MOTIFS DISTRIBUTION

IN DNA SEQUENCES

Stéphane ROBIN
robin@inapg.inra.fr

UMR INA-PG / INRA
Biométrie & Intelligence Artificielle

Lyon, October the 1st, 02

Contents

Biological interest of motif statistics

Vocabulary

Letter = nucleotide $\in \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$ (= base pair)

Word = short, exact sequence of letters:

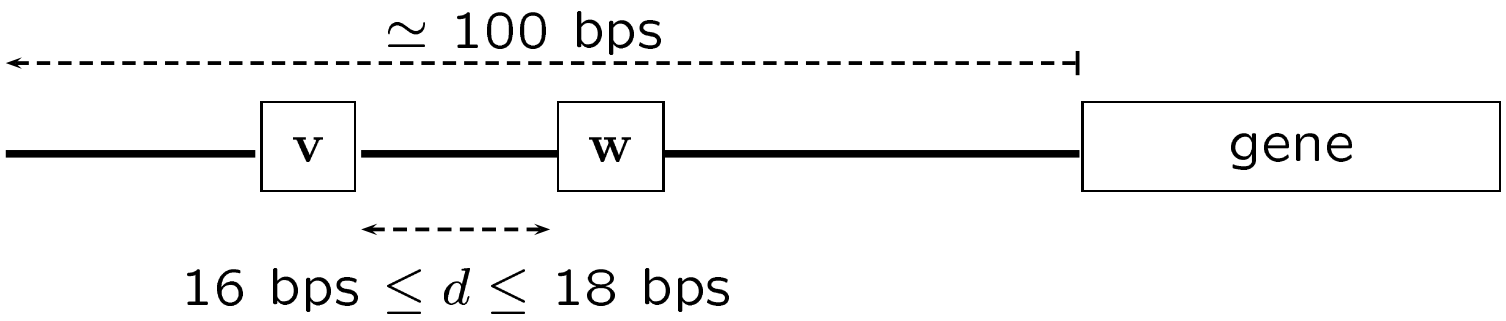
$$\mathbf{w} = \mathtt{gctggtgg}$$

Motif = set of words:

$$\mathbf{m} = \{\mathtt{gNtggtgg}\} = \{ \begin{array}{l} \mathbf{w}_1 = \mathtt{gatggtgg}, \\ \mathbf{w}_2 = \mathtt{gctggtgg}, \\ \mathbf{w}_3 = \mathtt{ggtggtgg}, \\ \mathbf{w}_4 = \mathtt{gttggtgg} \quad \} \end{array}$$

Three examples

Ex 1: Promoter motifs = structured motifs where polymerase binds to DNA



Which structured motifs are **unexpectedly frequent** in upstream regions of the genes of a given species?

Ex 2: CHI motifs in bacterial genomes

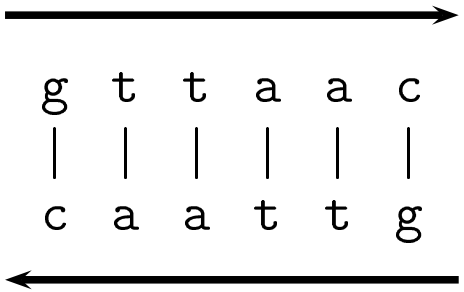
Crossover Hot-spot Initiator: **defense function** of the genome against the degradation activity of an enzyme

Exists in *E. coli* = `gctggtgg` and *H. influenza* = `gNtggtgg`

Is this motif **unexpectedly frequent** in some regions of the genome?

If so, these regions may contain crucial functions

Ex 3: Palindromes = self-complementary words



Palindromes of length 6 are restriction sites (i.e. **frailty sites**) of the genome of *E. coli*

If they are **especially avoided** in some regions, these regions may be of major importance for the organism

A model: what for ?

Model = Reference

To be able to decide if something is **unexpected**, one first need to know **what to expect**

- The model must be **well fitted** to the data to avoid artifacts and non-interesting discoveries
- but **not too well** fitted to let unexpected events happen

Interesting cases occur when the model is (very) **wrong**

The choice of the model depends on the problem

Overlapping structure of the word

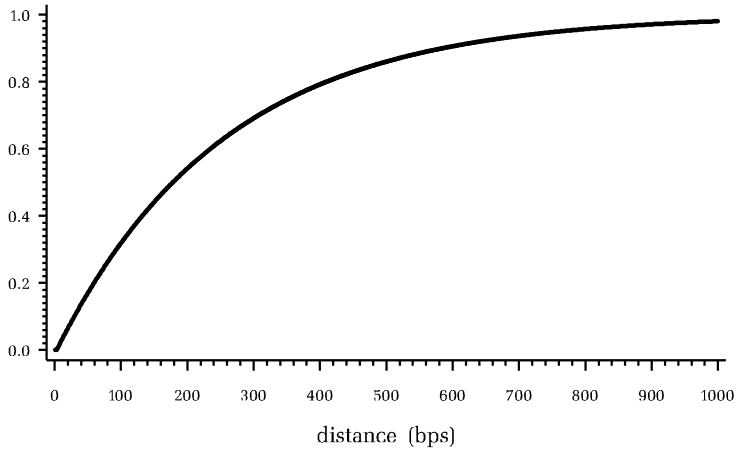
Some words can **overlap themselves** . . .

1 letter :								**g**	g	t	g	g	t	g	g
w		**g**	**g**	**t**	**g**	**g**	**t**	**g**	**g**						
5 letters :					**g**	**g**	**t**	**g**	**g**	t	g	g			
2 letters :								**g**	**g**	t	g	g	t	g	g

Conway (Gardner, 74); Guibas & Odlyzko, 81

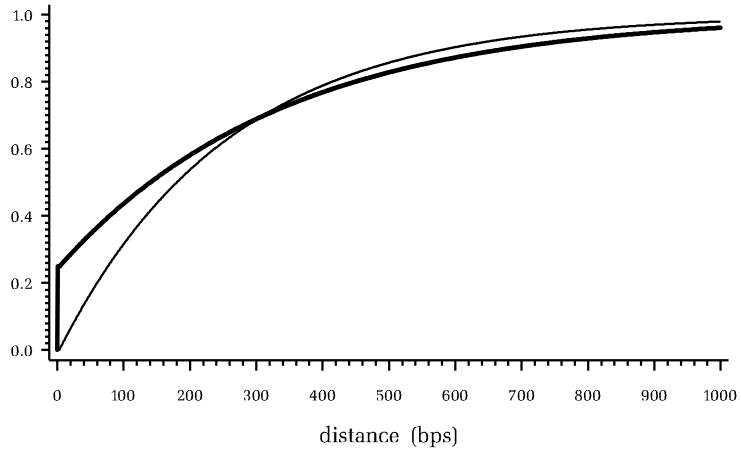
. . . and tend therefore to occur in **clumps**

$\mathbf{w} = (\mathtt{gatc})$



M00 : $\mathbb{E}(Y) = 256$ bps
$\mathbb{V}(Y) = (256.2 \text{ bps})^2$

$\mathbf{w} = (\mathtt{aaaa})$

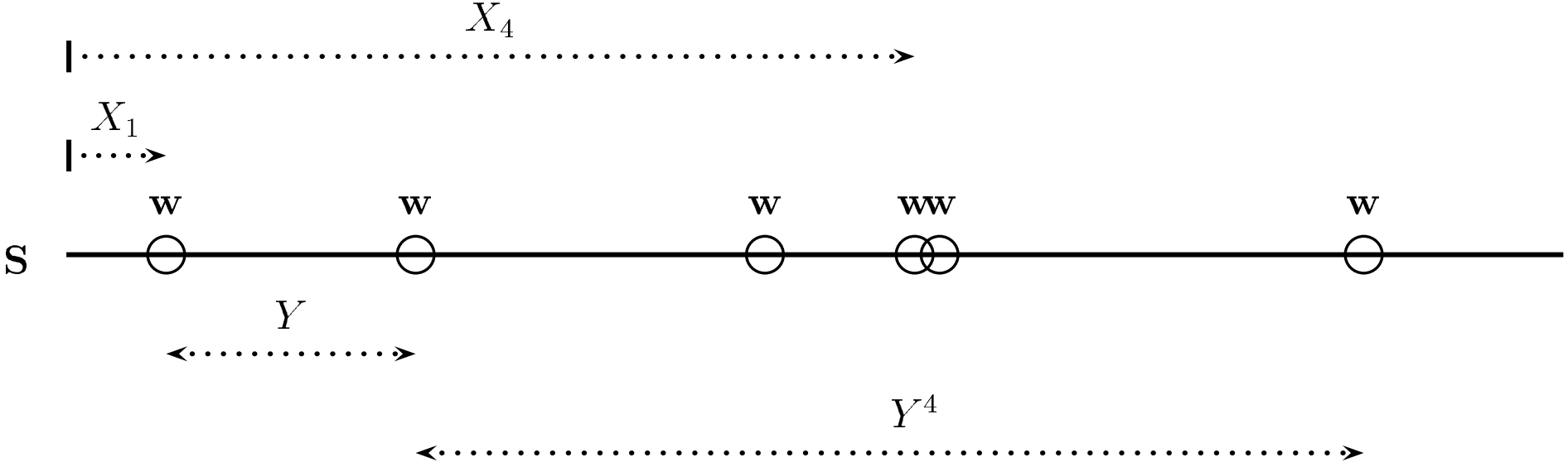


M00 : $\mathbb{E}(Y) = 256$ bps
$\mathbb{V}(Y) = (326.7 \text{ bps})^2$

Statistical issues

- **Probability for a motif to occur** in a sequence
  $\rightarrow$ promoter motifs
- Distribution of the **number** of occurrences (*Prum & al, Schbath, 95; Nicodème & al, 99; Regnier, 00; Nuel, 01*)
- **Distribution** of the occurrences along the sequence
  $\rightarrow$ CHI motifs, palindromes

Positions, distances, counts



Motifs occurrences in Markov chains

Markov chains = Discrete modeling

$\mathbf{S} = (S_1, \ldots, S_\ell)$ is an homogeneous stationary Markov chain

- of **order** m (Mm model)
- with **transition probabilities** $\pi(s_1, \ldots, s_m; s_{m+1}) =$

$$\Pr\{S_x = s_{m+1} | S_{x-m} = s_1, \ldots, S_{x-1} = s_m\}$$

Mm model is fitted to the frequencies of **all the words of length** $(m+1)$

$$\widehat{\pi}(s_1, \ldots, s_m; s_{m+1}) = \frac{N(s_1 \ldots s_m s_{m+1})}{N(s_1 \ldots s_m)}$$

Theoretically, properties derived under M1 can be generalized to Mm: M2 is equivalent to M1 on the alphabet $\mathcal{A}^2 = \{\mathtt{aa}, \mathtt{ac}, \ldots, \mathtt{tt}\}$

Principle for one word under M1

Blom & Thorburn, 82 (M0); *Robin & Daudin, 99* (M1)

Distribution of the distance Y

$$p(y) = \mathsf{Pr}\{Y = y\}$$

1. **Recursive formula** of order $y - 1$ $(\mathbf{O(y^2)})$

$$p(y) = f[p(1), \ldots, p(y-1)]$$

2. Probability **generating function**

$$\phi_Y(t) = \sum_{y \geq 1} p(y) t^y = U_Y(t)/V_Y(t)$$

3. **Taylor expansion** of ϕ_Y

$$p(y) = g[p(y - |\mathbf{w}|), \ldots, p(y-1)]$$

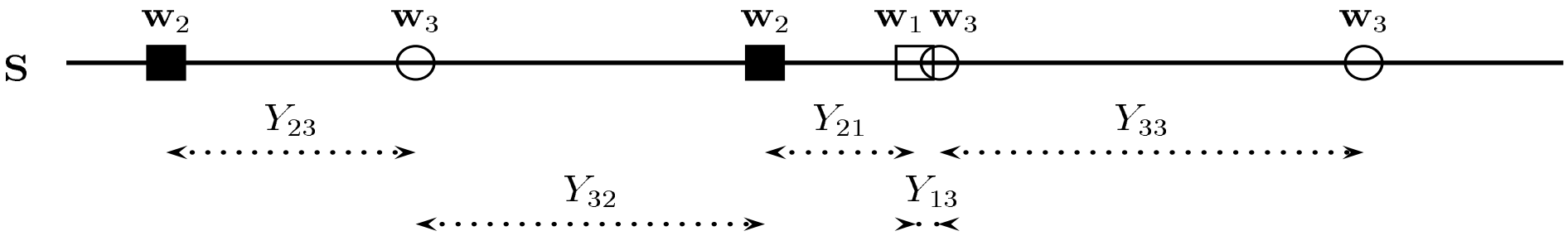
$\rightarrow$ recurrence of order $|\mathbf{w}|$ $(\mathbf{O(y)})$

Principle for a motif

Robin & Daudin, 01 (M1)

We are interested in the distribution of the occurrences of the motif $\mathbf{m} = \{\mathbf{w}_1, \ldots, \mathbf{w}_I\}$

But the distribution of the distances **depends on the words** themselves



Steps 1, 2, 3 follow the same principle as for one word but involve **generating matrices**

Denoting $\phi_{ij}(t) = \phi_{Y_{ij}}(t)$, $(i, j = 1..I)$

$$\underset{I\times I}{\mathbf{\Phi}(t)} = \begin{bmatrix} \phi_{11}(t) & \dots & \phi_{1I}(t) \\ \vdots & & \vdots \\ \phi_{I1}(t) & \dots & \phi_{II}(t) \end{bmatrix}, \qquad \phi_{ij}(t) = \frac{U_{ij}(t)}{V_{ij}(t)}$$

Step 2 requires the **formal inversion** of a generating matrix:

$$\mathbf{\Phi}(t) = \mathbf{F}(t)[\mathbf{I} - \mathbf{F}(t)]^{-1}$$

Complexity of this step: $O(I^3|\mathbf{m}|)$

Application to structured motifs

Difficulty: Complexity of the overlapping structure of structured motif

$$\mathbf{m} = \boxed{\mathbf{v}} \overset{d}{\longleftrightarrow} \boxed{\mathbf{w}}$$

Impossible to calculate the exact distribution of $X_1(\mathbf{m})$

Approximation (*Robin & al., 02*)

1. Probability for $\mathbf{m}$ **to occur at a given position** (using the distribution of the distances): $\mu(\mathbf{m})$

2. Approximation of **order 0** (geometric):

$$\Pr\{N(\mathbf{w}) \geq 1\} \approx 1 - [1 - \mu(\mathbf{m})]^{\ell - |\mathbf{m}| + 1}$$

   does not work (**simulations**)

3. Approximation of **order 1**:

$$\Pr\{N(\mathbf{w}) \geq 1\} \approx 1 - [1 - \mu(\mathbf{m})][1 - \mu_1(\mathbf{m})]^{\ell - |\mathbf{m}|}$$

   where $\mu_1(\mathbf{m}) = \Pr\{\mathbf{m} \text{ at } x | \mathbf{m} \text{ not at } x - 1\}$

Promoters in *B. subtilis*:

131 upstream regions of 100 bps

p-value $< 10^{-16}$

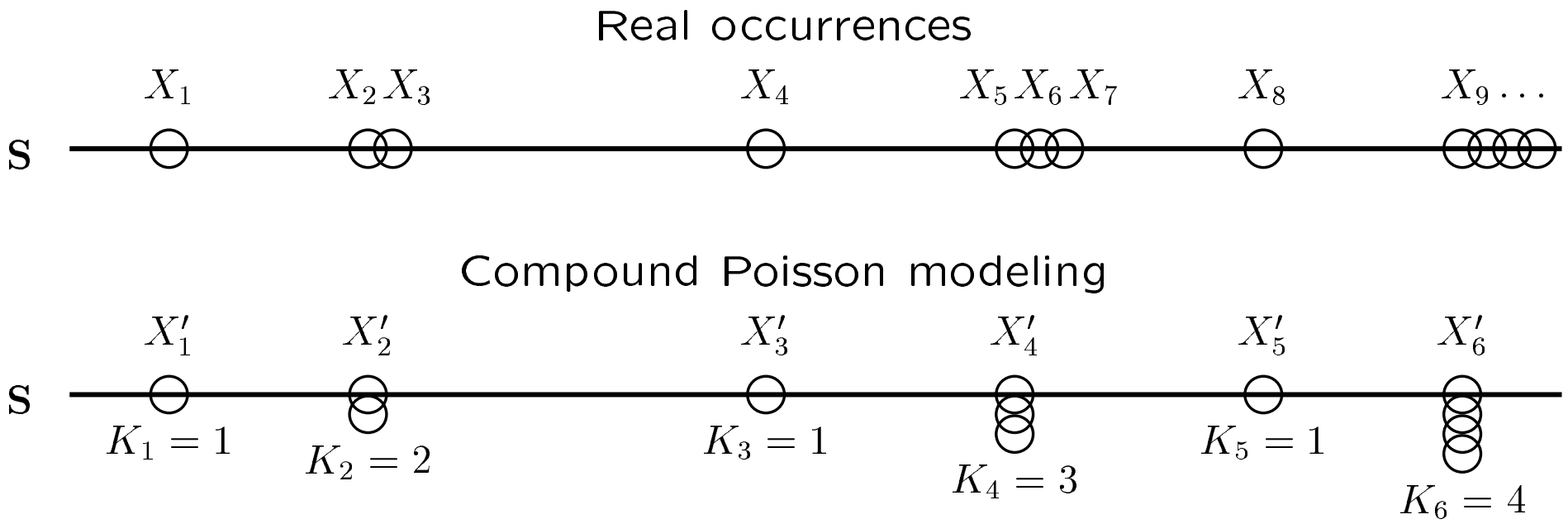
(putative alignment)

$\mathbf{m}$			number of regions containing $\mathbf{m}$	expected number
$\mathbf{v}$	$(d_1 : d_2)$	$\mathbf{w}$		
gttgaca	(16 : 18)	atataat	7	$2.43\ 10^{-2}$
gttgaca	(16 : 18)	tataata	8	$2.23\ 10^{-2}$
tgttgac	(16 : 18)	tataata	10	$2.12\ 10^{-2}$
ttgacaa	(16 : 18)	tacaat	9	$9.82\ 10^{-2}$
ttgacaa	(16 : 18)	tataata	10	$5.07\ 10^{-2}$
ttgacag	(16 : 18)	tataat	9	$7.12\ 10^{-2}$
ttgacaa	(17 : 19)	ataataa	9	$6.97\ 10^{-2}$
ttgttga	(17 : 19)	tataata	8	$5.17\ 10^{-2}$
gttgaca	(17 : 19)	ataataa	8	$3.09\ 10^{-2}$
gttgaca	(17 : 19)	tataata	8	$2.19\ 10^{-2}$
cttgaca	(17 : 19)	tataat	8	$6.04\ 10^{-2}$
tgttgac	(17 : 19)	tataata	12	$2.09\ 10^{-2}$
tgttgac	(17 : 19)	atataat	7	$2.29\ 10^{-2}$
ttgttga	(18 : 20)	tataata	8	$5.09\ 10^{-2}$
gttgaca	(18 : 20)	ataatga	7	$1.79\ 10^{-2}$
gttgttg	(18 : 20)	tataata	7	$2.53\ 10^{-2}$
tgttgac	(18 : 20)	ataataa	10	$2.90\ 10^{-2}$
tgttgac	(18 : 20)	atacta	7	$2.77\ 10^{-2}$
tgttgac	(19 : 21)	ataataa	10	$2.86\ 10^{-2}$
tgttgac	(19 : 21)	atacta	7	$2.73\ 10^{-2}$
tgttgac	(19 : 21)	tataat	10	$6.53\ 10^{-2}$
gttgact	(19 : 21)	ataata	8	$6.25\ 10^{-2}$

Compound Poisson model

Compound Poisson process = Continuous modeling

For rare words, the sequence S can be viewed as a **continuous line** $[0; \ell]$



Clump process $\{C(x)\} =$ Poisson process $\mathcal{P}(\lambda x)$

Clump sizes $\{K_1, K_2, \dots\}$ are i.i.d.

$$\Pr\{K = k\} = g(k)$$

Counting process of the occurrences $\{N(x)\} =$ compound Poisson process:

$$N(x) = \sum_{c=1}^{C(x)} K_c$$

Non overlapping word $\Rightarrow$ **simple Poisson** process

Interpretation: Poisson modeling implies that the clump are **uniformly distributed** along the genome

$\rightarrow$ **Null hypothesis** of the next part

Pólya-Aeppli model

When considering one single word $\mathbf{w}$, the clump sizes have a **geometric distribution**

$$g(k) = a^{k-1}(1-a) \quad \Rightarrow \quad \mathbb{E}(K) = 1/(1-a)$$

where a is the overlapping probability of $\mathbf{w}$

Parameter estimates: In a sequence of length ℓ

- $\widehat{\lambda}$ is the empirical frequency of the clumps:

$$\widehat{\lambda} = \frac{C(\ell)}{\ell}$$

- $\widehat{a}$ is the proportion of overlapped occurrences:

$$\widehat{a} = \frac{N(\ell) - C(\ell)}{C(\ell)}$$

Properties

- In the Pólya-Aeppli model, distances Y are **i.i.d.** with mixture distribution

$$\begin{array}{rcll} Y & = & 0 & \text{with probability } a \quad \text{(overlap)} \\ & \sim & \mathcal{E}(\lambda) & \text{with probability } 1-a \quad \text{(no overlap)} \end{array}$$

- Pólya-Aeppli is the **best approximation** of the distribution of the word count in the Markov model

- $\mathbb{E}[N(\ell)] = \ell \times \lambda \times \mathbb{E}(K)$

$$\Rightarrow \quad \widehat{N}(\ell) = \ell\widehat{\lambda}/(1-\widehat{a}) = N(\ell)$$

$$\Rightarrow \quad \textbf{no word has an ``unexpected'' count}$$

Clump size modeling

Robin, 02

In the general case (e.g. **motif** $\mathbf{m} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots\}$), the clump sizes do not have a geometric distribution

- Empirical estimates of an **arbitrary** distribution $g(k)$
- Empirical estimates of the **overlapping probabilities** between words $\mathbf{w}_1$, $\mathbf{w}_2$, ...
- Markov estimates of the same probabilities

However, the distances Y between words are **not i.i.d.**

Motifs distribution along a sequence

Two statistics

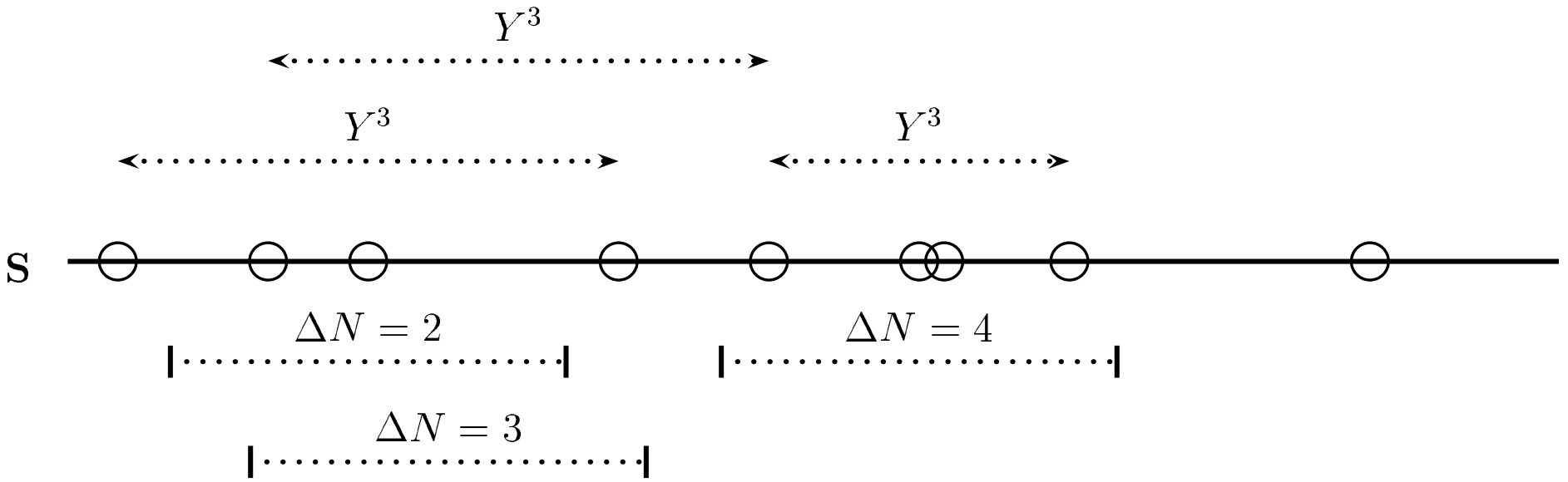
We aim to detect **poor** or **rich** regions in terms of occurrences of a given motif

A natural criterion for a given region is the **ratio**

$$\frac{\textbf{number of occurrences} \text{ in the region}}{\textbf{size} \text{ of the region}}$$

Cumulated distances: $\dfrac{\text{fixed numerator } r}{\text{random denominator } Y^r}$

(*Karlin & Macken, 91*: r-scans)



Moving windows: $\dfrac{\text{random numerator } \Delta N}{\text{fixed denominator } y}$

Distribution of the statistics

Robin, 02

Cumulated distance: the distribution of

$$Y_i^r = \sum_{j=i}^{i+r-1} Y_j = X_{i+r} - X_i$$

is known **when the distances Y_i are i.i.d.** (e.g. in the one word case) for Markov and compound Poisson models

Moving window: the distribution of

$$\Delta N(x) = N(x) - N(x-y)$$

is known for Markov and compound Poisson models

Extremal statistics

We are interested in the **richest** region, i.e.

$$Y^r_{\min} = \min_i \{Y^r_i\} \qquad \text{or} \qquad \Delta N_{\sup} = \sup_x \{\Delta N(x)\}$$

Poisson approximation

If the $\{Y^r_i\}$ or the $\{\Delta N(x)\}$ were **independent**

$$\Pr\{Y^r_{\min} \leq y\} \underset{n\to\infty}{\longrightarrow} \exp[-(n-r)\Pr\{Y^r \leq y\}]$$

$$\Pr\{\Delta N_{\sup} > n\} \underset{\ell\to\infty}{\longrightarrow} \exp[-(\ell-y)\Pr\{\Delta N > n\}]$$

Chen-Stein method

$\{Y_i^r\}$ and $\{\Delta N(x)\}$ are **not independent** but the quality of the Poisson approximation can be controlled in total variation distance (*Arratia & al, 89*):

$$\max_y \left| \Pr\{Y^r_{\min} \leq y\} - e^{-(n-r)\Pr\{Y^r \leq y\}} \right| \leq \text{bound}$$

Cumulated distances: an **explicit bound** can be calculated (*Dembo & Karlin, 92*)

Moving windows: **no explicit bound** can be derived, but this approximation is **optimal** (*Barbour & Brown, 92*)

Applications

CHI motif in *H. influenza*

In terms of overlap, $\mathbf{m} = (\mathtt{gNtggtgg})$ behaves as one single word

$\rightarrow$ **cumulated distances** can be used

Number of occurrences: $\ell = 1\,903\,356$ bps

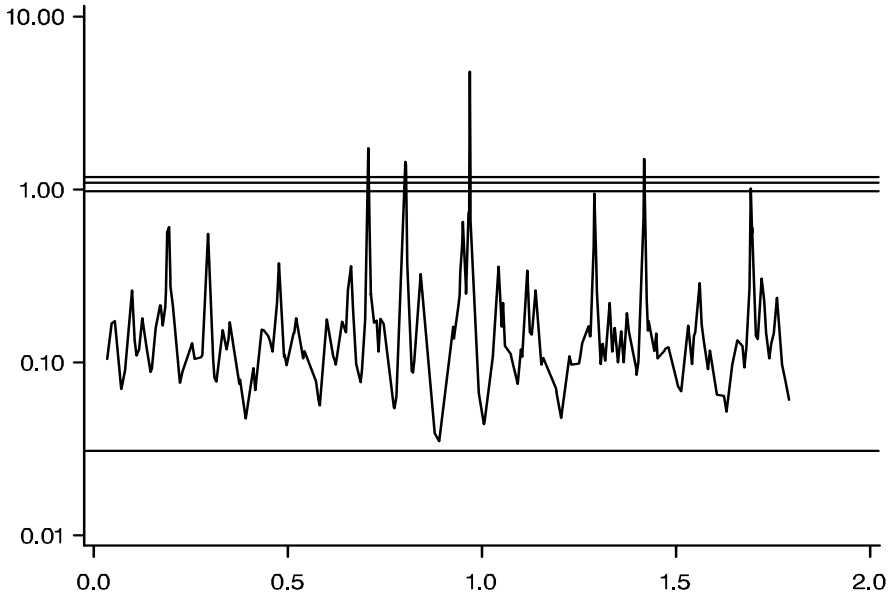
observed number of occurrences	=	223
expected under Markov (M1)	=	**58.5**
expected under compound Poisson	=	223

Significancy thresholds: for $\alpha = 5\%$

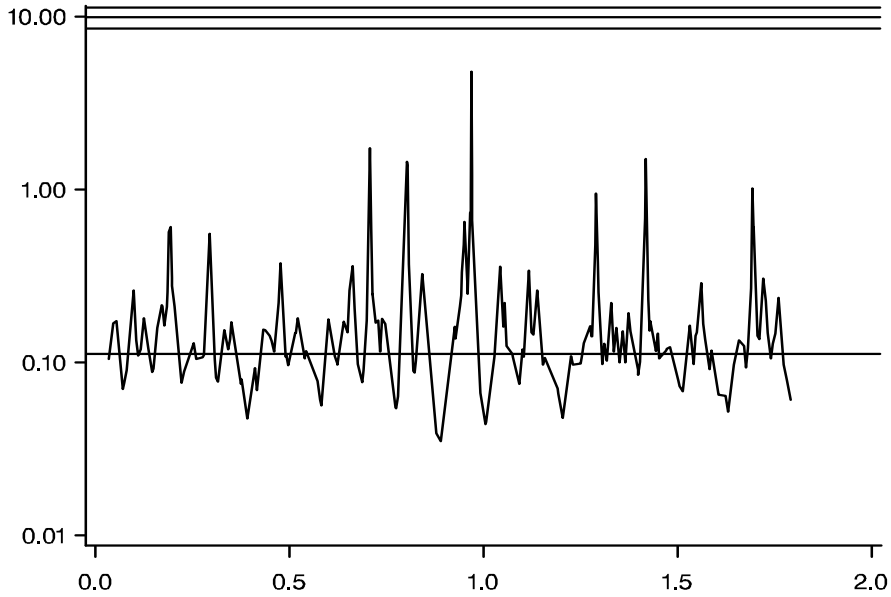
for Y^r:	6 312 bps
for $\min_{i=1\ldots222} Y_i^r$:	238 bps

Distribution: cumulated distances of order $r = 3$

plot of the **ratio** $3/Y^3$ $(\times 10^{-3})$ versus the **position** x



Markov (M1)
overall bias



compound Poisson
no significant peak

Remarks:

- **Markov model M7** would be unbiased (since $|\mathbf{m}| = 8$) but involves more than 12 000 **parameters**

- In the compound Poisson model, the peak around 1.0 Mb (replication termination) is significant **on its own**:

$$\Pr\{Y^3 \leq 208\} = 1.610^{-4}$$

$$\Pr\left\{\min_{i=1..220}\left(Y_i^3\right) \leq 208\right\} > 0.05$$

Palindromes in *E. coli* ($\ell = 4\,638\,868$)

There are 64 palindromes of length 6

They occur 54 724 times in 50 941 clumps

Clump size: Because of their overlapping structure, clumps can not be considered as geometric

$\rightarrow$ **Moving windows**

We use a **parsimonious modeling** based the overlapping probabilities given by the **M0 model** (4 parameters)

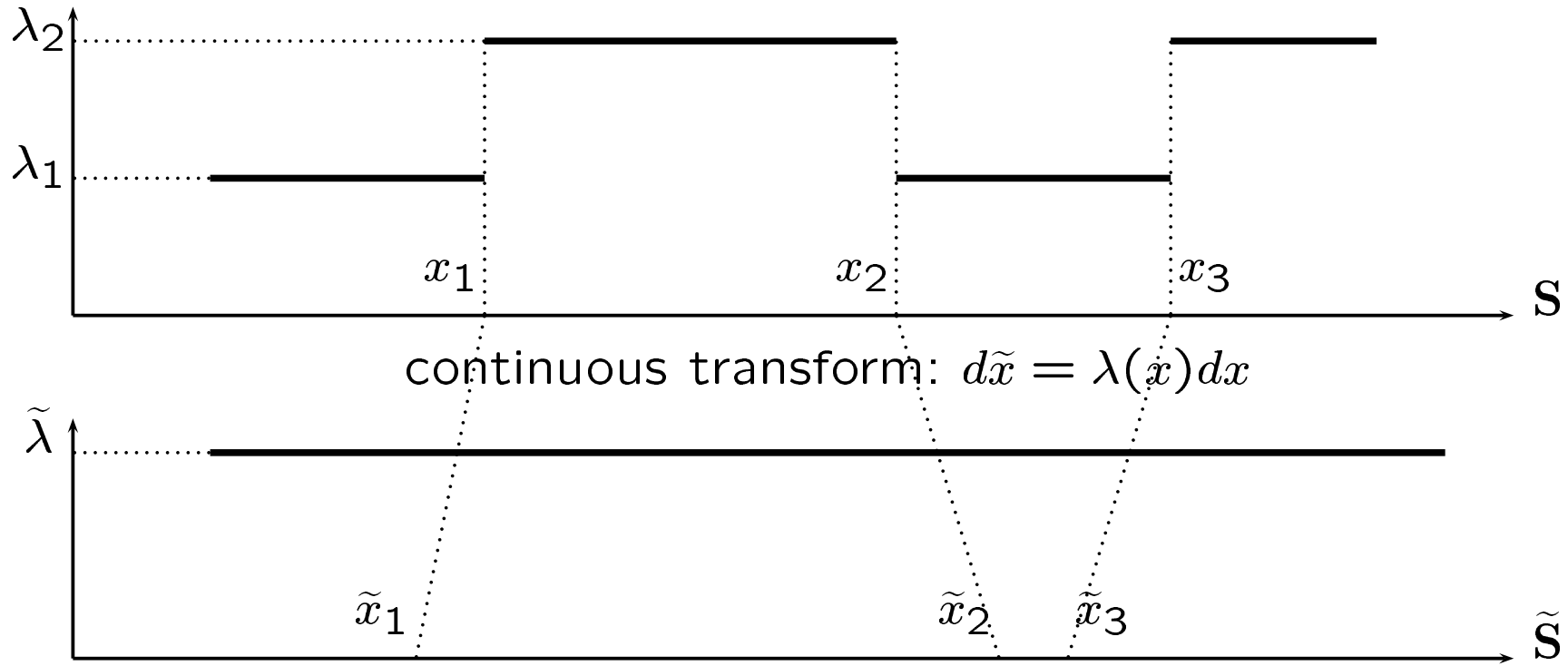
Results: Moving windows of width $y = 10\,000$ bps

- **Poorest region**: 73 occurrences (p-value $> 10\%$)
  non significant
- **Richest region**: 185 occurrences (p-value $< 5\%$)
  [2 460 567 bps; 2 461 566 bps]

... interpretation?

Distribution in heterogeneous sequences (*S. Ledent*)

1. Estimate the intensity of the **heterogeneous Poisson process** on the basis of a **prior heterogeneity** (coding / non-coding, g-c rich / poor regions, etc)
2. "**Homogenize**" the occurrences process



3. Apply the methods presented above to the homogenized sequence