Weighted word collector

Jérémie du Boisberranger* Danièle Gardy* Yann Ponty•

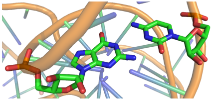
* PRiSM, CNRS/Université de Versailles, France

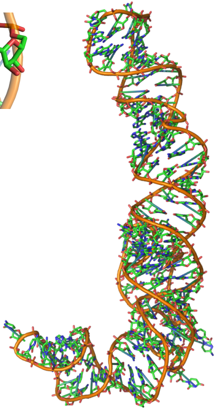
• LIX, CNRS/Ecole Polytechnique, France

Motivation: RNA folding prediction through statistical sampling

- RNA Structure

  RNA = Sequence over $\{A, C, G, U\}$.
  RNA folds, establishing hydrogen bonds.
  Such base-pairs stabilize structure.
  Free-energy E_S assigned to each structure S.

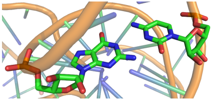


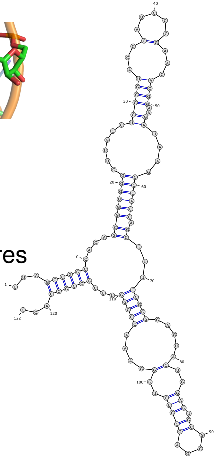


Motivation: RNA folding prediction through statistical sampling

- RNA Structure

  RNA = Sequence over $\{A, C, G, U\}$.
  RNA folds, establishing hydrogen bonds.
  Such base-pairs stabilize structure.
  Free-energy E_S assigned to each structure S.





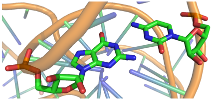
- Simplified energy model [Nussinov-Jacobson, 78]

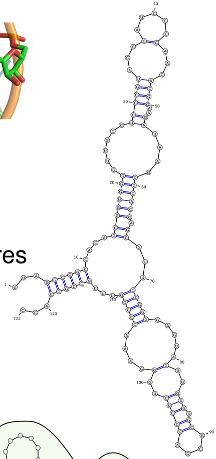
  Only non-crossing base-pairs allowed $\rightarrow$ Secondary Structures
  Free-Energy = $-\#$ Base-pairs

Motivation: RNA folding prediction through statistical sampling

- RNA Structure

  RNA = Sequence over $\{A, C, G, U\}$.
  RNA folds, establishing hydrogen bonds.
  Such base-pairs stabilize structure.
  Free-energy E_S assigned to each structure S.



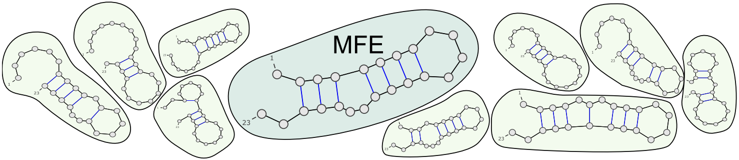


- Simplified energy model [Nussinov-Jacobson, 78]

  Only non-crossing base-pairs allowed $\rightarrow$ Secondary Structures
  Free-Energy = $-\#$ Base-pairs

- Boltzmann equilibrium [McCaskill, 90]

  Any structure S exists w.p. $p_S \propto e^{-\#E_S/RT}$.



Motivation: RNA folding prediction through statistical sampling

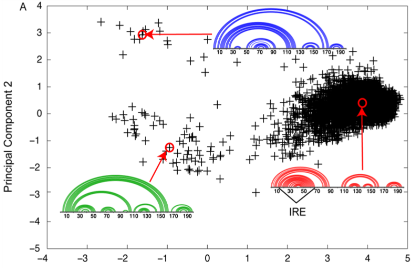
RNA *in silico* search: sequence $\rightarrow$ functional secondary structure

- MFE: Functional sec. str. = Most probable (aka min. free-energy) structure
  But approach lacks robustness to intrinsically uncertain energy models.

Motivation: RNA folding prediction through statistical sampling

RNA *in silico* search: sequence $\rightarrow$ functional secondary structure

- MFE: Functional sec. str. = Most probable (aka min. free-energy) structure
  But approach lacks robustness to intrinsically uncertain energy models.
- [Ding-Lawrence, 03] Functional sec. str. = Centroid structure

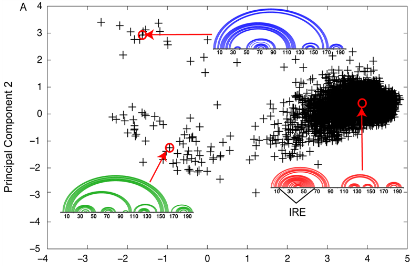


[Halvorsen *et al*, 10]

Motivation: RNA folding prediction through statistical sampling

RNA *in silico* search: sequence $\rightarrow$ functional secondary structure

- MFE: Functional sec. str. = Most probable (aka min. free-energy) structure
  But approach lacks robustness to intrinsically uncertain energy models.
- [Ding-Lawrence, 03] Functional sec. str. = Centroid structure



[Halvorsen *et al*, 10]

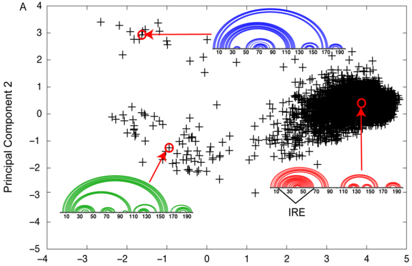
(Meta)-Algorithm:

1. Draw k sec. str. at random at Boltzmann equilibrium ($\neq$ Boltzmann sampling)
2. Cluster samples using your favorite machine learning method
3. Return centroid of most probable cluster

Motivation: RNA folding prediction through statistical sampling

RNA *in silico* search: sequence $\rightarrow$ functional secondary structure

- MFE: Functional sec. str. = Most probable (aka min. free-energy) structure
  But approach lacks robustness to intrinsically uncertain energy models.
- [Ding-Lawrence, 03] Functional sec. str. = Centroid structure



[Halvorsen *et al*, 10]

(Meta)-Algorithm:

1. Draw k sec. str. at random at Boltzmann equilibrium ($\neq$ Boltzmann sampling)
2. Cluster samples using your favorite machine learning method
3. Return centroid of most probable cluster

$\Rightarrow$ This simple idea greatly improved specificity of predictions.

Motivation: RNA folding prediction through statistical sampling

A closer look at the (meta)-algorithm:

1. Draw k sec. str. at random (with replacement!) in the Boltzmann distribution

Redundancy is uninformative, one should aim for k distinct secondary structures.

Which additional cost is induced by the redundant generation?

Motivation: RNA folding prediction through statistical sampling

A closer look at the (meta)-algorithm:

1. Draw k sec. str. at random (with replacement!) in the Boltzmann distribution

Redundancy is uninformative, one should aim for k distinct secondary structures.

Which additional cost is induced by the redundant generation?

$\downarrow$

Full-parameterized version difficult to study

Worst-case analysis: k =#Structures

Motivation: RNA folding prediction through statistical sampling

A closer look at the (meta)-algorithm:

1. Draw k sec. str. at random (with replacement!) in the Boltzmann distribution

Redundancy is uninformative, one should aim for k distinct secondary structures.

Which additional cost is induced by the redundant generation?

$\downarrow$

Full-parameterized version difficult to study

Worst-case analysis: k =#Structures

[Waterman, 78] Secondary structures are encoded by a simple context-free language.

Motivation: RNA folding prediction through statistical sampling

A closer look at the (meta)-algorithm:

1. Draw k sec. str. at random (with replacement!) in the Boltzmann distribution

Redundancy is uninformative, one should aim for k distinct secondary structures.

Which additional cost is induced by the redundant generation?

$\downarrow$

Full-parameterized version difficult to study

Worst-case analysis: k =#Structures

[Waterman, 78] Secondary structures are encoded by a simple context-free language.

$\downarrow$

How many generations before each word of length n in the language is obtained at least once?

Motivation: RNA folding prediction through statistical sampling

A closer look at the (meta)-algorithm:

1. Draw k sec. str. at random (with replacement!) in the Boltzmann distribution

Redundancy is uninformative, one should aim for k distinct secondary structures.

Which additional cost is induced by the redundant generation?

$\downarrow$

Full-parameterized version difficult to study

Worst-case analysis: k =#Structures

[Waterman, 78] Secondary structures are encoded by a simple context-free language.

$\downarrow$

How many generations before each word of length n in the language is obtained at least once?

$\downarrow$

Coupon collector problem!

The Coupon Collector problem (uniform)

The Coupon Collector problem (uniform)

The Coupon Collector problem (uniform)

The Coupon Collector problem (uniform)

How many stickers must she (I!) buy, on average, to get the full collection?

The Coupon Collector problem (uniform)

How many stickers must she (I!) buy, on average, to get the full collection?

$m \rightarrow$ Number of coupons
$C_m \rightarrow$ #purchased coupons (waiting time) for full collection (random variable).

Uniform distribution: Each coupon drawn with probability $1/m$.

The Coupon Collector problem (uniform)

How many stickers must she (I!) buy, on average, to get the full collection?

$m \rightarrow$ Number of coupons
$C_m \rightarrow$ #purchased coupons (waiting time) for full collection (random variable).

Uniform distribution: Each coupon drawn with probability $1/m$.

Average cost: - It takes a first coupon to get hooked!

$$E[C_m] = 1 \qquad .$$

The Coupon Collector problem (uniform)

How many stickers must she (I!) buy, on average, to get the full collection?

$m \rightarrow$ Number of coupons
$C_m \rightarrow$ #purchased coupons (waiting time) for full collection (random variable).

Uniform distribution: Each coupon drawn with probability $1/m$.

Average cost: - It takes a first coupon to get hooked!
- One must then buy, on the average, $\frac{m}{m-1}$ coupons to get the second coupon...

$$E[C_m] = 1 + \frac{m}{m-1} \qquad .$$

The Coupon Collector problem (uniform)

How many stickers must she (I!) buy, on average, to get the full collection?

$m \rightarrow$ Number of coupons
$C_m \rightarrow$ #purchased coupons (waiting time) for full collection (random variable).

Uniform distribution: Each coupon drawn with probability $1/m$.

Average cost: - It takes a first coupon to get hooked!
- One must then buy, on the average, $\frac{m}{m-2}$ coupons to get the third coupon...

$$E[C_m] = 1 + \frac{m}{m-1} + \frac{m}{m-2} \qquad .$$

The Coupon Collector problem (uniform)

10¢ per sticker
$m = 539$ stickers

$\Rightarrow$ 339 €!!!

How many stickers must she (I!) buy, on average, to get the full collection?

$m \rightarrow$ Number of coupons
$C_m \rightarrow$ #purchased coupons (waiting time) for full collection (random variable).

Uniform distribution: Each coupon drawn with probability $1/m$.

Average cost: - It takes a first coupon to get hooked!
- One must then buy, on the average, $\frac{m}{m-k}$ coupons to get the $k+1$-th coupon...

$$E[C_m] = 1 + \frac{m}{m-1} + \frac{m}{m-2} + \ldots + \frac{m}{m-k} + \ldots = m \cdot \mathcal{H}_m \underset{m\to\infty}{\sim} m \ln m.$$

Coupon Collector (non-uniform version)

Everybody knows that the dice are loaded, Everybody rolls with their fingers crossed...
Leonard Cohen, Everybody Knows

Coupon Collector (non-uniform version)

Everybody knows that the dice are loaded, Everybody rolls with their fingers crossed...
Leonard Cohen, Everybody Knows

Each sticker v_i is now drawn with probability p_i, $\sum_{i=1}^{m} p_i = 1$.

Theorem (Flajolet-Gardy-Thimonier,92)

$$E[C_m] = \int_0^\infty \left(1 - \prod_{i=1}^{m} \left(1 - e^{-p_i t}\right)\right) dt.$$

Coupon Collector (non-uniform version)

Everybody knows that the dice are loaded, Everybody rolls with their fingers crossed...
Leonard Cohen, Everybody Knows

Each sticker v_i is now drawn with probability p_i, $\sum_{i=1}^{m} p_i = 1$.

Theorem (Flajolet-Gardy-Thimonier,92)

$$E[C_m] = \int_0^{\infty} \left(1 - \prod_{i=1}^{m} \left(1 - e^{-p_i t}\right)\right) dt.$$

Difficult to evaluate exactly in general $\Rightarrow$ **Asymptotically?**

Coupon Collector (non-uniform version)

Everybody knows that the dice are loaded, Everybody rolls with their fingers crossed... *Leonard Cohen, Everybody Knows*

Each sticker v_i is now drawn with probability p_i, $\sum_{i=1}^{m} p_i = 1$.

Theorem (Flajolet-Gardy-Thimonier,92)

$$E[C_m] = \int_0^{\infty} \left(1 - \prod_{i=1}^{m} \left(1 - e^{-p_i t}\right)\right) dt.$$

Difficult to evaluate exactly in general $\Rightarrow$ **Asymptotically?**

Some results, derived on a case per case basis:

Coupon Collector (non-uniform version)

Everybody knows that the dice are loaded, Everybody rolls with their fingers crossed... *Leonard Cohen, Everybody Knows*

Each sticker v_i is now drawn with probability p_i, $\sum_{i=1}^{m} p_i = 1$.

Theorem (Flajolet-Gardy-Thimonier,92)

$$E[C_m] = \int_0^\infty \left(1 - \prod_{i=1}^{m} \left(1 - e^{-p_i t}\right)\right) dt.$$

Difficult to evaluate exactly in general $\Rightarrow$ **Asymptotically?**

Some results, derived on a case per case basis:

(i) [David-Barton,62] $\quad p_i = \dfrac{2i}{m(m+1)} \quad \Rightarrow \quad E[C_m] \underset{m\to\infty}{\sim} \left(\frac{2\pi}{\sqrt{3}} - 3\right) \cdot m \cdot (m+1).$

(ii) [Hildebrand,93] $\quad p_i = \dfrac{1}{iH_m} \quad \Rightarrow \quad E[C_m] \underset{m\to\infty}{\sim} m \cdot H_m \cdot \log m.$

A more general result

- Distribution defined by a sequence of positive numbers $\{a_1, ..., a_m\}$:

$$p_i = \frac{a_i}{\mu_m}, \qquad \text{with} \qquad \mu_m = \sum_{i=1}^{m} a_i.$$

A more general result

- Distribution defined by a sequence of positive numbers $\{a_1, ..., a_m\}$:

$$p_i = \frac{a_i}{\mu_m}, \qquad \text{with} \qquad \mu_m = \sum_{i=1}^{m} a_i.$$

Theorem (Boneh-Papanicolaou,94)

If $f(i) := 1/a_i$ satisfies:

$$\textit{(i)}\ f(x) \nearrow \infty, \qquad \textit{(ii)}\ \frac{f'(x)}{f(x)} \searrow, \qquad \textit{and} \qquad \textit{(iii)}\ \frac{f''(x)/f'(x)}{(f'(x)/f(x))\log(f'(x)/f)(x)} \to 0,$$

then

$$E[C_m] \underset{m\to\infty}{\sim} \mu_m \cdot f(m) \cdot \log \frac{f(m)}{f'(m)}$$

A more general result

- Distribution defined by a sequence of positive numbers $\{a_1, ..., a_m\}$:

$$p_i = \frac{a_i}{\mu_m}, \qquad \text{with} \qquad \mu_m = \sum_{i=1}^{m} a_i.$$

Theorem (Boneh-Papanicolaou,94)

If $f(i) := 1/a_i$ satisfies:

$$\text{(i) } f(x) \nearrow \infty, \qquad \text{(ii) } \frac{f'(x)}{f(x)} \searrow, \qquad \text{and} \qquad \text{(iii) } \frac{f''(x)/f'(x)}{(f'(x)/f(x))\log(f'(x)/f)(x)} \to 0,$$

then

$$E[C_m] \underset{m\to\infty}{\sim} \mu_m \cdot f(m) \cdot \log \frac{f(m)}{f'(m)}$$

What about sequences whose weights appear with multiplicities?
(e.g. unbounded #occurrences of some a_i as $m \to \infty$)

Random generation of words

Word collector?

Random generation of words

+

Coupon Collector

Random generation of words

+

Coupon Collector

=

Word collector

Which probability distribution on words?

Definition (Weighted language)

- **i** $\mathcal{L}$ *is a language over* $\Sigma = (a_1, ..., a_k)$*, and* $\mathcal{L}_n$ *its restriction to words of length* n*.*
- **ii** *Weight of a letter* $a_i \to \pi_{a_i} \in \mathbb{R}^+$.
- **iii** *Weight of a word* $\omega \in \mathcal{L}_n \to \pi(\omega) = \prod_{a \in \omega} \pi_a$
- **iv** *Weighted probability distribution, defined on* $\mathcal{L}_n$ *by:*

$$\forall \omega \in \mathcal{L}_n : \ \mathbb{P}[\omega] = \frac{\pi(\omega)}{\mu_n}, \qquad \text{with} \qquad \mu_n = \sum_{\omega \in \mathcal{L}_n} \pi(\omega).$$

Which probability distribution on words?

Definition (Weighted language)

- **i** *$\mathcal{L}$ is a language over $\Sigma = (a_1, ..., a_k)$, and $\mathcal{L}_n$ its restriction to words of length n.*
- **ii** *Weight of a letter $a_i \to \pi_{a_i} \in \mathbb{R}^+$.*
- **iii** *Weight of a word $\omega \in \mathcal{L}_n \to \pi(\omega) = \prod_{a \in \omega} \pi_a$*
- **iv** *Weighted probability distribution, defined on $\mathcal{L}_n$ by:*

$$\forall \omega \in \mathcal{L}_n : \ \mathbb{P}[\omega] = \frac{\pi(\omega)}{\mu_n}, \qquad \textit{with} \qquad \mu_n = \sum_{\omega \in \mathcal{L}_n} \pi(\omega).$$

Remark: Words having equal composition have equal probability:

$$\mathbb{P}[ababbabaaa] = \mathbb{P}[aababbbaaa] = \mathbb{P}[aaaaaabbbb] = \frac{\pi_a^6 \pi_b^4}{\mu_{10}}.$$

Which probability distribution on words?

Definition (Weighted language)

- **i** *$\mathcal{L}$ is a language over $\Sigma = (a_1, ..., a_k)$, and $\mathcal{L}_n$ its restriction to words of length n.*
- **ii** *Weight of a letter $a_i \to \pi_{a_i} \in \mathbb{R}^+$.*
- **iii** *Weight of a word $\omega \in \mathcal{L}_n \to \pi(\omega) = \prod_{a \in \omega} \pi_a$*
- **iv** *Weighted probability distribution, defined on $\mathcal{L}_n$ by:*

$$\forall \omega \in \mathcal{L}_n : \ \mathbb{P}[\omega] = \frac{\pi(\omega)}{\mu_n}, \qquad \textit{with} \qquad \mu_n = \sum_{\omega \in \mathcal{L}_n} \pi(\omega).$$

Remark: Words having equal composition have equal probability:

$$\mathbb{P}[ababbabaaa] = \mathbb{P}[aababbbaaa] = \mathbb{P}[aaaaaabbbb] = \frac{\pi_a^6 \pi_b^4}{\mu_{10}}.$$

Many coupons share equal weight, i.e. equal probability! $\Rightarrow$ Large multiplicities
$\Rightarrow$ None of existing results applies…

Asymptotics

Reminder: $$E[C_m] = \int_0^\infty \Phi(t)dt \qquad with \qquad \Phi(t) = 1 - \prod_{i=1}^{m}\left(1 - e^{-p_i t}\right).$$

Asymptotics of $E[C_m] \rightarrow$ **Evaluate integral?**

Asymptotics

Reminder: $$E[C_m] = \int_0^\infty \Phi(t)dt \qquad \text{with} \qquad \Phi(t) = 1 - \prod_{i=1}^{m}\left(1 - e^{-p_i t}\right).$$

Asymptotics of $E[C_m] \rightarrow$ **Evaluate integral?**

$$E[C_m] = \frac{\mu_m}{\omega(m)} \sum_{j=1}^{p} g_j(m) \int_0^\infty \Psi_m(t)\,dt \quad \text{with} \quad \Psi_m(t) := 1 - \prod_{i=1}^{|\mathbf{W}_m|}\left(1 - e^{-t\frac{W_{m,i}}{\omega(m)}\sum_{j=1}^{p} g_j(m)}\right)^{M_{m,i}}.$$

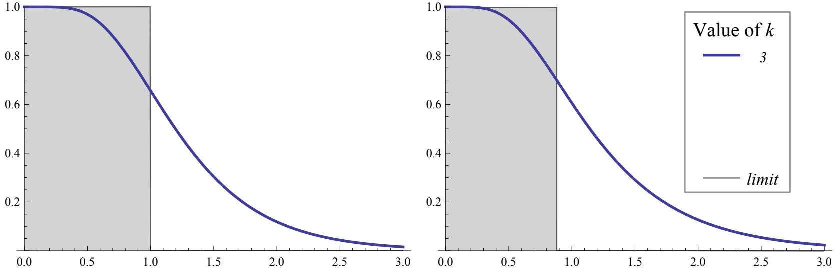
Asymptotics

Reminder: $E[C_m] = \int_0^\infty \Phi(t)dt$ *with* $\Phi(t) = 1 - \prod_{i=1}^{m}\left(1 - e^{-p_i t}\right)$.

Asymptotics of $E[C_m] \to$ **Evaluate integral?**

$$E[C_m] = \frac{\mu_m}{\omega(m)} \sum_{j=1}^{p} g_j(m) \int_0^\infty \Psi_m(t)\, dt \quad \text{with} \quad \Psi_m(t) := 1 - \prod_{i=1}^{|\mathbf{W}_m|}\left(1 - e^{-t\frac{W_{m,i}}{\omega(m)}\sum_{j=1}^{p} g_j(m)}\right)^{M_{m,i}}.$$

Example: Plot of $\Psi(t)$ for uniform and weighted distributions on $\{a, b\}^k$, $(m = 2^k)$.



Left: Uniform distribution. Right: Weighted distribution, $\pi_a/\pi_b = 2/3$.

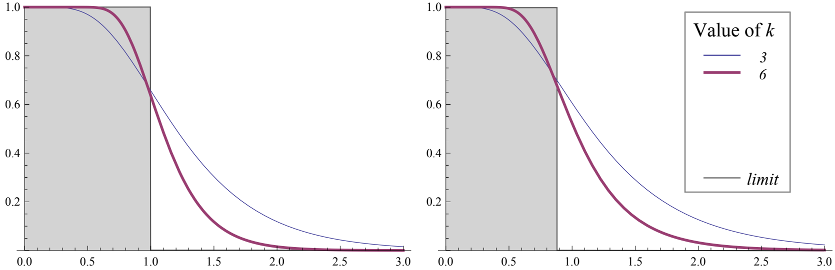
Asymptotics

Reminder: $$E[C_m] = \int_0^\infty \Phi(t)dt \quad \textit{with} \quad \Phi(t) = 1 - \prod_{i=1}^{m}\left(1 - e^{-p_i t}\right).$$

Asymptotics of $E[C_m] \rightarrow$ **Evaluate integral?**

$$E[C_m] = \frac{\mu_m}{\omega(m)} \sum_{j=1}^{p} g_j(m) \int_0^\infty \Psi_m(t)\, dt \quad \text{with} \quad \Psi_m(t) := 1 - \prod_{i=1}^{|\mathbf{w}_m|}\left(1 - e^{-t\frac{W_{m,i}}{\omega(m)}\sum_{j=1}^{p} g_j(m)}\right)^{M_{m,i}}.$$

Example: Plot of $\Psi(t)$ for uniform and weighted distributions on $\{a, b\}^k$, $(m = 2^k)$.



Left: Uniform distribution. Right: Weighted distribution, $\pi_a/\pi_b = 2/3$.

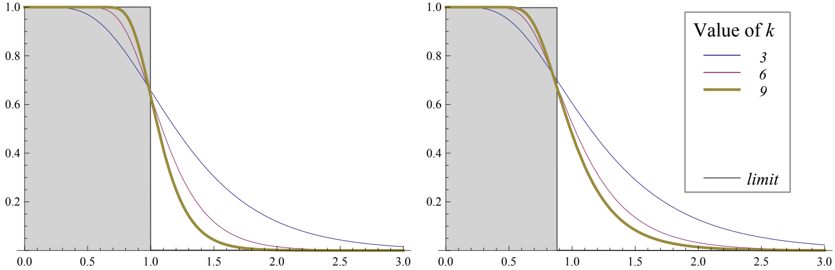
Asymptotics

Reminder: $$E[C_m] = \int_0^\infty \Phi(t)dt \quad \textit{with} \quad \Phi(t) = 1 - \prod_{i=1}^{m}\left(1 - e^{-p_i t}\right).$$

Asymptotics of $E[C_m] \to$ **Evaluate integral?**

$$E[C_m] = \frac{\mu_m}{\omega(m)} \sum_{j=1}^{p} g_j(m) \int_0^\infty \Psi_m(t)\, dt \quad \text{with} \quad \Psi_m(t) := 1 - \prod_{i=1}^{|\mathbf{W}_m|}\left(1 - e^{-t\frac{W_{m,i}}{\omega(m)}\sum_{j=1}^{p} g_j(m)}\right)^{M_{m,i}}.$$

Example: Plot of $\Psi(t)$ for uniform and weighted distributions on $\{a,b\}^k$, $(m = 2^k)$.



Left: Uniform distribution. Right: Weighted distribution, $\pi_a/\pi_b = 2/3$.

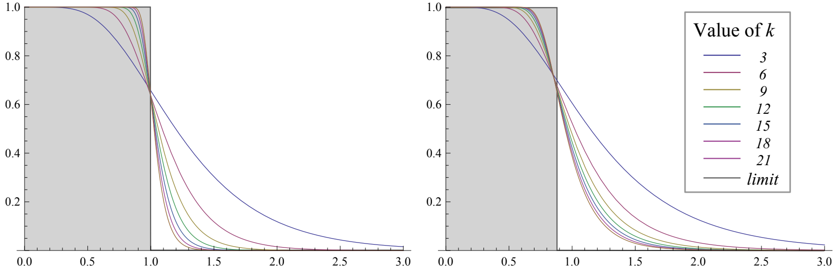
Asymptotics

Reminder: $E[C_m] = \int_0^\infty \Phi(t)dt$ *with* $\Phi(t) = 1 - \prod_{i=1}^{m}\left(1 - e^{-p_i t}\right).$

Asymptotics of $E[C_m] \to$ **Evaluate integral?**

$$E[C_m] = \frac{\mu_m}{\omega(m)} \sum_{j=1}^{p} g_j(m) \int_0^\infty \Psi_m(t)\, dt \quad \text{with} \quad \Psi_m(t) := 1 - \prod_{i=1}^{|\mathbf{W}_m|} \left(1 - e^{-t\frac{W_{m,i}}{\omega(m)}\sum_{j=1}^{p} g_j(m)}\right)^{M_{m,i}}.$$

Example: Plot of $\Psi(t)$ for uniform and weighted distributions on $\{a,b\}^k$, $(m = 2^k)$.



Left: Uniform distribution. Right: Weighted distribution, $\pi_a/\pi_b = 2/3$.

Asymptotics (continued)

Evaluating the integral? $\rightarrow$ **for distributions featuring large multiplicities**.

Asymptotics (continued)

Evaluating the integral? $\rightarrow$ **for distributions featuring large multiplicities**.

Definition (Distribution description)

- **i** *Smallest weight of a letter set to* 1.
- **ii** $\boldsymbol{W}_m$*: Vector of distinct weights, ordered increasingly.*
- **iii** $M_{m,i}$*: Multiplicity of weight* $W_{m,i}$.

Asymptotics (continued)

Evaluating the integral? $\rightarrow$ **for distributions featuring large multiplicities**.

Definition (Distribution description)

- **i** *Smallest weight of a letter set to* 1*.*
- **ii** $\mathbf{W}_m$*: Vector of distinct weights, ordered increasingly.*
- **iii** $M_{m,i}$*: Multiplicity of weight* $W_{m,i}$*.*

Hypotheses (simplified):

H1. *Multiplicity*. Control over (sufficient) growth of weight multiplicities $M_{m,i}$.
H2. *Regularity*. For large values of m, smallest weights of $\mathbf{W}_m$ no longer depend on m.
H3. *Growth*. Weight growth dominates multiplicity growth.

Asymptotics (continued)

Evaluating the integral? $\rightarrow$ **for distributions featuring large multiplicities**.

Definition (Distribution description)

- **i** *Smallest weight of a letter set to* 1*.*
- **ii** $\mathbf{W}_m$*: Vector of distinct weights, ordered increasingly.*
- **iii** $M_{m,i}$*: Multiplicity of weight* $W_{m,i}$*.*

Hypotheses (simplified):

H1. *Multiplicity*. Control over (sufficient) growth of weight multiplicities $M_{m,i}$.
H2. *Regularity*. For large values of m, smallest weights of $\mathbf{W}_m$ no longer depend on m.
H3. *Growth*. Weight growth dominates multiplicity growth.

Theorem (du Boisberranger-Gardy-P,2012)

If weight distribution satisfies hypotheses ***H1****,* ***H2*** *et* ***H3****, then*

$$E[C_m] \underset{m\to\infty}{\sim} \kappa \cdot \frac{\mu_m}{\omega(m)} \cdot g(m)$$

where κ*: explicit constant,*

Asymptotics (continued)

Evaluating the integral? $\rightarrow$ **for distributions featuring large multiplicities**.

Definition (Distribution description)

- **i** *Smallest weight of a letter set to* 1.
- **ii** $\mathbf{W}_m$*: Vector of distinct weights, ordered increasingly.*
- **iii** $M_{m,i}$*: Multiplicity of weight* $W_{m,i}$.

Hypotheses (simplified):

H1. *Multiplicity*. Control over (sufficient) growth of weight multiplicities $M_{m,i}$.
H2. *Regularity*. For large values of m, smallest weights of $\mathbf{W}_m$ no longer depend on m.
H3. *Growth*. Weight growth dominates multiplicity growth.

Theorem (du Boisberranger-Gardy-P,2012)

*If weight distribution satisfies hypotheses **H1**, **H2** et **H3**, then*

$$E[C_m] \underset{m\to\infty}{\sim} \kappa \cdot \frac{\mu_m}{\omega(m)} \cdot g(m)$$

where κ: explicit constant, $\omega(m)$: smallest weight in collection

Asymptotics (continued)

Evaluating the integral? $\rightarrow$ **for distributions featuring large multiplicities**.

Definition (Distribution description)

- **i** *Smallest weight of a letter set to* 1*.*
- **ii** $\mathbf{W}_m$*: Vector of distinct weights, ordered increasingly.*
- **iii** $M_{m,i}$*: Multiplicity of weight* $W_{m,i}$*.*

Hypotheses (simplified):

H1. *Multiplicity*. Control over (sufficient) growth of weight multiplicities $M_{m,i}$.
H2. *Regularity*. For large values of m, smallest weights of $\mathbf{W}_m$ no longer depend on m.
H3. *Growth*. Weight growth dominates multiplicity growth.

Theorem (du Boisberranger-Gardy-P,2012)

If weight distribution satisfies hypotheses ***H1**, **H2** et **H3**, then*

$$E[C_m] \underset{m\to\infty}{\sim} \kappa \cdot \frac{\mu_m}{\omega(m)} \cdot g(m)$$

where κ*: explicit constant,* $\omega(m)$*: smallest weight in collection and* $g(m)$*: log of rank-independent contribution in leading term of multiplicity as* $m \to \infty$*.*

Application: Σ^*

Description: Language $(a_1, ..., a_k)^*$ with $1 = \pi_{a_1} = ... = \pi_{a_j} < \pi_{a_{j+1}} \leq ... \leq \pi_{a_k}$.

Application: Σ^*

Description: Language $(a_1, ..., a_k)^*$ with $1 = \pi_{a_1} = ... = \pi_{a_j} < \pi_{a_{j+1}} \leq ... \leq \pi_{a_k}$.

. . . Verify Hypotheses H1, H2 and H3 (ouch!). . .

Application: Σ^*

Description: Language $(a_1, ..., a_k)^*$ with $1 = \pi_{a_1} = ... = \pi_{a_j} < \pi_{a_{j+1}} \leq ... \leq \pi_{a_k}$.

...Verify Hypotheses H1, H2 and H3 (ouch!)...

Reminder: $E[C_m] \sim \kappa \cdot \frac{\mu_m}{\omega(m)} \cdot g(m)$.

i $m \equiv$ #coupons $\equiv$ #words of length $n = k^n$.

ii $\omega(m) \equiv$ smallest weight $=$ weight of word $a_1^n = \pi_{a_1}^n \to 1$.

iii $\mu_m = (\pi_{a_1} + ... + \pi_{a_k})^n = m^{\log_k(\pi_{a_1}+\cdots+\pi_{a_k})}$.

iv $g(m) \sim \log\log m$ if $j = 1$, or $g(m) \sim \log m$ otherwise.

Application: Σ^*

Description: Language $(a_1, ..., a_k)^*$ with $1 = \pi_{a_1} = ... = \pi_{a_j} < \pi_{a_{j+1}} \leq ... \leq \pi_{a_k}$.

. . . Verify Hypotheses H1, H2 and H3 (ouch!). . .

Reminder: $E[C_m] \sim \kappa \cdot \frac{\mu_m}{\omega(m)} \cdot g(m)$.

- **i** $m \equiv \#\text{coupons} \equiv \#\text{words of length } n = k^n$.
- **ii** $\omega(m) \equiv$ smallest weight $=$ weight of word $a_1^n = \pi_{a_1}^n \to 1$.
- **iii** $\mu_m = (\pi_{a_1} + ... + \pi_{a_k})^n = m^{\log_k(\pi_{a_1}+\cdots+\pi_{a_k})}$.
- **iv** $g(m) \sim \log\log m$ if $j = 1$, or $g(m) \sim \log m$ otherwise.

Proposition (Waiting time – Full collection – Σ^*)

$$E[C_m] \sim \begin{cases} \kappa_1 \cdot m^p \cdot \log\log m & \text{if } j = 1, \\ \kappa_2 \cdot m^p \cdot \log m & \text{otherwise.} \end{cases}$$

where $p = \log_k(\pi_{a_1} + \cdots + \pi_{a_k}) > 1$.

Asymptotic waiting time differs from the uniform case.

Application: RNA Secondary Structures

Description: RNA sec. str. unambiguously generated by context-free grammar

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

where θ: minimal distance between matching parentheses ($\theta = 1$ or 3).

Boltzmann probability distribution $\Rightarrow \pi_{\bullet} = 1$ and $\pi_{(} \times \pi_{)} = e^{1/RT}$ ($\pi_{(} < \pi_{)}$).

Application: RNA Secondary Structures

Description: RNA sec. str. unambiguously generated by context-free grammar

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

where θ: minimal distance between matching parentheses ($\theta = 1$ or 3).

Boltzmann probability distribution $\Rightarrow \pi_{\bullet} = 1$ and $\pi_{(} \times \pi_{)} = e^{1/RT}$ ($\pi_{(} < \pi_{)}$).

...Verify Hypotheses H1, H2 and H3...

Application: RNA Secondary Structures

Description: RNA sec. str. unambiguously generated by context-free grammar

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

where θ: minimal distance between matching parentheses ($\theta = 1$ or 3).

Boltzmann probability distribution $\Rightarrow \pi_{\bullet} = 1$ and $\pi_{(} \times \pi_{)} = e^{1/RT}$ $(\pi_{(} < \pi_{)})$.

...Verify Hypotheses H1, H2 and H3...

Properties:

- **i** Gen. fun. + Singularity analysis $\Rightarrow \mu_m \sim \kappa \cdot m \cdot (\log m)^{-3/2}$
- **ii** Smallest weight = Weight of unpaired structure $\bullet^n = 1$
- **iii** Dominating term for multiplicity growth: $g(m) \sim \log\log m$.

Application: RNA Secondary Structures

Description: RNA sec. str. unambiguously generated by context-free grammar

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

where θ: minimal distance between matching parentheses ($\theta = 1$ or 3).

Boltzmann probability distribution $\Rightarrow \pi_\bullet = 1$ and $\pi_(\times \pi_) = e^{1/RT}$ $(\pi_(< \pi_))$.

...Verify Hypotheses H1, H2 and H3...

Properties:

- **i** Gen. fun. + Singularity analysis $\Rightarrow \mu_m \sim \kappa \cdot m \cdot (\log m)^{-3/2}$
- **ii** Smallest weight = Weight of unpaired structure $\bullet^n = 1$
- **iii** Dominating term for multiplicity growth: $g(m) \sim \log\log m$.

Proposition (Waiting time – Full collection – Boltzmann weighted Sec. Struct.)

$$E[C_m] \sim \kappa \cdot m^p \cdot (\log m)^{3p/2} \cdot \log\log m$$

where $p > 1$ depends on the dom. sing. of the cardinality and cumulated weight.

Again, asymptotic waiting time differs from the uniform case.

Corollary: On average, a sec. str. is generated $\Theta(m^{p-1}) = \mathcal{O}(\alpha^n)$ times, $\alpha > 1$.

Conclusion

- Words collector: new instance of the coupon collector.
- Original probability distribution: Large multiplicities.
- Asymptotic behaviors differ from uniform ones. . .
- . . . and arise as a trade-off between multiplicity and weight.
- Application: Still on a case-per-case basis. . .

Conclusion

- Words collector: new instance of the coupon collector.
- Original probability distribution: Large multiplicities.
- Asymptotic behaviors differ from uniform ones...
- ...and arise as a trade-off between multiplicity and weight.
- Application: Still on a case-per-case basis...

- Tighten and explain intuitively hypotheses.
- Variants: Partial collections? Waiting time before *k* distinct objects (alt. cumulated probability of $1 - \varepsilon$) are obtained.
- Language classes adopting ubiquitous asymptotic behaviors (regular/algebraic/strongly connected $\rightarrow$ Drmota/Lalley/Woods theorem...)
- Obtain second moment to get limiting distribution [Neal 08]

Conclusion

- Words collector: new instance of the coupon collector.
- Original probability distribution: Large multiplicities.
- Asymptotic behaviors differ from uniform ones. . .
- . . . and arise as a trade-off between multiplicity and weight.
- Application: Still on a case-per-case basis. . .

- Tighten and explain intuitively hypotheses.
- Variants: Partial collections? Waiting time before k distinct objects (alt. cumulated probability of $1-\varepsilon$) are obtained.
- Language classes adopting ubiquitous asymptotic behaviors (regular/algebraic/strongly connected $\to$ Drmota/Lalley/Woods theorem...)
- Obtain second moment to get limiting distribution [Neal 08]

Thanks for listening
Questions?

Acknowledgments: ANR **programs** MAGNUM **and** BOOLE