DNA evolution, Automata and Clumps

Pierre Nicodème

LIPN Team CALIN, University Paris 13, Villetaneuse

Problem setting

- Alphabet $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ (DNA)

$$\begin{array}{llll} \text{time} & = 0 & S_n(0) = & \mathtt{YYYYYYY......YYYYYYYY} \\ \vdots & & \vdots & \vdots \\ \text{time} & = T & S_n(T) = & \mathtt{YYYY...FF..FF..YYYYYY} \end{array}$$

- n = length of random sequences $S_n(0), \ldots, S_n(T)$ $(n \approx 2000)$

- $b = \mathtt{FF..FF} \in \mathcal{A}^k$ Transcription Factor $(5 \leq k = |b| \leq 10)$
  - b **does not occur** in $S_n(0)$
  - b **occurs for the first time by evolution** at time T in a sequence evolving from $S_n(0)$

Aim: Compute T

Initial $\nu(\alpha)$ and Substitution Probabilities $\pi_{\alpha\to\beta}$

α	$\nu(\alpha)$
A	0.23889
C	0.26242
G	0.25865
T	0.24004

$\rightsquigarrow$

substitution prob.
$\mathbb{P}(1) = \pi_{\alpha\to\beta}$
for **one** generation
(20 years)

A	$\rightsquigarrow$	A	0.9999999763
A	$\rightsquigarrow$	C	$4.54999994943 \times 10^{-9}$
A	$\rightsquigarrow$	G	$1.57499995613 \times 10^{-8}$
A	$\rightsquigarrow$	T	$3.4000001733 \times 10^{-9}$
C	$\rightsquigarrow$	A	$6.14999993408 \times 10^{-9}$
C	$\rightsquigarrow$	C	0.99999996495
C	$\rightsquigarrow$	G	$7.14999984731 \times 10^{-9}$
C	$\rightsquigarrow$	T	$2.17499993935 \times 10^{-8}$
G	$\rightsquigarrow$	A	$2.17499993935 \times 10^{-8}$
G	$\rightsquigarrow$	C	$7.14999984731 \times 10^{-9}$
G	$\rightsquigarrow$	G	0.99999996495
G	$\rightsquigarrow$	T	$6.14999993408 \times 10^{-9}$
T	$\rightsquigarrow$	A	$3.4000001733 \times 10^{-9}$
T	$\rightsquigarrow$	C	$1.57499995613 \times 10^{-8}$
T	$\rightsquigarrow$	G	$4.54999994943 \times 10^{-9}$
T	$\rightsquigarrow$	T	0.9999999763

Powers of $\mathbb{P}(1)$ remains close to the Identity Matrix

$$\mathbb{P}(1) \approx \begin{pmatrix} 1-3m & m & m & m \\ m & 1-3m & m & m \\ m & m & m & 1-3m \\ m & m & m & 1-3m \end{pmatrix} \qquad \text{with } m \approx 10^{-8}$$

$$\mathbb{P}^{N}(1) \approx \begin{pmatrix} 1-3mN & mN & mN & mN \\ mN & 1-3mN & mN & mN \\ mN & mN & mN & 1-3mN \\ mN & mN & mN & 1-3mN \end{pmatrix} + \mathcal{O}(m^2 N)$$

Therefore

$$P^{N}(1) \times \nu \approx \nu \qquad \text{for } N \approx 10^6 \text{ and } N < 10^6$$

$$P^{\infty}(1) \times \nu = (0.25, 0.25, 0.25, 0.25)^t$$

Geometric distribution of the Waiting Time

By **stationnarity** of ν, assuming $T \in \mathbb{N}$

$$\mathbf{P}\big(\textbf{no } b \text{ in } S_n(j+1) \mid \textbf{no } b \text{ in } S_n(j)\big)$$

$$= \mathbf{P}\big(\textbf{no } b \text{ in } S_n(1) \mid \textbf{no } b \text{ in } S_n(0)\big)$$

$$= 1 - \mathbf{P}\big(b \textbf{ occurs in } S_n(1) \mid \textbf{no } b \text{ in } S_n(0)\big)$$

Geometric distribution of the Waiting Time

By **stationnarity** of ν, assuming $T \in \mathbb{N}$

$$\mathbf{P}\big(\textbf{no } b \text{ in } S_n(j+1) \mid \textbf{no } b \text{ in } S_n(j)\big)$$
$$= \mathbf{P}\big(\textbf{no } b \text{ in } S_n(1) \mid \textbf{no } b \text{ in } S_n(0)\big)$$
$$= 1 - \mathbf{P}\big(b \textbf{ occurs in } S_n(1) \mid \textbf{no } b \text{ in } S_n(0)\big)$$

$$\mathbf{P}\big(b \textbf{ occurs in } S_n(T) \mid \textbf{no } b \text{ in } S_n(T-1)\big)$$
$$= \mathbf{P}\big(b \textbf{ occurs in } S_n(1) \mid \textbf{no } b \text{ in } S_n(0)\big)$$

Geometric distribution of the Waiting Time

By **stationnarity** of ν, assuming $T \in \mathbb{N}$

$$\mathbf{P}\big(\textbf{no } b \text{ in } S_n(j+1) \mid \textbf{no } b \text{ in } S_n(j)\big)$$
$$= \mathbf{P}\big(\textbf{no } b \text{ in } S_n(1) \mid \textbf{no } b \text{ in } S_n(0)\big)$$
$$= 1 - \mathbf{P}\big(b \textbf{ occurs in } S_n(1) \mid \textbf{no } b \text{ in } S_n(0)\big)$$

$$\mathbf{P}\big(b \textbf{ occurs in } S_n(T) \mid \textbf{no } b \text{ in } S_n(T-1)\big)$$
$$= \mathbf{P}\big(b \textbf{ occurs in } S_n(1) \mid \textbf{no } b \text{ in } S_n(0)\big)$$

Setting $\mathfrak{p}_n = \mathbf{P}\big(b \textbf{ occurs in } S_n(1) \mid \textbf{no } b \text{ in } S_n(0)\big)$,

$$\mathbf{E}(T) \approx i \sum_{i \geq 0} (1 - \mathfrak{p}_n)^i \times \mathfrak{p}_n = \frac{1}{\mathfrak{p}_n}$$

Renewing the Aim

We need now computing

$$\mathfrak{p}_n = \mathbf{P}\big(b \textbf{ occurs in } S_n(1) \mid \textbf{no } b \text{ in } S_n(0)\big)$$

$$= \frac{\mathbf{P}\big(b \textbf{ occurs in } S_n(1) \text{ AND } \textbf{ no } b \text{ in } S_n(0)\big)}{\mathbf{P}(\textbf{no } b \text{ in } S_n(0))}$$

Different computations of $\mathfrak{p}_n$

1. **Behrens-Vingron (2010)**
   - Approach **neglecting words correlation**.
   - **Efficient computation** of $\mathfrak{p}_n$ with respect to this assumption.
2. **Behrens-Nicaud-N (2012)**
   - **Rigorous and efficient approach by automata**.
3. **N (NCMA2012)**
   - Heuristic approach by **clump analysis**, either by **combinatorics of words** or by **automata** and generating functions.
4. **N (2013)**
   - Heuristic approach, adaptation of the **Régnier-Szpankowski equations** and **explicit formula** approximating $\mathfrak{p}_n$

Different computations of $\mathfrak{p}_n$

$$\begin{array}{llll} \text{time} & = 0 & S_n(0) = & \texttt{YYYYYYY......YYYYYYYY} \\ \vdots & & \vdots & \vdots \\ \text{time} & = 1 & S_n(1) = & \texttt{YYYY...FF..FF..YYYYYY} \end{array}$$

- Behrens-Vingron compute the probability that b **occurs** in $S_n(1)$ (**without allowing overlaps of occurrences**), and then the probability that $S_n(0)$ evolves to $S_n(1)$

Different computations of $\mathfrak{p}_n$

$$
\begin{array}{llll}
\text{time} & = 0 & S_n(0) = & \texttt{YYYYYYY......YYYYYYYY} \\
\vdots & & \vdots & \vdots \\
\text{time} & = 1 & S_n(1) = & \texttt{YYYY...FF..FF..YYYYYY}
\end{array}
$$

- Behrens-Vingron compute the probability that b **occurs** in $S_n(1)$ (**without allowing overlaps of occurrences**), and then the probability that $S_n(0)$ evolves to $S_n(1)$
- Behrens-Nicaud-N use an automaton on the alphabet $\mathcal{A} \times \mathcal{A}$ that scans **simultaneously** $S_n(0)$ **and** $S_n(1)$. This automaton is a kind of product of two **Knuth-Morris-Pratt automata**.

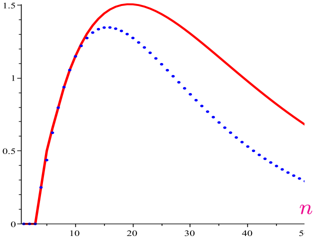
Different computations of $\mathfrak{p}_n$

$$\begin{array}{lll} \text{time } = 0 & S_n(0) = & \texttt{YYYYYYY......YYYYYYYY} \\ \vdots & \vdots & \vdots \\ \text{time } = 1 & S_n(1) = & \texttt{YYYY...FF..FF..YYYYYY} \end{array}$$

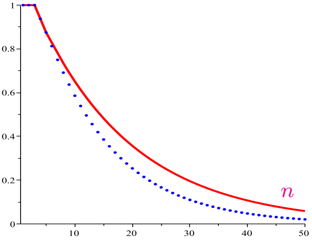
- Behrens-Vingron compute the probability that b **occurs** in $S_n(1)$ (**without allowing overlaps of occurrences**), and then the probability that $S_n(0)$ evolves to $S_n(1)$
- Behrens-Nicaud-N use an automaton on the alphabet $\mathcal{A} \times \mathcal{A}$ that scans **simultaneously** $S_n(0)$ **and** $S_n(1)$. This automaton is a kind of product of two **Knuth-Morris-Pratt automata**.
- N (2012) assumes that **a single mutation occurred** and considers the **clumps of neighbors** of b at **distance** 1 in $S_n(0)$.

An unexpected behaviour

$\frac{\mathfrak{p}_n}{\pi} \times \mathbf{P}(\text{no } b \text{ (or } b'\text{) in } S_n(0))$



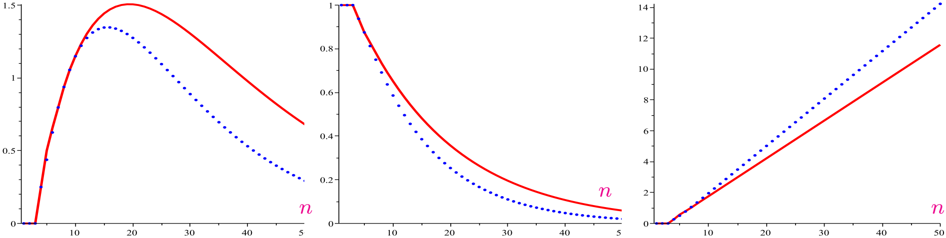
$\mathbf{P}(\text{no } b \text{ (or } b'\text{) in } S_n(0))$



$$b = \mathtt{ACAC} \quad b' = \mathtt{AACC}, \quad \nu(\mathtt{A}) = \nu(\mathtt{C}) = \frac{1}{2}, \quad \pi = \pi_{\mathtt{A}\to\mathtt{C}} = \pi_{\mathtt{C}\to\mathtt{A}}$$

An unexpected behaviour

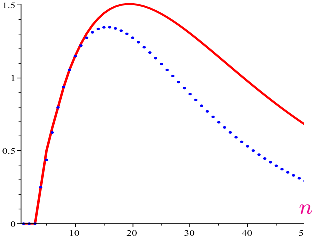
$\frac{\mathfrak{p}_n}{\pi} \times \mathbf{P}(\text{no } b \text{ (or } b'\text{) in } S_n(0))$ $\quad$ $\mathbf{P}(\text{no } b \text{ (or } b'\text{) in } S_n(0))$ $\quad$ $\frac{\mathfrak{p}_n}{\pi}$



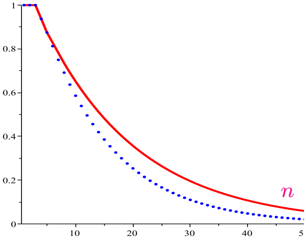
$$b = \texttt{ACAC} \quad b' = \texttt{AACC}, \quad \nu(\texttt{A}) = \nu(\texttt{C}) = \frac{1}{2}, \quad \pi = \pi_{\texttt{A}\to\texttt{C}} = \pi_{\texttt{C}\to\texttt{A}}$$

An unexpected behaviour

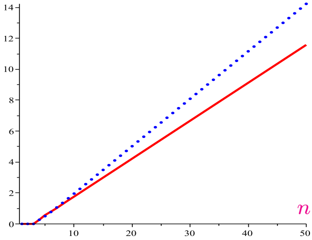
$\frac{\mathfrak{p}_n}{\pi} \times \mathbf{P}(\text{no } b \text{ (or } b'\text{) in } S_n(0))$



$\mathbf{P}(\text{no } b \text{ (or } b'\text{) in } S_n(0))$



$\frac{\mathfrak{p}_n}{\pi}$



$b = \texttt{ACAC} \quad b' = \texttt{AACC}, \quad \nu(\texttt{A}) = \nu(\texttt{C}) = \frac{1}{2}, \quad \pi = \pi_{\texttt{A}\to\texttt{C}} = \pi_{\texttt{C}\to\texttt{A}}$

$F(z,t)$ **rational function**

$$\mathbf{P}\big(b \in S_n(1) \mid b \notin S_n(0)\big) = \mathfrak{p}_n \times \mathbf{P}(\text{no } b \text{ in } S_n(0)) = [z^n] \left. \frac{\partial F(z,t)}{\partial t} \right|_{t=1}$$

$$\mathbf{P}(\text{no } b \text{ in } S_n(0)) = [z^n] F(z,1)$$

Formal Languages Approach - (N 2013)

(**Assuming a single mutation**)

$b =$ `AAAAA`

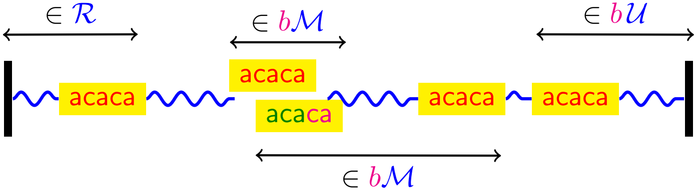
$$S(0) = \texttt{XXXX...XXXAAAAXAAAAXXXX.........XXX}$$
$$S(1) = \texttt{XXXX...XXXAAAAAAAAAXXXX.........XXX}$$
$$= \texttt{XX...} - \text{short clump of } \texttt{AAAAA} - \texttt{...XXX}$$

- **length of short clump** of b in $S(1)$ **must be less than** $2 \times |b| - 1$,
- else **there is at least one occurrence of** b **in** $S(0)$
- **no occurrences of** b in the `XXX...XXX`

if b **without self-overlap**, **short clump**$=b$

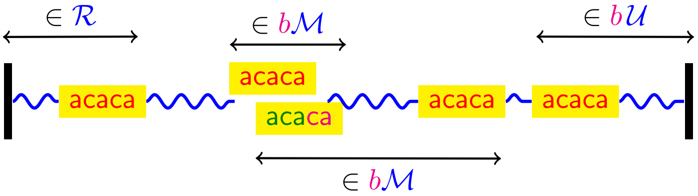
Guibas-Odlyzko decomposition - occurrences of a word b

$b = \texttt{acaca}$



Guibas-Odlyzko decomposition - occurrences of a word b

$b =$ acaca

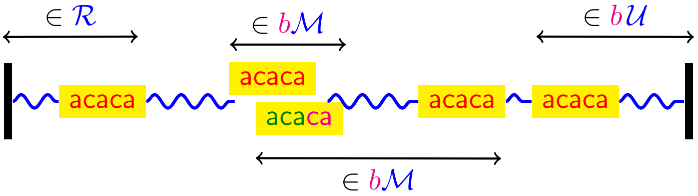


- Right $\mathcal{R}$: $= \{\ w = u.b$ et $\nexists r, s,\ w = r.b.s\ \}$

$$aaaaaacaca \subset \mathcal{R}, \quad cccccacacaca \not\subset \mathcal{R}$$

Guibas-Odlyzko decomposition - occurrences of a word b

$b =$ acaca



- **Right $\mathcal{R}$:** $= \{\ w = u.b$ et $\nexists r, s,\ w = r.b.s\ \}$

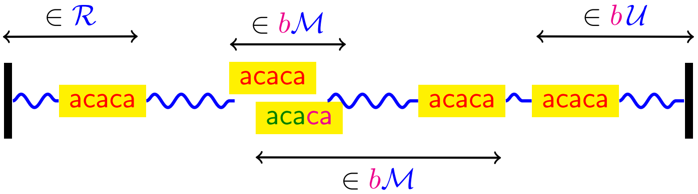
$$aaaaaacaca \subset \mathcal{R}, \quad cccccacacaca \not\subset \mathcal{R}$$

- **Minimal $\mathcal{M}$:** $= \{\ w,\quad b.w = u.b$ et $\nexists r, s,\ b.w = r.b.s\ \}$

$$acaca_{\,aaaaacaca} \subset \mathcal{M} \qquad ccaca_{\,cacccccccacaca} \not\subset \mathcal{M} \qquad ccaca_{\,ca} \subset \mathcal{M}$$

Guibas-Odlyzko decomposition - occurrences of a word b

$b =$ acaca



- Right $\mathcal{R}$: $= \{\ w = u.b$ et $\nexists r, s,\ w = r.b.s\ \}$

  $aaaaaacaca \subset \mathcal{R}, \quad ccccacacaca \not\subset \mathcal{R}$

- Minimal $\mathcal{M}$: $= \{\ w, \quad b.w = u.b$ et $\nexists r, s,\ b.w = r.b.s\ \}$

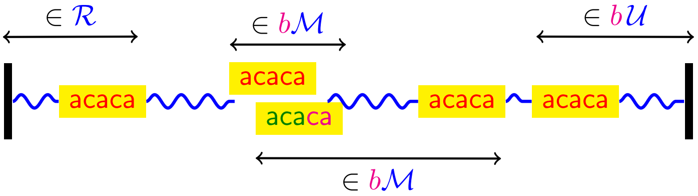
  $acaca\,{}_{aaaaacaca} \subset \mathcal{M} \quad ccaca\,{}_{cacccccccacaca} \not\subset \mathcal{M} \quad ccaca\,{}_{ca} \subset \mathcal{M}$

- Ultimate $\mathcal{U}$: $= \{w, \quad \nexists r, s,\ b.w = r.b.s\}$

  $acaca\,{}_{aacccaccccccc} \subset \mathcal{U} \qquad ccaca\,{}_{cacccccc} \not\subset \mathcal{U}$

Guibas-Odlyzko decomposition - occurrences of a word b

$b =$ acaca



- Right $\mathcal{R}$: $= \{\ w = u.b \quad \text{et} \quad \nexists r, s,\ w = r.b.s\ \}$

$$aaaaaacaca \subset \mathcal{R}, \qquad cccccacacaca \not\subset \mathcal{R}$$

- Minimal $\mathcal{M}$: $= \{\ w, \quad b.w = u.b \ \text{ et } \ \nexists r, s,\ b.w = r.b.s\ \}$

$${}^{acaca}aaaaacaca \subset \mathcal{M} \qquad {}^{ccaca}cacccccccacaca \not\subset \mathcal{M} \qquad {}^{ccaca}ca \subset \mathcal{M}$$

- Ultimate $\mathcal{U}$: $= \{w, \quad \nexists r, s,\ b.w = r.b.s\}$

$${}^{acaca}aacccacccccc \subset \mathcal{U} \qquad {}^{ccaca}caccccccc \not\subset \mathcal{U}$$

- Zero $\mathcal{Z} := \mathcal{A}^\star - \mathcal{A}^\star.b.\mathcal{A}^\star = \{w,\ \nexists r, s,\ w = r.b.s\}$

Régnier-Szpankowski Equations (see Lothaire)

- $\mathcal{A}^\star = \mathcal{U} + \mathcal{M}\mathcal{A}^\star$
- $\mathcal{A}^\star b = \mathcal{R}.\mathcal{C} + \mathcal{R}..\mathcal{A}^\star.b$
- $\mathcal{M}^+ = \mathcal{A}^\star.b + \mathcal{C} - \epsilon$
- $\mathcal{Z}.\sigma = \mathcal{R} + \mathcal{Z} - \epsilon$

Generating Functions of the Languages

$$R(z) = \frac{\mathbf{P}(b)z^{|b|}}{D(z)}, \quad M(z) = 1 - \frac{1-z}{D(z)},$$
$$U(z) = \frac{1}{D(z)}, \quad Z(z) = \frac{C(z)}{D(z)},$$

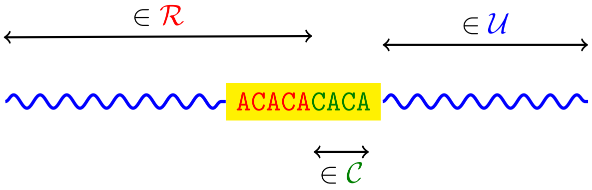
with $D(z) = (1-z)C(z) + \mathbf{P}(b)z^{|b|}$,

$\mathcal{C}$ **autocorrelation set** of the word b

$$\mathcal{C} = \{w; \quad b.w = u.b, \quad 0 \le |w| < |b|\} \qquad C(z) = \sum_{w \in \mathcal{C}} \mathbf{P}(w) z^{|w|}$$

What do we need in $S_n(1)$?

$b =$ ACACA



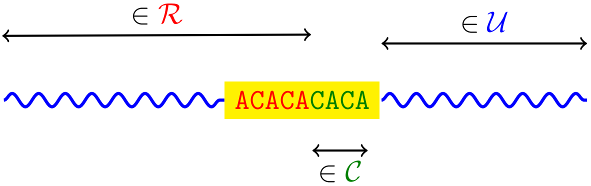
But **not any position** of the clump **can mutate**

ACACACACA	ACACACA	ACACA
NNNNYNNNN	NNYYYNN	YYYYY

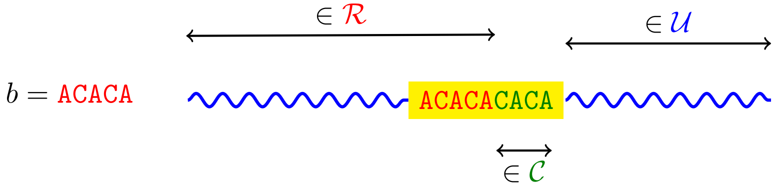
- **to avoid an occurrence** of b in $S(0)$
- if the short clump is $b.c$ with $c \in \mathcal{C}$
- **only** $t = |b| - |c|$ **positions can mutate**
- these positions are the t **last positions** of b

The right generating function

$b =$ ACACA



The right generating function

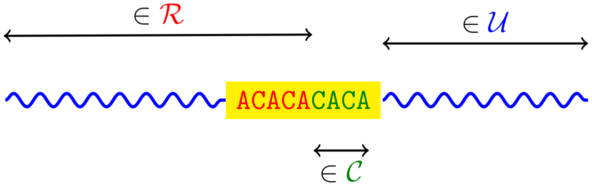


- **Gen.Fun.** $F(z)$ **of sequences with one short clump**

$$F(z) = R(z) \times \sum_{c \in \mathcal{C}} \mathbf{P}(c) z^{|c|} \times U(z) = \sum_{c \in \mathcal{C}} \frac{\mathbf{P}(b.c) z^{|b.c|}}{D^2(z)}$$

The right generating function

$b =$ ACACA



- **Gen.Fun.** $F(z)$ **of sequences with one short clump**

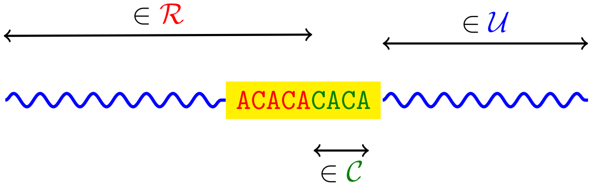
$$F(z) = R(z) \times \sum_{c \in \mathcal{C}} \mathbf{P}(c) z^{|c|} \times U(z) = \sum_{c \in \mathcal{C}} \frac{\mathbf{P}(b.c) z^{|b.c|}}{D^2(z)}$$

- **Gen.Fun** $Z(z)$ **of sequences with no occurrences of** b

$$Z(z) = \frac{C(z)}{D(z)}$$

The right generating function

$b =$ ACACA



- **Gen.Fun.** $F(z)$ **of sequences with one short clump**

$$F(z) = R(z) \times \sum_{c \in \mathcal{C}} \mathbf{P}(c) z^{|c|} \times U(z) = \sum_{c \in \mathcal{C}} \frac{\mathbf{P}(b.c) z^{|b.c|}}{D^2(z)}$$

- **Gen.Fun** $Z(z)$ **of sequences with no occurrences of** b

$$Z(z) = \frac{C(z)}{D(z)}$$

- $D(z) = (1 - z)C(z) + \mathbf{P}(b) z^{|b|}$

- $F(z)$ and $Z(z)$ have the **same dominant singularity** ω

Asymptotics of $\mathfrak{q}_n$ (**approximation** of $\mathfrak{p}_n$)

$$\mathfrak{q}_n = \frac{[z^n]F(z)}{[z^n]Z(z)} \qquad\qquad (\mathbf{P}(\epsilon) = 1)$$

ω dominant singularity of $D(z)$

$$\mathfrak{q}_n = \frac{\mathbf{P}(b)}{C(\omega)D'(\omega)}$$
$$\times \sum_{c \in \mathcal{C}} (|b| - |c|)\mathbf{P}(c)\omega^{|b.c|} \times \sum_{\substack{\beta \in \{b_{[|c|+1|]}, \ldots, b_{[|b|]}\} \\ \alpha \neq \beta}} \frac{\mathbf{P}(\alpha)}{\mathbf{P}(\beta)} \times \pi_{\alpha \to \beta}$$
$$\times \left((n - |b.c| + 1)\omega^{-1} + \frac{D''(\omega)}{D'(\omega)} \right) + o(\mathbf{P}(b)).$$

An even more approximated result

$$D(z) = (1 - z)C(z) + \mathbf{P}(b)z^{|b|}$$

by bootstrapping $\omega \approx 1 + \frac{\mathbf{P}(b)}{C(1) + |b|\mathbf{P}(b)} \approx 1$

Using $\omega \approx 1$ gives

$$\mathfrak{q}_n^{(\text{approx})} = \frac{\mathbf{P}(b)}{C^2(1)} \times \sum_{c \in \mathcal{C}} (|b| - |c|)\mathbf{P}(c)(n - |b.c| + 1) \times \sum_{\substack{\beta \in \left\{ b_{[|c|+1|]}, \dots, b_{[|b|]} \right\} \\ \alpha \neq \beta}} \frac{\mathbf{P}(\alpha)}{\mathbf{P}(\beta)} \times \pi_{\alpha \to \beta}$$

Theorem[N 2013]. The conditioned probability $\mathfrak{p}_n$ that a random sequence of length n that does not contain a k-mer b at time 0 evolves at time 1 to a random sequence that contains this k-mer verifies

$$\mathfrak{p}_n = \mathfrak{q}_n \times (1 + \mathcal{O}(n\psi)) + \mathcal{O}(n^2\psi^2)$$

where

$$\mathfrak{q}_n = \frac{\mathbf{P}(b)}{C(\omega)D'(\omega)}$$

$$\times \sum_{c\in\mathcal{C}}(|b| - |c|)\mathbf{P}(c)\omega^{|b.c|} \times \sum_{\substack{\beta\in\{b_{[|c|+1|]},\dots,b_{[|b|]}\} \\ \alpha\neq\beta}} \frac{\mathbf{P}(\alpha)}{\mathbf{P}(\beta)} \times \pi_{\alpha\to\beta}$$

$$\times \left((n - |b.c| + 1)\omega^{-1} + \frac{D''(\omega)}{D'(\omega)}\right) + o(\mathbf{P}(b)).$$

$$\psi = \frac{\max_{\alpha,\beta\in\mathcal{A};\alpha\neq\beta} p_{\alpha\to\beta}}{\min_{\alpha\in\mathcal{A}} p_{\alpha\to\alpha}}$$

Numerical validation

$\mathcal{A} = \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$ - uniform Bernoulli model for $S(0)$.

	$b = \texttt{AAAAA}$ and for $\alpha \neq \beta$, $p_{\alpha\to\beta} = 10^{-8}$			
Length n	$\mathfrak{p}_n \times 10^6$	$\mathfrak{h}_n \times 10^6$	$\mathfrak{q}_n \times 10^6$	$\mathfrak{q}_n^{(\text{approx})} \times 10^6$
10000	1.03335528	1.03335588	1.03335587	1.02703244
100000	10.3368481	10.3369021	10.3369021	10.2742439
10000000	1033.19278	1033.72699	1033.72698	1027.46750

- $\mathfrak{p}_n$ - Exact result by automata (Behrens-Nicaud-N 2012)
- Heuristic of a single mutation
  - $\mathfrak{h}_n$ clumps of neighbors at distance 1 of b in $S_n(0)$ (N 2012)
  - $\mathfrak{q}_n$, $\mathfrak{q}_n^{(\text{approx})}$ short clump approach on $S_n(1)$ (N 2103)