RNA as a combinatorial object
Asymptotics of RNA Shapes

Yann Ponty

Bioinformatics Team
École Polytechnique/CNRS/INRIA AMIB – France

November 27, 2009
RNA structure

Denition
Secondary structures of RNA =
Maximal non-crossing subset of canonical base-pairs.

Source: 5s rRNA (PDBID: 1K73:B)
RNA structure

UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCGCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC
U
U
A
G
C
G
G
C
G
C
C
A
C
A
G
C
G
G
U
G
G
GG
U
U
G
C
C
U
C
C
C
G
U
A
C
C
C
A
U
C
C
G
C
C
G
C
C
A
CC
1
10
20
30
40
50
60
70
80
90
100
110
120
122

Primary structure  Secondary+ structure  Tertiary structure

Source: 5s rRNA (PDBID: 1K73:B)

Definition

Secondary structures of RNA =
Maximal non-crossing subset of canonical base-pairs.
Outline

1. **Foreword**
   - Introduction
   - Motivation

2. **Enumerative combinatorics 101**
   - Generating functions
   - DSV/symbolic method
   - Singularity analysis

3. **RNA shapes**
   - Presentation
   - Motivation
   - $\pi$ shapes

Yann Ponty  RNA shapes combinatorics
Various representations for a versatile molecule

Outer planar graph

Well-parenthesized expression

Mountain view

Non intersecting arcs

Different objects
yet
Common combinatorial structure

Dot plot
Feynman’s diagram
Why use combinatorics?

Boltzmann ensemble is a (weighted) combinatorial class.

Studying it as such cleans out the details and helps:
- Assess asymptotic properties of sec. str.
- Investigate worst and average-case complexities
- Obtain better algorithms for RNA
Let $|\cdot|$ be a size function over objects (Sequences, trees, ...). Combinatorial classes are (infinite) sets $C$ of objects whose restrictions $C_n$ to objects of size $n$ are of finite cardinality.

**Definition (Generating functions)**

Let $C$ be a combinatorial class and $c_n = |C_n|$ the number of objects of size $n$ in $C$, then the generating function of $C$ is $C(z)$ s. t.

$$C(z) = \sum_{s \in C} z^{|s|} = \sum_{n \geq 0} c_n z^n$$

Closed forms for $C(z)$ are often easy to find . . .

**DNA example:** $\mathcal{D} := \{a, c, g, t\}^* \Rightarrow d_n = 4^n$
and $C(z) = 1 + 4z + 16z^2 + 64z^3 + \ldots = \sum_{n \geq 0} 4^n z^n = \frac{1}{1-4z}$

... and very often much simpler than for $c_n$!!!
From a **class specification**, one can directly establish the gen. fun.

Historically on languages, from Schützenberger’s observation that

*Gen. fun. are commutative images of languages*

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Generating function</th>
<th>Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C \rightarrow \varepsilon$</td>
<td>$C(z) = z^0 = 1$</td>
<td>$c_n = \mathbb{1}_{{0}}(n)$</td>
</tr>
<tr>
<td>$C \rightarrow t$</td>
<td>$C(z) = z^1 = z$</td>
<td>$c_n = \mathbb{1}_{{1}}(n)$</td>
</tr>
<tr>
<td>$C \rightarrow A \mid B$</td>
<td>$C(z) = A(z) + B(z)$</td>
<td>$c_n = a_n + b_n$</td>
</tr>
<tr>
<td>$C \rightarrow A.B$</td>
<td>$C(z) = A(z) \cdot B(z)$</td>
<td>$c_n = \sum_{k=0}^{n} a_k b_{n-k}$</td>
</tr>
</tbody>
</table>

**Remark:** One needs to ensure that unions are disjoint and concatenations unambiguous.

**DNA example:** $\{a, c, g, t\}^* \Leftrightarrow D \rightarrow a.D \mid c.D \mid g.D \mid t.D \mid \varepsilon$

\[ \Rightarrow D(z) = z \cdot D(z) + z \cdot D(z) + z \cdot D(z) + z \cdot D(z) + 1 \]
DSV/symbolic method

From a **class specification**, one can directly establish the gen. fun.

Historically on languages, from Schützenberger’s observation that

*Gen. fun. are commutative images of languages*

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Generating function</th>
<th>Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C \rightarrow \varepsilon$</td>
<td>$C(z) = z^0 = 1$</td>
<td>$c_n = \mathbb{1}_{{0}}(n)$</td>
</tr>
<tr>
<td>$C \rightarrow t$</td>
<td>$C(z) = z^1 = z$</td>
<td>$c_n = \mathbb{1}_{{1}}(n)$</td>
</tr>
<tr>
<td>$C \rightarrow A \mid B$</td>
<td>$C(z) = A(z) + B(z)$</td>
<td>$c_n = a_n + b_n$</td>
</tr>
<tr>
<td>$C \rightarrow A.B$</td>
<td>$C(z) = A(z) \cdot B(z)$</td>
<td>$c_n = \sum_{k=0}^{n} a_k b_{n-k}$</td>
</tr>
</tbody>
</table>

**Remark:** One needs to ensure that unions are disjoint and concatenations unambiguous.

**DNA example:** $\{a, c, g, t\}^* \iff D \rightarrow a.D \mid c.D \mid g.D \mid t.D \mid \varepsilon$

$\Rightarrow D(z) = 4z \cdot D(z) + 1$
DSV/symbolic method

From a **class specification**, one can directly establish the gen. fun.

Historically on languages, from Schützenberger’s observation that

*Gen. fun. are commutative images of languages*

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Generating function</th>
<th>Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C \rightarrow \varepsilon$</td>
<td>$C(z) = z^0 = 1$</td>
<td>$c_n = \mathbb{1}_{{0}}(n)$</td>
</tr>
<tr>
<td>$C \rightarrow t$</td>
<td>$C(z) = z^1 = z$</td>
<td>$c_n = \mathbb{1}_{{1}}(n)$</td>
</tr>
<tr>
<td>$C \rightarrow A \mid B$</td>
<td>$C(z) = A(z) + B(z)$</td>
<td>$c_n = a_n + b_n$</td>
</tr>
<tr>
<td>$C \rightarrow A.B$</td>
<td>$C(z) = A(z) \cdot B(z)$</td>
<td>$c_n = \sum_{k=0}^{n} a_k b_{n-k}$</td>
</tr>
</tbody>
</table>

**Remark:** One needs to ensure that unions are disjoint and concatenations unambiguous.

**DNA example:** $\{a, c, g, t\}^* \Leftrightarrow D \rightarrow a.D \mid c.D \mid g.D \mid t.D \mid \varepsilon$

\[ \Rightarrow D(z) = \frac{1}{1 - 4z} \]
Main principles

Disclaimer
What follows, although true in this context, is embarrassingly simplistic.
A rigorous presentation can (and must) be found in Flaj./Sedg. 08.

A singularity is a point \( z = \rho \) where \( C(z) \) is no longer analytic.
Asymptotics of coeff \( c_n \) are driven by the singularities of \( C(z) \).

1\(^{st}\) principle
Location of the dominant (smallest) singularity \( \rho \) dictates the exponential growth
\[ \Rightarrow \frac{c_n}{\rho^{-n}} = o(\alpha^n), \ \forall \alpha > 1. \]

DNA example: \( D(z) = 1/(1 - 4z) \Rightarrow \rho = 1/4 \Rightarrow d_n \sim 4^n P(n) \).
Basic scale

2nd principle

Nature of \( \rho \) dictates subexponential part \( P(n) \) s.t. \( c_n \sim \rho^{-n}P(n) \).

Basic scale: If one can rewrite \( C(z) \) as

\[
C(z) = f(z) + g(z)(1 - z/\rho)^\alpha
\]

where \( f \) and \( g \) are analytic \( \forall |z| < |\rho| \) and non-null at \( \rho \), then

\[
c_n \equiv [z^n]C(z) \sim \frac{g(\rho)\rho^{-n}}{\Gamma(-\alpha)n^{\alpha+1}}
\]

Example: \( D(z) = \frac{1}{1 - 4z} \Rightarrow c_n \sim 4^n \)

(\( \rho = 1/4, \alpha = -1, f(z) = 0, \) and \( g(z) = 1 \))
Asymptotic estimates are obtained using a 4 steps meta-algorithm:

1. **Find the right model**

2. **Translate into grammar**

3. **Translate into system and solve g. f.**

4. **Singularity analysis yields asymptotics**
Asymptotic estimates are obtained using a 4 steps meta-algorithm:

1. Find the right model
2. Translate into grammar
3. Translate into system and solve g. f.
4. Singularity analysis yields asymptotics
Asymptotic estimates are obtained using a 4 steps meta-algorithm:

1. Find the right model
2. Translate into grammar
3. Translate into system and solve g. f.
4. Singularity analysis yields asymptotics
Asymptotic estimates are obtained using a 4 steps meta-algorithm:

1. Find the right model
2. Translate into grammar
3. Translate into system and solve g. f.
4. Singularity analysis yields asymptotics
Let us count dot-bracket notations (Motzkin words)
Let us count dot-bracket notations (Motzkin words)

1. $\epsilon$

2. $M \rightarrow \bullet M \mid (M) M \mid \epsilon$
Let us count dot-bracket notations (Motzkin words)

1. \[ M = \bullet M \lor (M) M \lor \varepsilon \]

2. \[ M(z) = z \cdot M(z) + z \cdot M(z) \cdot z \cdot M(z) + 1 \]
   \[ = \frac{1 - z \pm \sqrt{1 - 2z - 3z^2}}{2z^2} \]
Let us count dot-bracket notations (Motzkin words)

1. \[ \bullet \cdots \bullet = \bullet \cdots \bullet \lor \bullet \cdots \bullet \lor \varepsilon \]

2. \[ M \rightarrow \bullet M \lor (M)M \lor \varepsilon \]

3. \[ M(z) = z \cdot M(z) + z \cdot M(z) \cdot z \cdot M(z) + 1 \]
   
   \[ = \left\{ \begin{array}{l}
   \frac{1 - z + \sqrt{1 - 2z - 3z^2}}{2z^2} = \frac{1}{z^2} - \frac{1}{z} - 1 - z - 2z^2 + O(z^3) \\
   \frac{1 - z - \sqrt{1 - 2z - 3z^2}}{2z^2} = 1 + z + 2z^2 + 4z^3 + 9z^4 + O(z^5)
   \end{array} \right. \]
Appetizer: Motzkin words

Let us count dot-bracket notations (Motzkin words)

1. $\begin{array}{c}
\bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \\
\end{array} = \begin{array}{c}
\bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \\
\end{array} \lor \begin{array}{c}
\bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \\
\end{array} \lor \varepsilon$

2. $M \rightarrow \bullet M \quad | \quad (M) M \quad | \quad \varepsilon$

3. $M(z) = z \cdot M(z) + z \cdot M(z) \cdot z \cdot M(z) + 1$

\[= \frac{1-z-\sqrt{1-2z-3z^2}}{2z^2}\]

4. $\rho = 1/3$, $M(z) = \frac{1-z}{2z^2} - g(z) \cdot \sqrt{1-z}/\rho$, and $g(z) := \frac{\sqrt{1+z}}{2z^2}$

\[\Rightarrow s_n \equiv [z^n]M(z) \sim \frac{g(\rho)\rho^{-n}}{\Gamma(-\alpha)n^{\alpha+1}} = \frac{3\sqrt{3}}{2\sqrt{\pi}} \cdot \frac{3^n}{n^{3/2}}(1 + O(1/n))\]
Consider RNA secondary structures (Waterman 78)
Consider RNA secondary structures (Waterman 78)

1. \[ \bullet \rightarrow \epsilon \]

2. \[ S \rightarrow \bullet S | (S_{>0}) S | \epsilon \]
Consider RNA secondary structures (Waterman 78)

1. $\cdot \cdots \cdot = \cdot \cdots \cdot \vee \geq 1 \vee \varepsilon$

2. $S \rightarrow \bullet S \mid (T) S \mid \varepsilon$
   $T \rightarrow \bullet S \mid (T) S$
Consider RNA secondary structures (Waterman 78)

1. \[ \bullet \ldots \bullet = \bullet \ldots \bullet \lor \bullet \ldots \bullet \lor \varepsilon \geq 1 \]

2. 
   - \( S \rightarrow \bullet S \mid (T) S \mid \varepsilon \)
   - \( T \rightarrow \bullet S \mid (T) S \)

3. \[ S(z) = \frac{1-z+z^2-\sqrt{1-2z-z^2-2z^3+z^4}}{2z^2} \]
Consider RNA secondary structures (Waterman 78)

1. $1 = \cdot \cdot \cdot \lor \cdot \cdot \cdot \lor \varepsilon$

2. $S \rightarrow \cdot S \mid (T) S \mid \varepsilon$
   $T \rightarrow \cdot S \mid (T) S$

3. $S(z) = \frac{1-z+z^2-\sqrt{1-2z-z^2-2z^3+z^4}}{2z^2}$

4. $\rho = \frac{3-\sqrt{5}}{2} = 1 - \phi$
   $[z^n]S(z) = \sqrt{\frac{15+7\sqrt{5}}{8\pi}} \cdot \left(\frac{3+\sqrt{5}}{2}\right)^n \frac{2}{n\sqrt{n}} (1 + O(1/n)) \sim 1.1 \cdot \frac{2.6^n}{n\sqrt{n}}$
Let us generalize the $\theta$ constraint

$1$

$\text{----------------------} = \text{----------------------} \geq \theta \text{----------------------} \lor \text{----------------------}$

$\text{----------------------} = \text{----------------------} \lor \varepsilon$
Let us generalize the $\theta$ constraint

1. 

2. 

$S \rightarrow U( S \geq \theta ) S | U$

$U \rightarrow \bullet U | \varepsilon$
Let us generalize the $\theta$ constraint

1. $\begin{array}{c}
\text{---} = \text{---} \geq \theta \text{---} \lor \text{---} \\
\text{---} = \text{---} \lor \varepsilon
\end{array}$

2. $\begin{array}{ll}
S & \rightarrow \ U ( T ) \ S \ | \ U \\
T & \rightarrow \ U ( T ) \ S \ | \ U^\theta
\end{array}$
Let us generalize the $\theta$ constraint

1. $\cdots \implies \cdots \geq \theta \implies \cdots$

2. $S \rightarrow U(T) S | U$
   $T \rightarrow U(T) S | \cdot^{\theta} U$
   $U \rightarrow \cdot U | \varepsilon$

3. $S(z) = \frac{1-2z+2z^2-z^{\theta+2}-\sqrt{1-4z+4z^2-2z^{\theta+2}+4z^{\theta+3}-4z^{\theta+4}+z^{2\theta+4}}}{(1-z)2z^2}$
Let us generalize the $\theta$ constraint

$$S(z) = \frac{1-2z+2z^2-z^{\theta+2}-\sqrt{1-4z+4z^2-2z^{\theta+2}+4z^{\theta+3}-4z^{\theta+4}+z^{2\theta+4}}}{(1-z)2z^2}$$

$$s_n \sim K. \frac{\beta^n}{n\sqrt{n}} (1 + \mathcal{O}(1/n))$$
Message #1
Finding the right decomposition (DP) is a combinatorial task.

Message #2
Applying automatic theorems gives precise asymptotic equivalents.

Message #3
There is a large exponential number of structures of size n:
Homopolymer model: $\Omega(2^n)$    Stickiness model: $\mathcal{O}(1.8^n/n^{3/2})$
Message #1
Finding the right decomposition (DP) is a combinatorial task.

Message #2
Applying automatic theorems gives precise asymptotic equivalents.

First order estimate:
\( n = 50 \Rightarrow 95\% \) accuracy
Message #1
Finding the right decomposition (DP) is a combinatorial task.

Message #2
Applying automatic theorems gives precise asymptotic equivalents.

$n = 20$ accuracies:
- $1^{st}$ order 88%
- $2^{nd}$ order 99.8%
- $3^{nd}$ order 99.992%

Secondary structures
Message #1
Finding the right decomposition (DP) is a combinatorial task.

Message #2
Applying automatic theorems gives precise asymptotic equivalents.

Message #3
There is a large exponential number of structures of size $n$:
Homopolymer model: $\Omega(2^n)$
Stickiness model: $O\left(1.8^n/n^{3/2}\right)$
Outline

1. Foreword
2. Enumerative combinatorics 101
3. RNA shapes
   - Presentation
   - Motivation
   - $\pi$ shapes
Definition (RNA shapes [Giegerich et al.])

*Coarse-grain* representation hierarchy for RNA sec. struct.

Based on the *underlying backbone* structure.

**Example**

Sec. str. ((((((.((((((....))))))))))))(((((......))))))

$\pi'$-shape [ [ [ [ ] ] ] [ [ ] ] ] ]

$\pi$-shape [ [ [ ] ] [ ] [ ] ]
Definition (RNA shapes [Giegerich et al])

*Coarse-grain* representation hierarchy for RNA sec. struct.

Based on the *underlying backbone* structure.

**Example**

Sec. str. ((((((.((((...((((.......))))))))((((.......)))))))....))))


Contract identical consecutive characters
Definition (RNA shapes [Giegerich et al])

*Coarse-grain* representation hierarchy for RNA sec. struct.

Based on the *underlying backbone* structure.

Example

Sec. str. ((((((...((((((((((((........))))))))))))))))))))((((((........))))))...))

\[\pi'-\text{shape} \ [ \bullet \ [ \bullet \ [ \bullet \ [ \bullet \ ] ] ] ] \ [ \bullet \ ] \bullet \ ]\]

\[\pi-\text{shape} \ [ \ [ \ [ \ - \ - \ ] ] \ [ \bullet \ ] \ ]\]

Remove unpaired regions

Contract nested helices
Motivation

RNA shapes allow a hierarchical search in the Boltzmann ensemble

10000 samples \( \Rightarrow \) 1727 Secondary structures...
Motivation

RNA shapes allow a hierarchical search in the Boltzmann ensemble

10000 samples ⇒ 1727 Secondary structures...

... 406 π'-shapes...
Motivation

RNA shapes allow a hierarchical search in the Boltzmann ensemble

10000 samples ⇒ 1727 Secondary structures...

... 406 π'-shapes...

... but only 9 π-shapes!
RNA shapes allow a hierarchical search in the Boltzmann ensemble

Is it reasonable to perform an exhaustive search of all possible shapes compatible with input structure?

10000 samples $\Rightarrow$ 1727 Secondary structures...

How many shapes must we investigate?

... 406 $\pi'$-shapes...

... but only 9 $\pi$-shapes!
Objective: Count $\pi$-shapes with $2n$ parentheses.

$\pi$-shapes are bracket words avoiding the $[[\ldots]]$ motif.
Objective: Count π-shapes with $2n$ parentheses.

1. π-shapes are bracket words avoiding the $[[\ldots]]$ motif.

2. $S \rightarrow [ S/\{[\ldots]\} ] S \mid [ S/\{[\ldots]\} ]$
**Objective:** Count $\pi$-shapes with $2n$ parentheses.

1. $\pi$-shapes are bracket words avoiding the $[[\ldots]]$ motif.
2. $S \rightarrow [T]S | [T]$  \hspace{1cm} $T \rightarrow [T]S | \varepsilon$
Objective: Count \( \pi \)-shapes with \( 2n \) parentheses.

1. \( \pi \)-shapes are bracket words avoiding the \([[[...]]]\) motif.

2. \[ \begin{align*} S & \rightarrow [T]S | [T] \\ T & \rightarrow [T]S | \varepsilon \end{align*} \]

3. \[ S(z) = \frac{1 - z^2 - \sqrt{1 - 2z^2 - 3z^4}}{2z^2} \]
Objective: Count \( \pi \)-shapes with \( 2n \) parentheses.

1. \( \pi \)-shapes are bracket words avoiding the \([\ldots]\) motif.

2. \( S \rightarrow [T]S | [T] \quad T \rightarrow [T]S | \varepsilon \)

3. \[
S(z) = \frac{1 - z^2 - \sqrt{1 - 2z^2 - 3z^4}}{2z^2}
\]

4. \[
s_{2n} \sim \frac{\sqrt{3}}{2 \sqrt{\pi}} \cdot \frac{3^n}{n \sqrt{\pi}} (1 + \mathcal{O}(1/n)) \quad \text{and} \quad s_{2n+1} = 0
\]

Remark: Doesn’t this look familiar???
Limitations

Number of $\pi$-shapes of size $n$

\[ \neq \]

Number of $\pi$-shapes compatible with RNA of size $n$

Reasons:

1. Shapes of size $\leq n$ should be considered
2. Forming a hairpin loop \([ \ ]\) takes at least $\theta + 2$ bases

2. \[ S \rightarrow [T]S | [T] \quad T \rightarrow [T]S | \varepsilon \]

3. \[ S(z) = \frac{1 - z^2 - \sqrt{1 - 2z^2 - 3z^4}}{2z^2} \]

4. For $n$ even: \[ s_{2n} \sim \frac{3\sqrt{3}}{4\sqrt{\pi}} \cdot \frac{3^n}{n \sqrt{\pi}} \left(1 + O(1/n)\right) \approx 0.48 \cdot \frac{3^n}{n \sqrt{n}} \]
Limitations

Number of $\pi$-shapes of size $n$ 
\[ \neq \]
Number of $\pi$-shapes compatible with RNA of size $n$

Reasons:

1. Shapes of size $\leq n$ should be considered
2. Forming a hairpin loop $[ ]$ takes at least $\theta + 2$ bases

\[ S \rightarrow [ T ] S | [ T ] \quad T \rightarrow [ T ] S | \bullet^\theta \]
\[ R \rightarrow \square S | \varepsilon \]

\[ R(z) = \frac{1 - z^2 - \sqrt{1 - 2z^2 - 3z^4}}{2z^2(1 - z)} \]

\[ r_{2n} \sim \frac{3\sqrt{3}}{4\sqrt{\pi}} \cdot \frac{3^n}{n\sqrt{n}}(1 + O(1/n)) \Rightarrow r_n \approx 2.07 \cdot \frac{1.73^n}{n\sqrt{n}} \]
Limitations

Number of $\pi$-shapes of size $n$

$\neq$

Number of $\pi$-shapes compatible with RNA of size $n$

Reasons:

1. Shapes of size $\leq n$ should be considered
2. Forming a hairpin loop $[]$ takes at least $\theta + 2$ bases

2. $S \rightarrow [T] SS | [T]$ $T \rightarrow [T] SS | \bullet^\theta$

$R \rightarrow \Box S | \varepsilon$

3. $R(z) = \frac{1 - z^{\theta+2} - \sqrt{1 - 2z^{\theta+2} - 4z^{\theta+4} + z^{2\theta+4}}}{2z^2(1 - z)}$

4. $\theta = 3 \Rightarrow r_n \approx 2.44 \frac{1.32^n}{n\sqrt{n}}$
A surprising bijection

Theorem

$\#\pi$ shapes of size $n = \#\text{Motzkin words of length } 2n + 2$

Proof.

\[
S(z) = \frac{1-z^2-\sqrt{1-2z^2-3z^4}}{2z^2}, \quad M(z) = \frac{1-z-\sqrt{1-2z-3z}}{2z^2}
\]

\[
S(z) = 1 + z^2 M(z^2) \implies s_n = m_{2n+2}
\]

These two classes are in bijection.
How to state it? Can we exploit it?
Explicit bijection

Let \( \psi, \phi : \{ [ , ] \}^* \rightarrow \{ ( , ) , \bullet \} \) such that

\[
\psi(( A ) B) = \begin{cases} 
\phi(A) & \text{if } B = \varepsilon \\
\phi(A) \bullet \psi(B) & \text{otherwise}
\end{cases}
\]

\[
\phi(( A ) B) = \phi(A)[\psi(B)] \\
\phi(\varepsilon) = \varepsilon.
\]

Then \( \psi \) is a bijection between \( s_{2n+2} \) and \( m_n \).
Explicit bijection

Let $\psi, \phi : \{ [ , ] \}^* \rightarrow \{ ( , ) , \bullet \}$ such that

\[
\psi(( A ) B) = \begin{cases} 
\phi(A) & \text{if } B = \varepsilon \\
\phi(A) \bullet \psi(B) & \text{otherwise}
\end{cases}
\]

\[
\phi(( A ) B) = \phi(A)[\psi(B)]
\]

$\phi(\varepsilon) = \varepsilon.$

Then $\psi$ is a bijection between $s_{2n+2}$ and $m_n$.

$2n + 2$ letters $\rightarrow 2n + 2$ edges $\rightarrow n$ edges
Explicit bijection

Let $\psi, \phi : \{[ , ]\}^* \rightarrow \{( , ) , \bullet\}$ such that

$$
\psi(( A ) B) = \begin{cases} 
\phi(A) & \text{if } B = \varepsilon \\
\phi(A) \bullet \psi(B) & \text{otherwise}
\end{cases}
$$

$$
\phi(( A ) B) = \phi(A)[\psi(B)]
$$

$\phi(\varepsilon) = \varepsilon.$

Then $\psi$ is a bijection between $s_{2n+2}$ and $m_n$.
Let $\psi, \phi : \{ [ , ] \}^* \to \{ ( , ) , \cdot \}$ such that

$$
\psi((A)B) = \begin{cases} 
\phi(A) & \text{if } B = \epsilon \\
\phi(A) \cdot \psi(B) & \text{otherwise}
\end{cases}
$$

$$
\phi((A)B) = \phi(A)[\psi(B)]
$$

$\phi(\epsilon) = \epsilon.$

Then $\psi$ is a bijection between $s_{2n+2}$ and $m_n$.

2n + 2 letters

2n + 2 edges

n edges
Let $\psi, \phi : \{ [ , ] \}^* \rightarrow \{ ( , ) , \bullet \}$ such that

\[
\psi(( A ) B) = \begin{cases} 
\phi(A) & \text{if } B = \varepsilon \\
\phi(A) \bullet \psi(B) & \text{otherwise}
\end{cases}
\]

\[
\phi(( A ) B) = \phi(A)[\psi(B)]
\]

$\phi(\varepsilon) = \varepsilon$.

Then $\psi$ is a bijection between $s_{2n+2}$ and $m_n$. 

Yann Ponty
RNA shapes combinatorics
Explicit bijection

Let \( \psi, \phi : \{ [ , ] \}^* \rightarrow \{ ( , ) , \bullet \} \) such that

\[
\psi(( A ) B) = \begin{cases} 
\phi(A) & \text{if } B = \varepsilon \\
\phi(A) \bullet \psi(B) & \text{otherwise}
\end{cases}
\]

\[
\phi(( A ) B) = \phi(A)[\psi(B)]
\]

\[
\phi(\varepsilon) = \varepsilon.
\]

Then \( \psi \) is a bijection between \( s_{2n+2} \) and \( m_n \).

\[
\begin{align*}
\text{[ [ [ [] [] ] ] ] } & \quad \longleftrightarrow \quad \text{[ [ [ [ ] ] ] ] [ ] ] [ ] [ ] [ ] [ ] [ [ ] ] [ ] } \\
2n + 2 \text{ letters} & \quad \longleftrightarrow \quad 2n + 2 \text{ edges} \quad \longleftrightarrow \quad n \text{ edges} \\
\end{align*}
\]
Explicit bijection

Let \( \psi, \phi : \{ [ , ] \}^* \rightarrow \{ ( , ) , \bullet \} \) such that

\[
\psi((A)B) = \begin{cases} 
\phi(A) & \text{if } B = \varepsilon \\
\phi(A) \bullet \psi(B) & \text{otherwise}
\end{cases}
\]

\[
\phi((A)B) = \phi(A)[\psi(B)]
\]

\[
\phi(\varepsilon) = \varepsilon.
\]

Then \( \psi \) is a bijection between \( s_{2n+2} \) and \( m_n \).
Limits of the bijection

Impacts of $\theta$ on shapes and Motzkin are drastically different.

**Theorem**

*Expectations of number of term. loops in Motzkin words and $\pi$-shapes scale like $m_n^t \sim \frac{n}{6} + \mathcal{O}(1)$ and $s_{2n+2}^t \sim \frac{2n}{3} + \mathcal{O}(1)$*
**Objective:** Count \( \pi' \)-shapes compatible with RNA of length \( n \).

1. \( \pi' \)-shapes = bracket words avoiding motifs \([[\cdots]]\) and \(\bullet\bullet\)

2. \( \begin{align*} R &\rightarrow \square R | S \\ S &\rightarrow U[T]S | U \\ U &\rightarrow \diamond | \varepsilon \\ T &\rightarrow U[T]U[T]S | \diamond[T] | [T] \diamond | \diamond[T] \diamond | \bullet^\theta \end{align*} \)

3. \( \theta = 3, \ R(z) = \frac{1 + 2z^2 + 2z^3 + z^4 - z^5}{z^6 - \sqrt{1 - 4z^3 - 2z^4 - 2z^5 + 2z^6 - 7z^8 - z^{10} + 2z^{11} + z^{12}} - \frac{2z^2(1 - z^2)}{2z^2}} \)

4. \( r_n \sim 1.27 \frac{1.81^n}{n\sqrt{n}} \)
Objective: Count $\pi'$-shapes compatible with RNA of length $n$. 

1. $\pi'$-shapes = bracket words avoiding motifs $[[\ldots]]$ and $\bullet\bullet$

2. 
   \[
   \begin{align*}
   R & \rightarrow \square R \mid S \\
   S & \rightarrow U [ T ] S \mid U \\
   U & \rightarrow \Diamond \mid \varepsilon \\
   T & \rightarrow U [ T ] U [ T ] S \mid \Diamond [ T ] [ T ] \Diamond \mid \Diamond [ T ] \Diamond \mid \bullet^\theta
   \end{align*}
   \]

3. \[\theta = 3, \quad R(z) = \frac{1+2z^2+2z^3+z^4-z^5-z^6-\sqrt{1-4z^3-2z^4-2z^5+2z^6-7z^8-z^{10}+2z^{11}+z^{12}}}{2z^2(1-z^2)}\]

4. \[r_n \sim 1.27 \frac{1.81^n}{n\sqrt{n}}\]
Objective: Count $\pi'$-shapes compatible with RNA of length $n$.

1. $\pi'$-shapes = bracket words avoiding motifs $[[\ldots]]$ and ••

2. $R \rightarrow \square R \mid S$
   $S \rightarrow U [ T ] S \mid U$
   $U \rightarrow \diamond \mid \varepsilon$

   $T \rightarrow U [ T ] U [ T ] S \mid \diamond [ T ] \mid [ T ]\diamond \mid \diamond [ T ]\diamond \mid \bullet^\theta$

3. $\theta = 3$, $R(z) = \frac{1 + 2z^2 + 2z^3 + z^4 - z^5 - z^6 - \sqrt{1 - 4z^3 - 2z^4 - 2z^5 + 2z^6 - 7z^8 - z^{10} + 2z^{11} + z^{12}}}{2z^2(1 - z^2)}$

4. $r_n \sim 1.27 \frac{1.81^n}{n\sqrt{n}}$
Objective: Count $\pi'$-shapes compatible with RNA of length $n$.

1. $\pi'$-shapes = bracket words avoiding motifs $[[\ldots]]$ and $\bullet\bullet$

2. $R \rightarrow \square R | S$
   $S \rightarrow U [ T ] S | U$
   $U \rightarrow \diamond | \varepsilon$

   $T \rightarrow U [ T ] U [ T ] S | \diamond [ T ] | [ T ] \diamond | \diamond [ T ] \diamond | \bullet^\theta$

3. $\theta = 3$, $R(z) = \frac{1 + 2z^2 + 2z^3 + z^4 - z^5 - z^7 - \sqrt{1 - 4z^3 - 2z^4 - 2z^5 + 2z^6 - 7z^8 - z^{10} + 2z^{11} + z^{12}}}{2z^2(1-z^2)}$

4. $r_n \sim 1.27 \frac{1.81^n}{n \sqrt{n}}$
<table>
<thead>
<tr>
<th>Model</th>
<th>Asymptotic number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sec. str. on $n$ – Combinatorial</td>
<td>$1.1 \cdot \frac{2.6^n}{n\sqrt{n}}$</td>
</tr>
<tr>
<td>Sec. str. on $n$ – Empirical</td>
<td>$0.04 \cdot \frac{1.4^n}{n\sqrt{n}}$</td>
</tr>
<tr>
<td>$\pi$-shapes of size $n$</td>
<td>$1.38 \cdot \frac{1.73^n}{n\sqrt{n}}$</td>
</tr>
<tr>
<td>$\pi$-shapes compatible with sec. str. on $n$</td>
<td>$2.44 \cdot \frac{1.32^n}{n\sqrt{n}}$</td>
</tr>
<tr>
<td>$\pi$-shapes – Empirical</td>
<td>$0.21 \cdot \frac{1.1^n}{n\sqrt{n}}$</td>
</tr>
<tr>
<td>$\pi'$-shapes of size $n$</td>
<td>$0.99 \cdot \frac{2.41^n}{n\sqrt{n}}$</td>
</tr>
<tr>
<td>$\pi'$-shapes compatible with sec. str. on $n$</td>
<td>$1.28 \cdot \frac{1.81^n}{n\sqrt{n}}$</td>
</tr>
</tbody>
</table>
Conclusion

- For context-free objects, finding gen. fun. is easy...
  ... and precise asymptotics estimates follow readily.
- Bijection between Motkzin words and π-shapes
- Way less many shapes than sec. str.!
- Homopolymer model overestimates number of shapes
  Need for a probabilistic model for base-pairing
  But stickiness is not enough...

Collaborators: W. A. Lorenz and P. Clote (Boston College)