Weighted random generation of context-free languages: Analysis of collisions in random urn occupancy models

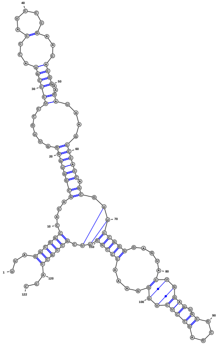
Danièle Gardy Yann Ponty

PRiSM – Université de Versailles St-Quentin en Yvelines – France
LIX – Polytechnique/CNRS/INRIA AMIB – France

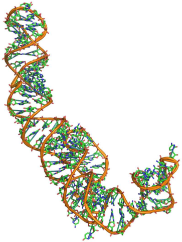
Three[1] levels of representation:

UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC

Primary structure



Secondary structure



Tertiary structures

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

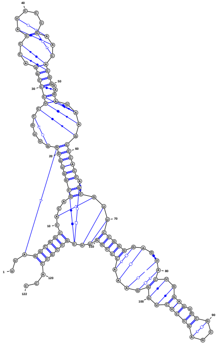
[1] Well, almost...

RNA structure(s)

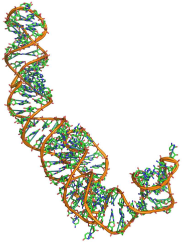
Three[1] levels of representation:

```
UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC
```

Primary structure



Secondary$^+$ structure

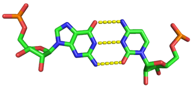


Tertiary structures

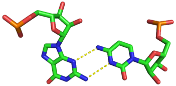
Source: 5s rRNA (PDB 1K73:B)

[1] Well, almost...

- Non-canonical base-pairs
  Any basepair other than {(A-U), (C-G), (G-U)}
  Or interacting using a non-standard edge/orientation (WC/WC-Cis) [LW01].

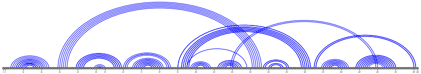


C/G canonical pair (WC/WC-Cis)



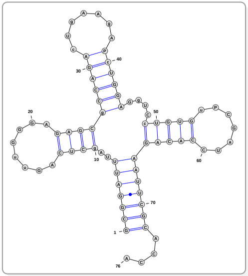
CG non-canonical pair (Sugar/WC-Trans)

- Pseudoknots



Pseudoknots within a group I Ribozyme (PDBID: 1Y0Q:A)

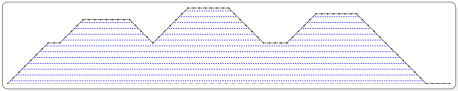
More expressive model, but *ab initio* folding with pseudoknots:
$\Rightarrow$ NP-Complete [LP00]... yet polynomial for restricted classes [CDR+04].



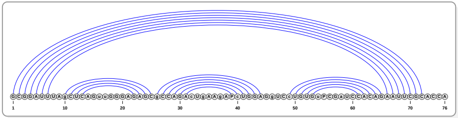
Outer planar graphs

```
(((((((..((((........))))((((((.......))))))....(((((.......)))))))))))).... 
```

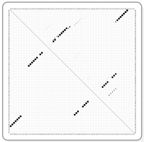
Well-parenthesized expressions



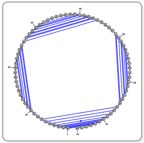
Mountain view



Linear



Dot plot

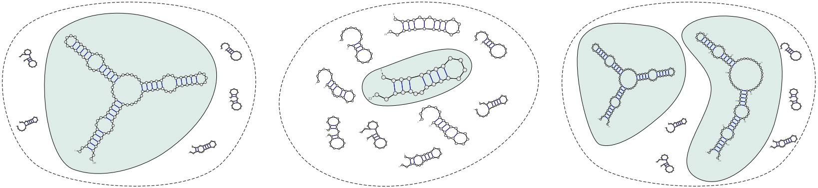


Feynman diagrams

Many representations **but**
Common combinatorial structure

Secondary structures = Motzkin words avoiding *plateaux* ($\bullet \cdots \bullet$) of width $< \theta$.

Driving hypothesis for the RNA folding community assumes a Boltzmann distribution $e^{\frac{-E}{RT}}$ based on free-energy on secondary structures compatible with base-pairing constraints.



Functional folding? | Ill-defined folding: mRNA? | Bistable RNA Kinetics?

Gold standard method: To build a consensus based on a representative sample of the Boltzmann ensemble of low-energy.
$\Rightarrow$ (Weighted)-random generation of **1000** structures and clustering.

Initial question: Is this magic number sufficient? (What is *sufficient*?)

Generalization: Drop base-pairing compatibility constraint...

Secondary structures	$\rightsquigarrow$	Context-free language
Boltzmann factor	$\rightsquigarrow$	Multiplicative weight

Starting point: (Weighted-)Random generation yields (huge) redundancy
Remark #1: Redundancy does not teach us anything.
Remark #2: Some structures/features might be obfuscated by heaviest structures.

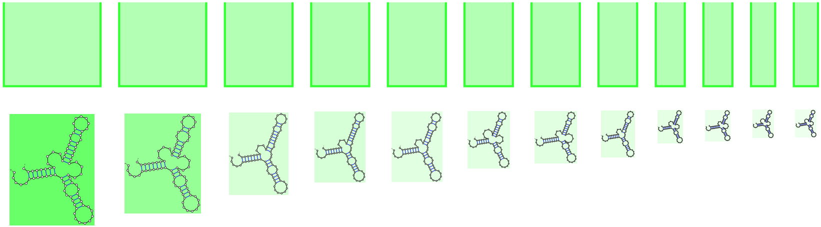
Natural questions:

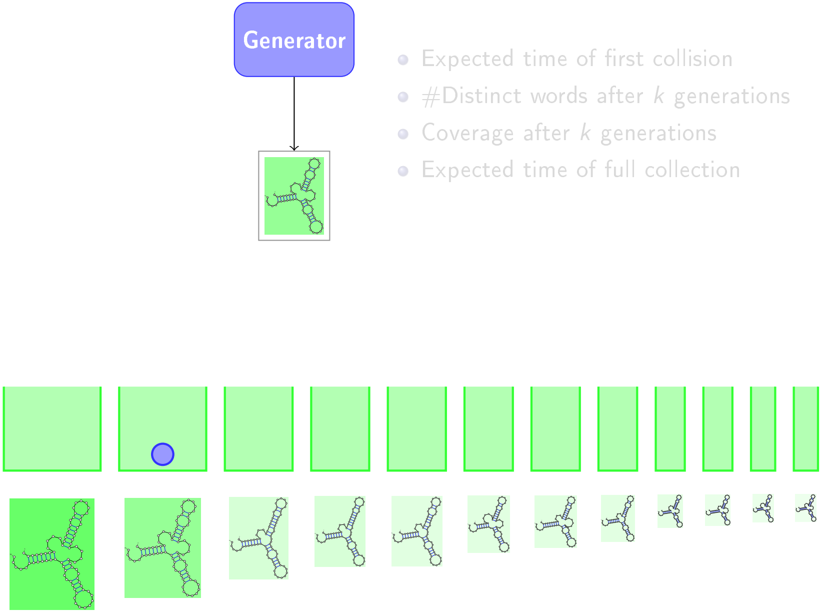
- **Q1** How many generations are required before some word is drawn twice?
- **Q2** How many words must be sampled before each word is encountered at least once?
- **Q3** How many distinct words are there after sampling k objects?
- **Q4** What is the cumulated non-redundant probability after k generations?

The random allocation analogy

Generator

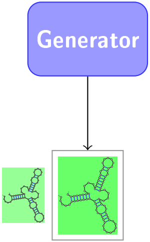
- Expected time of first collision
- #Distinct words after k generations
- Coverage after k generations
- Expected time of full collection



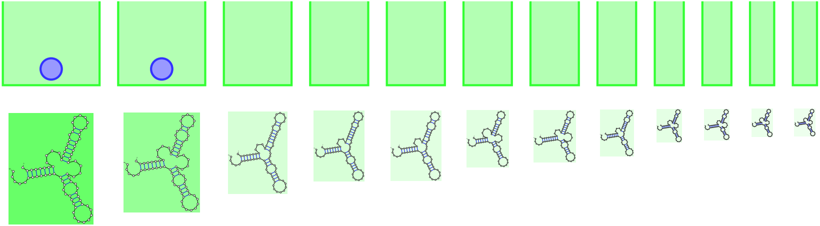


- Expected time of first collision
- #Distinct words after k generations
- Coverage after k generations
- Expected time of full collection

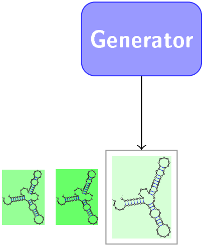
The random allocation analogy



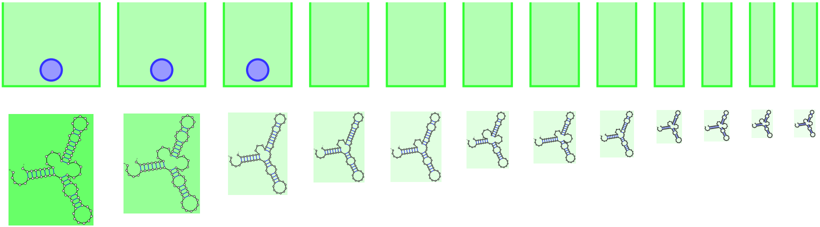
- Expected time of first collision
- #Distinct words after k generations
- Coverage after k generations
- Expected time of full collection

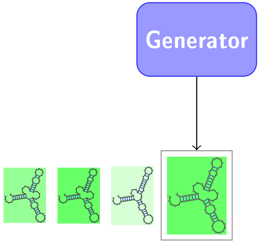


The random allocation analogy

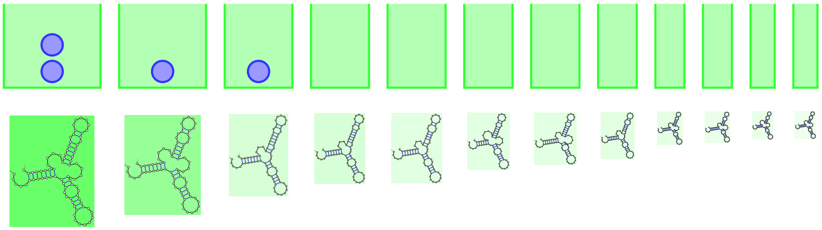


- Expected time of first collision
- #Distinct words after k generations
- Coverage after k generations
- Expected time of full collection

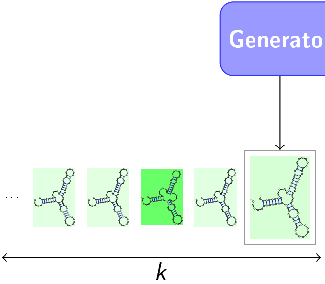




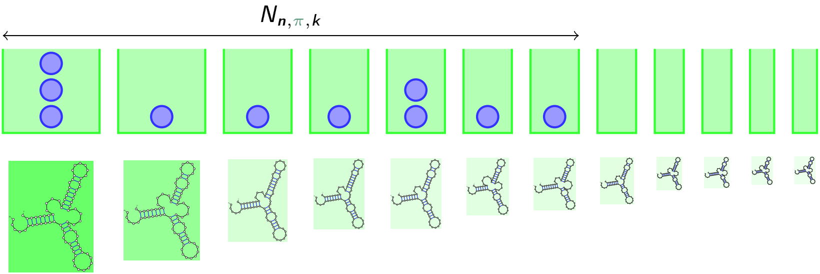
- Expected time of first collision
- #Distinct words after k generations
- Coverage after k generations
- Expected time of full collection



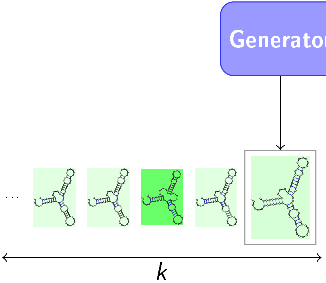
The random allocation analogy



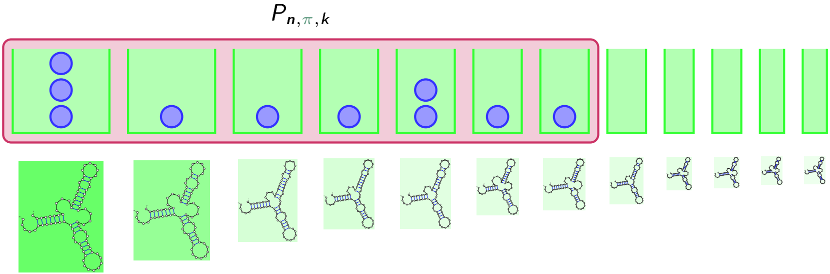
- Expected time of first collision
- **#Distinct words after k generations**
- Coverage after k generations
- Expected time of full collection

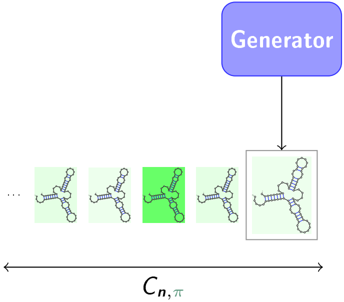


The random allocation analogy

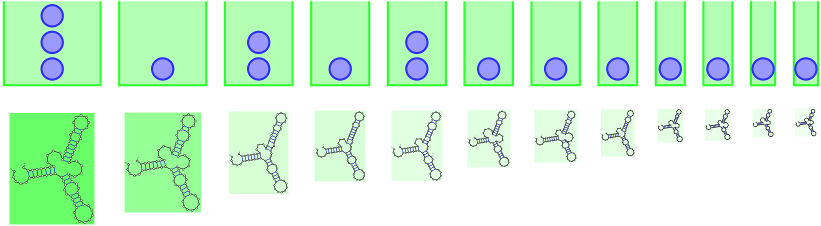


- Expected time of first collision
- #Distinct words after k generations
- Coverage after k generations
- Expected time of full collection





- Expected time of first collision
- #Distinct words after k generations
- Coverage after k generations
- Expected time of full collection



Definition (Context-free grammar)

Context-free grammar = 4-tuple $(\Sigma, \mathcal{N}, \mathcal{P}, \mathcal{S})$:

- Σ: Alphabet.
- $\mathcal{N}$: Non-terminal symbols.
- $\mathcal{P}$: Set of production rules $N \to X \in \mathcal{N} \times \{\Sigma \cup \mathcal{N}\}^*$.
- $\mathcal{S}$: Axiom, or initial non-terminal.

Alt.: Context-free grammar = **admissible specification** using:

- Operators $\{\times, +\}$
- Finite set of atoms $\{Z_1, Z_2, \ldots, Z_k\}$
- Empty structure 1

Definition (Weighted context-free grammar [DRT00])

A **weighted** context-free grammar is a **5**-tuple $\mathcal{G} = (\Sigma, \mathcal{N}, \mathcal{P}, \mathcal{S}, \boldsymbol{\pi})$:

- Σ, $\mathcal{N}$, $\mathcal{P}$, $\mathcal{S}$: Same as previously.
- π: **Weight function** $\pi : \Sigma \to \mathbb{R}$.

Consider the set $\mathcal{L}_n$ of words of length n generated by $\mathcal{G}$.

Definition (Weighted probability distribution)

A WCFG $\mathcal{G}$ implicitly defines a weighted probability distribution over $\mathcal{L}_n$:

$$\forall \omega \in \mathcal{L}_n,\ \mathbb{P}(\omega) = \frac{\pi(\omega)}{\mu_{n,\pi}}.$$

where $\mu_{n,\pi} = \sum_{w \in \mathcal{L}_n} \pi(w)$ is the **total weight** of $\mathcal{L}_n$ (*partition function*).

Generating k words of size n is in $\mathcal{O}\left(n^2 + n\log(n).k\right)^*$.

Furthermore, aiming at **observed** terminal frequencies:

$\Rightarrow$ Asymptotic weights can *sometimes* be derived analytically [DRT00]

$\Rightarrow$ Weights can be determined (Newton iteration)

Weighted grammars

Definition (Weighted context-free grammar [DRT00])

A **weighted** context-free grammar is a **5**-tuple $\mathcal{G} = (\Sigma, \mathcal{N}, \mathcal{P}, \mathcal{S}, \boldsymbol{\pi})$:

- Σ, $\mathcal{N}$, $\mathcal{P}$, $\mathcal{S}$: Same as previously.
- π: **Weight function** $\pi : \Sigma \to \mathbb{R}$.

Consider the set $\mathcal{L}_n$ of words of length n generated by $\mathcal{G}$.

Definition (Weighted probability distribution)

A WCFG $\mathcal{G}$ implicitly defines a weighted probability distribution over $\mathcal{L}_n$:

$$\forall \omega \in \mathcal{L}_n,\ \mathbb{P}(\omega) = \frac{\pi(\omega)}{\mu_{n,\pi}}.$$

where $\mu_{n,\pi} = \sum_{w \in \mathcal{L}_n} \pi(w)$ is the **total weight** of $\mathcal{L}_n$ (*partition function*).

Generating k words of size n is in $\mathcal{O}\left(n^2 + n\log(n).k\right)^*$.

Furthermore, aiming at **observed** terminal frequencies:

$\Rightarrow$ Asymptotic weights can *sometimes* be derived analytically [DRT00]

$\Rightarrow$ Weights can be determined (Newton iteration)

Definition (Weighted generating function)

A **weighted generating function** $L_\pi(z)$ can be defined as

$$L_\pi(z) \equiv \sum_{w\in\mathcal{L}} \pi(w) z^{|w|} = \sum_{n\geq 0} \mu_{\pi,n} z^n$$

G.f. is constructible as a solution of a weighted system of algebraic equations.

Assuming unicity of the dom. sing, the asymptotics of the total weight follow

$$[z^n]\ L_\pi(z) = \mu_{\pi,n} \sim \kappa_\pi \cdot \rho_\pi^{-n} \cdot n^{-k_\pi} \left(1 + \mathcal{O}\left(n^{-k'_\pi}\right)\right).$$

Definition (Asymptotics of total weights)

The *k*-**th moment** of a π-weighted distribution is given by

$$\alpha_{k,n} = \sum_{i=1}^{m_n} p_i^k = \frac{\sum_{w\in\mathcal{L}_n} \pi(w)^k}{{\mu_{\pi,n}}^k} = \frac{\mu_{\pi^k,n}}{{\mu_{\pi,n}}^k}.$$

C1 Diversity: The probability $p^{\triangle}_{n,\pi}$ of the most probable word within $\mathcal{L}_n$ decreases exponentially with n.

C2 Log-positive weights: For each terminal symbol $t \in \Sigma$, $\pi_t > 1$.
$\Rightarrow$ No loss of generality (weighted distribution stable through rescaling)

C3 Bounded dependency: For any rational number $k > 1$ and any weight vector π such that Condition **C2** holds, ${\rho_\pi}^k < \rho_{\pi^k}$ holds.
(Corollary of **C1**)

Theorem (First collision)

Under conditions **C1**, **C2** *and* **C3**, *the expected number of generations $E[B_{n,\pi}]$ before some word of $\mathcal{L}_n$ is drawn twice is such that*

$$E[B_{n,\pi}] \sim \frac{\sqrt{\pi}}{\sqrt{2\alpha_{2,n}}} = \frac{\mu_{\pi,n}\sqrt{\pi}}{\sqrt{2\mu_{\pi^2,n}}} \in \Omega\left(\gamma^n\right), \quad \gamma := \frac{\rho_\pi}{\sqrt{\rho_{\pi^2}}} > 1$$

Proof: Let $\lambda(t) = \prod_{i=1}^{m}(1 + p_i t)$, then Flajolet-Gardy-Thimonier [FGT92]

$$E[B] = \int_0^{+\infty} \lambda(t) e^{-t} dt.$$

C1 Diversity: The probability $p^{\triangle}_{n,\pi}$ of the most probable word within $\mathcal{L}_n$ decreases exponentially with n.

C2 Log-positive weights: For each terminal symbol $t \in \Sigma$, $\pi_t > 1$.
$\Rightarrow$ No loss of generality (weighted distribution stable through rescaling)

C3 Bounded dependency: For any rational number $k > 1$ and any weight vector π such that Condition **C2** holds, ${\rho_\pi}^k < \rho_{\pi^k}$ holds.
(Corollary of **C1**)

Theorem (First collision)

Under conditions **C1**, **C2** *and* **C3**, *the expected number of generations* $E[B_{n,\pi}]$ *before some word of* $\mathcal{L}_n$ *is drawn twice is such that*

$$E[B_{n,\pi}] \sim \frac{\sqrt{\pi}}{\sqrt{2\alpha_{2,n}}} = \frac{\mu_{\pi,n}\sqrt{\pi}}{\sqrt{2\mu_{\pi^2,n}}} \in \Omega\left(\gamma^n\right), \quad \gamma := \frac{\rho_\pi}{\sqrt{\rho_{\pi^2}}} > 1$$

Proof: Let $\lambda(t) = \prod_{i=1}^{m}(1 + p_i t)$, then Flajolet-Gardy-Thimonier [FGT92]

$$E[B] = \int_0^{+\infty} \lambda(t) e^{-t}\, dt.$$

C1 Diversity: The probability $p^{\triangle}_{n,\pi}$ of the most probable word within $\mathcal{L}_n$ decreases exponentially with n.

C2 Log-positive weights: For each terminal symbol $t \in \Sigma$, $\pi_t > 1$.
$\Rightarrow$ No loss of generality (weighted distribution stable through rescaling)

C3 Bounded dependency: For any rational number $k > 1$ and any weight vector π such that Condition **C2** holds, ${\rho_\pi}^k < \rho_{\pi^k}$ holds.
(Corollary of **C1**)

Theorem (First collision)

Under conditions **C1**, **C2** *and* **C3**, *the expected number of generations* $E[B_{n,\pi}]$ *before some word of* $\mathcal{L}_n$ *is drawn twice is such that*

$$E[B_{n,\pi}] \sim \frac{\sqrt{\pi}}{\sqrt{2\alpha_{2,n}}} = \frac{\mu_{\pi,n}\sqrt{\pi}}{\sqrt{2\mu_{\pi^2,n}}} \in \Omega\left(\gamma^n\right), \quad \gamma := \frac{\rho_\pi}{\sqrt{\rho_{\pi^2}}} > 1$$

Proof: Let $\lambda(t) = \prod_{i=1}^{m}(1 + p_i t)$, then Flajolet-Gardy-Thimonier [FGT92]

$$E[B] = \int_0^{+\infty} \lambda(t) e^{-t}\, dt.$$

C1 Diversity: The probability $p^{\triangle}_{n,\pi}$ of the most probable word within $\mathcal{L}_n$ decreases exponentially with n.

C2 Log-positive weights: For each terminal symbol $t \in \Sigma$, $\pi_t > 1$.

C3 Bounded dependency: For any rational number $k > 1$ and any weight vector π such that Condition **C2** holds, ${\rho_\pi}^k < \rho_{\pi^k}$ holds.

Theorem (First collision)

Under conditions **C1**, **C2** *and* **C3**, *the expected number of generations $E[B_{n,\pi}]$ before some word of $\mathcal{L}_n$ is drawn twice is such that*

$$E[B_{n,\pi}] \sim \frac{\sqrt{\pi}}{\sqrt{2\alpha_{2,n}}} = \frac{\mu_{\pi,n}\sqrt{\pi}}{\sqrt{2\mu_{\pi^2,n}}} \in \Omega\left(\gamma^n\right), \quad \gamma := \frac{\rho_\pi}{\sqrt{\rho_{\pi^2}}} > 1$$

Proof: Let $\lambda(t) = \prod_{i=1}^{m}(1 + p_i t)$, then Flajolet-Gardy-Thimonier [FGT92]

$$E[B] = \int_0^{+\infty} \lambda(t) e^{-t}\, dt.$$

C1 Diversity: The probability $p^{\triangle}_{n,\pi}$ of the most probable word within $\mathcal{L}_n$ decreases exponentially with n.

C2 Log-positive weights: For each terminal symbol $t \in \Sigma$, $\pi_t > 1$.

C3 Bounded dependency: For any rational number $k > 1$ and any weight vector π such that Condition **C2** holds, $\rho_\pi{}^k < \rho_{\pi^k}$ holds.

Theorem (First collision)

Under conditions **C1**, **C2** *and* **C3**, *the expected number of generations $E[B_{n,\pi}]$ before some word of $\mathcal{L}_n$ is drawn twice is such that*

$$E[B_{n,\pi}] \sim \frac{\sqrt{\pi}}{\sqrt{2\alpha_{2,n}}} = \frac{\mu_{\pi,n}\sqrt{\pi}}{\sqrt{2\mu_{\pi^2,n}}} \in \Omega\left(\gamma^n\right), \quad \gamma := \frac{\rho_\pi}{\sqrt{\rho_{\pi^2}}} > 1$$

Proof: Let $\lambda(t) = \prod_{i=1}^{m}(1 + p_i t)$, then Flajolet-Gardy-Thimonier [FGT92]

$$E[B] = \int_0^{\tau} \lambda(t)e^{-t}dt + \int_{\tau}^{+\infty} \lambda(t)e^{-t}dt.$$

C1 Diversity: The probability $p^{\triangle}_{n,\pi}$ of the most probable word within $\mathcal{L}_n$ decreases exponentially with n.

C2 Log-positive weights: For each terminal symbol $t \in \Sigma$, $\pi_t > 1$.

C3 Bounded dependency: For any rational number $k > 1$ and any weight vector π such that Condition **C2** holds, ${\rho_\pi}^k < \rho_{\pi^k}$ holds.

Theorem (First collision)

Under conditions **C1**, **C2** *and* **C3**, *the expected number of generations* $E[B_{n,\pi}]$ *before some word of* $\mathcal{L}_n$ *is drawn twice is such that*

$$E[B_{n,\pi}] \sim \frac{\sqrt{\pi}}{\sqrt{2\alpha_{2,n}}} = \frac{\mu_{\pi,n}\sqrt{\pi}}{\sqrt{2\mu_{\pi^2,n}}} \in \Omega\left(\gamma^n\right), \quad \gamma := \frac{\rho_\pi}{\sqrt{\rho_{\pi^2}}} > 1$$

Proof: Let $\lambda(t) = \prod_{i=1}^m (1 + p_i t)$, then Flajolet-Gardy-Thimonier [FGT92]

$$E[B] = \int_0^\tau e^{-\alpha_2 t^2/2} dt \cdot \left(1 + \mathcal{O}\left(\alpha_3 \tau^3\right)\right) + \int_\tau^{+\infty} \lambda(t) e^{-t} dt.$$

- **C1** Diversity: The probability $p^{\triangle}_{n,\pi}$ of the most probable word within $\mathcal{L}_n$ decreases exponentially with n.
- **C2** Log-positive weights: For each terminal symbol $t \in \Sigma$, $\pi_t > 1$.
- **C3** Bounded dependency: For any rational number $k > 1$ and any weight vector π such that Condition **C2** holds, ${\rho_\pi}^k < \rho_{\pi^k}$ holds.

Theorem (First collision)

Under conditions **C1**, **C2** *and* **C3**, *the expected number of generations* $E[B_{n,\pi}]$ *before some word of* $\mathcal{L}_n$ *is drawn twice is such that*

$$E[B_{n,\pi}] \sim \frac{\sqrt{\pi}}{\sqrt{2\alpha_{2,n}}} = \frac{\mu_{\pi,n}\sqrt{\pi}}{\sqrt{2\mu_{\pi^2,n}}} \in \Omega\left(\gamma^n\right), \quad \gamma := \frac{\rho_\pi}{\sqrt{\rho_{\pi^2}}} > 1$$

Proof: Let $\lambda(t) = \prod_{i=1}^{m}(1 + p_i t)$, then Flajolet-Gardy-Thimonier [FGT92]

$$E[B] = \sqrt{\frac{\pi}{2\alpha_2}}\left(1 + \mathcal{O}\left(e^{-\tau^2\alpha_2/2}\right) + \mathcal{O}\left(\alpha_3\tau^3\right)\right) + \int_\tau^{+\infty} \lambda(t)e^{-t}dt.$$

C1 Diversity: The probability $p^{\triangle}_{n,\pi}$ of the most probable word within $\mathcal{L}_n$ decreases exponentially with n.

C2 Log-positive weights: For each terminal symbol $t \in \Sigma$, $\pi_t > 1$.

C3 Bounded dependency: For any rational number $k > 1$ and any weight vector π such that Condition **C2** holds, ${\rho_\pi}^k < \rho_{\pi^k}$ holds.

Theorem (First collision)

Under conditions **C1**, **C2** *and* **C3**, *the expected number of generations* $E[B_{n,\pi}]$ *before some word of* $\mathcal{L}_n$ *is drawn twice is such that*

$$E[B_{n,\pi}] \sim \frac{\sqrt{\pi}}{\sqrt{2\alpha_{2,n}}} = \frac{\mu_{\pi,n}\sqrt{\pi}}{\sqrt{2\mu_{\pi^2,n}}} \in \Omega(\gamma^n), \quad \gamma := \frac{\rho_\pi}{\sqrt{\rho_{\pi^2}}} > 1$$

Proof: Let $\lambda(t) = \prod_{i=1}^m (1 + p_i t)$, then Flajolet-Gardy-Thimonier [FGT92]

$$E[B] = \sqrt{\frac{\pi}{2\alpha_2}}\left(1 + \mathcal{O}\left(e^{-\tau^2\alpha_2/2}\right) + \mathcal{O}\left(\alpha_3\tau^3\right)\right) + \mathcal{O}\left(\frac{\lambda(\tau)\, e^{-\tau}}{\sum_i \frac{p_i^2\tau}{1+p_i\tau}}\right).$$

C1 Diversity: The probability $p^{\triangle}_{n,\pi}$ of the most probable word within $\mathcal{L}_n$ decreases exponentially with n.

C2 Log-positive weights: For each terminal symbol $t \in \Sigma$, $\pi_t > 1$.

C3 Bounded dependency: For any rational number $k > 1$ and any weight vector π such that Condition **C2** holds, ${\rho_\pi}^k < \rho_{\pi^k}$ holds.

Theorem (First collision)

Under conditions **C1**, **C2** *and* **C3**, *the expected number of generations $E[B_{n,\pi}]$ before some word of $\mathcal{L}_n$ is drawn twice is such that*

$$E[B_{n,\pi}] \sim \frac{\sqrt{\pi}}{\sqrt{2\alpha_{2,n}}} = \frac{\mu_{\pi,n}\sqrt{\pi}}{\sqrt{2\mu_{\pi^2,n}}} \in \Omega\left(\gamma^n\right), \quad \gamma := \frac{\rho_\pi}{\sqrt{\rho_{\pi^2}}} > 1$$

Proof: Let $\lambda(t) = \prod_{i=1}^{m}(1 + p_i t)$, then Flajolet-Gardy-Thimonier [FGT92]

$$E[B] = \sqrt{\frac{\pi}{2\alpha_2}}\left(1 + \mathcal{O}\left(e^{-\tau^2\alpha_2/2}\right) + \mathcal{O}\left(\alpha_3\tau^3\right)\right) + \mathcal{O}\left(\frac{1}{\sqrt{\alpha_2}}\right)$$

for some τ chosen such that $\alpha_2\tau^2 \to +\infty$ and $\alpha_3\tau^3 \to 0$ (e.g. $\tau := \alpha_{5/2}$).

Theorem (Full weighted collection)

Let $W^{\nabla}_{\pi,n}$ be the weight of the least probable word in $\mathcal{L}_n$ and $m_n = |\mathcal{L}_n|$. The waiting time $E[C_{n,\pi}]$ of the full collection is such that

$$\frac{\mu_{\pi,n}}{W^{\nabla}_{\pi,n}} \leq E[C_{n,\pi}] \leq 2 \cdot \mathcal{H}_{m_n} \cdot \frac{\mu_{\pi,n}}{W^{\nabla}_{\pi,n}}$$

which, for large values of n, adopts the equivalent

$$\frac{\kappa_\pi \cdot \rho_\pi^{-n}}{W^{\nabla}_{\pi,n} \cdot n^{k_\pi}} \leq E[C_{n,\pi}] \leq \frac{2 \cdot \log(1/\rho) \cdot \kappa_\pi \cdot \rho_\pi^{-n}}{W^{\nabla}_{\pi,n} \cdot n^{k_\pi - 1}}.$$

Proof: Berenbrink and Sauerwald established that waiting time obeys

$$\frac{\mathcal{U}_m}{3e \cdot \log\log m} \leq E[C_\pi] \leq 2\mathcal{U}_m \quad \text{with} \quad \mathcal{U}_{m_n} = \sum_{i=1}^{m_n} \frac{1}{ip_i} \leq \frac{\mu_{\pi,n}}{W^{\nabla}_{\pi,n}} \left(\sum_{i=1}^{m_n} \frac{1}{i} = \mathcal{H}_m \right)$$

where p_i is the probability of the i-th least probable word.
Lower bound is simply the expected time for drawing least probable word.

Random generation= Allocation of undistinguishable balls into distinct urns.
= Sequence (urns) of sets (content) of balls.

$$\Rightarrow \Psi_{\pi}(x,y) = \sum_{j\geq 0}\sum_{k\geq 0} a_{j,k}\cdot x^j \cdot \frac{y^k}{k!} = \prod_{i=1}^{m}(1 + x\,(e^{p_i y} - 1))$$

where $a_{j,k}$ is the probability of reaching j distinct urns upon throwing k balls.

Theorem (Distinct samples – Hwang and Janson [HJ08])

The expected number $E[N_{n,\pi,k}]$ of distinct words after k generations obeys

$$E[N_{n,\pi,k}] = \sum_{i=1}^{|\mathbf{W}|} m_{n,i}\cdot\left(1 - \left(1 - \frac{W_{n,i}}{\mu_{\pi,n}}\right)^k\right) = \sum_{i=1}^{|\mathbf{W}|} m_{n,i}\cdot\left(1 - e^{-\frac{W_{n,i}}{\mu_{\pi,n}}k}\right) + \mathcal{O}(1).$$

where $\mathbf{W}$ is the set of weight classes.

Remark: Since there are at most $\mathcal{O}(n^{|\Sigma|})$ classes of distinct weights, this gives a polynomial-time algorithm for computing $E[N_{n,\pi,k}]$.

Similar analysis is performed for coverage, introducing the weight contribution

$$\Phi_{\pi}(x,y) = \sum_{j\geq 0}\sum_{k\geq 0} b_{j,k}\cdot x^j \cdot \frac{y^k}{k!} = \prod_{i=1}^{m}\left(1 + x^{W_i}\left(e^{p_i y} - 1\right)\right)$$

where $b_{j,k}$ is now the probability of reaching distinct urns of total weight j upon throwing k balls.

Theorem (Coverage)

In a weighted distribution, the expected cumulated probability $E[P_{n,\pi,k}] \in [0,1]$ *of the set of distinct words obtained after* k *generations is given by*

$$E[P_{n,\pi,k}] = \sum_{i=1}^{|W|} m_{n,i}\cdot \frac{W_{n,i}}{\mu_{\pi,n}}\cdot\left(1 - \left(1 - \frac{W_{n,i}}{\mu_{\pi,n}}\right)^k\right).$$

Secondary structures avoiding *plateaux* of length $\leq \theta$ are generated by

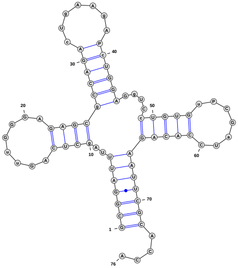
$$S \to (S_{\geq\theta})S \mid \bullet\, S \mid \varepsilon \qquad\qquad S_{\geq\theta} \to (S_{\geq\theta})S \mid \bullet\, S_{\geq\theta} \mid \bullet^{\theta}\,.$$

Weight function: $\pi(\boxed{\,)\,}) = \pi(\boxed{\bullet}) = 1$ and $\pi(\boxed{\,(\,}) = e^{\frac{-\Delta}{RT}}$ with $\Delta \in \{-1, -3\}$

Remark: Every base-pair can form in this homopolymer model.

Example: tRNA ($n = 80$)

Expectation	(θ, Δ)	#Samples
First collision	$(1, -1)$	$\sim 4.7\,.\,10^{13}$
	$(3, -3)$	~ 93.55
Full collection	$(1, -1)$	$\frac{0.64 \cdot 4.33^n}{n\sqrt{n}} \lesssim \cdot \lesssim \frac{1.24 \cdot 4.33^n}{\sqrt{n}}$
	$(3, -3)$	$\frac{0.065 \cdot 12.65^n}{n\sqrt{n}} \lesssim \cdot \lesssim \frac{0.11 \cdot 12.65^n}{\sqrt{n}}$



Secondary structures avoiding *plateaux* of length $\leq \theta$ are generated by

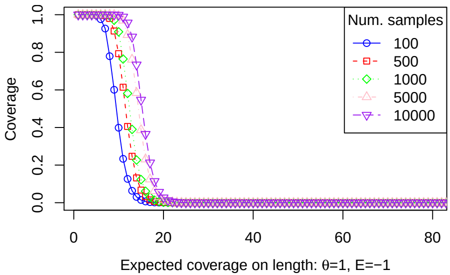
$$S \rightarrow (S_{\geq\theta})S \mid \bullet S \mid \varepsilon \qquad\qquad S_{\geq\theta} \rightarrow (S_{\geq\theta})S \mid \bullet S_{\geq\theta} \mid \bullet^{\theta}.$$

Weight function: $\pi(\boxed{)}) = \pi(\boxed{\bullet}) = 1$ and $\pi(\boxed{(}) = e^{\frac{-\Delta}{RT}}$ with $\Delta \in \{-1, -3\}$

Example:

Number $s_{n,k,i,\theta}$ of sec. str. of length n with i plateaux and $k \geq i$ bps obeys

$$s_{n,k,i,\theta} = \mathcal{N}(i,k)\binom{n-\theta k}{n-2i-\theta k} = \frac{1}{i}\binom{i}{k}\binom{i}{k-1}\binom{n-\theta k}{n-2i-\theta k}$$



Secondary structures avoiding *plateaux* of length $\leq \theta$ are generated by

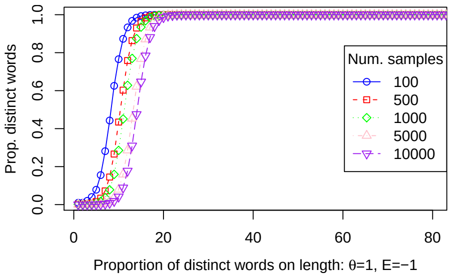
$$S \to (S_{\geq\theta})S \mid \bullet\, S \mid \varepsilon \qquad\qquad S_{\geq\theta} \to (S_{\geq\theta})S \mid \bullet\, S_{\geq\theta} \mid \bullet^{\theta}\,.$$

Weight function: $\pi(\boxed{)}) = \pi(\boxed{\bullet}) = 1$ and $\pi(\boxed{(}) = e^{\frac{-\Delta}{RT}}$ with $\Delta \in \{-1, -3\}$

Example:

Number $s_{n,k,i,\theta}$ of sec. str. of length n with i plateaux and $k \geq i$ bps obeys

$$s_{n,k,i,\theta} = \mathcal{N}(i,k)\binom{n-\theta k}{n-2i-\theta k} = \frac{1}{i}\binom{i}{k}\binom{i}{k-1}\binom{n-\theta k}{n-2i-\theta k}$$



Secondary structures avoiding *plateaux* of length $\leq \theta$ are generated by

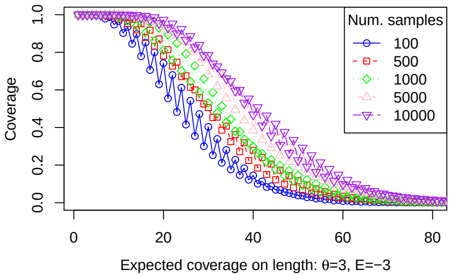
$$S \to (S_{\geq\theta})S \mid \bullet\, S \mid \varepsilon \qquad\qquad S_{\geq\theta} \to (S_{\geq\theta})S \mid \bullet\, S_{\geq\theta} \mid \bullet^{\theta}.$$

Weight function: $\pi(\boxed{)}) = \pi(\boxed{\bullet}) = 1$ and $\pi(\boxed{(}) = e^{\frac{-\Delta}{RT}}$ with $\Delta \in \{-1, -3\}$

Example:

Number $s_{n,k,i,\theta}$ of sec. str. of length n with i plateaux and $k \geq i$ bps obeys

$$s_{n,k,i,\theta} = \mathcal{N}(i,k)\binom{n-\theta k}{n-2i-\theta k} = \frac{1}{i}\binom{i}{k}\binom{i}{k-1}\binom{n-\theta k}{n-2i-\theta k}$$



Secondary structures avoiding *plateaux* of length $\leq \theta$ are generated by

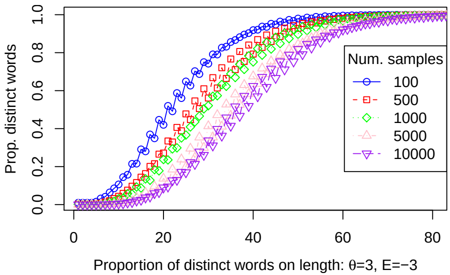
$$S \to (S_{\geq\theta})S \mid \bullet S \mid \varepsilon \qquad S_{\geq\theta} \to (S_{\geq\theta})S \mid \bullet S_{\geq\theta} \mid \bullet^{\theta}.$$

Weight function: $\pi(\boxed{)}) = \pi(\boxed{\bullet}) = 1$ and $\pi(\boxed{(}) = e^{\frac{-\Delta}{RT}}$ with $\Delta \in \{-1, -3\}$

Example:

Number $s_{n,k,i,\theta}$ of sec. str. of length n with i plateaux and $k \geq i$ bps obeys

$$s_{n,k,i,\theta} = \mathcal{N}(i,k)\binom{n-\theta k}{n-2i-\theta k} = \frac{1}{i}\binom{i}{k}\binom{i}{k-1}\binom{n-\theta k}{n-2i-\theta k}$$



Conclusion

We analyzed the level of redundancy within a sampled set of fixed size, and gave closed formulae/algorithms/asymptotic expansions for:

- Expected time of the first collision
- Expected time of the full collection
- Expected number of distinct samples after k generations
- Expected coverage of distinct samples

Yet certain questions remain open/partially addressed:

- Better characterization of suitable CFG languages.
  Which CFG satisfy the $p_1 \in o(\alpha^n), \alpha < 1$ property?
- Tighter bounds for the coupon collector ($\Theta(n)$ gap).
- Perform similar analysis for non-redundant generation.
- Waiting time for d distinct samples? For a desired coverage c?
  $\Rightarrow$ Determine for which k redundancy can be afforded (rejection).

Thank you!

Conclusion

We analyzed the level of redundancy within a sampled set of fixed size, and gave closed formulae/algorithms/asymptotic expansions for:

- Expected time of the first collision
- Expected time of the full collection
- Expected number of distinct samples after k generations
- Expected coverage of distinct samples

Yet certain questions remain open/partially addressed:

- Better characterization of suitable CFG languages.
  Which CFG satisfy the $p_1 \in o(\alpha^n), \alpha < 1$ property?
- Tighter bounds for the coupon collector ($\Theta(n)$ gap)
- Perform similar analysis for non-redundant generation.
- Waiting time for d distinct samples? For a desired coverage c?
  $\Rightarrow$ Determine for which k redundancy can be afforded (rejection).

Thank you!

Conclusion

We analyzed the level of redundancy within a sampled set of fixed size, and gave closed formulae/algorithms/asymptotic expansions for:

- Expected time of the first collision
- Expected time of the full collection
- Expected number of distinct samples after k generations
- Expected coverage of distinct samples

Yet certain questions remain open/partially addressed:

- Better characterization of suitable CFG languages. Which CFG satisfy the $p_1 \in o(\alpha^n), \alpha < 1$ property?
- Tighter bounds for the coupon collector ($\Theta(n)$ gap)
- Perform similar analysis for non-redundant generation.
- Waiting time for d distinct samples? For a desired coverage c? $\Rightarrow$ Determine for which k redundancy can be afforded (rejection).

Thank you!

Conclusion

We analyzed the level of redundancy within a sampled set of fixed size, and gave closed formulae/algorithms/asymptotic expansions for:

- Expected time of the first collision
- Expected time of the full collection
- Expected number of distinct samples after k generations
- Expected coverage of distinct samples

Yet certain questions remain open/partially addressed:

- Better characterization of suitable CFG languages.
  Which CFG satisfy the $p_1 \in o(\alpha^n), \alpha < 1$ property?
- Tighter bounds for the coupon collector ($\Theta(n)$ gap)
- Perform similar analysis for non-redundant generation.
- Waiting time for d distinct samples? For a desired coverage c?
  $\Rightarrow$ Determine for which k redundancy can be afforded (rejection).

Thank you!

A. Condon, B. Davy, B. Rastegari, S. Zhao, and F. Tarrant.
Classifying RNA pseudoknotted structures.
Theoretical Computer Science, 320(1):35–50, 2004.

A. Denise, O. Roques, and M. Termier.
Random generation of words of context-free languagesaccording to the frequencies of letters.
In D. Gardy and A. Mokkadem, editors, *Mathematics and Computer Science: Algorithms,Trees, Combinatorics and probabilities*, Trends in Mathematics, pages 113–125. Birkhaüser, 2000.

P. Flajolet, D. Gardy, and L. Thimonier.
Birthday paradox, coupon collectors, caching algorithms and self-organizing search.
Discrete Appl. Math., 39(3):207–229, 1992.

H.-K. Hwang and S. Janson.
Local limit theorems for finite and infinite urn models.
Ann. Probab., 36(3):992–1022, 2008.

R. B. Lyngsø and C. N. S. Pedersen.
RNA pseudoknot prediction in energy-based models.
Journal of Computational Biology, 7(3-4):409–427, 2000.

N. Leontis and E. Westhof.
Geometric nomenclature and classification of RNA base pairs.
RNA, 7:499–512, 2001.

Problem: Our bounds are not tight! $\Theta(n)$ factor between upper and lower bounds.

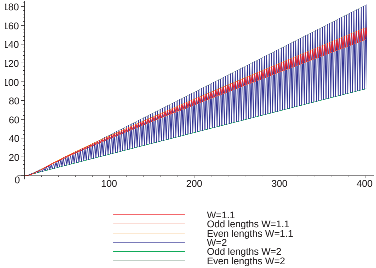


Figure: Plots of $\frac{W^{\nabla}_{\pi,\boldsymbol{n}}}{\mu_{\pi,\boldsymbol{n}}} \cdot \mathcal{U}_m$ for weighted Motzkin words exhibit a linear growth on n, suggesting that the upper bound is reached.