Statistique de mots en génomique : ce qui a été fait, ce qu'il reste à faire

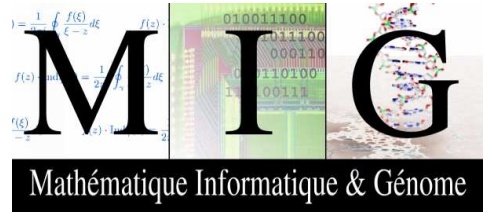
Sophie S. Schbath

Statistics of motifs: importance in genome analysis and open problems

S. Schbath

INRA, Jouy-en-Josas, France

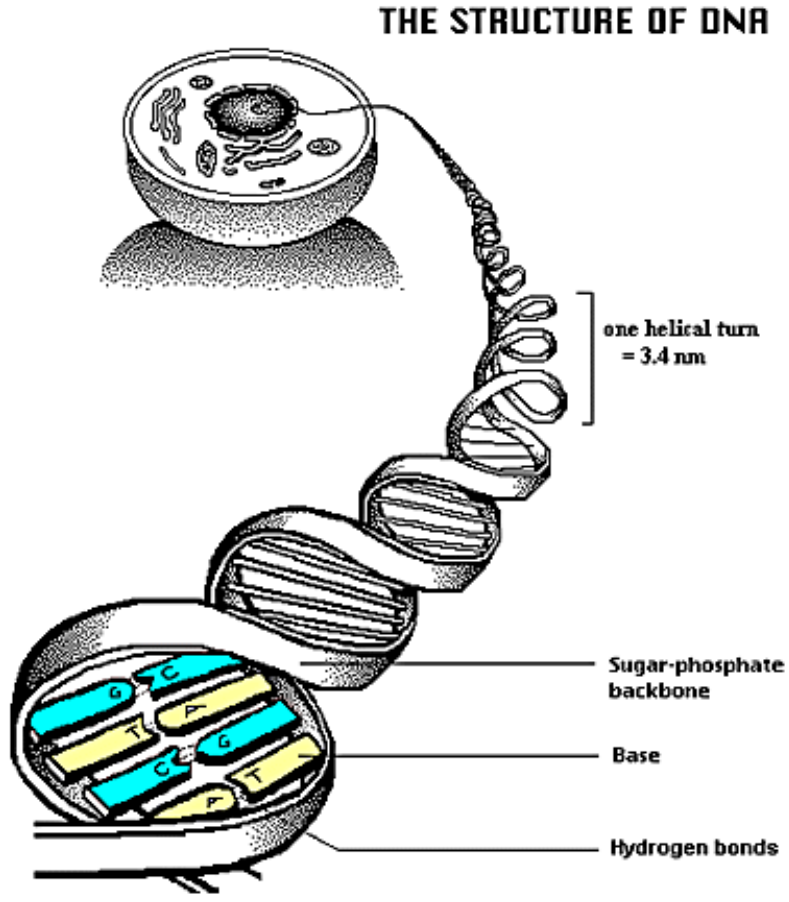
http://ssbgroup.fr



Introduction

DNA and motifs

- **DNA**: Long molecule, sequence of nucleotides
- **Nucleotides**: **A**(denine), **C**(ytosine), **G**(uanine), **T**(hymine).
- **Motif** (= oligonucleotides): short sequence of nucleotides, e.g. CAGTAG
- **Functional motif**: recognized by proteins or enzymes to initiate a biological process



TAGACAGATAGACGAT **CAGTAG** CCAGTAGACAGTAGGCATGA. . .

Some functional motifs

- **Restriction sites**: recognized by specific bacterial restriction enzymes $\Rightarrow$ double-strand DNA break.
  E.g. GAATTC recognized by *Eco*RI

- **Chi motif**: recognized by an enzyme which processes along DNA sequence and degrades it $\Rightarrow$ enzyme degradation activity stopped and DNA repair is stimulated by recombination.
  E.g. GCTGGTGG recognized by *RecBCD* (*E. coli*)

- ***parS***: recognized by the *Spo0J* protein $\Rightarrow$ organization of *B. subtilis* genome into macro-domains.

  $$\underset{c}{\mathrm{TGTT}}\overset{t}{\mathrm{A}}\underset{c}{\mathrm{A}}\underset{t}{\mathrm{C}}\mathrm{ACGTG}\underset{t}{\mathrm{A}}\mathrm{AACA}$$

- **promoter**: structured motif recognized by the RNA polymerase to initiate gene transcription.

  E.g. $\boxed{\mathrm{TTGAC}}\overset{(16;18)}{---}\boxed{\mathrm{TATAAT}}$ (*E. coli*).

Some functional motifs

- **Restriction sites**: recognized by specific bacterial restriction enzymes $\Rightarrow$ double-strand DNA break.
  E.g. GAATTC recognized by *Eco*RI
  very rare along bacterial genomes
- **Chi motif**: recognized by an enzyme which processes along DNA sequence and degrades it $\Rightarrow$ enzyme degradation activity stopped and DNA repair is stimulated by recombination.
  E.g. GCTGGTGG recognized by *RecBCD* (*E. coli*)
  very frequent along *E. coli* genome
- ***parS***: recognized by the *Spo0J* protein $\Rightarrow$ organization of *B. subtilis* genome into macro-domains.
  $\underset{c}{\text{TGTT}}\overset{t}{\underset{c}{\text{A}}}\underset{t}{\text{ACACGTG}}\underset{t}{\text{A}}\text{AACA}$
  very frequent into the ORI domain, rare elsewhere
- **promoter**: structured motif recognized by the RNA polymerase to initiate gene transcription.
  E.g. $\boxed{\text{TTGAC}}\overset{(16;18)}{---}\boxed{\text{TATAAT}}$ (*E. coli*).
  particularly located in front of genes

Prediction of functional motifs

Most of the functional motifs are unknown in the different species.

For instance,

- which would be the **Chi motif** of *S. aureus*? [Halpern *et al.* (08)]
- Is there an equivalent of ***parS*** in *E. coli*? [Mercier *et al.* (08)]

Statistical approach: to identify candidate motifs based on their statistical properties.

The most over-represented 8-letter words under M1

E. coli ($\ell = 4.6\ 10^6$)

word	obs	exp	score
gctggtgg	762	84.9	73.5
ggcgctgg	828	125.9	62.6
cgctggcg	870	150.8	58.6
gctggcgg	723	125.9	53.3
cgctggtg	619	101.7	51.3

The most over-represented families $a\text{n}bcdefg$ under M1

H. influenzae ($\ell = 1.8\ 10^6$)

motif	obs	exp	score
gntggtgg	223	55.3	22.33
anttcatc	469	180.3	21.59
anatcgcc	288	87.8	21.38
tnatcgcc	279	84.5	21.18
gnagaaga	270	83.6	20.10

Statistical questions on word occurrences

Here are some quantities of interest.

- **Number of occurrences** (overlapping or not):
  - Is $N^{\text{obs}}(\mathbf{w})$ significantly high in S?
  - Is $N^{\text{obs}}(\mathbf{w})$ significantly higher than $N^{\text{obs}}(\mathbf{w}')$ in S?
  - Is $N_1^{\text{obs}}(\mathbf{w})$ significantly more unexpected in S_1 than $N_2^{\text{obs}}(\mathbf{w})$ in S_2?
- **Distance** between motif occurrences:
  - Are there significantly rich regions with motif $\mathbf{w}$
  - Are two motifs $\mathbf{w}$ and $\mathbf{w}'$ significantly correlated?
- **Waiting time** till the first occurrence:
  - Is the presence of a motif $\mathbf{w}$ significant?

Models

Two kinds of models:

- **Models for random sequences of letters**
  Problems: count, distance between occurrences, waiting time till 1st occurrence.

Models

Two kinds of models:

- **Models for random sequences of letters**
  Problems: count, distance between occurrences, waiting time till 1st occurrence.
  - **Homogeneous and stationary Markov chain model of order m**
    $\hookrightarrow$ allows to fit the frequency of words of length 1 up to $m+1$,
    $\hookrightarrow$ possibility to consider the coding phase (3-periodicity),
    $\hookrightarrow$ possibility to consider heterogenous chains.

Models

Two kinds of models:

- **Models for random sequences of letters**
  Problems: count, distance between occurrences, waiting time till 1st occurrence.
  - **Homogeneous and stationary Markov chain model of order m**
    $\hookrightarrow$ allows to fit the frequency of words of length 1 up to $m+1$,
    $\hookrightarrow$ possibility to consider the coding phase (3-periodicity),
    $\hookrightarrow$ possibility to consider heterogenous chains.
- **Models for occurrences processes**
  Problems: distance between occurrences, modeling the dependence.

Models

Two kinds of models:

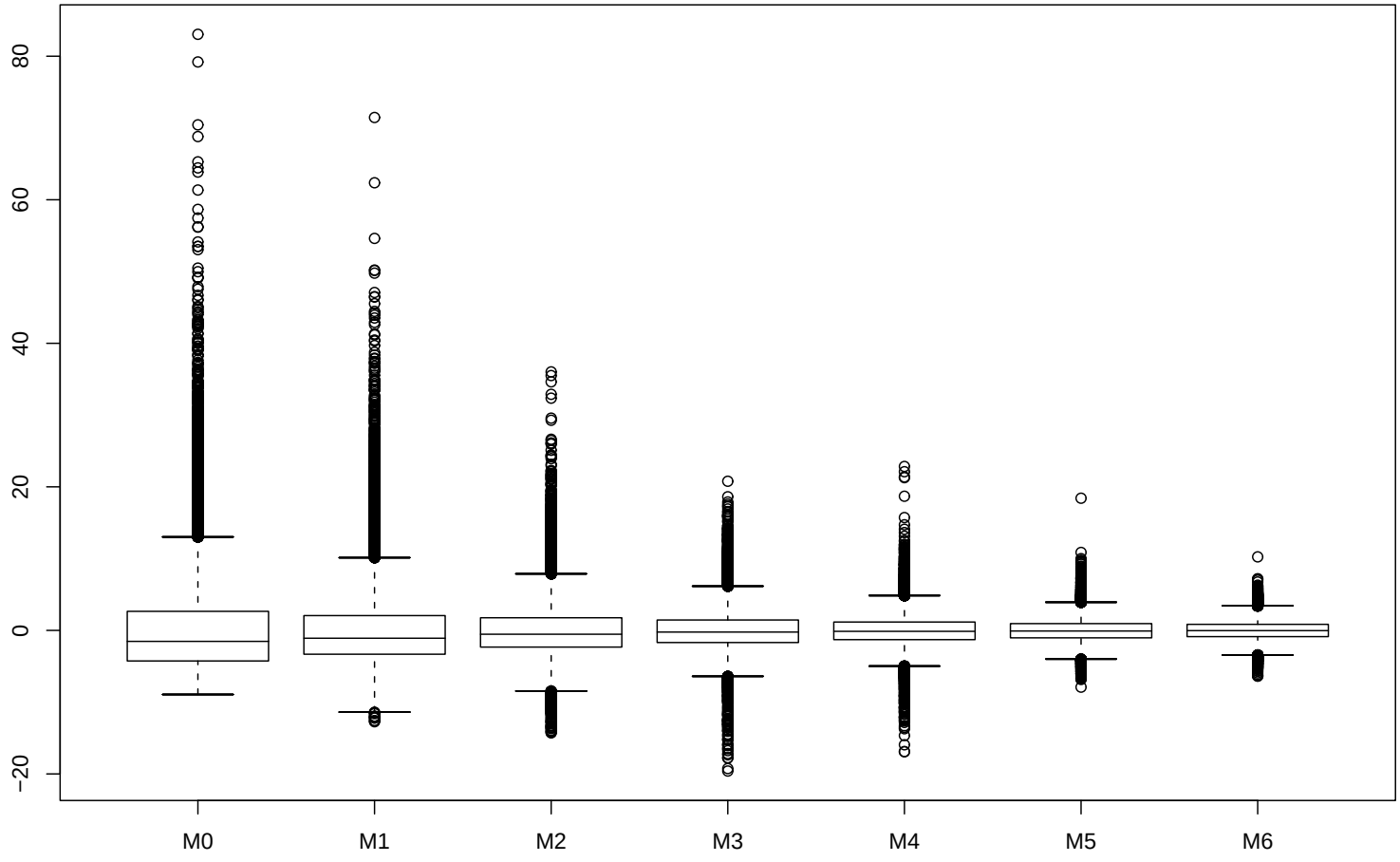
- **Models for random sequences of letters**
  Problems: count, distance between occurrences, waiting time till 1st occurrence.
  - **Homogeneous and stationary Markov chain model of order m**
    $\hookrightarrow$ allows to fit the frequency of words of length 1 up to $m+1$,
    $\hookrightarrow$ possibility to consider the coding phase (3-periodicity),
    $\hookrightarrow$ possibility to consider heterogenous chains.
- **Models for occurrences processes**
  Problems: distance between occurrences, modeling the dependence.
  - **(Compound) Poisson process**

Models

Two kinds of models:

- **Models for random sequences of letters**
  Problems: count, distance between occurrences, waiting time till 1st occurrence.
  - **Homogeneous and stationary Markov chain model of order m**
    $\hookrightarrow$ allows to fit the frequency of words of length 1 up to $m+1$,
    $\hookrightarrow$ possibility to consider the coding phase (3-periodicity),
    $\hookrightarrow$ possibility to consider heterogenous chains.
- **Models for occurrences processes**
  Problems: distance between occurrences, modeling the dependence.
  - **(Compound) Poisson process**
  - **Hawkes process**

Which model to use?



Scores of exceptionality for the 65,536 8-letter words in the *E.coli* backbone.

Influence of the model

	gctggtgg 762 occ.		**ggcgctgg** 828 occ.		**ccggccta** 71 occ.	
Model	$\mathbb{E}(N)$	p-value (rank)	$\mathbb{E}(N)$	p-value (rank)	$\mathbb{E}(N)$	p-value (rank)
M0	85.944	$< 10^{-323}$ (3)	85.524	$< 10^{-323}$ (2)	70.445	0.47 (25608)
M1	84.943	$< 10^{-323}$ (1)	125.919	$< 10^{-323}$ (2)	48.173	10^{-3} (13081)
M2	206.791	$< 10^{-323}$ (1)	255.638	10^{-283} (3)	35.830	10^{-8} (4436)
M3	355.508	$1.4\,10^{-107}$ (5)	441.226	10^{-78} (15)	14.697	10^{-49} (47)
M4	355.312	$2.3\,10^{-116}$ (2)	392.252	10^{-120} (1)	15.341	10^{-46} (21)
M5	420.867	$1.0\,10^{-86}$ (1)	633.453	10^{-22} (24)	27.761	10^{-18} (36)
M6	610.114	$1.5\,10^{-26}$ (3)	812.339	0.16 (14686)	25.777	10^{-26} (4)

Expected counts and p-values (rank) under models Mm, $m = 0, 1, \ldots, 6$, estimated from the *E. coli*'s genome (4 638 858 bps, leading strands).

Overlapping occurrences

Occurrences of words may overlap in DNA sequences (no space between words).

$\Rightarrow$ occurrences are intrinsically localy dependent.

- Occurrences of overlapping words will tend to occur in clumps.
  For instance, they are 3 overlapping occurrences of CAGCAG below :

  TAGACAGATAGACGAT **C**AG**C**AG**C**AGCAG ACAGTAGGCATGA. . .

- On the contrary, occurrences of non-overlapping words will never overlap.

Overlapping occurrences

Occurrences of words may overlap in DNA sequences (no space between words).

$\Rightarrow$ occurrences are intrinsically localy dependent.

- Occurrences of overlapping words will tend to occur in clumps.
  For instance, they are 3 overlapping occurrences of CAGCAG below :

  TAGACAGATAGACGAT **C**AG**C**AG**C**AGCAG ACAGTAGGCATGA. . .

- On the contrary, occurrences of non-overlapping words will never overlap.

All results on word occurrences will depend on the overlapping structure of the words.
Classically, this structure is described thanks to the periods of a word:

$$p \text{ is a period of } \mathbf{w} := w_1 w_2 \cdots w_h \text{ iff } w_i = w_{i+p}, \quad \forall i$$

It means that 2 occurrences of $\mathbf{w}$ can overlap on $h - p$ letters.

Counting motifs

Problem

Let $N(\mathbf{w})$ be the number of occurrences of the word $\mathbf{w} := w_1 w_2 \cdots w_h$ in the sequence $X_1 X_2 X_3 \cdots X_\ell$ (model M1):

$$N(\mathbf{w}) = \sum_{i=1}^{\ell-h+1} Y_i(\mathbf{w})$$

where

$$Y_i(\mathbf{w}) = \mathbf{1}\{\mathbf{w} \text{ starts at position } i\} \sim \mathcal{B}(\mu(\mathbf{w}))$$

and

$$\mu(\mathbf{w}) = \mu(w_1) \prod_{j=1}^{h-1} \pi(w_j, w_{j+1}).$$

Question: how to decide if $N^{\text{obs}}(\mathbf{w})$ is significantly unexpected (under model M1)?

Problem

Let $N(\mathbf{w})$ be the number of occurrences of the word $\mathbf{w} := w_1 w_2 \cdots w_h$ in the sequence $X_1 X_2 X_3 \cdots X_\ell$ (model M1):

$$N(\mathbf{w}) = \sum_{i=1}^{\ell-h+1} Y_i(\mathbf{w})$$

where

$$Y_i(\mathbf{w}) = \mathbf{1}\{\mathbf{w} \text{ starts at position } i\} \sim \mathcal{B}(\mu(\mathbf{w}))$$

and

$$\mu(\mathbf{w}) = \mu(w_1) \prod_{j=1}^{h-1} \pi(w_j, w_{j+1}).$$

Question: how to decide if $N^{\text{obs}}(\mathbf{w})$ is significantly unexpected (under model M1)?

Solution: to calculate the p-value $\mathbb{P}(N(\mathbf{w}) \geq N^{\text{obs}}(\mathbf{w}))$

$\Rightarrow$ What is the distribution of $N(\mathbf{w})$?

Overview

If the Y_i's were independent:

$$\sum_{i=1}^{\ell} Y_i(\mathbf{w}) \sim \mathcal{B}(\ell, \mu(\mathbf{w})) \text{ approximated by } \begin{cases} \mathcal{P}(\ell\mu(\mathbf{w})) & \text{if } \ell\mu(\mathbf{w}) = \text{O(1)} \\ \mathcal{N}(\ell\mu(\mathbf{w}), \ell\mu(\mathbf{w})(1-\mu(\mathbf{w}))) & \text{if } \ell\mu(\mathbf{w}) \sim \infty \end{cases}$$

Probabilistic overview

Since the Y_i's are locally dependent:

$\mathcal{PC}(\Lambda)$

$$\sum_{i=1}^{\ell} Y_i(\mathbf{w}) \sim \mathcal{B}(\ell, \mu(\mathbf{w})) \text{ approximated by } \begin{cases} \mathcal{P}(\ell\mu(\mathbf{w})) & \text{if } \ell\mu(\mathbf{w}) = \text{O(1)} \\ \mathcal{N}(\ell\mu(\mathbf{w}), \ell\mu(\mathbf{w})(1-\mu(\mathbf{w}))) & \text{if } \ell\mu(\mathbf{w}) \sim \infty \end{cases}$$

exact distribution

v^2

- **Exact distribution**: generating function or recursive formula
  [Robin & Daudin (99)], [Régnier (00)], [Stefanov (03)]
- **Asymptotic normality**: CLT for Markov chains.
- **Variance**: combinatorics on words.
  [Kleffe & Borodowsky (92)].
- **Compound Poisson approximation**: Chen-Stein method + combinatorics on words
  [S. (95)], [Roquain and S. (07)]

Probabilistic overview

Since the Y_i's are locally dependent:

$$\sum_{i=1}^{\ell} Y_i(\mathbf{w}) \sim \mathcal{B}(\ell, \mu(\mathbf{w})) \text{ approximated by } \begin{cases} \mathcal{P}(\ell\mu(\mathbf{w})) & \text{if } \ell\mu(\mathbf{w}) = \text{O(1)} \\ \mathcal{N}(\ell\mu(\mathbf{w}), \ell\mu(\mathbf{w})(1-\mu(\mathbf{w}))) & \text{if } \ell\mu(\mathbf{w}) \sim \infty \end{cases}$$

exact distribution

$\mathcal{PC}(\Lambda)$

v^2

- **Exact distribution**: generating function or recursive formula
  [Robin & Daudin (99)], [Régnier (00)], [Stefanov (03)]
- **Asymptotic normality**: CLT for Markov chains.
- **Variance**: combinatorics on words.
  [Kleffe & Borodowsky (92)].
- **Compound Poisson approximation**: Chen-Stein method + combinatorics on words
  [S. (95)], [Roquain and S. (07)]

p-values of exceptional words can also be approximated by using a **large deviation** principle ([Nuel (04)]) or a **moderate deviation** principle ([Behrens & Löwe (09)])

Statistical overview

If we now estimate the Markov chain parameters:

$\mathcal{PC}(\widehat{\Lambda})$

$$\sum_{i=1}^{\ell} Y_i(\mathbf{w}) \sim \mathcal{B}(\ell, \mu(\mathbf{w}))$$ approximated by $$\begin{cases} \mathcal{P}(\ell\widehat{\mu}(\mathbf{w})) & \text{if } \ell\widehat{\mu}(\mathbf{w}) = \text{O(1)} \\ \mathcal{N}(\ell\widehat{\mu}(\mathbf{w}), \ell\widehat{\mu}(\mathbf{w})(1-\widehat{\mu}(\mathbf{w}))) & \text{if } \ell\widehat{\mu}(\mathbf{w}) \sim \infty \end{cases}$$

exact distribution

$\widehat{v}^2$

$\widehat{\sigma}^2$

- **Asymptotic normality + variance**: δ-method, conditional approach, combinatorics [Lundstrum (90)], [Prum, Rodolphe, de Turckheim (95)]
- **Compound Poisson approximation**: Chen-Stein method + combinatoris on words [S. (95)]

Statistical overview

If we now estimate the Markov chain parameters:

$\mathcal{PC}(\widehat{\Lambda})$

$$\sum_{i=1}^{\ell} Y_i(\mathbf{w}) \sim \mathcal{B}(\ell, \mu(\mathbf{w}))$$ approximated by $$\begin{cases} \mathcal{P}(\ell\widehat{\mu}(\mathbf{w})) & \text{if } \ell\widehat{\mu}(\mathbf{w}) = \text{O(1)} \\ \mathcal{N}(\ell\widehat{\mu}(\mathbf{w}), \ell\widehat{\mu}(\mathbf{w})(1-\widehat{\mu}(\mathbf{w}))) & \text{if } \ell\widehat{\mu}(\mathbf{w}) \sim \infty \end{cases}$$

exact distribution

$\widehat{v}^2$

$\widehat{\sigma}^2$

- **Asymptotic normality + variance**: δ-method, conditional approach, combinatorics [Lundstrum (90)], [Prum, Rodolphe, de Turckheim (95)]
- **Compound Poisson approximation**: Chen-Stein method + combinatoris on words [S. (95)]

Both approximations can deal with a (finite) set of words and any order of Markov models.

Open problem: complex motifs (1/2)

Structured motifs

$$\mathbf{m} = \mathbf{w}_1(d_1 : D_1)\mathbf{w}_2(d_2 : D_2)\mathbf{w}_3 \dots \mathbf{w}_{b-1}(d_{b-1} : D_{b-1})\mathbf{w}_b.$$

Previous results need

- to explicitely enumerate the set of motifs and
- to describe all the possible overlaps between these motifs.

Here it is not possible due to the spacers.

The structure of the motif should then be taken into account.

Open problem: complex motifs (2/2)

Position Frequency Matrix

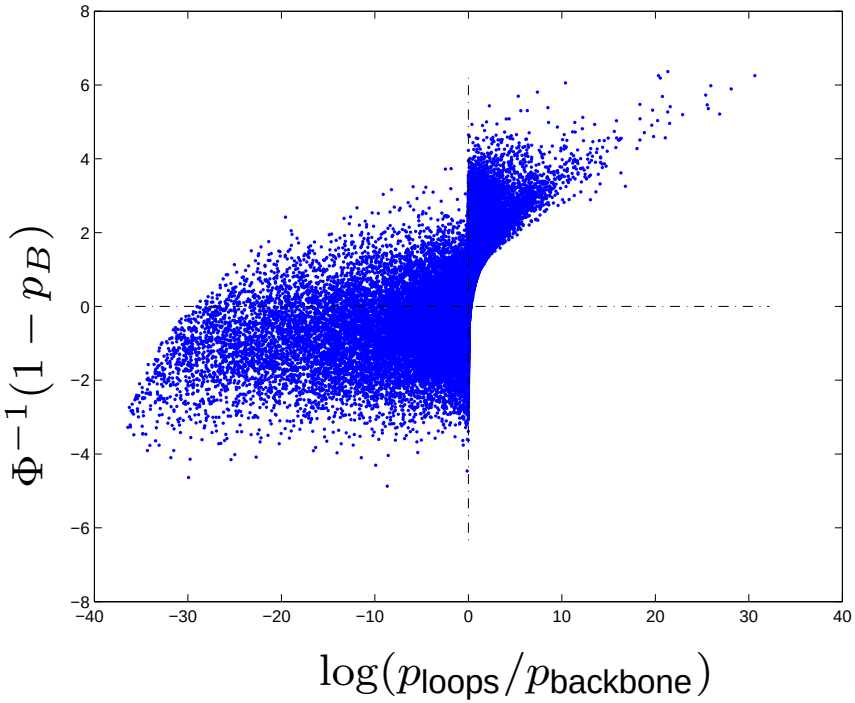
$$\begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathtt{1} & \mathtt{2} & \mathtt{3} & \mathtt{4} & \mathtt{5} \end{array} \\ \begin{pmatrix} 0.7 & 1 & 0.3 & 0 & 0.5 \\ 0.1 & 0 & 0.3 & 1 & 0.3 \\ 0 & 0 & 0.2 & 0 & 0 \\ 0.2 & 0 & 0.2 & 0 & 0.2 \end{pmatrix} \end{array}$$

How to define occurrences? Should we count weighted occurrences or occurrences with sufficiently high probability?

A compound Poisson approximation has been proposed in the later case with the problem of choosing the occurrence probability threshold [Pape, Rahmann, Sun, Vingron (08)]

Detecting specificities among sequences

- Some DNA motifs should be specific to part of a genome (domain, promotor regions, etc.).
- How to test if a given motif is more over-represented in sequence S_1 than in S_2?
- Basic idea: to compare both p-values e.g. $\log(p_1/p_2)$ but this is not satisfactory



Transformed binomial p-value $\Phi^{-1}(1-p_B)$ (y-axis) versus log ratio between the two p-values associated with the exceptionality of the motif in each sequence (x-axis) under model M1 for all octamers in the *E. coli* backbone/variable segments comparison

Need for a statistical test

Robin, Vandewalle, S. (07)

Model: 2 independent Poisson processes, N_1 and N_2

- with intensity $\lambda_1 = k_1 \times \mu_1$ along S_1 (length ℓ_1) $\quad \Rightarrow \mathbb{E}N_1 = k_1 \times \ell_1\mu_1$
- with intensity $\lambda_2 = k_2 \times \mu_2$ along S_1 (length ℓ_2) $\quad \Rightarrow \mathbb{E}N_2 = k_2 \times \ell_2\mu_2$

$\longrightarrow k_i$ stands for the exceptionality coefficient
$\longrightarrow \mu_i$ is the occurrence probability of the motif

Null hypothesis: $\mathbf{H}_0 = \{k_1 = k_2\}$

Test statistics: $N_1 \mid N_+ \sim \mathcal{B}(N_+, \pi)$ with

$$N_+ := N_1 + N_2 \quad \text{and} \quad \pi = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{(k_1/k_2)\ell_1\mu_1}{(k_1/k_2)\ell_1\mu_1 + \ell_2\mu_2}.$$

For long sequences, a LRT can also be performed.

Motifs with significant skew

- Some DNA motifs are specific to one DNA strand, say the leading strand.
- Both strands are dependent by definition.
- The quantity of interest is the skew defined by

$$B(\mathbf{w}) := \frac{N(\mathbf{w})}{\overline{N}(\mathbf{w})} = \frac{N(\mathbf{w})}{N(\overline{\mathbf{w}})}$$

  and one wants to evaluate the p-value $\mathbb{P}(B(\mathbf{w}) \geq B^{\text{obs}}(\mathbf{w}))$
- If the counts are asymptotically Gaussian, it is easy to derive an approximate p-value [Halpern *et al.* (07)]
- **Open problem**: what about if the counts are (compound) Poisson?

Waiting time and inter-arrival time

For simple motifs

The generating function of the following waiting times are known in Markovian sequences [Robin and Daudin (01)], [Stefanov (03)]:

- $T_{\alpha,\mathbf{w}}$, the waiting time to reach pattern $\mathbf{w}$ from state α
- $T_{\mathbf{w},\mathbf{w}'}$, the waiting time to reach pattern $\mathbf{w}'$ from pattern $\mathbf{w}$
- idem with competing occurrences from a pattern family $\mathcal{W}$: $X_{\alpha,\mathbf{w}}(\mathcal{W})$ and $X_{\mathbf{w},\mathbf{w}'}(\mathcal{W})$

Note the duality principle: $\mathbb{P}(N(\mathbf{w}) \geq n) = \mathbb{P}(T_n(\mathbf{w}) \leq \ell)$ where

$$T_n(\mathbf{w}) \stackrel{\mathcal{L}}{=} T_{\alpha,\mathbf{w}} + \sum_{j=1}^{n-1} T_{\mathbf{w},\mathbf{w}}^{(j)}$$

For structured motifs

$$\mathbf{m} = \mathbf{w}_1(d_1 : D_1)\mathbf{w}_2(d_2 : D_2)\mathbf{w}_3 \ldots \mathbf{w}_{b-1}(d_{b-1} : D_{b-1})\mathbf{w}_b.$$

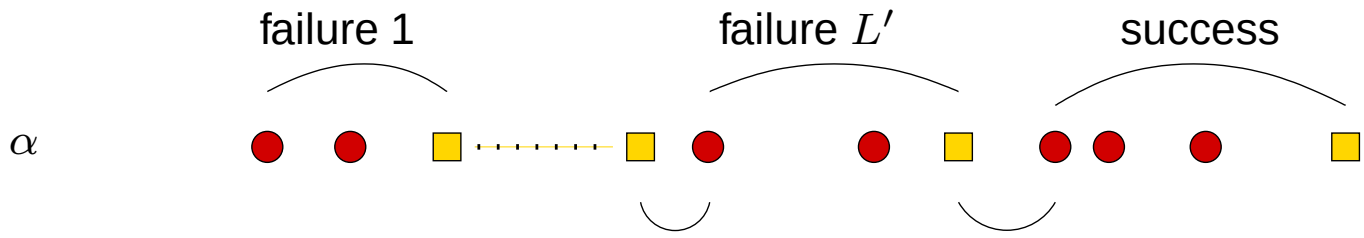
What is the distribution of $\tau_{\mathbf{m}}$, the waiting time to reach $\mathbf{m}$?

Random sum approach for 2 boxes

[Stefanov, Robin, S. (06)]

- Restriction: $\mathbf{w}_2$ should not occur in between the 2 boxes.
- The explicit formula for the pgf of $\tau_{\mathbf{m}}$ is given thanks to the following decomposition:

$$\tau_{\mathbf{m}} \stackrel{\mathcal{D}}{=} T_{\alpha,\mathbf{w}_1} + \sum_{b=1}^{L'} \left(\underbrace{\sum_{a=1}^{L_1} X^{(ab)}_{\mathbf{w}_1,\mathbf{w}_1} + F^{(b)}_{\mathbf{w}_1,\mathbf{w}_2}}_{\stackrel{\mathcal{D}}{=} T_{\mathbf{w}_1,\mathbf{w}_2} \mid \text{failure}} + T^{(b)}_{\mathbf{w}_2,\mathbf{w}_1} \right) + \underbrace{\sum_{c=1}^{L_2} X^{(c)}_{\mathbf{w}_1,\mathbf{w}_1} + S_{\mathbf{w}_1,\mathbf{w}_2}}_{\stackrel{\mathcal{D}}{=} T_{\mathbf{w}_1,\mathbf{w}_2} \mid \text{success}},$$



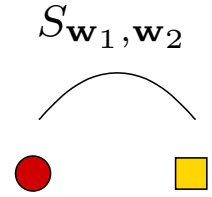
L_1, L_2 and L' are independent geometric variables.

Random sum approach for 2 boxes

[Stefanov, Robin, S. (06)]

- Restriction: $\mathbf{w}_2$ should not occur in between the 2 boxes.
- The explicit formula for the pgf of $\tau_{\mathbf{m}}$ is given thanks to the following decomposition:

$$\tau_{\mathbf{m}} \stackrel{\mathcal{D}}{=} T_{\alpha,\mathbf{w}_1} + \sum_{b=1}^{L'} \left(\underbrace{\sum_{a=1}^{L_1} X^{(ab)}_{\mathbf{w}_1,\mathbf{w}_1} + F^{(b)}_{\mathbf{w}_1,\mathbf{w}_2}}_{\stackrel{\mathcal{D}}{=} T_{\mathbf{w}_1,\mathbf{w}_2} \mid \text{failure}} + T^{(b)}_{\mathbf{w}_2,\mathbf{w}_1} \right) + \underbrace{\sum_{c=1}^{L_2} X^{(c)}_{\mathbf{w}_1,\mathbf{w}_1} + S_{\mathbf{w}_1,\mathbf{w}_2}}_{\stackrel{\mathcal{D}}{=} T_{\mathbf{w}_1,\mathbf{w}_2} \mid \text{success}},$$



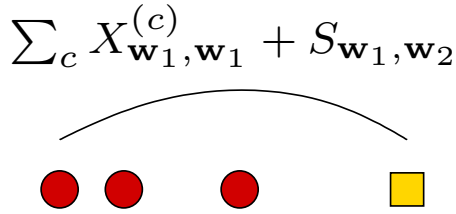
L_1, L_2 and L' are independent geometric variables.

Random sum approach for 2 boxes

[Stefanov, Robin, S. (06)]

- Restriction: $\mathbf{w}_2$ should not occur in between the 2 boxes.
- The explicit formula for the pgf of $\tau_{\mathbf{m}}$ is given thanks to the following decomposition:

$$\tau_{\mathbf{m}} \stackrel{\mathcal{D}}{=} T_{\alpha,\mathbf{w}_1} + \sum_{b=1}^{L'} \left(\underbrace{\sum_{a=1}^{L_1} X^{(ab)}_{\mathbf{w}_1,\mathbf{w}_1} + F^{(b)}_{\mathbf{w}_1,\mathbf{w}_2}}_{\stackrel{\mathcal{D}}{=} T_{\mathbf{w}_1,\mathbf{w}_2} \mid \text{failure}} + T^{(b)}_{\mathbf{w}_2,\mathbf{w}_1} \right) + \underbrace{\sum_{c=1}^{L_2} X^{(c)}_{\mathbf{w}_1,\mathbf{w}_1} + S_{\mathbf{w}_1,\mathbf{w}_2}}_{\stackrel{\mathcal{D}}{=} T_{\mathbf{w}_1,\mathbf{w}_2} \mid \text{success}},$$



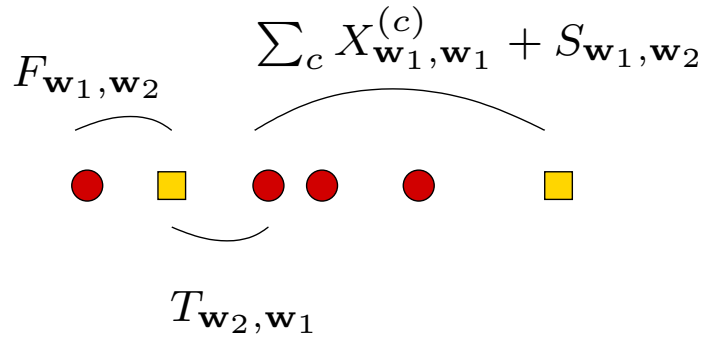
L_1, L_2 and L' are independent geometric variables.

Random sum approach for 2 boxes

[Stefanov, Robin, S. (06)]

- Restriction: $\mathbf{w}_2$ should not occur in between the 2 boxes.
- The explicit formula for the pgf of $\tau_{\mathbf{m}}$ is given thanks to the following decomposition:

$$\tau_{\mathbf{m}} \stackrel{\mathcal{D}}{=} T_{\alpha,\mathbf{w}_1} + \sum_{b=1}^{L'} \left(\underbrace{\sum_{a=1}^{L_1} X^{(ab)}_{\mathbf{w}_1,\mathbf{w}_1} + F^{(b)}_{\mathbf{w}_1,\mathbf{w}_2}}_{\stackrel{\mathcal{D}}{=} T_{\mathbf{w}_1,\mathbf{w}_2} \mid \text{failure}} + T^{(b)}_{\mathbf{w}_2,\mathbf{w}_1} \right) + \underbrace{\sum_{c=1}^{L_2} X^{(c)}_{\mathbf{w}_1,\mathbf{w}_1} + S_{\mathbf{w}_1,\mathbf{w}_2}}_{\stackrel{\mathcal{D}}{=} T_{\mathbf{w}_1,\mathbf{w}_2} \mid \text{success}},$$



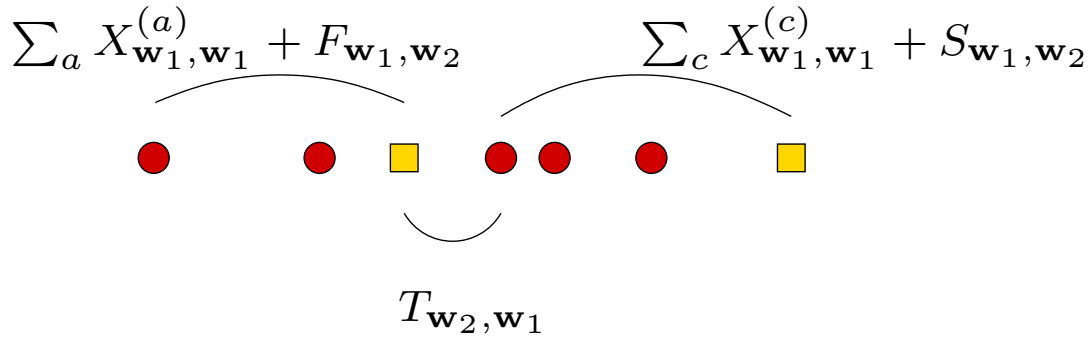
L_1, L_2 and L' are independent geometric variables.

Random sum approach for 2 boxes

[Stefanov, Robin, S. (06)]

- Restriction: $\mathbf{w}_2$ should not occur in between the 2 boxes.
- The explicit formula for the pgf of $\tau_{\mathbf{m}}$ is given thanks to the following decomposition:

$$\tau_{\mathbf{m}} \stackrel{\mathcal{D}}{=} T_{\alpha,\mathbf{w}_1} + \sum_{b=1}^{L'} \left(\underbrace{\sum_{a=1}^{L_1} X^{(ab)}_{\mathbf{w}_1,\mathbf{w}_1} + F^{(b)}_{\mathbf{w}_1,\mathbf{w}_2}}_{\stackrel{\mathcal{D}}{=} T_{\mathbf{w}_1,\mathbf{w}_2} \mid \text{failure}} + T^{(b)}_{\mathbf{w}_2,\mathbf{w}_1} \right) + \underbrace{\sum_{c=1}^{L_2} X^{(c)}_{\mathbf{w}_1,\mathbf{w}_1} + S_{\mathbf{w}_1,\mathbf{w}_2}}_{\stackrel{\mathcal{D}}{=} T_{\mathbf{w}_1,\mathbf{w}_2} \mid \text{success}},$$



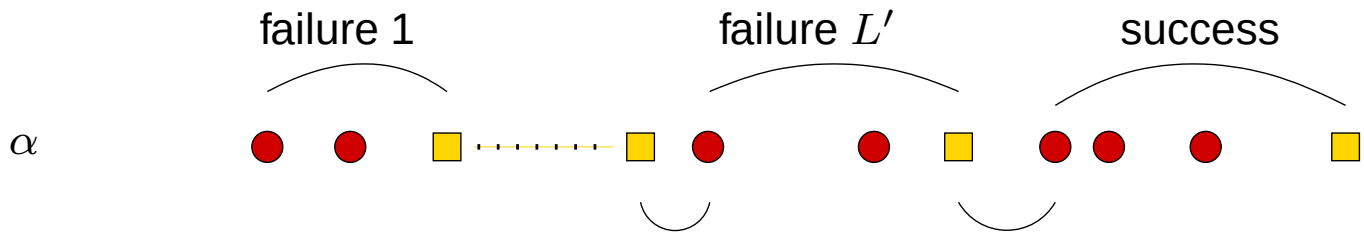
L_1, L_2 and L' are independent geometric variables.

Random sum approach for 2 boxes

[Stefanov, Robin, S. (06)]

- Restriction: $\mathbf{w}_2$ should not occur in between the 2 boxes.
- The explicit formula for the pgf of $\tau_{\mathbf{m}}$ is given thanks to the following decomposition:

$$\tau_{\mathbf{m}} \stackrel{\mathcal{D}}{=} T_{\alpha,\mathbf{w}_1} + \sum_{b=1}^{L'} \left(\underbrace{\sum_{a=1}^{L_1} X^{(ab)}_{\mathbf{w}_1,\mathbf{w}_1} + F^{(b)}_{\mathbf{w}_1,\mathbf{w}_2}}_{\stackrel{\mathcal{D}}{=} T_{\mathbf{w}_1,\mathbf{w}_2} \mid \text{failure}} + T^{(b)}_{\mathbf{w}_2,\mathbf{w}_1} \right) + \underbrace{\sum_{c=1}^{L_2} X^{(c)}_{\mathbf{w}_1,\mathbf{w}_1} + S_{\mathbf{w}_1,\mathbf{w}_2}}_{\stackrel{\mathcal{D}}{=} T_{\mathbf{w}_1,\mathbf{w}_2} \mid \text{success}},$$



L_1, L_2 and L' are independent geometric variables.

General case: semi-Markov process embedding

$$\mathbf{m} = \mathbf{w}_1(d_1 : D_1)\mathbf{w}_2(d_2 : D_2)\mathbf{w}_3 \ldots \mathbf{w}_{b-1}(d_{b-1} : D_{b-1})\mathbf{w}_b.$$

- Principle: to build a semi-Markov process Y_t such that

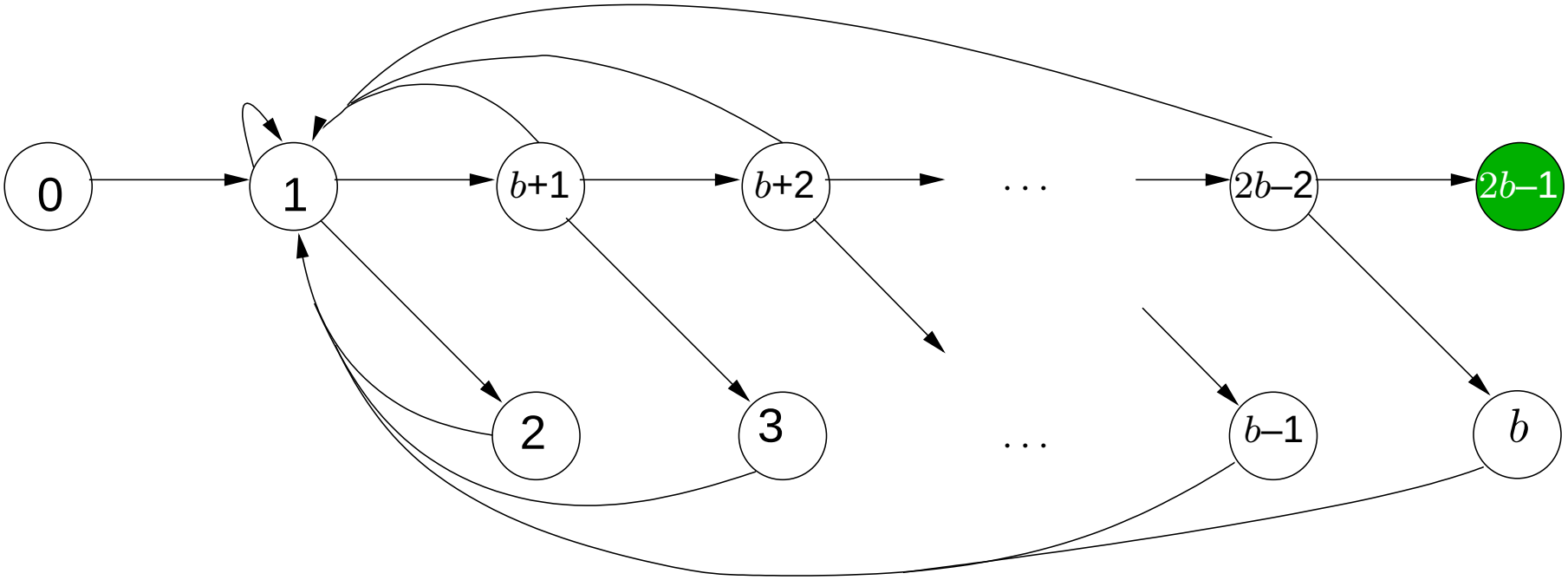
$$\tau_{\mathbf{m}} \stackrel{\mathcal{D}}{=} \text{first passage to a particular state in } Y_t$$

- Restrictions:
  - Pattern $\mathbf{w}_1$ appears only once in the structured motif $\mathbf{m}$;
  - For each $i = 2, \ldots, b$, pattern $\mathbf{w}_i$ appears only once in the sub-structured motif $\mathbf{w}_{i-1}(d_{i-1} : D_{i-1})\mathbf{w}_i$.

Semi-Markov process embedding (1/3)

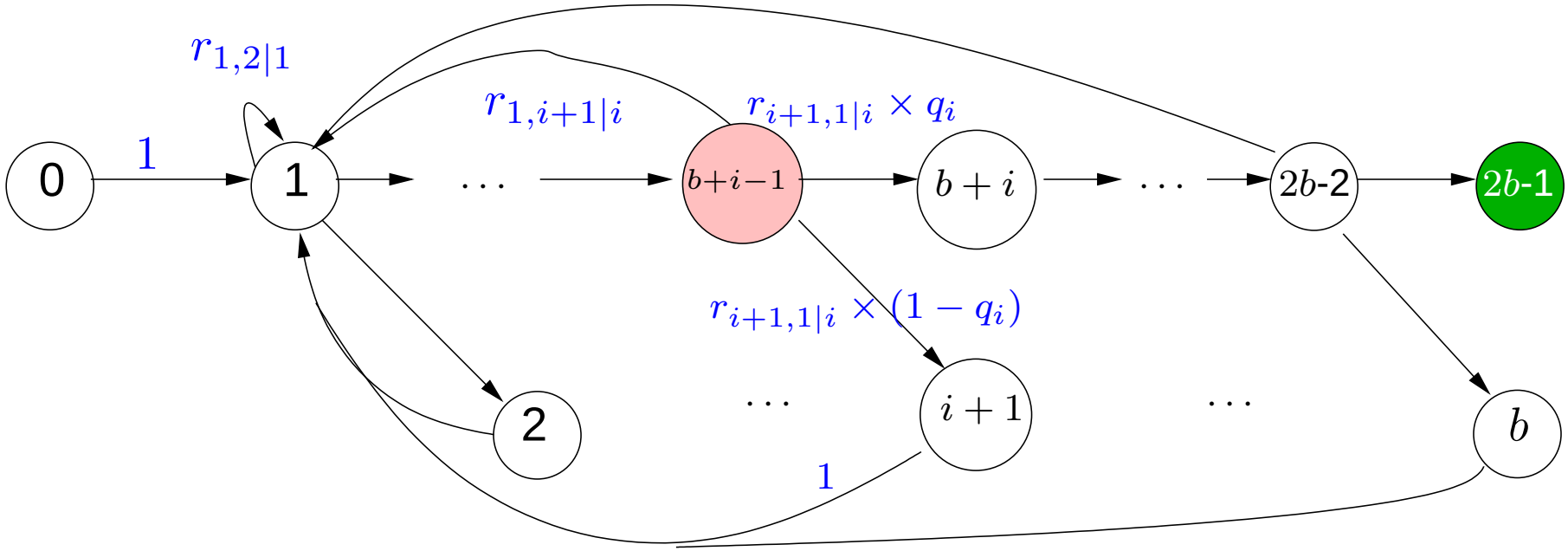
Let Y_t be the following semi-Markov process with states $0, 1, \ldots, 2b-1$:

- State interpretation:
  Set $\mathbf{m}_i = \mathbf{w}_1(d_1 : D_1)\mathbf{w}_2(d_2 : D_2)\mathbf{w}_3 \ldots \mathbf{w}_{i-1}(d_{i-1} : D_{i-1})\mathbf{w}_i$.
  state 0: neutral initial state
  state i: pattern $\mathbf{w}_i$ is reached without achieving $\mathbf{m}_i$, $i = 1, \ldots, b$,
  state $b+i$: sub-structured motif $\mathbf{m}_{i+1}$ is achieved, $i = 1, 2, \ldots, b-1$.
- Transition diagram:

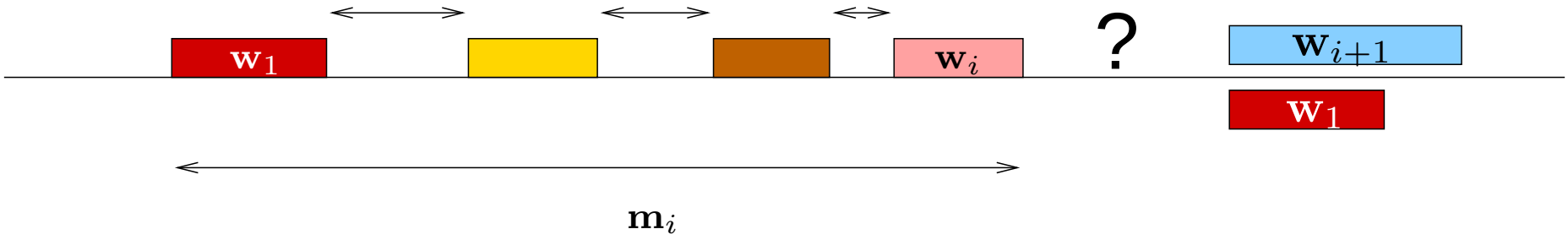


Semi-Markov process embedding (2/3)

- Transition probabilities:

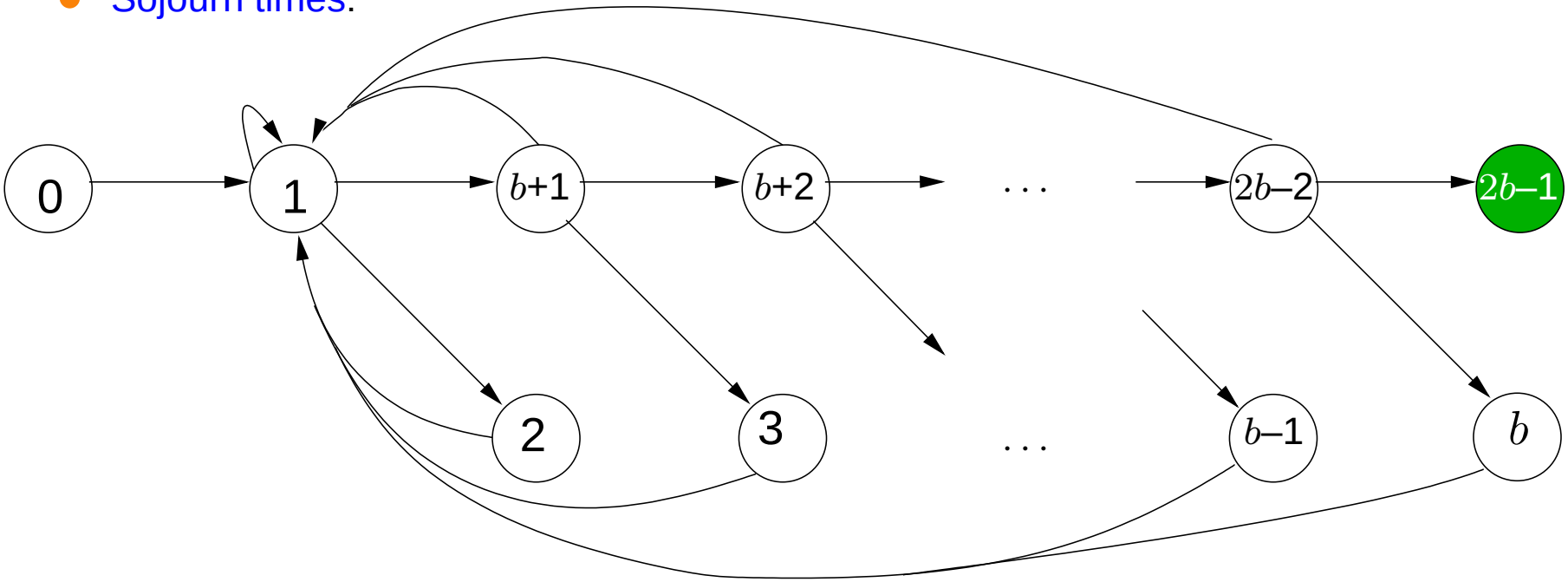


where $r_{i,j|k} = \mathbb{P}$(the first pattern from $\{\mathbf{w}_i, \mathbf{w}_j\}$ to be reach from $\mathbf{w}_k$ is $\mathbf{w}_i$)
and $q_i = \mathbb{P}(\mathbf{w}_{i+1}$ achieves motif $\mathbf{m}_{i+1})$



Semi-Markov process embedding (3/3)

- Sojourn times:



$$H_{0,1} \stackrel{\mathcal{D}}{=} T_{\alpha,\mathbf{w}_1} \qquad H_{i,1} \stackrel{\mathcal{D}}{=} T_{\mathbf{w}_1|\mathbf{w}_i}$$

$$H_{1,1} \stackrel{\mathcal{D}}{=} T_{\mathbf{w}_1,\mathbf{w}_2|\mathbf{w}_1} \qquad H_{b+i,1} \stackrel{\mathcal{D}}{=} T_{\mathbf{w}_1,\mathbf{w}_{i+2}|\mathbf{w}_{i+1}}$$

$$H_{1,2} \stackrel{\mathcal{D}}{=} F_{\mathbf{w}_2,\mathbf{w}_1|\mathbf{w}_1} \qquad H_{b+i,i+2} \stackrel{\mathcal{D}}{=} F_{\mathbf{w}_{i+2},\mathbf{w}_1|\mathbf{w}_{i+1}}$$

$$H_{1,b+1} \stackrel{\mathcal{D}}{=} S_{\mathbf{w}_2,\mathbf{w}_1|\mathbf{w}_1} \qquad H_{b+i,b+i+1} \stackrel{\mathcal{D}}{=} S_{\mathbf{w}_{i+2},\mathbf{w}_1|\mathbf{w}_{i+1}}$$

P.g.f. of the waiting time

- Let $S_{i,j}$ be the number of transitions from i to j until absorption of the semi-Markov process Y_t in state $2b-1$.
  The joint p.g.f. of $\mathbf{S} = (S_{i,j})_{ij}$, given initial state k, is defined like

$$G_{\mathbf{S}}^{(k)}(\mathbf{a}) := \mathbb{E}\left(\prod_{i,j=0}^{2b-1} a_{i,j}^{S_{i,j}} \mid Y_0 = k \right) \quad , \quad \mathbf{a} = (a_{i,j})_{ij}$$

- From Iosifescu (80),

$$(G_{\mathbf{S}}^{(k)}(\mathbf{a}))_k = (\mathbf{I} - \mathbf{V}(\mathbf{a}))^{-1}\mathbf{U}(\mathbf{a})$$

  where $v_{i,j} = p_{i,j}a_{i,j}$ and $u_i = p_{i,2b-1}a_{i,2b-1}$.

- From Stefanov (00),

$$G_{\tau_{\mathbf{m}}}(t) = G_{\mathbf{S}}^{(0)}(\mathbf{h}(t)),$$

  where $\mathbf{h}(t)$ is the matrix whose $h_{i,j}(t)$ is the p.g.f. of $H_{i,j}$.

Finally . . .

For 3 boxes, for instance, we get:

$$G_{\tau_{\mathbf{m}}}(t) = \frac{r_{2,1|1} q_{S,1} r_{3,1|2} q_{S,2} G_{T_1^{(s)}}(t) G_{S_{2,1|1}}(t) G_{S_{3,1|2}}(t)}{1 - r_{1,2|1} G_{X_{1,2|1}} - r_{2,1|1} \bar{q}_{S,1} G_{F_{2,1|1}} G_{T_{1|2}} - r_{2,1|1} q_{S,1} G_{S_{2,1|1}}(t) A(t)},$$

with

$$A(t) = r_{1,3|2} G_{X_{1,3|2}}(t) + r_{3,1|2} \bar{q}_{S,2} G_{T_{1|3}}(t) G_{F_{3,1|2}}(t).$$

In practice, . . .

Modeling occurrence dependencies

Present problem

Occurrences of a motif may be influenced by, or correlated to occurrences of other motifs (or itself) because these motifs are part of a common biological mechanism.

Occurrences of genes are associated to previous occurrences of promoters or transcription factor binding sites.

Present problem

Occurrences of a motif may be influenced by, or correlated to occurrences of other motifs (or itself) because these motifs are part of a common biological mechanism.

Occurrences of genes are associated to previous occurrences of promoters or transcription factor binding sites.

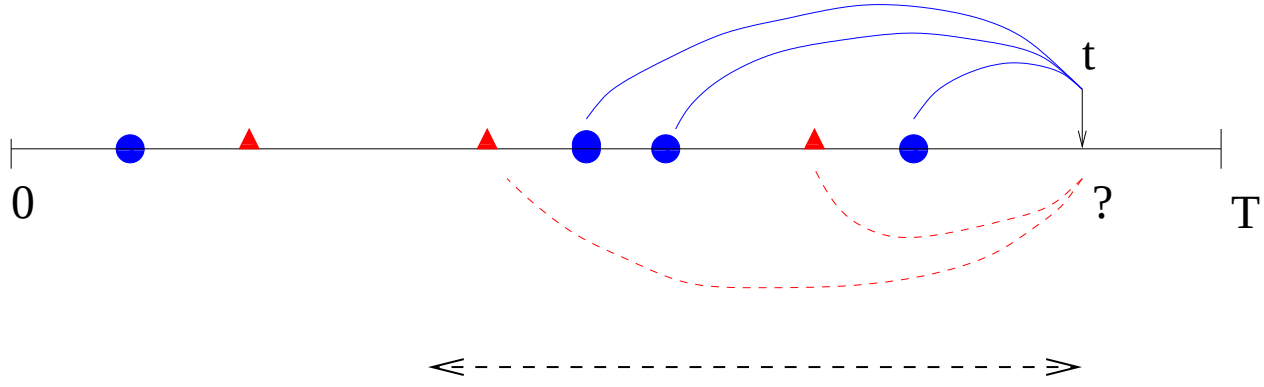
Protein interactions can induce positive or negative constraints on the spacing between motif occurrences

Can we find favored or avoided distances between occurrences of different objects?
This would help to understand various biological mechanisms.

Model: Hawkes process

We consider the ordered occurrences (U_i) of $\mathbf{u}$ (word, gene, etc.)

We model the intensity $\lambda(t)$ of the point process such that an occurrence of $\mathbf{u}$ can be influenced by a previous occurrence of $\mathbf{u}$ and/or a previous occurrence of $\mathbf{v}$

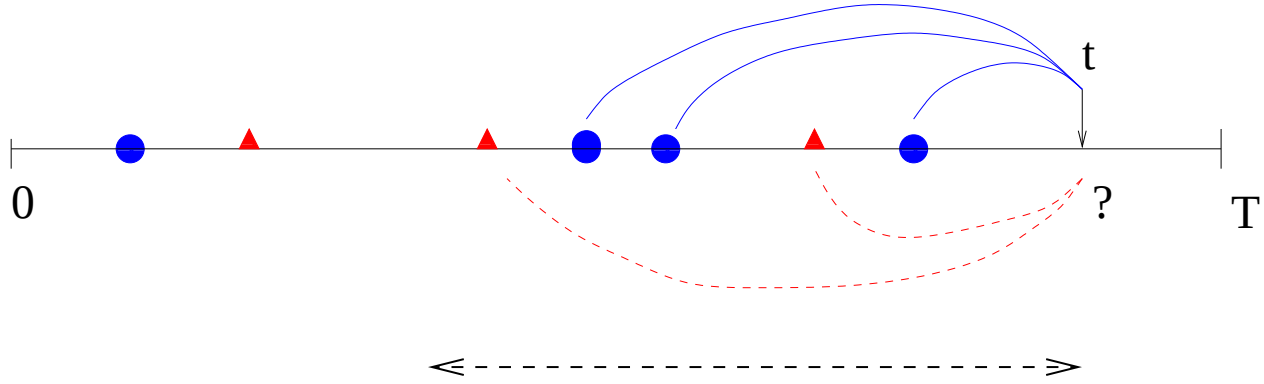


D: bounded memory

Model: Hawkes process

We consider the ordered occurrences (U_i) of $\mathbf{u}$ (word, gene, etc.)

We model the intensity $\lambda(t)$ of the point process such that an occurrence of $\mathbf{u}$ can be influenced by a previous occurrence of $\mathbf{u}$ and/or a previous occurrence of $\mathbf{v}$



D: bounded memory

Hawkes model:

$$\begin{aligned}\lambda(t) &= \mu + \int_{t-D}^{t} g(t-s)\, dU_s + \int_{t-D'}^{t} h(t-s)\, dV_s \\ &= \mu + \sum_{t-D \le U_i < t}^{t} g(t-U_i) + \sum_{t-D' \le V_j < t}^{t} h(t-V_j)\end{aligned}$$

Interpretation of the profiles g and h

$$\lambda(t) = \mu + \sum_{t-D \leq U_i < t} g(t - U_i) + \sum_{t-D' \leq V_j < t} h(t - V_j)$$

If $g \equiv h \equiv 0$: occ. of u do not depend on each other neither on occ. of v.

If $g \not\equiv 0$: occ. of u depend on each other (auto-exciting Poisson proc.).

If $h \not\equiv 0$: occ. of u depend on occ. of v (doubly Poisson proc.).

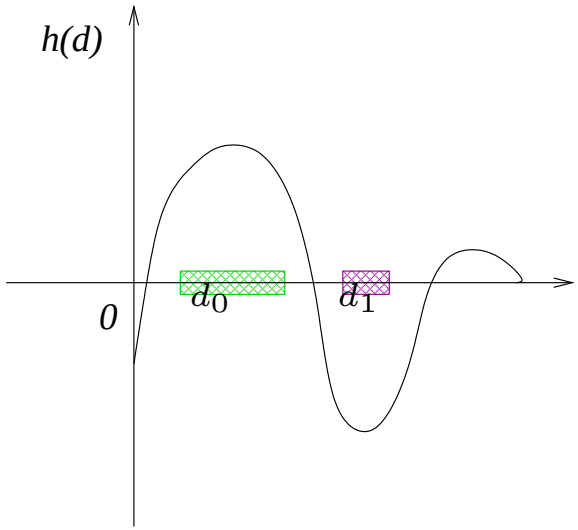
Interpretation of the profiles g and h

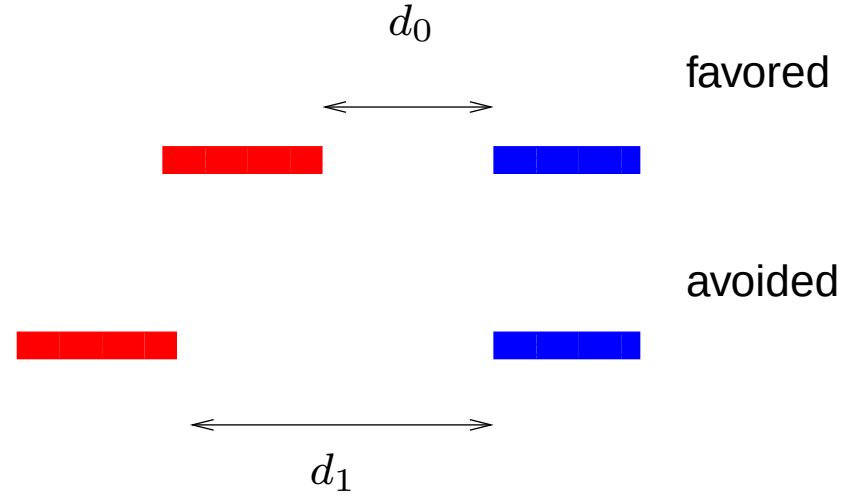
$$\lambda(t) = \mu + \sum_{t-D \leq U_i < t} g(t - U_i) + \sum_{t-D' \leq V_j < t} h(t - V_j)$$

If $g \equiv h \equiv 0$: occ. of u do not depend on each other neither on occ. of v.

If $g \not\equiv 0$: occ. of u depend on each other (auto-exciting Poisson proc.).

If $h \not\equiv 0$: occ. of u depend on occ. of v (doubly Poisson proc.).





Aim

The question is then, given the occurrences (U_i) and (V_j) along $[0, T]$, to estimate the parameter μ and the profiles g and h appearing in

$$\lambda(t) = \mu + \sum_{t-D \leq U_i < t} g(t - U_i) + \sum_{t-D' \leq V_j < t} h(t - V_j)$$

Estimation and model selection (*FADO*)

Gusto and S. (05)

Estimation:

- B-spline decompositions for g and h on $[0, D]$ (assume $D = D'$):

$$g(t) = \sum_{k=-d}^{n-1} a_k B_{d,k,\mathcal{E}_n} \qquad h(t) = \sum_{k=-d'}^{n'-1} b_k B_{d',k,\mathcal{E}_{n'}}$$

  where $B_{d,k,\mathcal{E}_n}$ denotes the k-th B-spline of degree d associated to the set of $(n+1)$ knots $\mathcal{E}_n = (\varepsilon_0, \ldots, \varepsilon_n)$ on $[0, D]$.

- estimation of the parameters $\theta_{n,n'} = (\mu, (a_k), (b_k))$ by maximum likelihood.
  The MLE $\widehat{\theta}$ exists and is unique as soon as $0 < \int_0^T |g(t)|\, dt < 1$ and $0 < \int_0^T |h(t)|\, dt < 1$ ([Ogata (78)], [Brémaud and Massoulié (96)]).

Estimation and model selection (*FADO*)

Gusto and S. (05)

Estimation:

- B-spline decompositions for g and h on $[0, D]$ (assume $D = D'$):

$$g(t) = \sum_{k=-d}^{n-1} a_k B_{d,k,\mathcal{E}_n} \qquad h(t) = \sum_{k=-d'}^{n'-1} b_k B_{d',k,\mathcal{E}_{n'}}$$

  where $B_{d,k,\mathcal{E}_n}$ denotes the k-th B-spline of degree d associated to the set of $(n+1)$ knots $\mathcal{E}_n = (\varepsilon_0, \ldots, \varepsilon_n)$ on $[0, D]$.

- estimation of the parameters $\theta_{n,n'} = (\mu, (a_k), (b_k))$ by maximum likelihood.
  The MLE $\widehat{\theta}$ exists and is unique as soon as $0 < \int_0^T |g(t)|\, dt < 1$ and $0 < \int_0^T |h(t)|\, dt < 1$ ([Ogata (78)], [Brémaud and Massoulié (96)]).

Model selection:

- The degrees d and d' of the B-splines are fixed.
- Assume the knots are regularly spaced.
- AIC criterion is used to select the number of knots (n, n').

Example

Two strands of *E. coli* genome: $T = 9\,288\,442$

$\mathbf{u}$: 4290 genes of *E. coli*

$\mathbf{v}$: 1036 occurrences of TATAAT

$D = D' = 10\,000, \quad d = d' = 2, \quad 3 \leq n, n' \leq 20$

Example

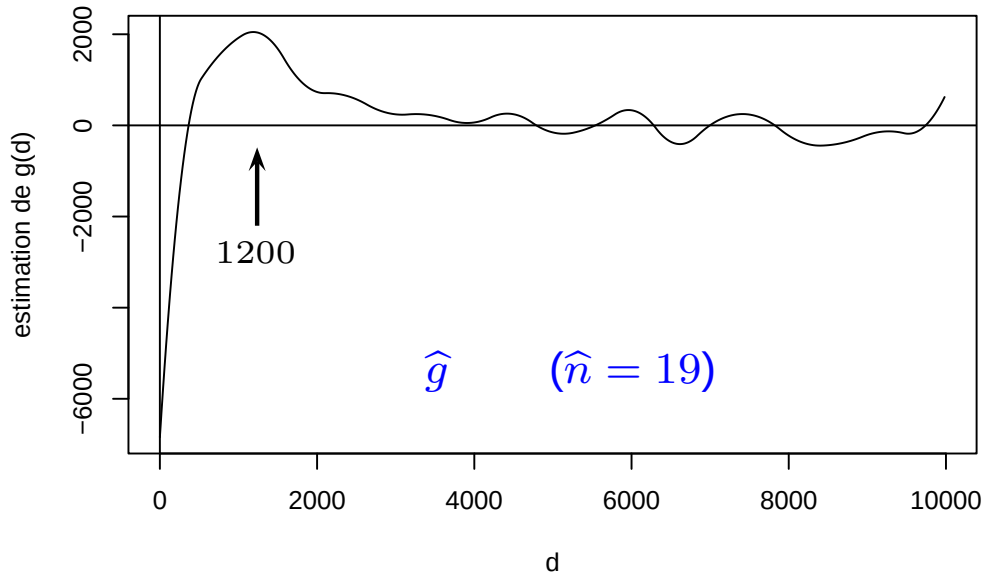
Two strands of *E. coli* genome: $T = 9\,288\,442$

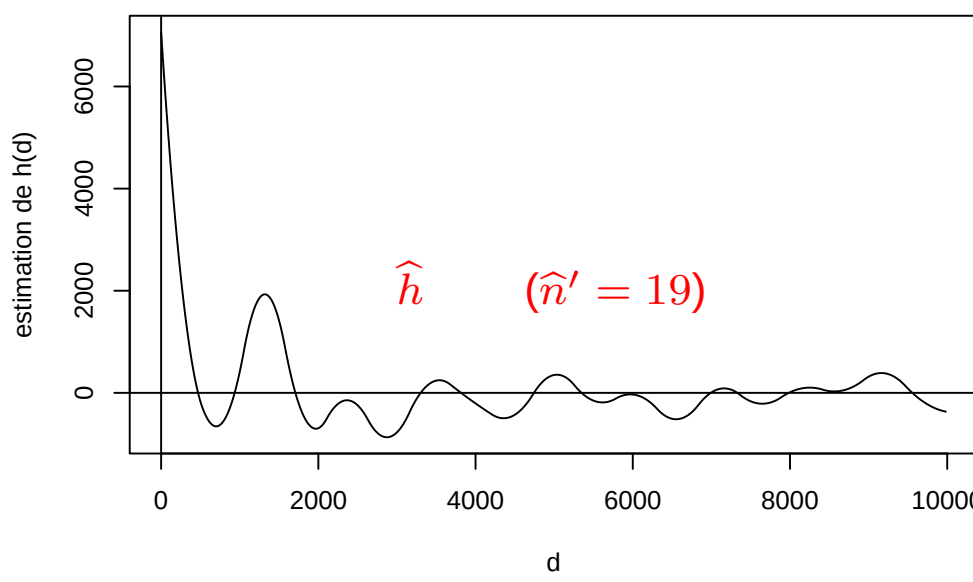
$\mathbf{u}$: 4290 genes of *E. coli*

$\mathbf{v}$: 1036 occurrences of TATAAT

$D = D' = 10\,000, \quad d = d' = 2, \quad 3 \leq n, n' \leq 20$

$$\widehat{\mu} = 3390.9/T$$





$H_0 = \{g \equiv 0\}$ rejected

$H_0 = \{h \equiv 0\}$ rejected

Generalization to multivariate Hawkes process

Carstensen, Sandelin, Winther, Hansen (10)

- **Model**: K processes (U_i^k), $k = 1, \ldots, K$, $i = 1, 2, \ldots$

$$\log(\lambda_k(t)) = \sum_{k'=1}^{K} \sum_{t-D \leq U_i^{k'} < t} h_{k,k'}(t - U_i^{k'})$$

- **Parametrization** for $h_{k,k'}$: cubic B-splines with 8 equidistant knots on $[0, D] \Rightarrow \theta$
- **Estimation**: maximum likelihood

$$\log(L(\theta)) = \int_0^T \log(\lambda_k(t)) dU_t - \int_0^T \lambda(t) dt$$

- No **model selection**

Penalized least square estimation ($h \equiv 0$)

Reynaud-Bouret and S. (10)

- **Model**: a unique process (U_i)

$$\lambda(t) = \mu + \sum_{t-D \leq U_i < t} g(t - U_i) = \mu + \int_{t-D}^{t} g(t-s) dU_s$$

- **Parametrization** for g: histogram basis
- **Estimation**: minimum least-square contrast

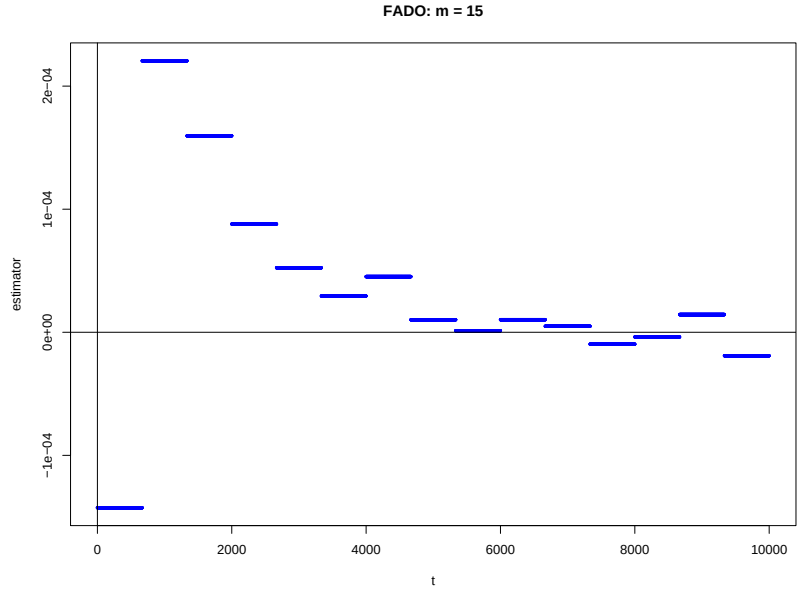
$$\gamma(\nu, f) = -\frac{2}{T} \int_0^T \left[\nu + \int_{t-D}^{t} f(t-s) dU_s \right] dU_t + \frac{1}{T} \int_0^T \left[\nu + \int_{t-D}^{t} f(t-s) dU_s \right]^2$$

- **Model selection**: penalized projection estimator over a particular family of histograms.

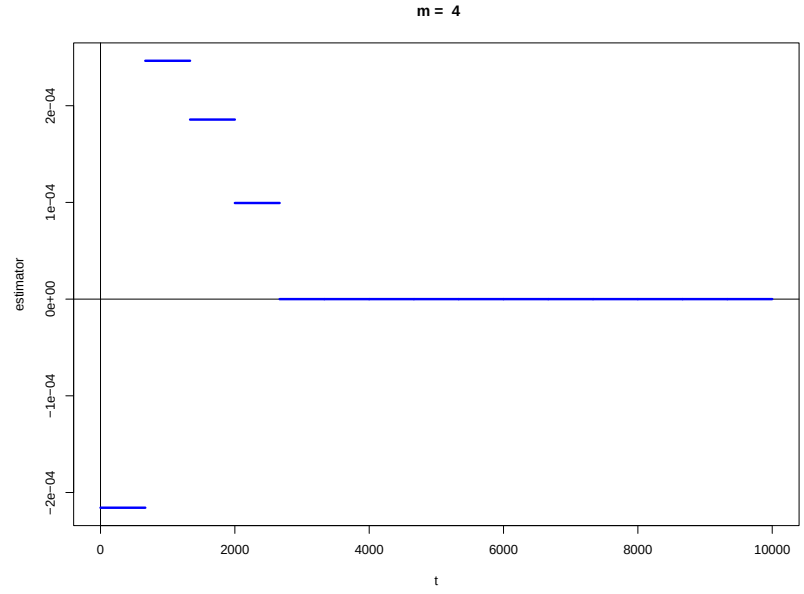
Comparison with *FADO*

Two strands of *E. coli* genome: $T = 9\,288\,442$

$\mathbf{u}$: 4290 genes of *E. coli*



Gusto and S. (05)



Reynaud-Bouret and S. (10)

Adaptive estimation ($g \equiv 0, \mu = 0$)

Sansonnet, Rivoirar, Reynaud-Bouret

- **Model**: two processes (U_i) and (V_i)

$$\lambda(t) = \sum_{t-D \leq V_i < t}^{t} h(t - V_i) = \int_{t-D}^{t} h(t-s) dV_s$$

- **Parametrization** for h: Haar wavelet basis $\Rightarrow \beta_{j,k}$
- **Estimation**: construction d'un estimateur sans biais $\widehat{\beta}_{j,k}$
- **Model selection**: thresholding rules

Open problems in Hawkes model

- Testing $\{g \equiv 0\}$ and/or $\{h \equiv 0\}$ only possible for now when working with likelihood.
- How to test for the significance of individual peaks?