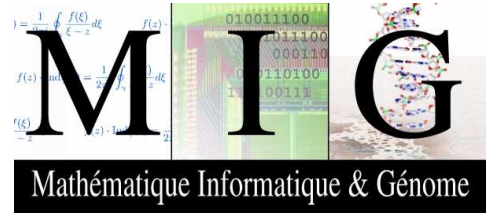
Exceptionnalité d'un motif dans une séquence hétérogène

Etienne Roquain, Sophie Schbath

Unité Mathématique, Informatique & Génome

INRA – Jouy-en-Josas



Introduction

Motivation : séquence d'ADN observée

a g a t g t t a g c g c a c a t g g ···

⇒ Trouver **mots** avec une fonction biologique particulière

mot = succession de lettres dans {a,c,g,t} (ex : gctggtgg)

Introduction

Motivation : séquence d'ADN observée

a g a t g t t a g c g c a c a t g g ···

$\Rightarrow$ Trouver **mots** avec une fonction biologique particulière

mot = succession de lettres dans {a,c,g,t} (ex : gctggtgg)

Approche statistique :

mots de fréquence significativement forte ou faible / modèle

Introduction

Motivation : séquence d'ADN observée

`a g a t g t t a g c g c a c a t g g ···`

$\Rightarrow$ Trouver **mots** avec une fonction biologique particulière

mot = succession de lettres dans $\{\texttt{a,c,g,t}\}$ (ex : `gctggtgg`)

Approche statistique :

mots de fréquence significativement forte ou faible / modèle

Mesure de l'exceptionnalité d'un mot $\mathbf{w}$:

$$p\textbf{-value de } \mathbf{w} = \mathbf{P}\ \ N(\mathbf{w}) \geq N_{obs}(\mathbf{w})\ \ ,$$

- $N(\mathbf{w})$ comptage de $\mathbf{w}$ dans séquence aléatoire
- $N_{obs}(\mathbf{w})$ comptage de $\mathbf{w}$ dans séquence d'ADN observée

Étapes générales de la démarche statistique

1. Choix du modèle pour la séquence.
   Exceptionnalité par rapport à quoi ?
   Classiquement modèle markovien homogène d'ordre m
   $\Rightarrow$ except. par rapport à la composition en $(m+1)$-mots

Étapes générales de la démarche statistique

1. Choix du modèle pour la séquence.
   Exceptionnalité par rapport à quoi ?
   Classiquement modèle markovien homogène d'ordre m
   $\Rightarrow$ except. par rapport à la composition en $(m+1)$-mots

2. Calcul p-value $\mathbf{P}\ \ N(\mathbf{w}) \geq N_{obs}(\mathbf{w})\ \ $ dans le modèle choisi.
   Calcul exact (séq. longueur $n < 10\,000$, $m \leq 1$) [Robin et al. 99]
   Grandes déviations, $n \to \infty$ [Nuel 04, Pudlo 04]
   Approx. loi $N(\mathbf{w})$, $n \to \infty$:
   - poissonniennes (mots rares) [Schbath 95]
   - gaussiennes (mots fréquents) [Prum et al. 95]

Étapes générales de la démarche statistique

1. Choix du modèle pour la séquence.
   Exceptionnalité par rapport à quoi ?
   Classiquement modèle markovien homogène d'ordre m
   $\Rightarrow$ except. par rapport à la composition en $(m+1)$-mots

2. Calcul p-value $\mathbf{P}\ \ N(\mathbf{w}) \geq N_{obs}(\mathbf{w})$ dans le modèle choisi.
   Calcul exact (séq. longueur $n < 10\,000$, $m \leq 1$) [Robin et al. 99]
   Grandes déviations, $n \to \infty$ [Nuel 04, Pudlo 04]
   Approx. loi $N(\mathbf{w})$, $n \to \infty$:
   - poissonniennes (mots rares) [Schbath 95]
   - gaussiennes (mots fréquents) [Prum et al. 95]

3. Selection des mots exceptionnels ?
   Beaucoup de p-values à traiter (long. $7 \to 16384$ mots)
   $\Rightarrow$ problème de tests multiples

Étapes générales de la démarche statistique

1. Choix du modèle pour la séquence.
   Exceptionnalité par rapport à quoi ?
   Classiquement modèle markovien homogène d'ordre m
   $\Rightarrow$ except. par rapport à la composition en $(m+1)$-mots

2. Calcul p-value $\mathbf{P}\ \ N(\mathbf{w}) \geq N_{obs}(\mathbf{w})$ dans le modèle choisi.
   Calcul exact (séq. longueur $n < 10\,000$, $m \leq 1$) [Robin et al. 99]
   Grandes déviations, $n \to \infty$ [Nuel 04, Pudlo 04]
   Approx. loi $N(\mathbf{w})$, $n \to \infty$:
   - poissonniennes (mots rares) [Schbath 95]
   - gaussiennes (mots fréquents) [Prum et al. 95]

3. Selection des mots exceptionnels ?
   Beaucoup de p-values à traiter (long. $7 \to 16384$ mots)
   $\Rightarrow$ problème de tests multiples

Puis validation des motifs candidats en laboratoire

Pb : prise en compte de l'hétérogénéité

Critique du modèle de Markov homogène d'ordre m :
Suppose la même composition en $(m+1)$-mots tout au long de la séquence.

$\Rightarrow$ pour des séquences hétérogènes : mesure erronée de l'exceptionnalité

Pb : prise en compte de l'hétérogénéité

Critique du modèle de Markov homogène d'ordre m :
Suppose la même composition en $(m+1)$-mots tout au long de la séquence.

$\Rightarrow$ pour des séquences hétérogènes : mesure erronée de l'exceptionnalité

But de l'exposé :

Prendre en compte une hétérogénéité connue dans le calcul de l'exceptionnalité d'un mot

Pour cela :

- modèle de Markov hétérogène à hétérogénéité connue
- approximations poissonniennes de la loi de $N(\mathbf{w})$ dans ce modèle. Généralisation approx. de Poisson composée de [Schbath 95]

Rappel : méthode homogène (1)

Chaîne de Markov homogène (ordre 1) :
Séquence $X_1 \cdots X_n$ "de mémoire d'ordre 1" :

$$\begin{aligned}\mathbf{P}(X_i = x_i \mid X_{i-1} = x_{i-1}, X_j, j < i-1) &= \mathbf{P}(X_i = x_i \mid X_{i-1} = x_{i-1}) \\ &= \pi(x_{i-1}, x_i)\end{aligned}$$

où π probabilité de transition indépendante de la position i (loi statio. μ).

Rappel : méthode homogène (1)

Chaîne de Markov homogène (ordre 1) :
Séquence $X_1 \cdots X_n$ "de mémoire d'ordre 1" :

$$\begin{aligned}\mathbf{P}(X_i = x_i \mid X_{i-1} = x_{i-1}, X_j, j < i-1) &= \mathbf{P}(X_i = x_i \mid X_{i-1} = x_{i-1}) \\ &= \pi(x_{i-1}, x_i)\end{aligned}$$

où π probabilité de transition indépendante de la position i (loi statio. μ).

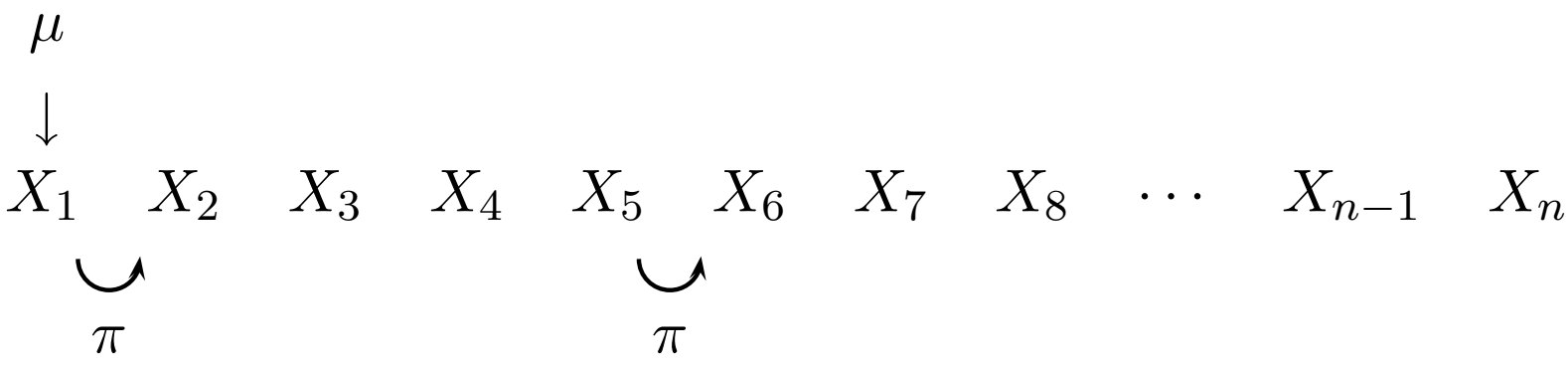
$$X_1 \quad X_2 \quad X_3 \quad X_4 \quad X_5 \quad X_6 \quad X_7 \quad X_8 \quad \cdots \quad X_{n-1} \quad X_n$$

Rappel : méthode homogène (1)

Chaîne de Markov homogène (ordre 1) :
Séquence $X_1 \cdots X_n$ "de mémoire d'ordre 1" :

$$\begin{aligned}\mathbf{P}(X_i = x_i \mid X_{i-1} = x_{i-1}, X_j, j < i-1) &= \mathbf{P}(X_i = x_i \mid X_{i-1} = x_{i-1}) \\ &= \pi(x_{i-1}, x_i)\end{aligned}$$

où π probabilité de transition indépendante de la position i (loi statio. μ).

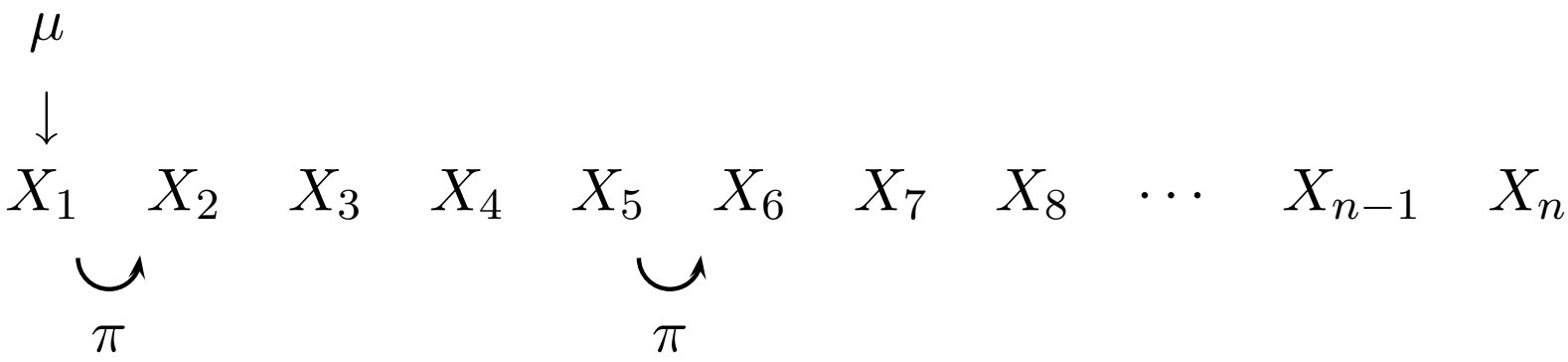


Rappel : méthode homogène (1)

Chaîne de Markov homogène (ordre 1) :
Séquence $X_1 \cdots X_n$ "de mémoire d'ordre 1" :

$$\mathbf{P}(X_i = x_i \mid X_{i-1} = x_{i-1}, X_j, j < i-1) = \mathbf{P}(X_i = x_i \mid X_{i-1} = x_{i-1}) = \pi(x_{i-1}, x_i)$$

où π probabilité de transition indépendante de la position i (loi statio. μ).

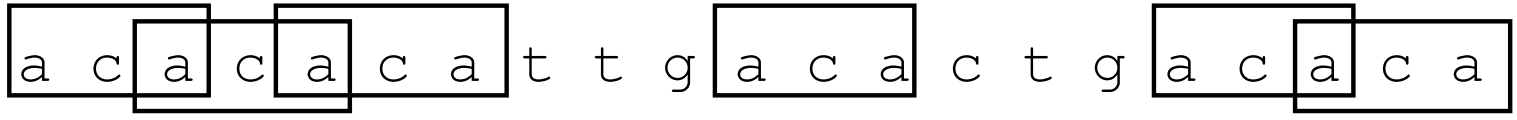


Pour un mot $\mathbf{w}$, comptage $N(\mathbf{w})$ dans $X_1 \cdots X_n$ variable aléatoire

$\Rightarrow$ approcher la loi de $N(\mathbf{w})$ dans ce modèle homogène ?

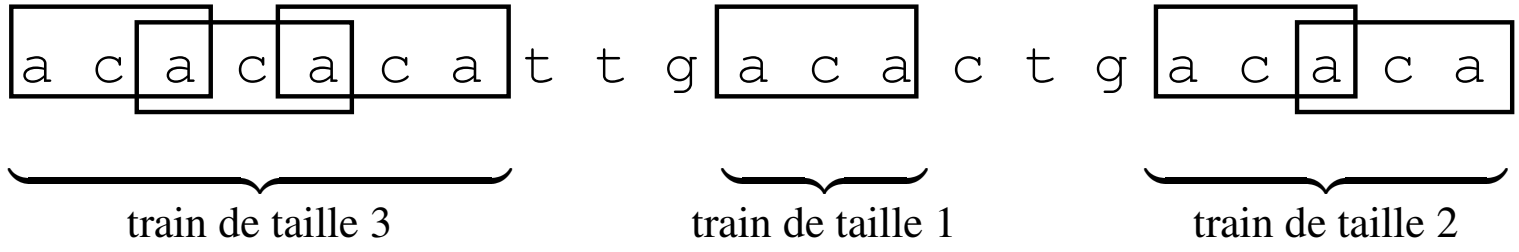
Rappel : méthode homogène (2)

Par ex. $\mathbf{w} = \texttt{aca}$:



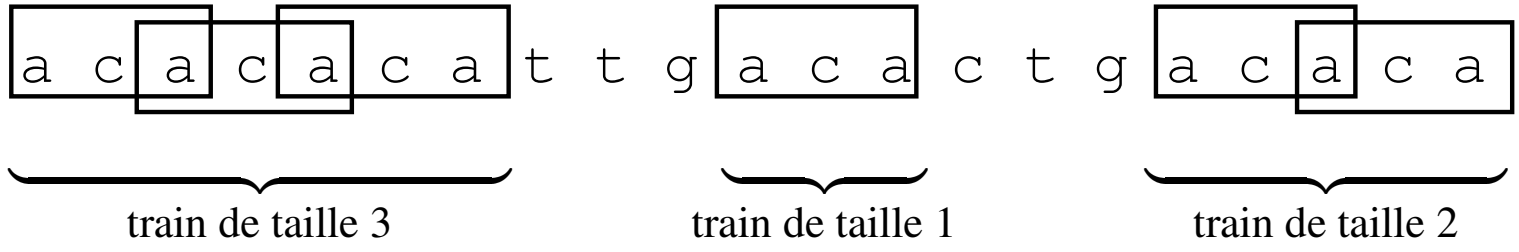
Rappel : méthode homogène (2)

Par ex. $\mathbf{w} = \texttt{aca}$:



Rappel : méthode homogène (2)

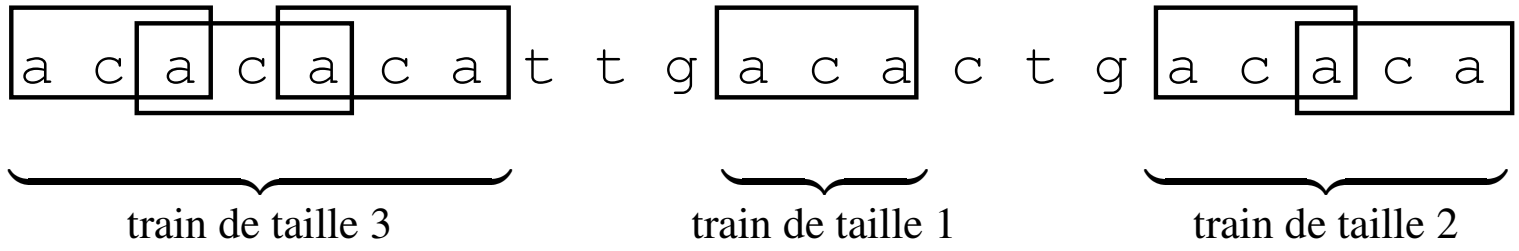
Par ex. $\mathbf{w} = \texttt{aca}$:



Les paquets sont appelés des trains :
successions d'occ. chevauchantes de $\mathbf{w}$ **maximales** i.e. ne chevauchant pas d'autres occ. de $\mathbf{w}$ ni avant ni après.

Rappel : méthode homogène (2)

Par ex. $\mathbf{w} = \texttt{aca}$:

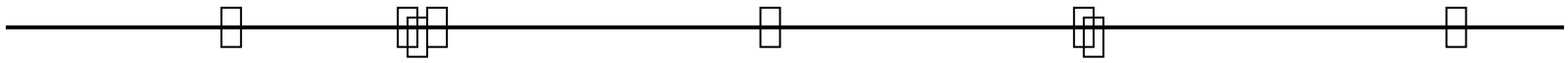


Les paquets sont appelés des trains :
successions d'occ. chevauchantes de $\mathbf{w}$ **maximales** i.e. ne chevauchant pas d'autres occ. de $\mathbf{w}$ ni avant ni après.

$$\begin{aligned} N(\mathbf{w}) &= 1 \times \text{nb de trains de taille } 1 \\ &+ 2 \times \text{nb de trains de taille } 2 \\ &+ 3 \times \text{nb de trains de taille } 3 + \dots \\ &= \sum_{k \geq 1} k N_k(\mathbf{w}) \end{aligned}$$

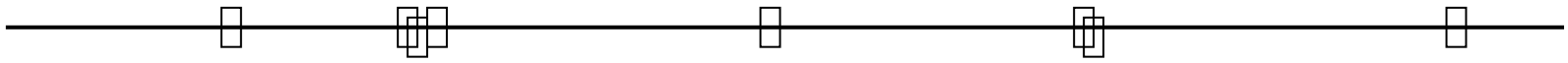
Rappel : méthode homogène (3)

Situation lorsque $n \to \infty$ et $\mathbb{E}N(\mathbf{w}) = O(1)$ (mot rare) :



Rappel : méthode homogène (3)

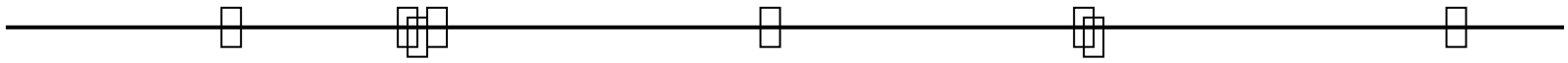
Situation lorsque $n \to \infty$ et $\mathbb{E}N(\mathbf{w}) = O(1)$ (mot rare) :



Par la méthode de Chen-Stein, on montre que lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$,
loi du nombre de k-trains $N_k(\mathbf{w}) \simeq$ loi de Poisson de para. $\mathbb{E}N_k(\mathbf{w})$.
et les $N_k(\mathbf{w})$ sont asymptotiquement indépendants

Rappel : méthode homogène (3)

Situation lorsque $n \to \infty$ et $\mathbb{E}N(\mathbf{w}) = O(1)$ (mot rare) :

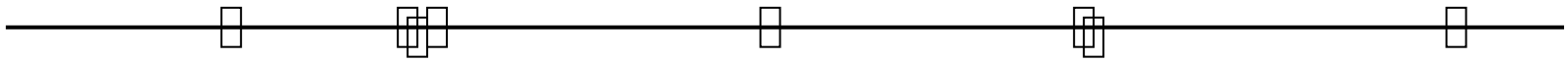


Par la méthode de Chen-Stein, on montre que lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$,
loi du nombre de k-trains $N_k(\mathbf{w}) \simeq$ loi de Poisson de para. $\mathbb{E}N_k(\mathbf{w})$.
et les $N_k(\mathbf{w})$ sont asymptotiquement indépendants

Comme $N(\mathbf{w}) = \sum_{k \geq 1} k N_k(\mathbf{w})$,
on a $\mathcal{L}(N(\mathbf{w})) \simeq \mathcal{L}\left(\sum_{k \geq 1} k Z_k\right)$ où Z_k suit $\mathcal{P}\ \mathbb{E}N_k(\mathbf{w})$ indep.

Rappel : méthode homogène (3)

Situation lorsque $n \to \infty$ et $\mathbb{E}N(\mathbf{w}) = O(1)$ (mot rare) :



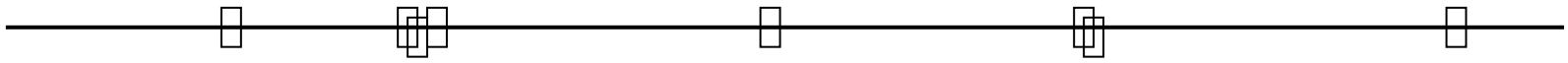
Par la méthode de Chen-Stein, on montre que lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$,
loi du nombre de k-trains $N_k(\mathbf{w}) \simeq$ loi de Poisson de para. $\mathbb{E}N_k(\mathbf{w})$.
et les $N_k(\mathbf{w})$ sont asymptotiquement indépendants

Comme $N(\mathbf{w}) = \sum_{k \geq 1} k N_k(\mathbf{w})$,
on a $\mathcal{L}(N(\mathbf{w})) \simeq \mathcal{L}\left(\sum_{k \geq 1} k Z_k\right)$ où Z_k suit $\mathcal{P}\ \mathbb{E}N_k(\mathbf{w})$ indep.

$$\mathcal{L}(N(\mathbf{w})) \simeq \mathcal{CP}(\mathbb{E}N_k(\mathbf{w}), k \geq 1).$$

Rappel : méthode homogène (3)

Situation lorsque $n \to \infty$ et $\mathbb{E}N(\mathbf{w}) = O(1)$ (mot rare) :



Par la méthode de Chen-Stein, on montre que lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$,
loi du nombre de k-trains $N_k(\mathbf{w}) \simeq$ loi de Poisson de para. $\mathbb{E}N_k(\mathbf{w})$.
et les $N_k(\mathbf{w})$ sont asymptotiquement indépendants

Comme $N(\mathbf{w}) = \sum_{k \geq 1} k N_k(\mathbf{w})$,
on a $\mathcal{L}(N(\mathbf{w})) \simeq \mathcal{L}\left(\sum_{k \geq 1} k Z_k\right)$ où Z_k suit $\mathcal{P}\ \mathbb{E}N_k(\mathbf{w})$ indep.

$$\mathcal{L}(N(\mathbf{w})) \simeq \mathcal{CP}(\mathbb{E}N_k(\mathbf{w}), k \geq 1).$$

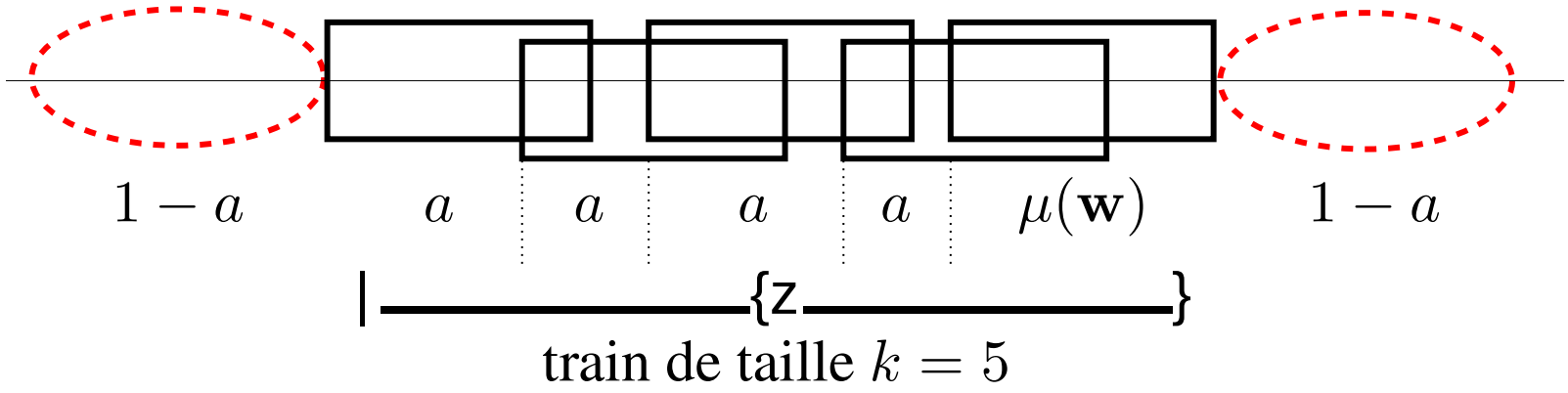
Calcul de $\mathbb{E}N_k(\mathbf{w})$ comptage attendu d'un k-train ?

Rappel : méthode homogène (4)

$$\mathbf{P}(\text{un train de taille } k \text{ commence en position } i) \ ?$$

Rappel : méthode homogène (4)

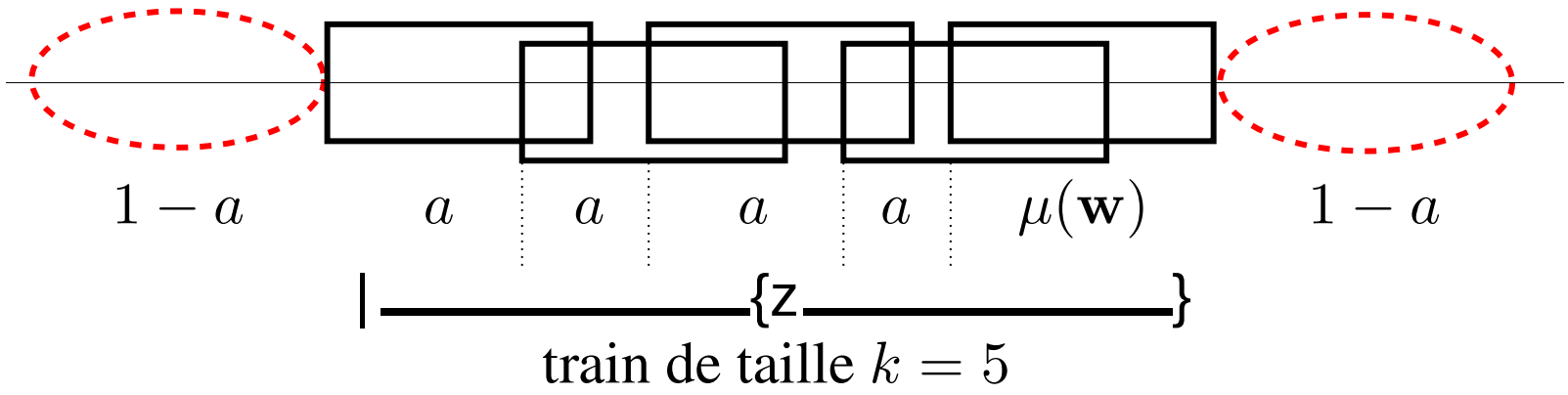
$\mathbf{P}(\text{un train de taille } k \text{ commence en position } i)$?



On note :
a la probabilité d'auto-recouvrement de $\mathbf{w}$
$\mu(\mathbf{w})$ la probabilité d'occ. de $\mathbf{w}$

Rappel : méthode homogène (4)

$\mathbf{P}$(un train de taille k commence en position i) ?



On note :
a la probabilité d'auto-recouvrement de $\mathbf{w}$
$\mu(\mathbf{w})$ la probabilité d'occ. de $\mathbf{w}$

$$\Rightarrow \mathbb{E}N_k(\mathbf{w}) = (n-h+1)(1-a)^2 a^{k-1}\mu(\mathbf{w}).$$

Rappel : méthode homogène (5)

Le résultat homogène :

Théorème [Schbath 95] :
Dans un modèle markovien homogène stationnaire, si $\mathbb{E}N(\mathbf{w}) = O(1)$,

$$\mathcal{L}(N(\mathbf{w})) \simeq \mathcal{CP}_{hom},$$

$$\mathcal{CP}_{hom} := \mathcal{CP} \quad (n-h+1)(1-a)^2 a^{k-1}\mu(\mathbf{w}), k \geq 1$$

Rappel : méthode homogène (5)

Le résultat homogène :

Théorème [Schbath 95] :
Dans un modèle markovien homogène stationnaire, si $\mathbb{E}N(\mathbf{w}) = O(1)$,

$$\mathcal{L}(N(\mathbf{w})) \simeq \mathcal{CP}_{hom},$$

$$\mathcal{CP}_{hom} := \mathcal{CP} \quad (n-h+1)(1-a)^2 a^{k-1}\mu(\mathbf{w}), k \geq 1$$

Résultat analogue dans un modèle hétérogène ?

Méthode hétérogène

Segmentation :

Chaque position i attachée à un état connu $s_i \in \mathcal{S}$. Ici $\mathcal{S} = \{1, 2\}$.

1	1	2	1	2	2	2	1	$\cdots$	1	1
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	$\cdots$	X_{n-1}	X_n

Méthode hétérogène

Segmentation :

Chaque position i attachée à un état connu $s_i \in \mathcal{S}$. Ici $\mathcal{S} = \{1, 2\}$.

1	1	2	1	2	2	2	1	$\cdots$	1	1
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	$\cdots$	X_{n-1}	X_n

La segmentation $s_1 \cdots s_n$ est **déterministe et connue a priori**

Méthode hétérogène

Segmentation :

Chaque position i attachée à un état connu $s_i \in \mathcal{S}$. Ici $\mathcal{S} = \{1, 2\}$.

1	1	2	1	2	2	2	1	$\cdots$	1	1
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	$\cdots$	X_{n-1}	X_n

La segmentation $s_1 \cdots s_n$ est **déterministe et connue a priori**

Elle peut représenter différentes informations biologiques :

- région codante / non-codante
- région conservée / variable
- ...

Méthode hétérogène

Segmentation :

Chaque position i attachée à un état connu $s_i \in \mathcal{S}$. Ici $\mathcal{S} = \{1, 2\}$.

1	1	2	1	2	2	2	1	$\cdots$	1	1
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	$\cdots$	X_{n-1}	X_n

La segmentation $s_1 \cdots s_n$ est **déterministe et connue a priori**

Elle peut représenter différentes informations biologiques :

- région codante / non-codante
- région conservée / variable
- ...

Notation : séquence $X_1 \cdots X_n$ coloriée selon les états 1 ou 2

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	$\cdots$	X_{n-1}	X_n

Méthode hétérogène : modèle

Modèle de Markov hétérogène par morceaux :

concaténation (selon la segmentation) de modèles markoviens homogènes indépendants

- dans l'état 1 : probabilité de transition π_1 (loi statio. μ_1)
- dans l'état 2 : probabilité de transition π_2 (loi statio. μ_2)

Méthode hétérogène : modèle

Modèle de Markov hétérogène par morceaux :

concaténation (selon la segmentation) de modèles markoviens homogènes indépendants

- dans l'état 1 : probabilité de transition π_1 (loi statio. μ_1)
- dans l'état 2 : probabilité de transition π_2 (loi statio. μ_2)

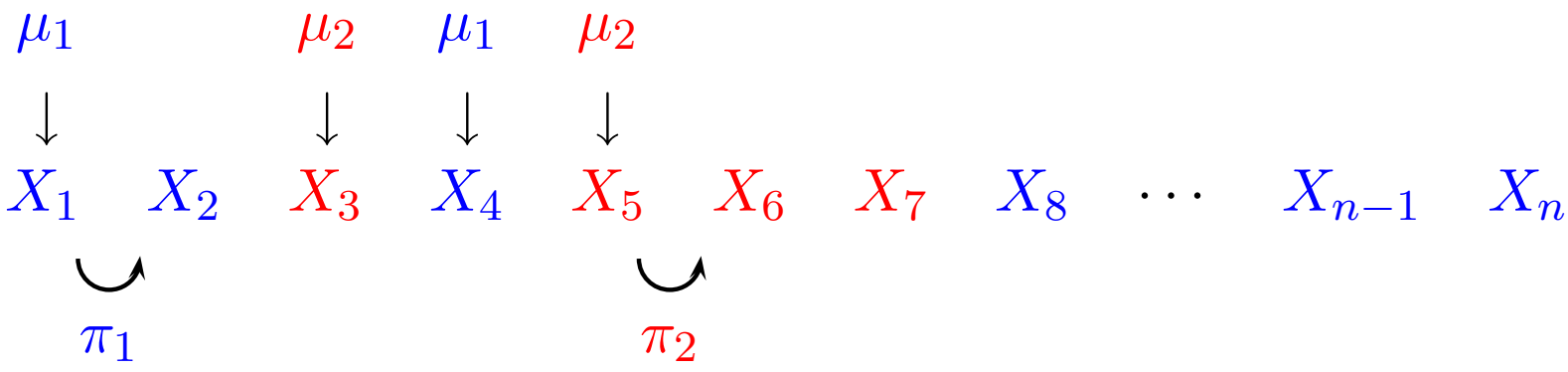
$$X_1 \quad X_2 \quad X_3 \quad X_4 \quad X_5 \quad X_6 \quad X_7 \quad X_8 \quad \cdots \quad X_{n-1} \quad X_n$$

Méthode hétérogène : modèle

Modèle de Markov hétérogène par morceaux :

concaténation (selon la segmentation) de modèles markoviens homogènes indépendants

- dans l'état 1 : probabilité de transition π_1 (loi statio. μ_1)
- dans l'état 2 : probabilité de transition π_2 (loi statio. μ_2)

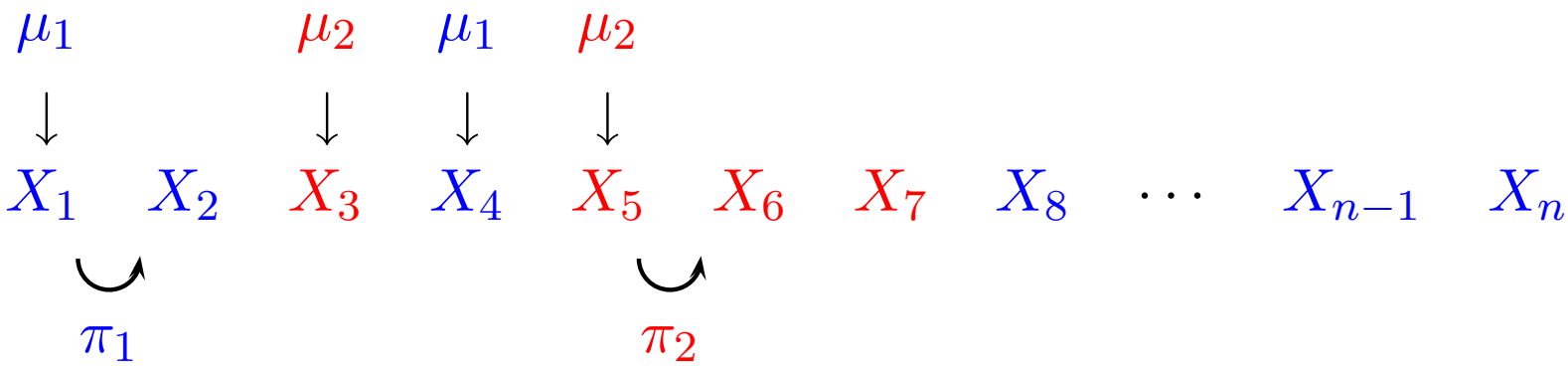


Méthode hétérogène : modèle

Modèle de Markov hétérogène par morceaux :

concaténation (selon la segmentation) de modèles markoviens homogènes indépendants

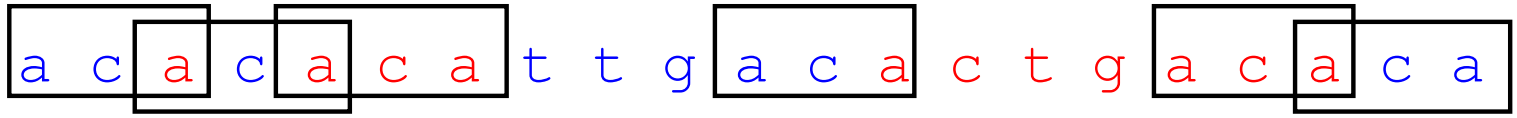
- dans l'état 1 : probabilité de transition π_1 (loi statio. μ_1)
- dans l'état 2 : probabilité de transition π_2 (loi statio. μ_2)



$\Rightarrow$ pour un mot $\mathbf{w}$, loi du comptage $N(\mathbf{w})$ dans ce modèle hétérogène ?

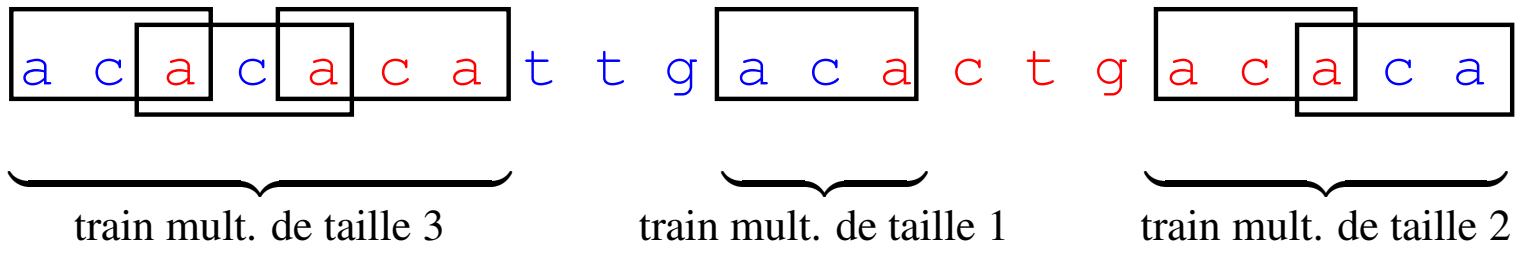
Méthode hétérogène : comptage multicolore

$N(\mathbf{w})$ compte les occurrences multicolores :



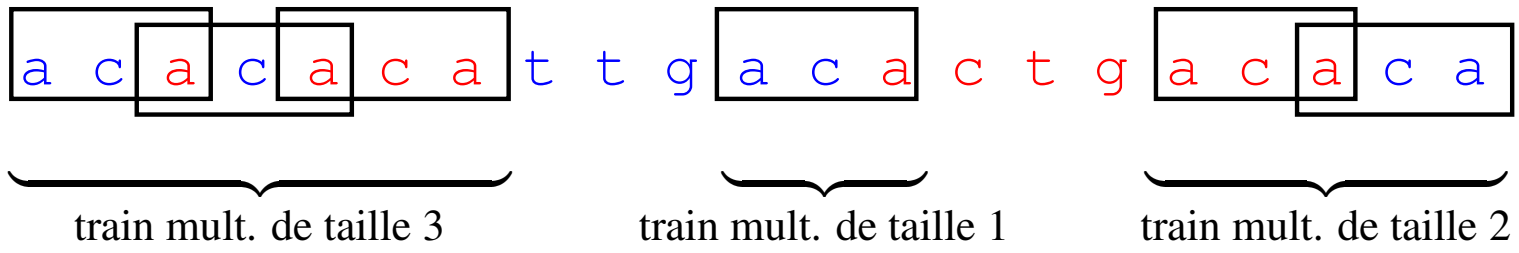
Méthode hétérogène : comptage multicolore

$N(\mathbf{w})$ compte les occurrences multicolores :



Méthode hétérogène : comptage multicolore

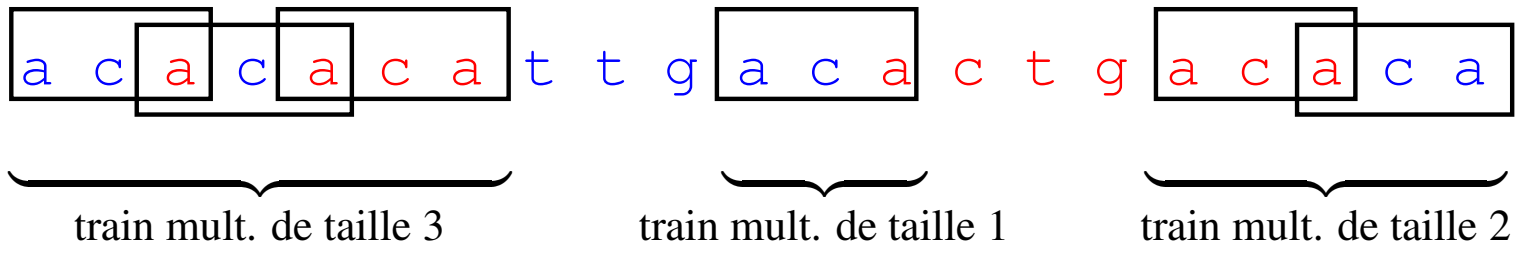
$N(\mathbf{w})$ compte les occurrences multicolores :



Probabilité d'occurrence d'un train multicolore $\Rightarrow$ trop complexe !

Méthode hétérogène : comptage multicolore

$N(\mathbf{w})$ compte les occurrences multicolores :



Probabilité d'occurrence d'un train multicolore $\Rightarrow$ trop complexe !

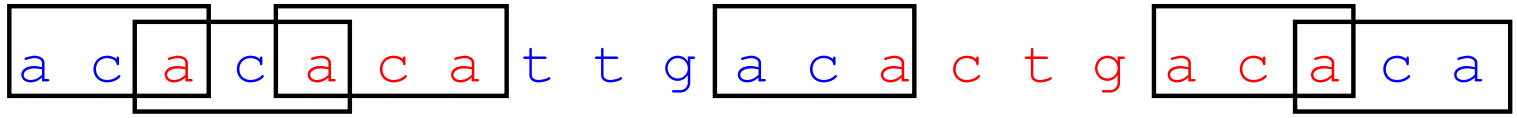
Idée : réduire $N(\mathbf{w})$ en négligeant certaines occ. coloriées de $\mathbf{w}$

Compromis : plus on compte d'occurrences

- plus approx. **précise** (comptage proche de $N(\mathbf{w})$)
- plus approx. **complexe** (difficulté de calcul)

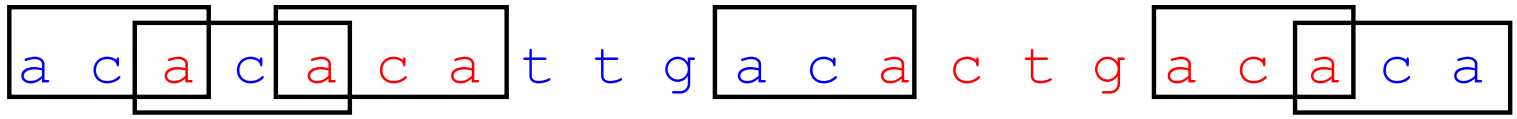
Méthode hétérogène : les comptages

$N(\mathbf{w})$ compte les occurrences multicolores : ici $N(\mathbf{w}) = 6$



Méthode hétérogène : les comptages

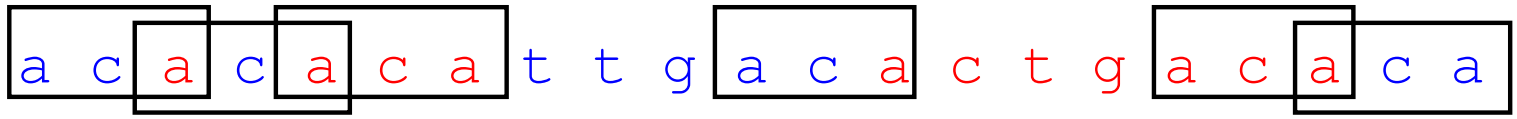
$N(\mathbf{w})$ compte les occurrences multicolores : ici $N(\mathbf{w}) = 6$



$N_{uni}(\mathbf{w})$ compte les occ. unicolores ($\in$ un segment) : ici $N_{uni}(\mathbf{w}) = 2$

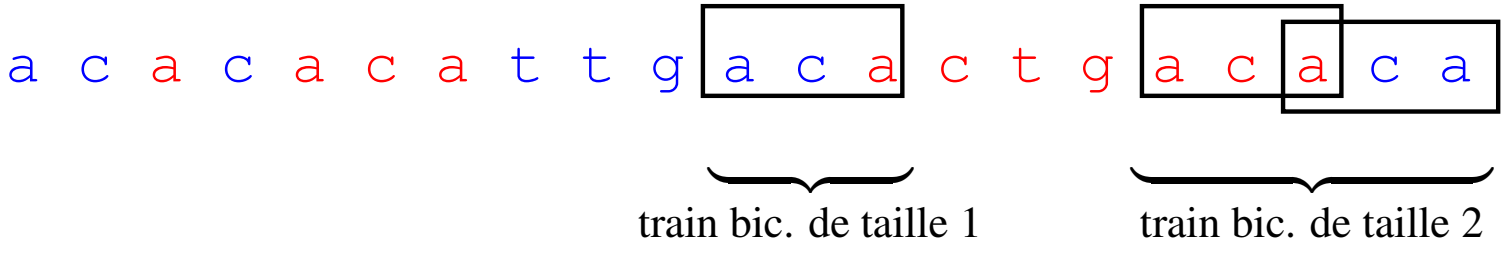
Méthode hétérogène : les comptages

$N(\mathbf{w})$ compte les occurrences multicolores : ici $N(\mathbf{w}) = 6$



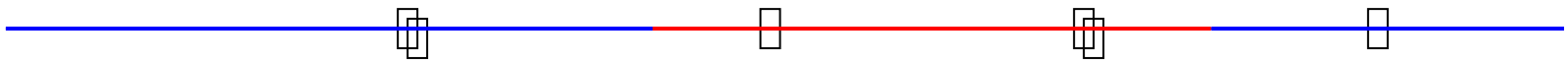
$N_{uni}(\mathbf{w})$ compte les occ. unicolores ($\in$ un segment) : ici $N_{uni}(\mathbf{w}) = 2$

$N_{bic}(\mathbf{w})$ compte les occ. dans les trains bicolores (au plus à cheval sur deux segments) : ici $N_{bic}(\mathbf{w}) = 3$



Méthode hétérogène : approx. avec $N_{uni}(\mathbf{w})$

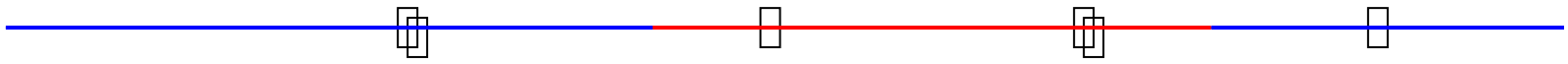
Situation lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$ et nombre de ruptures $\rho = O(1)$:



$$\Rightarrow \mathcal{L}(N(\mathbf{w})) \simeq \mathcal{L}(N_{uni}(\mathbf{w}))$$

Méthode hétérogène : approx. avec $N_{uni}(\mathbf{w})$

Situation lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$ et nombre de ruptures $\rho = O(1)$:



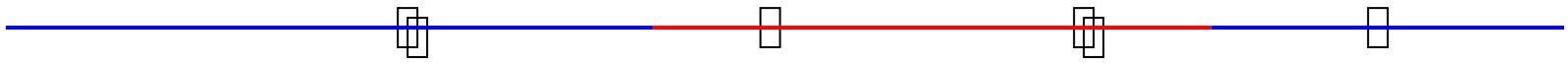
$$\Rightarrow \mathcal{L}(N(\mathbf{w})) \simeq \mathcal{L}(N_{uni}(\mathbf{w}))$$

Démarche pour approcher la loi de $N_{uni}(\mathbf{w})$:

1. Utiliser la méthode homogène de [Schbath 95] sur chaque segment
2. Regrouper les différentes lois de Poisson composées (indépendantes)

Méthode hétérogène : approx. avec $N_{uni}(\mathbf{w})$

Situation lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$ et nombre de ruptures $\rho = O(1)$:



$$\Rightarrow \mathcal{L}(N(\mathbf{w})) \simeq \mathcal{L}(N_{uni}(\mathbf{w}))$$

Démarche pour approcher la loi de $N_{uni}(\mathbf{w})$:

1. Utiliser la méthode homogène de [Schbath 95] sur chaque segment
2. Regrouper les différentes lois de Poisson composées (indépendantes)

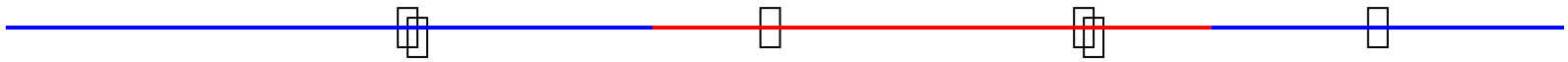
Proposition : Lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$ et $\rho h = o(n)$ (peu de segments),

$$\mathcal{L}(N(\mathbf{w})) \simeq \mathcal{CP}_{uni},$$

avec $\mathcal{CP}_{uni} := \mathcal{CP}_{hom} * \mathcal{CP}_{hom}$

Méthode hétérogène : approx. avec $N_{uni}(\mathbf{w})$

Situation lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$ et nombre de ruptures $\rho = O(1)$:



$$\Rightarrow \mathcal{L}(N(\mathbf{w})) \simeq \mathcal{L}(N_{uni}(\mathbf{w}))$$

Démarche pour approcher la loi de $N_{uni}(\mathbf{w})$:

1. Utiliser la méthode homogène de [Schbath 95] sur chaque segment
2. Regrouper les différentes lois de Poisson composées (indépendantes)

Proposition : Lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$ et $\rho h = o(n)$ (peu de segments),

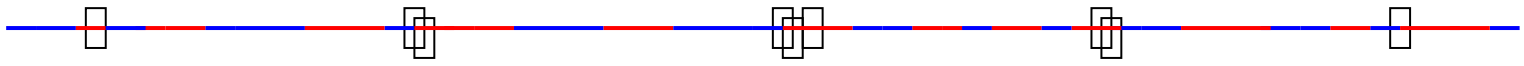
$$\mathcal{L}(N(\mathbf{w})) \simeq \mathcal{CP}_{uni},$$

avec $\mathcal{CP}_{uni} := \mathcal{CP}_{hom} * \mathcal{CP}_{hom}$

Approx. par $\mathcal{CP}_{uni}$ mauvaise si beaucoup de ruptures !

Méthode hétérogène : approx. avec $N_{bic}(\mathbf{w})$

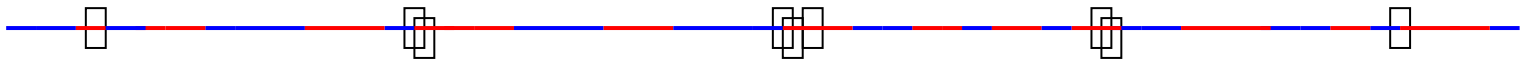
Situation lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$ et $+$ de ruptures mais $L_{\min}$ grand :



$$\Rightarrow \mathcal{L}(N(\mathbf{w})) \simeq \mathcal{L}(N_{bic}(\mathbf{w}))$$

Méthode hétérogène : approx. avec $N_{bic}(\mathbf{w})$

Situation lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$ et $+$ de ruptures mais $L_{\min}$ grand :



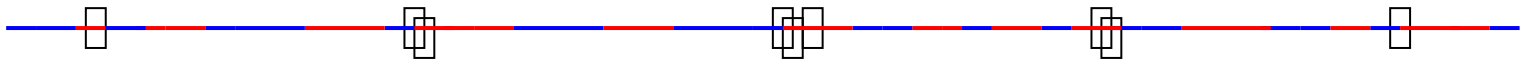
$$\Rightarrow \mathcal{L}(N(\mathbf{w})) \simeq \mathcal{L}(N_{bic}(\mathbf{w}))$$

Démarche :

1. Utiliser la méthode de Chen-Stein pour $N(\mathbf{w}) \simeq \mathcal{CP}(\mathbb{E}N_k(\mathbf{w}), k \geq 1)$
2. $\mathbb{E}N_k(\mathbf{w}) \simeq \mathbb{E}N_{k,bic}(\mathbf{w})$ comptage attendu des k-trains bicolores
3. Calcul de $\mathbb{E}N_{k,bic}(\mathbf{w})$(assez technique)

Méthode hétérogène : approx. avec $N_{bic}(\mathbf{w})$

Situation lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$ et + de ruptures mais $L_{\min}$ grand :



$$\Rightarrow \mathcal{L}(N(\mathbf{w})) \simeq \mathcal{L}(N_{bic}(\mathbf{w}))$$

Démarche :

1. Utiliser la méthode de Chen-Stein pour $N(\mathbf{w}) \simeq \mathcal{CP}(\mathbb{E}N_k(\mathbf{w}), k \geq 1)$
2. $\mathbb{E}N_k(\mathbf{w}) \simeq \mathbb{E}N_{k,bic}(\mathbf{w})$ comptage attendu des k-trains bicolores
3. Calcul de $\mathbb{E}N_{k,bic}(\mathbf{w})$(assez technique)

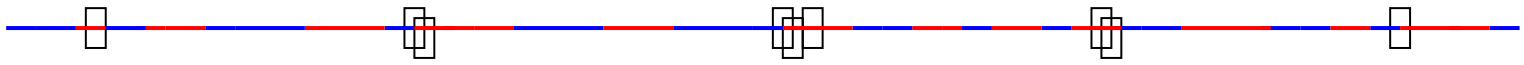
Théorème : Lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$ et $L_{min}/h \to \infty$,

$$\mathcal{L}(N(\mathbf{w})) \simeq \mathcal{CP}_{bic}$$

Avec une expression explicite (mais compliquée) pour $\mathcal{CP}_{bic}$.

Méthode hétérogène : approx. avec $N_{bic}(\mathbf{w})$

Situation lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$ et + de ruptures mais $L_{\min}$ grand :



$$\Rightarrow \mathcal{L}(N(\mathbf{w})) \simeq \mathcal{L}(N_{bic}(\mathbf{w}))$$

Démarche :

1. Utiliser la méthode de Chen-Stein pour $N(\mathbf{w}) \simeq \mathcal{CP}(\mathbb{E}N_k(\mathbf{w}), k \geq 1)$
2. $\mathbb{E}N_k(\mathbf{w}) \simeq \mathbb{E}N_{k,bic}(\mathbf{w})$ comptage attendu des k-trains bicolores
3. Calcul de $\mathbb{E}N_{k,bic}(\mathbf{w})$(assez technique)

Théorème : Lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$ et $L_{min}/h \to \infty$,

$$\mathcal{L}(N(\mathbf{w})) \simeq \mathcal{CP}_{bic}$$

Avec une expression explicite (mais compliquée) pour $\mathcal{CP}_{bic}$.

Rq : condition $L_{min}/h \to \infty$ forte ($L_{min} \geq 5h$ suffit en pratique)

Méthode hétérogène : petit bilan

Deux approx. loi $N(\mathbf{w})$ lorsque la séquence est hétérogène (et $\mathbf{w}$ rare) :

- par $\mathcal{CP}_{uni}$: bonne si “peu" de ruptures i.e. $\rho h/n$ “petit"
- par $\mathcal{CP}_{bic}$: bonne si segments “longs" i.e. L_{min} “grand"

Méthode hétérogène : petit bilan

Deux approx. loi $N(\mathbf{w})$ lorsque la séquence est hétérogène (et $\mathbf{w}$ rare) :

- par $\mathcal{CP}_{uni}$: bonne si “peu" de ruptures i.e. $\rho h/n$ “petit"
- par $\mathcal{CP}_{bic}$: bonne si segments “longs" i.e. L_{min} “grand"

Remarques :
- Généralisation à l'ordre $m \geq 2$ possible
- Temps de calcul raisonnable pour $\mathcal{CP}_{uni}$ et plus long pour $\mathcal{CP}_{bic}$

Méthode hétérogène : petit bilan

Deux approx. loi $N(\mathbf{w})$ lorsque la séquence est hétérogène (et $\mathbf{w}$ rare) :

- par $\mathcal{CP}_{uni}$: bonne si “peu" de ruptures i.e. $\rho h/n$ “petit"
- par $\mathcal{CP}_{bic}$: bonne si segments “longs" i.e. L_{min} “grand"

Remarques :
- Généralisation à l'ordre $m \geq 2$ possible
- Temps de calcul raisonnable pour $\mathcal{CP}_{uni}$ et plus long pour $\mathcal{CP}_{bic}$

Qualités des approx. $\mathcal{CP}_{uni}, \mathcal{CP}_{bic}$?
On propose une étude de simulation

Données simulées

Segmentation **régulière** $n = 100\,000$ avec **nombre de ruptures** ρ

$\rho = 1$

$\rho = 3$

$\rho = 8$

Données simulées

Segmentation **régulière** $n = 100\,000$ avec **nombre de ruptures** ρ

$\rho = 1$

$\rho = 3$

$\rho = 8$

Modèle hétérogène d'ordre 0 avec **"niveau d'hétérogénéité"** ε

$$\mu_1(\mathtt{a}) = \mu_1(\mathtt{g}) = 0.25 + \varepsilon \qquad \mu_1(\mathtt{c}) = \mu_1(\mathtt{t}) = 0.25 - \varepsilon$$

$$\mu_2(\mathtt{a}) = \mu_2(\mathtt{g}) = 0.25 - \varepsilon \qquad \mu_2(\mathtt{c}) = \mu_2(\mathtt{t}) = 0.25 + \varepsilon$$

Données simulées

Segmentation **régulière** $n = 100\,000$ avec **nombre de ruptures** ρ

$\rho = 1$

$\rho = 3$

$\rho = 8$

Modèle hétérogène d'ordre 0 avec **"niveau d'hétérogénéité"** ε

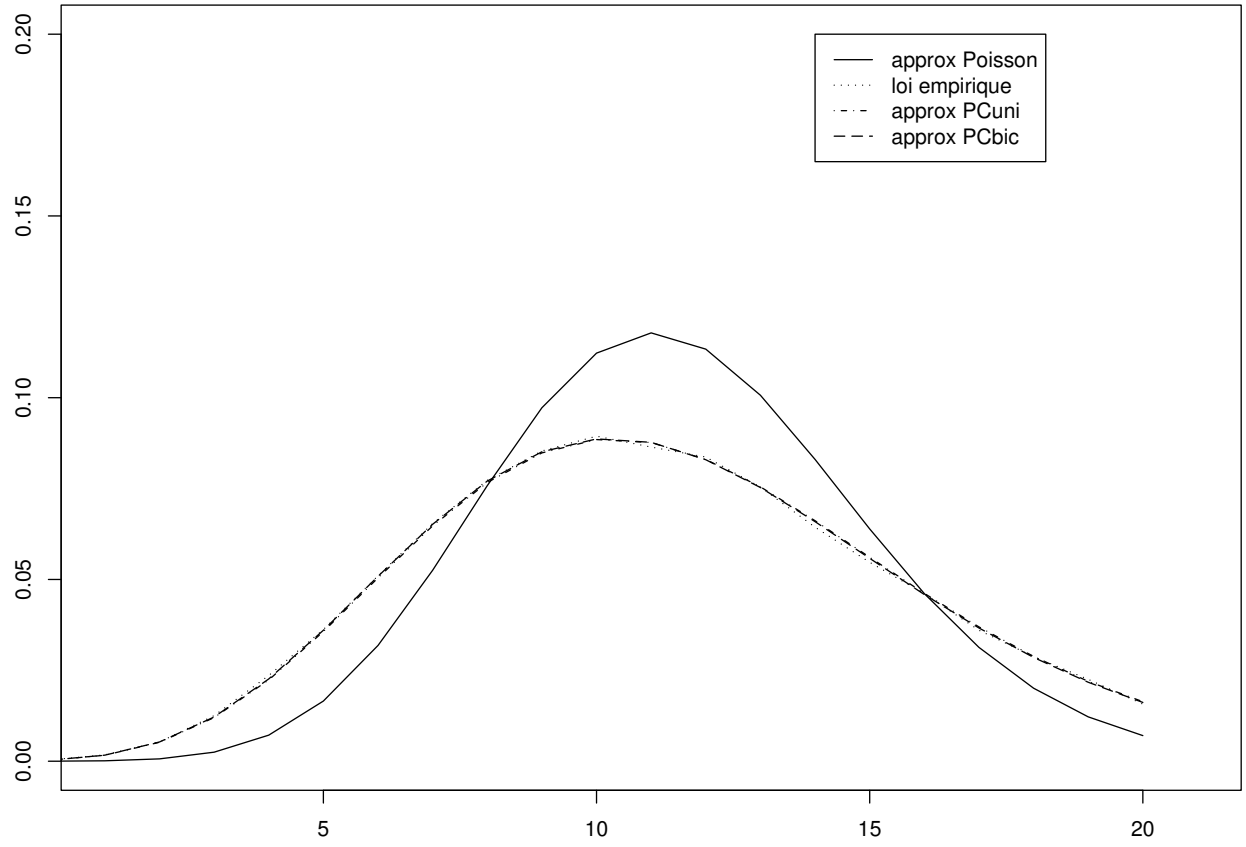
$$\mu_1(\mathtt{a}) = \mu_1(\mathtt{g}) = 0.25 + \varepsilon \qquad \mu_1(\mathtt{c}) = \mu_1(\mathtt{t}) = 0.25 - \varepsilon$$
$$\mu_2(\mathtt{a}) = \mu_2(\mathtt{g}) = 0.25 - \varepsilon \qquad \mu_2(\mathtt{c}) = \mu_2(\mathtt{t}) = 0.25 + \varepsilon$$

On considère : loi empirique de $N(\mathbf{w})$ (100 000 simulations)
$\mathcal{CP}_{uni}$, $\mathcal{CP}_{bic}$ et loi de Poisson ajustée

Avec $\mathbf{w} = \mathtt{aaaaaaa}$

Données simulées

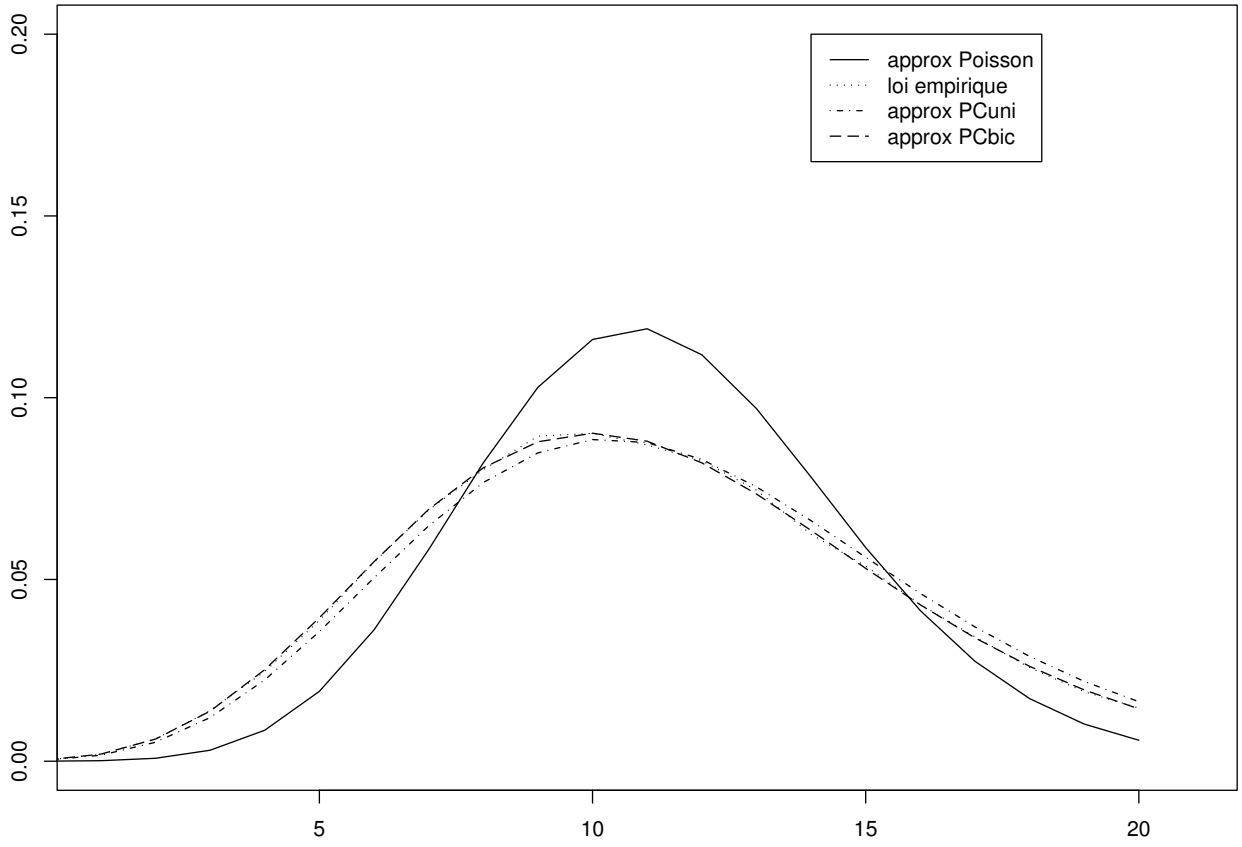
$\varepsilon = 0.05$, $\rho = 100$



- Approx. par $\mathcal{CP}_{uni}$ et $\mathcal{CP}_{bic}$ bonnes $h\rho/n \simeq 0.006$
- Approx. par Poisson pas bonne (mot très recouvrant)

Données simulées

$\varepsilon = 0.05$, $\rho = 1000$

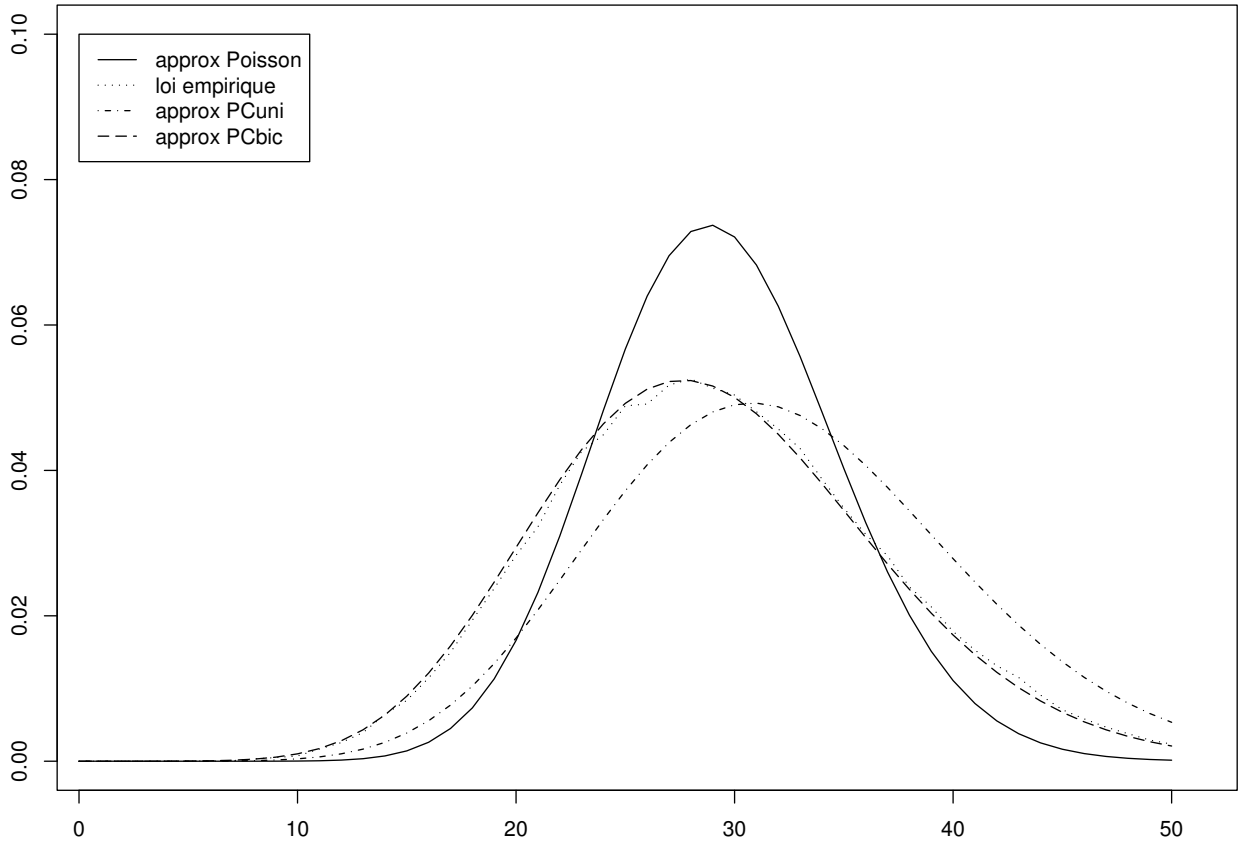


- Approx. par $\mathcal{CP}_{bic}$ bonne

- Approx. par $\mathcal{CP}_{uni}$ moins bonne $h\rho/n \simeq 0.056$

Données simulées

$\varepsilon = 0.01$, $\rho = 2000$



- Approx. par $\mathcal{CP}_{bic}$ bonne
- Approx. par $\mathcal{CP}_{uni}$ mauvaise $h\rho/n \simeq 0.11$

Données simulées

Conclusion des simulations :

- Qualité de l'approx. par $\mathcal{CP}_{uni}$ assez bien évaluée par $h\rho/n$
- Approx. par $\mathcal{CP}_{bic}$ valable sur une plus large bande

Données simulées

Conclusion des simulations :

- Qualité de l'approx. par $\mathcal{CP}_{uni}$ assez bien évaluée par $h\rho/n$
- Approx. par $\mathcal{CP}_{bic}$ valable sur une plus large bande

$\Rightarrow$ En pratique :

- si $h\rho/n$ petit (par ex : $h\rho/n \leq 0.01$), on utilise $\mathcal{CP}_{uni}$
- sinon, on utilise $\mathcal{CP}_{bic}$ (calcul plus long)

Données simulées

Conclusion des simulations :

- Qualité de l'approx. par $\mathcal{CP}_{uni}$ assez bien évaluée par $h\rho/n$
- Approx. par $\mathcal{CP}_{bic}$ valable sur une plus large bande

$\Rightarrow$ En pratique :

- si $h\rho/n$ petit (par ex : $h\rho/n \leq 0.01$), on utilise $\mathcal{CP}_{uni}$
- sinon, on utilise $\mathcal{CP}_{bic}$ (calcul plus long)

Comportement sur des vraies données ?

Notamment comparaison méthode hétérogène / méthode homogène

Données réelles : exemple 1

Génome du phage *Lambda*
Séquence long. $n = 48502$:

```
gggcggcgacctcgcgggttttcgctatttatgaaaattttccggtttaa
ggcgtttccgttcttcttcgtcataacttaatgttttttatttaaaatacc
ctctgaaaagaaaggaaacgacaggtgctgaaagcgaggctttttggcct...
```

Données réelles : exemple 1

Génome du phage *Lambda*
Séquence long. $n = 48502$:

```
gggcggcgacctcgcgggttttcgctatttatgaaaattttccggtttaa
ggcgtttccgttcttcttcgtcataacttaatgttttttatttaaaatacc
ctctgaaaagaaaggaaacgacaggtgctgaaagcgaggctttttggcct...
```

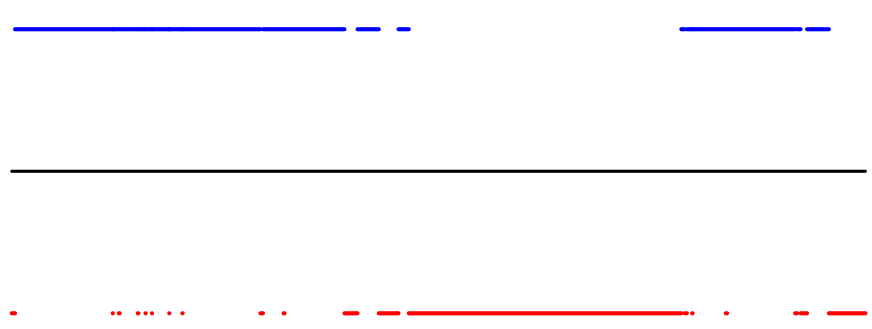
Segmentation codant dans le sens direct / non codant dans le sens direct :

Données réelles : exemple 1

Génome du phage *Lambda*
Séquence long. $n = 48502$:

```
gggcggcgacctcgcgggttttcgctatttatgaaaattttccggtttaa
ggcgtttccgttcttcttcgtcataacttaatgttttttatttaaaatacc
ctctgaaaagaaaggaaacgacaggtgctgaaagcgaggctttttggcct...
```

Segmentation codant dans le sens direct / non codant dans le sens direct :



Mots except de longueur $h = 5$.
- méthode **homogène** $\mathcal{CP}_{hom}$
- méthode **hétérogène** $\mathcal{CP}_{uni}$ ($\rho h/n \simeq 0.003$)

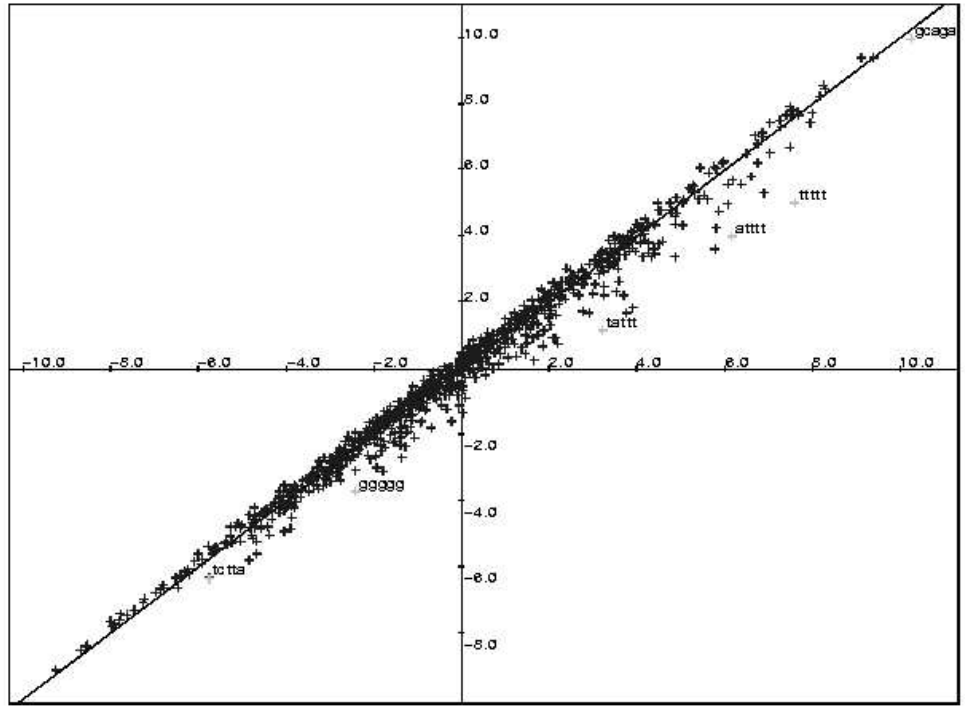
Données réelles : exemple 1

p-values dans $[0, 1] \longrightarrow$ score dans $\mathbb{R}$ (transfo. quantile)

Données réelles : exemple 1

p-values dans $[0, 1] \longrightarrow$ score dans $\mathbb{R}$ (transfo. quantile)

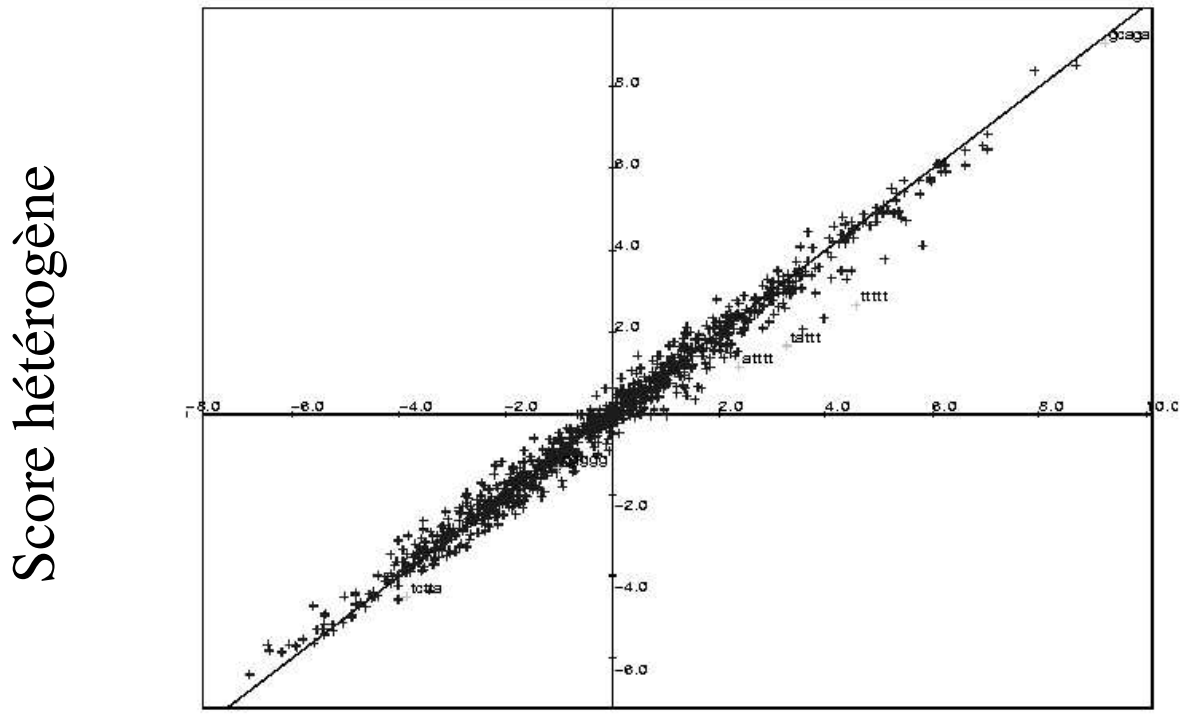
$$m = 0$$



Données réelles : exemple 1

p-values dans $[0, 1] \longrightarrow$ score dans $\mathbb{R}$ (transfo. quantile)

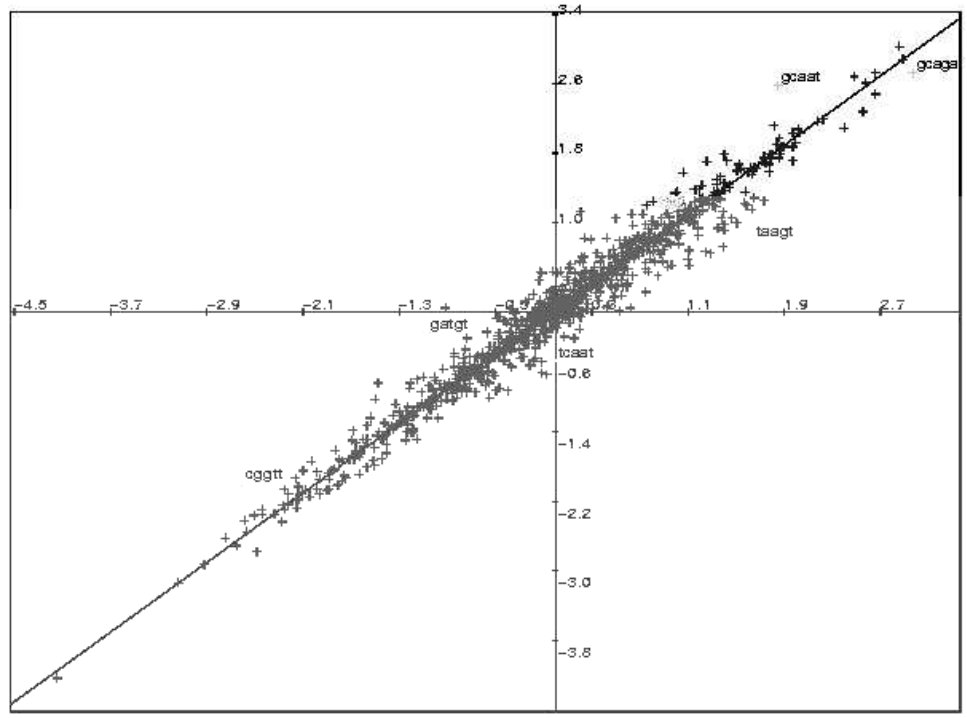
$m = 1$



Données réelles : exemple 1

p-values dans $[0, 1] \longrightarrow$ score dans $\mathbb{R}$ (transfo. quantile)

$$m = 3$$



Données réelles : exemple 2

Séquence : concaténation *E. Coli* + *H. influenzae*
(100 000 premières bases)

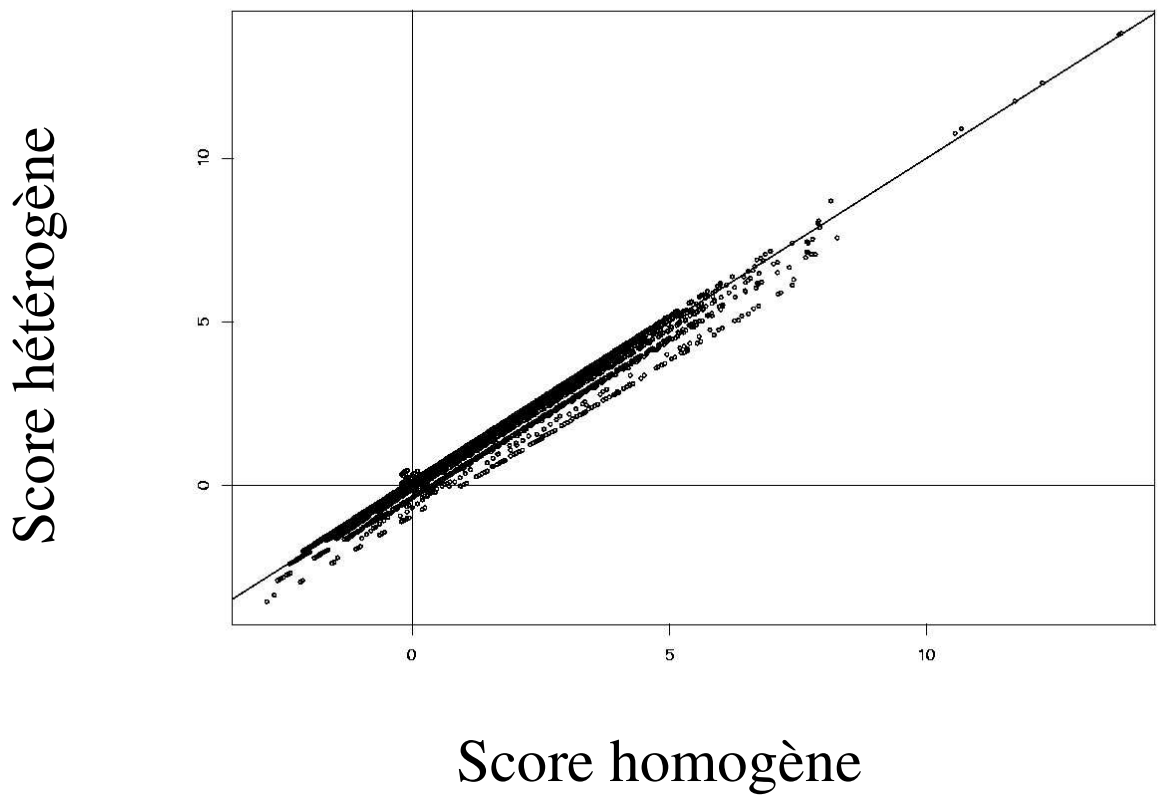
Segmentation : 1 pour *E. Coli* / 2 pour *H. influenzae*

$\Rightarrow$ mots exceptionnels dans les deux génomes simultanément ?

Données réelles : exemple 2

Mots de longueur $h = 8$, concaténation *E. Coli* + *H. influenzae*

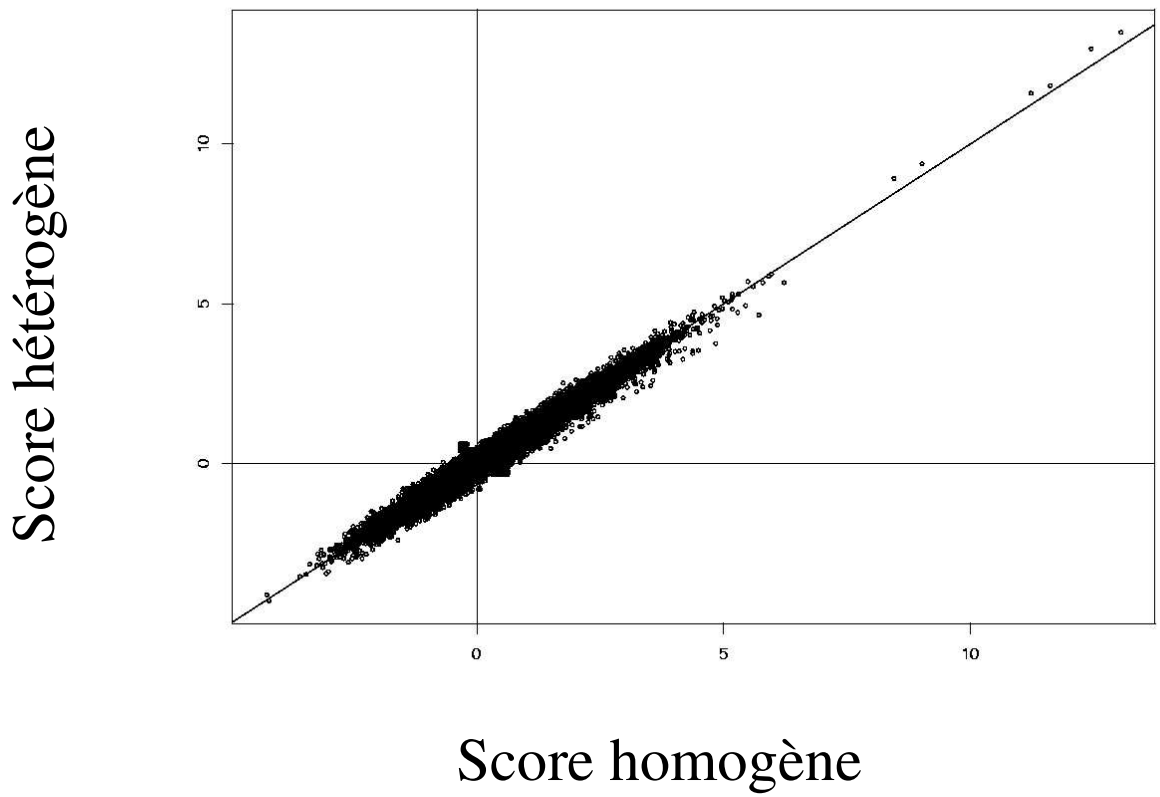
$m = 0$



Données réelles : exemple 2

Mots de longueur $h = 8$, concaténation *E. Coli* + *H. influenzae*

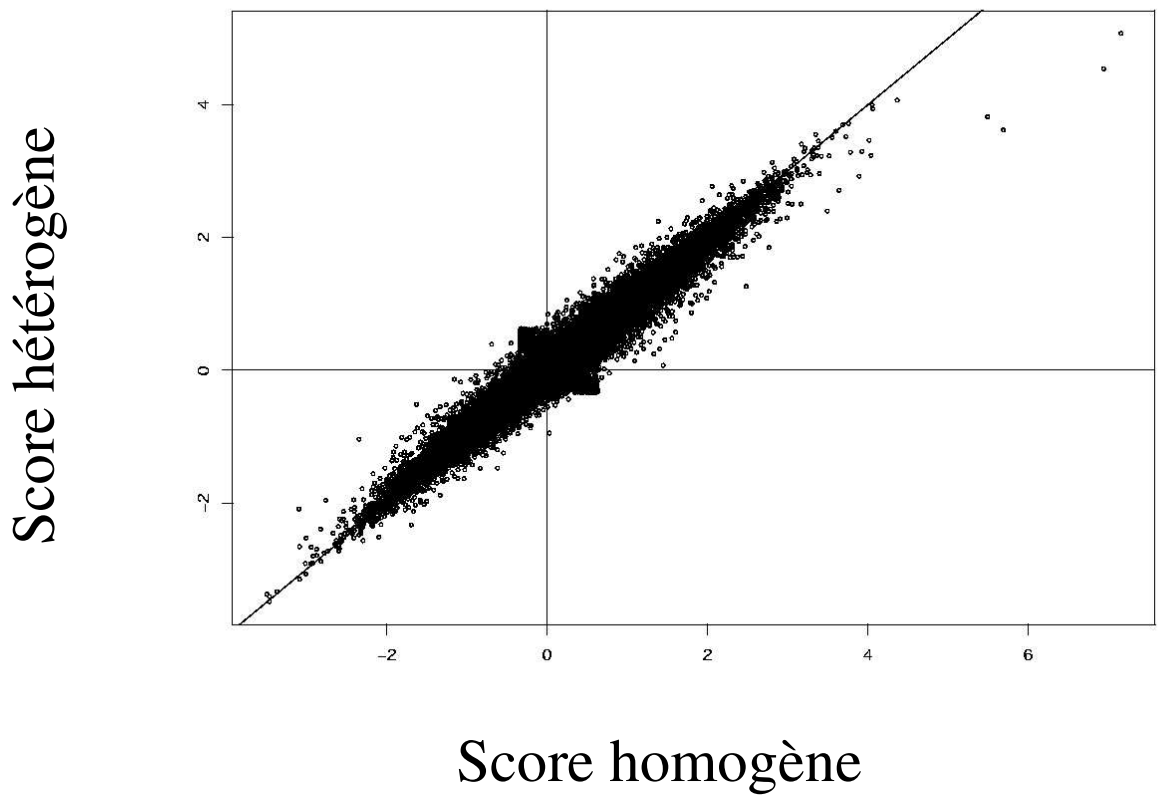
$m = 2$



Données réelles : exemple 2

Mots de longueur $h = 8$, concaténation *E. Coli* + *H. influenzae*

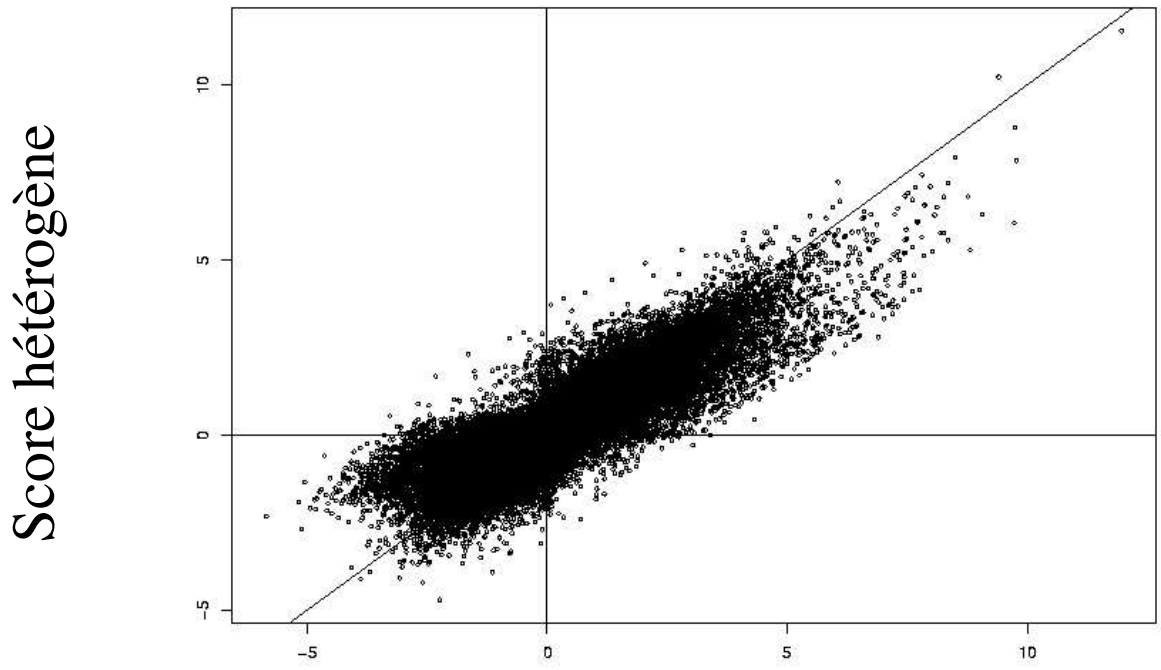
$m = 5$



Données réelles : exemple 3

Mots de longueur $h = 8$, concaténation
Clostridium botulinum (28% GC) + *Pseudomonas aeruginosa* (66% GC)

$m = 2$



Conclusion

Pour rechercher les mots de fréquence exceptionnelle dans un génome :

Deux approx. pour la loi de $N(\mathbf{w})$:
lorsque la séquence est markovienne hétérogène pour $\mathbf{w}$ rare

- par $\mathcal{CP}_{uni}$: bonne si $\rho h/n$ “petit"
- par $\mathcal{CP}_{bic}$: plus précise mais plus longue à calculer

Conclusion

Pour rechercher les mots de fréquence exceptionnelle dans un génome :

Deux approx. pour la loi de $N(\mathbf{w})$:
lorsque la séquence est markovienne hétérogène pour **w** rare

- par $\mathcal{CP}_{uni}$: bonne si $\rho h/n$ “petit"
- par $\mathcal{CP}_{bic}$: plus précise mais plus longue à calculer

En pratique :

- pour mesurer l’exceptionnalité d’un mot : nouvelle méthode qui prend en compte une segmentation donnée a priori
- À préférer à la méthode homogène si la séquence est **réellement hétérogène**

Conclusion

Pour rechercher les mots de fréquence exceptionnelle dans un génome :

Deux approx. pour la loi de $N(\mathbf{w})$:
lorsque la séquence est markovienne hétérogène pour **w** rare

- par $\mathcal{CP}_{uni}$: bonne si $\rho h/n$ “petit"
- par $\mathcal{CP}_{bic}$: plus précise mais plus longue à calculer

En pratique :

- pour mesurer l’exceptionnalité d’un mot : nouvelle méthode qui prend en compte une segmentation donnée a priori
- À préférer à la méthode homogène si la séquence est **réellement hétérogène**

Autre conséquence : recherche dans plusieurs génomes simultanément

Conclusion

Pour rechercher les mots de fréquence exceptionnelle dans un génome :

Deux approx. pour la loi de $N(\mathbf{w})$:
lorsque la séquence est markovienne hétérogène pour **w** rare

- par $\mathcal{CP}_{uni}$: bonne si $\rho h/n$ “petit"
- par $\mathcal{CP}_{bic}$: plus précise mais plus longue à calculer

En pratique :

- pour mesurer l’exceptionnalité d’un mot : nouvelle méthode qui prend en compte une segmentation donnée a priori
- À préférer à la méthode homogène si la séquence est **réellement hétérogène**

Autre conséquence : recherche dans plusieurs génomes simultanément

Disponibilité : en cours d’implémentation dans R’MES
(http://genome.jouy.inra.fr/ssb/rmes/)

Merci de votre attention!