Seminar

# Exponentially few RNA structures are designable

YAO, Hua-Ting

Ecole Polytechnique, France
McGill University, Canada

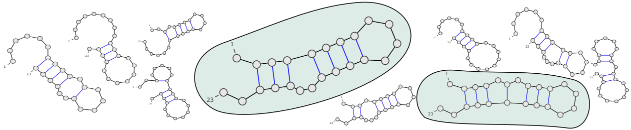
In collaboration with:
Cedric Chauve, Simon Fraser University, Canada

Supervised by:
Mireille Régnier, Ecole Polytechnique, France
Yann Ponty, Ecole Polytechnique, France

35th TBI Winterseminar, Bled — February 12, 2020

# In a nutshell (TL;DR)

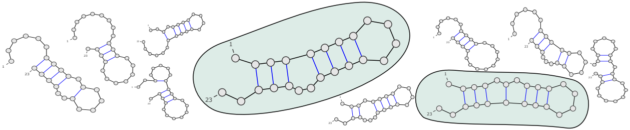
- Adoption of a given structure essential for many RNA function(s)
- #Secondary structure grows exponentially with RNA size $n$ ($\approx 2.6^n$)
- but many structures are too unstable for any sequence




How many RNA structures ($\rightarrow$ functions) can be evolved?

# In a nutshell (TL;DR)

- Adoption of a given structure essential for many RNA function(s)
- #Secondary structure grows exponentially with RNA size $n$ ($\approx 2.6^n$)
- but many structures are too unstable for any sequence

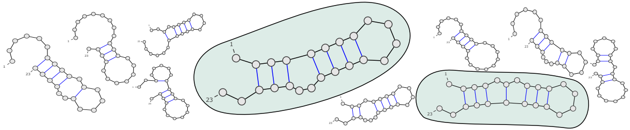


How many RNA structures ($\rightarrow$ functions) can be evolved?

Working hypothesis: Nature solves (at least) a design problem

# In a nutshell (TL;DR)

- Adoption of a given structure essential for many RNA function(s)
- #Secondary structure grows exponentially with RNA size $n$ ($\approx 2.6^n$)
- but many structures are too unstable for any sequence



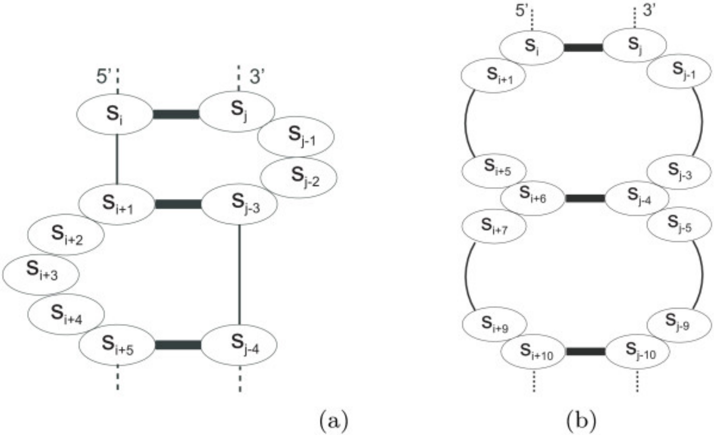
How many RNA structures ($\to$ functions) can be evolved?

Working hypothesis: Nature solves (at least) a design problem

Main results:

- (Algorithmic) discovery of undesignable local motifs
- Proportion of designable structures exponentially decreasing on size

# Some undesignable motifs




(Aguirre-Hernández *et al*, 2007)

# RNA negative design definition

- A sequence $w$ is a negative design for a structure $S^*$ if and only if
  $\rightarrow$ Unique minimum free energy structure, $\text{MFE}(w) = \{S^*\}$
  $\rightarrow$ No other competitive structures, defect $\mathcal{D}(w, S^*) \leq \varepsilon$

# RNA negative design definition

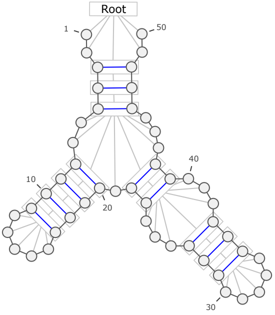
- A sequence $w$ is a negative design for a structure $S^*$ if and only if
  $\rightarrow$ Unique minimum free energy structure, MFE$(w) = \{S^*\}$
  $\rightarrow$ No other competitive structures, defect $\mathcal{D}(w, S^*) \leq \varepsilon$

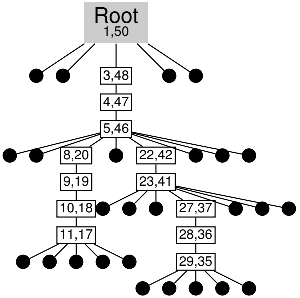
- Classical defects:
  $\rightarrow$ Suboptimal Defect $\mathcal{D}_S$, free-energy dist. to first suboptimal
  $\rightarrow$ Probability Defect $\mathcal{D}_P$, Boltzmann prob. of alternative structures
  $\rightarrow$ Ensemble Defect $\mathcal{D}_E$, expected BP dist. to a random structure

Existence of a negative design NP-hard (Bonnet *et al*, RECOMB 2018)

$\rightarrow$ Counting at least as hard $\rightarrow$ Upper bounds

# RNA secondary structure

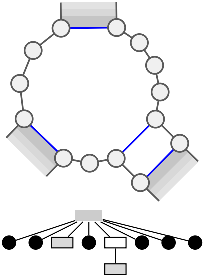


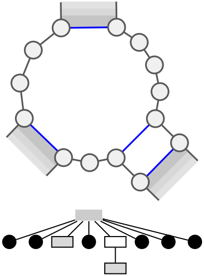

Leaf ● : unpaired base

Internal node ▭ : base pair

# Local motif



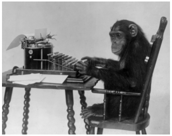
# Local motif



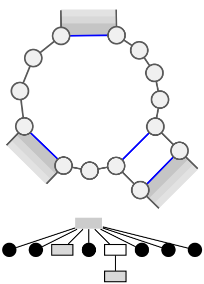
Local motif exceeds defect tolerance
⇒ No structures containing the motif can be designed

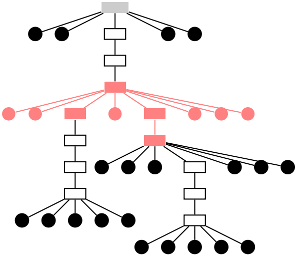
But random RNA structures asymptotically contain every motif

Monkeys and (tree-generating) typewriters paradox. . .



# Local motif





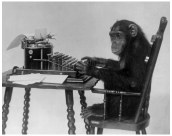
Undesignable

Local motif exceeds defect tolerance

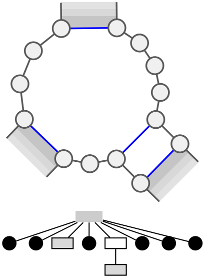
$\Rightarrow$ No structures containing the motif can be designed

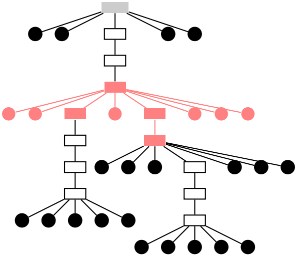
But random RNA structures asymptotically contain every motif

Monkeys and (tree-generating) typewriters paradox. . .

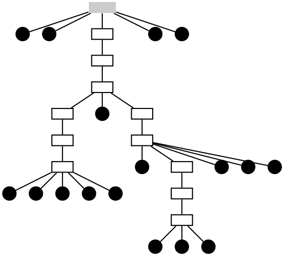


# Local motif





Undesignable



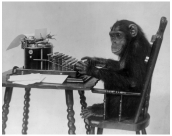
?

Local motif exceeds defect tolerance

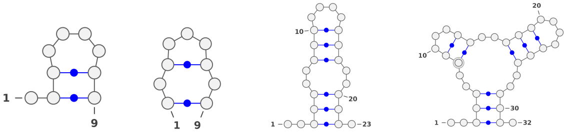
⇒ No structures containing the motif can be designed

But random RNA structures asymptotically contain every motif

Monkeys and (tree-generating) typewriters paradox. . .

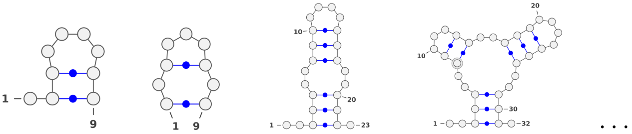


# Object counting

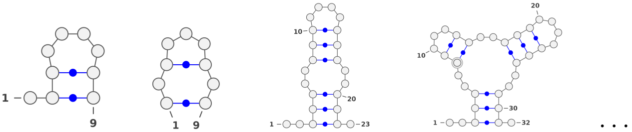



. . .

# Object counting




$z^9$ $\quad$ $z^9$ $\quad$ $z^{23}$ $\quad$ $z^{32}$ $\quad$ $\dots$




$z^9$ $z^9$ $z^{23}$ $z^{32}$ $\dots$

$$\begin{aligned} S(z) &= z^9 + z^9 + z^{23} + z^{32} + \cdots \\ &= 2z^9 + z^{23} + z^{32} + \cdots \end{aligned}$$

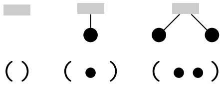
$$S(z) = \sum_{n \geq 0} s_n z^n$$

- $S(z)$: Generating function of structures avoiding undesignable motifs $\mathcal{F}$
  $s_n = [z^n]\, S(z)$ : #Structures of size $n$ avoiding $\mathcal{F}$

$$S(z) = \sum_{n \geq 0} s_n z^n$$

- $S(z)$: Generating function of structures avoiding undesignable motifs $\mathcal{F}$
  $s_n = [z^n]\, S(z)$ : #Structures of size $n$ avoiding $\mathcal{F}$

$\mathcal{F} =$



# Analytic combinatorics

$$S(z) = \sum_{n \geq 0} s_n z^n$$

- $S(z)$: Generating function of structures avoiding undesignable motifs $\mathcal{F}$
  $s_n = [z^n]\, S(z)$ : #Structures of size $n$ avoiding $\mathcal{F}$

$\mathcal{F} =$ 

( )   ( • )   ( • • )

$$S \quad = \quad (\, T_0\, )\, S \mid \bullet\, S \mid \varepsilon$$

# Analytic combinatorics

$$S(z) = \sum_{n \geq 0} s_n z^n$$

- $S(z)$: Generating function of structures avoiding undesignable motifs $\mathcal{F}$
  $s_n = [z^n]\, S(z)$ : #Structures of size $n$ avoiding $\mathcal{F}$

$\mathcal{F} =$

( )    ( • )    ( • • )

$$\begin{aligned} S &= (\, T_0 \,)\, S \mid \bullet S \mid \varepsilon \\ T_0 &= (\, T_0 \,)\, S \mid \bullet T_1 \end{aligned}$$

$$S(z) = \sum_{n \geq 0} s_n z^n$$

- $S(z)$: Generating function of structures avoiding undesignable motifs $\mathcal{F}$
  $s_n = [z^n]\, S(z)$ : #Structures of size $n$ avoiding $\mathcal{F}$

$\mathcal{F} =$

( )    ( • )    ( • • )

$$\begin{aligned} S &= (\, T_0 \,)\, S \mid \bullet S \mid \varepsilon \\ T_0 &= (\, T_0 \,)\, S \mid \bullet T_1 \\ T_1 &= (\, T_0 \,)\, S \mid \bullet T_2 \\ T_2 &= (\, T_0 \,)\, S \mid \bullet S \end{aligned}$$

$$S(z) = \sum_{n \geq 0} s_n z^n$$

- $S(z)$: Generating function of structures avoiding undesignable motifs $\mathcal{F}$
  $s_n = [z^n]\, S(z)$ : #Structures of size $n$ avoiding $\mathcal{F}$

$\mathcal{F} =$ 

( )  ( • )  ( • • )

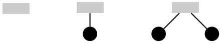
$$\begin{aligned} S &= (\, T_0 \,)\, S \mid \bullet\, S \mid \varepsilon \\ T_0 &= (\, T_0 \,)\, S \mid \bullet\, T_1 \\ T_1 &= (\, T_0 \,)\, S \mid \bullet\, T_2 \\ T_2 &= (\, T_0 \,)\, S \mid \bullet\, S \end{aligned}$$

$$\begin{aligned} S(z) &= z^2\, T_0(z)\, S(z) + z\, S(z) + 1 \\ T_0(z) &= z^2\, T_0(z)\, S(z) + z\, T_1(z) \\ T_1(z) &= z^2\, T_0(z)\, S(z) + z\, T_2(z) \\ T_2(z) &= z^2\, T_0(z)\, S(z) + z\, S(z) \end{aligned}$$

# Analytic combinatorics

$$S(z) = \sum_{n \geq 0} s_n z^n$$

- $S(z)$: Generating function of structures avoiding undesignable motifs $\mathcal{F}$
  $s_n = [z^n]\, S(z)$ : #Structures of size $n$ avoiding $\mathcal{F}$

$\mathcal{F} =$ 

( )   ( • )   ( • • )

$$\begin{aligned} S &= (\, T_0 \,)\, S \mid \bullet S \mid \varepsilon \\ T_0 &= (\, T_0 \,)\, S \mid \bullet T_1 \\ T_1 &= (\, T_0 \,)\, S \mid \bullet T_2 \\ T_2 &= (\, T_0 \,)\, S \mid \bullet S \end{aligned}$$

$$\begin{aligned} S(z) &= z^2\, T_0(z)\, S(z) + z\, S(z) + 1 \\ T_0(z) &= z^2\, T_0(z)\, S(z) + z\, T_1(z) \\ T_1(z) &= z^2\, T_0(z)\, S(z) + z\, T_2(z) \\ T_2(z) &= z^2\, T_0(z)\, S(z) + z\, S(z) \end{aligned}$$

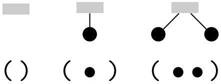
$$z^2 S(z)^2 - (z^4 + z^3 + z^2 - z + 1) S(z) + 1 = 0$$

# Analytic combinatorics

$$S(z) = \sum_{n\geq 0} s_n z^n = \frac{z^4+z^3+z^2-z+1-\sqrt{(z^4+z^3+z^2-z+1)^2-4z^2}}{2z^2}$$

- $S(z)$: Generating function of structures avoiding undesignable motifs $\mathcal{F}$
  $s_n = [z^n]\,S(z)$ : #Structures of size $n$ avoiding $\mathcal{F}$

$\mathcal{F} =$




$$\begin{aligned} S &= (\,T_0\,)\,S \mid \bullet S \mid \varepsilon \\ T_0 &= (\,T_0\,)\,S \mid \bullet T_1 \\ T_1 &= (\,T_0\,)\,S \mid \bullet T_2 \\ T_2 &= (\,T_0\,)\,S \mid \bullet S \end{aligned}$$

$$\begin{aligned} S(z) &= z^2\,T_0(z)\,S(z) + z\,S(z) + 1 \\ T_0(z) &= z^2\,T_0(z)\,S(z) + z\,T_1(z) \\ T_1(z) &= z^2\,T_0(z)\,S(z) + z\,T_2(z) \\ T_2(z) &= z^2\,T_0(z)\,S(z) + z\,S(z) \end{aligned}$$

$$z^2S(z)^2 - (z^4+z^3+z^2-z+1)S(z) + 1 = 0$$

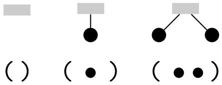
$$S(z) = \sum_{n\geq 0} s_n z^n = \frac{z^4 + z^3 + z^2 - z + 1 - \sqrt{(z^4 + z^3 + z^2 - z + 1)^2 - 4z^2}}{2z^2}$$

- $S(z)$: Generating function of structures avoiding undesignable motifs $\mathcal{F}$
  $s_n = [z^n]\, S(z)$ : #Structures of size $n$ avoiding $\mathcal{F}$
- Dominant singularity $\rho$ of $S(z)$ drives asymptotics

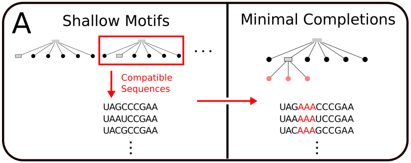
$$[z^n]\, S(z) \in \Theta\left(\frac{\rho^{-n}}{n\sqrt{n}}\right)$$

Example: For motifs below, $s_n \equiv 2.289^n$ (vs $2.618^n$ for all 2D structs)

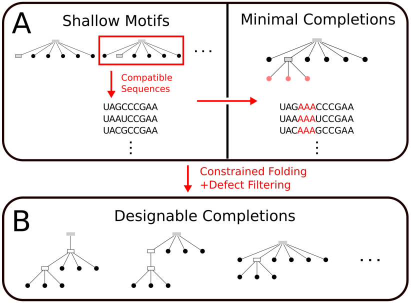
$\mathcal{F} =$

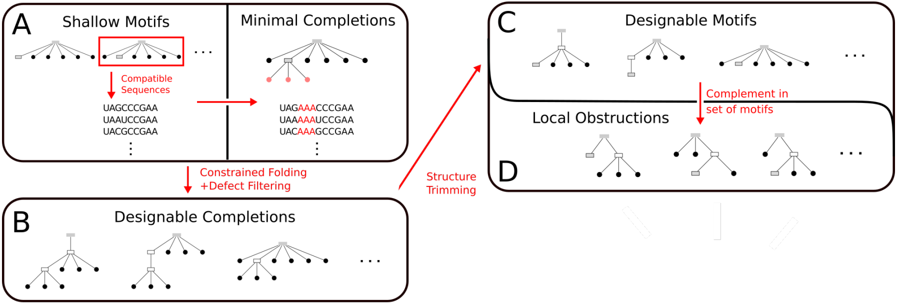


# Workflow

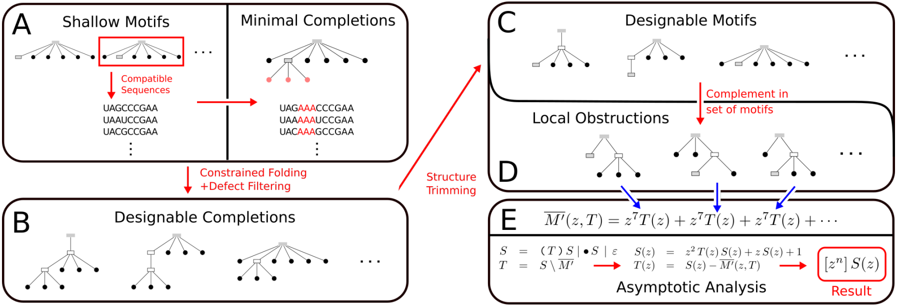

A
Shallow Motifs
Compatible Sequences
UAGCCCGAA
UAAUCCGAA
UACGCCGAA
Minimal Completions
UAGAAACCCGAA
UAAAAAUCCGAA
UACAAAGCCGAA

# Workflow


A
Shallow Motifs
Compatible Sequences
UAGCCCGAA
UAAUCCGAA
UACGCCGAA
Minimal Completions
UAGAAACCCGAA
UAAAAAUCCGAA
UACAAAGCCGAA
Constrained Folding
+Defect Filtering
B
Designable Completions


A
Shallow Motifs
Compatible
Sequences
UAGCCCGAA
UAAUCCGAA
UACGCCGAA
Minimal Completions
UAGAAACCCGAA
UAAAAAUCCGAA
UACAAAGCCGAA
Constrained Folding
+Defect Filtering
B
Designable Completions
Structure
Trimming
C
Designable Motifs
Complement in
set of motifs
Local Obstructions
D

# Workflow

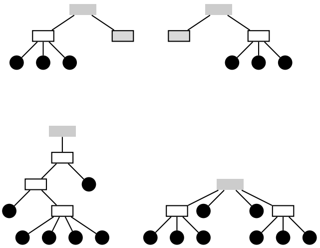


A
Shallow Motifs
Compatible Sequences
UAGCCCGAA
UAAUCCGAA
UACGCCGAA
Minimal Completions
UAGAAACCCGAA
UAAAAAUCCGAA
UACAAAGCCGAA
Constrained Folding +Defect Filtering
B
Designable Completions
Structure Trimming
C
Designable Motifs
Complement in set of motifs
Local Obstructions
D
E
$\overline{M'}(z,T) = z^7T(z) + z^7T(z) + z^7T(z) + \cdots$
$S = (T)S \mid \bullet S \mid \varepsilon$
$T = S \setminus \overline{M'}$
$S(z) = z^2T(z)S(z) + zS(z) + 1$
$T(z) = S(z) - \overline{M'}(z,T)$
Asymptotic Analysis
$[z^n]\,S(z)$
Result

# Undesignable motifs

A sequence $w$ is a negative design for a structure $S^*$ if and only if
$\rightarrow$ Unique minimum free energy structure, MFE$(w) = \{S^*\}$
$\rightarrow$ No other competitive structures, $\mathcal{D}(w, S^*) \leq \varepsilon$

- $\mathcal{D}_S \leq 1, 104$ local motifs



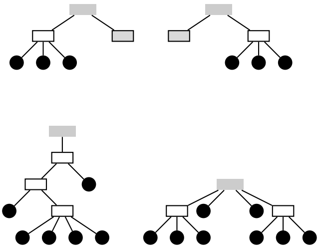
# Undesignable motifs

A sequence $w$ is a negative design for a structure $S^*$ if and only if

$\rightarrow$ Unique minimum free energy structure, MFE$(w) = \{S^*\}$

$\rightarrow$ No other competitive structures, $\mathcal{D}(w, S^*) \leq \varepsilon$

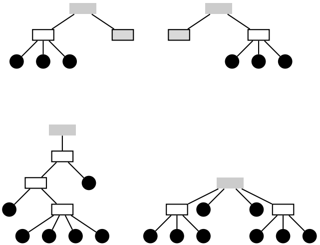
- $\mathcal{D}_S \leq 1$, $104$ local motifs
- $\mathcal{D}_P \leq 0.5$, $117$ local motifs



# Undesignable motifs

A sequence $w$ is a negative design for a structure $S^*$ if and only if
$\rightarrow$ Unique minimum free energy structure, MFE$(w) = \{S^*\}$
$\rightarrow$ No other competitive structures, $\mathcal{D}(w, S^*) \leq \varepsilon$

- $\mathcal{D}_S \leq 1$, $104$ local motifs
- $\mathcal{D}_P \leq 0.5$, $117$ local motifs
- $\mathcal{D}_P \leq 0.1$, $152$ local motifs
- $\mathcal{D}_P \leq 0.01$, $174$ local motifs



# Asymptotic results

| Defect | $\varepsilon$ | Asymptotic equivalent | Proportion (vs $2.289^n$) | | |
|---|---|---|---|---|---|
| | | | $P_{50}$ (%) | $P_{100}$ (%) | $P_{1000}$ (%) |
| $\mathcal{D}_S$ | 1 | $\Theta\left(\frac{2.226^n}{n\sqrt{n}}\right)$ | 25.4 | 6.48 | $1.30 \cdot 10^{-10}$ |
| $\mathcal{D}_P$ | .5 | $\Theta\left(\frac{2.224^n}{n\sqrt{n}}\right)$ | 24.2 | 5.84 | $4.64 \cdot 10^{-11}$ |
| $\mathcal{D}_P$ | .1 | $\Theta\left(\frac{2.176^n}{n\sqrt{n}}\right)$ | 7.69 | 0.59 | $5.29 \cdot 10^{-21}$ |
| $\mathcal{D}_P$ | .01 | $\Theta\left(\frac{2.078^n}{n\sqrt{n}}\right)$ | 0.80 | $6.44 \cdot 10^{-3}$ | $1.22 \cdot 10^{-40}$ |

Note: Asymptotic equivalents are upper bound

Exact proportion of designable structures could be even lower...

# Conclusions/perspectives

https://gitlab.com/htyao/countingdesign/

- Proportion of designable structures decreases exponentially
  → Library-based approaches for design (Bellaousov*et al*, RNA 2018)
  → Revisit neutral networks theory

# Conclusions/perspectives

https://gitlab.com/htyao/countingdesign/

- Proportion of designable structures decreases exponentially
  $\rightarrow$ Library-based approaches for design (Bellaousov*et al*, RNA 2018)
  $\rightarrow$ Revisit neutral networks theory
- Extends to pseudoknotted structures
  $\rightarrow$ Multiple grammars $\rightarrow$ Same combinatorial prop.

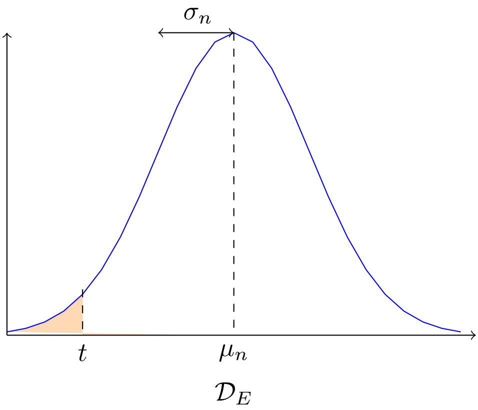
# Conclusions/perspectives

https://gitlab.com/htyao/countingdesign/

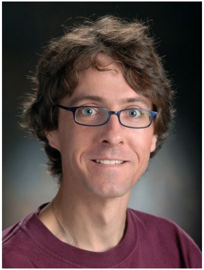
- Proportion of designable structures decreases exponentially
  → Library-based approaches for design (Bellaousov*et al*, RNA 2018)
  → Revisit neutral networks theory
- Extends to pseudoknotted structures
  → Multiple grammars → Same combinatorial prop.
- Better upper bounds for popular ensemble defect
  → Bivariate generating functions

https://gitlab.com/htyao/countingdesign/

- Better upper bounds for popular ensemble defect
  → Bivariate generating functions

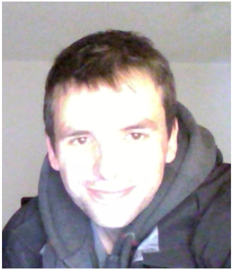


# Acknowledgement



Cedric Chauve
Simon Fraser University
Canada

Mireille Régnier
Ecole Polytechnique
France



Yann Ponty
Ecole Polytechnique
France

# Backup slides

# Defect and RNA negative design

- Defect: $\mathcal{D} : \Sigma^* \times \mathcal{S} \to \mathbb{R}$
  - Suboptimal Defect $\mathcal{D}_S$

$$\log \mathcal{D}_S(w, S^*) := - \min_{\substack{S \in \mathcal{S}_{|w|} \\ S \neq S^*}} E(w, S) - E(w, S^*);$$

  - Probability Defect $\mathcal{D}_P$

$$\mathcal{D}_P(w, S^*) := \sum_{\substack{S \in \mathcal{S}_{|w|} \\ S \neq S^*}} \mathbb{P}(S \mid w) = 1 - \mathbb{P}(S^* \mid w);$$

# Defect and RNA negative design

- Defect: $\mathcal{D} : \Sigma^* \times \mathcal{S} \to \mathbb{R}$
  - Suboptimal Defect $\mathcal{D}_S$

$$\log \mathcal{D}_S(w, S^*) := -\min_{\substack{S \in \mathcal{S}_{|w|} \\ S \neq S^*}} E(w, S) - E(w, S^*);$$

  - Probability Defect $\mathcal{D}_P$

$$\mathcal{D}_P(w, S^*) := \sum_{\substack{S \in \mathcal{S}_{|w|} \\ S \neq S^*}} \mathbb{P}(S \mid w) = 1 - \mathbb{P}(S^* \mid w);$$

- Given $\varepsilon \geq 0$ and a defect $\mathcal{D}$, a sequence $w$ is a (negative) $(\mathcal{D}, \varepsilon)$-design for a structure $S^*$ if and only if

$$\text{MFE}(w) = \{S^*\} \quad \text{and} \quad \mathcal{D}(w, S^*) \leq \varepsilon$$

# Methods

$$
\begin{aligned}
S &= (\, T\,)\, S \mid \bullet\, S \mid \varepsilon \\
T &= S \setminus \overline{M'}
\end{aligned}
$$

where

$$\overline{M'} := \{m' \mid \forall m \in \overline{\mathcal{M}}, m = (\, m'\,)\}$$

# Methods

$$
\begin{aligned}
S &= (\, T \,)\, S \mid \bullet\, S \mid \varepsilon \\
T &= S \setminus \overline{M'}
\end{aligned}
$$

where

$$\overline{M'} := \{ m' \mid \forall m \in \overline{\mathcal{M}}, m = (\, m' \,) \}$$

$$
\begin{aligned}
S(z) &= z^2\, T(z)\, S(z) + z\, S(z) + 1 \\
T(z) &= S(z) - \overline{M'}(z, T)
\end{aligned}
$$

where

$$\overline{M'}(z, T) = \sum_{m' \in \overline{\mathcal{M}'}} z^{\gamma(m')}\, T^{\delta(m')} - c(z, T)$$