the line such that if two arcs share an endpoint p, then p is a left endpoint of both the arcs.

7 Ways of relating $n + 1$ points in the plane lying on a horizontal line by non crossing arcs above the line such that no arc connects adjacent points and the right endpoints of the arcs are all distinct.

We see that the examples from 4 to 7 will help to understand the behaviour of Catalan numbers, this behaviour is very closed to RNA secondary structure as introduced in next section and more detailed later.

1.4. Ribonucleic acid

The Ribonucleic Acid (RNA) is the workhorse chemical in the body. It is in charge for making the proteins and other biochemicals that the body needs to function. Thus RNA is a single stranded compound of nucleic acids held together on a backbone of polysaccharides. See in [7]. A secondary structure is single stranded, these long RNA molecules that can bend without crossing over and attach to themselves.

RNA molecules are particularly interesting since they represent both genotypic legislative via their primary sequence and phenotypic executive via their functionality associated to $2D$ or $3D$-structures, respectively, [7]. Accordingly, it is believed that RNA may have been instrumental for early evolution-before proteins emerged. The primary sequence of an RNA molecule is formed by the sequence of nucleotides A, G, C, U which is paired like $(A - U, G - C)$ and $(U - G)$ according Watson-Crick. Single stranded RNA molecules form helical structures whose bonds satisfy the above base pairing rules and which, in many cases, determine their function. For instance, RNA ribozymes are capable of catalytic activity, cleaving other RNA molecules. RNA secondary structure prediction is of polynomial complexity [22] which is the result from the fact that in secondary structures no two bonds can cross.

For better understanding RNA let us define the following terms as it is given in [14]

1. An RNA molecule is a sequence of nucleotides of four possible types, denoted by the letters A,C,G and U, connected by a backbone and is called RNA *Primary Structure.*

2. Two nucleotides that are connected via hydrogen bonds are called a *base pair.* In the Watson-Crick base pairing, A always forms a base pair with U, as does G with C. In the Wobble base pairing, G forms a base pair with U.

3. *An RNA structure* is a set S of base pairs $i.j$ for $1 < i < j < n$ such that for an $i_1.j_1, i_2.j_2 \in S : i_1 = i_2 \Leftrightarrow j_1 = j_2$

4. The set S is called *secondary structure* if for all $i_1.j_1, i_2.j_2 \in S$ they are nested, i.e., $i_1 < i_2 < j_2 < j_1$, or disjoint, i.e., $i_1 < j_1 < i_2 < j_2$.

The definition bellow helps us for understanding the difference between Watson-Crick base pair and Wobble base pair [14].

Definition 1.4.1. A *Wobble base pair* is a pairing between two nucleotides in RNA molecules which do not respect Watson-Crick base pairing rules. The four main Wobble base pairs are guanine-uracil ($G-U$), hypoxanthine-uracil ($I-U$), hypoxanthine-adenine ($I-A$), and hypoxanthine-cytosine ($I-C$)

Those properties and definitions of Catalan numbers and RNA give us the impression of dealing with the most important problem as given in Statement of the problem

1.5. Statement of the problem

The theoretical biophysics nowadays has the most important problem and the greatest challenge of deciphering the code that transforms sequences of biopolymers into spatial molecular structures. Those sequences are properly visualized as a string of symbols which together with the environment encodes the molecular architecture of the biopolymer, one of the particular class of biopolymers, the ribonucleic acid (RNA) molecules, decoding of information stored in the sequence can be properly decomposed into two steps such us the transformation of the string into a planar graph and folding of the string into a three-dimensional structure under conservation of the neighborhood relation determined by the graph. Our main objective is the representation of a secondary structure ribonucleic acid in the sequence of the string into planar graphs and folding of the string into a three-dimensional structure without changing any of its property. we achieve the objective first of all proving some identities leading to Catalan number by using power Series.

The rest of thesis is organized as follows. In the second chapter we discuss and give some theorems, definitions, proofs and representation of some RNA secondary structure from literature that will help us understand the relation to the Catalan numbers. In chapter 3 we turn to mathematical treatment of the Catalan numbers from different perspective to show possible research directions. Then we study RNA secondary structure and deciphering the code in chapter 4. We conclude the thesis by some further discussions.

Chapter 2

Literature Review on RNA structure

In this section we introduce definitions, state theorems and proofs done in the literature. Also we give the representation of some ribonucleic acids secondary structure that will help us to achieve the objectives seat. We introduce the characteristic equations and the inequalities that relates to the sequence of RNA secondary structure. We deal also with graph-theoretic properties of secondary structure with no consideration of specific pairing rule or properties of a specific single strangled nucleic acid, anywhere the object is to give a precise definition of secondary structure and of the component secondary structure for an arbitrary sequence of length n. The total number of secondary structure for a sequence of length n is considered.

2.1. The concept of secondary structure

The concept of secondary structures is given in various ways.

Definition 2.1.1. A *secondary structure* is a graph of the set of n labeled point $1, 2, ..., n$ such that the adjacency matrix $A = (a_{ij})$ has the following three properties:

i) $(a_{i,i+1}) = 1$ for $1 \leq i \leq n - 1$.

ii) For each fixed i,$1 \leq i \leq n$ there is at most one $(a_{ij}) = 1$ where $j \neq i \pm 1$

iii) if $(a_{ij}) = (a_{kl}) = 1$, where $i < k < j$ then $i \leq l \leq j$. if $(a_{ij}) = 1$, i and j are said to be bonded

We explain each part in Definition 2.1.1. Item i) requires adjacent point to be bonded. Item ii) says that each point can be bonded to at most one other point. Then item iii) assures that if i and j are bonded then all bonding of the points $i < k < j$ is with points l between i and j. Item iii) is a crucial point of the definition as it keeps the structure form "folding" and becoming three-dimensional or tertiary structure.

The couple (i, j) is an edge which is bond or base pair if and only if $|i-j| \neq 1$. A vertex i connected only to $i-1$ and $i+1$ is called unpaired while a vertex i is said to be interior to the base pair (k, l) if $k < i < l$ as illustrated by Waterman [22] and by [8].

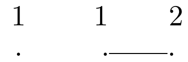
More explanation on Definition 2.1.1 is given by Waterman as shown bellow.

Theorem 2.1.1. Let $S(n)$ be the number of secondary structures for n points. Then $S(1) = S(2) = 1$ and for $n > 2$, $S(n)$ satisfies

$$S(n+1) = S(n) + \sum_{k=0}^{n-2} S(k)S(n-k-1), \tag{2.1}$$

where $S(0) \equiv 1$. Also $S(n) \geq 2^{n-2}$ For $n \geq 2$.

Proof. Let us prove the theorem by induction. By definition the only secondary structures for $n = 1, n = 2$ are



$(n = 1)$ $(n = 2)$

Obviously $S(1) = 1, S(2) = 1$. Assume we know $S(k)$ for all $1 \leq k \leq n$. We want to show that

$$S(n+1) = S(n) + \sum_{k=0}^{n-2} S(k)S(n-k-1).$$

To this end we argue how a sequence of $n+1$ points can be paired. There are two scenarios: $n+1$ is not paired or $n+1$ is paired with j for $j = 1, ..., n-1$. In the first case n points from 1 to n can form all possible secondary structures. In the second case, each of the points 1 to $j-1$ and $j+1, ..., n$ can form all possible secondary structures. All together they form secondary structures for $n+1$ points from 1 to $n+1$ by the definition. Now we can count how many possible secondary structure there are:

$$S(n+1) = S(n) + S(n-1) + S(1)S(n-2) + \ldots + S(n-2)S(1)$$

we know that $S(0) = 1$, so equation above takes the form for $(n \geq 2)$ of

$$S(n+1) = S(n) + \sum_{k=0}^{n-2} S(k)S(n-k-1).$$

This is the proof of the first part.

After one more iteration we obtain

$$S(n+1) = S(n) + S(n-1) + \sum_{k=1}^{n-2} S(k)S(n-k-1)$$

and equivalent to

$$S(n+1) = S(n) + S(n-1) + \sum_{k=0}^{n-3} S(k)S(n-k-2)$$

by shifting the index in the sum. Since $S(k+1) \geq S(k)$,

$$S(n+1) \geq S(n) + S(n-1) + \sum_{k=0}^{n-3} S(k)S(n-k-2) = S(n) + S(n) = 2S(n).$$

Now $S(2) = 1$ so that the above inequality implies that $S(n) \geq 2^{n-2}$, completing the proof. □

Next we need to show that 2^{n-2} is an unsatisfactory bound. To this end , we assume that $\lambda_2\alpha^n \leq S(n) < \lambda_1\alpha^n$ where $\lambda_{2>0}$, then we have the following inequalities

$$\begin{aligned} \lambda_1\alpha^{n+1} &\geq \lambda_2\Big(\alpha^n + \lambda_2\sum_{k=0}^{n-1}\alpha^{n-1}\Big) \\ &= \lambda_2\Big(\alpha^n + \lambda_2(n-2)\alpha^{n-1}\Big) \end{aligned} \tag{2.2}$$

and

$$(\lambda_1/\lambda_2)\alpha^2 \geq \alpha + \lambda_2(n-2)$$

which is a contradiction. This shows that the rate of growth of $S(n)$ is not geometric. therefore it will be shown that $S(n)$ is bounded by a geometric growth rate. Let us consider $\phi(x)$ as generating function, defined by $\phi(x) = \sum_{n=0}^{\infty} S(n)x^n$,the recursion formula in Theorem 2.1.1 can be multiplied by x^{n+1} and gives

$$x^2\phi^2(x) + (x-1-x^2)\phi(x) + 1 = 0.$$

This equation can be solved and the solution is

$$\phi(x) = \frac{x^2 - x + 1 - [1 + x(x^3 - 2x^2 - x - 2)]^{1/2}}{2x^2}.$$

Corollary 2.1.1.1. For $n \geq 2$ there is a fixed $M > 0$ such that

$$2^{n-2} \leq S(n) \leq M4^n.$$

Proof. Note that we have already proved the first inequality. It remains to show the second inequality.
Let $S(n+1) \leq \sum_{k=0}^{n} S(k)S(n+1-k)$ so that if $g(0) = g(1) = 1$ and

$$g(n) = \sum_{k=0}^{n-1} g(k)g(n-k)$$

therefore

$$S(n) \leq g(n) \Rightarrow g(n) = \frac{1}{2^{n-1}} \binom{2n}{n} = O(4^n)$$

as shown in [22]. □

This shows several of the features that can be identified from the adjacency matrix.

Definition 2.1.2. A secondary structure deals with the following structure elements

i) A *stack* consists of subsequent base pairs $(p-k, q+k), (p-k+1, q+k-1)$,...,$(p, q)$ such that neither $(p-k-1, q+k+1)$ nor $(p+1, q-1)$ is a base pair. k+1 is the length of the stack,$(p-k, q+k)$ is the *terminal* base pair of the stack.

ii) A *loop* consists of all unpaired vertices which are immediately interior to some base pair (p, q)

iii) An *external* vertex is an unpaired vertex which does not belong to a loop. A collection of adjacent external vertices is called an external element. If it contains the vertex 1 or n it is a free end, otherwise it is called joint.

The following lemma gives an additional result of secondary structure.

Lemma 2.1.1. Any secondary structure S can be uniquely decomposed into stacks, loops, and external elements.

Definition 2.1.2 and proof bellow is given in [8]

Proof. Each vertex which is contained in a base pair belongs to a unique stack. Since an unpaired vertex is either external or immediately interior to a unique base pair the decomposition is unique: Each loop is characterized uniquely by its "closing" base pair. □

Definition 2.1.3. Suppose A is the adjacency matrix for a secondary structure on $(1, 2, \ldots, n)$

i) The point j is said to be *paired* if there is some point $i \neq j \pm 1$ such that $a_{ij} = 1$.

ii) The region $(i+1, i+2, \ldots, (j-1))$ is a *loop* if $i+1, i+2, \ldots, j-1$ are all unpaired and $a_{ij} = 1$. The pair (i, j) is said to be the foundation of the loop.

iii) The sequence $(i+1, i+2, \ldots, (j-1))$. is a *bulge* if $(i+1, i+2, \ldots, (j-1))$ are all unpaired, i and j are both paired, and $a_{ij} \neq 1$

iv) An *interior loop* is two bulges $(i+l, i+2, \ldots, (j-1))$ and $(k+1, k+2, \ldots, (l-1))$ such that $a_{il} = 1$ and $a_{jk} = 1$ (Here $i < j < k < l$.)

v) A *join* is a bulge $i, i+1, \ldots, j$ such that $a_{kl} = 1$ for $k < i$ implies $l \leq i$ and $a_{kl} = 1$ for $k > j$ implies $l \geq j$.

vi) A *tail* is a sequence $(1, 2, \ldots, j)$ where $1, 2, \ldots, j$ are unpaired and $j+1$ is paired.

vii) A *ladder* is two sequences $(i+1, i+2, \ldots, i+j)$ and $(k+1, k+2, \ldots, e-k+j)$ such that $i+j+1 < k, a_{i+1,k+j-l+1} = 1$ for $1 \leq l \leq j$ and $a_{i,k+j+l} = a_{i+j+l,k=0}$. . I f $i+j+3 = k+1$,this last requirement is dropped

viii) A *hairpin* is the longest sequence $(i+1-i+2-\ldots-(j-1))$ containing exactly one loop such that $a_{i+1,j-1} = 1$ and $a_{i,j} = 0$. The paired points $i+1$ and $j-1$ will be called the foundation of the hairpin. We find this definition in [22].

The definition of secondary structure given above in Definition 2.1.3 is good enough to include the elementary structures. The structure can be easily identified from the graph or from the adjacency matrix. It has been the basis of algorithm to predict secondary structure therefore these algorithms rely on the examination of all possible secondary structure.

The next theorem gives full information on decomposition of secondary structure and the meaning of the above definition.

Theorem 2.1.2. Any secondary structure can be uniquely decomposed into loops, ladders, bulges, and tails.

Proof. If $a_{ij} = 1$ where $i \neq j \pm 1$ then i and j are members of sequences which are a ladder. Thus, assume i is an unpaired point.
Then let $(i-j), \ldots, i, \ldots, i+k$ be the longest sequence of unpaired point that i is the member off. if $i-j = 1$ or $i+k = n$ then i belongs to a tail. Elsewhere $i-j-1$ and $i+k+1$ are paired. If $a_{i-j-1,i+k+1} = 1$ then i belongs to a loop. If $a_{i-j-1,i+k+1} = 0$ then i belongs to a bulge. □

From above theorem we see that the ladders of two sequences making up the ladder does not make a new sequence. Since there are a finite number of ladders and every paired point belongs to a ladder then there exists a ladder such that the non-empty sequence of points between the two sequences, for making up the ladder has the property that they are all unpaired. By definition this sequence is a loop [22].

The following definition deals with the loops and degrees.

Definition 2.1.4. The *degree* of a loop is given by 1 plus the number of terminal base pairs of stacks which are interior to the closing bond of the loop. A loop of degree 1 is called hairpin (loop), a loop of a degree larger than 2 is called multi-loop. A loop of *degree* 2 is called bulge if the closing pair of the loop and the unique base pair immediately interior to it are adjacent; otherwise a loop of *degree* 2 is termed interior loop seen in [22]. More recursion formula of this definition is given in 2.3.3 subsection.

Definition 2.1.5. Let S an arbitrary secondary structure. For all S let us denote by $\omega(S)$ the unique secondary structure which is obtained from S the following procedure:

(i) For each hairpin, open its stack and add the corresponding bases to the hairpin loop.

(ii) If a bulge or interior loop follows, then add its digits also to the hairpin and continue by opening its stack.

(iii) If a multi-loop or a joint follows, then add the now unpaired digits to the multi-loop and stop.

From this definition Waterman [22] deduce that $\Omega(S)$ is the order of a secondary structure as the smallest number of repetitions of Ω necessary to obtain the open structure. Thus the open structure has order $\Omega = 0$ and any structure without a multi-loop has order $\Omega = 1$. In the following subsection we deal with the representation of secondary structure.

2.2. Biological motivation: RNA secondary structures

Latest discoveries emphasize the important regulatory and catalytic function performed by RNA molecules, their traditional role in mediating the production of protein from DNA to RNA as illustrated by Christine [7]. Like proteins, the functionality of an RNA molecule is determined by its overall three-dimensional structure. The primary structure of an RNA molecule is is oriented in biochemical chain of sequence of nucleotides or bases of which there are four types: Adenine (A),Guanine (G), Cytosine (C) and Uracil (U). Thus the chemically distinct ends called the $5^{'}$and $3^{'}$ When an RNA molecule folds, bonds may form between certain pairs of bases, where each base may pair with at most one other base. The resulting structure depends on environmental conditions, such as temperature and salt concentration of the solution in which the molecule resides.

A secondary structure R is a set of pairs i, j , $1 \leq i \leq j \leq n$ such that no index occurs in more than one pair. However, the unlike the canonical double stranded DNA helix or protein or the protein substructures created by more subtle amino acid interactions, a single stranded RNA molecule self bonds to create a set of intra molecule or base pairs called secondary structure, a function of single stranded RNA molecules are significantly related to be base pairing of their structure, by making the design, analysis and prediction of RNA secondary structure vital problems in computational molecular biology.

Mathematically, we treat an RNA sequence as a string R of consecutive symbols from the four letters A, C, G, U. For our purposes, we define a nested secondary structure of R as a set of intra-sequence base pairs, although a more standard definition includes additional constraints based on the energetic of RNA base pairing as said Christine [7]. Let us give a full definition:

Definition 2.2.1. Let $R = b_1 b_2 \ldots b_n \in (A, C, G, U)^+$. Denote the pairing of base b_i with b_j by $b_i - b_j$ for $1 \leq i < j \leq n$. **A nested secondary structure** of R is a set of base pairs $S(R) = b_i - b_j | 1 \leq 1 \leq n$ such that either $i < j < i' < j'$ or $i < i' < j' < j$, for any two base pairs $b_i - b_j, b'_i - b'_j \in S(R)$.

We see that the absence of base triples and of pseudo-knots, or base pairings with $i < j < i' < j'$, are fundamental assumptions in the current thermodynamic model of RNA base pairing. A nested RNA secondary structure decomposes into local substructures with nearest neighbor energetic interactions is determined by experimental researchers such as the Turner group [17]. For instance, the energy value assigned to a single-stranded region, or loop, is a function of the number and type of base pairs contained in the loop.

2.3. RNA Representation of Secondary Structure

The existence of RNA secondary structure will help us to deduce the three dimension representation

2.3.1. The graphs

A graphical representation of secondary structure S can be translated into a rooted ordered tree Υ by introducing an auxiliary root and representing a base pair (p, q) by a vertex x such that the sons $Y_1, ..., Y_k$ of x correspond to the base pairs $(p_1, q_1)...(p_k q_k)$ immediately interior to (p, q)[6].

The researchers like Waterman found that the property of secondary structure is best understood by considering a structure as diagram, which is obtained as follows: To draw the primary sequence of nucleotides horizontally and ignores all chemical bonds of backbone. For other side one draws all bonds satisfying the Watson-Crick base pairing rules (and $G - U$ pairs) as arcs in the upper half plane. In this representation we identify the RNA secondary structures have the following property: there exist no two arcs $(i_1, j_1), (i_2, j_2)$, where $i_1 < j_1$ and $i_2 < j_2$ with the property $i_1 < i_2 < j_1 < j_2$ and all arcs have at least length 2. Thus there exist no two arcs that cross in the diagram representation of the structure as shown in the following representation given in [6]. in the figure bellow

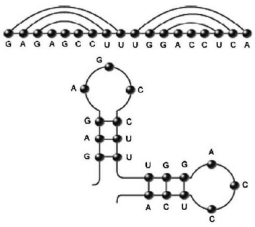


Figure 2.1: RNA secondary structures.

The are many types of representation like circular, Tree and no-crossing representation as shown by the matlab code in appendix from [10]. The figure are given from (2.3.1-2.3.1).

2.3.2. RNA Secondary Structure Design

A basic assumption underlying current understanding secondary structure is that the base sequences fold minimizes free energy. Under this hypothesis, the fundamental problem with RNA secondary structure design is to ensure

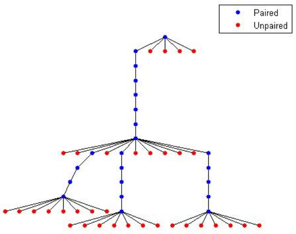


Figure 2.2: tree

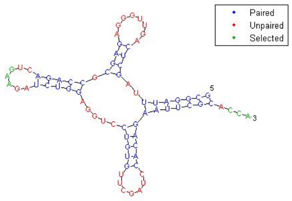


Figure 2.3: Non crossing

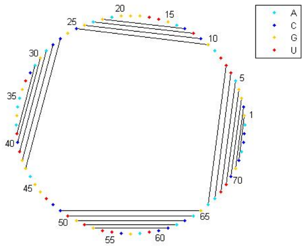


Figure 2.4: circular

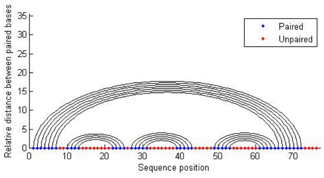


Figure 2.5: linear

the desired minimal free energy configuration of the constructed sequence. Intuitively, a sequence will fold to a configuration which minimizes loop costs while maximizing the beneficial stacked pairs.

Combinatorial design efforts have focused on precluding alternative configurations by ensuring that any improvement in loop energies is offset by the penalty of lost base pairs. These results employed a simple design strategy which still captured many essential aspects of this difficult problem.

To overcome this problem there are two algorithm solutions to the standard hypothesis excluding pseudoknots from RNA secondary structure. First way is to restrict the generating sequences which satisfy the loop-protecting property. This consists of maintaining the unpaired in any alternate configuration. This is done by enforcing restriction to a three alphabets $[A, C, G]$ and assigning A exactly to the unpaired segments. The second algorithmic solution is that the design must be sufficiently good to preclude any alternate minimal energy configurations from a particular subclass of structures as shown in [7].

In the loop-protecting RNA model, there can be no interaction between the intended A loop segments and the $C-G$, $G-C$ base pairs forming the helical stretches. Using two algorithmic solutions we may get possible secondary structure configurations.

2.3.3. RNA recursion

Based on above definitions 2.1.1-2.1.5 of RNA and Catalan number properties we have the basic recursion such that: A secondary structure on $n+1$ digits may be obtained from a structure on n digits either by adding a free end at the right hand or by inserting a base pair $1 \equiv k+2$, [8].

In the second case the substructure enclosed by this pair is an arbitrary structure on k digits, and the remaining part of length $n-k-1$ is also an arbitrary valid secondary structure. It can be obtained by the following recursion

$$S_{n+1} = S_n + \sum_{k=1}^{n-2} S(k)S(n-k-1) \qquad n \geq m+1 \tag{2.3}$$

$$\textit{where}$$
$$S_0 = S_1 = \ldots = S_{m+1} = 1$$

while the secondary structure with certain property is given by the recursion

$$J_{n+1}(b) = J_n(b) + \sum_{k=1}^{n-2} S_k J_{n-k-1}(b-1) \qquad b > 0, n > m+1, \tag{2.4}$$
$$J_n(b) = 0, b > 0, n \leq m+1, \qquad J_n(0) = 1, n \geq 0$$

where we denote $J_n(b)$ the number of structure with exact b.

Adding an unpaired digits to a structure on n digits does not affect the number of the components while introducing an added bracket makes the bracketed part of length k a single component and does not affect the remainder of the sequence. By consider the number of structures with exactly b base pair (bonds)on n vertices $H_n(b)$

$$H_{n+1}(b) = H_n(b) + \sum_{k=m}^{n-1} \sum_{l=0}^{b-1} H_k(l) H_{(n-k-1)}(b-l-1) \qquad b > 0, n \geq m+1 \tag{2.5}$$

$$H_n(b) = 0, b > 0, n \leq m+1, H^n(0) = 1, n \geq 0$$

for special cases of $m = 1$ the number of structure with exact number b was established in [15]:

$$H_n(b) = \frac{1}{b} \binom{n-1}{b+1} \binom{n-b-1}{b-1}$$

To obtain the number of of structures with b external digits $E_n(b)$ we have

$$E_{n+1}(b) = E_n(b-1) + \sum_{k=m}^{n-1} S_k E_{n-k-1}(b) \qquad b \geq 0, n \geq m+1 \tag{2.6}$$

$$E_n(n) = 1, E_n(b) = 0.b \neq n, \leq m+1, E_n(-1) = 0.$$

The recursion for the number $N_n(b)$ of the sequence with given number of stacks is obtained by introducing the auxiliary variable $Z_n(b)$ denoting the number of secondary structure with exactly b stacks given that it $3'$ and $5'$ end are paired, then we obtain

$$N_{n+1}(b) = N_n(b) + \sum_{k=m}^{n-1} \sum_{l=0}^{b} Z_{k+2}(l) E_{n-k-1}(b-l) \qquad b \geq 0, n \geq m+1 \tag{2.7}$$

$$N_n(n) = 1, \qquad N_n(b) = 0, \qquad b \neq n, \leq m+1$$

The auxiliary variable will be obtained by

$$Z_n(b) = Z_{n-2}(b) + N_{n-2}(b-1) - Z_{n-2}(b-1), \qquad Z_0(b) \quad = Z_1(b) = 0 \tag{2.8}$$

The structure with exactly hairpins is given by

$$A_{n+1}(b) = A_n + \sum_{k=m}^{n-1} \Big[\sum_{l=1}^{b} A_k(l) A_{n-k-1}(b-l) + A_{n-k-1}(b-1) \Big] \tag{2.9}$$

$n \geq m+1, \qquad A_n(b) = \delta_{0,b}, n \leq m+1$

where $\delta_{0,b}$ is Kronecker's symbol, i.e., $\delta_{0,0} = 1$ and $\delta_{0,b} = 0, b \neq 0$. This important recursion formula is given in [8].

Normally there are two main approaches to the folding problem; first is to predict RNA structure based on thermodynamics stability of the molecule and look for a thermodynamics optimum. The second approach is based on probabilistic models which try to find the state of RNA molecule in probabilistic optimum.

Chapter 3

Study of some properties of Catalan numbers

Very often we have to determine difficulty-looking sum or prove such identities. With generating function we can make it much easier. A *generating function* $f(x)$ is a formal power series

$$g(x) = \sum_{n \geq 0} a_n x^n$$

for given sequence $\{a_0, a_1, a_2, ...\}$. Note that we do not discuss the issues like convince and convergence region in this thesis.

Here are some examples of generating function.

1. For the sequence $\{0, 1, 1, ...\}$ we have the generating function

$$g(x) = x + x^2 + x^3 + \cdots = x(1 + x + x^2 + \cdots) = x \cdot \frac{1}{1-x} = \frac{x}{1-x}.$$

2. The sequence $\{1, 1, 2, 3, 5, ...\}$ (the Fibonacci number sequence) has the generating function

$$F(x) = \frac{1}{1-x-x^2}$$

using the recursive of the sequence $F_n = F_{n-1} + F_{n-2}$ because

$$\begin{aligned} F(x) &= \sum_{n \geq 0} F_n x^n = 1 + z + \sum_{n \geq 2} (F_{n-1} + F_{n-2}) x^n \\ &= 1 + x + x \sum_{n \geq 1} F_n x^n + x^2 \sum_{n \geq} F_n x^n \\ &= 1 + x + x(F(x) - 1) + x^2 F(x) = 1 + xF(x) + x^2 F(x). \end{aligned}$$

Solving this equation we obtain the desired function.

3. The generating function for the Catalan numbers $\{C_0, C_1, C_3, ...\}$ is

$$C(x) = \frac{1-\sqrt{1-4x}}{2x} = \frac{2}{1+\sqrt{1-4x}}.$$

We want to determine the function $C(x)$ through the series $\sum_{n\geq 0} C_n x^n$. Note that

$$C_n = \sum_{n=0}^{n-1} C_k C_{n-1-k}.$$

Then

$$\begin{aligned} C(x) &= \sum_{n\geq 0} C_n x^n = \sum_{n\geq 0}\sum_{k=0}^{n-1} C_k C_{n-1-k} x^n \\ &= 1 + \sum_{n\geq 1}\sum_{k=0}^{n-1} C_k C_{n-1-k} x^n = 1 + x\sum_{n\geq 0}\sum_{k=0}^{n} C_k C_{n-k} x^n \\ &= 1 + x\left(\sum_{k\geq 0} C_k x^k\right)\left(\sum_{n-k\geq 0} C_{n-k} x^{n-k}\right) = 1 + xC(x)^2. \end{aligned}$$

This shows that

$$C(x) = \frac{1\pm\sqrt{1-4x}}{2x}$$

To determine which sign we should take we note that

$$\lim_{x\to 0} C(x) = \lim_{x\to 0}\sum_{n\geq 0} C_x^n = C_0 = 1.$$

So we have to take the negative sign.

If we want to have the closed closed forms for F_n and C_n respectively, we meet with a problem to find the function that has the power series with the given sequence. So we see that it will be helpful if we can recognize some power series. For our purpose we list some power series

$$\frac{1}{1-x} = \sum_{n\geq 0} x^n \tag{3.1}$$

$$(1+x)^\alpha = \sum_{n\geq 0} (\alpha_n) x^n \tag{3.2}$$

$$\frac{1}{(1-x)^{k+1}} = \sum_{n} \binom{n+k}{n} x^n \tag{3.3}$$

$$\frac{x^k}{(1-x)^{k+1}} = \sum_{r\geq 0} \binom{r}{k} x^r \tag{3.4}$$

$$C(x) = \frac{1}{2x}(1-\sqrt{1-4x}) = \sum_{n\geq 0} C_n x^n \tag{3.5}$$

$$\frac{1}{\sqrt{1-4x}} = \sum_{n\geq 0} \binom{2n}{n} x^n \tag{3.6}$$

$$\frac{1}{\sqrt{1-4x}}\Big(\frac{1-\sqrt{1-4x}}{2x}\Big)^k = \sum_{n} \binom{2n+k}{n} x^n \tag{3.7}$$

$$\Big(\frac{1-\sqrt{1-4x}}{2x}\Big)^k = \sum_{n\geq 0} \frac{k(2n+k-1)!}{n!(n+k)!} x^n \qquad k\geq 1 \tag{3.8}$$

3.1. Generating function approach to some identities involving Catalan numbers

First we have the following proposition for the operations of differentiation and integration of the generating function, which can be found in any combinatoric text book. Again we do not discuss the interchange of the limit processes here. The identities dealt with in this section are taken from the literature[1] I have studied, but could not recall exact source at the writing moment.

Proposition 3.1.1. Let $\{a_0, a_1, a_2, \ldots\}$ be a given sequence and

$$A(x) = \sum_{n\geq 0} a_n x^n$$

be its generating function. Then

(i) The generating function for $\{na_n\}_{n=0}^{\infty}$ is $x\frac{d}{dx}(A(x))$.

(ii) The generating function for $\left\{\dfrac{a_n}{n+1}\right\}_{n=0}^{\infty}$ is $\displaystyle\int_0^x A(t)dt$.

Let us show the usefulness of these simple observations in computing difficulty-looking sums.

Example 3.1.1. Compute $\sum_{k=0}^{n} \frac{1}{k+1}\binom{n}{k}$.
Note that this is a less difficulty-looking one but we illustrate how generating function can be effectively applied to solve the problem. Note also that we can think of this sum as the value of the function

$$f(x) = \sum_{k=0}^{n} \frac{1}{k+1}\binom{n}{k} x^k$$

[1] In the revision I added now the reference [23] which I was not aware of. This book was mentioned by the reviewer who pointed out that some of identities below are in this book.

at $x = 1$. Clearly $\sum_{k=0}^{\infty} \frac{1}{k+1}\binom{n}{k}x^k$ is integral of

$$\sum_{k=0}^{n} \binom{n}{k} x^k = (1+x)^n$$

Thus

$$\begin{aligned}\sum_{k=0}^{n} \frac{1}{k+1}\binom{n}{k} x^k &= \int_0^x (1+t)^n dt \\ &= \frac{(1+x)^{n+1}-1}{n+1}\end{aligned}$$

Set $x = 1$ we have

$$\sum_{k=0}^{n} \frac{1}{k+1}\binom{n}{k} = \frac{2^{n+1}-1}{n+1}$$

Remark: Since we have to deal with the binomial coefficients and sums often we make the following conventions and the range of summation variables: (i) $\binom{x}{m} = 0$ if $m < 0$ or if x is a nonnegative integer less than m; (ii) a summation variable whose range is not explicitly stated is understood to be summed from $-\infty$ to ∞. For example $\sum_k \binom{n}{k} = 2^n$ should be understood in the sense that the sum ranges over all positive and negative and 0 values of k, the summand vanishes unless $0 \leq k \leq n$, and the sum has the value 2^n. Occasionally we will also use the notation $[x^n]f(x) =: f_n$ for the coefficient of x^n. We will also use the following identities:

$$\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r} \qquad n \geq r \tag{3.9}$$

$$\binom{-n}{r} = (-1)^r \binom{n+r-1}{r} \tag{3.10}$$

Now we prove some properties of Catalan number using generating function provided that we can recognize to which function a power series is associated. To this end we need some conventions about binomial coefficients and the range of summation variables:

Proposition 3.1.2.

$$nC_n = \sum_{j=0}^{n-1} C_j \binom{2n-2j-1}{n-j}$$

in other words

$$\sum_{j=0}^{n-1} \frac{1}{j+1}\binom{2j}{j}\binom{2n-2j-1}{n-j} = \frac{n}{n+1}\binom{2n}{n}$$

Proof. Let

$$f_n = \sum_{j=0}^{n-1} C_j \binom{2n-2j-1}{n-j}.$$

Consider its generating function $\sum_{n\geq 1} f_n x^n$ Note that

$$\binom{2n-2j-1}{n-j} = \binom{2n-2j-1}{n-j-1}$$

so

$$f_n = \sum_{j=0}^{n-1} C_j \binom{2n-2j-1}{n-j-1}$$

multiply this equation by x^n and sum over $n \geq 1$ we have

$$\begin{aligned}
&\sum_{j=0}^{n-1} \left(C_j \binom{2n-2j-1}{n-j-1} \right) x^n \\
&= \sum_j C_j x^{j+1} \sum_{n\geq 1} \binom{2n-2j-1}{n-j-1} x^{n-j-1} \qquad \text{(by (3.10))} \\
&= \sum_j C_j x^{j+1} \sum_n (-1)^{n-j-1} \binom{-n+j-1}{n-j-1} x^{n-j-1} \\
&= \sum_j C_j x^{j+1} \frac{1}{(1-x)^{n-j-1}} \\
&= \frac{x}{(1-x)^{n+1}} \sum_{j\geq 0} C_j x^j (1-x)^j \\
&= \frac{x}{(1-x)^{n+1}} \sum_{j\geq 0}^{\infty} C_j (x(1-x))^j \\
&= \frac{x}{(1-x)n+1} C(x(1-x)) \\
&= \frac{x}{(1-x)^{n+1}} \frac{1-\sqrt{1-4x(1-x)}}{2x(1-x)} \\
&= \frac{1-\sqrt{1-4x(1-x)}}{2(1-x)^{n+2}}
\end{aligned}$$

What we need is the $(n+1)^{th}$ coefficients of this function. Note that

$$\frac{1-\sqrt{1-4x(1-x)}}{2} = 1 + O(x^n)$$

for any $m \geq 0$. So the $(n-1)$ coefficient is the same as the $(n-1)$-th coefficient of $\frac{1}{(1-x)^{n+2}}$. It is

$$(-1)^{n-1} \binom{-(n+2)}{n-1} = (-1)^{n-1}(-1)^{n-1} \binom{2n}{n-1}$$

$$
\begin{aligned}
&= \binom{2n}{n-1} \\
&= \binom{2n}{n}\frac{n}{n+1}
\end{aligned}
$$

as desired. □

Proposition 3.1.3.

$$\sum_{k=0}^{n}(-1)^k\binom{n+k}{2k}C_k = 0$$

for all positive integers n.

Proof. We multiply the sum by x^n and sum over n. Then

$$
\begin{aligned}
\sum_{n\geq 0}\sum_{k=0}^{n}(-1)^k\binom{n+k}{2k}C_k x^n &= \sum_{k\geq 0}(-1)^k C_k x^k \sum_{n-k\geq 0}\binom{n+k}{n-k}x^{n-k} \qquad \{r := n-k\} \\
&= \sum_{k\geq 0}(-1)^k C_k x^k \sum_{r\geq 0}\binom{r+2k}{2k}x^r \qquad \text{(by (3.3))} \\
&= \sum_{k\geq 0}(-1)^k C_k x^k \frac{1}{(1-x)^{2k+1}} \\
&= \frac{1}{(1-x)}\sum_{k\geq 0}C_k\Big(\frac{-x}{(1-x)^2}\Big)^k \\
&= \frac{1}{1-x}C\Big(\frac{-x}{(1-x)^2}\Big) \\
&= \frac{1}{1-x}\frac{1-\sqrt{1-4(\frac{-x}{(1-x)^2})}}{2(\frac{-x}{1-x})^2} \\
&= \frac{1}{1-x}\frac{1-\frac{1+x}{1-x}}{2(-\frac{x}{(1-x)^2})} \\
&= \frac{-2x}{-2x} = 1
\end{aligned}
$$

From this we see that the generating function is identically 1. So all the coefficient with the index larger than 0 is 0. This is indeed the required statement. □

Proposition 3.1.4.

$$\sum_{k}(-1)^k\binom{n+k}{m+2k}C_k = \binom{n-1}{m-1} \qquad (m, n \geq 0)$$

Proof. Consider the generating function

$$\begin{aligned}
f(x) &= \sum_{n\geq 0}\Big(\sum_{k}(-1)^k\binom{n+k}{m+2k}C_k\Big)x^n\\
&= \sum_{k}(-1)^kC_kx^{-k}\sum_{n\geq 0}\binom{n+k}{m+2k}x^{n+k} \qquad \{r := n+k\}\\
&= \sum_{k}(-1)^kC_kx^{-k}\sum_{r\geq k}\binom{r}{m+2k}x^{r} \quad \text{(by (3.4))}\\
&= \sum_{k}(-1)^kC_kx^{-k}\frac{x^{m+2k}}{(1-x)^{m+2k+1}}\\
&= \frac{x^m}{(1-x)^{m+1}}\sum_{k}\frac{(-1)^kC_kx^k}{(1-x)^{2k}}\\
&= \frac{x^m}{(1-x)^{m+1}}\sum_{k}C_k\Big(\frac{-x}{(1-x)^2}\Big)^k\\
&= \frac{x^m}{(1-x)^{m+1}}.\frac{1-\sqrt{1-4(\frac{-x}{(1-x)^2})}}{2(\frac{-x}{1-x})^2}\\
&= \frac{-x^{m-1}}{2(1-x)^{m-1}}\Big(1-\sqrt{1+\frac{4x}{(1-x)^2}}\Big)\\
&= \frac{-x^{m-1}}{2(1-x)^{m-1}}\Big(1-\frac{1+x}{1-x}\Big) = \frac{x^m}{(1-x)^m}
\end{aligned}$$

so the coefficient of x^n in this last function is the original sum which is $\binom{n-1}{m-1}$, since the coefficient of x^{n-1} in $\frac{x^{m-1}}{(1-x)^{(m-1)+1}}$ is $\binom{n-1}{m-1}$. □

The main reason for looking at above identities came from a desire of proving the following identity [14] where there is no proof.

Proposition 3.1.5. ([14]) For $C_k, k = 1, 2, \ldots$ we have

$$\sum_{k=3}^{t-1}\binom{k-1}{k-2}C_{t-k}C_{k-2} = \binom{t-3}{1}C_{t-2}.$$

Proof. The left hand side is the coefficients of those from convolution of two series. To include c_0 we rewrite the LHS as follows:

$$\sum_{k=2}^{t}\binom{k-1}{k-2}C_{t-k}C_{k-2} - C_{t-2} - (t-1)C_{t-2} = \sum_{k=2}^{t}\binom{k-1}{k-2}C_{t-k}C_{k-2} - tC_{t-2}$$

Now we compute the first term by looking at the series

$$\sum_{t\geq 2}\sum_{k=3}^{t-1}\binom{k-1}{k-2}C_{t-k}C_{k-2}x^t$$

$$
\begin{aligned}
&= x^2\Big(\sum_{k\geq 2}\binom{k-1}{k-2}C_{k-2}x^{k-2}\Big)\Big(\sum_{t\geq k}C_{t-k}x^{t-k}\Big)\\
&= x^2\Big(\frac{1}{\sqrt{1-4x}}\Big)\Big(\frac{1-\sqrt{1-4x}}{2x}\Big)\\
&= x^2\sum_{n\geq 0}\binom{2n+1}{n}x^n = \sum_{n\geq 0}\binom{2n+1}{n}x^{n+2} \qquad \text{(by (3.7))}
\end{aligned}
$$

Since we want the to pick up the coefficient from the term x^t i.e $n = t-2$, the desired sum should be

$$\binom{2(t-2)+1}{t-2} = \binom{2t-3}{t-2}$$

Finally

$$
\begin{aligned}
&\sum_{k=2}^{t}\binom{k-1}{k-2}C_{t-k}C_{k-2}\\
&= \binom{2t-3}{t-2} - tC_{t-2}\\
&= \frac{(2t-3)(2t-4)!}{(t-2)!(t-1)!} - tC_{t-2}\\
&= (2t-3)C_{t-2} - tC_{t-2}\\
&= (t-3)C_{t-2}
\end{aligned}
$$

That is the proof of the desired identity. □

3.2. Relation between the Catalan number and the Chebyshev polynomials

Let us consider the following path graph on n nodes

$$
\begin{array}{ccccccccc}
\bullet & - & \bullet & - & \bullet & - & \cdots\bullet & - & \bullet\\
1 & & 2 & & 3 & & n-1 & & n
\end{array}
$$

Define the adjacency $n \times n$ matrix $A_n = (a_{ij})$ by $a_{ij} = 1$ if (i,j) is the edge and $a_{ij} = 0$ otherwise, that is,

$$
A_n = \begin{pmatrix}
0 & 1 & 0 & 0 & \dots & 0\\
1 & 0 & 1 & 0 & \dots & 0\\
0 & 1 & 0 & 1 & \dots & 0\\
\vdots & \vdots & \vdots & \ddots & \vdots & \\
0 & 0 & 0 & \dots & 1 & 0
\end{pmatrix}
$$

Let $P_n(\lambda)$ is the Characteristic polynomial of A. Then by determinant expansion by first row, we obtain the following three-term recurrence[2]

$$\begin{cases} P_{n+1}(\lambda) = \lambda P_n(\lambda) - P_{n-1}(\lambda) \\ P_0(\lambda) = 1 \quad P_1(\lambda) = \lambda \end{cases} \tag{3.11}$$

We recognize immediately the following fact: This recurrence is related to the Chebyshev polynomials of the second kind by variable change $U_n(\lambda) := P_n(2\lambda)$. U_n satisfies

$$\begin{cases} U_{n+1}(\lambda) = 2\lambda U_n(\lambda) - U_{n-1}(\lambda) \\ U_0 = 1 \quad U_1(\lambda) = \lambda \end{cases} \tag{3.12}$$

It is not difficult to show that U_n satisfies

$$U_n(\cos\theta) = \frac{\sin((n+1)\theta)}{\sin\theta}.$$

Now we compute the generating function for the sequence $\{P_0(\lambda), \ldots P_n(\lambda), \ldots\}$

$$F(x,y) = \sum_{n=0}^{\infty} P_n(x) y^n.$$

The recurrence relation for $P_n(x)$ yields

$$F(x,y) = xyF(x,y) - y^2F(x,y).$$

Solving for $F(x,y)$ we have the generating function

$$F(x,y) = \frac{1}{1 - xy + y^2}.$$

Next we will relate Chebyshev polynomial to Catalan numbers. Remember that the generating function $C(x)$ of the Catalan numbers satisfies the relation

$$C(x) = 1 + xC(x)^2$$

we can therefore iterate $C(x)$ in the following manner

$$\begin{aligned} C(x) &= \frac{1}{1 - xC(x)} \\ &= \frac{1}{1 - \frac{x}{1 - xC(x)}} \end{aligned}$$

[2] After I got the reviewer's advice [23] I noticed that three-term recursion is treated in this book.

$$= \frac{1}{1 - \frac{x}{1 - \frac{x}{1 - \frac{x}{1 - \frac{x}{1 - \dots}}}}}$$

The convergents to this continued fraction are defined by the recurrence

$$Q_0(x) = 1, \quad Q_n(x) = \frac{1}{1 - xQ_{n-1}(x)}, n \geq 1.$$

Assume for some polynomials $q_n(x)$ satisfying

$$\begin{cases} q_{n+1}(x) = q_n(x) - xq_{n-1}(x) \\ q_0(x) = q_1(x) = 1 \end{cases} \tag{3.13}$$

we have

$$Q_n(x) = \frac{q_{n-1}(x)}{q_n(x)}.$$

In particular $q_n(-1) = F_n$, the Fibonacci sequence, and $\frac{F_{n-1}}{F_n}$ are convergents to the continued fraction expression of golder ratio

$$\phi = C(-1) = \frac{1 + \sqrt{5}}{2.}$$

Associate $\{q_n(x)\}$ to the generating function

$$G(x, y) = \sum_{n=0}^{\infty} q_n(x) y^n$$

we have

$$G(x, y) = yG(x, y) - xy^2 G(x, y),$$

i.e., the generating function for $q_n(x)$ is

$$G(x, y) = \frac{1}{1 - y + x^2}$$

Make substitution $x = \frac{1}{u^2}, y = uv$ we then relate the generating function $F(u, v)$ to the sequence $\{p_n(x)\}$ and the generating function $G(x, y)$ for the sequence $q_n(x)$ as follows

$$G\left(\frac{1}{u^2}, uv\right) = F(u, v)$$

This implies that

$$p_n(u) = u^n q_n \left(\frac{1}{u^2}\right).$$

Since $q_n(x)$ are related to the convergents of Catalan's continued fraction and $p_n(x)$ are related to the Chebyshev polynomials of the second kind (which are related to the Chebyshev polynomials of the first kind), we obtain the relation between the Catalan number and Chebyshev polynomials.

Remark 3.2.1. The Chebyshev polynomials are orthogonal polynomials. They have wide applications ranging from classical function theory to numerical analysis and modern control theory and engineering. More about relation between the Catalan numbers and orthogonal polynomial systems are discussed in next section.

3.3. Some Hankel Determinants

Let

$$H_n^0 = \begin{pmatrix} C_0 & C_1 & C_2 & \dots & C_n \\ C_1 & C_2 & C_3 & \dots & C_{n+1} \\ C_2 & \dots & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_n & C_{n+1} & C_{n+2} & \dots & C_{2n} \end{pmatrix} \quad H_n^1 = \begin{pmatrix} C_1 & C_2 & C_3 & \dots & C_{n+1} \\ C_2 & C_3 & C_4 & \dots & C_{n+2} \\ C_2 & \dots & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{n+1} & C_{n+2} & C_{n+3} & \dots & C_{2n+1} \end{pmatrix}$$

be the Hankel matrices (of order 0 and 1, respectively) of the Catalan numbers $\{C_0, C_1, \dots\}$. In [11] it is shown, by a counting result for disjoint path system in acyclic directed graphs that

$$\det H_n^0 = \det H_n^1 = 1.$$

However, we will not follow this counting argument. Instead we give a matrix theoretic approach, in the same spirit as those in [4]. The reason is to use the orthogonal polynomials introduced above and avoid more graph theory due to the limitation of the volume of the current thesis.

Our starting point is the three-term recurrence. Given two sequences

$$\{s_0, s_1, s_2, \dots\}, \{t_1, t_2, \dots\} \quad \text{and } t_i \neq 0, \text{ for all } i.$$

Define the matrix $A = (a_{n,k})$ recursively by

$$\begin{cases} a_{0,0}(x) = 1 \quad a_{0,k} = 0 & (k > 0) \\ a_{n,k} = a_{n-1,k-1} + a_{n-1,k-1} & (n \geq 1) \end{cases} \tag{3.14}$$

($a_{ij} = 0$, when some index is negative). Explicitly we can see the Catalan number in this matrix

n/k	0	1	2	3	4	5	6
0	$\underline{1}$						
1	0	1					
2	$\underline{1}$	0	1				
3	0	2	0	1			
4	$\underline{2}$	0	3	0	1		
5	0	5	0	4	0	1	
6	$\underline{5}$	0	9	0	5	0	1

Then we would guess that $C_n = a_{2n,0}$. Obviously

Observation 3.3.1. *A is a lower triangular matrix with diagonal equal to 1*

Observation 3.3.2. *For $s_i = 0$, $t_i = 1$ $\forall i$, A includes the Catalan numbers in the first column at even position i.e $C_n = a_{2n,0}$*

The proof is postponed to later discussion.

Observation 3.3.3. *$a_{n+1,n} = s_0 + s_1 + \ldots + s_n$.*

Proof. For $n = 0$, we have

$$a_{1,0} = a_{0,-1} + s_0 a_{0,0} + t_1 a_{0,1} = s_0.$$

Assume the equality holds for $n = l - 1$, i.e., $a_{l,l-1} = s_0 + \cdots + s_{l-1}$. Then, for $n = l$,

$$\begin{aligned} a_{l+1,l} &= a_{l,l-1} + s_l a_{l,l} + t_{l+1} a_{l,l+1} \text{ (by the induction assumption)} \\ &= (s_0 + \cdots + s_{l-1}) + s_l \cdot 1 + t_{l+1} \cdot 0 = s_0 + \cdots + s_l \end{aligned}$$

as required. □

Given $t_1, \ldots, t_n$, we define $T_0 = 1$, $T_n = t_1 \cdot t_2 \cdots t_n$ for $(n \geq 1)$. We will show the following theorem which will be the key to prove $\det H_n^0 = \det H_n^1 = 1$. Denote $b_n := a_{n,0}$.

Theorem 3.3.4. *Let $A = (a_{n,k})$ be defined by (3.3.3). Then for all m, n the following equality holds*

$$\sum_k a_{m,k} a_{n,k} T_k = a_{m+n,0} = b_{m+n}$$

In matrix form, it is equivalent to $ATA^t = H$ where A^t is the transpose of A, $T = \text{diag}(T_0, T_1, \ldots, T_n, \ldots)$ and $H = (b_{i+j})_{i,j=0}^{\infty}$

Proof. For $n = 0$ and any m, we have

$$\sum_k a_{m,k} a_{0,k} T_k = a_{m,0},$$

since $a_{0,k} = 0$ except $a_{0,0} = 1$, and $T_0 = 1$. Now assume it is true for $n - 1$ and any m then

$$\begin{aligned} &\sum_k a_{m,k} a_{n,k} T_k \\ &= \sum_k (a_{n-1,k-1} + s_k a_{n-1,k} + t_{k+1} a_{n-1,k+1}) a_{m,k} T_k \end{aligned}$$

Changing index in the first term by $k-1$ to k and in the last $k+1$ to k together with $T_{k+1} = t_{k+1}T_k$ yield

$$\begin{aligned}
&\sum_k a_{m,k}a_{n,k}T_k \\
&= \sum_k a_{n-1,k}(a_{n,k+1}t_{k+1} + s_k a_{m,k} + a_{m,k-1})T_k \\
&= \sum_k a_{n-1,k}a_{m+1,k}T_{k+1} \text{ (by induction assumption)} \\
&= a_{n+m,0}.
\end{aligned}$$

Similarly we can show this formula for arbitrary n and use induction on m. □

Note that $\det(A) = 1$ we have $\det H_n = T_0T_1T_2\cdots T_n \neq 0$ provided that $T_i \neq 0$ for all i.

Corollary 3.3.4.1. The matrix A be defined recursively as above if and only if $ATA^t = H$ with $T_n \neq 0$ for all n,$T_0 = 1$. The sequences $\{s_0, s_1 \ldots, s_n, ...\}$ and $\{t_1, t_2, ...\}$ are given by

$$s_k = a_{k+1,k} - a_{k,k-1}, \qquad t_k = \frac{T_k}{T_{k-1}}.$$

Corollary 3.3.4.2. $b_n = a_{n,0}$ if and only if $\det H_n \neq 0$, $\forall n$, where $\{a_{n,0}\}$ are defined recursively by (3.14).

Since A is invertible, we consider $V = A^{-1} = (v_{ij})$ which is still a lower triangular matrix with diagonal equal to 1. We can prove that

$$\begin{cases} v_{0,0}(x) = 1 \quad v_{0,k} = 0 \quad (k > 0) \\ v_{n+1,k} = v_{n,k-1} - s_n v_{n,k} - t_n v_{n-1,k} \quad (n \geq 0) \end{cases} \tag{3.15}$$

Let $p_n(x) = \sum_{k=0}^n v_{n,k}x^k$. Then p_i satisfies

$$\begin{cases} p_0(x) = 1 \\ p_{n+1}(x) = (x - s_n)p_n(x) - t_n p_{n-1}(x) \quad (n \geq 0). \end{cases} \tag{3.16}$$

Therefore we obtain the following equivalent statements:

$$\begin{array}{ccc}
A \text{ is recursively by (3.14)} & \Leftrightarrow & ATA^Tt = H \\
\Updownarrow & & \Updownarrow \\
p_{n+1}(x) = (x - s_n)p_n(x) - t_n p_{n-1}(x) & & VHV^t = T
\end{array}$$

Then we have shown

Corollary 3.3.4.3. $\{p_n(x)\}$ satisfies (3.16) if and only if $VHV^t = T$ for some Hankel matrix.

What does $VHV^t = T$ mean in terms of the polynomial sequence $\{p_n(x)\}$? The answer can be found in the theory of orthogonal polynomial systems.

Definition 3.3.1. A sequence of real polynomial $\{p_n(x)\}$ with degree n, for all n, is said to be *orthogonal system* if there exists a linear functional $\mathbb{R}[x] \to \mathbb{R}$ and numbers $T_n \neq 0 \quad (n \geq 1), \quad T_0 = 1$ such that

$$L(p_m(x)p_n(x)) = \delta_{mn}T_n.$$

Suppose $VHV^t = T$ Which H being the Hankel matrix of the sequence $\{b_n\}$. Define

$$L : X^n \to b_n, \text{ for all } n.$$

Then

$$\begin{aligned} VHV^t = T &\Leftrightarrow \sum_{ik} v_{m,i}v_{n,k}b_{i+k} = \delta_{mn}T_n \\ &\Leftrightarrow L((\sum_i v_{m,i}x^i)(\sum_k v_{n,k}x^k)) = \delta_{mn}T_n \\ &\Leftrightarrow L(p_m(x)p_n(x)) = \delta_{mn}T_n \qquad \forall m, n \end{aligned}$$

This implies that $\{p_n(x)\}$ forms an orthogonal polynomial system. The converse can also be established easily. Hence we have shown

Corollary 3.3.4.4. A sequence $\{p_n(x)\}$ forms an orthogonal system if and only if

$$\begin{cases} p_{n+1}(x) = (x - S_n)p_n(x) - t_n p_{n-1}(x) \quad (n \geq 0) \\ p_0(x) = 1 \end{cases} \tag{3.17}$$

for some pair of sequences $\{s_k\}$,$\{t_k\}$.

For more studies in this direction we refer to the reference for example [2].

Example 3.3.1. Let $s_i \equiv 0$, $t_i \equiv 1$. Then $T_n = 1$, $\forall n$. Therefore

$$\det H_n = \det(b_{i+j})_{0<i.j\leq n} = 1.$$

by (3.16)

$$\begin{cases} p_{n+1}(x) = xp_n(x) - p_{n-1}(x) \\ p_0(x) = 1 \end{cases} \tag{3.18}$$

This was discussed in the previous subsection.

As we observed, not yet proved, that $a_{2n,0}$, generated by the three-term recurrence, are Catalan-numbers for lower values of n. Now we will show that this is indeed true in general. We are also going to look for the Catalan numbers in the sequence generated by the three-term recurrence to compute the determinants of the Hankel matrix $(C_{i+j})_{0\leq i,j\leq n}$ because we have already obtained the determinant of the Hankel matrix for $(b_{i+j})_{0\leq i,j\leq n}$. To this end we consider a special case of parameters s_k, t_k. Let $s_0 = a, s_k = s$ and $t_k = 1$ for all k. Then the three-term recursion is

$$\begin{cases} a_{n,k} = a_{n-1,k-1} + s_k a_{n-1,k} + a_{n-1,k+1}, & n \geq 1 \\ a_{0,0} = 1, \quad a_{0,k} = 0 \end{cases} \tag{3.19}$$

Consider the k-th generating function $A_k(x)$ associated with $\{a_{n,k}\}$ for $k \geq 0$. Let

$$A_k(x) := \sum_{n\geq 0} a_{n,k} x^n.$$

By the recurrence (3.19), for $k \geq 1$,

$$\begin{aligned} A_k(x) &= \sum_{n\geq 0}(a_{n-1,k-1} + s a_{n-1,k} + a_{n-1,n+1})x^n \\ &= x \sum_{n-1\geq 0} a_{n-1,k-1}x^{n-1} + sx \sum_{n-1\geq 0} a_{n-1,k}x^{n-1} + x \sum_{n-1\geq 0} a_{n-1,k+1}x^{n-1} \\ &= x(A_{k-1}(x) + sA_k(x) + A_{k+1}(x)) \quad \text{and} \\ A_0(x) &= x(aA_0(x) + A_1(x)) + 1. \end{aligned}$$

So we have the following equations

$$\begin{cases} A_k(x) = x(A_{k-1}(x) + sA_k(x) + A_{k+1}(x)), & k \geq 1, \\ A_0(x) = x(aA_0(x) + A_1(x)) + 1. \end{cases} \tag{3.20}$$

Now we claim that $A_k(x) = f(x)^k A_0(x)$ where $f(x)$ is the generating function satisfying

$$f(x) = x(1 + sf(x) + f(x)^2).$$

By (3.20) we have

$$f(x)^k A_0(x) = x(f(x)^{k-1}A_0(x) + sf(x)^k A_0(x) + f(x)^{k+1}A_0(x))$$

That is $f(x)^k$ satisfied the recursion (3.20). So we have found solution

$$A_k(x) = f(x)^k A_0.$$

It remains to determine the function $A_0(x)$. From the definition for the generating function $f(x)$ above we can easily find that

$$f(x) = \frac{1 - sx - \sqrt{1 - 2sx + (s^2-4)x^2}}{2x}.$$

Substituting the formula into the relation to $A_0(x)$

$$A_0(x) = x(aA_0(x) + f(x)A_0(x)) + 1$$

gives the

$$A_0(x) = \frac{1 - (2a - s)x - \sqrt{1 - 2sx + (s^2 - 4)x^2}}{2(s - a)x + 2(a^2 - as + 1)x^2}.$$

Hence we have proved

Proposition 3.3.1. Let $b_n(a, s) := b_n = a_{n,0}$ are the numbers generated by the (3.20) corresponding to the sequences. The generating function for the sequence of $b_n(a, s)$ is given by

$$A_0(x) = \frac{1 - (2a - s)x - \sqrt{1 - 2sx + (s^2 - 4)x^2}}{2(s - a)x + 2(a^2 - as + 1)x^2}$$

Remark: The numbers $b_n(a, s)$ are called Catalan-like numbers, introduced by Aigner [1]. Note that we only work on the Catalan number's so we do a special case.

Now it is apparent that

1. For $a = s = 0$, the generating function is $A_0(x) = \dfrac{1 - \sqrt{1 - 4x^2}}{2x^2} = A_0(x^2)$, resulting the numbers

   $$\{C_0, 0, C_1, 0, C_2, 0, C_3 \ldots\}.$$

   i.e. the observation 2 is proved.

2. For $a = 1, s = 2$ the generating function is $A_0(x) = \dfrac{1 - \sqrt{1 - 4x}}{2x} =: C(x)$, the generating function of the Catalan numbers so the resulting sequence is

   $$\{C_0, C_1, C_3, ...\}.$$

For other special sequences depending on other choices of s and a, we refer to [1].

Note that $\{C_n\}$ (i.e. $\{a_{n,0}\}$) is generated by

$$\begin{aligned} &a_{n,k} = a_{n-1,k-1} + 2a_{n-1,k} + a_{n-1,k+1}, \quad n \geq 1, k \geq 1 \\ &a_{n,0} = a_{n-1,0} + a_{n-1,1}, \quad n \geq 1, k = 0 \\ &a_{0,0} = 1, \quad a_{0,k} = 0 \end{aligned}$$

As before we write this in matrix form

$$AA^t = (C_{i+j})_{i,j \geq 0}$$

and truncate it at n, then it takes the form

$$A_n A_n^t = (C_{i+j})_{0 \le i,j \le n}.$$

Obviously

$$\det((C_{i+j})_{0 \le i,j \le n}) = 1 \quad \forall n \ge 0,$$

since the matrix A_n is triangular matrix with diagonal equal to 1.

Next we try to compute $\det((C_{i+j+1})_{0 \le i,j \le n})$. Let σ be the shift operator with the matrix representation

$$\begin{pmatrix} 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix}$$

From the above matrix form for our iteration, it is clear that

$$\sigma(A)A^t = (C_{i+j+1})_{i,j \ge 0}$$

Again truncate this at n we get

$$\det((C_{i+j+1})_{0 \le i,j \le n}) = \det((\sigma(A))_n).$$

By inspection of of the recurrence we have

$$\sigma(A) = AJ$$

where $J = \begin{pmatrix} 1 & 1 & & & \\ 1 & 2 & 1 & & 0 \\ & 1 & 2 & 1 & \\ & 0 & \ddots & \ddots & \ddots \end{pmatrix}$

So

$$\det((\sigma(A))_n) = \det(J_n) = 1.$$

Hence we have proved, as in M.E Mays and J. Wojchiechowski or [1],

Theorem 3.3.5. The determinants of the Hankel matrix of order 0 $(C_{i+j})_{0 \le i,j \le n}$ and of order 1 resulted from the Catalan numbers $(C_{i+j+1})_{0 \le i,j \le n}$ are 1, i.e.

$$\det((C_{i+j})_{0 \le i,j \le n}) = \det((C_{i+j+1})_{0 \le i,j \le n}) = 1, \qquad \text{for all } n \ge 0.$$

In fact the Catalan numbers are uniquely determined by these two determinants, a new way of presenting the Catalan numbers. See [1] for details.

Remark: We can further compute the determinants of higher order Hankel matrices. For example by observation $\sigma(A)(\sigma(A))^t = (C_{i+j+2})_{i,j \ge 0}$. However we leave it as it is.

Chapter 4

RNA secondary structure

4.1. RNA folding problem

This section deals with predicting the secondary structure of the RNA by transforming sequences of biopolymers into spatial molecular structure. We would like to find the the optimal structure using dynamics programming as done in [9]. To this end we would need a scoring scheme, then we need to create a structure with minimum free energy. We achieve our goal by using two known algorithms such as the *Nussinov* and *Zucker*.

Let us start with the Nussinov's algorithm, the recursion formula for this problem is first described by Nussinov in 1978 and the name of algorithm is derived from him. the algorithm is based of calculation the best substructure for the subsequences till it finds the structure of the whole structure of the whole sequence. The two cases to get from one step to the next in the recursion are based eiher by newly added base to unpaired or is based with base k. For the latter case the base pair (i, j) divided the problem into subproblem which can be then recursively solved in the same way.

We calculate the minimum free energy that is interested in the sequence corresponding to this particular energy a helper matrix is filled to back-tracking over the sequence we find the codes bellow

$$E_{ij} = min\Big\{E_{i+1,j}, min_{k,\sqcap_{ik}=1}\{E_{i+1,k-l} + E_{k+1,j} + \beta_{ik}\}\Big\}$$

Where E_{ij}: minimum energy of subsequence $i, ..., j$ β_{ik} energy contribution of pair (i, j) $\sqcap_{ik}$ is 1 if the bases i and j can pair and 0 otherwise. The recursion formula for Nussinov algorithm along with a graphical depiction of how it works.

The helper array K_{ij} is filled when the recursion that holds the optimal secondary structure when k is paired with i for a sub-sequence i, j. If i is unpaired in the optimal structure structure $K_{ij} = 0$ The Zucker algorithm is also the most important for this section.

The **Zucker algorithm** is a variance of *Nusssinov* which includes stacking energy to calculate the RNA structure. Some modern RNA folding algorithm for RNA structure prediction. In Zucker algorithm we have four cases to deal with: The procedure requires four matrices such as: F_{ij} contains the free energy of the overall optimal structure of the sub-sequence x_{ij}. The newly added base can be unpaired or paired. For the latter case, we introduce the helper Matrix C_{ij} that contains the free energy of the optimal substructure of x_{ij} under the constraint that i and j are paired. This structure closed by a base-pair can either be a hairpin, an interior loop or a multi-loop as given in [13].

μ_{ij} holds the free energy of the optimal structure of x_{ij} under the constraint x_{ij} that i and j is part of a multi-loop with at least one component. μ^1_{ij} holds the free energy of the optimal structure of x_{ij} under the constraint that x_{ij} is part of a multi-loop and has exactly one component closed by pair $(i;k)$ with $i < k < j$. The ideal behind is to decompose a multiloop into in two arbitrary parts of which the first is a multi-loop with at least one component and the second a multi-loop with exactly one component and starting with a base-pair. These two parts corresponding to μ and μ^1 can further be decomposed into substructures that we already know, i.e. unpaired intervals, substructures closed by a base-pair,or multi-loops. In reality, in room temperature RNA is not actually in one single state, but rather it varies in a Thermodynamic ensemble of structure. Base pairs can break their bonds quite easily, and although we might find an absolute optimum in terms of free energy, it might be the case that there is another sub-optimal structure which is very different from what are predicted and has an important role in the cell also is stated by [13].

To fix the problem we can calculate the base pair probabilities to get the ensemble of structures, and t en we would have a much better idea of what the RNA structure probably looks like. In order to do this, we utilize the Boltzman factor:

$$\text{prob}(S) = \frac{\exp(-\triangle G(S)/RT)}{Z}$$

Which gives us the probability of a given structure, in a thermodynamic system. We need to normalize the temperature using the partition function Z,which is the weighted sum of all structures, based on their Boltzman factor:

$$Z = \sum_{S} \exp(-\triangle G(S)/RT)$$

We can also represent this ensemble graphically, using a dot plot to visualize the base pair probabilities. To calculate the specific probability for a base pair $(i;j)$, we need to calculate the partition function, which is given by the following formula :

$$p_{ij} = \frac{\hat{Z}_{ij} Z_{i+1,j-1} \exp(-\beta_{ij})/RT}{Z}.$$

To calculate Z we use the recursion similar to the Nussinovs Algorithm The inner partition function is calculated using the formula:

$$Z_{ij} = Z_{i+1,j} + \sum_{i+1<k<j} Z_{i+1,k-1} Z_{k+1,j} \exp(-\beta_{ij})/RT$$

With each of the additions corresponding to a different split as founded in [9]

4.2. The evolution of RNA

Let us consider two methods such as dynamic programming and probabilistic method, we start by probabilistic as seen in [21] there are different ways that can be used in evolution of RNA we have: It is very interesting to know about its evolution structure, since it unveils valuable data, and can also give us hints to treat our structure predictions. When we look into functionally important RNA over time, although their nucleotides have changed at some parts, but the structure is not changed. In RNA there are a lot of compensatory or consistent mutations, in a consistent mutation the structure does not change. There are the mutation between U and A which makes the pairs AU and the mutation of C and G then makes the CG pairs. To do this we may calculate the probability and compare it with of two base pair structures agreeing randomly to be paired the information content is calculated using the formula bellow

$$M_{ij} = \sum_{xy} f_{ij}(XY) \log \frac{XY}{f_i(X) f_j(Y)}$$

if we normalize these probabilities, we can plot it in a $3D$ model and track the evolutionary signatures.

4.3. Probabilistic folding

It is very hard problem of finding RNA coding sequence inside the genome there are the way to do it. One way is to combine the thermodynamic stability information with normalized RNA fold score and then we can do a Support Vector Machine classification and compare the thermodynamic stability of the sequence to some random sequence of the same GC content and the same length by combining it with the evolutionary measure and see if the RNA is more conserved or not. This gives us an idea if the genomic sequence is actually coding an RNA [9].

4.4. Thermodynamics folding

The thermodynamic approach is a good way of folding the RNAs since it use the algorithm solution like Zuker's mfoldprogram and implement an efficient recursive calculation of the minimum free energy configuration under certain assumptions, these energy calculation is assumed to decompose into the sum over independent loop energies [16]

4.5. Representation of Secondary Structure

From Waterman's definition [22]we have seen that the are many ways of representing secondary structure as shown in 2.3.1 section. There are the rules that governing sequence of matching brackets and dots. A secondary structures implies that each branch of the corresponding Υ tree representation has at least one terminal half-vertex, or equivalently,each matching pair of brackets contains at least one dot. The number of unpaired positions is at least 3, implying at least 3 dots within each pair of matching brackets .

From the combinatorial point of view it makes perfect sense to consider the general problem with a minimum number $m \geq 0$ of unpaired vertices in each hairpin loop. In fact, for $m = 0$ one recovers three well known Motzkin families [20].

For Some application it is useful to work with simplified representations by comparing RNA secondary structures using tree comparison *A* tree *T* is obtained by denoting a stack by single vertex. In terms of the representation Υ this means that each vertex of degree 2 not carrying a half-vertex is merged with its son and then the half-vertices are removed thus The number of vertices in *T* is then just the number of stacks in *S*, the number of components of *S* coincides with the number of sons of the root in *T*. An alternative "coarse grained" representation of a secondary structures is the homeomorphically irreducible tree *H* corresponding to Υ which is obtained by removing all vertices of degree 2 and all half-vertices. Again the number of components of *S* equals the number of sons of the root. Waterman's degree ω coincides with the height of *H* [22].

4.5.1. Secondary Structures of a Given Order

The Secondary Structures of a Given Order is founded in [8], as follows let start by setting $D_n(c, \omega)$ as the number of secondary structures with c components and order ω. Furthermore let $D^*_n(\omega)$ be the number of structures which yield a structure of order ω when enclosed by an additional base pair.

The recursion holds as following

$$D_{n+1}(c,\omega) = D_n(c,\omega) + \sum_{k=m}^{n-1}\{D_n^*(\omega)\sum_{l=0}^{\omega-1} D_{n-k-1}(c-1,l) +$$
$$D_{n-k-1}(c-1,\omega)\sum_{l=0}^{\omega-1} D_k^*(l) + D_k^*(\omega)D_{n-k-1}(c-1,\omega)\}$$

Where

$$D_n(0,0) = 1, D_n(0,d) = D_n(c,0)$$

for $n \leq m+1$ a structure with a base pair $1 \equiv k+2$ has order d and c components if and only if either the bracketed part has order ω and the tail has a order at most ω and $c-1$ components or the bracketed part has a degree smaller than ω and the tail has $c-1$ components and the order ω. Where $D_n^*(\omega)$ is obtained by

$$D_n^*(0) = 0$$

$$D_n^*(0) = 1 + D_n(1,1)$$

$$D_n^*(\omega) = D_n(1,\omega) + \sum_{l=2}^{\infty} D_k(l,\omega-1), \omega \geq 2$$

While for $n \leq m$ we have $D_n^*(\omega) = 0$ There is no structure of order 0 with a bracket in it; order one is obtained by either bracketing the open structure or by bracketing a structure with a single component and order 1. If the bracketed part has only a single components its order is preserved by adding a terminal bracket. If it consists of more than one components, the addition of the multiloop increases the order by one.

Theorem 4.5.1. For any finite order ω there is a positive constant ϵ such that

$$\lim_{n\to\infty} \frac{\tilde{D}_n(\omega-1)e^{\epsilon n}}{\tilde{D}_n(\omega)} = 0 \tag{4.1}$$

Proof. To prove this we need to use the generating function

$$\triangle_{\omega} = \sum_{n=0}^{\infty} \tilde{D}_n(\omega)x^n \qquad \triangle_{\omega^*} = \sum_{n=0}^{\infty} D_{\dot{n}(\omega)}x^n \qquad \triangle_{\omega}^{'} = \sum_{n=0}^{\infty} D_n(1,\omega)x^n \tag{4.2}$$

where $\tilde{D}_n(\omega)$ be the number of structures with the given order, by using the recursion formula bellow

$$\tilde{D}_{n+1}(\omega) = \tilde{D}_n + \sum_{k=m}^{n-1}\left\{D_k^*\sum_{l=0}^{\omega-1}\tilde{D}_{n-k-1}(l) + \tilde{D}_{n-k-1}(d)\sum_{l=0}^{\omega} D_k^*(l)\right\}$$

$$
\begin{aligned}
D_k^*(\omega) = \tilde{D}_k(\omega-1) + D_k(1,\omega) - D_k(1,\omega-1) \qquad & n \geq m+2 \\
D_{n+1}(1,\omega) = D_n(1,\omega) + \sum_{k=m}^{n-1} D_k^*(\omega)\tilde{D}_n(0) = 1 & \\
\tilde{D}_n(\omega) = 0 \qquad for \quad \omega \geq 1 \qquad & n \leq m+1
\end{aligned}
$$

yields the following system of coupled functional equations for the above generating functions

$$
\begin{aligned}
\triangle_\omega = x \ \triangle_\omega + x^2 \ \triangle_\omega^* \sum_{i=0}^{\omega-1} \triangle_i + x^2 \ \triangle_\omega \sum_{i=0}^{\omega} \triangle_i^* & \\
\triangle_\omega^* = \triangle_{\omega-1} + \triangle_\omega^{'} - \triangle_{\omega-1}^{'} \qquad \omega \geq 2 & \\
\triangle_\omega^{'} = x \ \triangle_\omega^{'} + x^2 \ \triangle_\omega^* \ \frac{1}{1-x} &
\end{aligned}
$$

For $\omega = 0$ we have $\triangle_0 = \frac{1}{1-x}$ and for $\omega = 1$ we find the explicitly

$$
\begin{aligned}
\triangle_1^* \ (x) \frac{1-x}{1-2x} x^m & \\
\triangle_1 \ (x) = \frac{x^{m+2}}{1-x} \frac{1}{1-2x-x^{m+2}} & \qquad (4.3)
\end{aligned}
$$

By eliminating $\triangle_\omega^{'}$ we find explicitly for $\omega \geq 2$

$$
\begin{aligned}
\triangle_\omega^* = \frac{(1-x)^2}{1-2x} \ \triangle_{\omega-1} - \frac{x^2}{1-2x} \ \triangle_{\omega-1}^* & \\
\triangle_\omega = \frac{x^2 \ \triangle_\omega^* \ \sum_{i=0}^{\omega-1} \ \triangle_i}{1-x-x^2 \sum_{i=0}^{\omega} \ \triangle_i^*} & \qquad (4.4)
\end{aligned}
$$

we find the result of $\triangle_\omega \ (x)$ by using Mathematica [8]. □

We need to know the total number like unpaired , the paired, vertices, stacks and the loops bases of RNA secondary structure so that we can predict the folding that can exist as seen [8]. Let us denotes U_n the total unpaired obtained by summing over k an unpaired base to each structure on n digits plus them the S_n.

$$
\begin{aligned}
U_{n+1} = (U_n + S_n) + \sum_{k=m}^{n-1} [S_k U_{n-k-1} + S_{n-k-1} U_k], \qquad & n \geq m+1 \qquad (4.5) \\
U_n = n, \qquad & n \leq m+1
\end{aligned}
$$

Let us denote the total number of base pairs by P_n we have the following recursion

$$
P_{n+1} = P_n + \sum_{k=m}^{n-1} S_k P_{n-k-1} + S_{n-k-1}(P_k + S_k), \qquad (4.6)
$$

$$n \leq m+1, P_n = 0$$

The same way the number of vertices I_n is given by

$$I_{n+1} = I_n + \sum_{k=m}^{n-1} S_k I_{n-k-1} + S_{n-k-1} \qquad I_n = 0 \qquad n \leq m+1 \tag{4.7}$$

The number of stack in the set of structure on N_{n+1} digits of all stacks and add all the number of structure with new introduced base pair of all stacks we have

$$N_{n+1} = N_n + \sum_{k=m}^{n-1} \Big\{ S_k N_{n-k-1} + S_{n-k-1}(N_k + S_k) \Big\} - \sum_{k=m+2}^{n-1} S_{K-2} S_{n-k-1}, \tag{4.8}$$

$$n \geq m+1, N_n = 0, n \leq m+1$$

Let us consider $Q_n(b)$ denote the number of loop with b u unpaired digits in the set of all secondary structures we have

$$Q_{n+1}(b) = Q_n(b) + \sum_{k=m}^{n-1} \Big\{ Q_{n-k-1}(b) S_k + S_{n-k-1}[Q_k(b) + E_k(b)] \Big\}, \tag{4.9}$$

$$n \leq m+1, b > 0 \quad Q_n(b) = 0, n \geq m+1$$

where b unpaired vertices remains unchanged and additionally each with exactly b external vertices within the new base pair gives rise to an additional loop with b unpaired digit [8]. The RNA secondary structure with certain given in this section help to deduce the relationship between them and Catalan number.

4.6. Catalan number and RNA Secondary structure

The Catalan number and the RNA secondary structure have quite relationship. An RNA molecule is the sequence of four possible letters A, C, G, U connected by backbone and is called RNA primary structure this can be pairing by two nucleotide in the following way, according to *Watson-Crick* A pairs with U and G pairs with C but in the *Wobble* base pairing G forms pair with U [14]. Any sequence of RNA can be represented as Catalan numbers in terms of parenthesis and dots as shown in the example bellow

Example 4.6.1. The sequence $GAGAGCCUUUGGACCUCA$ can be represented in parenthesis and dots like (((..((...))..))).

This can be mathematically denoted as: An RNA sequence of length n is assumed as sequence of n points each point i is connected to $i-1$ and to

$i+1$ such that $1 < i < n$. The notation $i.j$ stands for the nucleotide i is pairing with j and $i < j$ Therefore an RNA structure is a set S of base pairs $i.j$ with $1 < i < j < n$ such that $i_1.j_1,\ i_2.j_2 \in S : i_1 = i_2 \Leftrightarrow j_1 = j_2$ while S is called *secondary structure* if for all $i_1.j_1,\ i_2.j_2 \in S$ they are nested or disjoint [14].

The relationship between Catalan numbers and RNA secondary structure can be given as: Let us consider counting non crossing matching of base pair edges. Let C_n denote the number of non crossing perfect matching in the complete graph L_{2n}. We know that $C_0 = C_1 = 1$ for the general case n we say that the nodes L_{2n} are labeled with the positive integer from 1 to $2n$. there are $2L-2$ nodes on one side and $2n-2L$ nodes on other side, we can form different ways $C_{k-1}.C_{n-k}$ a perfect matching remaining codes of L_{2n}. If we let m varies over all possible $n-1$ choices of even numbers between 1 and $2n$ then we have the recurrence relation $C_n = \sum_{k=1}^{n} C_{k-1}C_{n-k}$. Thus the resulting number C_n counting non crossing perfect matching in L_{2n} are called The Catalan number found in $rosalind.info/problems/cat/$

Based on the theorems 4.6.1 and 4.6.2 we show relationship between the recursion formula and Catalan number as given in [8]

Theorem 4.6.1. Let

$$y(x) = \sum_{n=0}^{\infty} y_n x^n \tag{4.10}$$

be of the form

$$y(x) = \beta(x) + \sum_{k} g_k(x)\Big(1 - \frac{x}{\alpha}\Big)^{\omega_k}$$

where β, g_k are analytic on a circle larger than the circle of convergence of $y(x)$, ω_k real but not a non-negative integer. Suppose y has only a single singularity at $x = \alpha$ Denote by ω the smallest exponent ω and by $g(x)$ the corresponding analytic factor. Then

$$y_n \sim \frac{g(\alpha)}{\Gamma(-\omega)} n^{-1-\omega} \Big(\frac{1}{\alpha}\Big)^n \tag{4.11}$$

and the theorem

Theorem 4.6.2. The total number of structures with b base pairs is

$$H_n(b) \sim \frac{1}{(b+1)!b!} n^{2n} \tag{4.12}$$

also the recursion formula

$$h_b(x) = \eta_b(x)\frac{1}{1-x}\Big(\frac{x}{1-x}\Big)^{2b} \tag{4.13}$$

where

$$\eta_b(x) = \sum_{k=1}^{b} \eta_k(x)\eta_{p-k-1}(x) + x^m\eta_{b-1} \tag{4.14}$$

The theorem 4.6.1 assures that

$$H_n(b) \sim \frac{\eta_b(1)}{\Gamma(2b+1)}n^{2b}$$

since $\eta_0(1) = 0$ the recursion 4.14 becomes the well known recursion for the Catalan number with

$$\eta_b(1) = C_b = \frac{1}{b+1}\binom{2b}{b}$$

Also we can see the relationship between Catalan number and RNA secondary structure. Let us consider the generating function of the form

$$\begin{aligned} \nu_b(x) &= \mu_b(x)\frac{1}{(x+1)^b}\frac{1}{(x-1)^{3b+1}} \\ \zeta_b &= \xi_b(x)\frac{1}{(x+1)^b}\frac{1}{(x-1)^{3b+1}} \end{aligned} \tag{4.15}$$

where $\mu_b(x)$ and $\xi_b(x)$ are polynomials, the theorem 4.6.1 thus yields

$$N_n(b) \sim \frac{1}{2^b}\frac{\mu_b(1)}{\Gamma(3b+1)}n^{3b} \tag{4.16}$$

where $\mu_b(1)$ and $\xi_b(1)$ fulfill the recursions

$$\xi_{b-1}(1) = \mu_{b-1}(1) \qquad \mu_n(1) = \sum_{l=1}^{b} \xi_l(1)\mu_{b-1}(1) = \sum_{l=0}^{b-1} \mu_l(1)\mu_{b-l-1}(1) \tag{4.17}$$

so the coefficient $\mu_b(1)$ coincide with Catalan numbers. The following theorem also gives that relationship.

Theorem 4.6.3. Let C_n be the number of secondary structures for n points. Then $C_1 = C_2 = 1$, and for $n > 2$, C_n satisfies

$$C_n = C_{n-1} + \sum_{k=1}^{n-1} C_k C_{n-k-1} \tag{4.18}$$

Where $C_0 = 1$

This theorem coincides exactly with the first property of Catalan numbers seen in section 1.2, [14] where the authors claim the following four identities are true for the Catalan number C_k for $k = 1, 2, ...$ and give the crucial relationship between Catalan numbers and RNA secondary structure seen in [14].

1)

$$\sum_{k=3}^{t-1}\binom{k-1}{k-2}C_{t-k}C_{k-2}=\binom{t-3}{1}C_{t-2}$$

2)

$$\begin{aligned}&\sum_{k=3}^{t-1}\binom{k-1}{k-2}C_{t-k}\sum_{l=3}^{k-1}\binom{l-1}{1-2}C_{k-l}C_{l-2}\\&-\sum_{k=3}^{t-1}\binom{k-1}{k-3}\sum_{i=k+2}^{t}C_{t+l-i}C_{i-l-k}C_{k-2}=\binom{t-3}{2}C_{t-2};\end{aligned}$$

3)

$$\begin{aligned}&\sum_{k=3}^{t-1}\binom{k-1}{k-2}\binom{k-3}{3}C_{t-2}C_{t-k}C_{k-2}\\&-\sum_{k=3}^{t-1}\binom{k-1}{k-2}\binom{k-3}{1}\sum_{i=k+2}^{t}C_{t+1-i_1}C_{i_1-1-k}C_{k-2}\\&+\sum_{k=3}^{t-1}\binom{k-1}{k-4}\sum_{i_2=k+3}^{t}\sum_{i_1=k+3}^{i_2}C_{t+1-i_2}C_{i_2+1-i_1}C_{i_1-k-2}C_{k-2}=\binom{t-3}{3}C_{t-2};\end{aligned}$$

4)

$$\begin{aligned}&\sum_{k=3}^{t-1}\binom{k-1}{k-2}\binom{k-3}{3}C_{t-k}C_{k-2}\\&-\sum_{k=3}^{t-1}\binom{k-1}{k-3}\binom{k-3}{2}\sum_{i=k+2}^{t}C_{t+1-i_1}C_{i_1-1-k}C_{k-2}\\&+\sum_{k=3}^{t-1}\binom{k-1}{k-4}\binom{k-3}{1}\sum_{i_2=k+3}^{t}\sum_{i_1=k+3}^{i_2}C_{t+1-i_2}C_{i_2+1-i_1}C_{i_1-k-2}C_{k-2}\\&-\sum_{k=3}^{t-1}\binom{k-1}{k-5}\sum_{i_3=k+4}^{t}\sum_{i_2=k+4}^{i_3}\sum_{i_1=k+4}^{i_2}C_{t+1-i_3}C_{i_3+1-i_2}C_{i_2+1-i_1}C_{i_1-k-2}C_{k-2}\\&=\binom{t-3}{4}C_{t-2};\end{aligned}$$

This can generalised as

$$
\begin{aligned}
&A_k(t, j_1, j_2, \ldots, j_{k-1}) = \sum_{j_1=3}^{t-1} \binom{j_1-1}{j_1-2} C_{t-j_1} A_{k-1}(t, j_1, j_2, \ldots, j_{k-2}) \\
&- \sum_{j_1=3}^{t-1} \binom{j_1-1}{j_1-3} \sum_{i_1=j_1+2}^{t} C_{t+1-i_1} C_{i_1-1-j_1} A_{k-2}(t, j_1, j_2, \ldots, j_{k-3}) \\
&+ \sum_{j_1=3}^{t-1} \binom{j_1-1}{j_1-4} \sum_{i_2=j_1+3}^{t} \sum_{i_1=j_1+3}^{i_2} C_{t+1-i_2} C_{i_2+1-i_1} C_{i_1-k-2} A_{k-3}(t, j_1, j_2, \ldots, j_{k-4}) \\
&+ \ldots + (-1)^k \sum_{j_1=3}^{t-1} \binom{j_1-1}{j_1-k} \sum_{i_{k-2}=j_1+k-1}^{t} \sum_{i_{k-3}=j_1+k-1}^{i_{k-2}} \\
&\ldots \sum_{i_1=j_1+(k-1)}^{i_2} C_{t+1-i_{k-2}} C_{i_{k-1}+1-i_{k-3}} C_{i_2+1-i_1} C_{i_1-k-2} C_{k-2} \ldots C_{i_2+1-i_1} C_{i_1-j_1-2} A_1(j_1) \\
&+ (-1)^{k+1} \sum_{j_1=3}^{t-1} \binom{j_1-1}{j_1-(k+1)} \sum_{i_{k-2}=j_1+k}^{t} \sum_{i_{k-3}=j_1+k}^{i_{k-2}} \cdots \sum_{i_1=j_1+k}^{i_2} C_{t+1-i_{k-2}} C_{t_{k-1}+1-i_{k-3}} \\
&\ldots C_{i_2+1-i_1} C_{i_1-j_1-2} C_{j_1-2} \\
&= \binom{t-3}{k} C_{t-2};
\end{aligned}
$$

Note that the proof of first relation is given in 3.1.5.

By using these new properties of Catalan number will help us to achieve the goals of deciphering RNA and represent them in $3D$.

Chapter 5

Conclusions

We have studied some nice properties of the Catalan numbers in this thesis. It covers material from enumerative combinatorics, function theory and applications in RNA second structure. In particular, we found that generating functions are powerful in proving some difficulty-looking sums, which appear in counting subjects. Many parts of this thesis do not involve combinatoric arguments. We tried to use classical function theory and linear algebra to prove and illustrate theorems. We gave proofs to some identities whose proofs could not be found in the literature to our knowledge. We recommend the researchers to continue the proof of the rest of the identities mentioned in the end of the preceding chapter because we do not prove all due to the limitation of time and volume. We recommend also the researchers to give a proper code to do a concrete problem in protein folding so that they reply to the problem of theoretical biophysics face for .

Bibliography

[1] M. Aigner, (1999) *Catalan-like numbers and deterinants*, J. Combinatorial Theory, Series A **87**,33–51 .

[2] N. Akhiezer (1965) *The classical moment problem* , Olivier Boyd, .

[3] P. J. Cameron (2010) *Notes on Counting: An Introduction to Enumerative Combinatorics* School of Mathematical Sciences Queen Mary, University of London,.

[4] J. Cigler (2009) *A simple approach to some Hinkel determinants* ArXiv:0902.1650,

[5] T. Došlić and Darko Veljan (2007) *Secondary structures, plane trees and Motzkin numbers*, University of Zagreb, pp. 163-172

[6] W. Fontana, T. Griesmacher, W. Schnabl, P. F. Stadler, and P. Schuster,(1991) *Statistics of Landscapes Based on Free Energies Replication and Degradation Rate Constants of RNA Secondary Structures*.Mh. Chern., 122, pp. 795-819.

[7] C.E. Heitsch (2012), *Combinatorics on Plane Trees, Motivated by RNA Secondary Structure Configurations* University of Wisconsin, Madison,

[8] I.L. Hofacker, P. Schuster and P.F. Stadler,(1994), *Combinatorics of RNA Secondary Structures*, SANTA FE INSTITUTE.

[9] M. Kellis,(2012), *Computational Biology: Genomes, Networks, Evolution* MIT, 132-139

[10] Mathworks, `http://www.mathworks.com/help/bioinfo/examples/predicting-and-visualizing-the-secondary-structure-of-rna-sequences.html`

[11] M.E Mays and J. Wojchiechowski,(2000), *A determinant property of Catalan numbers* Discrete Mathematics, vol.211.Issunes 1-3,p.125-133.

[12] P.J. Larcombe and P.D.C. Wilson,(1998), *On the tail of the Catalan sequence* Mathematics today 34, pp.114-117

[13] W. Stefan,(2011), *Graph Representations and Algorithms in Computational Biology of RNA Secondary Structure* Structural Analysis of Complex Networks, pp. 421-437

[14] I.S. Rakhimov and K.A. M. Atan,(2014), *On some new properties of Catalan numbers* Serdang, Selangor Darul Ehsan, Malaysia.Vol. 9

[15] W.R. Schmitt W. R. AND M.S. Waterman,(1992) *Plane Trees and RNA Secondary Structure* Preprint.

[16] P. Schuster, W. Fontana, P. Stadler, and I. Hofacker,(1994), *From sequences to shapes and back: a case study in RNA secondary structures*,Proc R Soc Lond B Biol Sci, 255(1344):279-284,

[17] R. Simion,(2000), *Non crossing partitions* Discrete Math., 217(1-3):367-409.

[18] R.P. Stanlay,(1999), *Enumerative Combinatorics*, volume 2, Cambrige University Press.

[19] R.P. Stanlay, *Catalan addendum*, `http://www-math.mit.edu/~rstan/ec/catadd.pdf`

[20] P. R. Stein and M.S. Waterman,(1978), *On Some New Sequences Generalizing the Catalan and Motzkin Numhers*, J. Diser. Math., 26 , pp. 261-272

[21] Z. Wang, M. Gestein, and M. Snyder,(2009), *Rna-seq: a revolutionary tool for transcriptomics* Nat Rev Genet., 10(1):57-63.

[22] M.S. Waterman,(1978), *Secondary Structure of Single-Stranded Nucleic Acids* Los Alamos Scientific Lahorarory Los Alamos. New Mexico.

[23] H. Wilf,(1990), *generating functionology* Academic Press.

Appendix: Matlab Code for representation of secondary structure

```
phe_seq = ’GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCU
        GUGUUCGAUCCACAGAAUUCGCACCA’;
phe_str = rnafold(phe_seq)
% === Plot RNA secondary structure as tree
rnaplot(phe_str, ’seq’, phe_seq, ’format’, ’tree’);
% === Plot the secondary structure using the dot diagram representation
rnaplot(phe_str, ’seq’, phe_seq, ’format’, ’dot’);
text(500, 200, ’T-stem’);
text(100, 600, ’Anticodon stem’);
text(550, 650, ’D-stem stem’);
text(700, 400, ’Acceptor stem’);
aag_pos = 34:36;
cca_pos = 74:76;
rnaplot(phe_str, ’sequence’, phe_seq, ’format’, ’diagram’, ...
    ’selection’, [aag_pos, cca_pos]);
rnaplot(phe_str, ’sequence’, phe_seq, ’format’, ’graph’);
[ha, H] = rnaplot(phe_str, ’sequence’, phe_seq, ’format’, ’circle’, ...
    ’colorby’, ’state’);
H.Unpaired.Visible = ’off’;
legend off;
[ha, H] = rnaplot(phe_str, ’sequence’, phe_seq, ’format’, ’circle’,
         ’colorby’, ’residue’);
```