Analytic properties of RNA secondary structures and their representations

Asymptotics of RNA Shapes

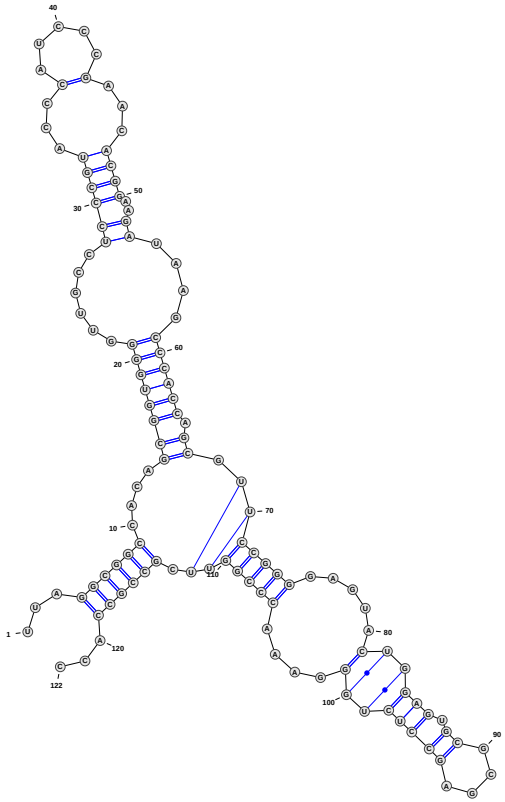
Yann Ponty

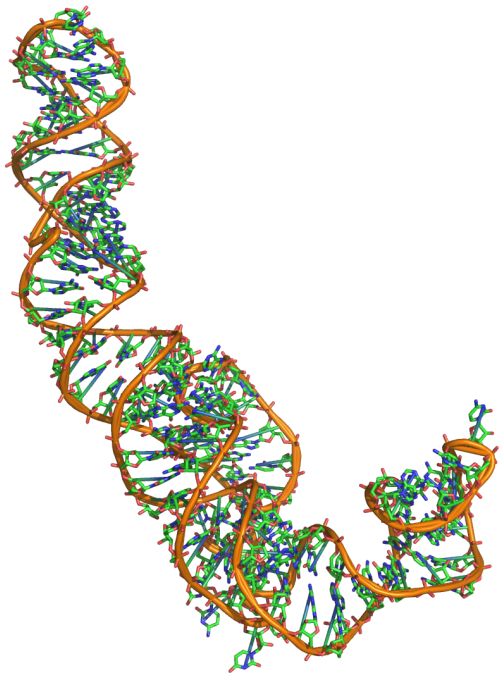
Bioinformatics Team
École Polytechnique/CNRS/INRIA AMIB – France

April 15, 2014

RNA structure

UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC





Primary structure | Secondary structure | Tertiary structure

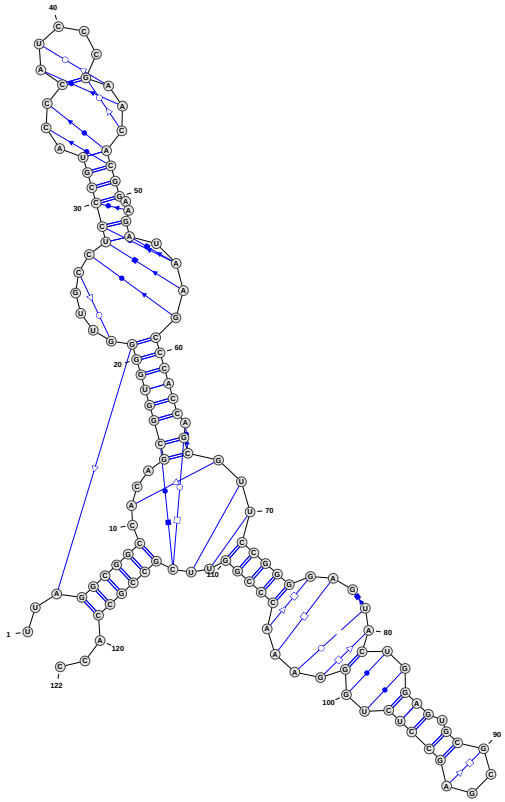
Source: 5s rRNA (PDBID: 1K73:B)

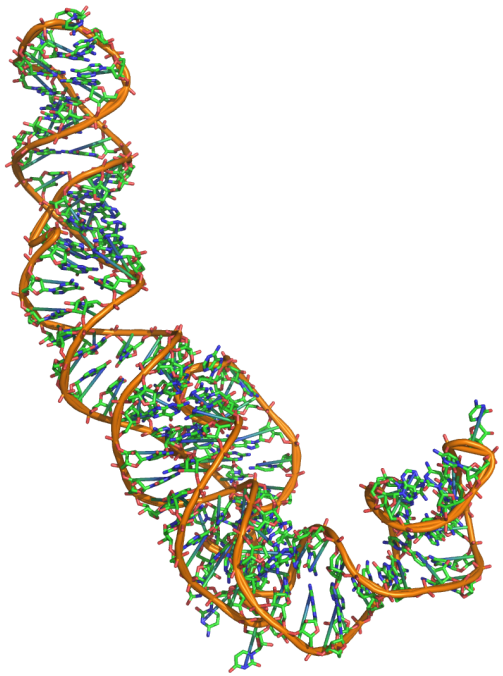
Definition

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

RNA structure

UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC





Primary structure Secondary$^+$ structure Tertiary structure

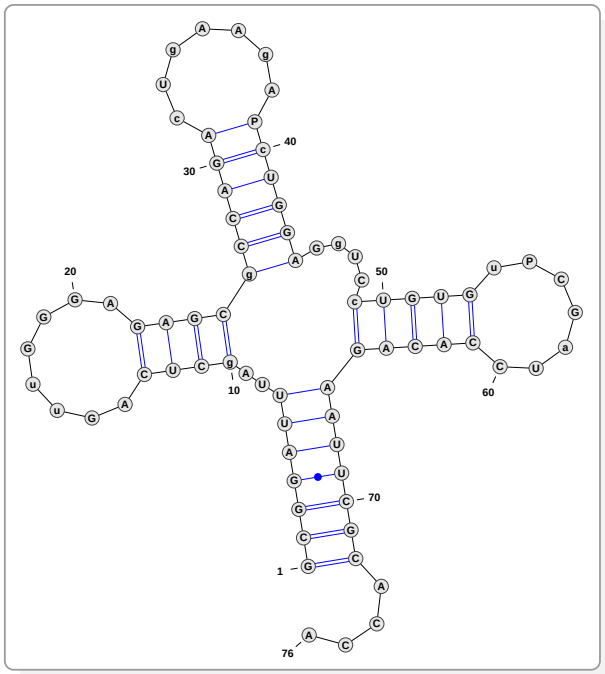
Source: 5s rRNA (PDBID: 1K73:B)

Definition

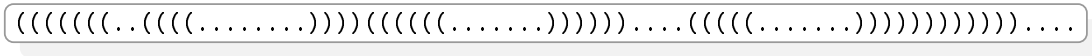
Secondary structures of RNA =
Maximal non-crossing subset of **canonical** base-pairs.

Outline

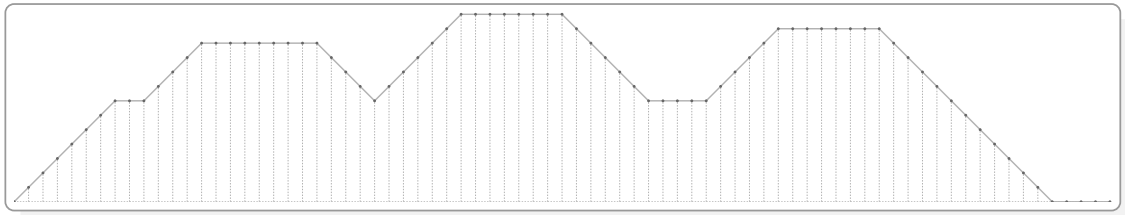
Various representations for a versatile molecule



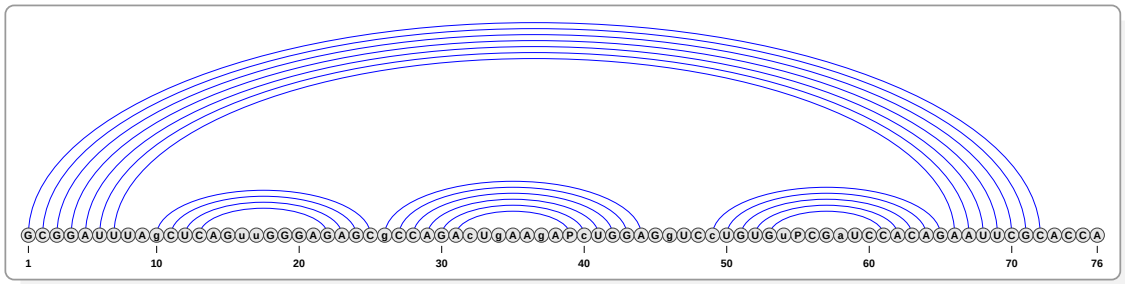
Outer planar graph



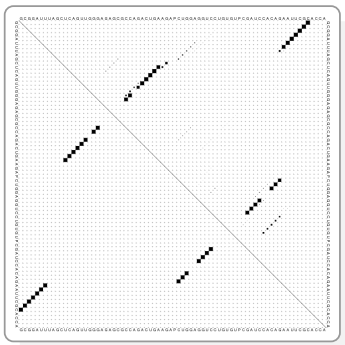
Well-parenthesized expression



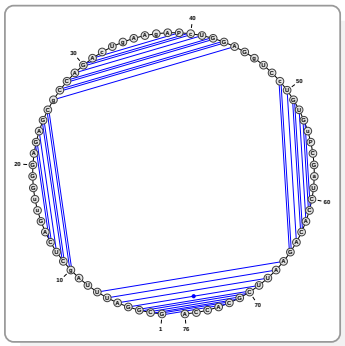
Mountain view



Non intersecting arcs



Dot plot

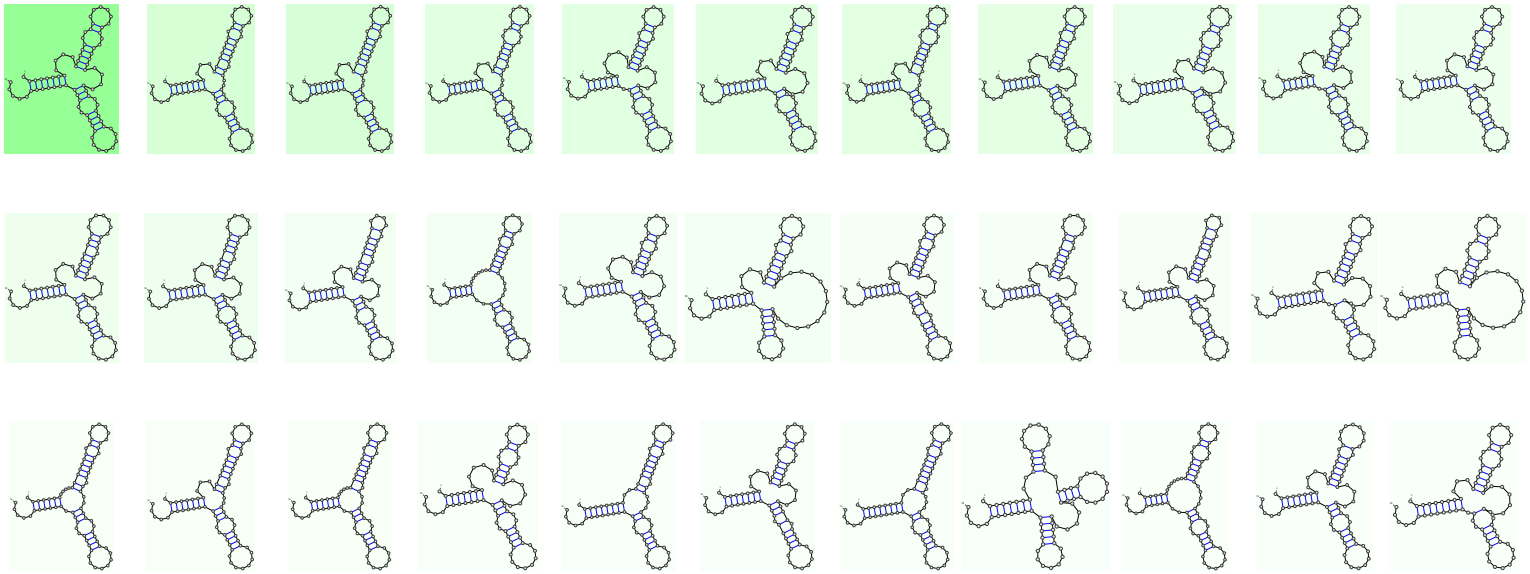


Circular diagram

Different objects
yet
Common combinatorial structure

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

Generating functions

Let $|.|$ be a **size function** over **objects** (Sequences, trees, ...).
Combinatorial classes are (infinite) sets $\mathcal{C}$ of objects whose restrictions $\mathcal{C}_n$ to objects of size n are of **finite cardinality**.

Definition (Generating functions)

Let $\mathcal{C}$ be a combinatorial class and $c_n = |\mathcal{C}_n|$ the number of objects of size n in $\mathcal{C}$, then the **generating function** of $\mathcal{C}$ 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 ...

Example (DNA)

$\mathcal{D} := \{a, c, g, t\}^* \Rightarrow d_n = 4^n$
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 finding closed-form for c_n!!!

DSV/symbolic method

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

Historically on languages, stems from Schützenberger's observation

Gen. fun. are commutative images of languages

Grammar	Generating function	Coefficients
$C \to \varepsilon$	$C(z) = z^0 = 1$	$c_n = \mathbb{1}_{\{0\}}(n)$
$C \to t$	$C(z) = z^1 = z$	$c_n = \mathbb{1}_{\{1\}}(n)$
$C \to A \mid B$	$C(z) = A(z) + B(z)$	$c_n = a_n + b_n$
$C \to A.B$	$C(z) = A(z) \cdot B(z)$	$c_n = \sum_{k=0}^{n} a_k b_{n-k}$

Remark: Disjoint unions and unambiguous concatenations.

Example (DNA)

DNA = $\{a, c, g, t\}^* \Leftrightarrow D \to a.D \mid c.D \mid g.D \mid t.D \mid \varepsilon$

$$\Rightarrow D(z) = z \cdot D(z) + z \cdot D(z) + z \cdot D(z) + z \cdot D(z) + 1$$

DSV/symbolic method

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

Historically on languages, stems from Schützenberger's observation

Gen. fun. are commutative images of languages

Grammar	Generating function	Coefficients
$C \to \varepsilon$	$C(z) = z^0 = 1$	$c_n = \mathbb{1}_{\{0\}}(n)$
$C \to t$	$C(z) = z^1 = z$	$c_n = \mathbb{1}_{\{1\}}(n)$
$C \to A \mid B$	$C(z) = A(z) + B(z)$	$c_n = a_n + b_n$
$C \to A.B$	$C(z) = A(z) \cdot B(z)$	$c_n = \sum_{k=0}^{n} a_k b_{n-k}$

Remark: Disjoint unions and unambiguous concatenations.

Example (DNA)

DNA = $\{a, c, g, t\}^* \Leftrightarrow D \to a.D \mid c.D \mid g.D \mid t.D \mid \varepsilon$

$$\Rightarrow D(z) = 4z \cdot D(z) + 1$$

DSV/symbolic method

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

Historically on languages, stems from Schützenberger's observation

Gen. fun. are commutative images of languages

Grammar	Generating function	Coefficients
$C \to \varepsilon$	$C(z) = z^0 = 1$	$c_n = \mathbb{1}_{\{0\}}(n)$
$C \to t$	$C(z) = z^1 = z$	$c_n = \mathbb{1}_{\{1\}}(n)$
$C \to A \mid B$	$C(z) = A(z) + B(z)$	$c_n = a_n + b_n$
$C \to A.B$	$C(z) = A(z) \cdot B(z)$	$c_n = \sum_{k=0}^{n} a_k b_{n-k}$

Remark: Disjoint unions and unambiguous concatenations.

Example (DNA)

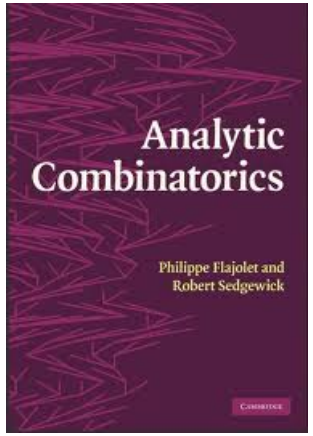
DNA = $\{a, c, g, t\}^* \Leftrightarrow D \to a.D \mid c.D \mid g.D \mid t.D \mid \varepsilon$

$$\Rightarrow D(z) = \frac{1}{1 - 4z}$$

Analytic combinatorics: Main principles

Disclaimer

What follows, although true in this context, is **embarassingly simplistic**. A rigorous presentation can (and should) be sought in the **Flajolet/Sedgewick Bible**.



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 ρ dictates the **exponential growth** $\Rightarrow \frac{c_n}{\rho^{-n}} = o(\alpha^n),\ \forall \alpha > 1.$

Example (DNA)

$D(z) = 1/(1-4z) \Rightarrow \rho = 1/4 \Rightarrow d_n \sim 4^n P(n).$

Analytic combinatorics: Basic transfer theorem

2^{nd} principle

Nature of ρ 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 ρ, then

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

Example (DNA)

$$D(z) = \frac{1}{1-4z} \Rightarrow c_n \sim 4^n$$
$$(\rho = 1/4, \alpha = -1, f(z) = 0, \text{ and } g(z) = 1)$$

General methodology

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

General methodology

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

General methodology

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

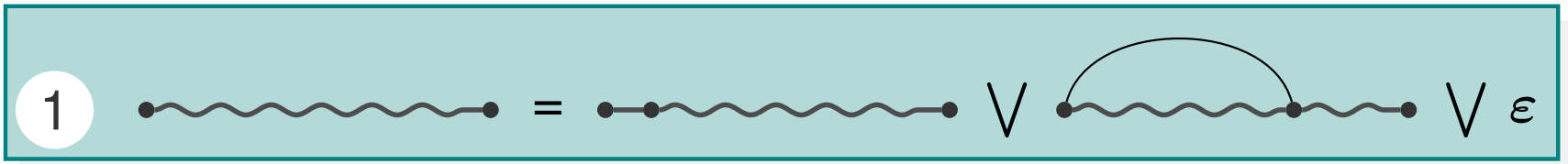
General methodology

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

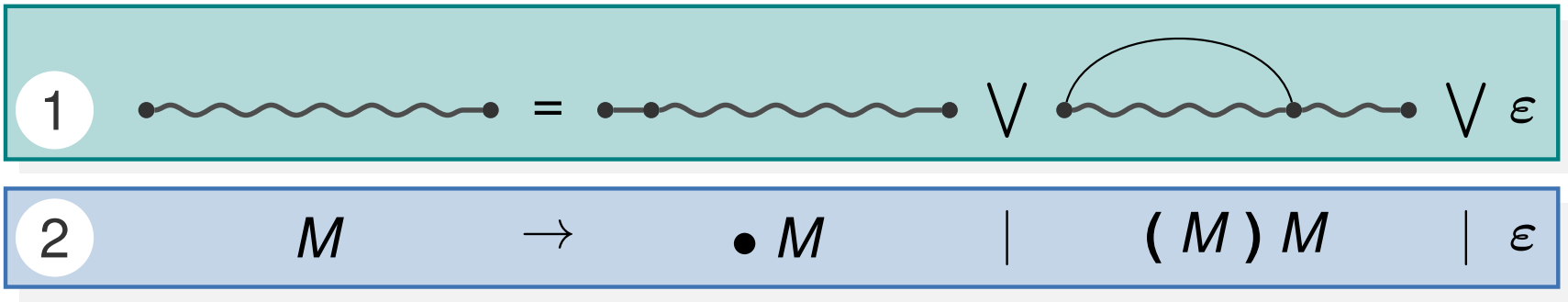
Appetizer: Motzkin words

Let us count dot-bracket notations (Motzkin words)



Appetizer: Motzkin words

Let us count dot-bracket notations (Motzkin words)



Appetizer: Motzkin words

Let us count dot-bracket notations (Motzkin words)

1. 

2. $$M \quad \to \quad \bullet\, M \quad | \quad (\, M\,)\, M \quad | \quad \varepsilon$$

3. $$M(z) = z \cdot M(z) + z \cdot M(z) \cdot z \cdot M(z) + 1$$
$$= \frac{1 - z \pm \sqrt{1 - 2z - 3z^2}}{2z^2}$$

Appetizer: Motzkin words

Let us count dot-bracket notations (Motzkin words)

1. 

2. $$M \quad \to \quad \bullet\, M \quad | \quad (\, M\,)\, M \quad | \quad \varepsilon$$

3. $$M(z) = z \cdot M(z) + z \cdot M(z) \cdot z \cdot M(z) + 1$$
$$= \begin{cases} \frac{1-z+\sqrt{1-2z-3z^2}}{2z^2} = \frac{1}{z^2} - \frac{1}{z} - 1 - z - 2z^2 + \mathcal{O}(z^3) \\ \frac{1-z-\sqrt{1-2z-3z^2}}{2z^2} = 1 + z + 2z^2 + 4z^3 + 9z^4 + \mathcal{O}(z^5) \end{cases}$$

Appetizer: Motzkin words

Let us count dot-bracket notations (Motzkin words)

1. 

2. $$M \quad \to \quad \bullet\, M \quad | \quad (\, M\,)\, M \quad | \quad \varepsilon$$

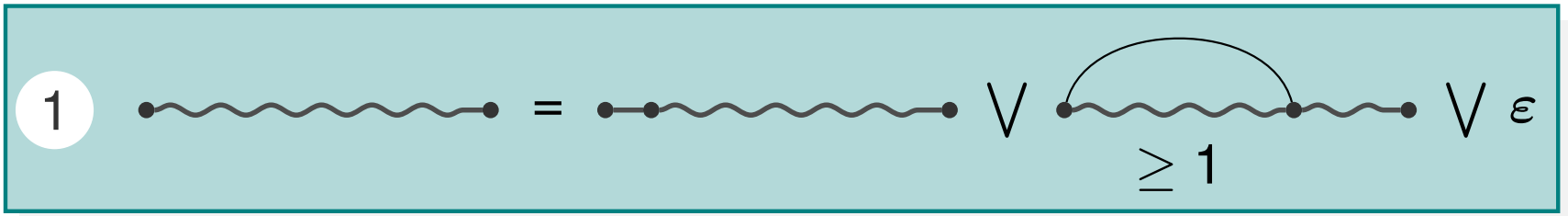
3. $$M(z) = z \cdot M(z) + z \cdot M(z) \cdot z \cdot M(z) + 1$$
$$= \frac{1-z-\sqrt{1-2z-3z^2}}{2z^2}$$

4. $$\rho = 1/3,\ M(z) = \frac{1-z}{2z^2} - g(z)\cdot\sqrt{1-z/\rho},\ \text{and}\ g(z) := \frac{\sqrt{1+z}}{2z^2}$$
$$\Rightarrow s_n \equiv [z^n]M(z) \sim \frac{g(\rho)\rho^{-n}}{\Gamma(-\alpha)n^{\alpha+1}} = \frac{3\sqrt{3}}{2\sqrt{\pi}} \cdot \frac{3^n}{n\sqrt{n}}(1 + \mathcal{O}(1/n))$$

RNA secondary structures

Consider RNA secondary structures

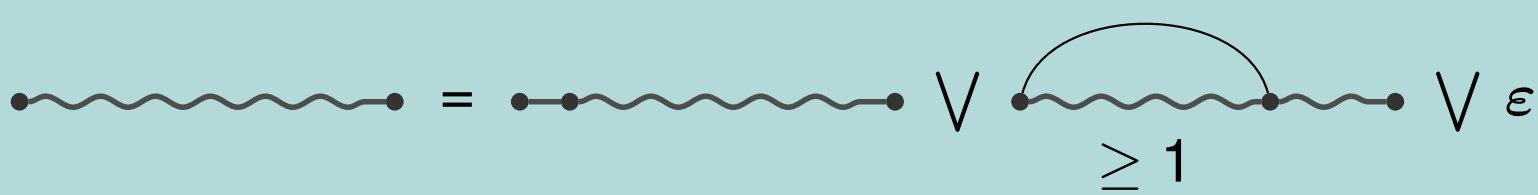
[Waterman 78 + Viennot/Vauchaussade de Chaumont 78]



RNA secondary structures

Consider RNA secondary structures

[Waterman 78 + Viennot/Vauchaussade de Chaumont 78]

1. 

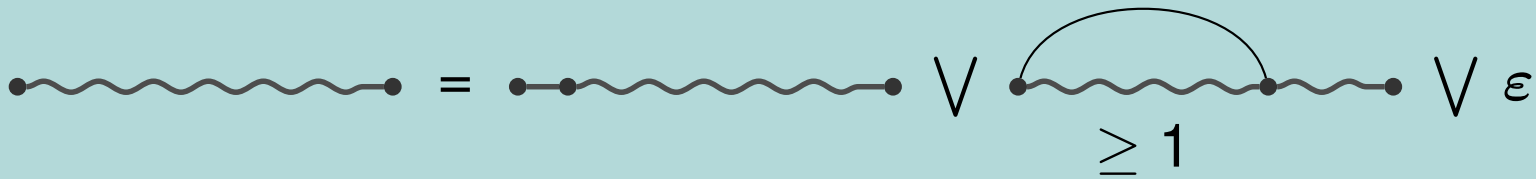
2. $S \quad \rightarrow \quad \bullet\, S \mid (\, S_{>0}\,)\, S \mid \varepsilon$

RNA secondary structures

Consider RNA secondary structures

[Waterman 78 + Viennot/Vauchaussade de Chaumont 78]

1



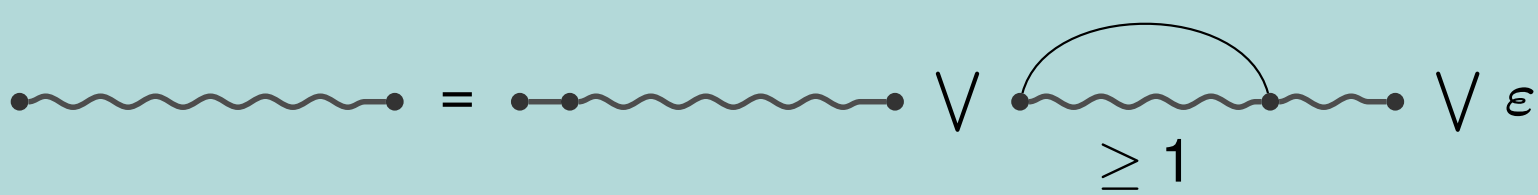
2

$$S \rightarrow \bullet\, S \mid (\, \mathbf{T}\,)\, S \mid \varepsilon$$
$$\mathbf{T} \rightarrow \bullet\, S \mid (\, \mathbf{T}\,)\, S$$

RNA secondary structures

Consider RNA secondary structures

[Waterman 78 + Viennot/Vauchaussade de Chaumont 78]

1. 

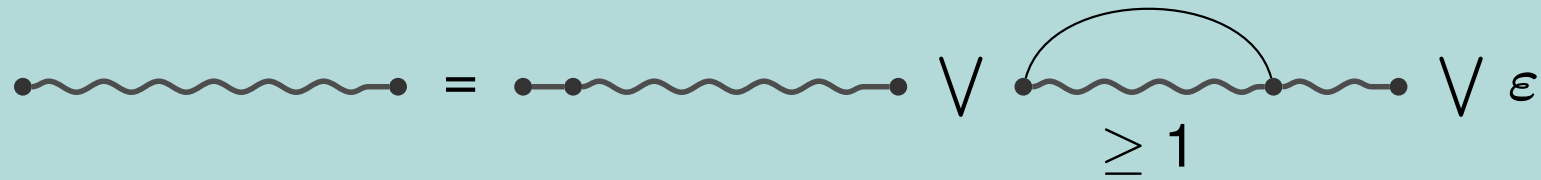
2. $$\begin{aligned} S &\to \bullet\, S \mid (\, \mathbf{T}\,)\, S \mid \varepsilon \\ \mathbf{T} &\to \bullet\, S \mid (\, \mathbf{T}\,)\, S \end{aligned}$$

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

RNA secondary structures

Consider RNA secondary structures

[Waterman 78 + Viennot/Vauchaussade de Chaumont 78]

1. 

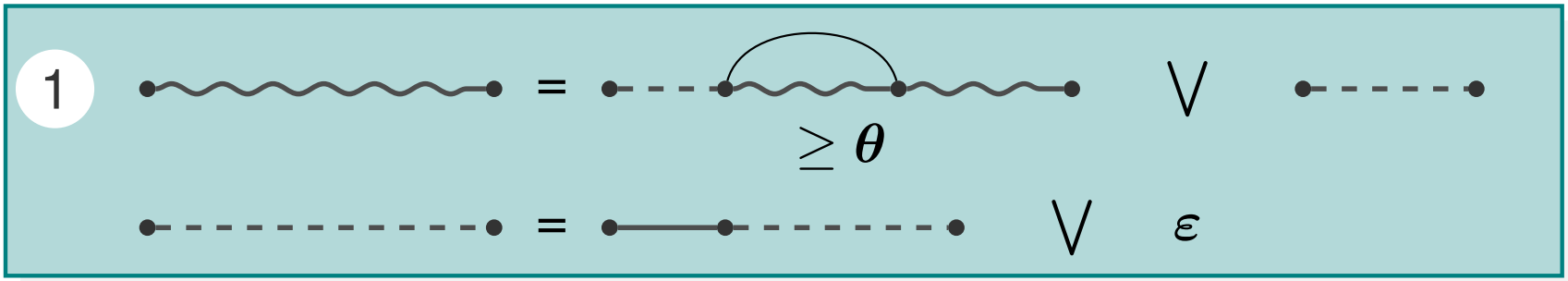
2. $$\begin{aligned} S &\to \bullet\, S \mid (\, \mathbf{T}\,)\, S \mid \varepsilon \\ \mathbf{T} &\to \bullet\, S \mid (\, \mathbf{T}\,)\, S \end{aligned}$$

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

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

RNA secondary structures

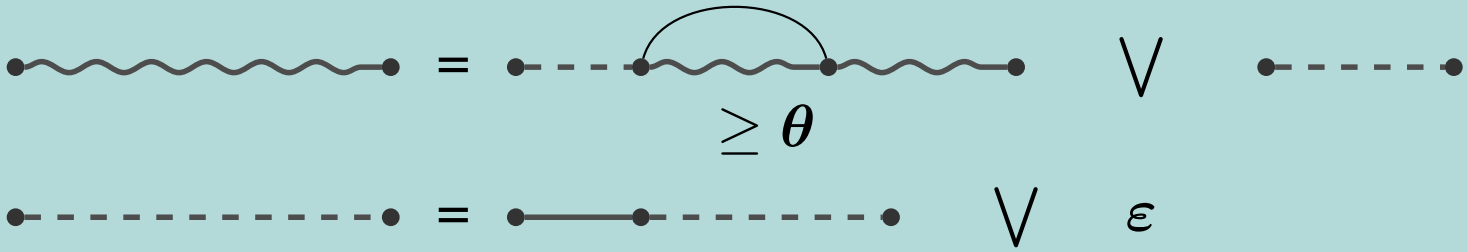
Let us generalize the θ constraint



RNA secondary structures

Let us generalize the θ constraint

1



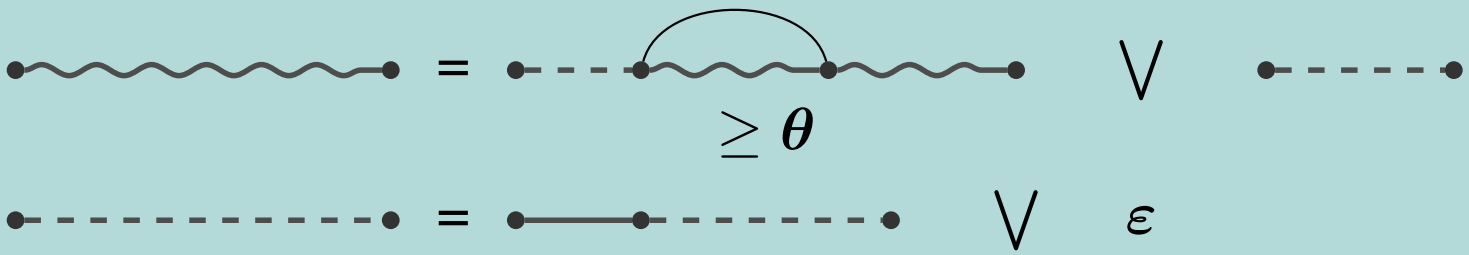
2

$$S \to U(S_{\geq\theta})S \mid U \qquad U \to \bullet U \mid \varepsilon$$

RNA secondary structures

Let us generalize the θ constraint

1



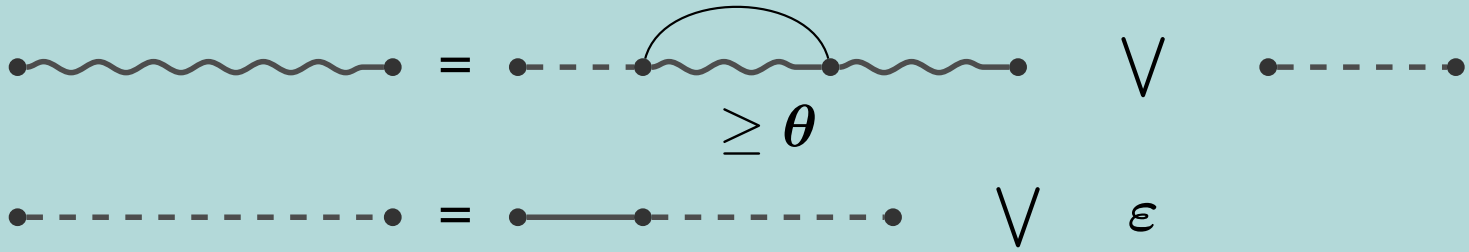
2

$$S \to U(\mathbf{T})S \mid U \qquad U \to \bullet U \mid \varepsilon$$
$$\mathbf{T} \to U(\mathbf{T})S \mid \bullet^{\theta} U$$

RNA secondary structures

Let us generalize the θ constraint

1



2

$$S \to U(\mathbf{T})S \mid U \qquad U \to \bullet U \mid \varepsilon$$
$$\mathbf{T} \to U(\mathbf{T})S \mid \bullet^{\theta} U$$

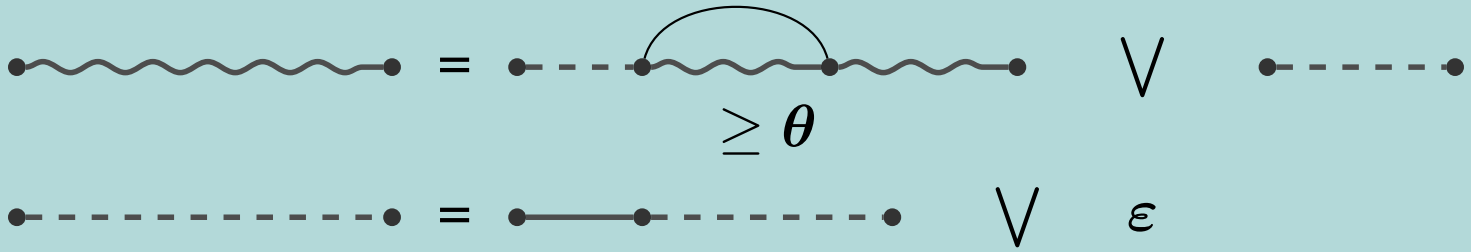
3

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

RNA secondary structures

Let us generalize the θ constraint

1



2

$$S \to U(\mathbf{T})S \mid U \qquad U \to \bullet U \mid \varepsilon$$
$$\mathbf{T} \to U(\mathbf{T})S \mid \bullet^{\theta} U$$

3

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

4

$$s_n \sim K.\frac{\beta^n}{n\sqrt{n}}(1+\mathcal{O}(1/n))$$

θ	0	1	3	10
β	3.	2.62	2.29	2.02

Half-time report

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:** $\mathcal{O}(1.8^n/n^{3/2})$

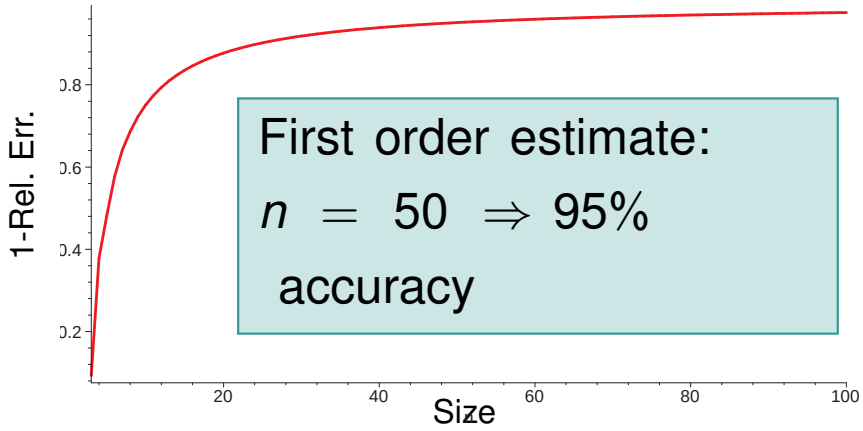
Half-time report

Message #1

Finding the right decomposition (DP) is a combinatorial task.

Message #2

Applying **automatic theorems** gives **precise** asymptotic equivalents.



Secondary structures

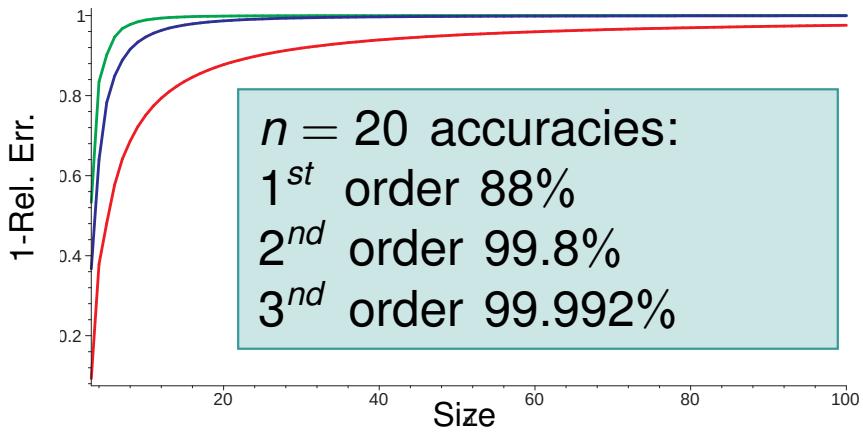
Half-time report

Message #1

Finding the right decomposition (DP) is a combinatorial task.

Message #2

Applying **automatic theorems** gives **precise** asymptotic equivalents.



Secondary structures

Half-time report

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:** $\mathcal{O}(1.8^n/n^{3/2})$

Outline

Presentation

Definition (RNA shapes [Giegerich *et al*])

Coarse-grain representation hierarchy for RNA sec. struct.

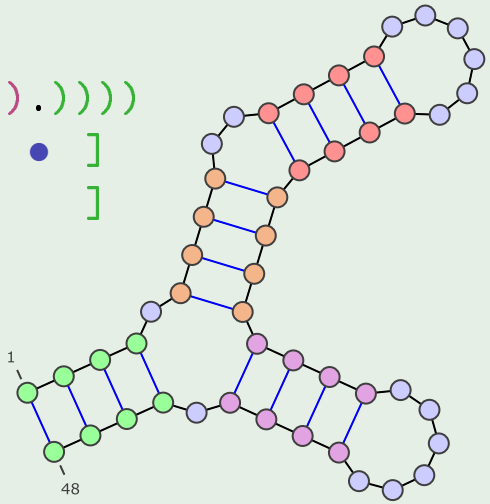
Based on the *underlying backbone* structure.

Example

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

π'-shape [• [• [•]] [•] •]

π-shape [[- -] [ ]]



Presentation

Definition (RNA shapes [Giegerich *et al*])

Coarse-grain representation hierarchy for RNA sec. struct.

Based on the *underlying backbone* structure.

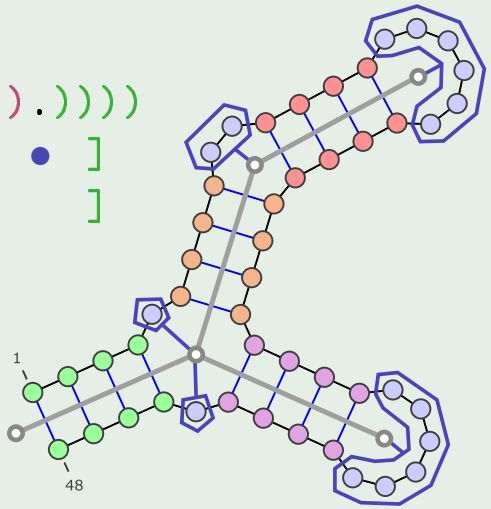
Example

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

π'**-shape** [• [• [•]] [•] •]

π-shape [[- -] [ ]]

Contract identical consecutive characters



Presentation

Definition (RNA shapes [Giegerich *et al*])

Coarse-grain representation hierarchy for RNA sec. struct.

Based on the *underlying backbone* structure.

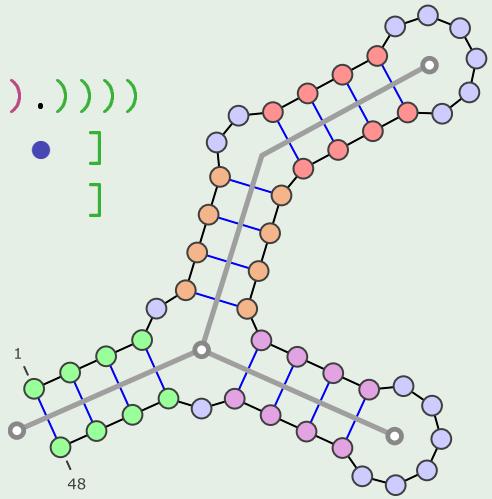
Example

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

π'-shape [• [• [•]] [•] •]

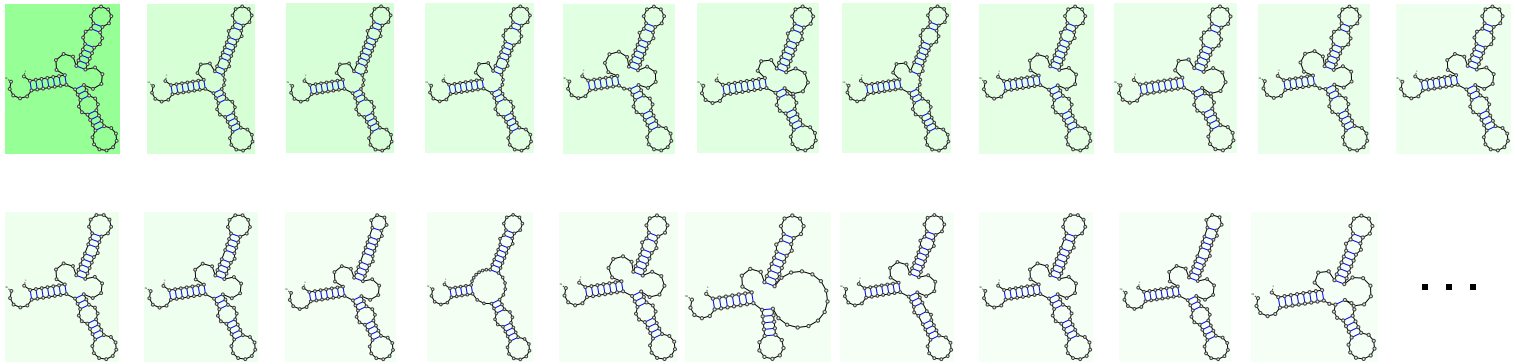
π-shape [[- -] [ ]]

Remove unpaired regions
Contract nested helices



Motivation

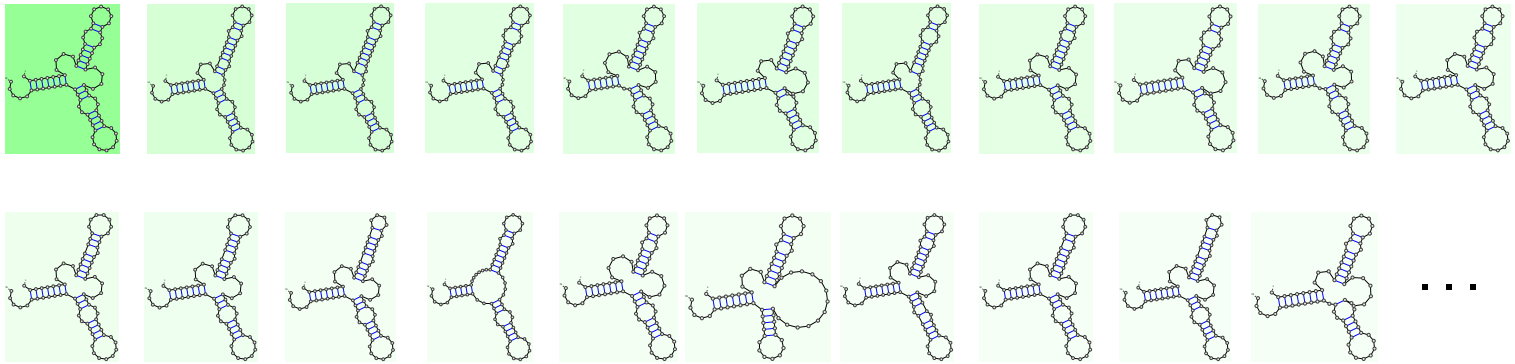
RNA shapes allow for a hierarchical search in the Boltzmann ensemble



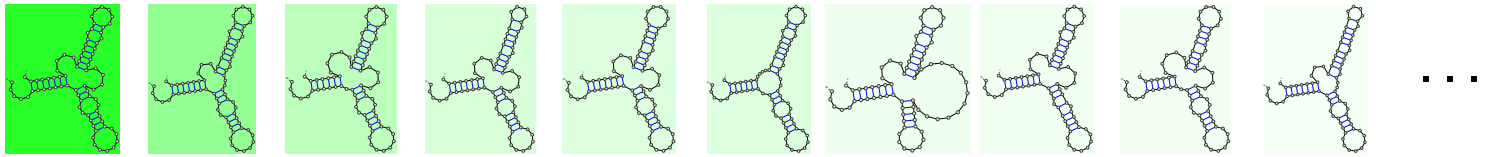
10000 samples $\Rightarrow$ 1727 Secondary structures. . .

Motivation

RNA shapes allow for a hierarchical search in the Boltzmann ensemble



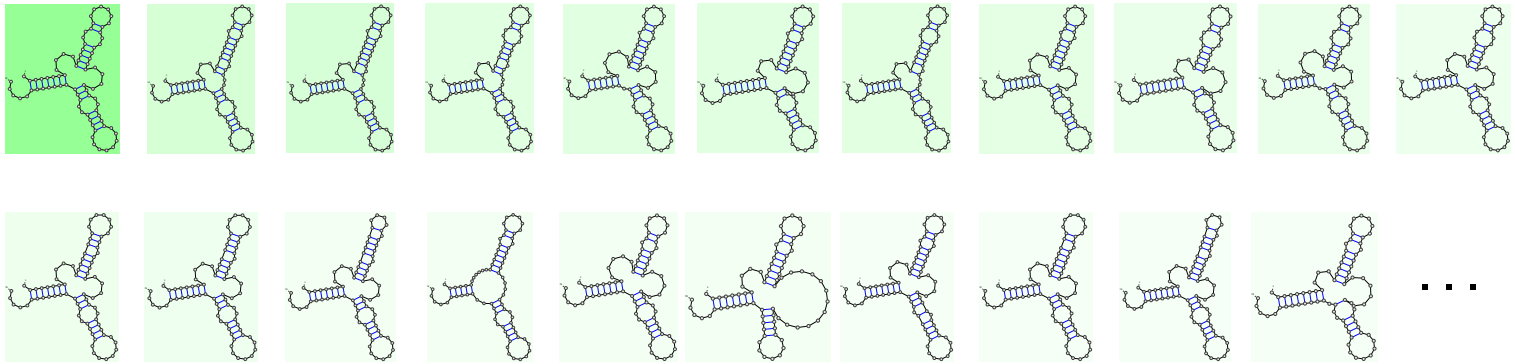
10000 samples $\Rightarrow$ 1727 Secondary structures...



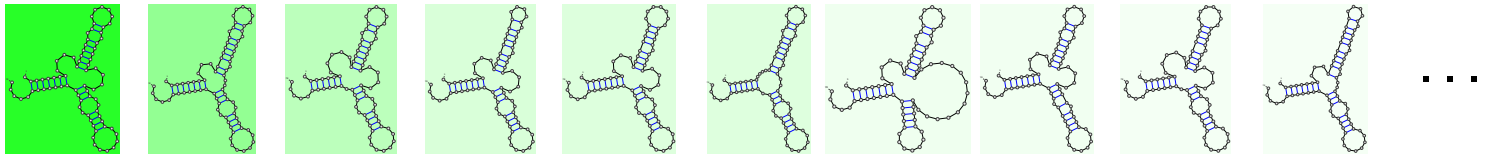
...406 π'-shapes...

Motivation

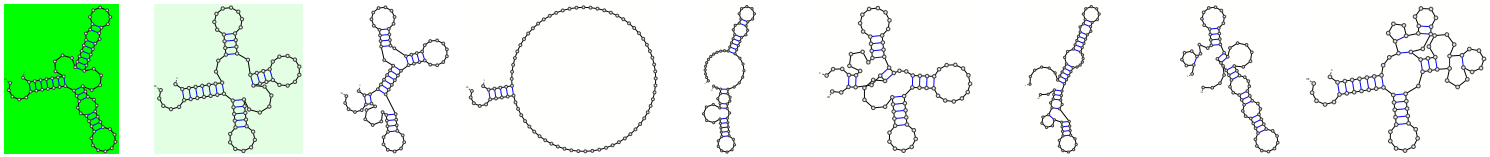
RNA shapes allow for a hierarchical search in the Boltzmann ensemble



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



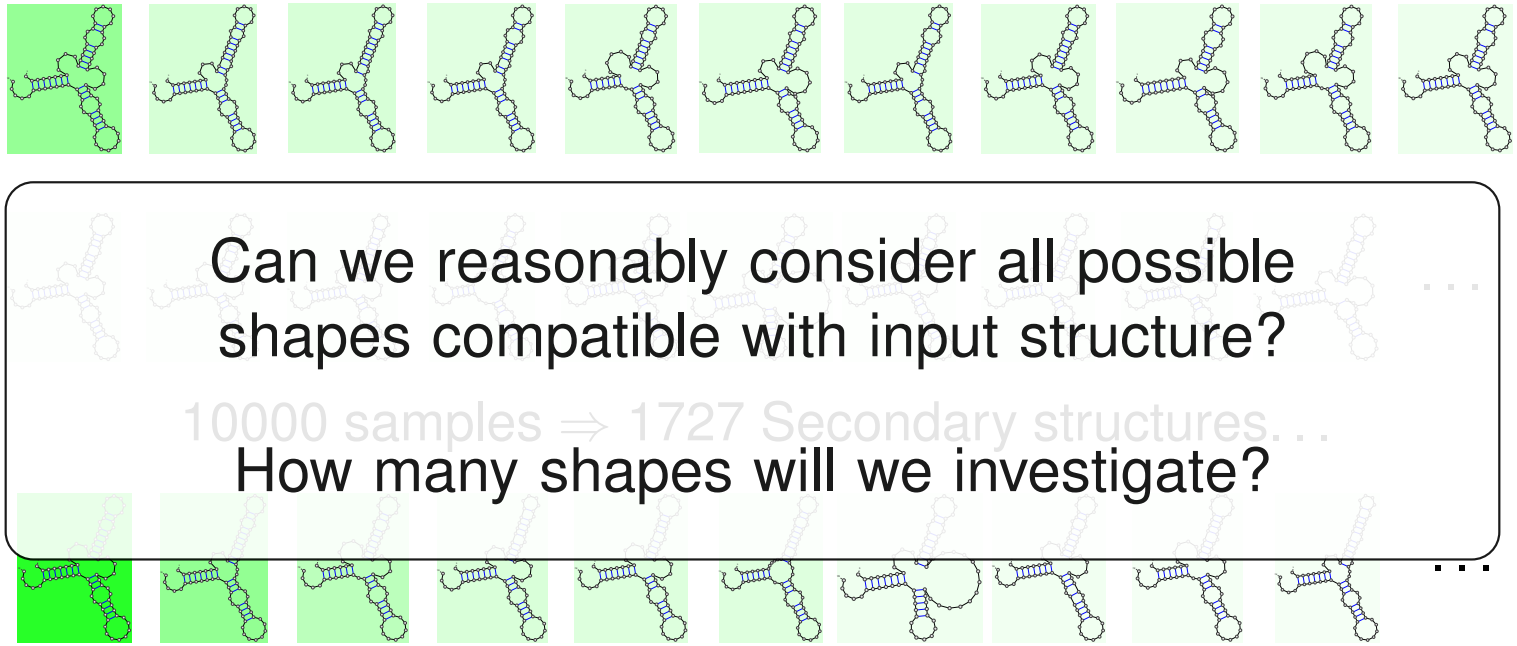
...406 π'-shapes...



...but only 9 π-shapes!

Motivation

RNA shapes allow for a hierarchical search in the Boltzmann ensemble

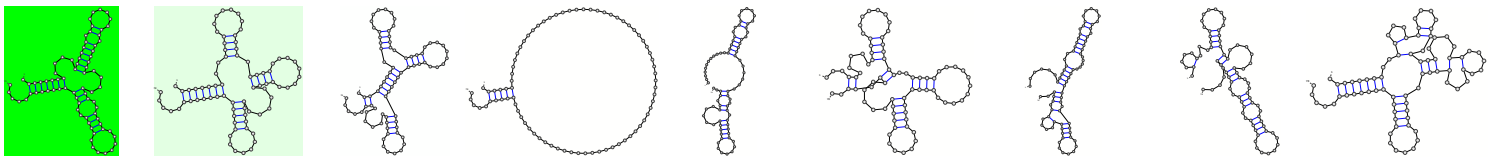


Can we reasonably consider all possible shapes compatible with input structure?

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

How many shapes will we investigate?

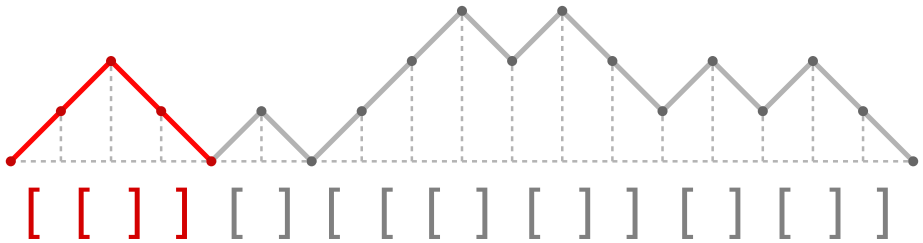
...406 π'-shapes...



...but only 9 π-shapes!

π-shapes

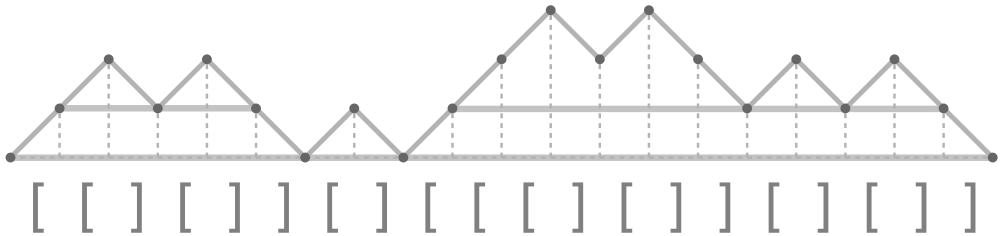
Objective: Count π-shapes with $2n$ parentheses.



1. π-shapes are bracket words avoiding the `[[...]]` motif.

π-shapes

Objective: Count π-shapes with $2n$ parentheses.

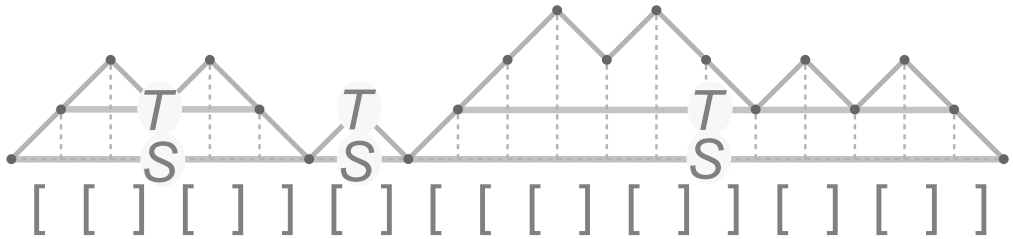


1. π-shapes are bracket words avoiding the `[[...]]` motif.

2. $S \rightarrow [\, S_{/\{[\ldots]\}} \,]\, S \mid [\, S_{/\{[\ldots]\}} \,]$

π-shapes

Objective: Count π-shapes with $2n$ parentheses.

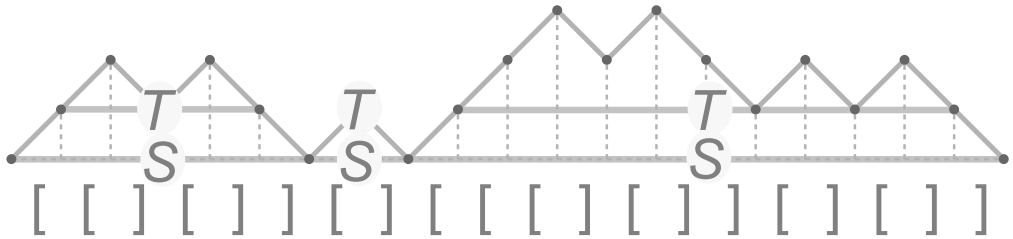


1. π-shapes are bracket words avoiding the `[[...]]` motif.

2. $S \rightarrow [\,\mathbf{T}\,]\,S \mid [\,\mathbf{T}\,] \qquad \mathbf{T} \rightarrow [\,\mathbf{T}\,]\,S \mid \varepsilon$

π-shapes

Objective: Count π-shapes with $2n$ parentheses.



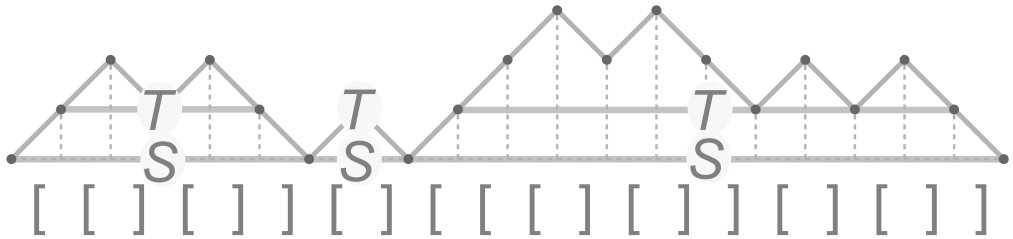
1. π-shapes are bracket words avoiding the `[[...]]` motif.

2. $S \to [\,\mathbf{T}\,]\,S \mid [\,\mathbf{T}\,] \qquad \mathbf{T} \to [\,\mathbf{T}\,]\,S \mid \varepsilon$

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

π-shapes

Objective: Count π-shapes with $2n$ parentheses.



1. π-shapes are bracket words avoiding the `[[...]]` motif.

2. $S \to [\,\mathbf{T}\,]\,S \mid [\,\mathbf{T}\,] \qquad \mathbf{T} \to [\,\mathbf{T}\,]\,S \mid \varepsilon$

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 π-shapes of size n

$\neq$

Number of π-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 $S \rightarrow [\ T\ ]\ S \mid [\ T\ ] \qquad T \rightarrow [\ T\ ]\ S \mid \varepsilon$

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

4 For n even: $s_{2n} \sim \frac{3\sqrt{3}}{4\sqrt{\pi}} \cdot \frac{3^n}{n\sqrt{\pi}}(1 + \mathcal{O}(1/n)) \approx 0.48 \cdot \frac{3^n}{n\sqrt{n}}$

Limitations

Number of π-shapes of size n

$\neq$

Number of π-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 $$S \rightarrow [\,T\,]\,S \mid [\,T\,] \qquad T \rightarrow [\,T\,]\,S \mid \bullet^{\theta}$$
$$R \rightarrow \square\,\mathrm{S} \mid \varepsilon$$

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

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

Limitations

Number of π-shapes of size n

$\neq$

Number of π-shapes compatible with RNA of size n

Reasons:

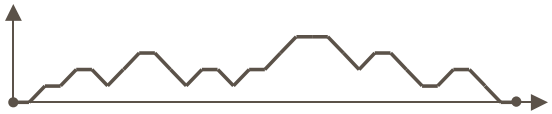
1. Shapes of size $\leq n$ should be considered
2. Forming a hairpin loop [ ] *takes* at least $\theta + 2$ bases

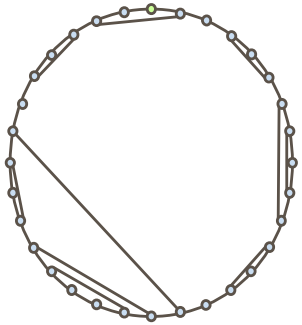
2 $$S \rightarrow [\ T\ ]\ S \mid [\ T\ ] \qquad T \rightarrow [\ T\ ]\ S \mid \bullet^{\theta}$$
$$R \rightarrow \square\, S \mid \varepsilon$$

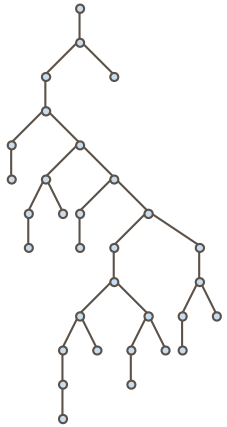
3 $$R(z) = \frac{1 - z^{\theta+2} - \sqrt{1 - 2z^{\theta+2} - 4z^{\theta+4} + z^{2\theta+4}}}{2z^2(1-z)}$$

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

A surprising bijection







Theorem

$\#\pi$ *shapes of size* $n = \#$*Motzkin words of length* $2n+2$

Proof.

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

$$S(z) = 1 + z^2 M(z^2) \quad \Rightarrow \quad s_n = m_{2n+2}$$

□

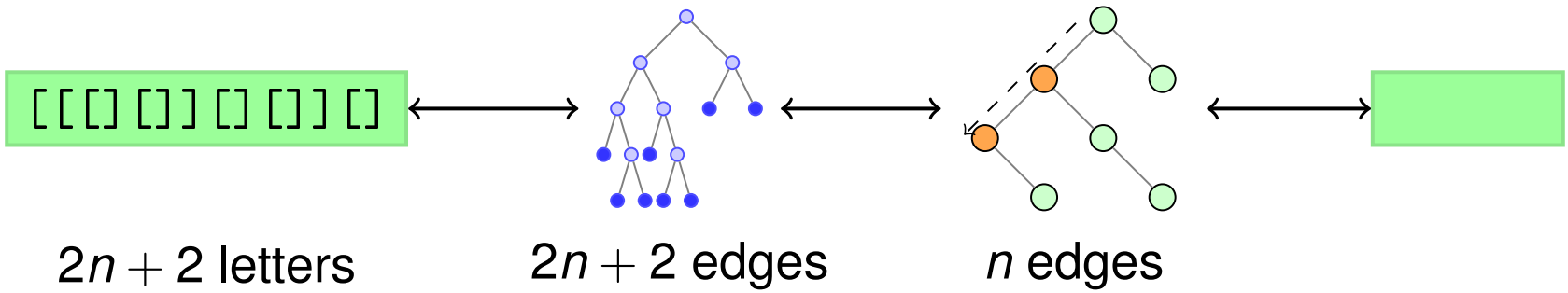
These two classes are in bijection. → **How**? Can we **exploit** it?

Explicit bijection

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

$$\begin{aligned} \psi((\ A\ )\ B) &= \begin{cases} \phi(A) & \text{If } B = \varepsilon \\ \phi(A) \bullet \psi(B) & \text{Otherwise} \end{cases} \\ \phi((\ A\ )\ B) &= \phi(A)[\psi(B)] \\ \phi(\varepsilon) &= \varepsilon. \end{aligned}$$

Then ψ is a **bijection** between s_{2n+2} and m_n.

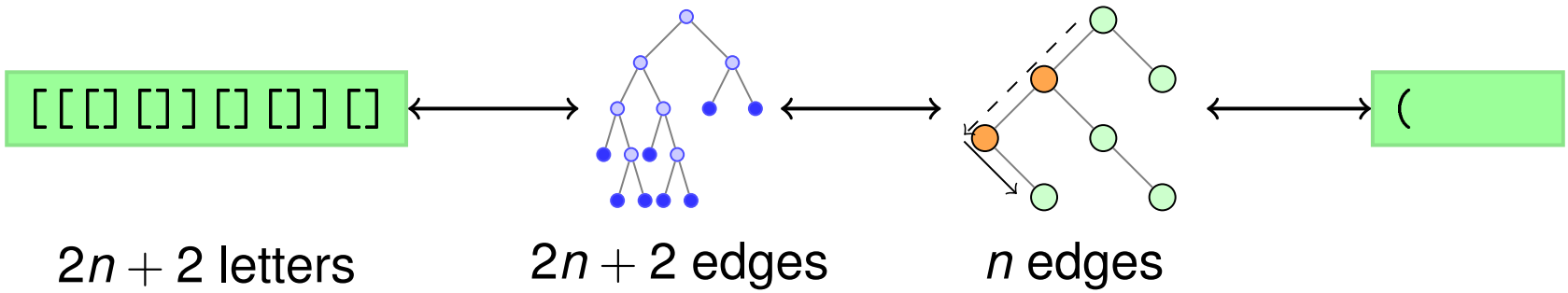


Explicit bijection

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

$$\begin{aligned} \psi((\ A\ )\ B) &= \begin{cases} \phi(A) & \text{If } B = \varepsilon \\ \phi(A) \bullet \psi(B) & \text{Otherwise} \end{cases} \\ \phi((\ A\ )\ B) &= \phi(A)[\psi(B)] \\ \phi(\varepsilon) &= \varepsilon. \end{aligned}$$

Then ψ is a **bijection** between s_{2n+2} and m_n.



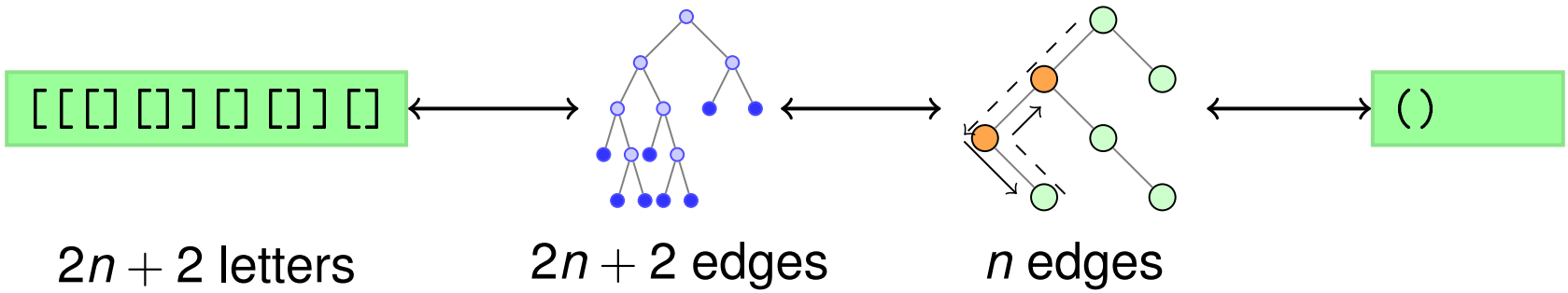
$2n+2$ letters $\quad$ $2n+2$ edges $\quad$ n edges

Explicit bijection

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

$$\begin{aligned} \psi((\ A\ )\ B) &= \begin{cases} \phi(A) & \text{If } B = \varepsilon \\ \phi(A) \bullet \psi(B) & \text{Otherwise} \end{cases} \\ \phi((\ A\ )\ B) &= \phi(A)[\psi(B)] \\ \phi(\varepsilon) &= \varepsilon. \end{aligned}$$

Then ψ is a **bijection** between s_{2n+2} and m_n.

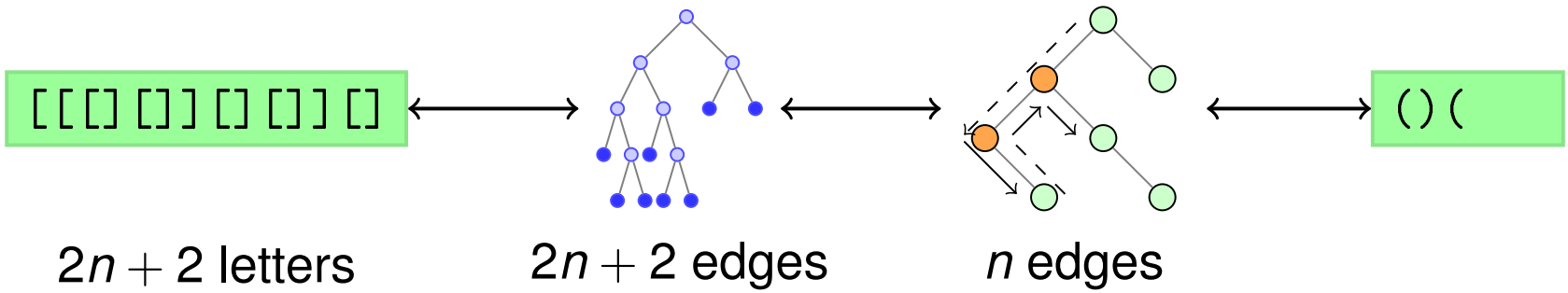


Explicit bijection

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

$$\begin{aligned} \psi((\ A\ )\ B) &= \begin{cases} \phi(A) & \text{If } B = \varepsilon \\ \phi(A) \bullet \psi(B) & \text{Otherwise} \end{cases} \\ \phi((\ A\ )\ B) &= \phi(A)[\psi(B)] \\ \phi(\varepsilon) &= \varepsilon. \end{aligned}$$

Then ψ is a **bijection** between s_{2n+2} and m_n.

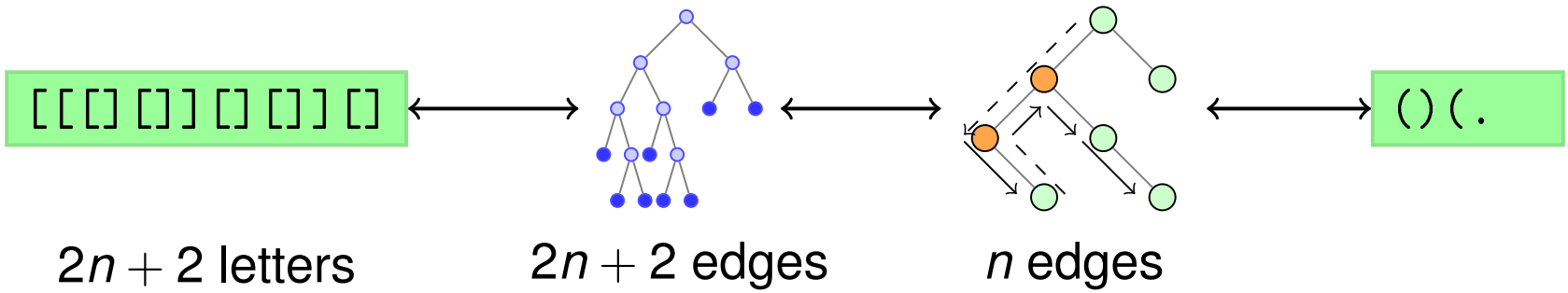


Explicit bijection

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

$$\begin{aligned} \psi((\ A\ )\ B) &= \begin{cases} \phi(A) & \text{If } B = \varepsilon \\ \phi(A) \bullet \psi(B) & \text{Otherwise} \end{cases} \\ \phi((\ A\ )\ B) &= \phi(A)[\psi(B)] \\ \phi(\varepsilon) &= \varepsilon. \end{aligned}$$

Then ψ is a **bijection** between s_{2n+2} and m_n.

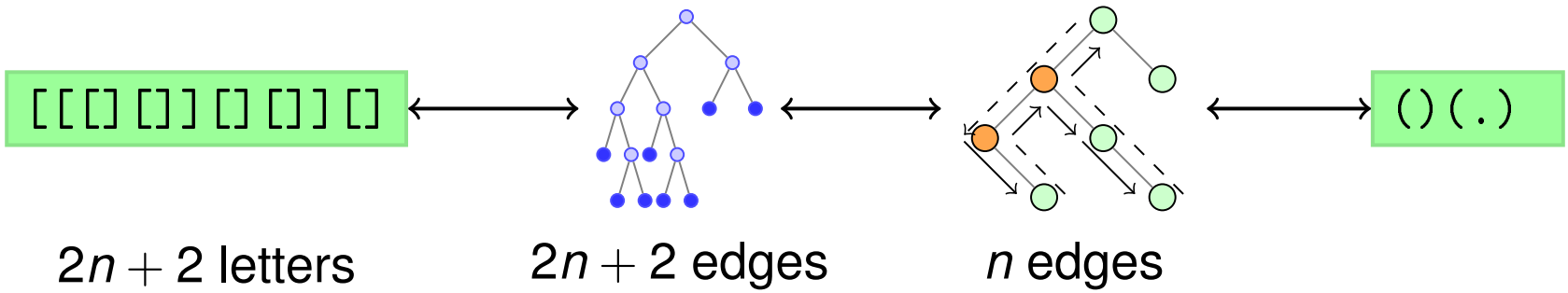


Explicit bijection

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

$$\begin{aligned} \psi((\ A\ )\ B) &= \begin{cases} \phi(A) & \text{If } B = \varepsilon \\ \phi(A) \bullet \psi(B) & \text{Otherwise} \end{cases} \\ \phi((\ A\ )\ B) &= \phi(A)[\psi(B)] \\ \phi(\varepsilon) &= \varepsilon. \end{aligned}$$

Then ψ is a **bijection** between s_{2n+2} and m_n.

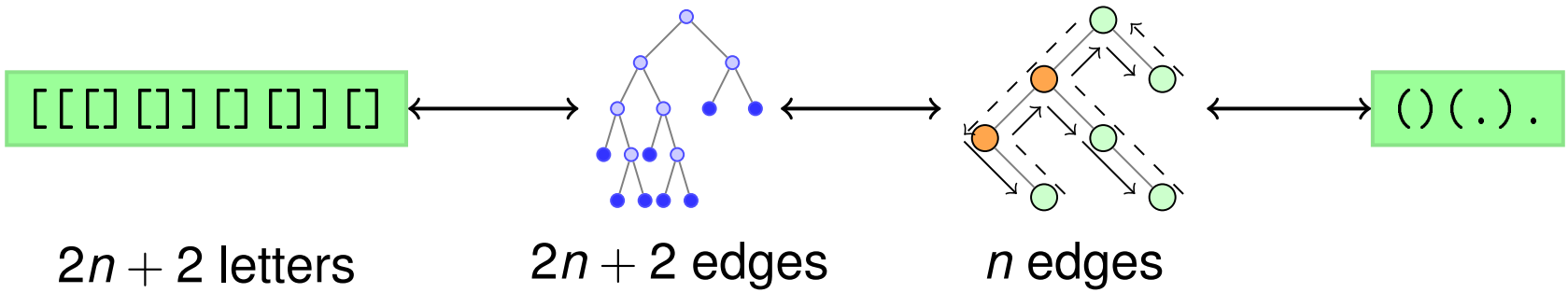


Explicit bijection

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

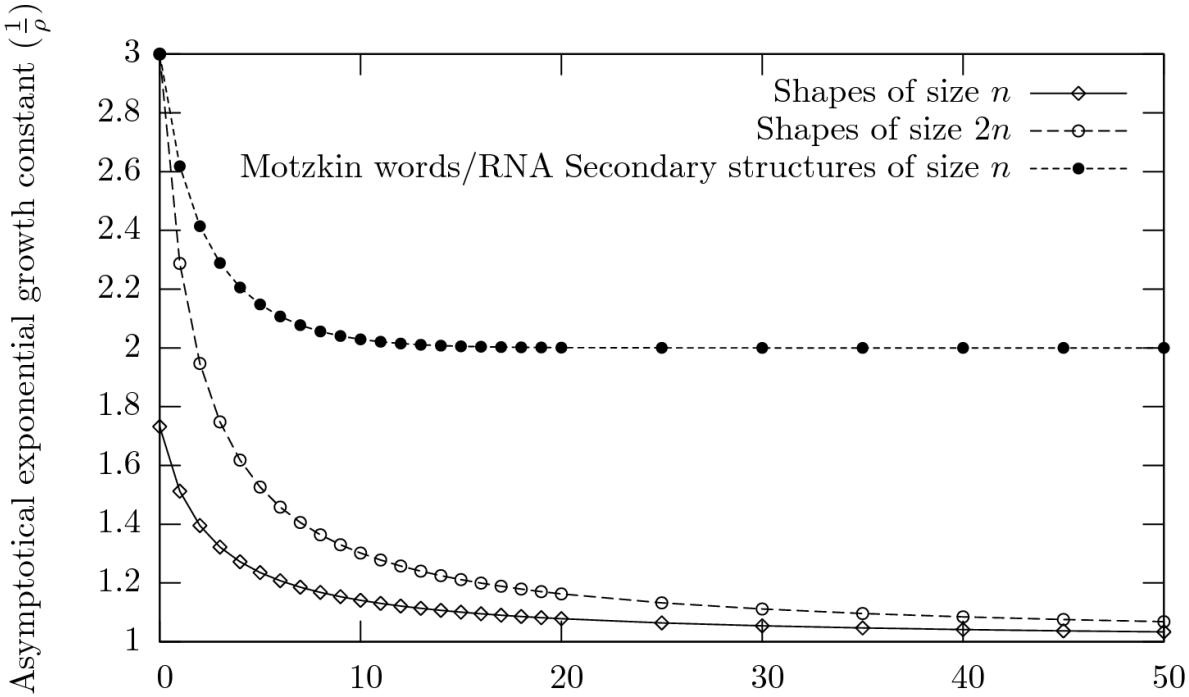
$$\begin{aligned} \psi((\ A\ )\ B) &= \begin{cases} \phi(A) & \text{If } B = \varepsilon \\ \phi(A) \bullet \psi(B) & \text{Otherwise} \end{cases} \\ \phi((\ A\ )\ B) &= \phi(A)[\psi(B)] \\ \phi(\varepsilon) &= \varepsilon. \end{aligned}$$

Then ψ is a **bijection** between s_{2n+2} and m_n.



$2n+2$ letters $\quad$ $2n+2$ edges $\quad$ n edges

Limits of the bijection



Impacts of θ on shapes and Motzkin are drastically different.

Theorem

Expectations of number of term. loops in Motzkin words and π-shapes scale like $m_n^t \sim \frac{n}{6} + \mathcal{O}(1)$ *and* $s_{2n+2}^t \sim \frac{2n}{3} + \mathcal{O}(1)$

π'-shapes

Objective: Count π'-shapes compatible with RNA of length n.

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

2. $$R \to \square\, R \mid S \qquad S \to U\,[\,T\,]\,S \mid U \qquad U \to \diamond \mid \varepsilon$$
$$T \to U\,[\,T\,]\,U\,[\,T\,]\,S \mid \diamond\,[\,T\,] \mid [\,T\,]\,\diamond \mid \diamond\,[\,T\,]\,\diamond \mid \bullet^{\theta}$$

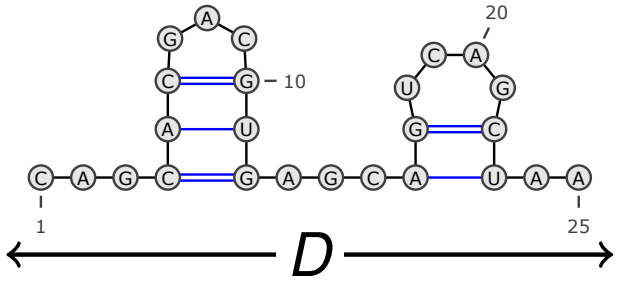
3. $$\theta = 3,\ R(z) = \frac{1+2z^2+2z^3+z^4-z^5-z^6-\sqrt{1-4z^3-2z^4-2z^5+2z^6-7z^8-z^{10}+2z^{11}+z^{12}}}{2z^2(1-z^2)}$$

4. $$r_n \sim 1.27\,\frac{1.81^n}{n\sqrt{n}}$$

Summary

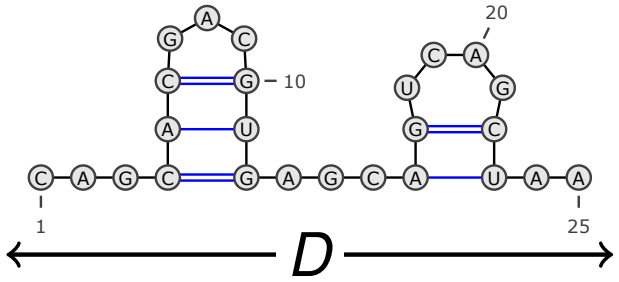
Model	Asymptotic number
Sec. str. on n – Combinatorial	$1.1 \cdot \frac{\mathbf{2.6}^n}{n\sqrt{n}}$
Sec. str. on n – Empirical	$0.04 \cdot \frac{\mathbf{1.4}^n}{n\sqrt{n}}$
π-shapes of size n	$1.38 \cdot \frac{\mathbf{1.73}^n}{n\sqrt{n}}$
π-shapes compatible with sec. str. on n	$2.44 \cdot \frac{\mathbf{1.32}^n}{n\sqrt{n}}$
π-shapes – Empirical	$0.21 \cdot \frac{\mathbf{1.1}^n}{n\sqrt{n}}$
π'-shapes of size n	$0.99 \cdot \frac{\mathbf{2.41}^n}{n\sqrt{n}}$
π'-shapes compatible with sec. str. on n	$1.28 \cdot \frac{\mathbf{1.81}^n}{n\sqrt{n}}$

5'-3' Distance: Theory vs Empirical evidences



[Clote-P-Steyaert 12]

5'-3' Distance: Theory vs Empirical evidences

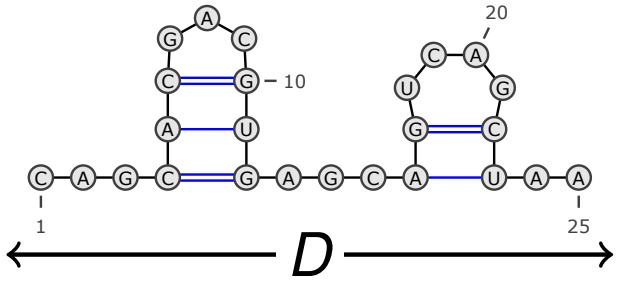


[Clote-P-Steyaert 12]

2

$$T \to [S_{\geq\theta}]T| \bullet T|\varepsilon \qquad S \to (S_{\geq\theta})S| \circ S|\varepsilon$$
$$S_{\geq\theta} \to (S_{\geq\theta})S| \circ S_{\geq\theta}|\circ^{\theta}$$

5'-3' Distance: Theory vs Empirical evidences



[Clote-P-Steyaert 12]

2

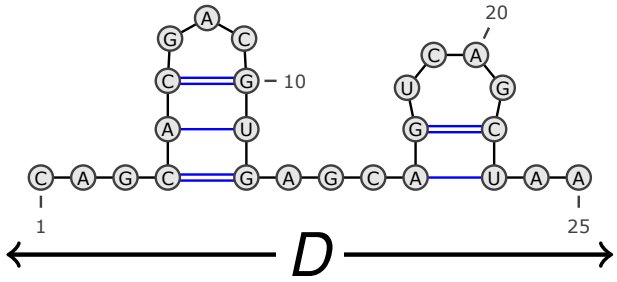
$$T \to [S_{\geq\theta}]T| \bullet T|\varepsilon \qquad S \to (S_{\geq\theta})S| \circ S|\varepsilon$$
$$S_{\geq\theta} \to (S_{\geq\theta})S| \circ S_{\geq\theta}|\circ^{\theta}$$

3

$$E_\theta(z) = \frac{\begin{array}{c} 2 - 9z + 14z^2 - 8z^3 + 2z^5 \\ +z^{\theta+2}(-4 + 10z - 10z^2 + 2z^3) + z^{2\theta+4}(2 - z) \\ -(2 - 5z + 4z^2 - 2z^{\theta+2} + z^{\theta+3})\sqrt{\Delta_\theta} \end{array}}{2(1-z)^2 z^4}$$

$$\Delta_\theta := 1 - 4z + 4z^2 - 2z^{\theta+2} + 4z^{\theta+3} - 4z^{\theta+4} + z^{2\theta+4}$$

5'-3' Distance: Theory vs Empirical evidences



[Clote-P-Steyaert 12]

2

$$T \to [S_{\geq\theta}]T| \bullet T|\varepsilon \qquad S \to (S_{\geq\theta})S| \circ S|\varepsilon$$
$$S_{\geq\theta} \to (S_{\geq\theta})S| \circ S_{\geq\theta}|\circ^{\theta}$$

3

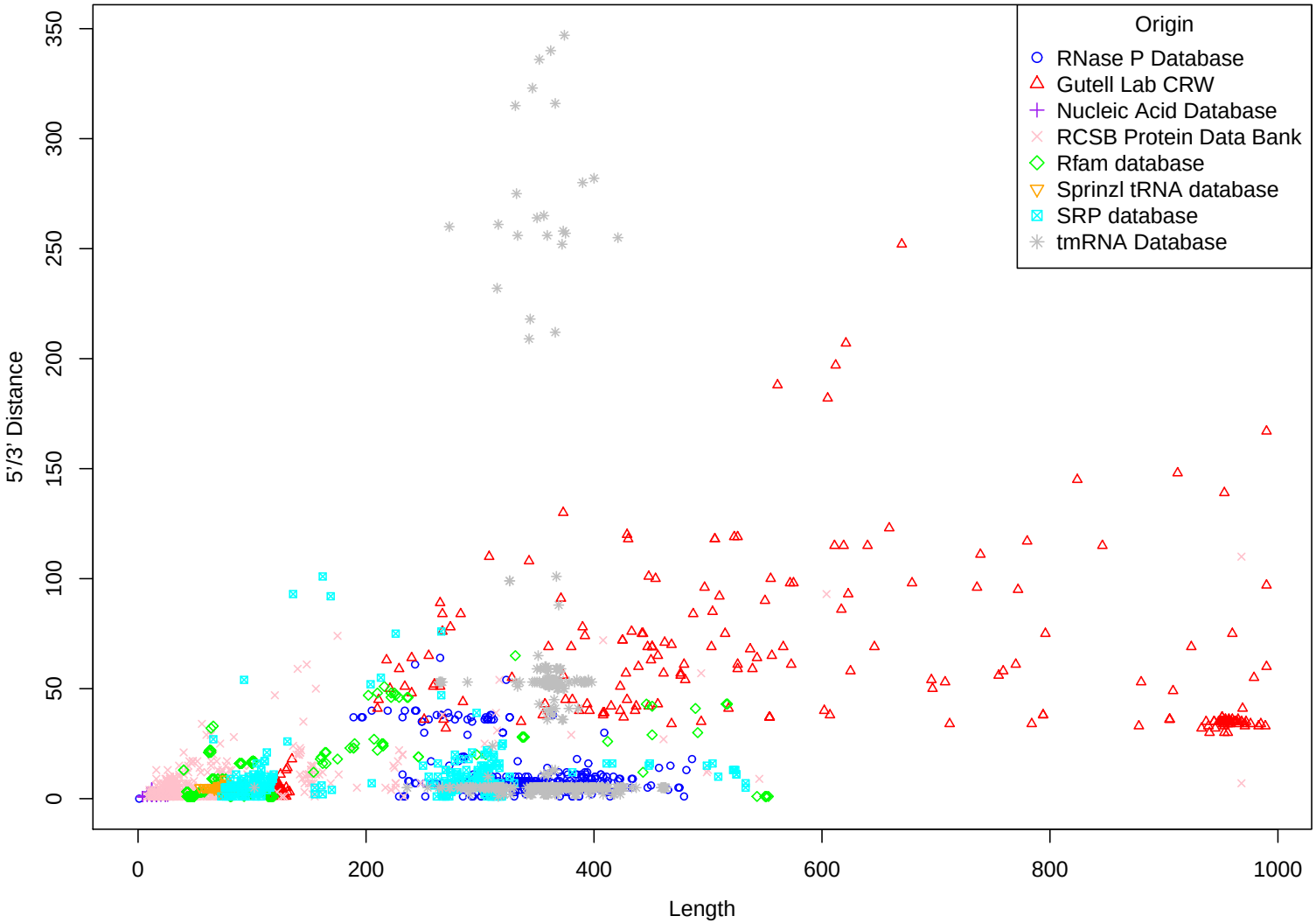
$$E_\theta(z) = \frac{\begin{array}{c} 2 - 9z + 14z^2 - 8z^3 + 2z^5 \\ +z^{\theta+2}(-4 + 10z - 10z^2 + 2z^3) + z^{2\theta+4}(2 - z) \\ -(2 - 5z + 4z^2 - 2z^{\theta+2} + z^{\theta+3})\sqrt{\Delta_\theta} \end{array}}{2(1-z)^2 z^4}$$

$$\Delta_\theta := 1 - 4z + 4z^2 - 2z^{\theta+2} + 4z^{\theta+3} - 4z^{\theta+4} + z^{2\theta+4}$$

4

$$D_n \sim \frac{2-5\rho+4\rho^2-2\rho^{\theta+2}+\rho^{\theta+3}}{(1-\rho)\rho^2} - 1,\ \rho \text{ smallest sol. of } \Delta_\theta = 0$$

5’-3’ Distance: Theory vs Empirical evidences



Empirical distribution of 5’-3’ distance (Strand database)

$\Rightarrow$ Linear correlation Length/Distance [Clote-P-Steyaert 12]. . . **uh?!**

Analytic combinatorics: Limiting distribution

3^{rd} principle (Drmota-Lalley-Woods)

Strongly connected grammar/specification + Technical conditions
$\Rightarrow$ Dominant singularity in $\sqrt{1 - z/\rho}$
$\Rightarrow$ #Occurrences of any letter t follows **Normal** distribution
Expectation: $\mathbb{E}(\#t) \in \Theta(\mu \cdot n)$ **Variance:** $\mathbb{V}(\#t) \in \Theta(\sigma^2 \cdot n)$
+ μ and σ are easy to compute **symbolically** or **numerically**.

Results robust to **weights**...
...or **almost connected** specifications (subcritical composition)...
...generalizes to **higher dimensions** (joint distribution)...
...but any limiting distrib. can be induced by general grammars!

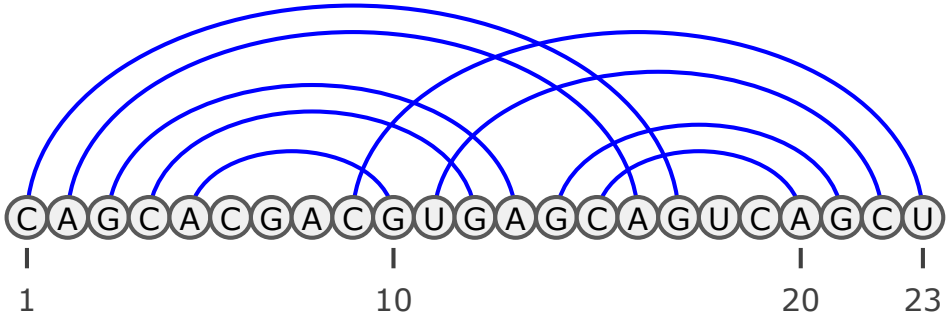
Example (RNA)

Generated by a strongly connected grammar (Nussinov)

$\Rightarrow$ #Base pairs, #Unpaired bases... are Normally distributed

Pseudoknots and partition function

Recursive Pseudoknots, Akutsu-style [Saule et al 10] [Nebel Weinberg 12]



2 $$S \to (S)S \mid \bullet S \mid \Phi\,(\mathrm{Seq}(a.a' \mid b.b'))\,SS \mid \varepsilon$$

Strongly connected grammar

$\Rightarrow$ Partition function approach will always predict PKs for large structures...

...but PK-free structures exist.

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...

Hammer (analytic combinatorics) $\rightarrow$ **nail** (RNA+homopolymer model)
$\Rightarrow$ Time for a **new hammer**?

Collaborators: W. A. Lorenz and P. Clote (Boston College), A. Denise (Univ. Paris Sud), J.-M. Steyaert (Ecole Polytechnique)