MOTIFS DISTRIBUTION IN DNA SEQUENCES

Stéphane ROBIN
robin@inapg.inra.fr

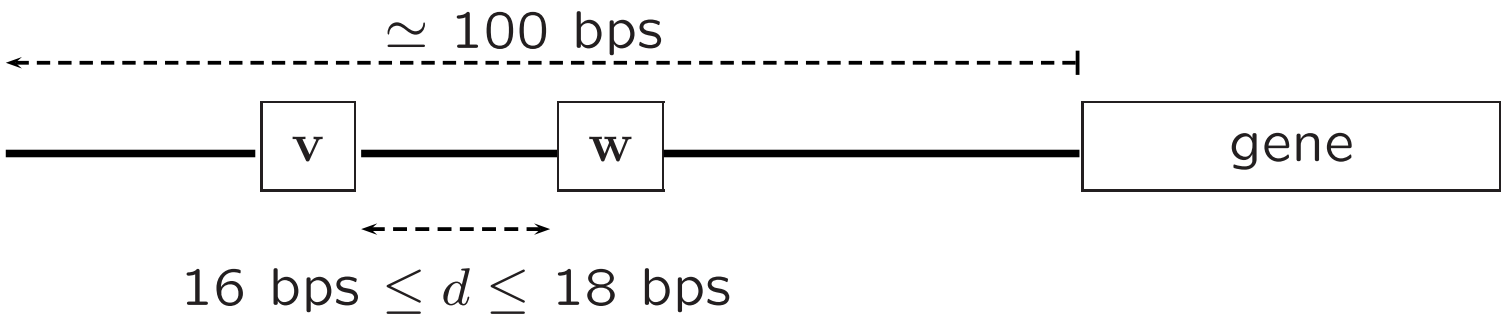
UMR INA-PG / INRA, Paris
Mathématique et Informatique Appliquées

Bio-Info-Math Workshop, Tehran, April 2005

Biological interest of motif statistics

Four examples

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



Which structured motifs occur almost (*too*?) systematically in upstream regions of the genes of a given species ?

Ex 2 : CHI motifs in bacterial genomes

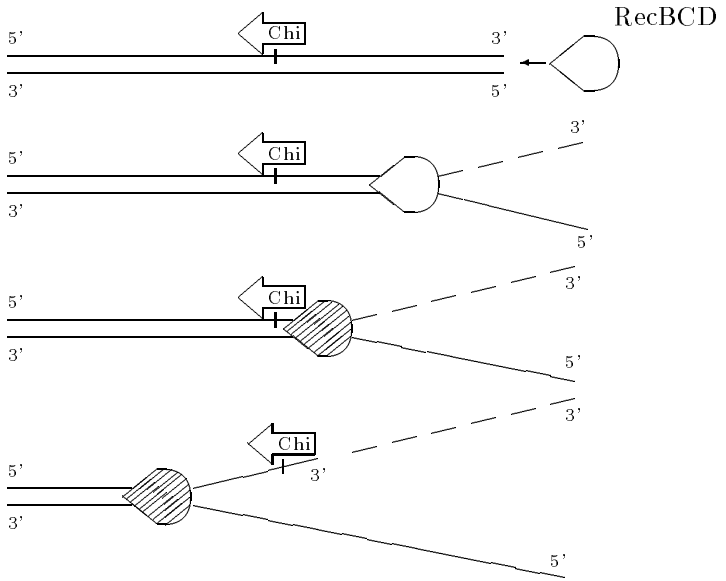
Crossover Hot-spot Initiator : defense function of the genome against the degradation activity of an enzyme

Known in several bacterial genomes :

E. coli : `gctggtgg`

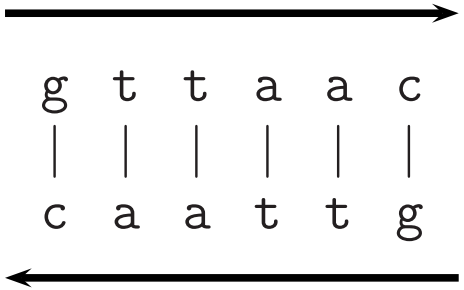
H. influenza : `gNtggtgg`

(*Figure : Schbath, 95*)



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.

Ex 4 : Detection of unknown motifs

– Motifs with favorable functions should be *unexpectedly frequent*,
– Motifs with damaging functions should be *unexpectedly rare*

Even when we know nothing about them (except their length) , such motifs may be detected only because they have *unexpected frequencies*

A model : what for ?

Model = Reference

To be able to decide if something is unexpected, we first need to know what to expect.

To avoid artifacts, the model should typically account for

- the frequencies of nucleotides, or di-, or tri-nucleotides in the sequence,
- the overlapping structure of the word,
- eventually, the overall frequency of the word in the sequence

The choice of the model (Markov chain / compound Poisson process) depends on the question.

(*R., Rodolphe & Schbath ; 05*)

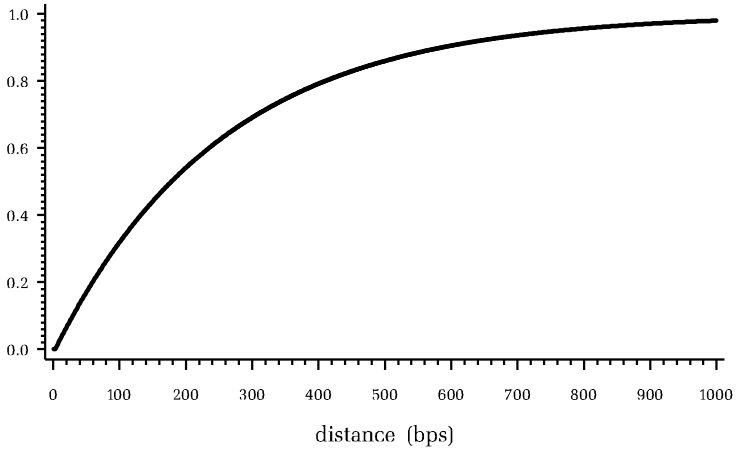
Overlapping structure of the word

Some words can overlap themselves (see *Conway (Gardner, 74)* ; *Guibas & Odlyzko, 81*).

Such words tend to occur in *clumps* and have a less regular distribution along the sequence.

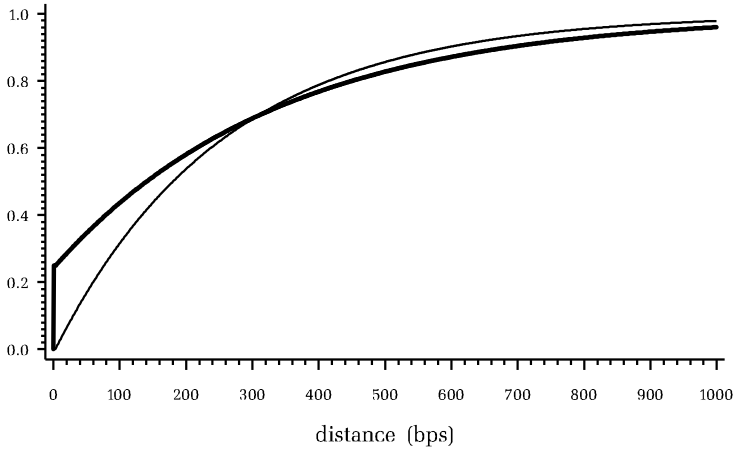
Cdf of the distance between two occurrences under model M00 :

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



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

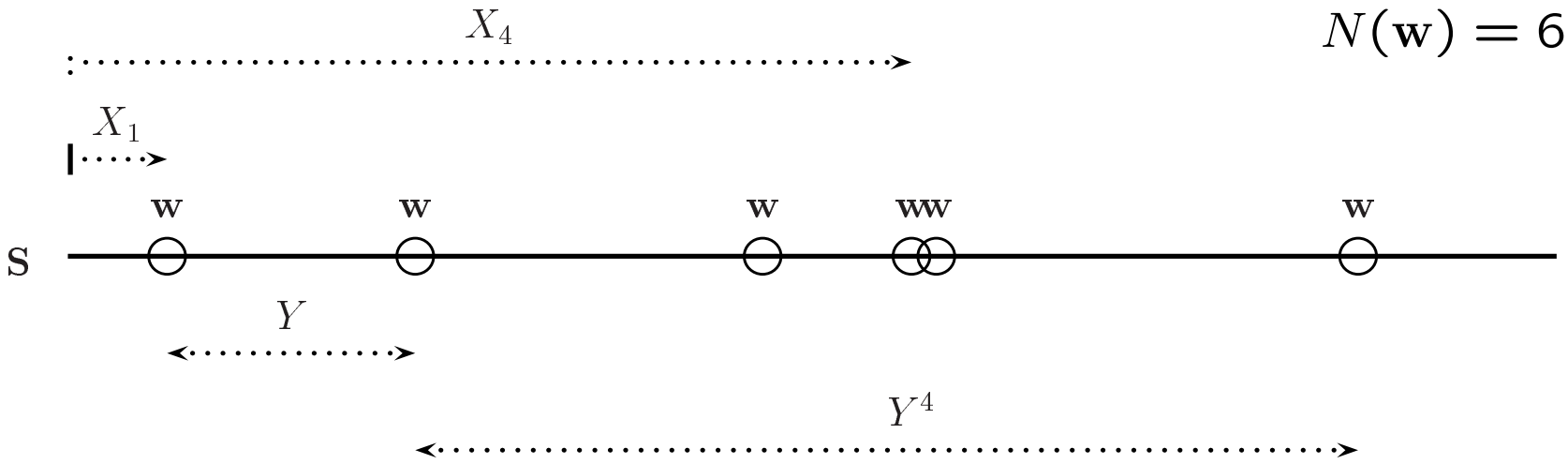
$\mathbf{w} = (\mathtt{aaaa})$



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

Probabilities and distributions of interest

Positions, distances, counts



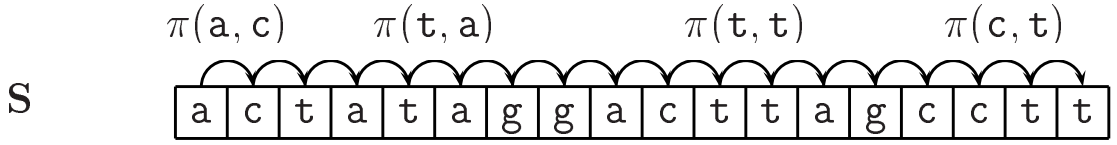
- Probability for a motif to occur in a sequence : X_1
  $\longrightarrow$ promoter motifs
- Distribution of the number of occurrences : N
- Distribution of the occurrences along the sequence : $Y^r, N(x) - N(x-y) \longrightarrow$ CHI motifs, palindromes

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) over the alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$
- with transition probabilities $\pi(s_1, \ldots, s_m; s_{m+1})$.



The 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 $\mathcal{A}^2 = \{\mathtt{aa}, \mathtt{ac}, \ldots, \mathtt{tt}\}$

Distribution of the count

The (ficticious) word $\mathbf{w} = \texttt{gctt}$ occurs 56 times in a given genome, is it significantly high ?

M1 model. Occurrence probability (at any position) :

$$\mu(\mathbf{w}) = \mu(w1) \times \pi(w_1, w_2) \times \cdots \times \mu(w_{|\mathbf{w}|-1}, w_{|\mathbf{w}|})$$

Expected count (sequence of length ℓ) : $\mathbb{E}N(\mathbf{w}) = (\ell - k + 1)\mu(\mathbf{w})$
Kleffe & Borodowsky, 92 : $\mathbb{E}N(\mathbf{w}), \mathbb{V}N(\mathbf{w})$

Distribution of the count. The exceptionality of the observed frequency is measured by the p-value

$$\Pr_{M1}\{N(\mathbf{w}) \geq n_{obs}(\mathbf{w})\} = \Pr_{M1}\{N(\texttt{gctt}) \geq 56\}$$

Gaussian approximation. If $\mathbf{w}$ is "frequent" , $\mathbb{E}N(\mathbf{w}) = \mathcal{O}(\ell)$ (*Prum & al, 95*),

$$U(\mathbf{w}) = \frac{N(\mathbf{w}) - \widehat{\mathbb{E}}N(\mathbf{w})}{\sqrt{\widehat{\mathbb{V}}N(\mathbf{w})}} \approx \mathcal{N}(0,1)$$

Poisson approximation. If $\mathbf{w}$ is "rare" : $\mathbb{E}N(\mathbf{w}) = \mathcal{O}(\log \ell)$ (*Schbath, 95*),

$$N(\mathbf{w}) \approx \mathcal{P}[\mathbb{E}N(\mathbf{w})]$$

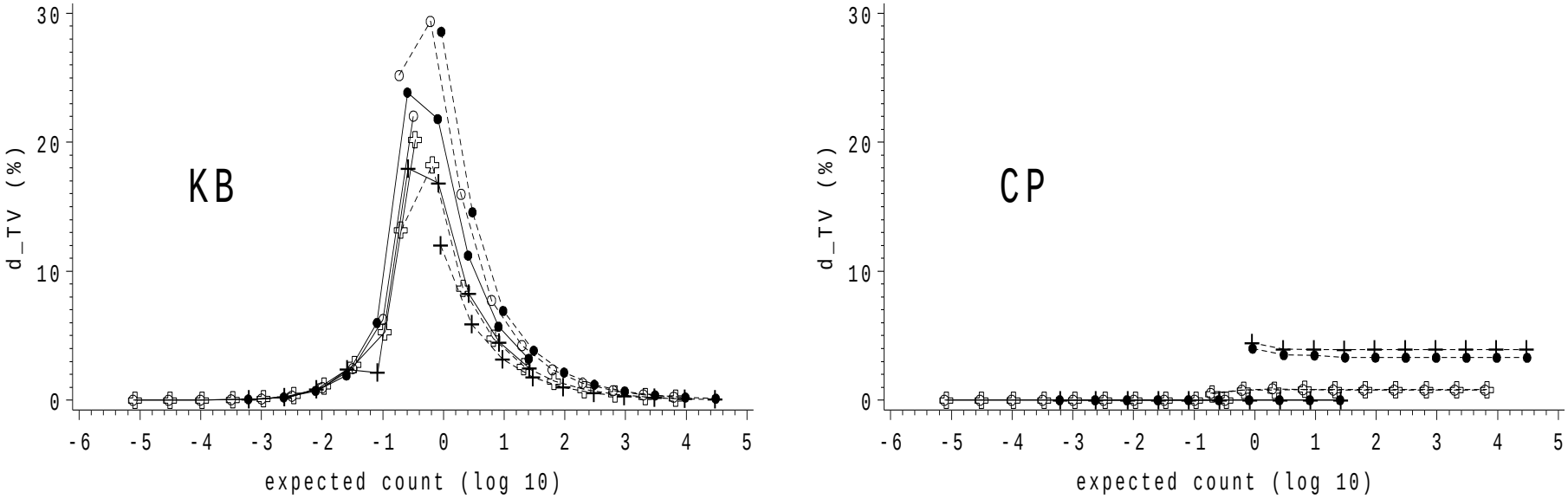
For overlapping words : compound Poisson approximation.

Binomial approximation : *Van Helden, 99*

Exact distribution of $N(\mathbf{w})$: *R. & Daudin, 99 ; Nicodème & al, 99 ; Regnier, 00*

Large deviation : *Nuel, 01*

Quality of the approximations. The (compound) Poisson approximation turns out to perform very well, in many situations (*R. & Schbath, 01*) :



Even for rather frequent words.

CP approximation fails for frequent and short words.

Influence of the order of the Markov chain. The exceptionality of a word's frequency strongly depends on the chosen model :

	W =GGCGCTGG $N(W) = 77$		W =CGCTGGCG $N(W) = 68$		W =GCCAGCA $N(W) = 57$	
Modèle	$U(W)$	Rang	$U(W)$	Rang	$U(W)$	Rang
M0	24.5041	**2**	22.2126	**3**	19.6367	8
M1	19.2064	**2**	14.2766	13	18.5421	**3**
M2	11.2313	6	7.4762	78	8.2979	49
M3	6.0286	112	1.7250	6896	5.2716	208
M4	7.4302	49	0.2647	23754	2.0605	4043
M5	3.4283	577	0.6474	17247	3.5052	521
M6	-0.4911	42656	0.0313	30053	1.7435	4558

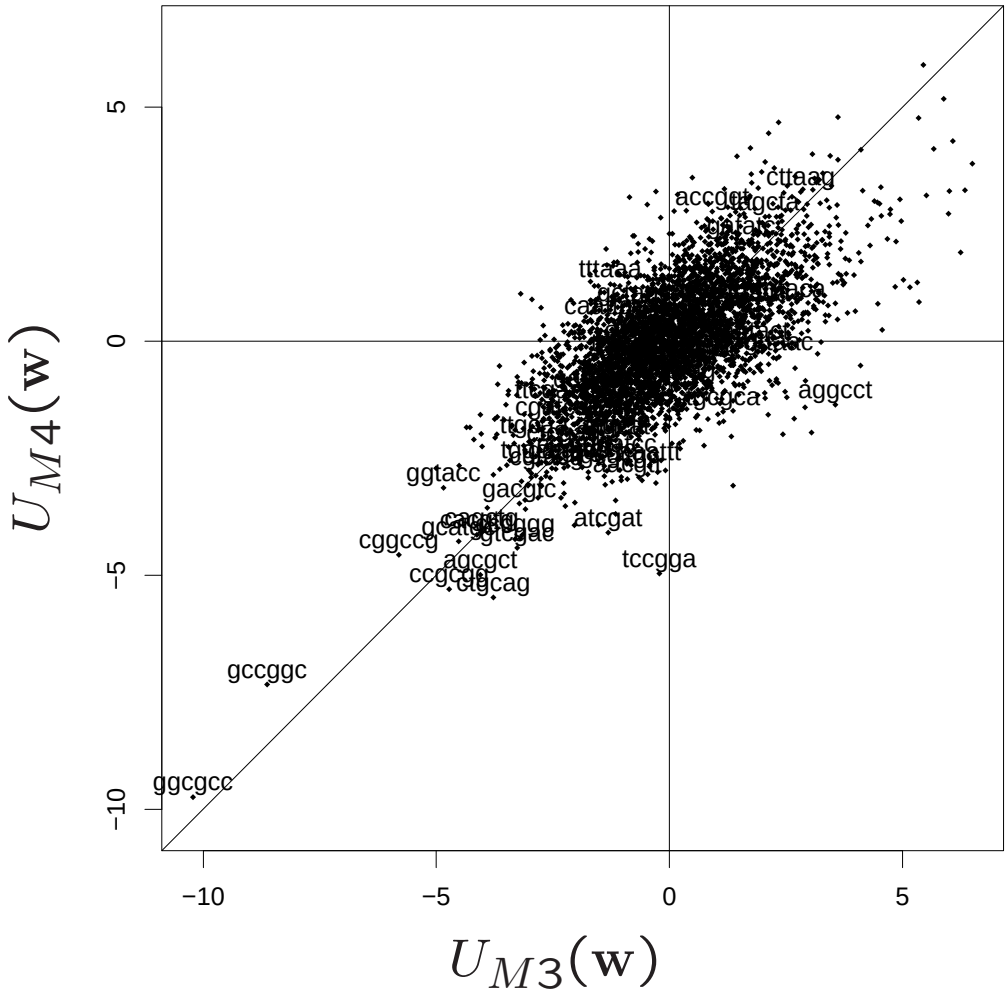
(R'mes software : *Bouvier & al., 99*)

The CHI motif $\mathbf{w}$ = gctggtgg appears at the top of the list for almost all orders.

Palindromes of length 6 :

In both model M3 and M4, most of them seem to be avoided in the genome of *E. coli*

Most of them are restriction sites : possible defense system of *E. coli*'s genome.



Distribution of the distance : one word

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

Distribution of the distance Y

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

1. Linear recursive formula of order $y - 1$ (complexity $= O(y^2)$)

$$p(y) = \sum_{z=1}^{y-1} c_z p(y - z)$$

2. Derive the 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 with a *new* linear recursive formula of order $|\mathbf{w}|$ (complexity $= O(y)$)

$$p(y) = \sum_{z=1}^{|\mathbf{w}|} c'_k p(y - z)$$

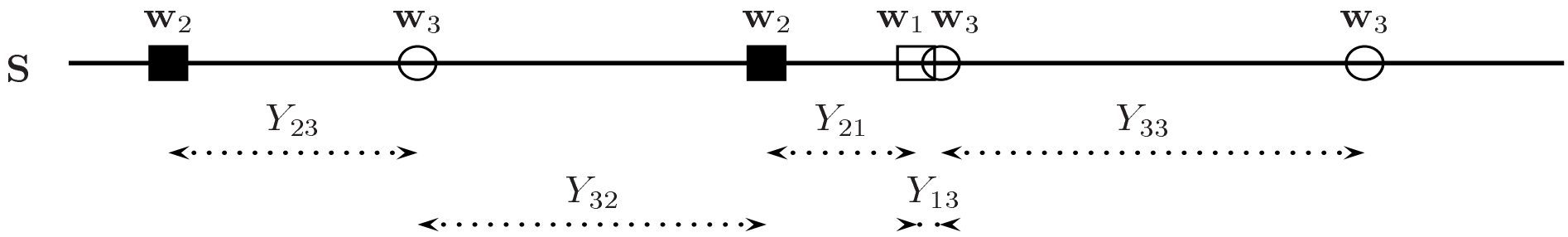
Principle for a set of words

R. & Daudin, 01 (M1)

Consider the distribution of the occurrences of the motif

$$\mathbf{m} = \{\mathbf{w}_1, \ldots, \mathbf{w}_I\}$$

The distribution of the distances depends on the words themselves (semi-Markov process)



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 inversion of a generating matrix :

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

Limitations :

- Complexity of this last step : $O(I^3|\mathbf{m}|)$
- Numerical instability except if $[\mathbf{I} - \mathbf{F}(t)]$ is inverted formally $\Longrightarrow$ small set of short words (small I and $|\mathbf{m}|$)

Other approaches : algorithmic (*Nicodème, 00*), embedded Markov chain (*Fu & Koutras, 94*, *Koutras, 97*), properties of the exponential family (*Stefanov & Pakes, 99*), etc.

Application to structured motifs

Difficulty : Complexity of the overlapping structure of structured motif

$$\mathbf{m} = \boxed{\mathbf{v}} \xleftarrow{\quad d \quad} \boxed{\mathbf{w}}$$

$\Longrightarrow$ impossible to calculate the exact distribution of $X_1(\mathbf{m})$ with the method presented above

Approximation (*R. & 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) does not work (simulations) :

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

3. Approximation of order 1 ($\mu_1(\mathbf{m}) = \Pr\{\mathbf{m} \text{ at } x | \mathbf{m} \text{ not at } x-1\}$) :

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

Promoters in *B. subtilis* :

131 upstream regions of 100 bps

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

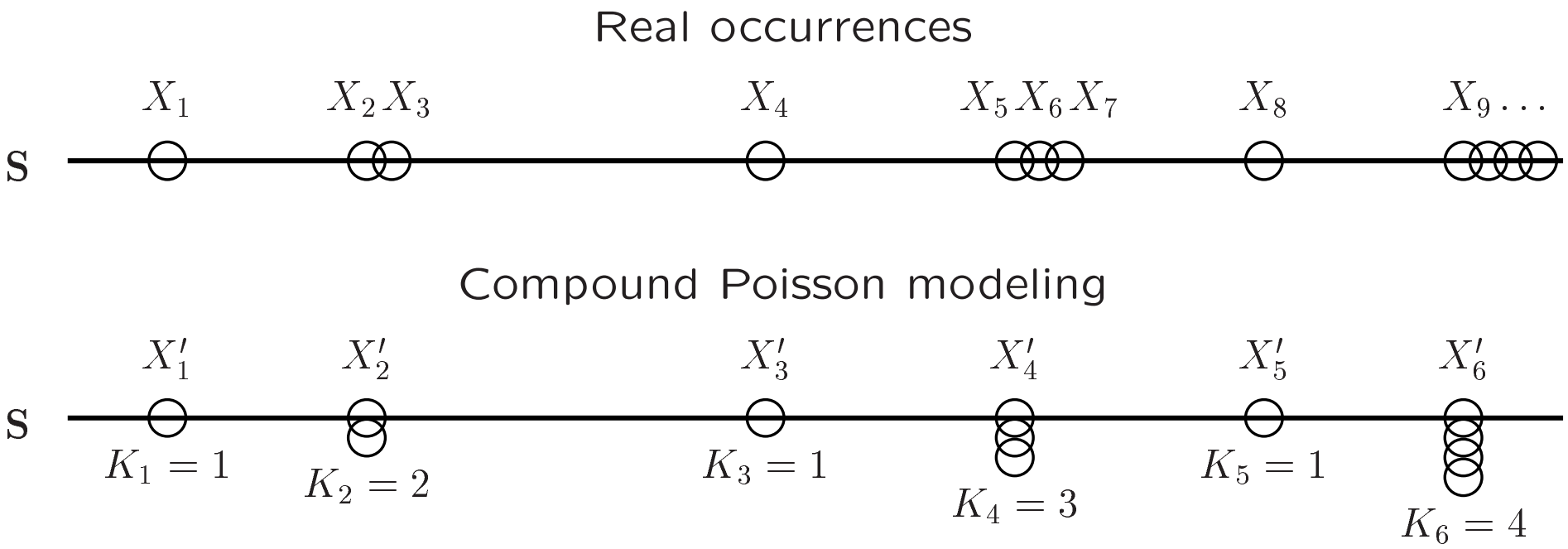
(putative alignment)

$\overbrace{\mathbf{v} \quad (d_1 : d_2) \quad \mathbf{w}}^{\mathbf{m}}$	number of regions containing $\mathbf{m}$	expected number
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 $\mathbf{S}$ can be viewed as a continuous line $[0; \ell]$



Clump process $\{C(x)\}$ = Poisson process with intensity $\equiv \lambda$

Clump sizes $\{K_1, K_2, \ldots\}$ are iid

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

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

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

Non overlapping word $\Longrightarrow$ simple Poisson process

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

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

Pólya-Aeppli model

When considering one single word $\mathbf{w}$, the clump size has a geometric distribution

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

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 : $\widehat{\lambda} = C(\ell)/\ell$
- $\widehat{a}$ is the proportion of overlapped occurrences : $\widehat{a} = \frac{N(\ell)-C(\ell)}{N(\ell)}$

Properties

- Pólya-Aeppli is the best approximation of the distribution of the word count in the Markov model (*R. & Schbath, 01*)
- $\mathbb{E}[N(\ell)] = \ell \times \lambda \times \mathbb{E}(K) \quad \Longrightarrow \quad \widehat{\mathbb{E}}N(\ell) = \ell\widehat{\lambda}/(1-\widehat{a}) = N(\ell)$
  $\Longrightarrow$ *no word has an "unexpected" count*

Clump size modeling

R., 02

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

We may use

- empirical estimates of an arbitrary distribution $g(k)$
- empirical estimates of the overlapping probabilities between words $\mathbf{w}_1, \mathbf{w}_2, \ldots \Longrightarrow I^2$ parameters to be estimated
- Markov estimates of the overlapping probabilities $\longrightarrow$ even M00 may provide a good fit

However, distances Y between words are *not iid*

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

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

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

Local counts in a window of width y : $\dfrac{\text{random numerator } \Delta N}{\text{fixed denominator } y}$

Distribution of the statistics

R., 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 iid* (e.g. in the one word case) for Markov and compound Poisson models

Local count : the distribution of the count

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

is known for Markov and compound Poisson models (*Barbour & al, 92*)

Extremal statistics

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

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

Chen-Stein approximation (*Arratia & al, 89*)

Cumulated distances : an explicit bound distance can be calculated (*Dembo & Karlin, 92*) for the distribution of Y^r_{min} :

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

Local counts : no explicit bound can be derived, but this approximation

$$\Pr\{\Delta N_{\text{sup}} > n\} \simeq \exp[-(\ell - y)\Pr\{\Delta N > n\}]$$

is optimal (*Barbour & Brown, 92*)

Applications

CHI motif in *H. influenza*

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

$\Longrightarrow$ cumulated distances can be used

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

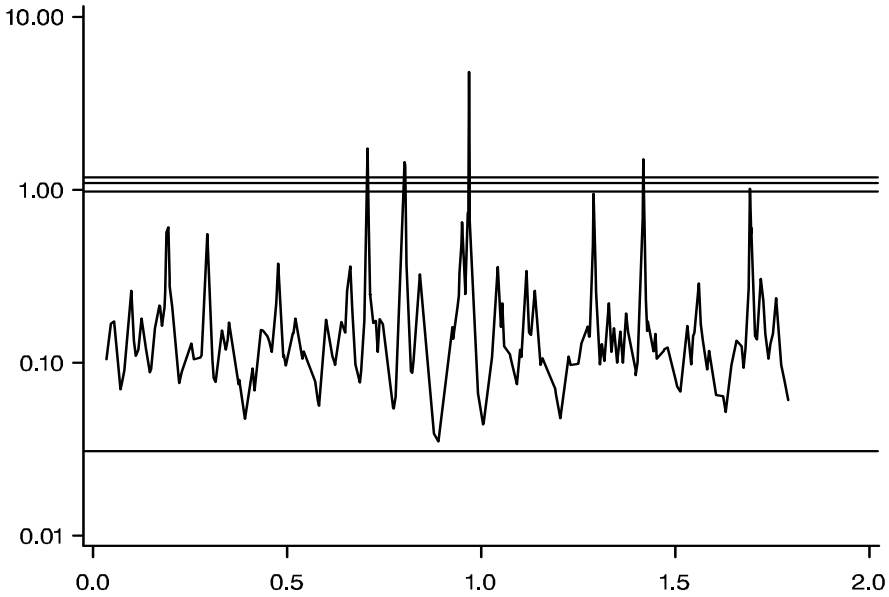
observed number of occurrences	=	223
expected under Markov (M1)	=	*58.5*
expected under compound Poisson	=	223

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

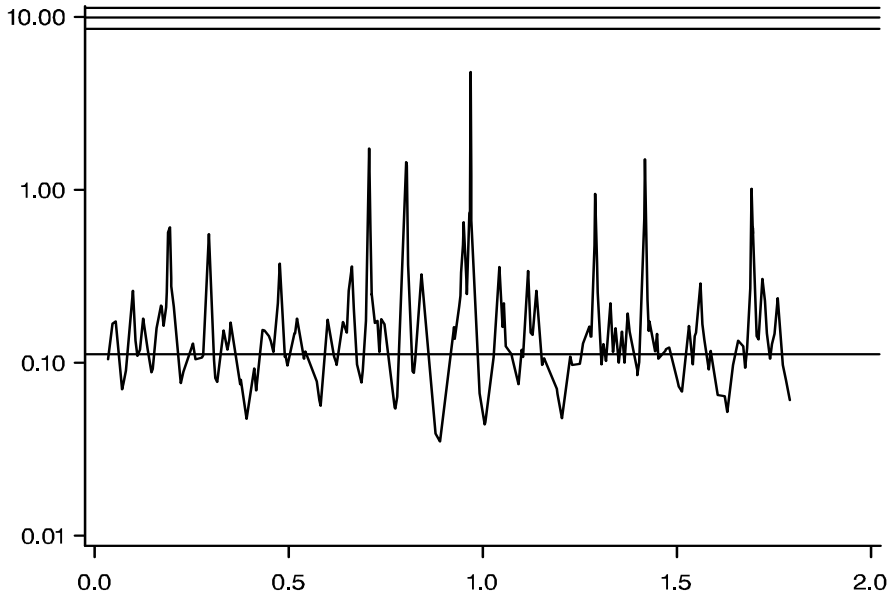
for Y^r :	6 312 bps
for $\min_{i=1\ldots222} Y_i^r$:	238 bps

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

  The compound Poisson model has a better fit with much less 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

$\Longrightarrow$ Local counts should be used

We use a parsimonious modeling of $g(k)$ based the overlapping probabilities given by the M0 model (4 parameters)

Results : 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 : horizontal transfer ?

Distribution in heterogeneous sequences

Ledent & R., 04

An exogenous information about the heterogeneity of the sequence is sometimes available.

It can be summarize in the quantity $\pi_s(x) =$

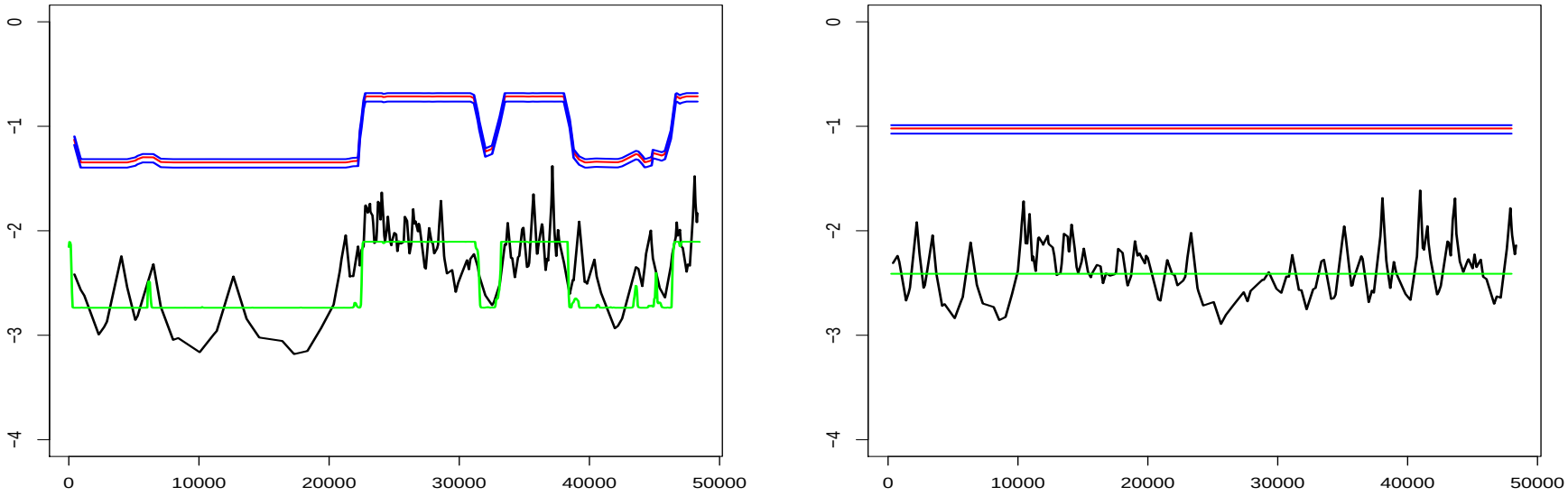
- binary (0/1) variable indicating if position x belongs to state s, where states can be : coding / non coding,
- posterior probability of being in state s at position x provided by an HMM model

The intensity $\lambda(x)$ can be modeled according to this information :

$$\lambda(x) = \sum_s \lambda_s \pi_s(x),$$

so does the distribution of the clump.

Three steps estimation procedure. Occurrences of aatt in the genome of phage *Lambda* ($\ell = 48\,500$ bps)



3 steps :

1. Estimate the intensity $\lambda(x)$ (left : green line)
2. "Homogenize" the clump process and calculate thresholds (right : red line + blue lines for the bounds)
3. come back to the original process (left)