DNA evolution, Automata and Clumps

Pierre Nicodème

CNRS - LIPN Team CALIN, University Paris-North, Villetaneuse

Biological Motivation

- **Promoters** are DNA sequences located upstream of the gene they regulate; regulation can be positive for enhancers or negative for repressors.
- The promoters contain binding sites for regulatory proteins such as **Transcription Factors** (**TFs**) that are **short stretches of DNA**.

Biological Motivation

- **Promoters** are DNA sequences located upstream of the gene they regulate; regulation can be positive for enhancers or negative for repressors.
- The promoters contain binding sites for regulatory proteins such as **Transcription Factors** (**TFs**) that are **short stretches of DNA**.

- **Waiting time:** **how long** it takes for a **Transcription Factor** to **appear** in a **promoter** under a **probabilistic model of evolution** helps understanding the **overall evolution of promoters** within species and between species?

From infinitesimal to discrete evolution model

- $\mathbb{Q}(t)dt$ evolution matrix for **infinitesimal time**
- $\mathbb{P}(t)$ evolution matrix **from time** x and **time** $x+t$

$$\mathbb{P}(t) = e^{\mathbb{Q}(t)} \qquad \text{(Karlin-Taylor 1975)}$$

- $\mathbb{P}(1) = (\pi_{\alpha\to\beta})$ evolution matrix for **one generation** (**20 years**), $\quad \alpha, \beta \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$

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 probability $\pi_{\alpha\to\beta}$ 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	$3.40000001733 \times 10^{-9}$
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	$3.40000001733 \times 10^{-9}$
T	$\rightsquigarrow$	C	$1.57499995613 \times 10^{-8}$
T	$\rightsquigarrow$	G	$4.54999994943 \times 10^{-9}$
T	$\rightsquigarrow$	T	0.9999999763

Probability of occurrence of a k-mer at time 1

- $S_n(0)$ **random DNA sequence** of **length** n at **time 0**
- $S_n(1)$ sequence obtained **from** $S_n(0)$ **by evolution** at **time 1**
- b a k-**mer** (word of **length** k over $\mathcal{A} = \{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\}$)

- $\mathfrak{P}_n(b)$ probability that b
  - **occurs at time 1**
  - **while not occurring at time 0**

$$\mathfrak{P}_n(b) = \mathbb{P}(b \in S_n(1) \mid b \notin S_n(0))$$

Probability of occurrence of a k-mer at time 1

- $S_n(0)$ **random DNA sequence** of **length** n at **time 0**
- $S_n(1)$ sequence obtained **from** $S_n(0)$ **by evolution** at **time 1**
- b a k-**mer** (word of **length** k over $\mathcal{A} = \{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\}$)

- $\mathfrak{P}_n(b)$ probability that b
  - **occurs at time 1**
  - **while not occurring at time 0**

$$\mathfrak{P}_n(b) = \mathbb{P}(b \in S_n(1) \mid b \notin S_n(0))$$

Expectation of the Waiting time $\mathfrak{E}_n(b)$

-

$$\mathfrak{E}_n(b) \approx \frac{1}{\mathfrak{P}_n(b)} \quad (\textbf{geometric distribution} - \text{BehVin2010})$$

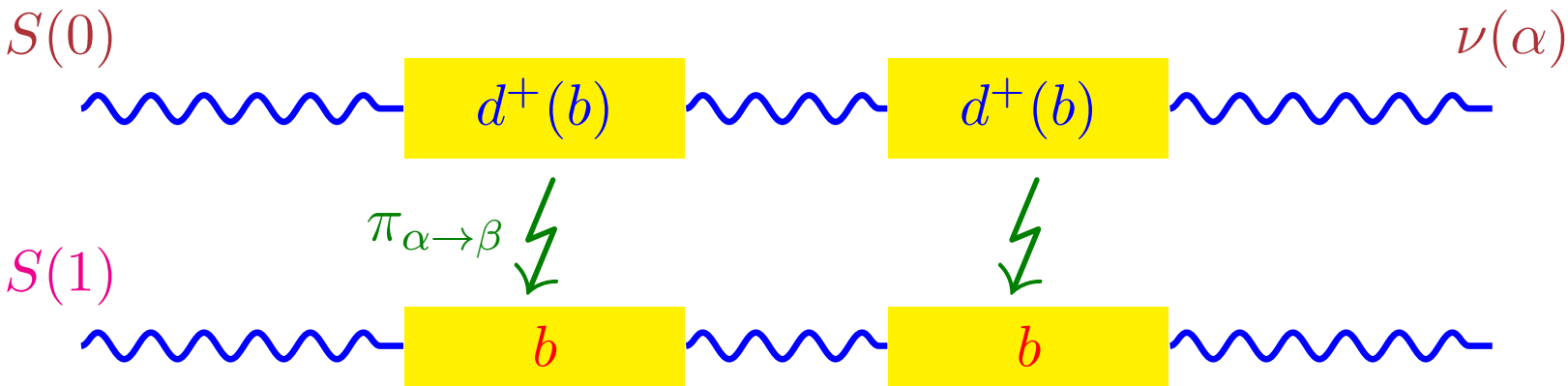
Plan of the talk

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-P.N. (2012)**
   - **Rigorous and efficient approach by automata**.
   - Approach **hiding the quasi-linear behaviour** of $\mathfrak{P}_n$
3. **P.N. (NCMA2012)**
   - **Non-efficient** approach by **clump analysis**, either by **combinatorics of words** or by **automata**.
   - **Proof by singularity analysis** of the **quasi-linear behaviour** of $\mathfrak{P}_n$

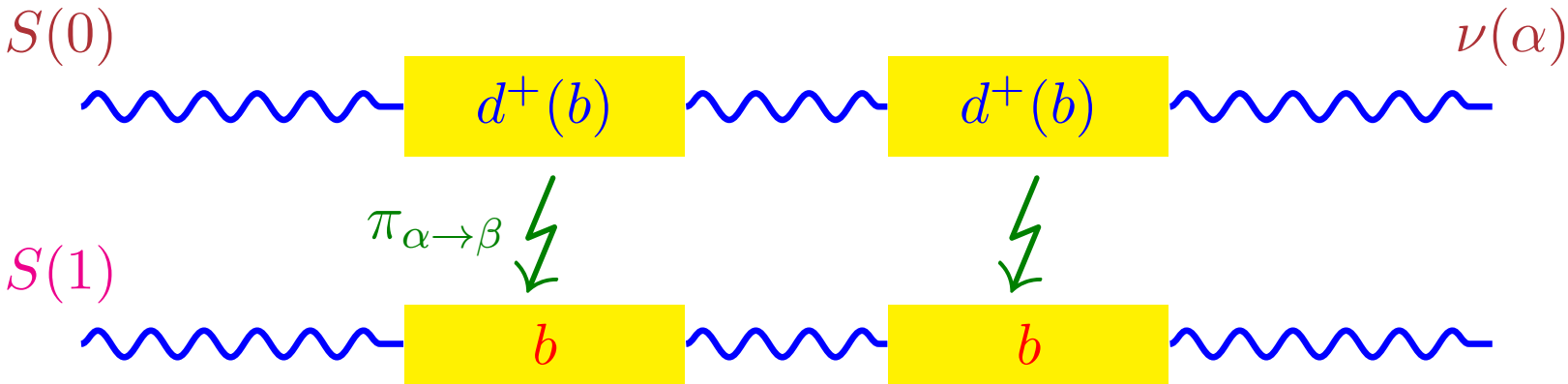
Behrens-Vingron 2010

- $d^+(b)$ neighbors of b by substitution



Behrens-Vingron 2010

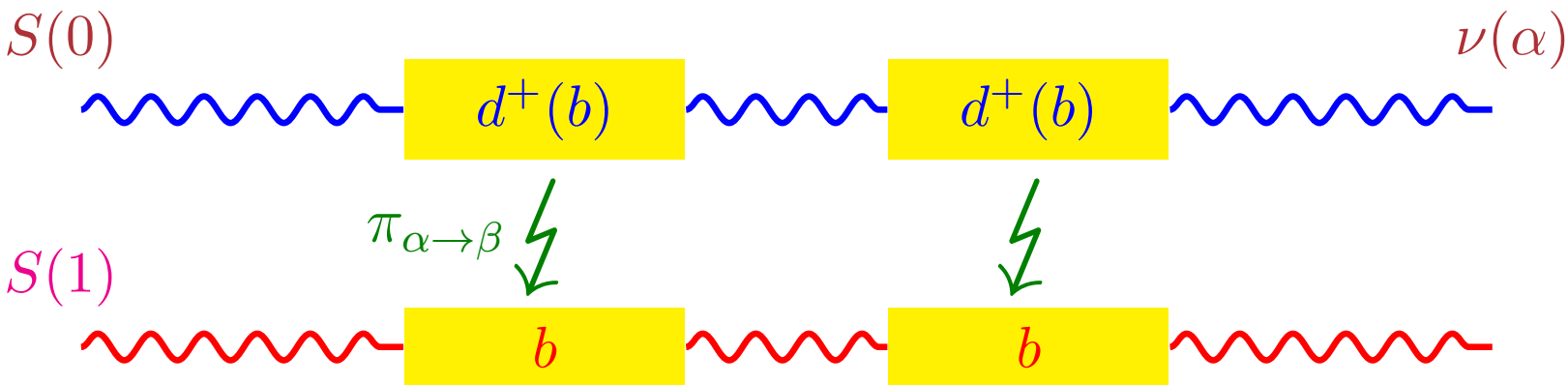
- $d^+(b)$ neighbors of b by substitution



$$\begin{cases} \mathfrak{P}_n \approx \sum_{i=1}^{\lfloor n/k \rfloor} (-1)^{i+1} \binom{n-i(k-1)}{i} \Phi^i \\ \Phi = \sum_{(a_1,\ldots,a_k)\in\mathcal{A}^k\setminus\{b_1,\ldots,b_k\}} \nu(a_1)\times\cdots\times\nu(a_k)\cdot\prod_{j=1}^{k}\pi_{a_i\to b_i}(1) \end{cases}$$

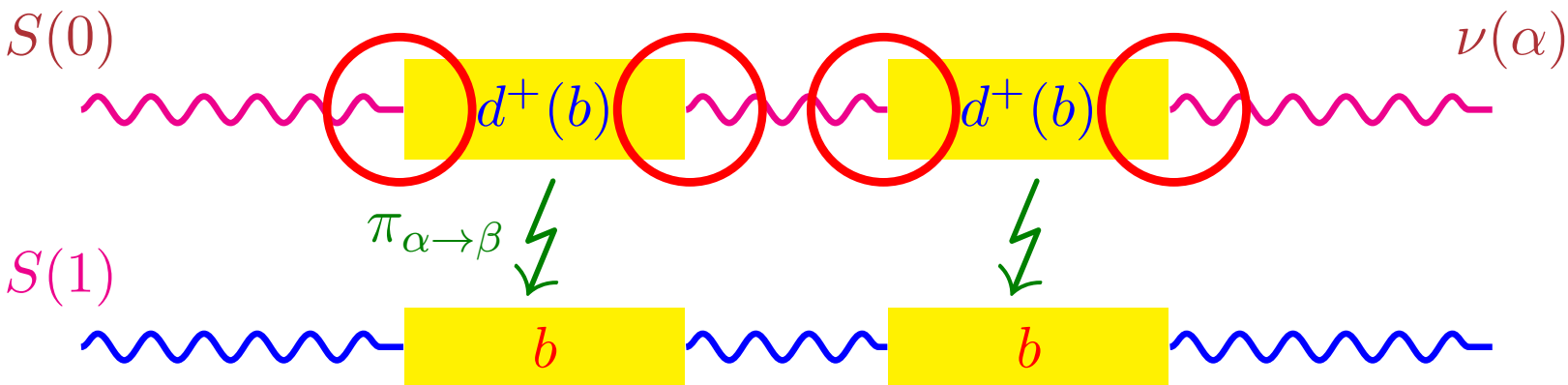
Approximations of Behrens-Vingron 2010

- occurrences of b in $S(1)$ **do not overlap**



Approximations of Behrens-Vingron 2010

- occurrences of b in $S(1)$ **do not overlap**
- possible **unwanted occurrences** of b at **junctions** in $S(0)$

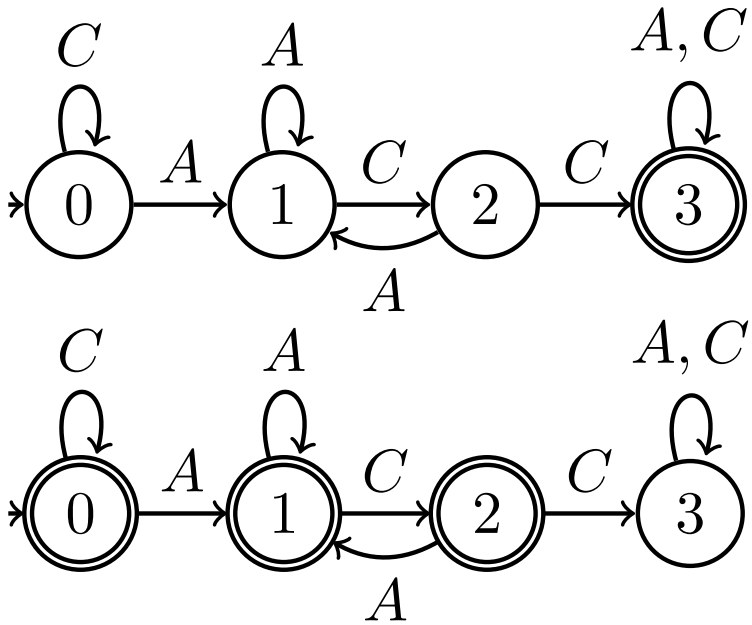


Behrens-Nicaud-P.N. 2012

Construct an **automaton**

- on the **alphabet** $\Sigma = \mathcal{A} \times \mathcal{A}$ with $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$
- **recognizing sequences** $\mathsf{S}(b) = S(0) \otimes S(1)$
- **such that**
  1. $b \notin S(0)$
  2. $b \in S(1)$

Using the Knuth-Morris-Pratt automaton



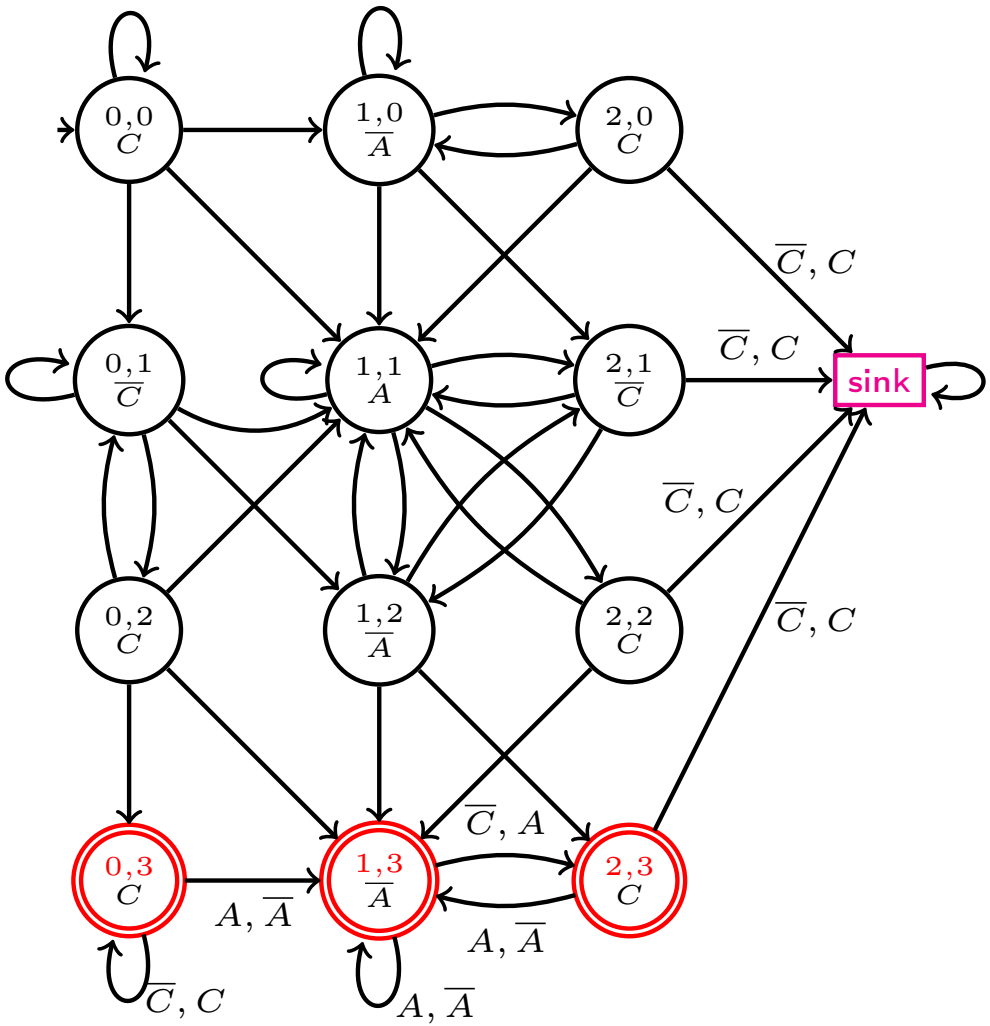
$$\mathcal{M}_{\mathsf{ACC}} = \{Q, \delta, s = 0, F\}$$

$$\overline{\mathcal{M}}_{\mathsf{ACC}} = \{Q, \delta, s = 0, Q \setminus F\}$$

$$\begin{cases} \mathcal{M}_b = (Q = \{0, \ldots, k\}, \delta_b, 0, \{k\}) \\ \overline{\mathcal{M}}_b = (Q = \{0, \ldots, k\}, \delta_b, 0, \{0, \ldots, k-1\}) \\ \mathcal{N}_b = \overline{\mathcal{M}}_b \otimes \mathcal{M}_b = (Q \times Q, \Delta, q'_0 = (0,0), F' = \{0,\ldots,k-1\} \times \{k\}) \end{cases}$$

$$\Delta((r,s),(\alpha,\beta)) = (\delta_b(r,\alpha), \delta_b(s,\beta))$$

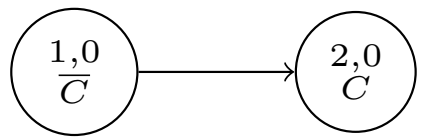
The automaton $\mathcal{N}_{\mathsf{ACC}} = \overline{\mathcal{M}}_{\mathsf{ACC}} \otimes \mathcal{M}_{\mathsf{ACC}}$ with matrix $\mathbb{P}$



Notations for the transitions:

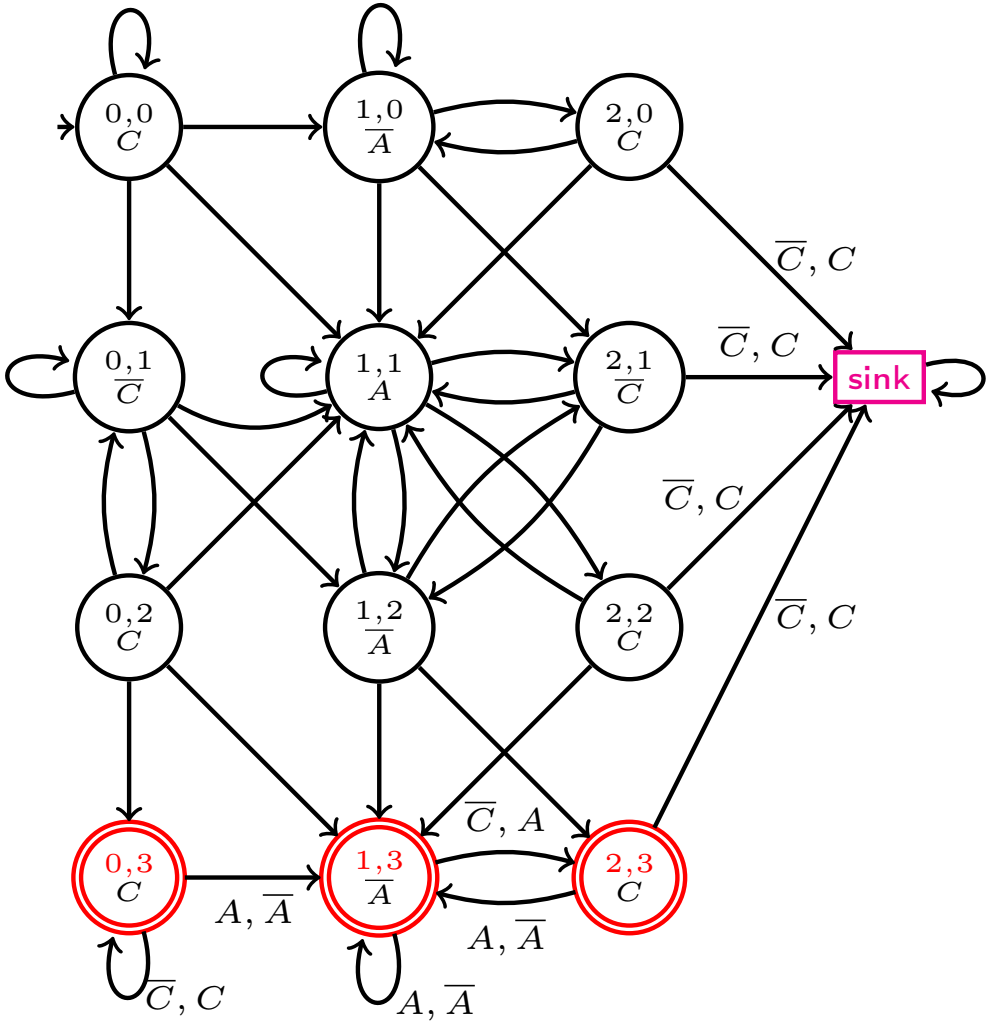
$$\begin{cases} A = \binom{A}{A}, & C = \binom{C}{C} \\ \overline{A} = \binom{A}{C}, & \overline{C} = \binom{C}{A} \end{cases}$$

a missing label of a transition is set to the letter at the bottom of its ending state



is labelled by C

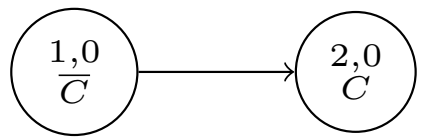
The automaton $\mathcal{N}_{\text{ACC}} = \overline{\mathcal{M}}_{\text{ACC}} \otimes \mathcal{M}_{\text{ACC}}$ with matrix $\mathbb{P}$



Notations for the transitions:

$$\begin{cases} A = \binom{A}{A}, & C = \binom{C}{C} \\ \overline{A} = \binom{A}{C}, & \overline{C} = \binom{C}{A} \end{cases}$$

a missing label of a transition is set to the letter at the bottom of its ending state



is labelled by C

$$\mathfrak{P}_n = \mathbf{P}(S_n(1) \in \mathcal{A}^\star b \mathcal{A}^\star | S_n(0) \notin \mathcal{A}^\star b \mathcal{A}^\star) = \frac{V_{q_0'} \mathbb{P}^n {V_{F'}}^t}{1 - V_{q_0'} \mathbb{P}^n V_{\text{sink}}^t}$$

Results for 5-mers of DNA

	BNN		BV		
	$\mathbf{E}_{\mathrm{BNN}}(T_{1000})/10^6$	Rank	$\mathbf{E}_{\mathrm{BV}}(T_{1000})/10^6$	Rank	$\frac{\mathbf{E}_{\mathrm{BNN}}(T_{1000})}{\mathbf{E}_{\mathrm{BV}}(T_{1000})}$
CCCCC	9,105	1021	6,304	1	1.44
GGGGG	9,570	1022	6,666	142	1.44
TTTTT	10,401	1023	7,457	993	1.39
AAAAA	10,656	1024	7,654	1024	1.39
CGCGC	7,047	699	6,446	11	1.09
TCCCC	7,076	737	6,477	17	1.09
CCCCT	7,076	738	6,477	21	1.09
GCGCG	7,127	787	6,518	31	1.09
CTCTC	7,263	883	6,679	148	1.09
...	...	...	...	...	...

$$\left\{\begin{array}{l} 4\% \text{ of the } 5\text{-mers} \\ 0.2\% \text{ of the } 7\text{-mers} \\ 0.002\% \text{ of the } 10\text{-mers} \end{array}\right. \Bigg| \text{ verify } \frac{\mathbf{E}_{\mathrm{BNN}}(T_{1000})}{\mathbf{E}_{\mathrm{BV}}(T_{1000})} > 1.05\%$$

Numerical remarks

- **length** of promoters $n \in [500 - 2000]$
- **Mutation probability** $\pi = \max(\pi_{\alpha \to \beta}) \approx 10^{-9}$

Numerical remarks

- **length** of promoters $n \in [500 - 2000]$
- **Mutation probability** $\pi = \max(\pi_{\alpha \to \beta}) \approx 10^{-9}$

We have

- p_r: **probability of Mutation** to b from a r-neighbour of b with $r \geq 2$
  $p_r \leq n \times \pi^r \leq 2000 \times 10^{-18} < 2.10^{-6} \times \pi$
- q_s: **probability** that s **1-neighbours simultaneously** mutate to b with $s \geq 2$
  $q_s \leq n \times \pi^s \leq 2000 \times 10^{-18} < 2.10^{-6} \times \pi$

Numerical remarks

- **length** of promoters $n \in [500 - 2000]$
- **Mutation probability** $\pi = \max(\pi_{\alpha \to \beta}) \approx 10^{-9}$

We have

- p_r: **probability of Mutation** to b from a r-neighbour of b with $r \geq 2$
  $p_r \leq n \times \pi^r \leq 2000 \times 10^{-18} < 2.10^{-6} \times \pi$
- q_s: **probability** that s **1-neighbours simultaneously** mutate to b with $s \geq 2$
  $q_s \leq n \times \pi^s \leq 2000 \times 10^{-18} < 2.10^{-6} \times \pi$

Therefore assuming a single mutation in the promoter is numerically sound

Putative-hit positions.

- Given a **sequence $S(0)$ not containing a k-mer b**,
- a **putative-hit position** is any position of $S(0)$ that can **lead by a mutation to an occurrence of b in $S(1)$**,
- where we assume that a **single** mutation has occurred.

$$S(0) = \texttt{CCCAACAC}, \quad b = \texttt{ACC} \quad \rightsquigarrow \quad \underline{S}(0) = \underline{\texttt{C}}\texttt{CCA}\underline{\texttt{A}}\texttt{CAC},$$

putative-hit positions **underlined** in $\underline{S}(0)$.

Putative-hit positions.

- Given a **sequence $S(0)$ not containing a k-mer b**,
- a **putative-hit position** is any position of $S(0)$ that can **lead by a mutation to an occurrence of b in $S(1)$**,
- where we assume that a **single** mutation has occurred.

$$S(0) = \texttt{CCCAACAC}, \quad b = \texttt{ACC} \qquad \rightsquigarrow \qquad \underline{S}(0) = \underline{\texttt{C}}\texttt{CCA}\underline{\texttt{A}}\texttt{CAC},$$

putative-hit positions **underlined** in $\underline{S}(0)$.

In a random sequence of length n with $\mathcal{A} = \{\texttt{A}, \texttt{C}\}$, let

- $H^{(n)}_{\texttt{A}\to\texttt{C}}$ number of putative-hit-positions $\texttt{A} \to \texttt{C}$,
- $H^{(n)}_{\texttt{C}\to\texttt{A}}$ number of putative-hit-positions $\texttt{C} \to \texttt{A}$,

Then

$$\mathfrak{P}_n \approx \mathbf{E}\big(H^{(n)}_{\texttt{A}\to\texttt{C}}\big) \times \pi_{\texttt{A}\to\texttt{C}} + \mathbf{E}\big(H^{(n)}_{\texttt{C}\to\texttt{A}}\big) \times \pi_{\texttt{C}\to\texttt{A}}$$

Computing via generating functions

Aim:

Compute

$$F_b(z, t_{\mathtt{A}\to\mathtt{C}}, t_{\mathtt{C}\to\mathtt{A}}) = \sum_{n\geq 0} \sum_{0\leq i\leq n-|b|} \sum_{0\leq j\leq n-|b|} f_{n,i,j}\, t^i_{\mathtt{A}\to\mathtt{C}}, t^j_{\mathtt{C}\to\mathtt{A}} z^n$$

where $f_{n,i,j}$ is the probability that a sequence $S_n(0)$ **with no** b, of length n, contains

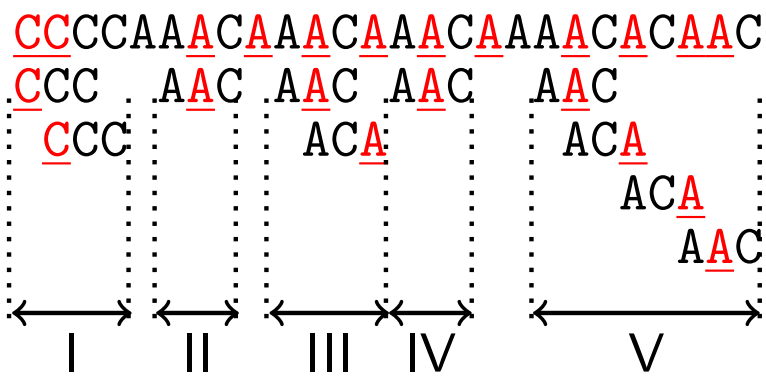
- i putative-hit positions $\mathtt{A}\to\mathtt{C}$
- and j putative-hit positions $\mathtt{C}\to\mathtt{A}$

We have

$$\mathfrak{P}_n = [z^n]\left(\pi_{\mathtt{A}\to\mathtt{C}} \left.\frac{\partial F(z, t_{\mathtt{A}\to\mathtt{C}}, 1)}{\partial t_{\mathtt{A}\to\mathtt{C}}}\right|_{t_{\mathtt{A}\to\mathtt{C}}=1} + \pi_{\mathtt{C}\to\mathtt{A}} \left.\frac{\partial F(z, 1, t_{\mathtt{C}\to\mathtt{A}})}{\partial t_{\mathtt{C}\to\mathtt{A}}}\right|_{t_{\mathtt{C}\to\mathtt{A}}=1}\right)$$

Putative-Hit-Positions and clump analysis

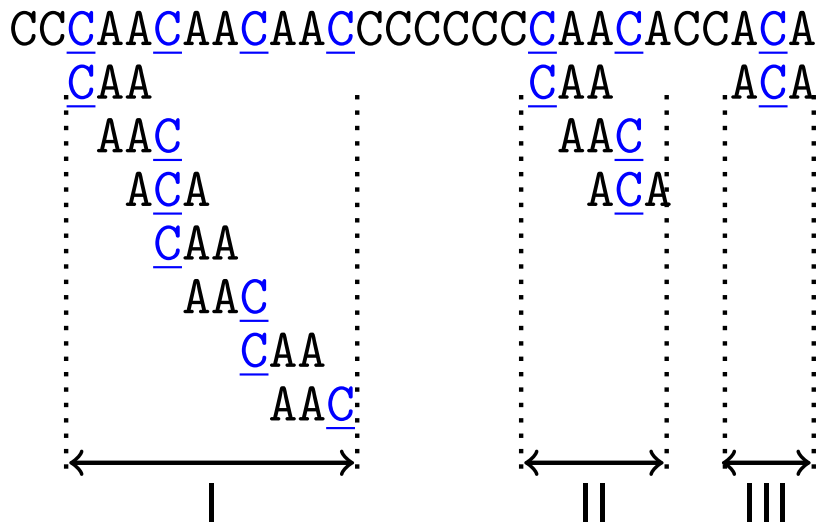
$\mathcal{A} = \{\mathtt{A}, \mathtt{C}\}$ $b = \mathtt{ACC} \longrightarrow d(\mathtt{ACC}, 1) = \{\underline{\mathtt{C}}\mathtt{CC}, \mathtt{A}\underline{\mathtt{A}}\mathtt{C}, \mathtt{AC}\underline{\mathtt{A}}\}$



- (left) $b = \mathtt{ACC}$ - in clump I, when the right extension of a clump adds a new putative-hit position, this position is not necessarily in the extension, but possibly backwards left

Putative-Hit-Positions and clump analysis

$\mathcal{A} = \{\mathtt{A}, \mathtt{C}\} \quad b' = \mathtt{AAA} \longrightarrow d(\mathtt{AAA}, 1) = \{\underline{\mathtt{C}}\mathtt{AA}, \mathtt{A}\underline{\mathtt{C}}\mathtt{A}, \mathtt{AA}\underline{\mathtt{C}}\}$



- (right)$b' = \mathtt{AAA}$ - clump I contains 7 occurrences of $d(\mathtt{AAA})$, but only 4 putative-hit positions for $b' = \mathtt{AAA}$. The number of word occurrences is not the relevant statistics for counting putative-hit positions

I - Automaton approach

Clumps of the set of words $\mathcal{U} = \{aabaa, baab\}$

$\mathcal{E}_{w_1,w_2}$ correlation set from w_1 to w_2

$$\mathcal{E}_{aabaa,aabaa} = \{baa, abaa\} \qquad \mathcal{E}_{baab,baab} = \{aab\}$$
$$\mathcal{E}_{aabaa,baab} = \{b\} \qquad \mathcal{E}_{baab,aabaa} = \{aa\}$$

Clumps of the set of words $\mathcal{U} = \{aabaa, baab\}$

$\mathcal{E}_{w_1,w_2}$ correlation set from w_1 to w_2

$$\begin{array}{llll} \mathcal{E}_{aabaa,aabaa} & = \{baa, abaa\} & \mathcal{E}_{baab,baab} & = \{aab\} \\ \mathcal{E}_{aabaa,baab} & = \{b\} & \mathcal{E}_{baab,aabaa} & = \{aa\} \end{array}$$

Algorithm

1. Build the set of strings

$$X = \{aabaa.(\epsilon + \mathcal{E}_{aabaa,aabaa})\} \cup \{aabaa.\mathcal{E}_{aabaa,baab}\} \cup \{baab.(\epsilon + \mathcal{E}_{baab,baab})\} \cup \{baab.\mathcal{E}_{baab,aabaa}\}$$

Clumps of the set of words $\mathcal{U} = \{aabaa, baab\}$

$\mathcal{E}_{w_1,w_2}$ correlation set from w_1 to w_2

$$\begin{array}{llll} \mathcal{E}_{aabaa,aabaa} & = \{baa, abaa\} & \mathcal{E}_{baab,baab} & = \{aab\} \\ \mathcal{E}_{aabaa,baab} & = \{b\} & \mathcal{E}_{baab,aabaa} & = \{aa\} \end{array}$$

Algorithm

1. Build the set of strings

$$\begin{array}{rlll} X = & \{aabaa.(\epsilon + \mathcal{E}_{aabaa,aabaa})\} & \cup & \{aabaa.\mathcal{E}_{aabaa,baab}\} \\ \cup & \{baab.(\epsilon + \mathcal{E}_{baab,baab})\} & \cup & \{baab.\mathcal{E}_{baab,aabaa}\} \end{array}$$

2. Build a trie $\mathcal{T}$ on X

Clumps of the set of words $\mathcal{U} = \{aabaa, baab\}$

$\mathcal{E}_{w_1,w_2}$ correlation set from w_1 to w_2

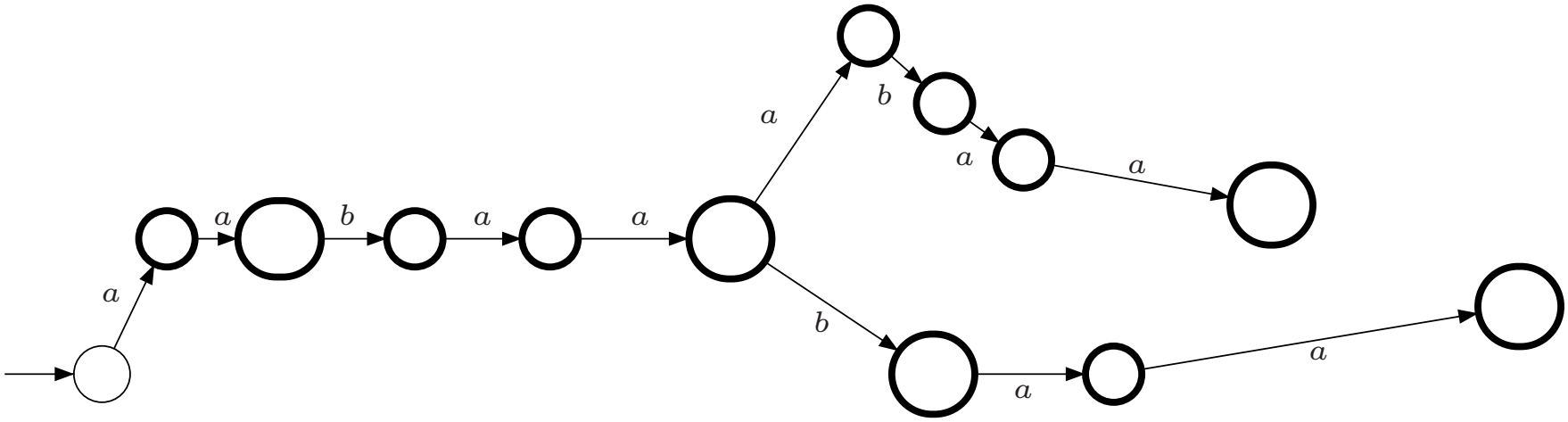
$$\begin{array}{ll} \mathcal{E}_{aabaa,aabaa} = \{baa, abaa\} & \mathcal{E}_{baab,baab} = \{aab\} \\ \mathcal{E}_{aabaa,baab} = \{b\} & \mathcal{E}_{baab,aabaa} = \{aa\} \end{array}$$

Algorithm

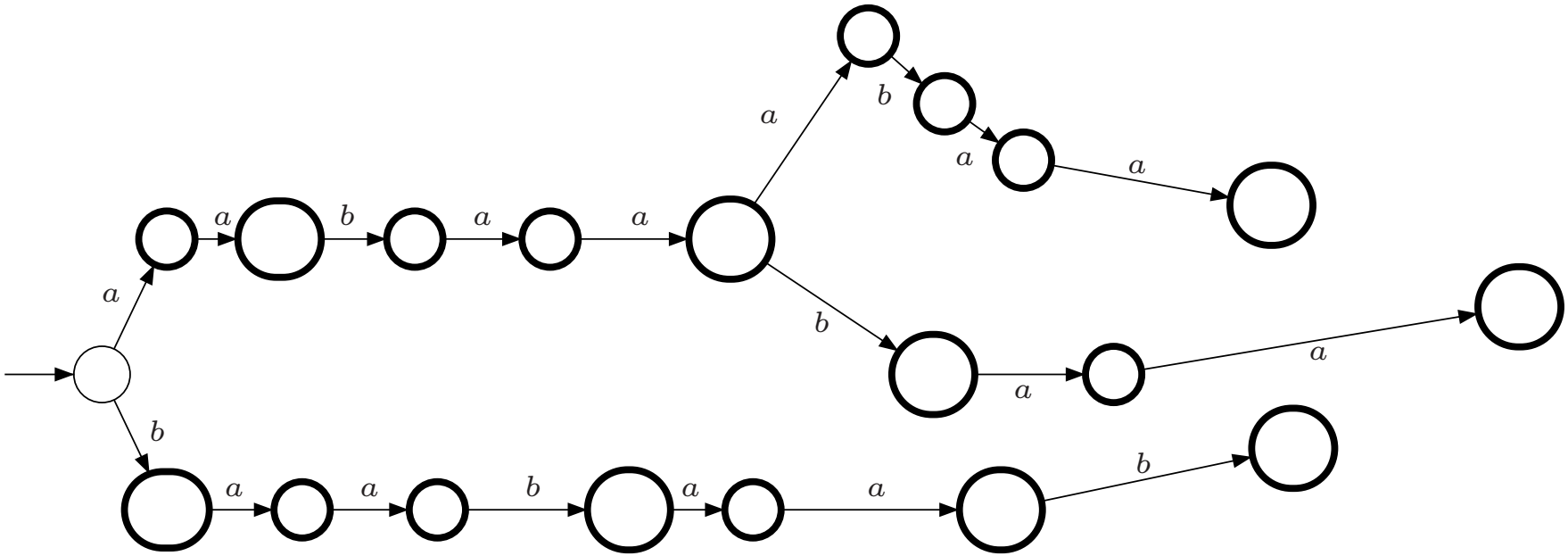
1. Build the set of strings
$$X = \{aabaa.(\epsilon + \mathcal{E}_{aabaa,aabaa})\} \cup \{aabaa.\mathcal{E}_{aabaa,baab}\} \cup \{baab.(\epsilon + \mathcal{E}_{baab,baab})\} \cup \{baab.\mathcal{E}_{baab,aabaa}\}$$
2. Build a trie $\mathcal{T}$ on X
3. Build a **Aho-Corasick like automaton** upon $\mathcal{T}$. For each node ν of $\mathcal{T}$ with "access word" v, use the transition function δ

   $\delta(\nu, \ell) =$ node accessed by the **longuest prefix** in X that is **suffix** of $v.\ell$

$X = \{aabaa, aabaabaa, aabaaabaa, aabaab \quad \}$

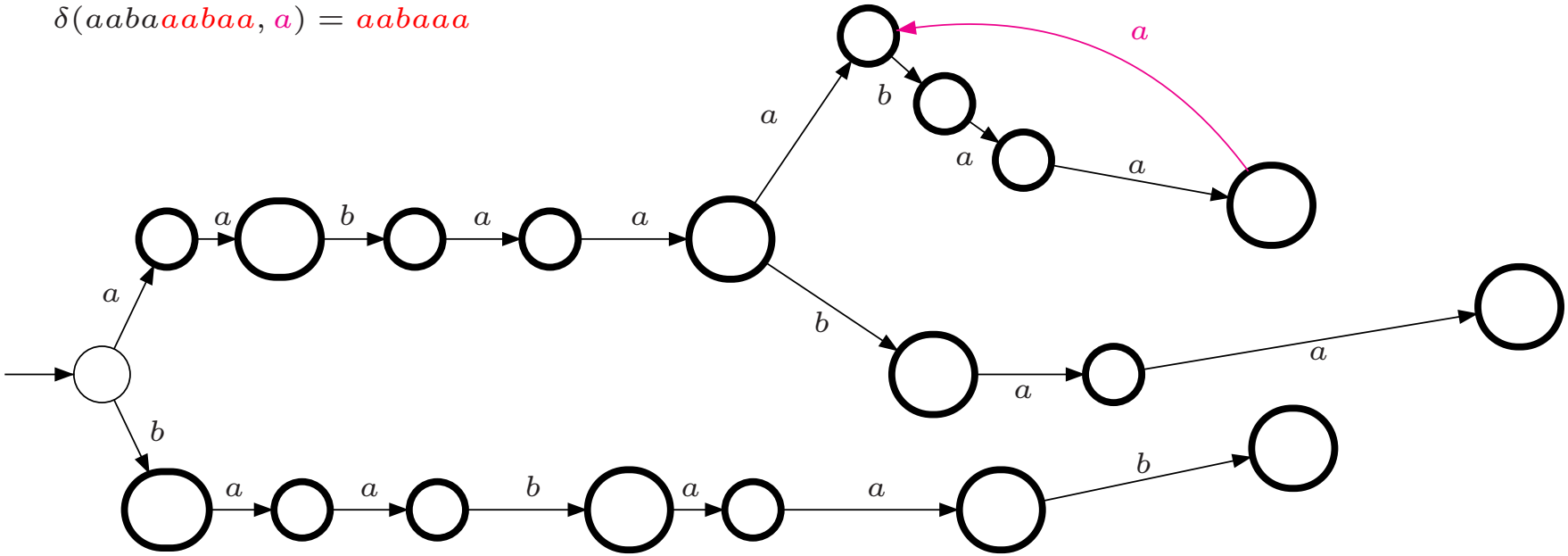


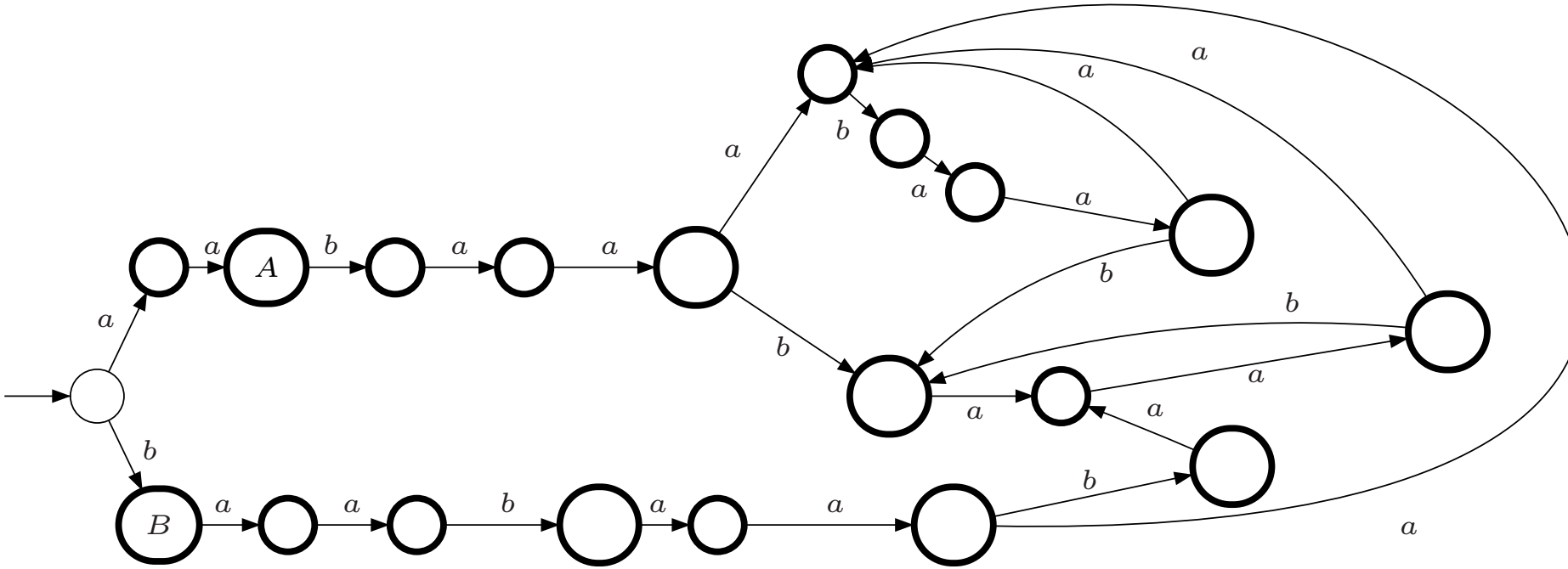
$X = \{aabaa, aabaabaa, aabaaabaa, aabaab, baab, baabaab\}$



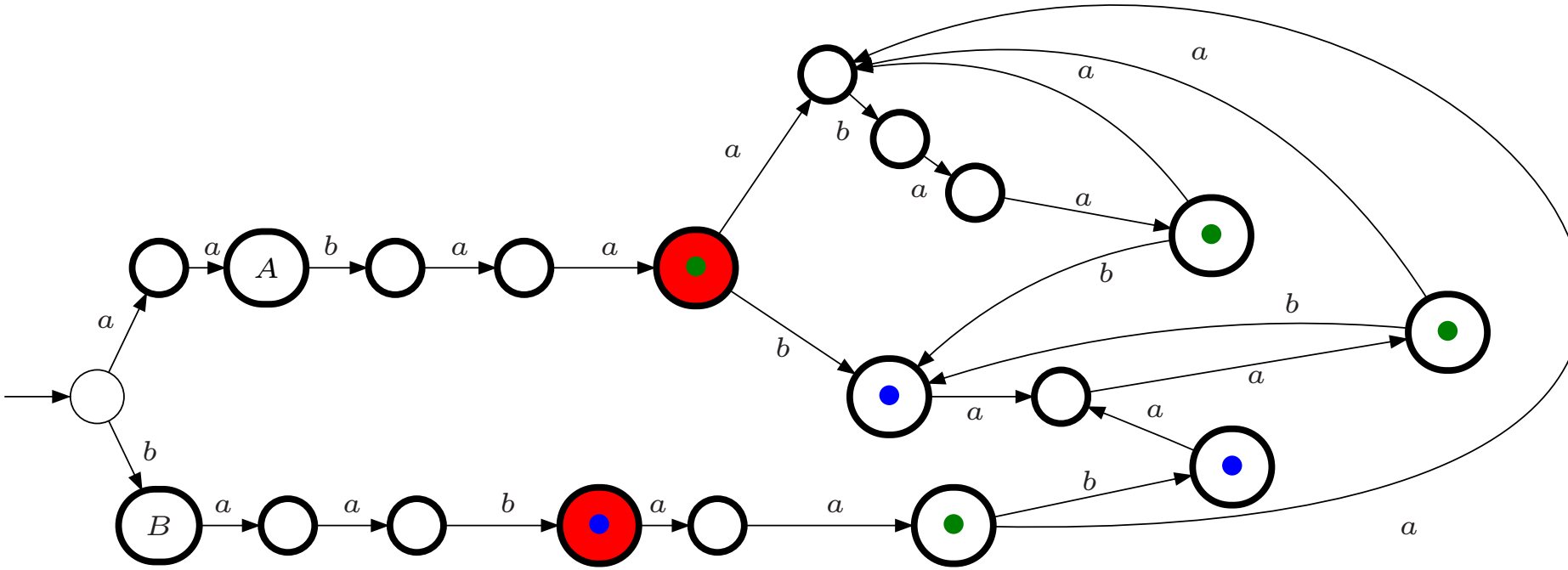
$X = \{aabaa, aabaabaa, aabaaabaa, aabaab, baab, baabaab\}$

$\delta(aabaaabaa, a) = aabaaa$



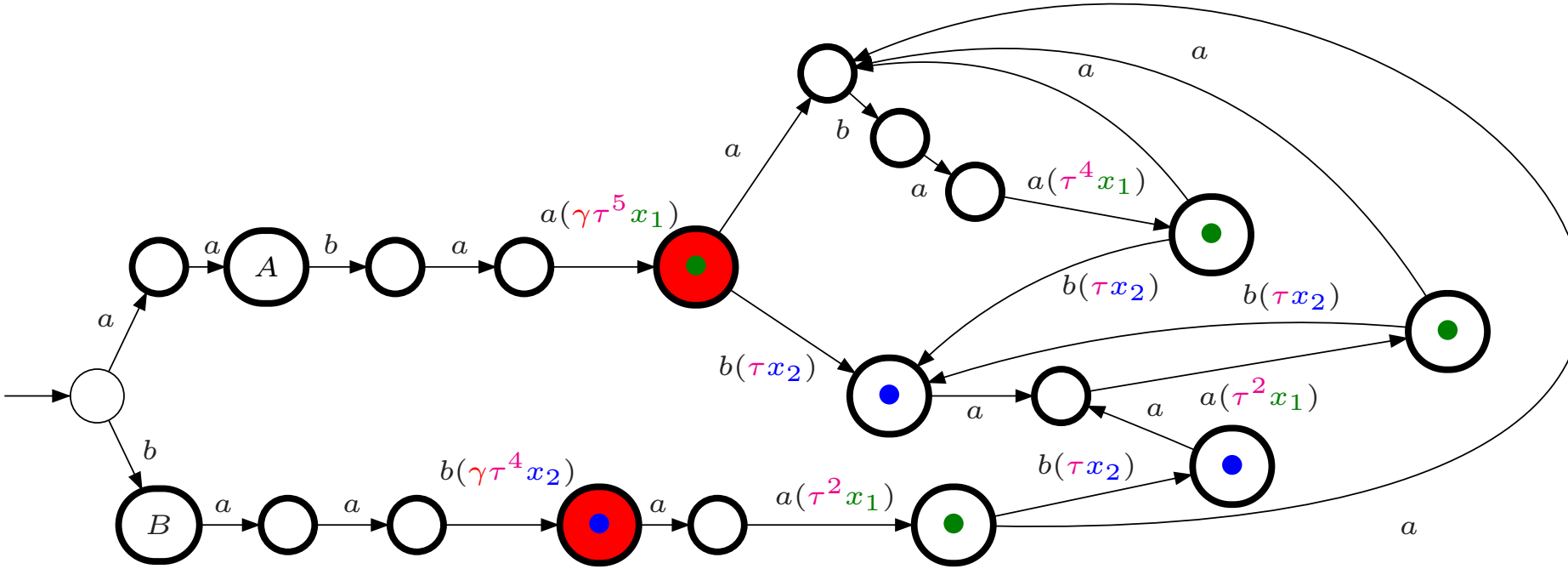


An automaton for $\mathcal{V} = \{v_1 = aabaa, v_2 = baab\}$. All transitions labeled by a and b ending respectively on state A and B are omitted.



An automaton for $\mathcal{V} = \{v_1 = aabaa, v_2 = baab\}$. All transitions labeled by a and b ending respectively on state A and B are omitted.

- ●, ● → the corresponding prefix (or state) ends with some occurrence of $aabaa$, $baab$.
- **red states** → states where we have entered a **new clump**



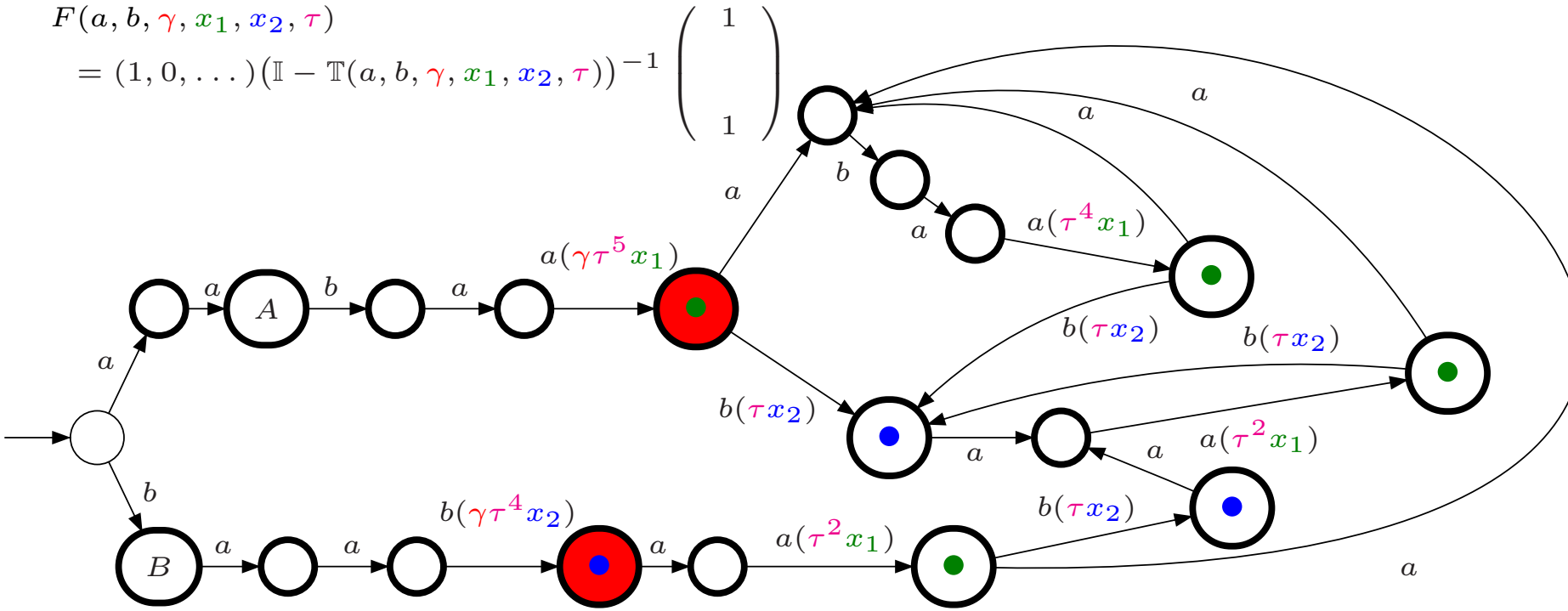
An automaton for $\mathcal{V} = \{v_1 = aabaa, v_2 = baab\}$. All transitions labeled by a and b ending respectively on state A and B are omitted.

- $\bullet$, $\bullet$ $\rightarrow$ the corresponding prefix (or state) ends with some occurrence of $aabaa, baab$.
- **red states** $\rightarrow$ states where we have entered a **new clump**

Formal weights on transitions

- $\gamma \rightarrow$ the number of clumps
- $\tau \rightarrow$ total length of clumps
- $x_1, x_2 \rightarrow$ occurrences of $aabaa, baab$

$$F(a,b,\gamma,x_1,x_2,\tau) = (1,0,\dots)\big(\mathbb{I}-\mathbb{T}(a,b,\gamma,x_1,x_2,\tau)\big)^{-1}\begin{pmatrix}1\\ \\ 1\end{pmatrix}$$

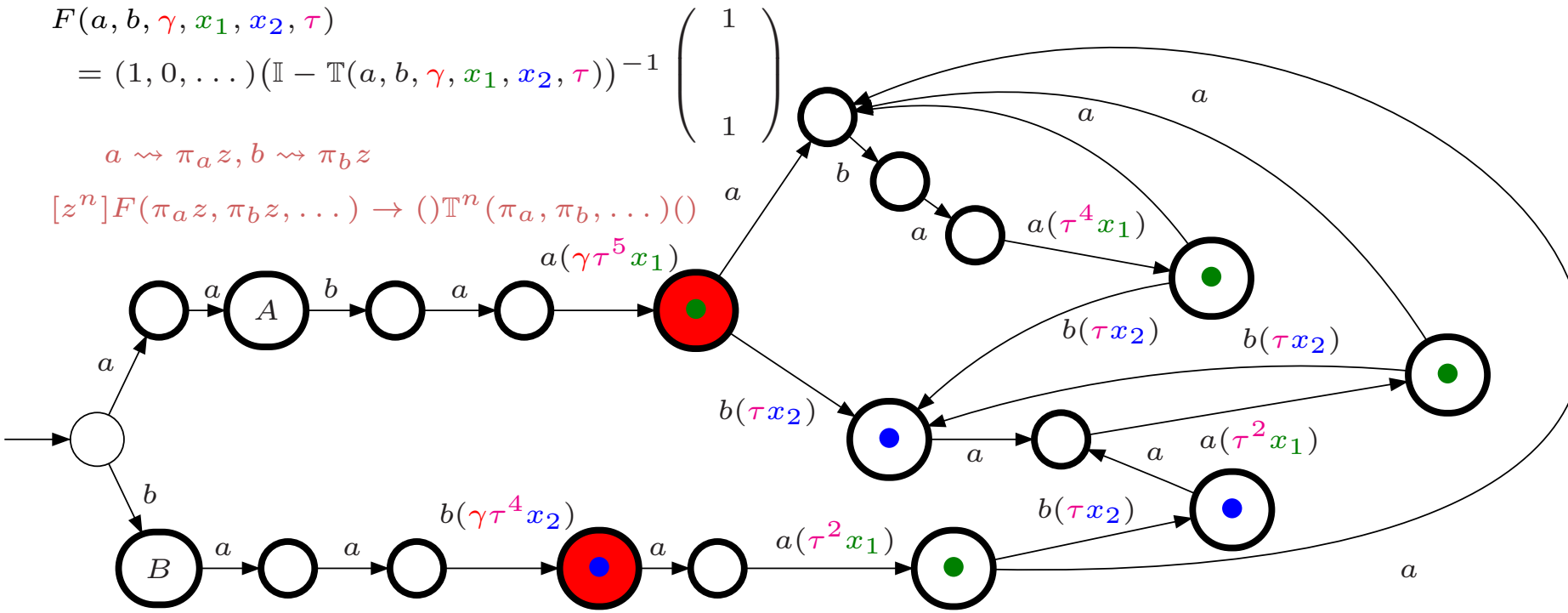


An automaton for $\mathcal{V} = \{v_1 = aabaa, v_2 = baab\}$. All transitions labeled by a and b ending respectively on state A and B are omitted.

- ●, ● → the corresponding prefix (or state) ends with some occurrence of $aabaa, baab$.
- **red states** → states where we have entered a **new clump**

Formal weights on transitions

- γ → the number of clumps
- τ → total length of clumps
- x_1, x_2 → occurrences of $aabaa, baab$



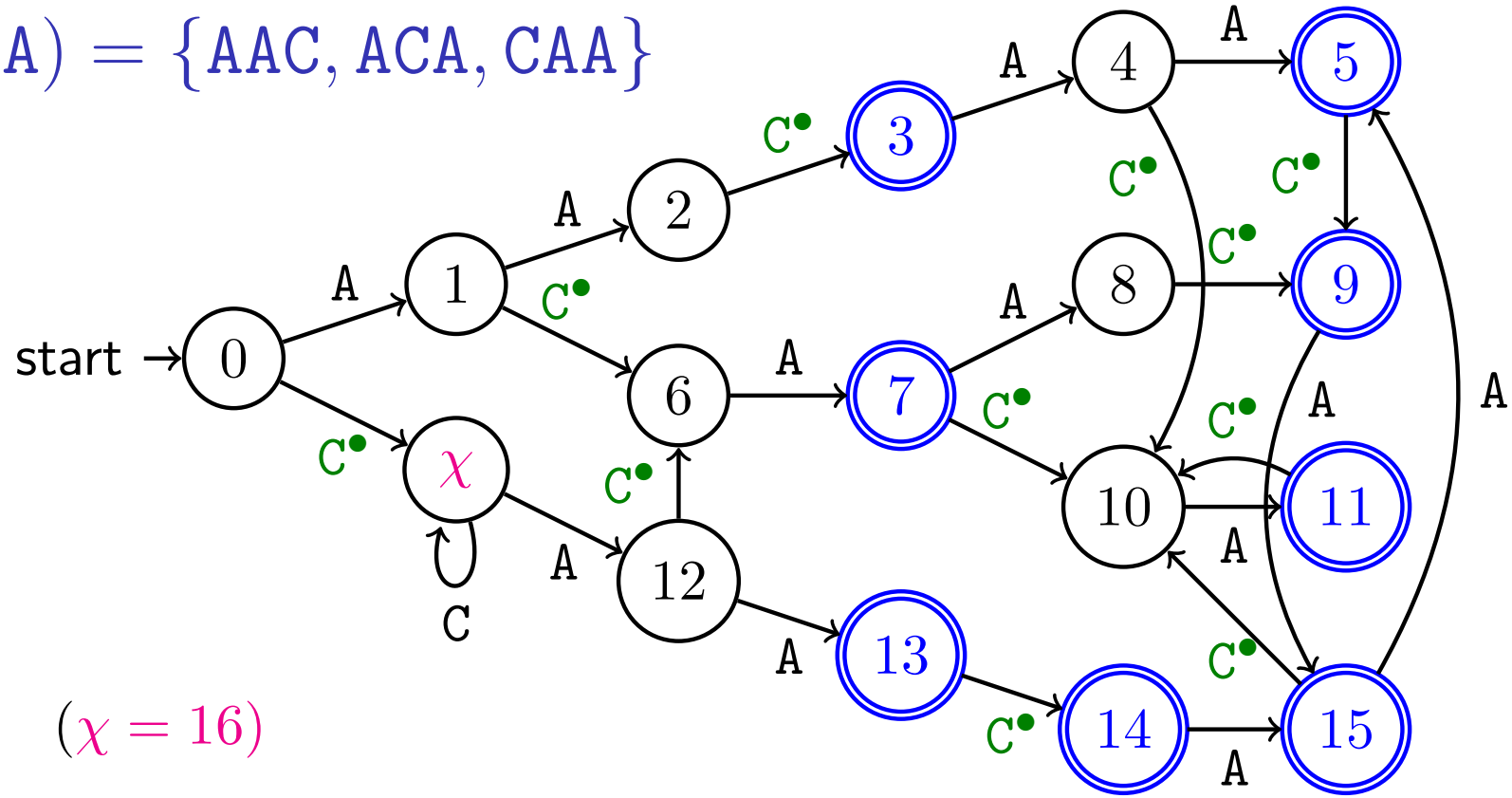
An automaton for $\mathcal{V} = \{v_1 = aabaa, v_2 = baab\}$. All transitions labeled by a and b ending respectively on state A and B are omitted.

- ●, ● → the corresponding prefix (or state) ends with some occurrence of $aabaa, baab$.
- **red states** → states where we have entered a **new clump**

Formal weights on transitions

- γ → the number of clumps
- τ → total length of clumps
- x_1, x_2 → occurrences of $aabaa, baab$

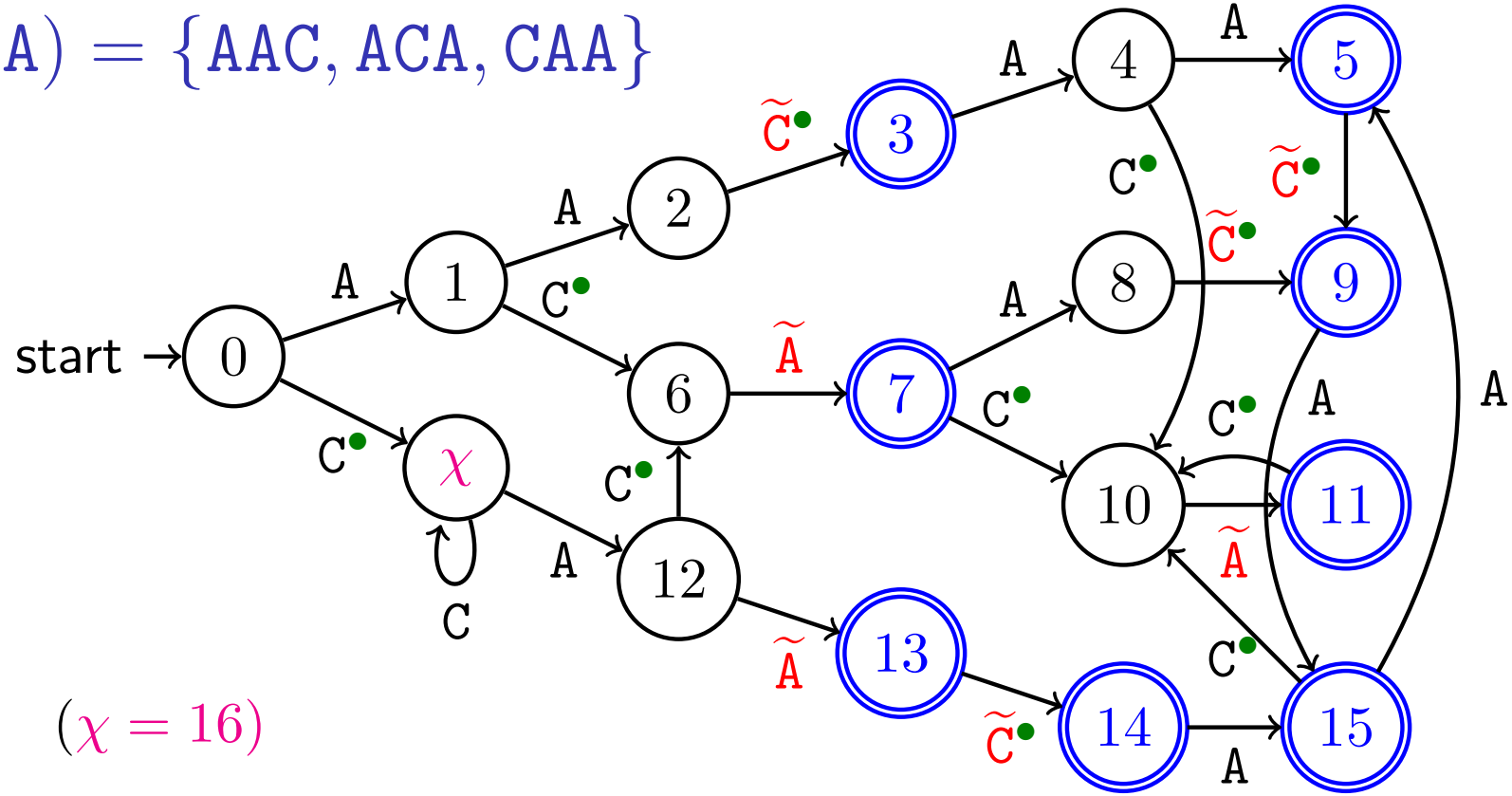
Automaton for constrained clumps of $d(\mathtt{AAA}) = \{\mathtt{AAC}, \mathtt{ACA}, \mathtt{CAA}\}$



$(\chi = 16)$

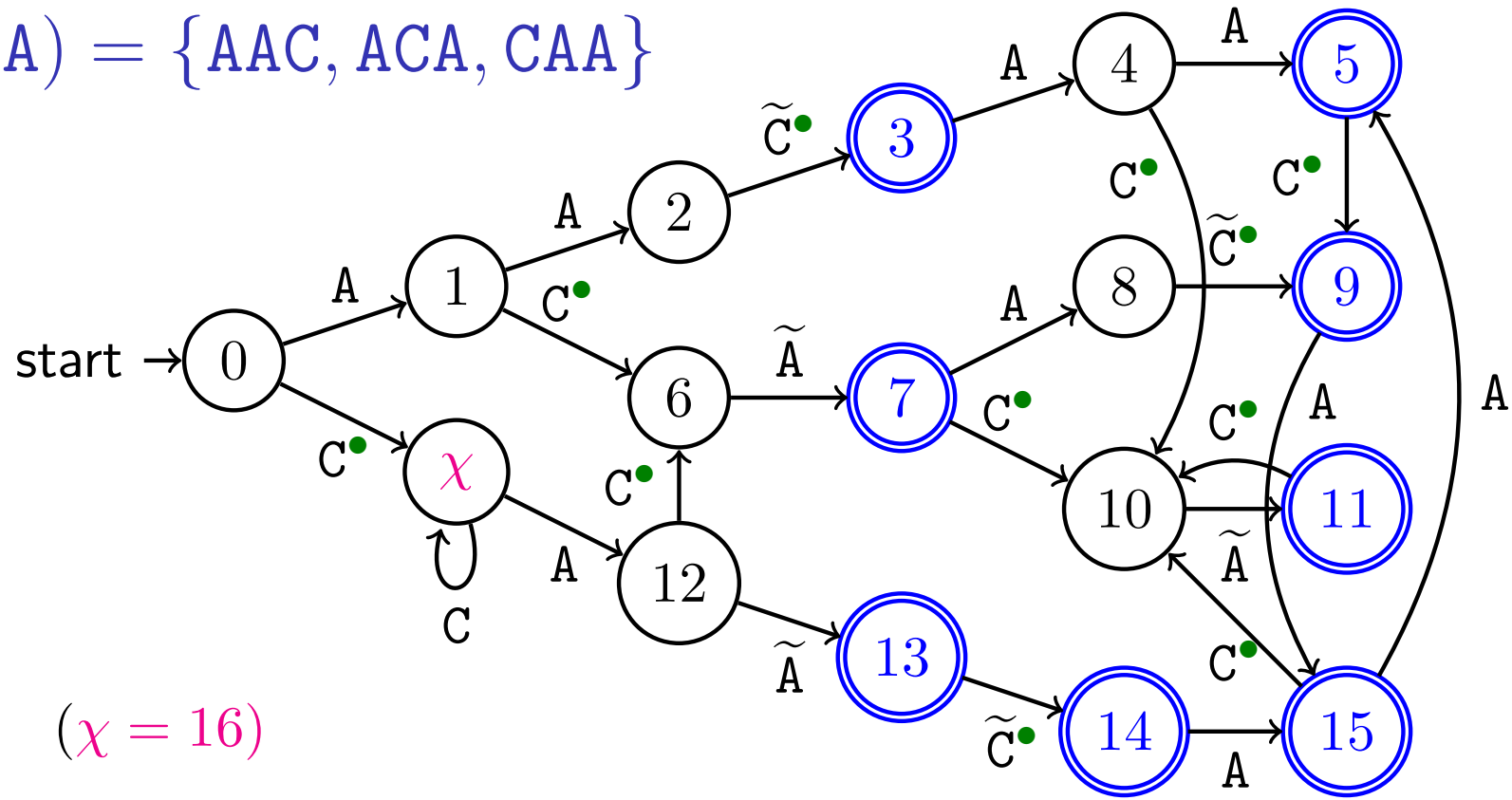
- **Double circles** signals an occurrence of a word of $d(aaa)$.
- **Avoiding** AAA leads to **missing transitions** A
- The **missing transitions** C **point to** the **state** χ.
- • **characters** mark **putative-hit-positions**

Automaton for constrained clumps of $d(\mathtt{AAA}) = \{\mathtt{AAC}, \mathtt{ACA}, \mathtt{CAA}\}$



- **Double circles** signals an occurrence of a word of $d(aaa)$.
- **Avoiding** `AAA` leads to **missing transitions** `A`
- The **missing transitions** `C` **point to** the **state** χ.
- • **characters** mark **putative-hit-positions**
- Transitions covered by tildes ($\widetilde{\mathtt{A}}, \widetilde{\mathtt{C}}$) emits a signal counting a putative-hit position.

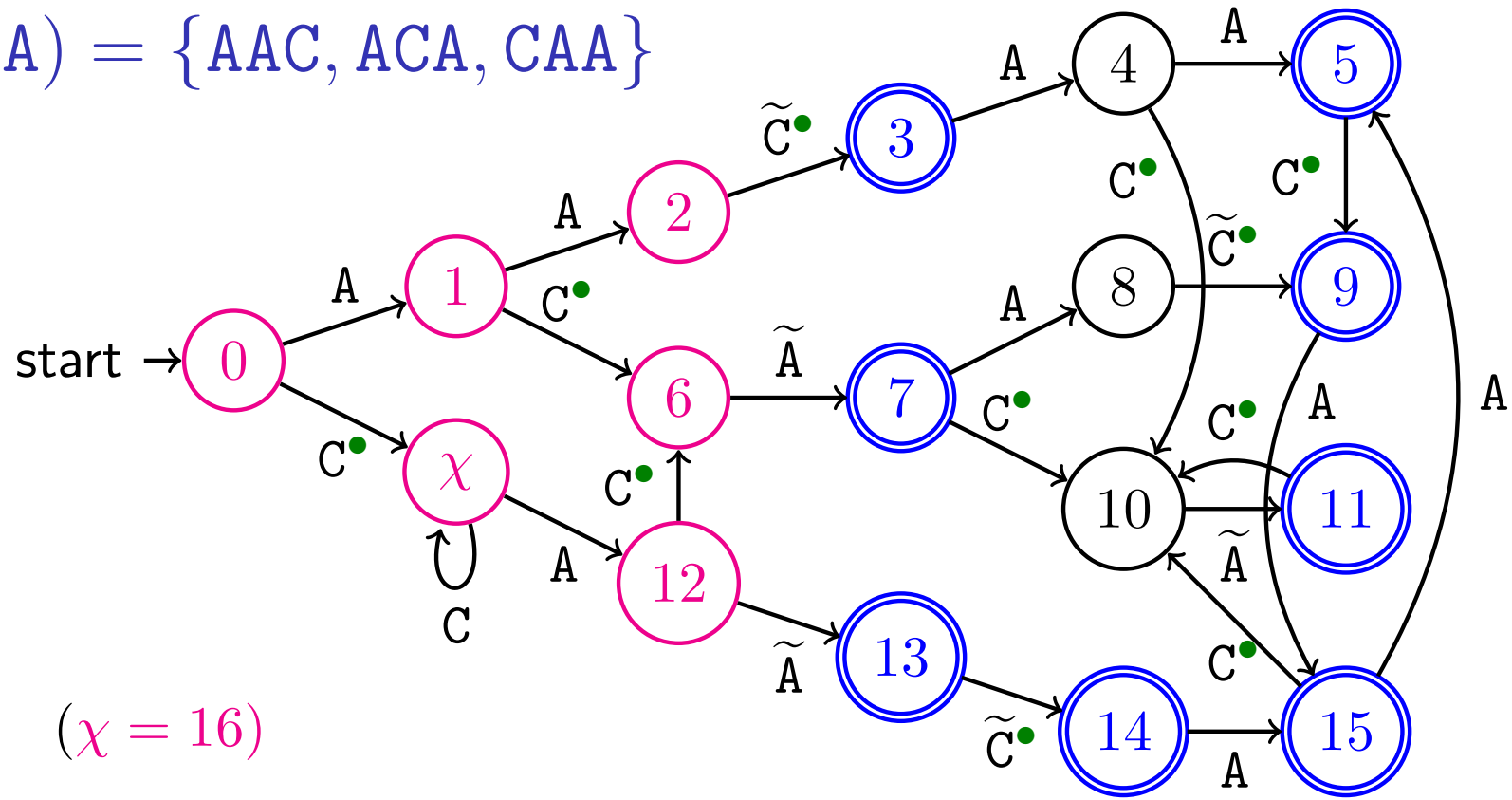
Automaton for constrained clumps of $d(\mathtt{AAA}) = \{\mathtt{AAC}, \mathtt{ACA}, \mathtt{CAA}\}$



$(\chi = 16)$

- $O = \{q, \quad \delta(0, w) = q, \ w \in X\}$, (**occurrence** of a word of $d(aaa)$).

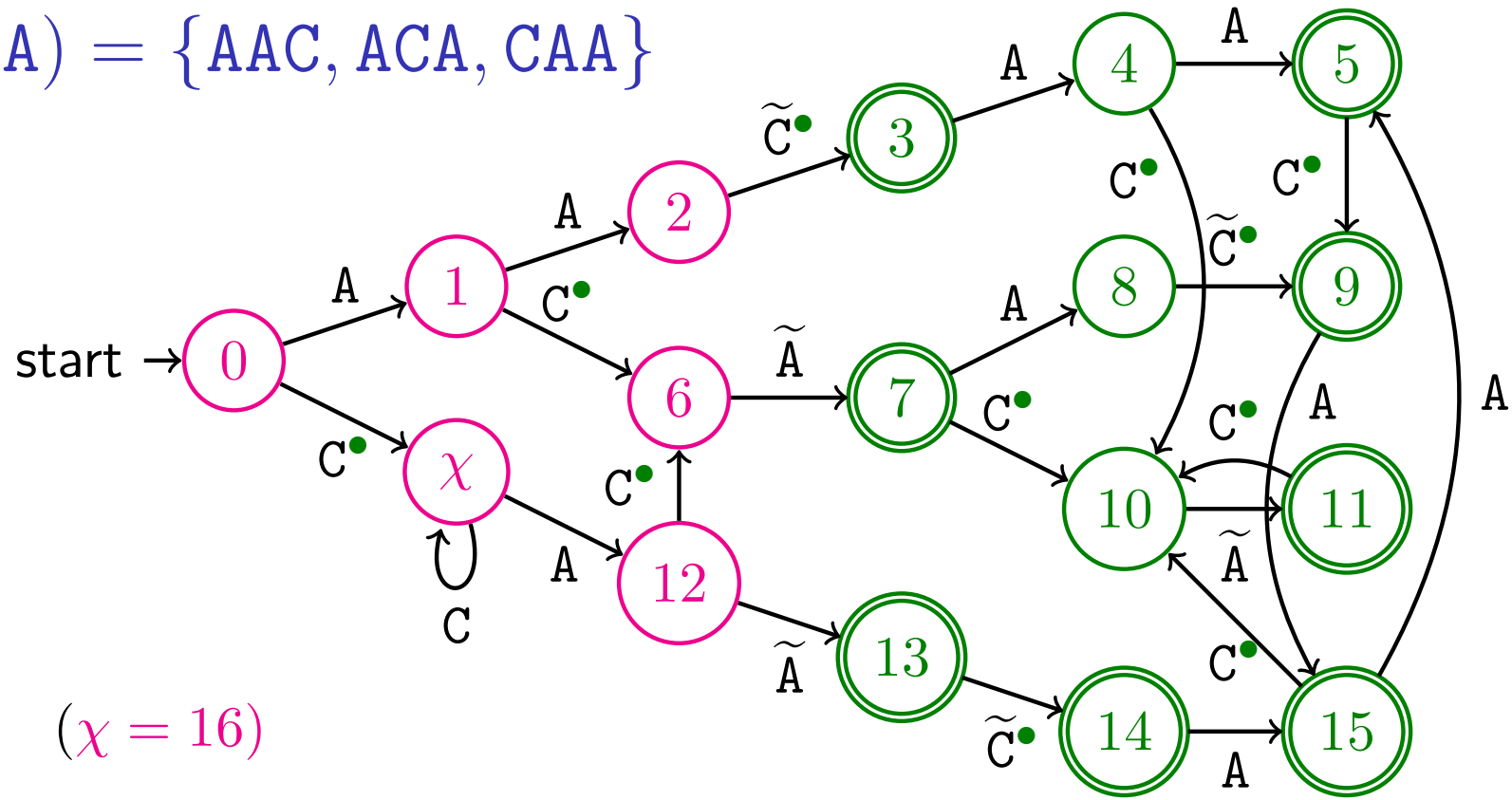
Automaton for constrained clumps of $d(\mathtt{AAA}) = \{\mathtt{AAC}, \mathtt{ACA}, \mathtt{CAA}\}$



$(\chi = 16)$

- $O = \{q, \quad \delta(0, w) = q, \ w \in X\}$, (**occurrence** of a word of $d(aaa)$).
- $\overline{E} = \{q, \quad \delta(0, w) = q, \ w \in \widehat{\mathrm{Pref}}(d(b))\}$, with $\widehat{\mathrm{Pref}}(d(b))$ set of **strict prefixes** of words of $d(b)$.

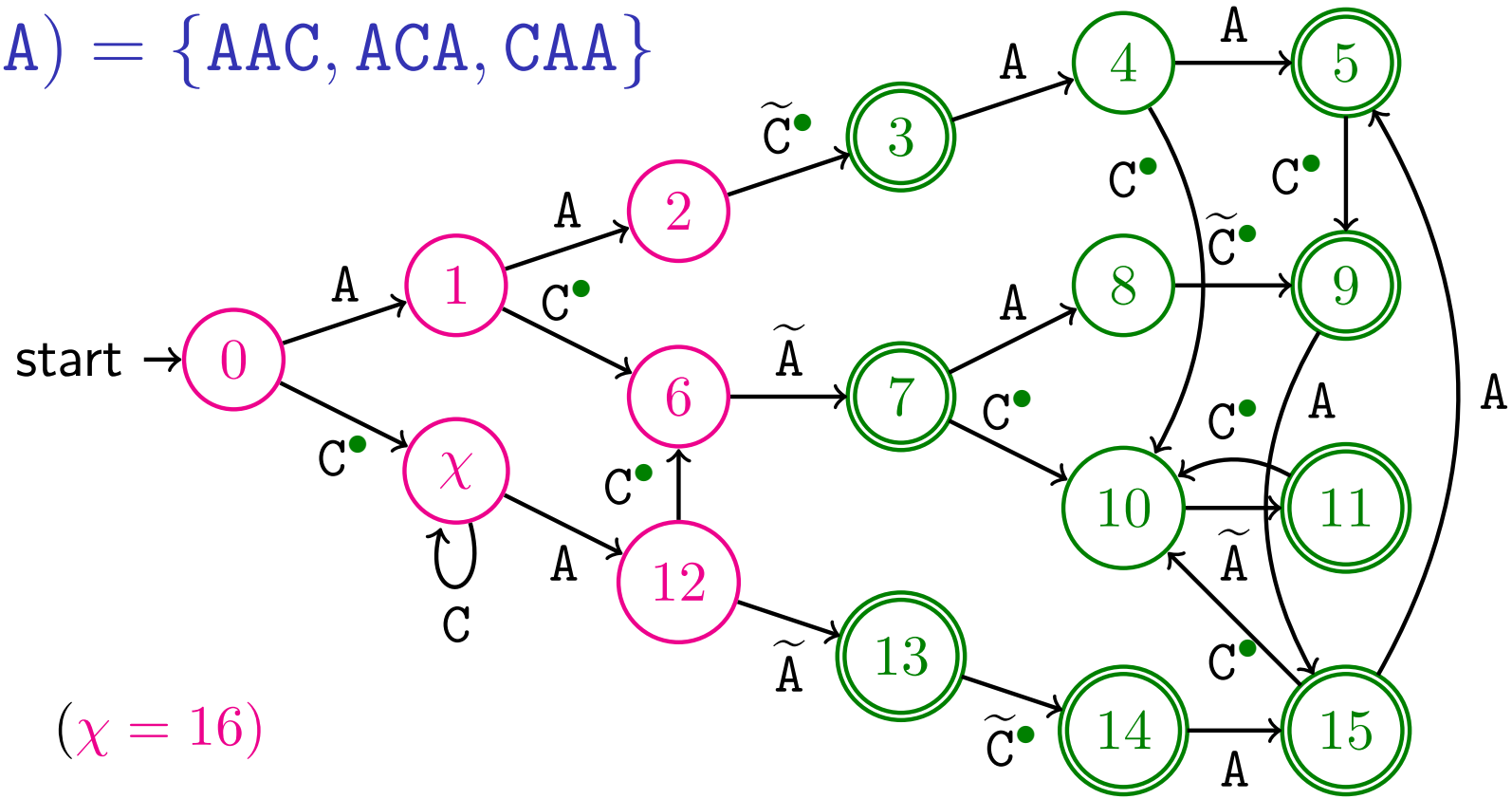
Automaton for constrained clumps of $d(\mathtt{AAA}) = \{\mathtt{AAC}, \mathtt{ACA}, \mathtt{CAA}\}$



$(\chi = 16)$

- $O = \{q, \quad \delta(0, w) = q, \ w \in X\}$, (**occurrence** of a word of $d(aaa)$).
- $\overline{E} = \{q, \quad \delta(0, w) = q, \ w \in \widehat{\mathrm{Pref}}(d(b))\}$, with $\widehat{\mathrm{Pref}}(d(b))$ set of **strict prefixes** of words of $d(b)$.
- **Clump-Core** of the automaton $E = Q \setminus \overline{E}$

Automaton for constrained clumps of $d(\mathtt{AAA}) = \{\mathtt{AAC}, \mathtt{ACA}, \mathtt{CAA}\}$

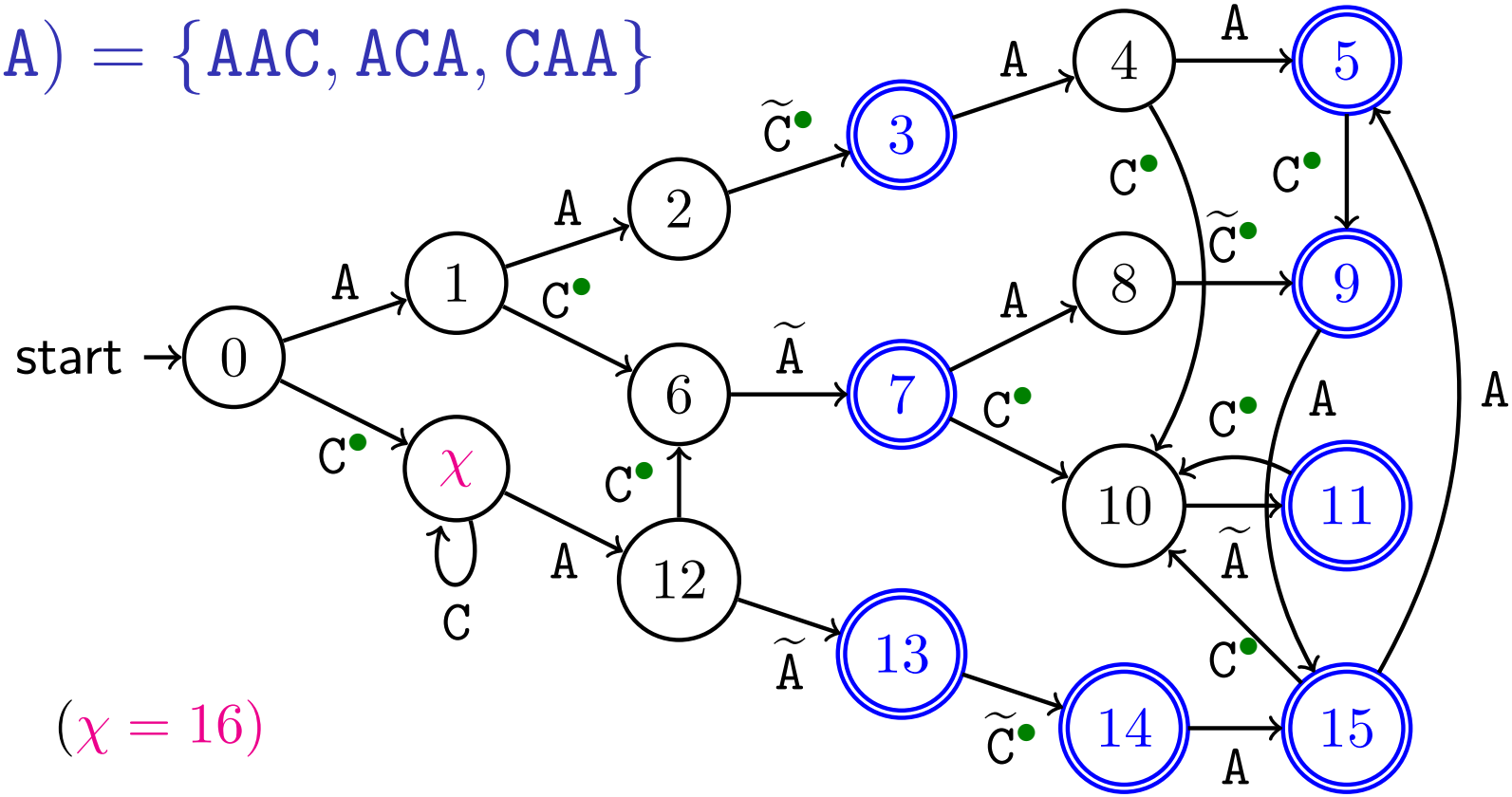


$(\chi = 16)$

- $O = \{q, \quad \delta(0, w) = q, \ w \in X\}$, (**occurrence** of a word of $d(aaa)$).
- $\overline{E} = \{q, \quad \delta(0, w) = q, \ w \in \widehat{\mathrm{Pref}}(d(b))\}$, with $\widehat{\mathrm{Pref}}(d(b))$ set of **strict prefixes** of words of $d(b)$.
- **Clump-Core** of the automaton $E = Q \setminus \overline{E}$
- **Markov property:** $\forall q \in E, \qquad |\{w \in \mathcal{A}^{|b|}; \delta(x, w) = q\}| = 1$

Definition of an auxiliary function θ

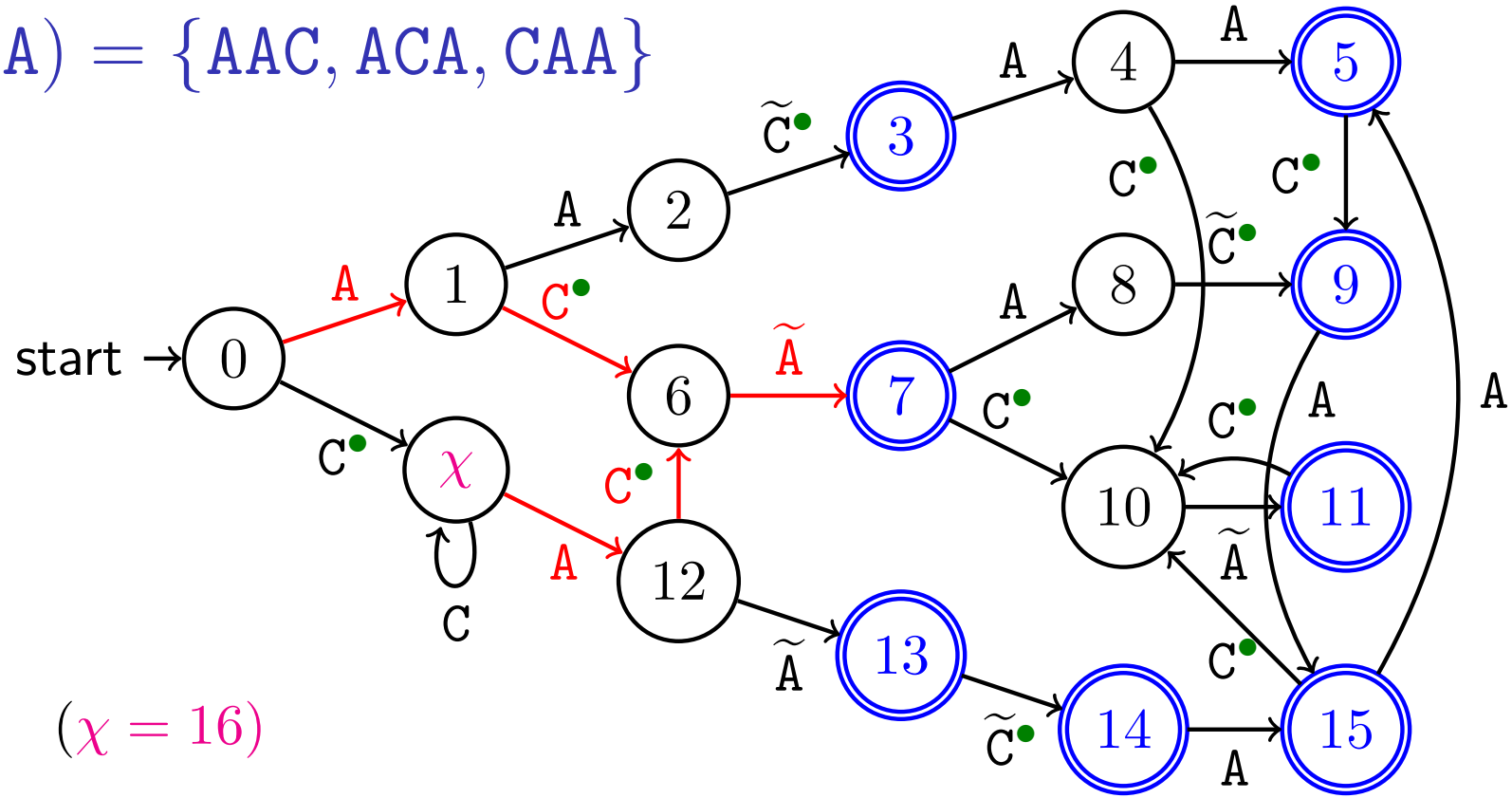
$d(\mathtt{AAA}) = \{\mathtt{AAC}, \mathtt{ACA}, \mathtt{CAA}\}$



$(\chi = 16)$

Definition of an auxiliary function θ

$d(\mathtt{AAA}) = \{\mathtt{AAC}, \mathtt{ACA}, \mathtt{CAA}\}$

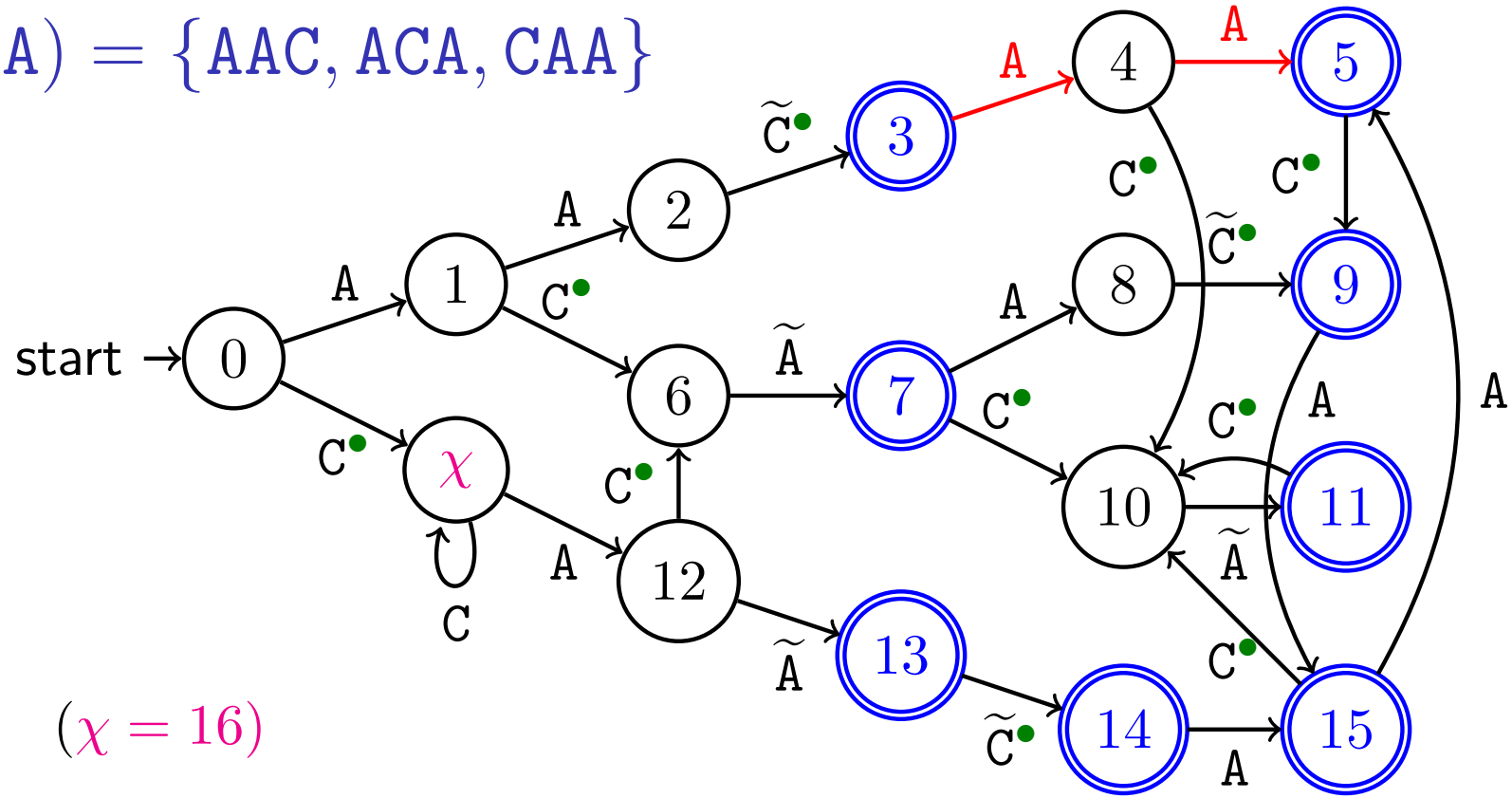


$(\chi = 16)$

$\theta(7) = \mathtt{ACA}$

Definition of an auxiliary function θ

$d(\mathtt{AAA}) = \{\mathtt{AAC}, \mathtt{ACA}, \mathtt{CAA}\}$

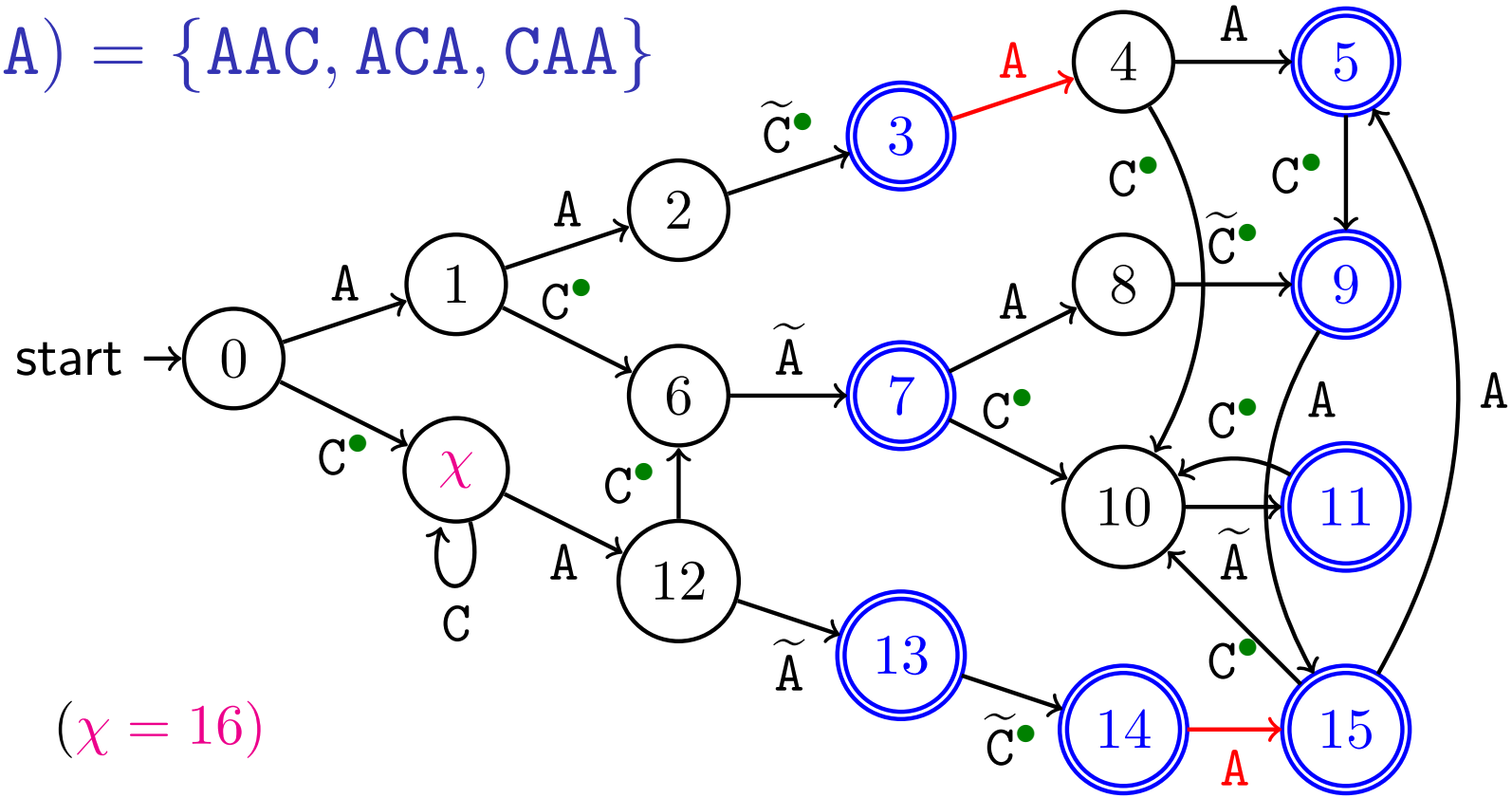


$(\chi = 16)$

$\theta(7) = \mathtt{ACA}\ , \quad \theta(5) = \mathtt{AA}$

Definition of an auxiliary function θ

$d(\mathtt{AAA}) = \{\mathtt{AAC}, \mathtt{ACA}, \mathtt{CAA}\}$



$(\chi = 16)$

$$\theta(7) = \mathtt{ACA}\ , \quad \theta(5) = \mathtt{AA}\ , \quad \theta(4) = \theta(15) = \mathtt{A}$$

Formal definition of θ

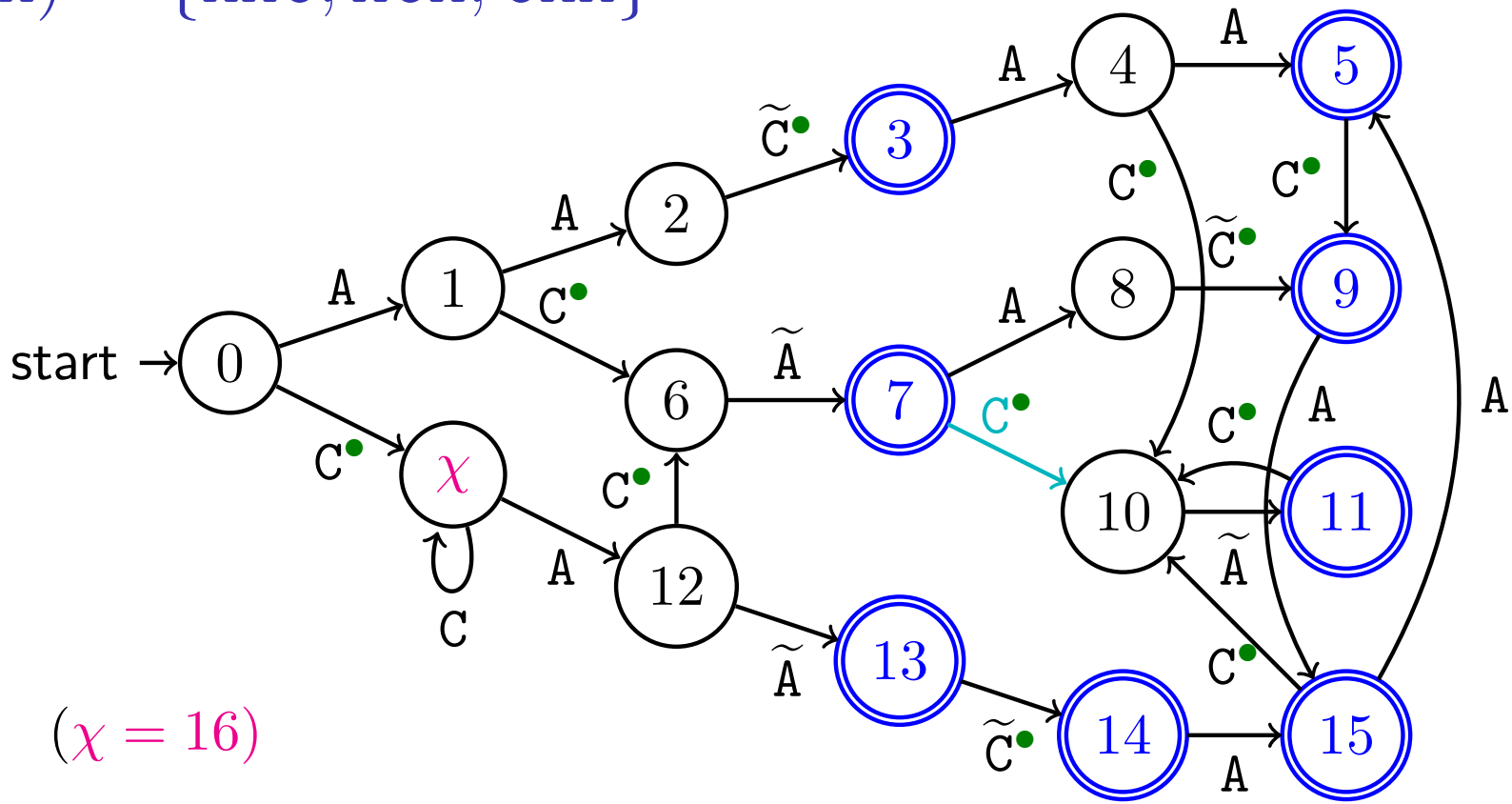
For each state $o \in O$ (recognizing an occurrence of $d(b)$),

$$\theta(o) = \left\{ \begin{array}{l} w \text{ with } |w| \leq |b|, \text{ of } \textbf{maximal length}, \\ \text{verifying} \left| \begin{array}{l} (a) \text{ there exists } q \text{ such that } \delta(q, w) = o, \\ (b) \textbf{ there is no } u \in \widehat{\mathrm{Pref}}(w) \\ \qquad \text{such that } \delta(q, u) \in O \end{array} \right. \end{array} \right.$$

By the **Markov** property, $\theta(o)$ defines a **unique word**

Adjacency matrix $\mathbb{H}(t) = (h_{ij}(t))$

$d(\mathtt{AAA}) = \{\mathtt{AAC}, \mathtt{ACA}, \mathtt{CAA}\}$

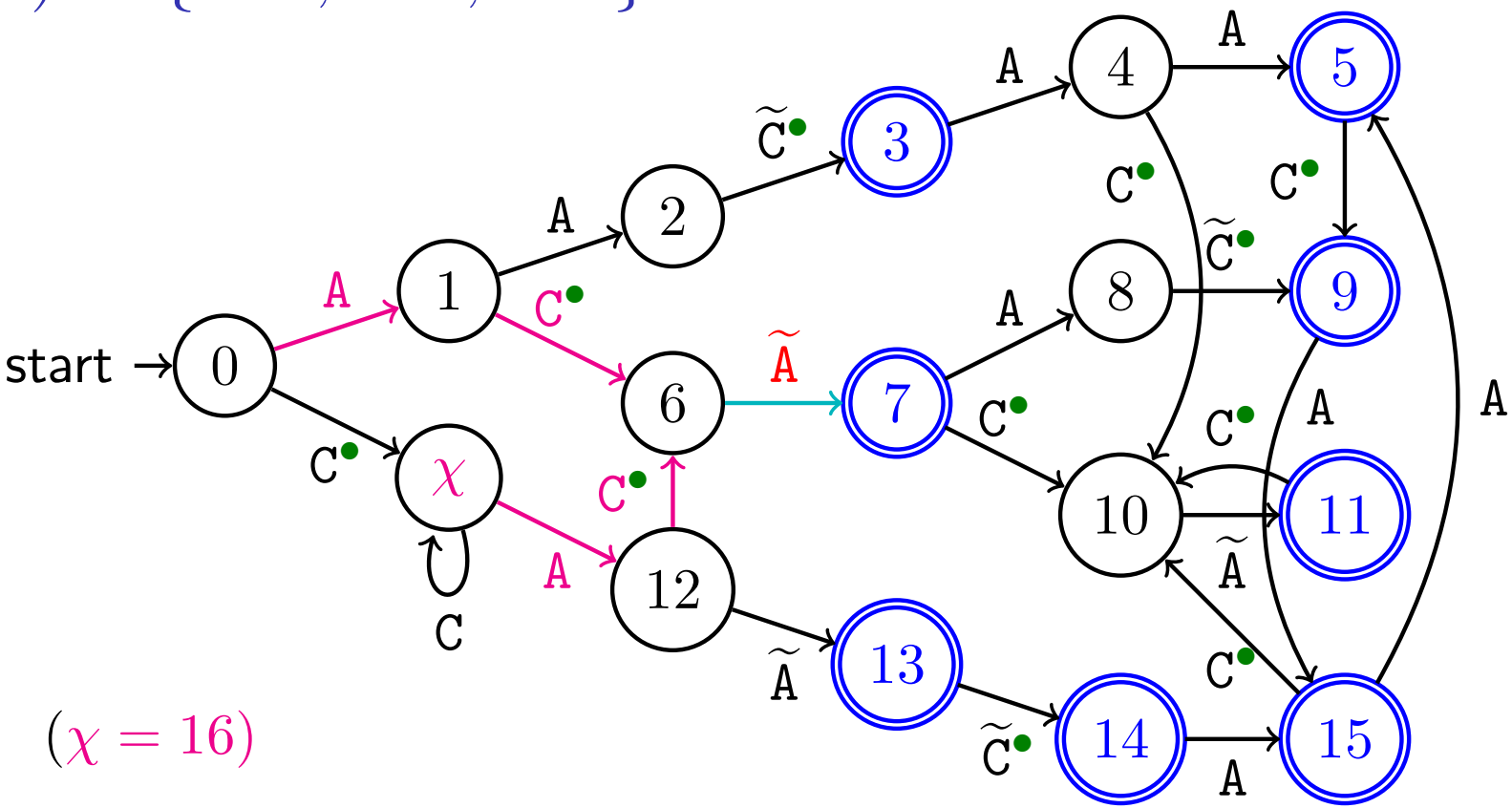


$(\chi = 16)$

- $h_{7,10}(t_{\mathtt{C}\to\mathtt{A}}) = \nu_{\mathtt{C}}$

Adjacency matrix $\mathbb{H}(t) = (h_{ij}(t))$

$d(\mathtt{AAA}) = \{\mathtt{AAC}, \mathtt{ACA}, \mathtt{CAA}\}$



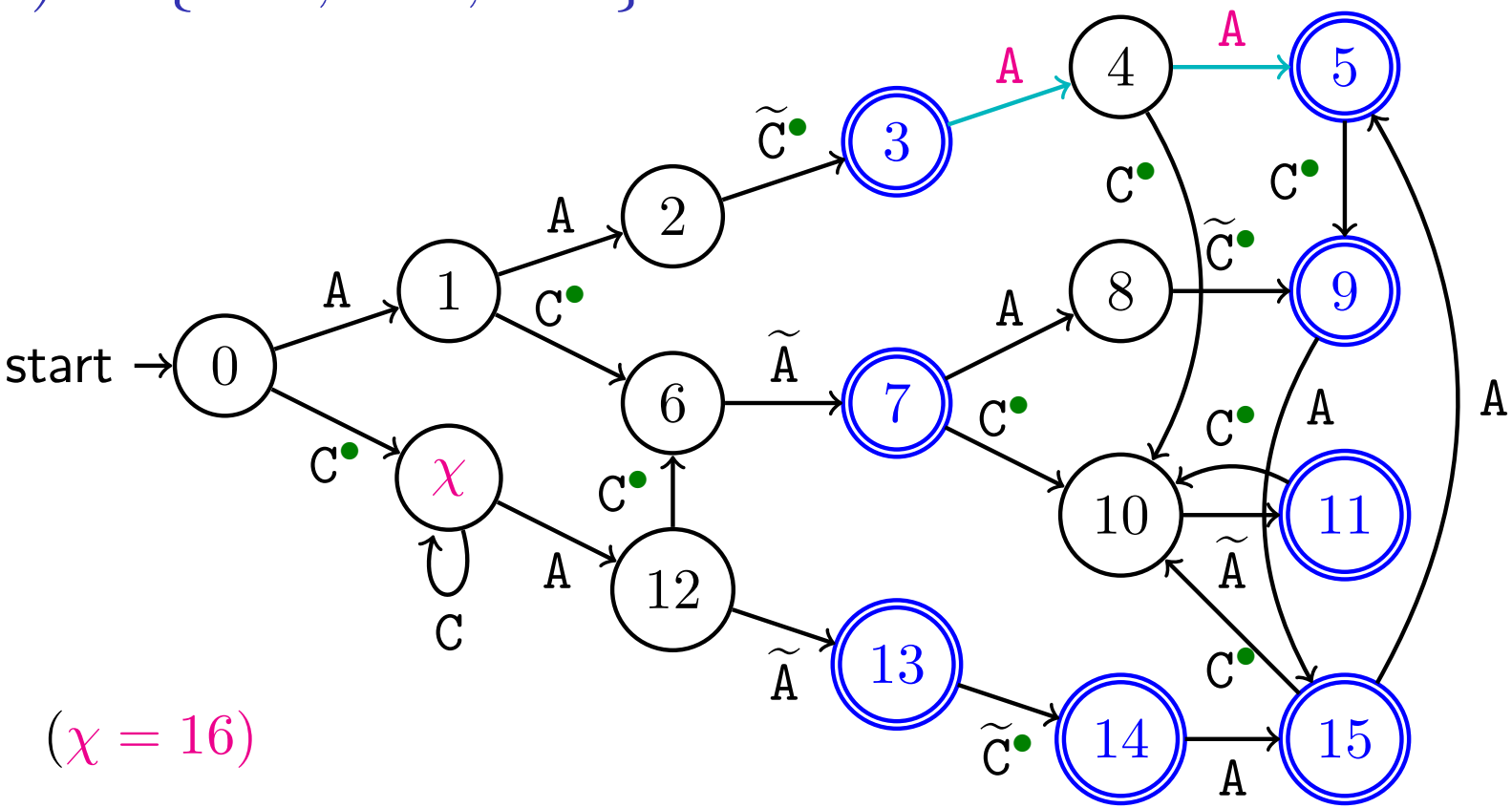
$(\chi = 16)$

- $h_{7,10}(t_{\mathtt{C}\to\mathtt{A}}) = \nu_{\mathtt{C}}$
- $h_{6,7}(t_{\mathtt{C}\to\mathtt{A}}) = \nu_{\mathtt{A}} t_{\mathtt{C}\to\mathtt{A}}$

$(\theta(7) = \mathtt{AC}^{\bullet}\mathtt{A})$

Adjacency matrix $\mathbb{H}(t) = (h_{ij}(t))$

$d(\mathtt{AAA}) = \{\mathtt{AAC}, \mathtt{ACA}, \mathtt{CAA}\}$

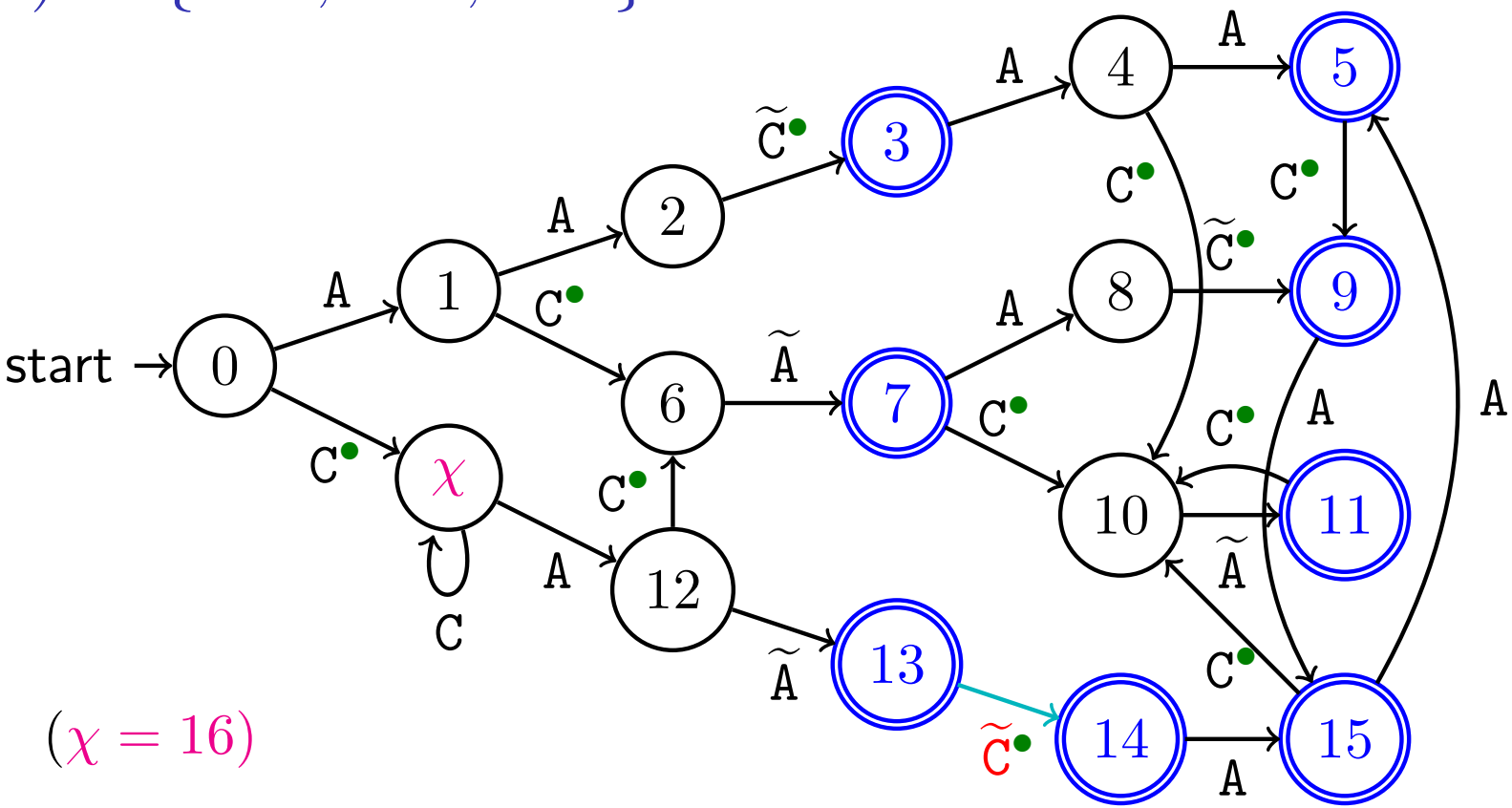


$(\chi = 16)$

- $h_{7,10}(t_{\mathtt{C}\to\mathtt{A}}) = \nu_{\mathtt{C}}$
- $h_{6,7}(t_{\mathtt{C}\to\mathtt{A}}) = \nu_{\mathtt{A}} t_{\mathtt{C}\to\mathtt{A}}$ $\quad (\theta(7) = \mathtt{AC^{\bullet}A})$
- $h_{3,4}(t_{\mathtt{C}\to\mathtt{A}}) = h_{4,5}(t_{\mathtt{C}\to\mathtt{A}}) = \nu_{\mathtt{A}}$ $\quad (\theta(5) = \mathtt{AA})$

Adjacency matrix $\mathbb{H}(t) = (h_{ij}(t))$

$d(\mathtt{AAA}) = \{\mathtt{AAC}, \mathtt{ACA}, \mathtt{CAA}\}$



$(\chi = 16)$

- $h_{7,10}(t_{\mathtt{C}\to\mathtt{A}}) = \nu_{\mathtt{C}}$
- $h_{6,7}(t_{\mathtt{C}\to\mathtt{A}}) = \nu_{\mathtt{A}} t_{\mathtt{C}\to\mathtt{A}}$ $\quad (\theta(7) = \mathtt{AC}^{\bullet}\mathtt{A})$
- $h_{3,4}(t_{\mathtt{C}\to\mathtt{A}}) = h_{4,5}(t_{\mathtt{C}\to\mathtt{A}}) = \nu_{\mathtt{A}}$ $\quad (\theta(5) = \mathtt{AA})$
- $h_{13,14}(t_{\mathtt{C}\to\mathtt{A}}) = \nu_{\mathtt{C}} t_{\mathtt{C}\to\mathtt{A}}$ $\quad (\theta(14) = \mathtt{C}^{\bullet})$

Formal definition of the adjacency matrix $\mathbb{H}(t)$

(a) $h_{ij}(t) = 0$ if there is no transition from i to j

(b) With $\delta(i, \alpha) = j$,

$$h_{i,j}(t) = \begin{cases} \nu(\alpha) \text{ if } \left| \begin{array}{l} j \notin O, \\ j \in O \text{ and } \theta(j) \text{ contains } \textbf{no putative-hit position} \end{array} \right. \\ \nu(\alpha) \times t \text{ elsewhere} \end{cases}$$

Formal definition of the adjacency matrix $\mathbb{H}(t)$

(a) $h_{ij}(t) = 0$ if there is no transition from i to j

(b) With $\delta(i, \alpha) = j$,

$$h_{i,j}(t) = \begin{cases} \nu(\alpha) \text{ if } \left| \begin{array}{l} j \notin O, \\ j \in O \text{ and } \theta(j) \text{ contains } \textbf{no putative-hit position} \end{array} \right. \\ \nu(\alpha) \times t \text{ elsewhere} \end{cases}$$

From matrix to generating function

$$\begin{aligned} F_b(z,t) &= (1,0,\ldots,0) \times \left(\mathbb{I} + z\mathbb{H}(t) + \cdots + z^n\mathbb{H}^n(t) + \ldots\right) \times \mathbf{1}^t \\ &= (1,0,\ldots,0) \times (\mathbb{I} - z\mathbb{H}(t))^{-1} \times \mathbf{1}^t. \end{aligned}$$

Entries of $(\mathbb{I} - z\mathbb{H}(t))^{-1}$ **rational functions** in z and t

Rational functions and gfun

rational function $\frac{f(z)}{g(z)}$ $\rightarrow$ `gfun[diffeqtorec]` $\rightarrow$ **recurrence** equations

recurrence equations $\rightarrow$ `gfun[rectoproc]` $\rightarrow$ **procedure** Proc(n)=$[z^n]\frac{f(z)}{g(z)}$

Rational functions, Taylor coefficient of order n and `gfun`

rational function $\frac{f(z)}{g(z)}$ $\rightarrow$ `gfun[diffeqtorec]` $\rightarrow$ **recurrence** equations

recurrence equations $\rightarrow$ `gfun[rectoproc]` $\rightarrow$ **procedure** Proc(n)=$[z^n]\frac{f(z)}{g(z)}$

$$F_b(z,t)=\frac{P(z,t)}{Q(z,t)} \quad \text{and} \quad F_b(z,1)=\sum_{n\geq 0}\widehat{f}_n^{(b)}z^n=\frac{P(z,1)}{Q(z,1)},$$

$$E(z)=\sum_n \eta_n z^n=\frac{\partial}{\partial t}F_b(z,t)\Big|_{t=1}=\frac{P_t'(z,1)}{Q(z,1)}-\frac{P(z,1)Q_t'(z,1)}{Q^2(z,1)}$$

Rational functions, Taylor coefficient of order n and `gfun`

rational function $\frac{f(z)}{g(z)}$ $\rightarrow$ `gfun[diffeqtorec]` $\rightarrow$ **recurrence** equations

recurrence equations $\rightarrow$ `gfun[rectoproc]` $\rightarrow$ **procedure** Proc(n)=$[z^n]\frac{f(z)}{g(z)}$

$$F_b(z,t)=\frac{P(z,t)}{Q(z,t)} \quad \text{and} \quad F_b(z,1)=\sum_{n\geq 0}\widehat{f}_n^{(b)}z^n=\frac{P(z,1)}{Q(z,1)},$$

$$E(z)=\sum_n \eta_n z^n=\frac{\partial}{\partial t}F_b(z,t)\Big|_{t=1}=\frac{P_t'(z,1)}{Q(z,1)}-\frac{P(z,1)Q_t'(z,1)}{Q^2(z,1)}$$

where, $P(z,t)$ and $Q(z,t)$ are **polynoms**, and,
in a **random** sequence $S_n(0)$ of length n with **no occurrence** of b,

Rational functions, Taylor coefficient of order n and `gfun`

rational function $\frac{f(z)}{g(z)}$ $\rightarrow$ `gfun[diffeqtorec]` $\rightarrow$ **recurrence** equations

recurrence equations $\rightarrow$ `gfun[rectoproc]` $\rightarrow$ **procedure** Proc(n)=$[z^n]\frac{f(z)}{g(z)}$

$$F_b(z,t)=\frac{P(z,t)}{Q(z,t)} \quad \text{and} \quad F_b(z,1)=\sum_{n\geq 0}\widehat{f}_n^{(b)}z^n=\frac{P(z,1)}{Q(z,1)},$$

$$E(z)=\sum_n \eta_n z^n=\frac{\partial}{\partial t}F_b(z,t)\Big|_{t=1}=\frac{P_t'(z,1)}{Q(z,1)}-\frac{P(z,1)Q_t'(z,1)}{Q^2(z,1)}$$

where, $P(z,t)$ and $Q(z,t)$ are **polynoms**, and,
in a **random** sequence $S_n(0)$ of length n with **no occurrence** of b,

- $\widehat{f}_n^{(b)}=\mathbf{P}(S_n(0))=\mathbf{P}(\textbf{not going into sink})$

Rational functions, Taylor coefficient of order n and `gfun`

rational function $\frac{f(z)}{g(z)}$ $\rightarrow$ `gfun[diffeqtorec]` $\rightarrow$ **recurrence** equations

recurrence equations $\rightarrow$ `gfun[rectoproc]` $\rightarrow$ **procedure** Proc(n)=$[z^n]\frac{f(z)}{g(z)}$

$$F_b(z,t) = \frac{P(z,t)}{Q(z,t)} \quad \text{and} \quad F_b(z,1) = \sum_{n\geq 0} \widehat{f}_n^{(b)} z^n = \frac{P(z,1)}{Q(z,1)},$$

$$E(z) = \sum_n \eta_n z^n = \frac{\partial}{\partial t} F_b(z,t)\Big|_{t=1} = \frac{P'_t(z,1)}{Q(z,1)} - \frac{P(z,1)Q'_t(z,1)}{Q^2(z,1)}$$

where, $P(z,t)$ and $Q(z,t)$ are **polynoms**, and,
in a **random** sequence $S_n(0)$ of length n with **no occurrence** of b,

- $\widehat{f}_n^{(b)} = \mathbf{P}(S_n(0)) = \mathbf{P}(\textbf{not going into sink})$
- η_n is the **unconditionned probability** of the expectation of the count of putative-hit positions

Rational functions, Taylor coefficient of order n and `gfun`

rational function $\frac{f(z)}{g(z)}$ $\rightarrow$ `gfun[diffeqtorec]` $\rightarrow$ **recurrence** equations

recurrence equations $\rightarrow$ `gfun[rectoproc]` $\rightarrow$ **procedure** Proc(n)=$[z^n]\frac{f(z)}{g(z)}$

$$F_b(z,t)=\frac{P(z,t)}{Q(z,t)} \quad \text{and} \quad F_b(z,1)=\sum_{n\geq 0}\widehat{f}_n^{(b)}z^n=\frac{P(z,1)}{Q(z,1)},$$

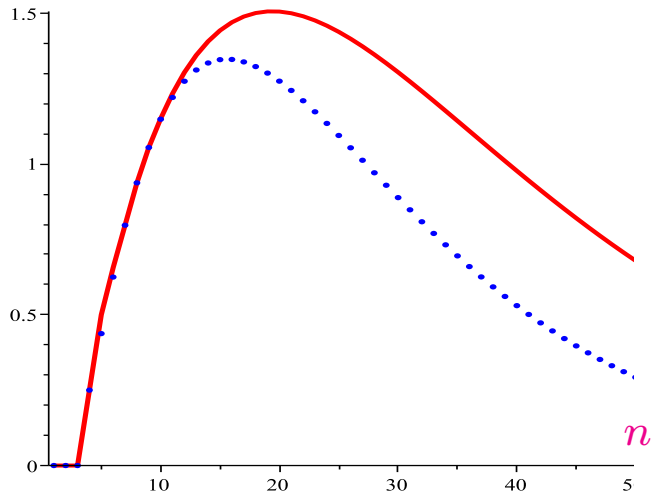
$$E(z)=\sum_n \eta_n z^n=\frac{\partial}{\partial t}F_b(z,t)\Big|_{t=1}=\frac{P'_t(z,1)}{Q(z,1)}-\frac{P(z,1)Q'_t(z,1)}{Q^2(z,1)}$$

where, $P(z,t)$ and $Q(z,t)$ are **polynoms**, and,
in a **random** sequence $S_n(0)$ of length n with **no occurrence** of b,

- $\widehat{f}_n^{(b)}=\mathbf{P}(S_n(0))=\mathbf{P}(\textbf{not going into sink})$
- η_n is the **unconditionned probability** of the expectation of the count of putative-hit positions
- **Conditionned expectation:** $\widetilde{\eta}_n=\eta_n/\widehat{f}_n^{(b)}$

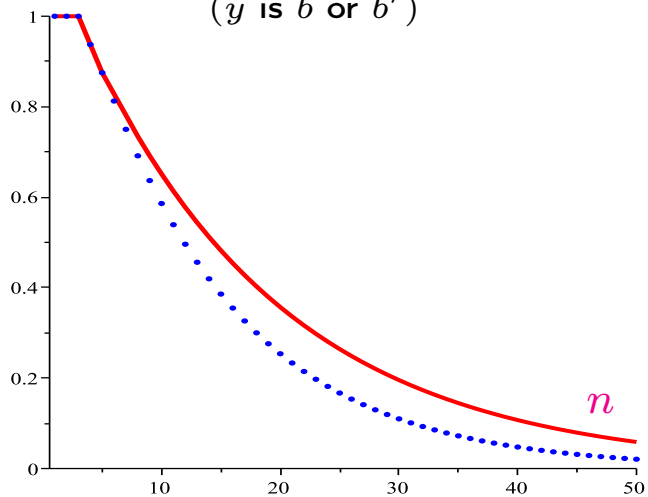
An unexpected behaviour

$\eta_n = \mathbf{E}(H_n^{(\mathtt{A}\to\mathtt{C})}) + \mathbf{E}(H_n^{(\mathtt{C}\to\mathtt{A})})$



$\widehat{f}_n^{(y)} = \mathbf{P}(|S_n(0)|_y = 0)$

(y is b or b')



$b = \mathtt{ACAC} \qquad b' = \mathtt{AACC}$

$\nu(\mathtt{A}) = \nu(\mathtt{C}) = \dfrac{1}{2}$

$\mathbf{E}(H_n^{(\mathtt{A}\to\mathtt{C})}) + \mathbf{E}(H_n^{(\mathtt{C}\to\mathtt{A})}) = \left.\dfrac{\partial F_b(z,t)}{\partial t}\right|_{t=1}$

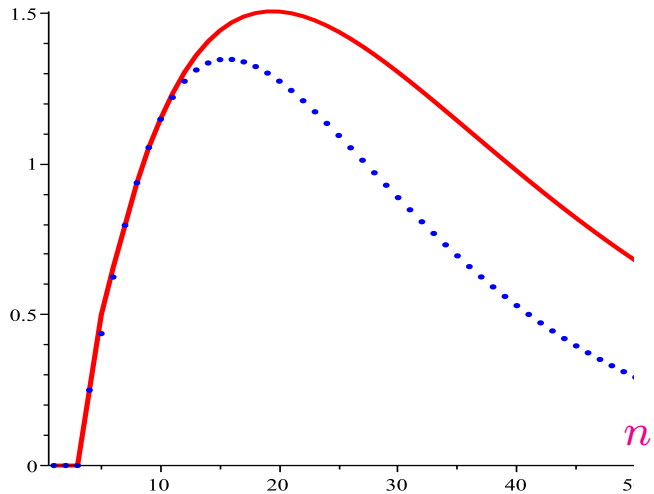
$\pi_{\mathtt{A}\to\mathtt{C}} = \pi_{\mathtt{C}\to\mathtt{A}}$

$t = t_{\mathtt{A}\to\mathtt{C}} = t_{\mathtt{C}\to\mathtt{A}}$

$\pi_{\mathtt{A}\to\mathtt{A}} = \pi_{\mathtt{C}\to\mathtt{C}}$

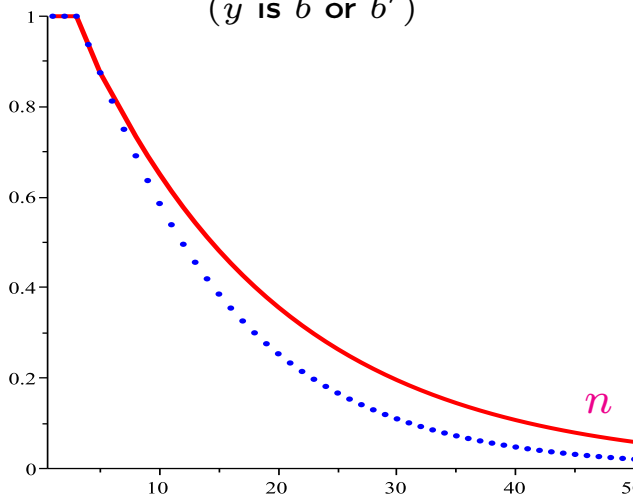
An unexpected behaviour

$\eta_n = \mathbf{E}(H_n^{(\mathtt{A}\to\mathtt{C})}) + \mathbf{E}(H_n^{(\mathtt{C}\to\mathtt{A})})$

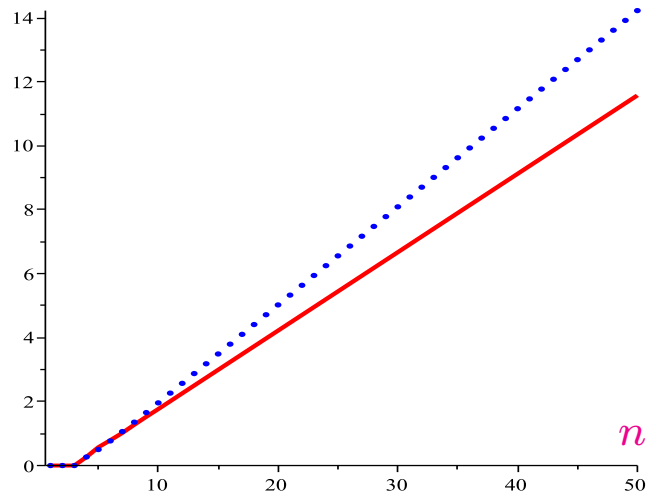


$\widehat{f}_n^{(y)} = \mathbf{P}(|S_n(0)|_y = 0)$

(y is b or b')



$\widetilde{\eta}_n = \eta_n \Big/ \widehat{f}_n^{(y)}$



$b = \mathtt{ACAC} \qquad b' = \mathtt{AACC}$

$\nu(\mathtt{A}) = \nu(\mathtt{C}) = \dfrac{1}{2}$

$\pi_{\mathtt{A}\to\mathtt{C}} = \pi_{\mathtt{C}\to\mathtt{A}}$

$\pi_{\mathtt{A}\to\mathtt{A}} = \pi_{\mathtt{C}\to\mathtt{C}}$

$$\mathbf{E}(H_n^{(\mathtt{A}\to\mathtt{C})}) + \mathbf{E}(H_n^{(\mathtt{C}\to\mathtt{A})}) = \left.\frac{\partial F_b(z,t)}{\partial t}\right|_{t=1}$$

$t = t_{\mathtt{A}\to\mathtt{C}} = t_{\mathtt{C}\to\mathtt{A}}$

A proof by singularity analysis

$$F_b(z,t) = \frac{P(z,t)}{Q(z,t)} \qquad P(z,t) \text{ and } Q(z,t) \text{ polynomials}$$

$$F_b(z,1) = \sum_{n\geq 0} \widehat{f}_n^{(b)} z^n = \frac{P(z,1)}{Q(z,1)}$$

$\widehat{f}_n^{(b)}$ probability that $S_n(0)$ has **no occurrence** of b.

$$E(z) = \sum_{n\geq 0} \mathbf{E}(H_n) z^n = \frac{P'_x(z,1)}{Q(z,1)} - \frac{P(z,1)Q'_x(z,1)}{Q^2(z,1)}$$

A proof by singularity analysis

$$F_b(z,t) = \frac{P(z,t)}{Q(z,t)} \qquad P(z,t) \text{ and } Q(z,t) \text{ polynomials}$$

$$F_b(z,1) = \sum_{n\geq 0} \widehat{f}_n^{(b)} z^n = \frac{P(z,1)}{Q(z,1)}$$

$\widehat{f}_n^{(b)}$ probability that $S_n(0)$ has **no occurrence** of b.

$$E(z) = \sum_{n\geq 0} \mathbf{E}(H_n) z^n = \frac{P'_x(z,1)}{Q(z,1)} - \frac{P(z,1)Q'_x(z,1)}{Q^2(z,1)}$$

The **dominant singularity** τ is the smallest positive solution of $Q(z,1) = 0$. Use suitable Cauchy integrals

$$\begin{cases} \widehat{f}_n^{(b)} = \psi \times \tau^{-(n-1)} \left(1 + \mathcal{O}\left(B^n\right)\right), \qquad (B<1) \\ \mathbf{E}(H_n) = [z^n]E(z) = \tau^{-n}(\phi_1 \times n + \phi_2) \times \left(1 + \mathcal{O}\left(B^n\right)\right) \end{cases}$$

$$\Longrightarrow \mathbf{E}(\widetilde{H}_n) = \frac{\mathbf{E}(H_n)}{\widehat{f}_n^{(b)}} = (c_1 \times n + c_2) \times \left(1 + \mathcal{O}\left(B^n\right)\right), \ (B<1).$$

General case

Compute $F_b(z, t_{\mathtt{A}\to\mathtt{C}}, t_{\mathtt{A}\to\mathtt{G}}, t_{\mathtt{A}\to\mathtt{T}}, t_{\mathtt{C}\to\mathtt{A}}, \dots, t_{\mathtt{T}\to\mathtt{C}}, t_{\mathtt{T}\to\mathtt{G}})$

$$\widehat{f}_n^{(b)} = [z^n] F_b(z, 1, 1, \dots, 1, 1)$$

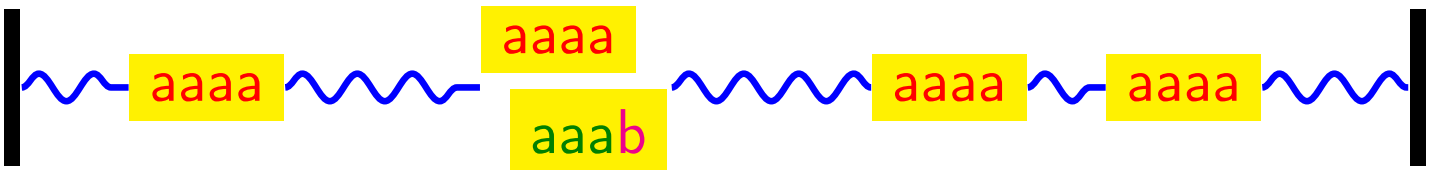
$$\mathfrak{P}_n \approx [z^n] \sum_{\alpha\neq\beta\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \left.\frac{\partial F_b(z, 1, \dots, 1, \pi_{\alpha\to\beta} t_{\alpha\to\beta}, 1, \dots)}{\partial t_{\alpha\to\beta}}\right|_{t_{\alpha\to\beta}=1} \Big/ \widehat{f}_n^{(b)}$$

- The **dominant singularities** of **all the terms of the sum** are **equal** to the **dominant singularity** of $F_b(z, 1, 1, \dots, 1, 1)$
- $\mathfrak{P}_n$ behaves **quasi-linearly**

II - Formal Languages Approach

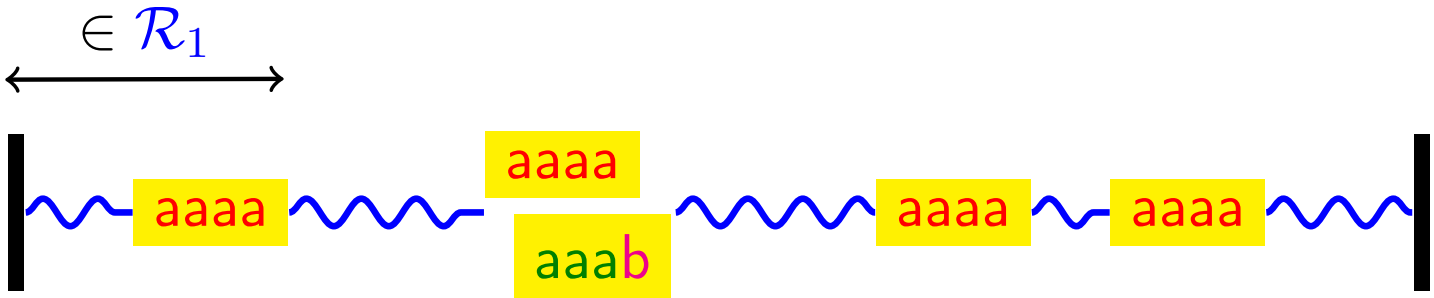
Guibas-Odlyzko decomposition - occurrences of a word u

$U = (\texttt{aaaa}, \texttt{aaab})$ $\quad \begin{cases} u_1 = \texttt{aaaa} \\ u_2 = \texttt{aaab} \end{cases}$



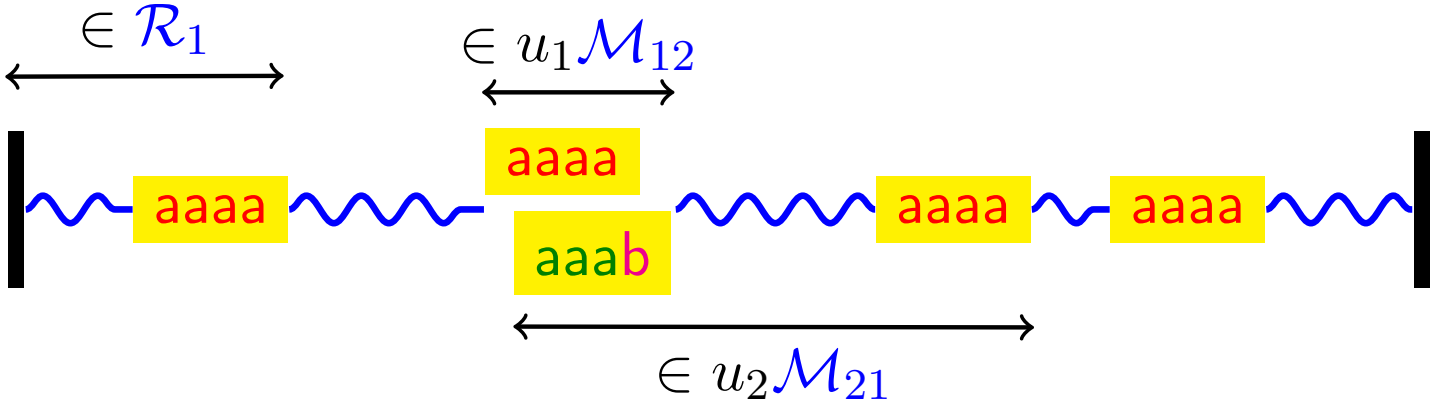
Guibas-Odlyzko decomposition - occurrences of a word u

$U = (\texttt{aaaa}, \texttt{aaab})$ $\quad \begin{cases} u_1 = \texttt{aaaa} \\ u_2 = \texttt{aaab} \end{cases}$



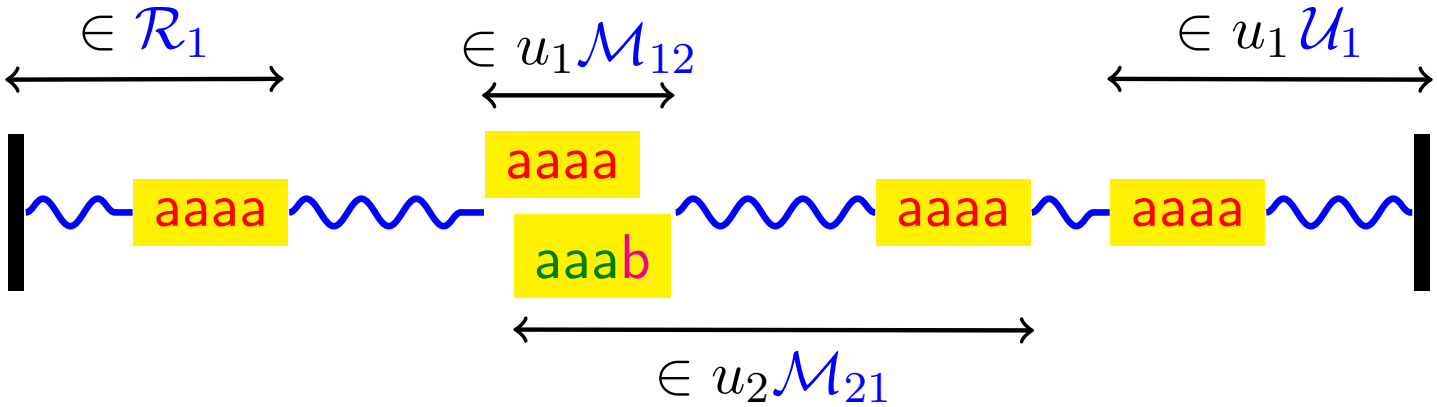
Guibas-Odlyzko decomposition - occurrences of a word u

$U = (\texttt{aaaa}, \texttt{aaab})$ $\quad \begin{cases} u_1 = \texttt{aaaa} \\ u_2 = \texttt{aaab} \end{cases}$



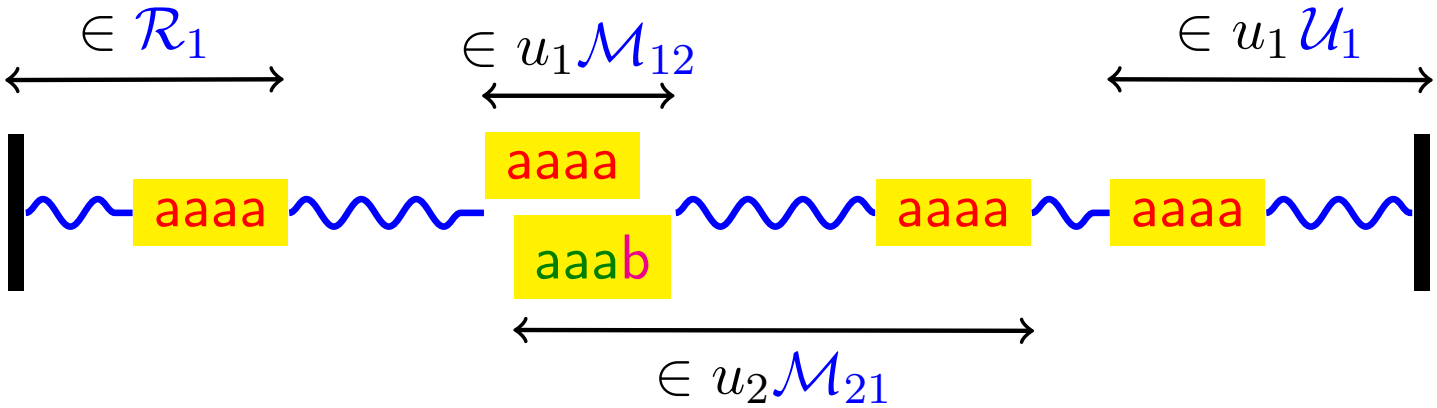
Guibas-Odlyzko decomposition - occurrences of a word u

$U = (\texttt{aaaa}, \texttt{aaab})$ $\quad \begin{cases} u_1 = \texttt{aaaa} \\ u_2 = \texttt{aaab} \end{cases}$



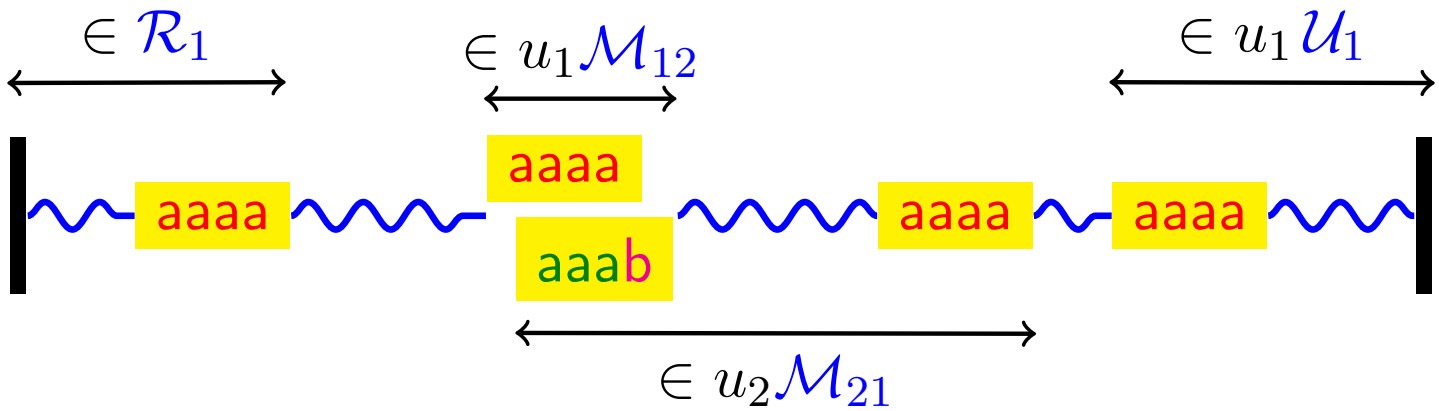
Guibas-Odlyzko decomposition - occurrences of a word u

$U = (\texttt{aaaa}, \texttt{aaab})$ $\quad \begin{cases} u_1 = \texttt{aaaa} \\ u_2 = \texttt{aaab} \end{cases}$



Guibas-Odlyzko decomposition - occurrences of a word u

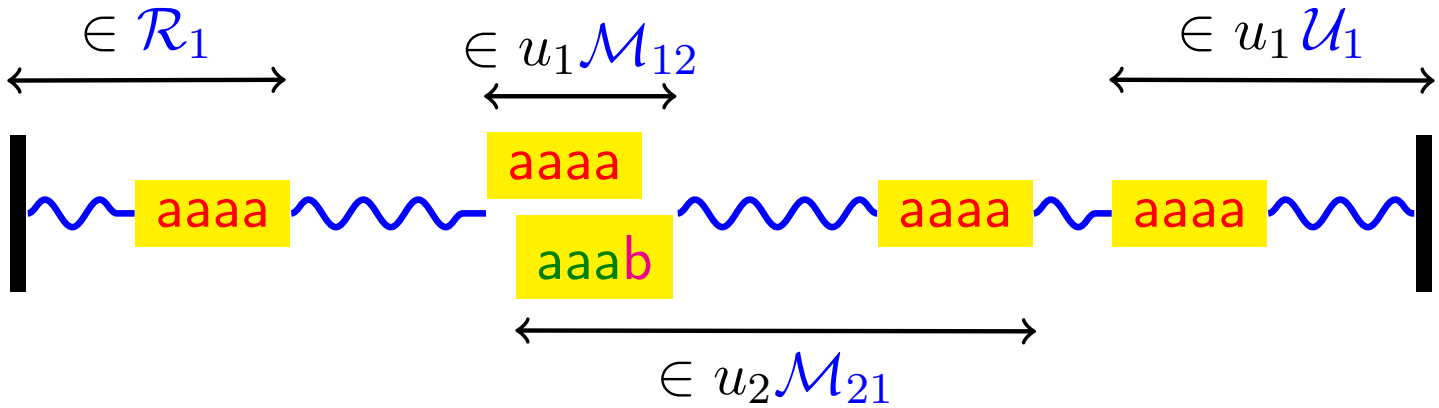
$U = (\text{aaaa}, \text{aaab})$ $\quad \begin{cases} u_1 = \text{aaaa} \\ u_2 = \text{aaab} \end{cases}$



- The "Right" language $\mathcal{R}_i$ associated to the word u_i is the set of words $\mathcal{R}_i = \{r \,|\, r = e \cdot u_i$ and there is no $v \in U$ such that $r = xvy$ with $|y| > 0\}$.
- The "Minimal" language $\mathcal{M}_{ij}$ leading from a word u_i to a word u_j is the set of words $\mathcal{M}_{ij} = \{m \,|\, u_i \cdot m = e \cdot u_j$ and there is no $v \in U$ such that $u_i \cdot m = xvy$ with $|x| > 0, |y| > 0\}$.
- The "Ultimate" language $\mathcal{U}_i$ of words following the last occurrence of the word u_i (such that this occurrence is the last occurrence of U in the text) is the set of words $\mathcal{U}_i = \{u \,|\,$ there is no $v \in U$ such that $u_i \cdot u = xvy$ with $|x| > 0\}$.
- The "Not" language $\mathcal{N}$ is the set of words with no occurrences of U, $\mathcal{N} = \{n \mid$ there is no $v \in U$ such that $n = xvy\}$.

Guibas-Odlyzko decomposition - occurrences of a word u

$U = (\texttt{aaaa}, \texttt{aaab})$ $\quad \begin{cases} u_1 = \texttt{aaaa} \\ u_2 = \texttt{aaab} \end{cases}$



$$F(z, x_1, x_2) = \mathcal{N}(z) + (\mathcal{R}_1(z)x_1, \mathcal{R}_2(z)x_2) \begin{pmatrix} \mathcal{M}_{11}(z)x_1 & \mathcal{M}_{12}(z)x_2 \\ \mathcal{M}_{21}(z)x_1 & \mathcal{M}_{22}(z)x_2 \end{pmatrix} \begin{pmatrix} \mathcal{U}_1(z) \\ \mathcal{U}_2(z) \end{pmatrix}$$

Computing the languages

- $\mathcal{C}_{u_1,u_2}$ **correlation set** of two words u_1 and u_2
  $\mathcal{C}_{u_1,u_2} = \{\ e \mid \exists\ e' \in \mathcal{A}^+,\ u_1 e = e' u_2$ with $|e| < |u_2|\ \}$.
- $\mathcal{C}_u = \mathcal{C}_{u,u}$ **autocorrelation set** $\quad (\epsilon \in \mathcal{C}_u)$

Computing the languages

- $\mathcal{C}_{u_1,u_2}$ **correlation set** of two words u_1 and u_2
  $\mathcal{C}_{u_1,u_2} = \{\ e \mid \exists\ e' \in \mathcal{A}^+,\ u_1 e = e' u_2 \text{ with } |e| < |u_2|\ \}.$
- $\mathcal{C}_u = \mathcal{C}_{u,u}$ **autocorrelation set** $\quad (\epsilon \in \mathcal{C}_u)$

- **Régnier-Szpankowski Equations**

$$\bigcup_{k\geq 1} \left(\mathbb{M}^k\right)_{i,j} = \mathcal{A}^\star \cdot u_j + \mathcal{C}_{ij} - \delta_{ij}\epsilon, \quad \mathcal{U}_i \cdot \mathcal{A} = \bigcup_j \mathcal{M}_{ij} + \mathcal{U}_i - \epsilon,$$

$$\mathcal{A}\cdot\mathcal{R}_j - (\mathcal{R}_j - u_j) = \bigcup_i u_i \mathcal{M}_{ij}, \qquad \mathcal{N}\cdot u_j = \mathcal{R}_j + \bigcup_i \mathcal{R}_i \left(\mathcal{C}_{ij} - \delta_{ij}\epsilon\right),$$

- **Automaton Computation**

$$\mathcal{R}_i = \bigotimes_{1\leq r\leq k} \overline{\mathcal{A}^\star u_r \mathcal{A}^\star} \cdot \mathcal{A} \bigotimes \mathcal{A}^\star u_i$$

$$u_i \mathcal{M}_{ij} = u_i \mathcal{A}^\star \bigotimes \mathcal{A}^\star u_j \bigotimes_{1\leq r\leq k} \mathcal{A}\overline{\mathcal{A}^\star u_r \mathcal{A}^\star}\mathcal{A}$$

$$u_j \mathcal{U}_j = u_j \mathcal{A}^\star \bigotimes_{1\leq r\leq k} \mathcal{A} \cdot \overline{\mathcal{A}^\star u_r \mathcal{A}^\star}$$

$$\mathcal{N} = \mathrm{NOT}\left(\bigotimes_{1\leq r\leq k} \mathcal{A}^\star u_r \mathcal{A}^\star\right)$$

Constrained Guibas-Odlyzko languages

Example: $b = \mathtt{AA}, \qquad d_\ell(b) = (\mathtt{AC}, \mathtt{CA})$

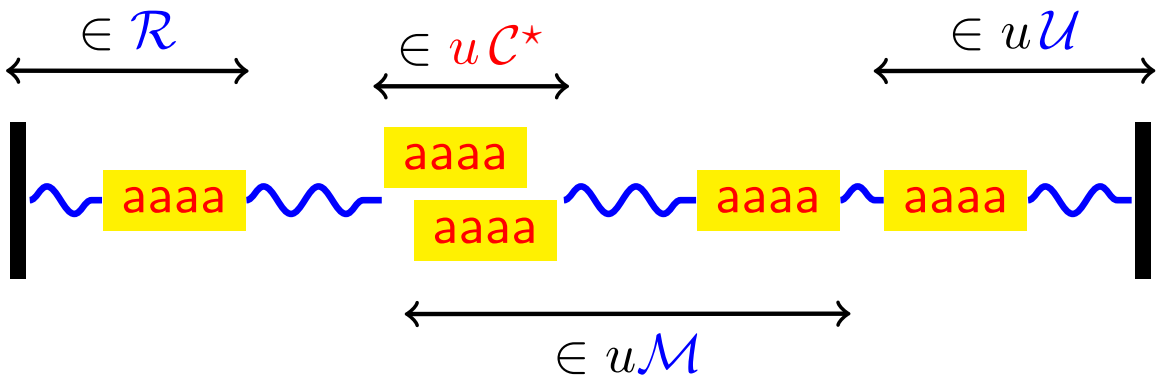
- We need **avoiding** $\mathtt{AA}$ in $S(0)$ and therefore **in the Right**, **Minimal** and **Ultimate** languages
- **Build** the **Régnier-Szpankowski languages** **for the pattern** $(\mathtt{AC}, \mathtt{CA}, \mathtt{AA})$

$$\mathcal{L} = \mathcal{N} + (\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3) \begin{pmatrix} \mathcal{M}_{11} & \mathcal{M}_{12} & \mathcal{M}_{13} \\ \mathcal{M}_{21} & \mathcal{M}_{22} & \mathcal{M}_{23} \\ \mathcal{M}_{31} & \mathcal{M}_{32} & \mathcal{M}_{33} \end{pmatrix} \begin{pmatrix} \mathcal{U}_1 \\ \mathcal{U}_2 \\ \mathcal{U}_3 \end{pmatrix}$$

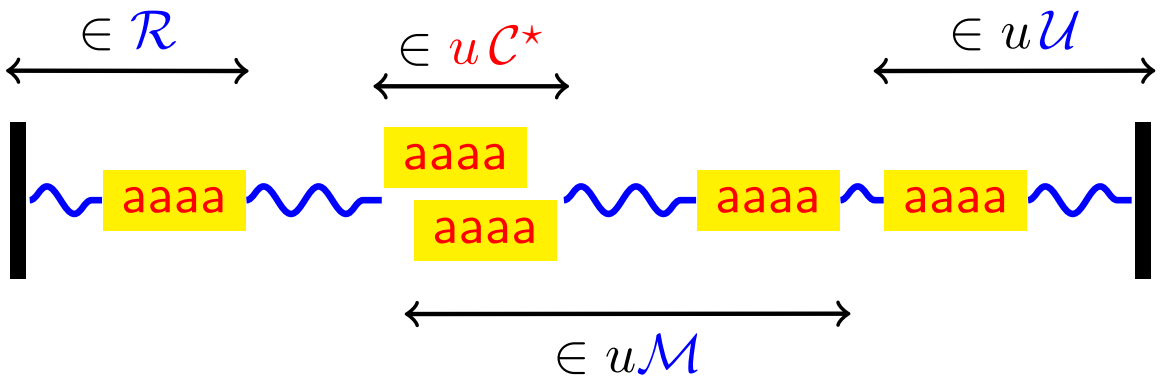
$$\widehat{\mathcal{L}} = \mathcal{N} + (\mathcal{R}_1, \mathcal{R}_2) \begin{pmatrix} \mathcal{M}_{11} & \mathcal{M}_{12} \\ \mathcal{M}_{21} & \mathcal{M}_{22} \end{pmatrix} \begin{pmatrix} \mathcal{U}_1 \\ \mathcal{U}_2 \end{pmatrix}$$

Notations: write $\widehat{\mathcal{N}}, \widehat{\mathcal{R}_i}, \widehat{\mathcal{M}_{ij}}, \widehat{\mathcal{U}_j}$ for **constrained languages**

Clump Analysis (Bassino-Clément-Fayolle-P.N. 2008)



Clump Analysis (Bassino-Clément-Fayolle-P.N. 2008)

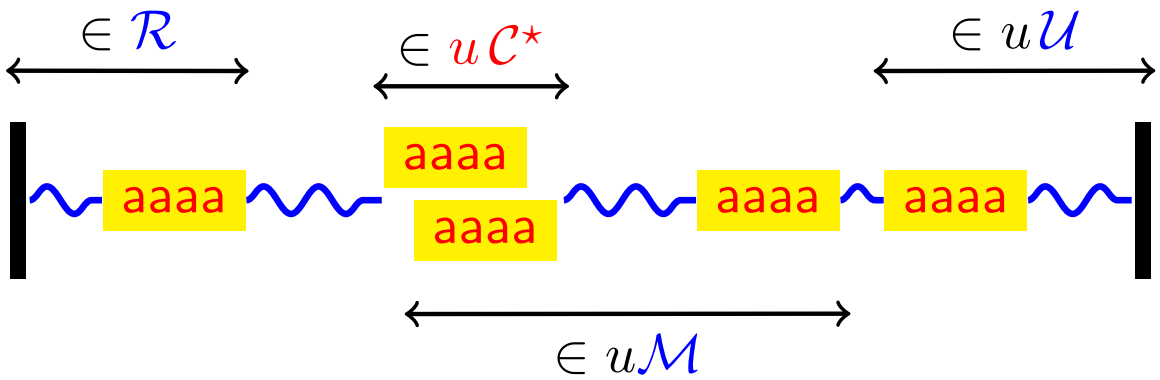


- *residual* language $\mathcal{D} = \mathcal{L}.u^- : \quad \mathcal{D} = \{h,\ h \cdot u \in \mathcal{L}\}$
- $\mathcal{L}_2 - \mathcal{L}_1 = \mathcal{L}_2 \setminus \mathcal{L}_1 = \{h;\ h \in \mathcal{L}_2, h \notin \mathcal{L}_1\}$

Combinatorial decomposition (one word)

$$\mathcal{A}^\star = \mathcal{N} + \mathcal{R}u^- u\,\mathcal{C}^\star \Big((\mathcal{M} - \mathcal{K})u^- u\,\mathcal{C}^\star \Big)^\star \mathcal{U}$$

Clump Analysis (Bassino-Clément-Fayolle-P.N. 2008)



- *residual* language $\mathcal{D} = \mathcal{L}.u^-$: $\mathcal{D} = \{h,\ h \cdot u \in \mathcal{L}\}$
- $\mathcal{L}_2 - \mathcal{L}_1 = \mathcal{L}_2 \setminus \mathcal{L}_1 = \{h;\ h \in \mathcal{L}_2, h \notin \mathcal{L}_1\}$

Combinatorial decomposition (one word)

$$\mathcal{A}^\star = \mathcal{N} + \mathcal{R}u^- u\,\mathcal{C}^\star \Big((\mathcal{M} - \mathcal{K}) u^- u\,\mathcal{C}^\star \Big)^\star \mathcal{U}$$

$$= \mathcal{N} + \mathcal{R}u^- u\,\mathcal{K}^\star \Big((\mathcal{M} - \mathcal{K}) u^- u\,\mathcal{K}^\star \Big)^\star \mathcal{U}$$

$$= \mathcal{N} + \mathcal{R}u^- \mathbf{S} \Big((\mathcal{M} - \mathcal{K}) u^- \mathbf{S} \Big)^\star \mathcal{U}$$

Clumps: $\mathbf{S} = u\,\mathcal{C}^\star = u\,\mathcal{K}^\star$

Some combinatorial properties

$u = aaaaa$
$\mathcal{C} - \{\epsilon\} = \{a, aa, aaa, aaaa\}$
$\mathcal{K} = \{a\}$
$\mathcal{M} = \{a, b(b + ab + aab + aaab + aaaab)^\star aaaaa\}$

Properties

- $\mathcal{K} \subset \mathcal{M}$
- $\mathcal{M} - \mathcal{K} = \mathcal{L}u$

Lemma.
Let $\mathcal{C}_\circ = \mathcal{C} - \{\epsilon\}$ be the strict autocorrelation set of a word u

- the Prefix code $\mathcal{K} = \mathcal{C}_\circ - \mathcal{C}_\circ\mathcal{A}^+$ generates **unambiguously** $\mathcal{C}^+ - \{\epsilon\}$, which implies that $\mathcal{K}^\star = {\mathcal{C}_\circ}^\star$

Clumps of reduced sets of words

Minimal Correlation Language: $\mathcal{K}_{ij} = \left(\mathcal{C}_{ij} - \mathcal{C}_{ij}\mathcal{A}^{+}\right) \bigcap \mathcal{M}_{ij}$

Lemma: $\mathcal{M}_{ij} - \mathcal{K}_{ij} = \mathcal{L}v_j$

Decomposition of a text by clumps:

$$\mathbb{K} = \begin{pmatrix} \mathcal{K}_{11} & \mathcal{K}_{12} \\ \mathcal{K}_{21} & \mathcal{K}_{22} \end{pmatrix}, \qquad \mathbb{S} = \mathbb{K}^{\star} \qquad \mathbb{G} = \begin{pmatrix} v_1\mathbb{S}_{11} & v_1\mathbb{S}_{12} \\ v_2\mathbb{S}_{21} & v_2\mathbb{S}_{22} \end{pmatrix}$$

$$\mathcal{A}^{\star} = \mathcal{N} + (\mathcal{R}_1 v_1^{-1}, \mathcal{R}_2 v_2^{-1})\,\mathbb{G}\left(\left((\mathcal{M}_{ij} - \mathcal{K}_{ij})v_j^{-1}\right)\mathbb{G}\right)^{\star}\begin{pmatrix} \mathcal{U}_1 \\ \mathcal{U}_2 \end{pmatrix}$$

Constrained clumps

- **finite** code languages $\mathcal{K}_{ij}$ **easy to compute**
- we must however **avoid** the **forbidden word** b while **extending clumps**
- $v_i,\ v_j \in d_\ell(b) \quad \rightsquigarrow \quad \widehat{\mathcal{K}}_{ij} = \{h \in \mathcal{K}_{ij}; \qquad |v_i.h|_b = 0\}$

  sets $\mathcal{K}_{ij}$ finite $\quad \Longrightarrow \quad$ computation of $\widehat{\mathcal{K}}_{ij}$ by string-matching

Constrained clumps

- **finite** code languages $\mathcal{K}_{ij}$ **easy to compute**
- we must however **avoid** the **forbidden word** b while **extending clumps**
- $v_i,\ v_j \in d_\ell(b) \quad \rightsquigarrow \quad \widehat{\mathcal{K}}_{ij} = \{h \in \mathcal{K}_{ij}; \qquad |v_i.h|_b = 0\}$

  sets $\mathcal{K}_{ij}$ finite $\quad \Longrightarrow \quad$ computation of $\widehat{\mathcal{K}}_{ij}$ by string-matching

- **Decomposition by constrained clumps**

$$\widehat{\mathcal{A}_b^\star} = \widehat{\mathcal{N}} + (\widehat{\mathcal{R}}_1 v_1^-, \ldots, \widehat{\mathcal{R}}_r v_r^-)\ \widehat{\mathbb{G}}\Big((\widehat{\mathbb{M}} - \widehat{\mathbb{K}})^- \widehat{\mathbb{G}}\Big)^\star \begin{pmatrix} \widehat{\mathcal{U}}_1 \\ \vdots \\ \widehat{\mathcal{U}}_r \end{pmatrix}$$

$$\text{with} \begin{cases} \widehat{\mathbb{K}} = (\widehat{\mathcal{K}}_{ij}), \\ \widehat{\mathbb{S}} = \widehat{\mathbb{K}}^\star, \\ \widehat{\mathbb{G}} = \left(v_i \widehat{\mathbb{S}}_{ij}\right) \end{cases}$$

Generating function of the number of putative-hit positions

- $v_i(z,t) = \nu(v_i) t z^{|v_i|}$ for each $v_i \in d(b)$.
- for each $\widehat{\mathcal{K}}_{ij}$, we can compute by string matching the number of putative-hit positions in each word of $v_i.\widehat{\mathcal{K}}_{ij}$.

$$\widehat{\mathcal{K}}_{ij}(z,t) = \sum_{w \in \widehat{\mathcal{K}}_{ij}} \nu(w) t^{\text{put-hit-pos}(v_i.w)-1} z^{|w|},$$

Generating function of the number of putative-hit positions

- $v_i(z,t) = \nu(v_i) t z^{|v_i|}$ for each $v_i \in d(b)$.
- for each $\widehat{\mathcal{K}}_{ij}$, we can compute by string matching the number of putative-hit positions in each word of $v_i.\widehat{\mathcal{K}}_{ij}$.

$$\widehat{\mathcal{K}}_{ij}(z,t) = \sum_{w \in \widehat{\mathcal{K}}_{ij}} \nu(w) t^{\text{put-hit-pos}(v_i.w)-1} z^{|w|},$$

$$\widehat{\mathbb{K}}(z,t) = \left(\widehat{\mathcal{K}}_{ij}(z,t)\right), \quad \widehat{\mathbb{S}}(z,t) = \left(\mathbb{I} - \widehat{\mathbb{K}}(z,t)\right)^{-1},$$

$$\widehat{\mathbb{G}}(z,t) = \left(v_i(z,t)\widehat{\mathbb{S}}_{ij}(z,t)\right).$$

Generating function of the number of putative-hit positions

- $v_i(z,t) = \nu(v_i) t z^{|v_i|}$ for each $v_i \in d(b)$.
- for each $\widehat{\mathcal{K}}_{ij}$, we can compute by string matching the number of putative-hit positions in each word of $v_i.\widehat{\mathcal{K}}_{ij}$.

$$\widehat{\mathcal{K}}_{ij}(z,t) = \sum_{w \in \widehat{\mathcal{K}}_{ij}} \nu(w) t^{\text{put-hit-pos}(v_i.w)-1} z^{|w|},$$

$$\widehat{\mathbb{K}}(z,t) = \Big(\widehat{\mathcal{K}}_{ij}(z,t)\Big), \quad \widehat{\mathbb{S}}(z,t) = \Big(\mathbb{I} - \widehat{\mathbb{K}}(z,t)\Big)^{-1},$$
$$\widehat{\mathbb{G}}(z,t) = \Big(v_i(z,t)\widehat{\mathbb{S}}_{ij}(z,t)\Big).$$

$$F_b(z,t) = \widehat{\mathcal{A}^\star_b}(z,t)$$
$$= \widehat{\mathcal{N}}(z) + (\widehat{\mathcal{R}}_1 v_1^-(z), \ldots, \widehat{\mathcal{R}}_r v_r^-(z))\ \widehat{\mathbb{G}}(z,t)\Big((\widehat{\mathbb{M}} - \widehat{\mathbb{K}})^-(z)\widehat{\mathbb{G}}(z,t)\Big)^\star \begin{pmatrix} \widehat{\mathcal{U}}_1(z) \\ \vdots \\ \widehat{\mathcal{U}}_r(z) \end{pmatrix}$$

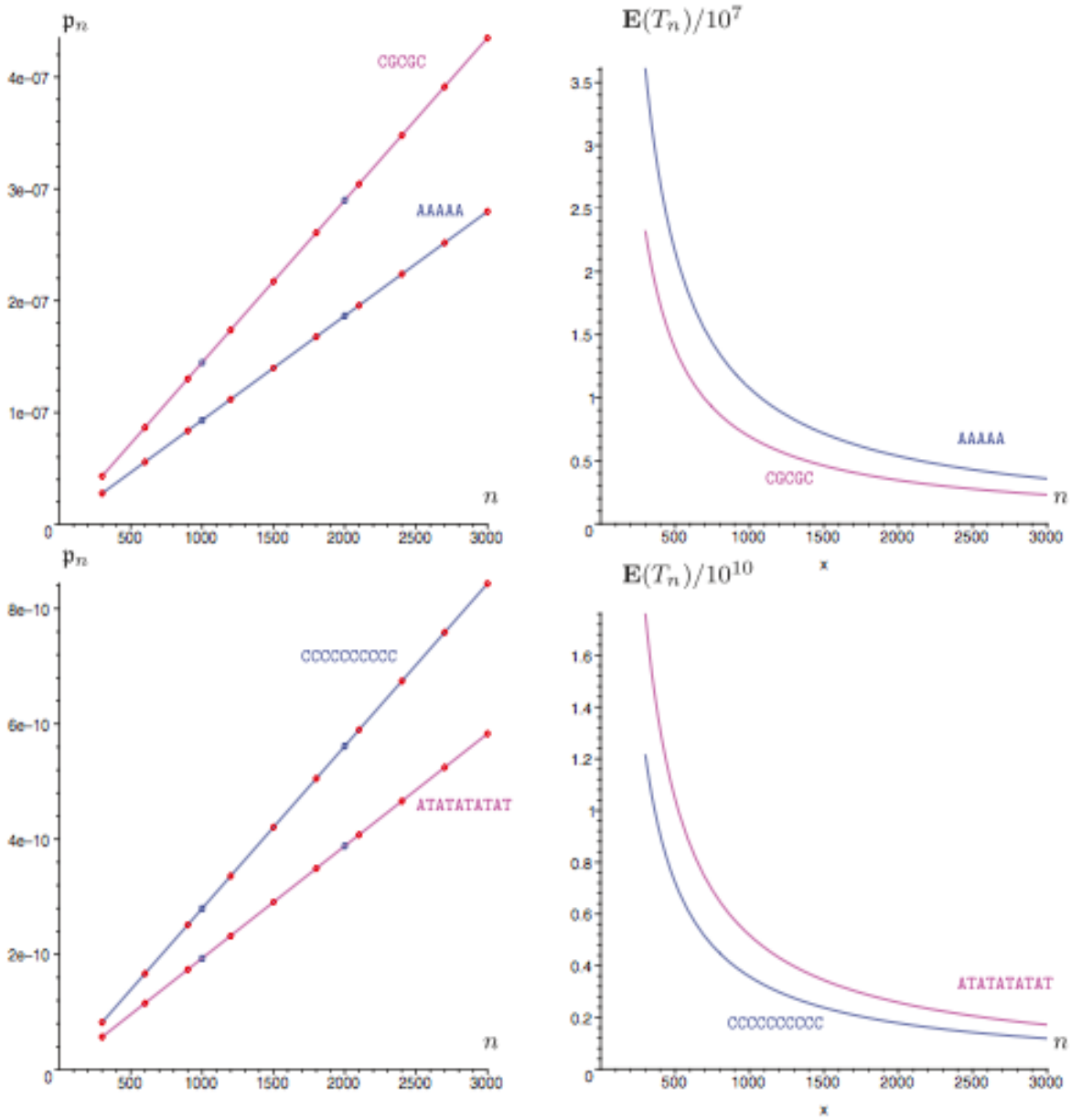


FIG. 5. Plots of the probability p_n **(left)** and of the expected waiting time $E(T_n)$ **(right)**. **(Top)** b = AAAAA (blue) and b' = CGCGC (magenta). **(Botton)** b = CCCCCC CCCC (blue) and b' = ATATATA TAT (magenta). In the linear plots of the probability, the anchors values for $n = 1000$ and $n = 2000$ (computed by automata) are represented by boxes; the straight lines are the straight lines going through the corresponding points and the circles are test values also computed by automata. The fit is perfect as expected from singularity analysis.

(from Behrens-Nicaud-P.N., JCB 19,5, 2012)