Détection de courts segments inversés dans les génomes : méthodes et applications

David Robelin

dirigé par

Bernard Prum

Laboratoire Statistique et Génome,
Génopole, EVRY,
France

Motivations Biologiques

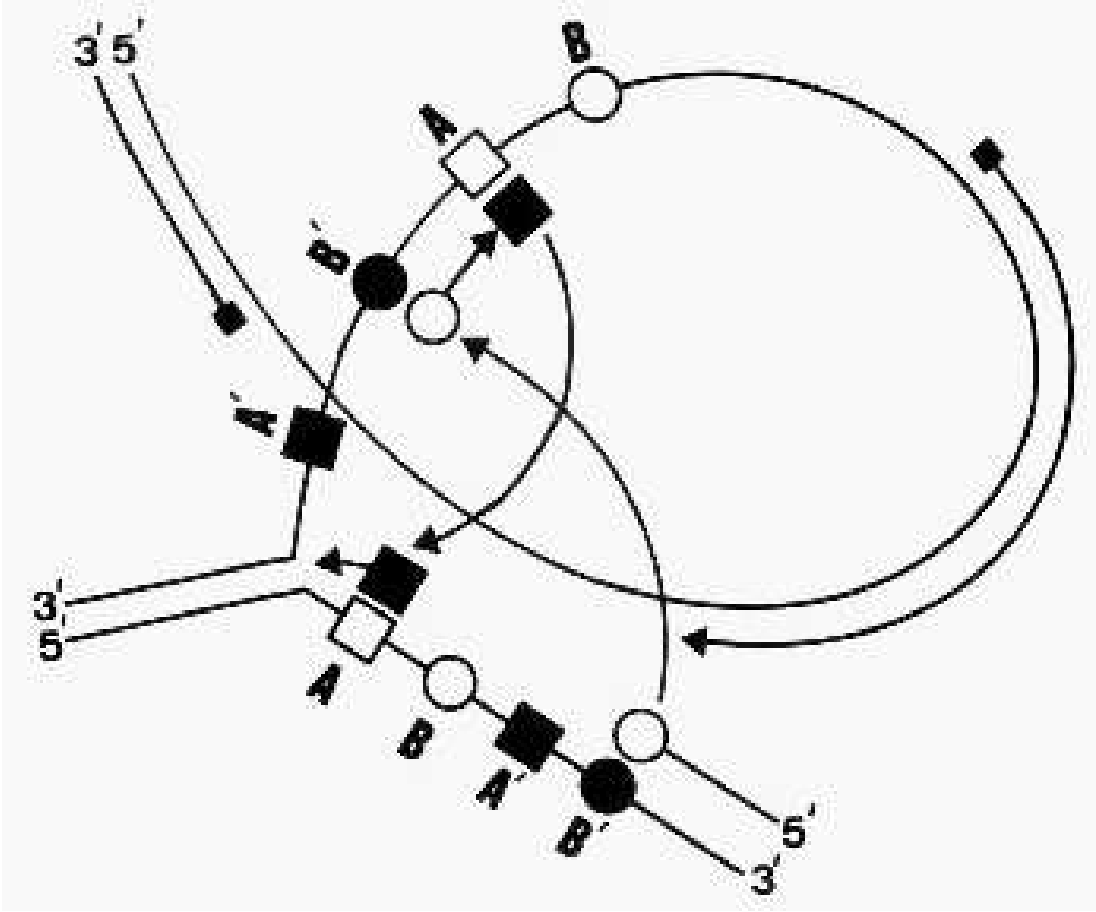
• **Mutations ponctuelles** ne suffisent pas à expliquer la variabilité génétique observée aujourd'hui (Ochman et al. 2000) :
→ **duplications, inversions, transferts genetiques horizontaux, transpositions, rearrangements**

• Mutations à grande échelle peuvent être initiées par des mutations à petites échelles:
Ex. : chez les mamifères, les motifs courts répétés peuvent initier la formation de grands palindromes, eux mêmes associés à l'amplification génique (Tanaka et al., 2002).

• **Inversion de petit fragments d'ADN** (5Bp to 100Bp) est suspectée être un vecteur important de diversité génétique. (Goldstein et al. 2000, 2003).

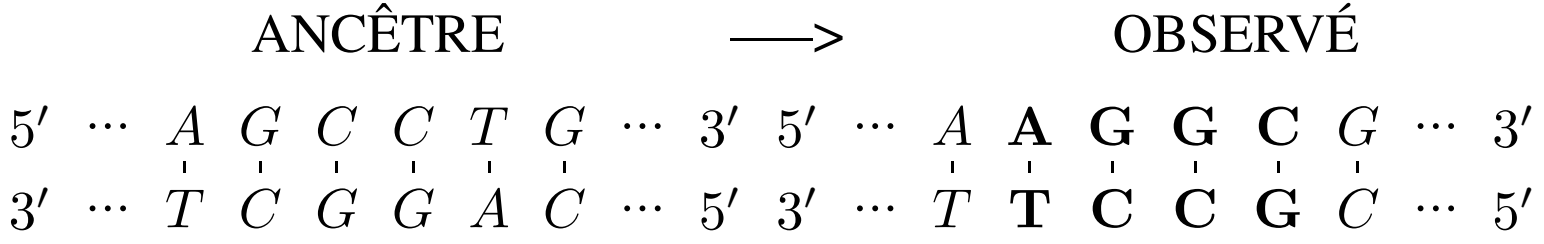
Processus biologique hypothétique (Gordon et Halliday, 1995).

Erreur lors de la réplication de l'ADN.



Dincom : **D**NA **In**verse **Com**plémentaire

Figure 1: Un exemple de **dincom**.



Objectif

- **Détecter** et **localiser** les éventuels *dincoms* d'une séquence donnée
- Associer **un niveau de confiance** à chaque *dincom* détecté

Plan de l'exposé

1. Modélisation Markovienne des séquences
2. Méthodes de détection des segments inversée
3. r plus grandes valeurs du score local
4. Applications et logiciel SIC

Modélisation Markovienne des séquences

Modèle de Markov

La séquence est modélisée par une **chaîne de Markov** homogène supposée à l'état stationnaire X: $\forall u, v \in \mathcal{A}$

$$\mathrm{P}(u, v) = \mathbb{P}(X_{i+1} = v | X_i = u)$$

$$\mu(u) = \mathbb{P}(X_i = u)$$

Chaîne de Markov d'ordre m :

$$\mathbb{P}(X_{i+1} = v | X_1 ... X_i) = \mathbb{P}(X_{i+1} = v | X_{i-m} ... X_i)$$

Nombre de paramètres linéairement indépendants : $(|\mathcal{A}| - 1) \times |\mathcal{A}|^m$

La **chaîne inversée** X^- est également Markovienne (alphabet complémentaire) : $\forall u, v \in \{a, c, g, t\}$,

$$\mu^-(u) = \mu(\bar{u})$$

$$\mathrm{P}^-(u, v) = \mathrm{P}(\bar{v}, \bar{u}) \frac{\mu(\bar{v})}{\mu^-(u)}$$

car, $\mathbb{P}\left((X_i^-, X_{i+1}^-) = (u, v)\right) = \mathbb{P}\left((X_i, X_{i+1}) = (\bar{v}, \bar{u})\right)$

Méthodes de détection de segments inversés

Taille de segments connues

Méthode par fenêtre glissante

Séquence d'intérêt : $s_1, ..., s_n$

Pour chaque fenêtre de taille l, le rapport de vraisemblance suivant :

$$T_i = \log\left(\frac{\mathbb{P}^-(s_i, .., s_{i+l-1})}{\mathbb{P}^+(s_i, .., s_{i+l-1})}\right),\ i = 1, ..., n-l+1$$

où $\mathbb{P}^+(s_i, .., s_{i+l-1})$ (resp. $\mathbb{P}^-(s_i, .., s_{i+l-1})$) est la probabilité d'observer $(s_1, .., s_l)$ sous le modèle de Markov X^+ (resp. X^-).

l est choisi en fonction de **connaissances biologiques a priori**

Taille de segments connues (2)

Distribution de T_i quand il n'y a pas de dincom

• l est "petit" :

On considère les 4^l segments différents dont on calcule la probabilité sous chacun des modèles

$\rightarrow$ **Distribution Exacte** de T_i

Problème : complexité exponentielle en l.

Taille de segments connues (3)

Distribution de T_i quand il n'y a pas de dincom (2)

• l est grand :

$$T_i = \log\left(\frac{\mu^-(s_i)}{\mu^+(s_i)}\right) + \sum_{u,v\in\{a,c,g,t\}} \log\left(\frac{\mathrm{P}^-(v|u)}{\mathrm{P}^+(v|u)}\right) \times N_{i,i+l-1}(uv)$$

où $N_{i,i+l-1}(uv)$: **comptage du mot "uv"** dans la séquence $s_i, .., s_{i+l-1}$.

$\{N_{i,i+l-1}(uv), \forall(u,v) \in \mathcal{A}^2\}$ est asymptotiquement **gaussien**

$\rightarrow T_i$ **est asymptotiquement gaussien** quand $l \rightarrow \infty$

$\mathbb{E}[T_i]$ et $\mathbb{V}[T_i]$ calculées avec $\mathbb{E}[N_{i,i+l-1}(uv)]$ de
$\mathbb{C}\mathrm{ov}[N_{i,i+l-1}(uv), N_{i,i+l-1}(u'v')]$

Taille de segments connues (4)

Test sur la globalité de la séquence

H_0 : Il n'y a pas **de dincom** dans la séquence

H_1 : **Au moins un dincom** dans la séquence

On cherche des valeurs élevées de T_i :

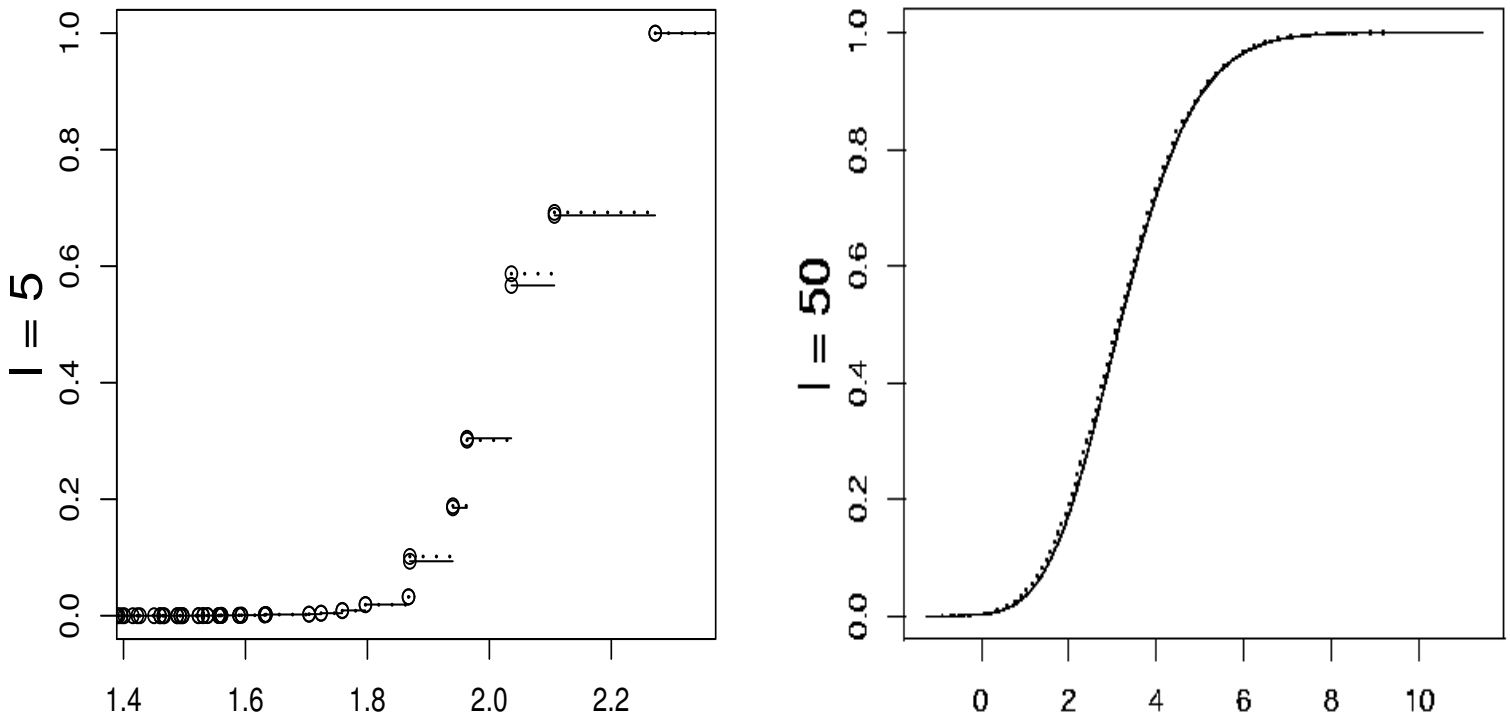
$$S_l^n = \max_{i=1,\ldots,n-l+1} T_i$$

Distribution de S_l^n quand il n'y a pas de dincom :

- **Approximation "Product-type"** (Glaz and Balakrishnan,1999)

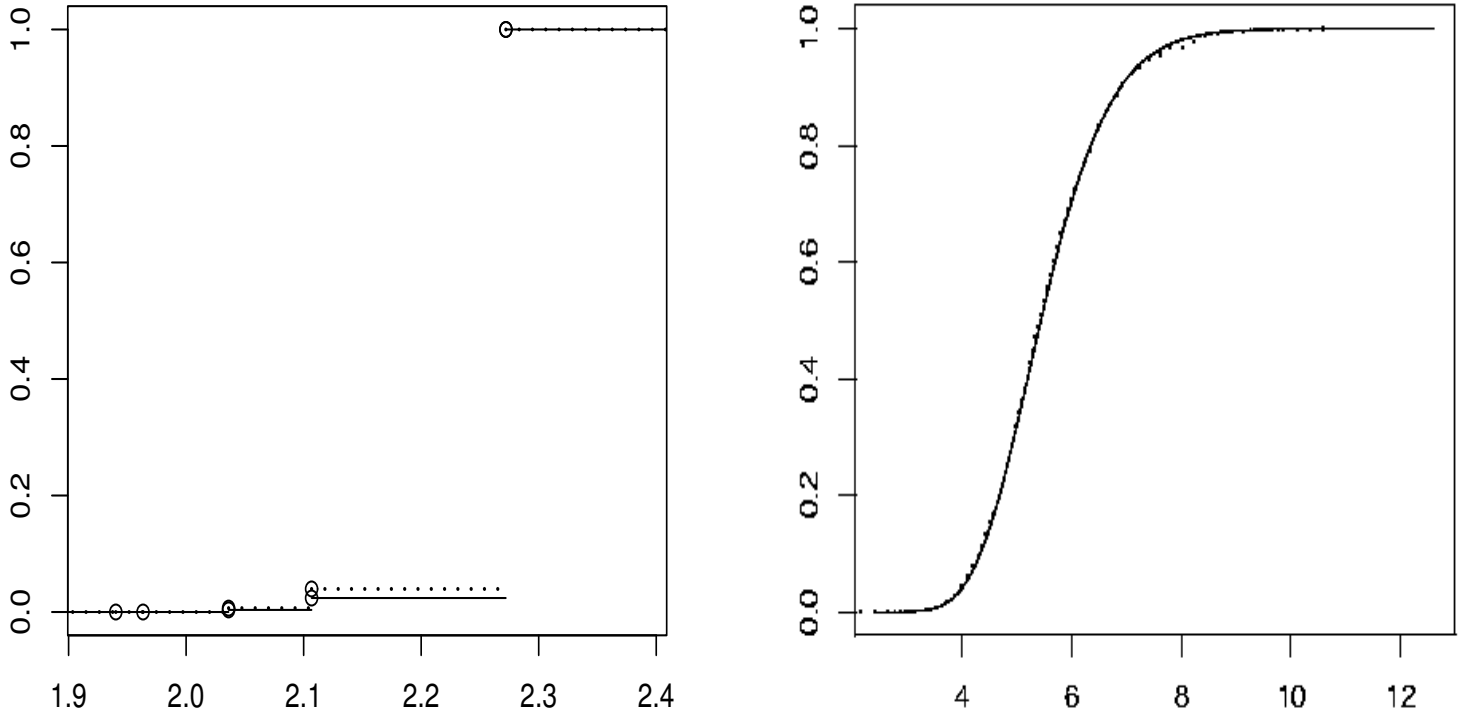
$$\mathbb{P}(S_l^n \leq s) \approx \mathbb{P}(S_l^{3l} < s) \left(\frac{\mathbb{P}(S_l^{3l} < s)}{\mathbb{P}(S_l^{2l} < s)} \right)^{n/l-3}$$

$\mathbb{P}(S_l^{3l} < s)$, $\mathbb{P}(S_l^{2l} < s)$: obtenus par Monte-Carlo (Importance Sampling).



Fonction de répartition de S_l^n pour $n = 1000$ et $l = 5$ et 50.

— : Monte Carlo ; ... : Approximation product-type



Fonction de répartition de S_l^n pour $n = 10000$ et $l = 5$ et 50.

— : Monte Carlo ; ... : Approximation product-type

Taille de segments inconnues

But : Ne pas fixer a priori la taille du dincom

Méthode de score local

On note $Y_i = \log\left(\frac{\mathrm{P}(X_i,X_{i+1})}{\mathrm{P}^-(X_i,X_{i+1})}\right)$

$$Y_i + Y_{i+1} + ... + Y_j = \log\left(\frac{\mathbb{P}^-(X_i,..,X_j|X_i)}{\mathbb{P}^+(X_i,..,X_j|X_i)}\right)$$

On cherche la sous séquence associée à :

$$H^n = \max_{1\leq i\leq j\leq n-1} Y_i + ... + Y_j$$

Méthode de Score local (2)

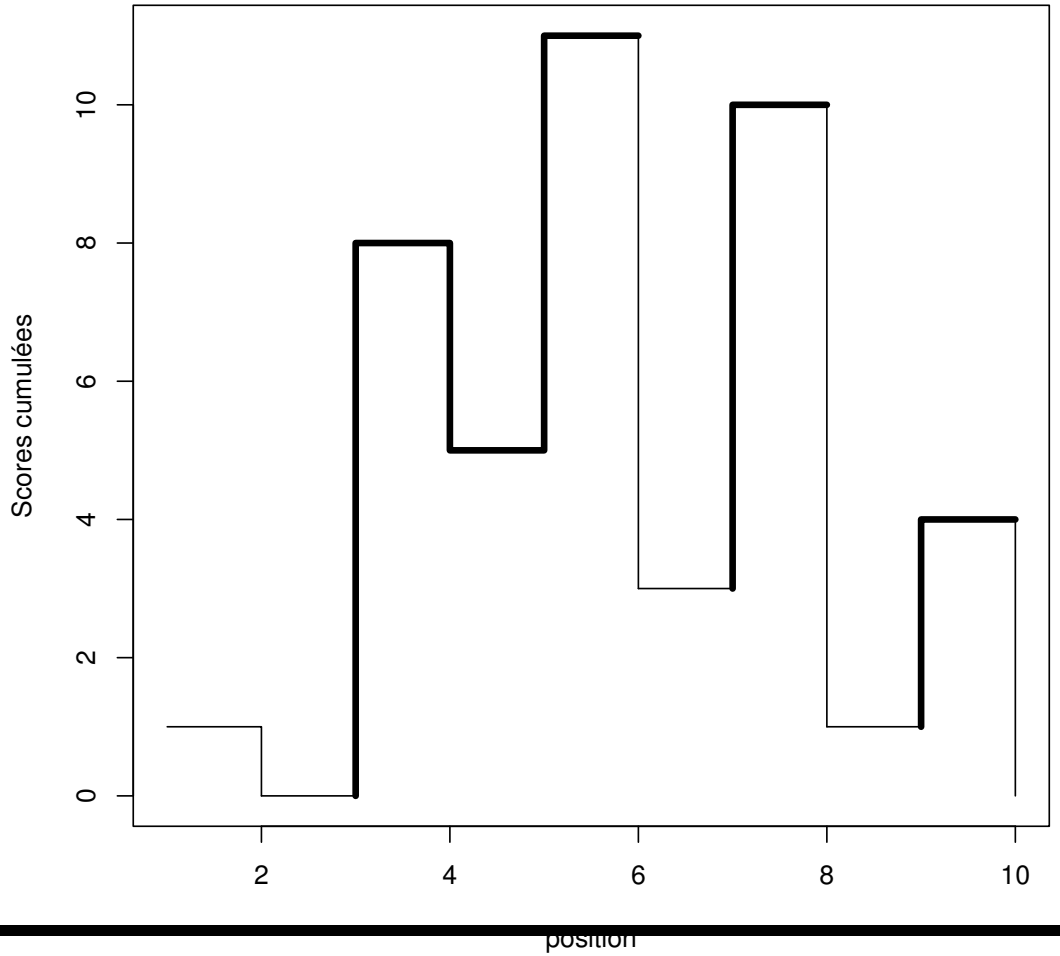
En pratique
(algorithme en $O(n)$) :

$$H^n = \max_{i=1,\ldots n-1} (S_i - \min_{j<i} S_j)$$

où

$$S_i = \sum_{k=1}^{i} Y_k$$

est la somme cumulée au rang i
$S_{i+k} - S_i = Y_{i+1} + ... + Y_{i+k}$



Méthode de score local (3)

Distribution de H^n quand il n'y a pas de dincom

$H^n \sim$ Gumbel (Karlin and Dembo, 1992), car $\mathbb{E}^+[Y_i] < 0$

$$\lim_{n \to +\infty} \mathbb{P}\left(H_n \leq \frac{\ln n}{\lambda} + x \right) = \exp\left(-K^* \exp\left(-\lambda x\right)\right)$$

Estimation de K^* et λ

$$\ln\left(-\ln\left(F(y)\right)\right) \approx \ln K^* - \lambda\, y + \ln n$$

$F(y)$ est estimée par Monte Carlo sur des séquences de tailles convenables.
K^* et λ sont déduits à l'aide d'une régression linéaire.

Méthode de score local (4)

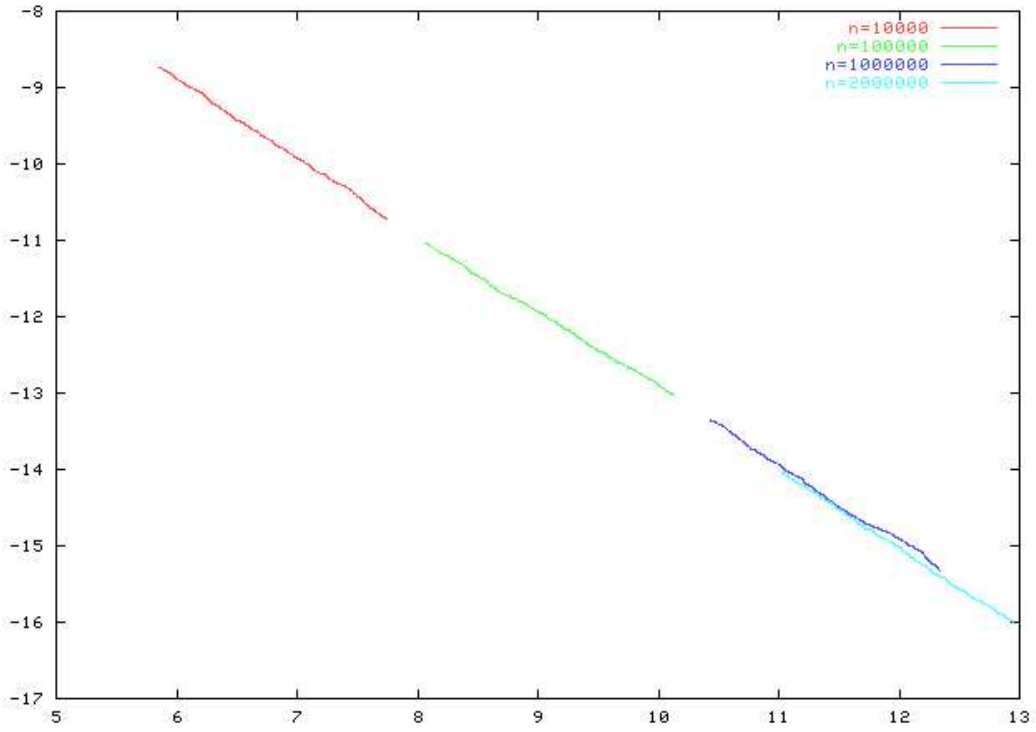


Figure 2: Fonction de répartition linéarisée de H_n normalisée par $\log(n)$ pour $n = 1000, 10000, 100000, 200000$

Performances des méthodes : Fenêtre vs Score local

Chaîne Reversible $\rightarrow$ puissance nulle !

Besoin d'une distance entre les modèles P^+ et P^-

$\rightarrow$ Mesure de degré d'orientation de la chaîne de Markov

Taux d'entropie relative :

$$Er(X^+, X^-) = \sum_{u,v \in \{a,c,g,t\}} \mathbb{P}^+(u,v) \log \left(\frac{\mathrm{P}^+(v|u)}{\mathrm{P}^-(v|u)} \right)$$

Remarque :

$$Er(X^+, X^-) = -\mathbb{E}^+[Y_i]$$

Evaluation comparative des performances des méthodes - Etude de simulation

Trois composantes :

• **Taille de la séquence** $n = 1000$, 5000 et 10000

• **Taille du dincom** $l = 5$, 10, 15, 20, 30, 40, 50, 75, 100 et 150

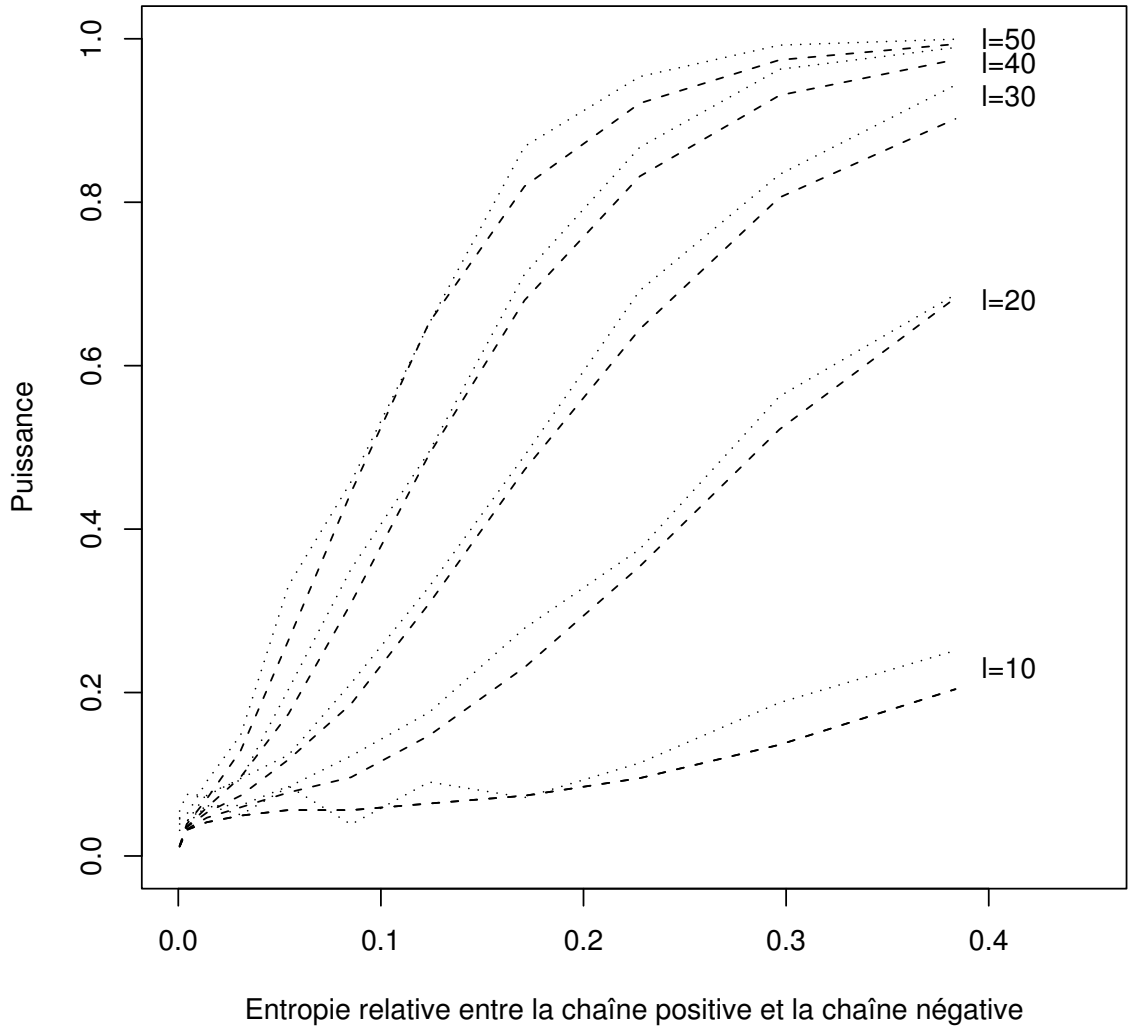
• **Degré d'orientation** de la chaîne de Markov simulée :

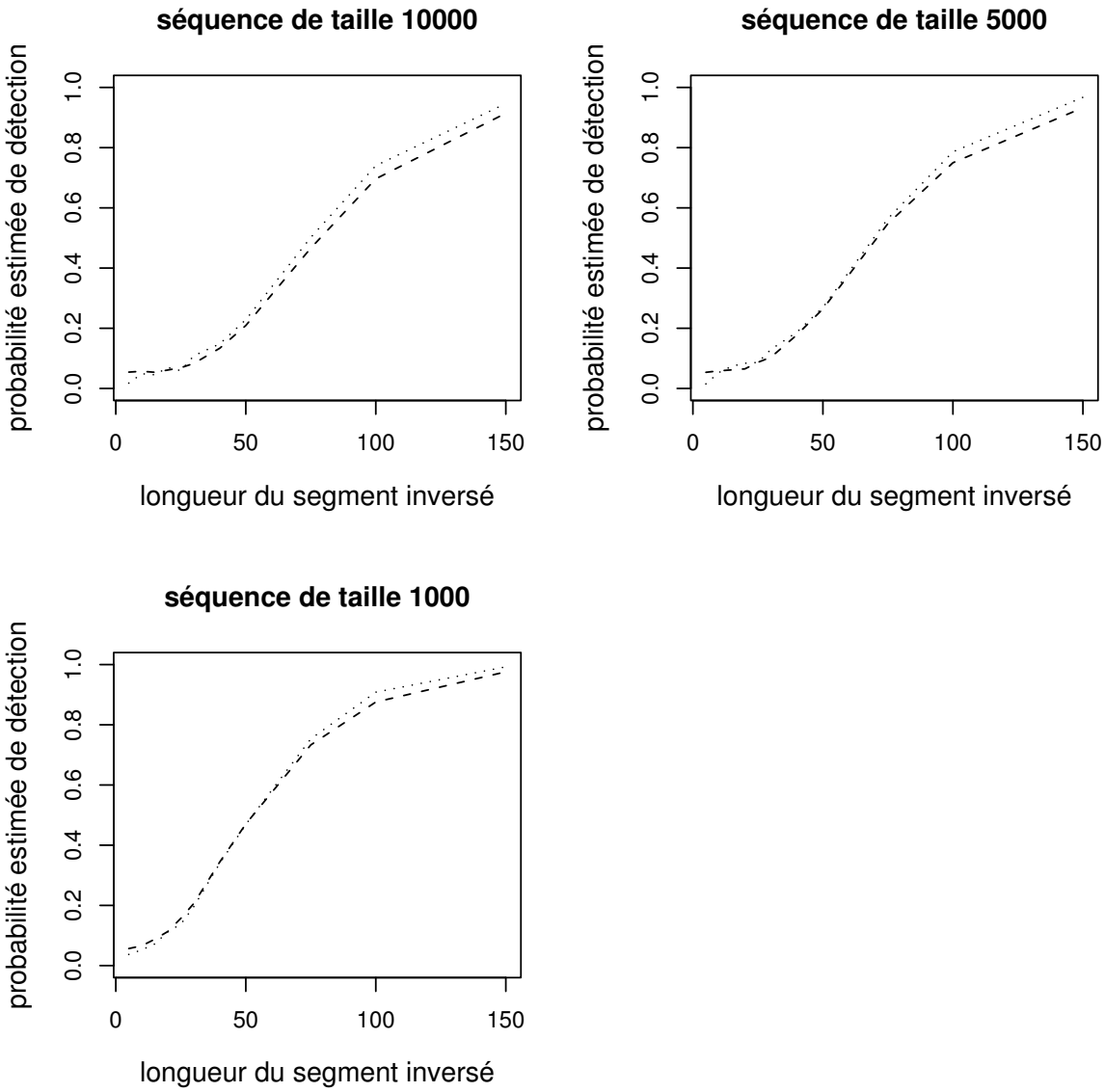
P^+ matrice estimée sur la séquence HIV1.

P^- matrice de la chaîne inverse.

On définit $\mathrm{P}_p^+ = p * \mathrm{P}^+ + (1-p) * \mathrm{P}^-$

p	0.5	0.6	0.7	0.9	**1.0**	1.1	1.2	1.3	1.4	1.5
$Er(\mathrm{P}_p^+, \mathrm{P}_p^+)$	7.10^{-4}	4.10^{-3}	0.01	0.05	**0.09**	0.12	0.17	0.23	0.30	0.38


Puissance
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
l=50
l=40
l=30
l=20
l=10
Entropie relative entre la chaîne positive et la chaîne négative


séquence de taille 10000
probabilité estimée de détection
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
longueur du segment inversé
séquence de taille 5000
probabilité estimée de détection
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
longueur du segment inversé
séquence de taille 1000
probabilité estimée de détection
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
longueur du segment inversé

Score Local : r plus grandes valeurs

Définition - Exemple

La $r^{ième}$ plus grandes valeurs de score local : plus grande valeur de score local dans la séquence **privée** des segments réalisant les $r-1$ plus valeurs supérieurs de score local.

position :	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
score :	1	-1	-1	1	1	-1	1	1	-1	-1	1	1	-1	-1	-1	1
score cumulée:	1	0	-1	0	1	0	1	2	1	0	1	2	1	0	-1	0

4 Segments de score maximal :

Position :	$\{1\}$	$\{4,5,6,7,8\}$	$\{11,12\}$	$\{16\}$
Valeur :	1	3	2	1

Algorithme de recherche de **complexité linéaire** avec la taille de la séquence (Ruzzo et Tompa, 1999).

Loi de la plus grande valeur de score local

Temps de records des sommes cumulées :

$$T_0 = 0 \text{ et } T_{k+1} = \inf\{i : i > T_k \text{ et } S_i \leq S_{T_k}\}$$

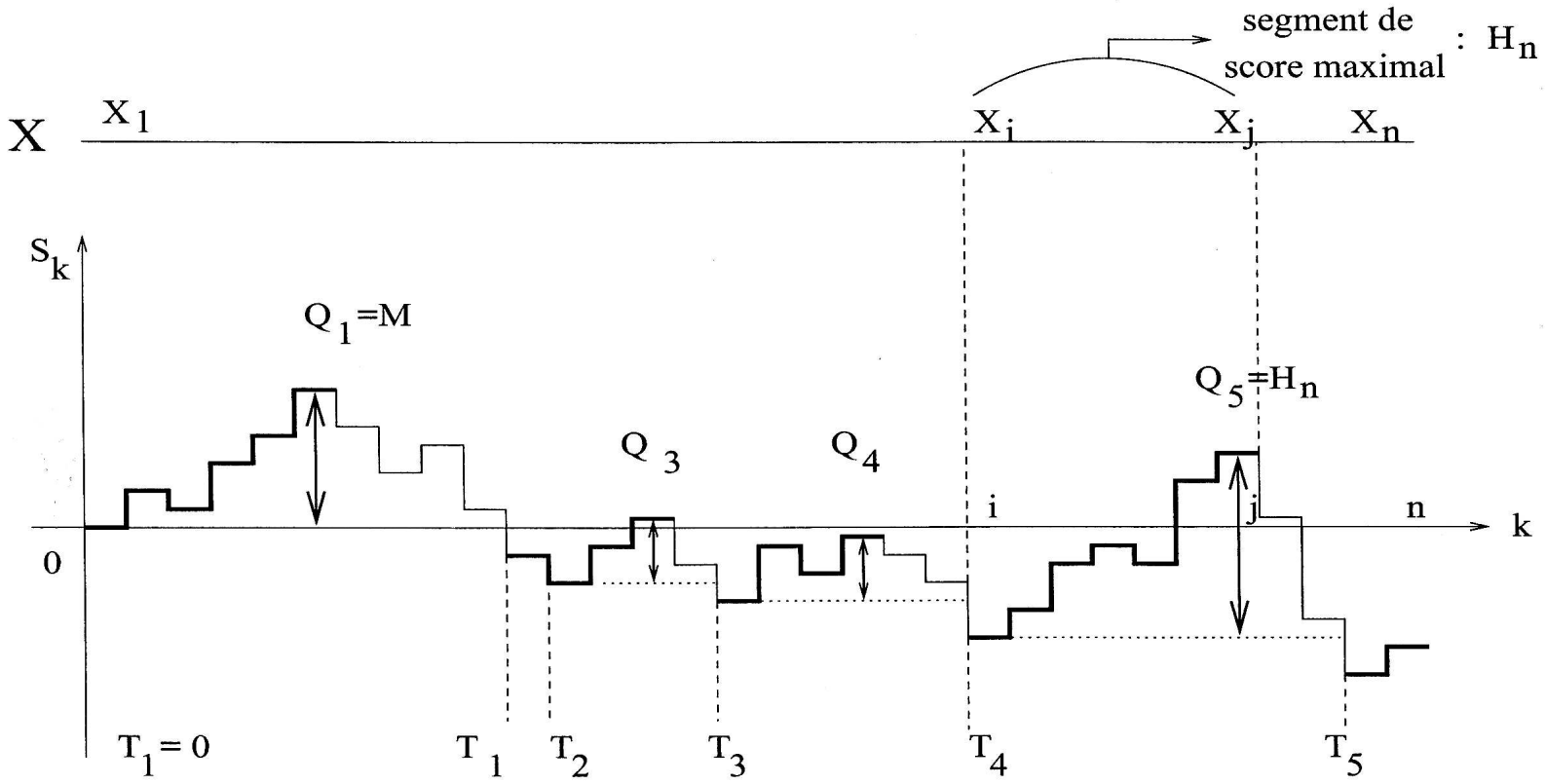
Maximum des sommes cumulées sur un “bloc”.

$$Q_i = \max_{T_i \leq k < T_{i+1}} (S_k - S_{T_i})$$

Le score local se réécrit : $H_n = \max_{i=1,\ldots,N_n} Q_i$

Loi de la plus grande valeur de score local (2)

Démarche de Karlin et Dembo (1992)



Loi de la plus grande valeur de score local (3)

1. Les v.a. Q_i sont **indépendantes et identiquement distribuées**
2. Loi de Q_i est dans le domaine d'attraction de la loi de **Gumbel**. On note $M_k = \max_{i=1,\ldots,k} Q_i$

$$\lim_{k\to\infty} \mathbb{P}((M_k - b_k)/a_k < x) = e^{-e^{-x}}$$

3. N_n/n tend presque sûrement vers une **constante**
4. Finalement, $\lim_{n\to\infty} \mathbb{P}((H_n - b'_n)/a'_n < x) = e^{-e^{-x}}$

$$\begin{aligned} a'_n &= \frac{1}{\lambda} \\ b'_n &= \frac{\ln n + \ln(C/\mu)}{\lambda} \end{aligned}$$

Loi des r plus grandes valeurs de score local

Loi jointe asymptotique des r plus grandes valeurs d'un échantillon connue (cas attraction Gumbel) :

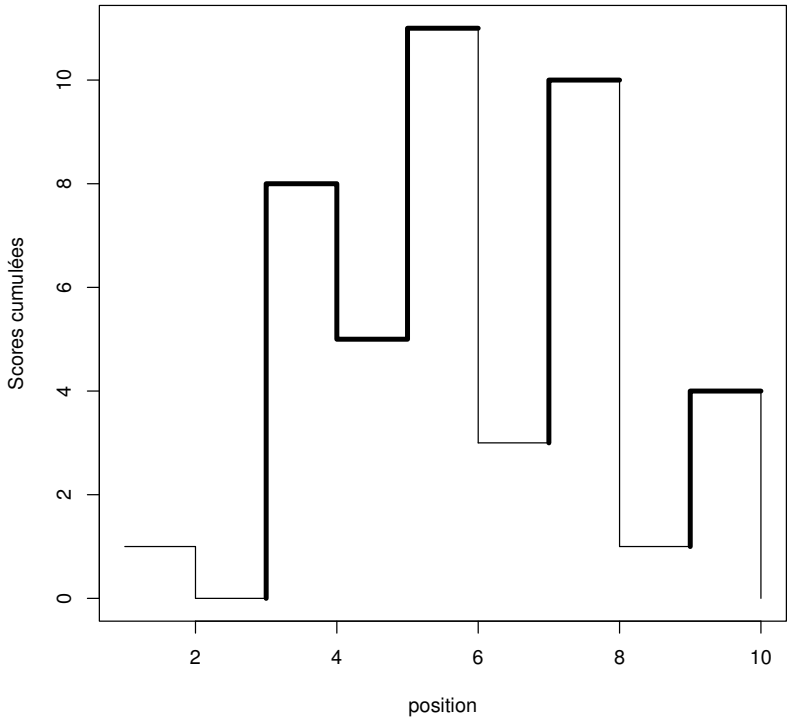
$$\left[\frac{M_n^{(1)} - b_n}{a_n}, ..., \frac{M_n^{(r)} - b_n}{a_n}\right] \rightarrow$$

$$\exp\left\{-\exp\left(-\frac{M_n^{(r)} - \mu}{\sigma}\right)\right\} \prod_{i=1}^{r} \sigma^{-1} \exp\left(-\frac{M_n^{(i)} - \mu}{\sigma}\right)$$

avec $M_n^{(k)}$ la $k^{ième}$ plus grande valeur.

Loi des r plus grandes valeurs de score local

Problème : deuxième plus grande valeur n'est pas nécessairement $Q^{(2)}$



Détermination des scores locaux significatifs

H_0 : les données $Y_1 ... Y_n$ sont indépendantes et identiquement distribuées.

Deux hypothèses alternatives possibles :

- H_1 : les données $Y_1 ... Y_n$ ne sont pas indépendantes et identiquement distribuées. Il y a au moins $\mathbf{1}$ accumulation de valeurs élevées.
- H_1' : les données $Y_1 ... Y_n$ ne sont pas indépendantes et identiquement distribuées. Il y a au moins $\mathbf{r}$ accumulations de valeurs élevées.

Deux formes de régions de rejet :

$$H_1 : \quad 1 - \mathbb{P}(M_n^{(1)} < s_n^{(1)}, ..., M_n^{(r)} < s_n^{(r)} \mid H_0) = \alpha$$
$$H_1' : \quad \mathbb{P}(M_n^{(1)} > s_n'^{(1)}, ..., M_n^{(r)} > s_n'^{(r)} \mid H_0) = \alpha$$

Détermination des scores locaux significatifs (2)

Démarche classique : Le seuil de significativité s est déterminé en fonction de la loi de $H_n^{(1)}$

$\rightarrow$ Tous les scores dépassant s sont considérés.

- Assure que le risque global est inférieur à α
- Démarche conservative

D'un point de vue biologique, plus la valeur d'un score local est élevée, plus le segment qui l'a engendré est caractéristique du phénomène étudié.

$\rightarrow$ Permet de "hierarchiser" les hypothèses

Détermination des scores locaux significatifs (3)

On définit la série d'hypothèses suivantes :

$H_0^{'(i)}$: Il y a (au plus) $i-1$ accumulations de valeurs élevées

contre

$H_1^{'(i)}$: Il y a au moins i accumulations de valeurs élevées

Démarche :

1. Effectuer le test de l'hypothèse $H_0^{'(i)}$ contre $H_1^{'(i)}$.
2. Si le test est significatif au risque α_i choisi, alors
   $i \leftarrow i+1$ et continuer au point 1.
3. Sinon conclure :
   - si $i > 1$ alors "Les $i-1$ plus grands scores sont significatifs
   - sinon "Il n'y a pas de score significatif dans la séquence".

Détermination des scores locaux significatifs (4)

Remarque : Le risque de première espèce global de cette démarche vaut α_1

$$\begin{aligned}\mathbb{P}(\text{rejet de } H_0 \mid H_0) &= \mathbb{P}\left(\text{rejeter } H_0^{(1)} \mid H_0\right) \\ &= \alpha_1\end{aligned}$$

Risque de première espèce partiel

$$\alpha'_i = \mathbb{P}\left(\bigcup_{j\geq i}\left\{\text{Rejeter à tort } H_0^j\right\} \mid \bigcap_{j<i}\left\{\text{Rejeter } H_0^j\right\}\right)$$

Le risque de première espèce global est alors α'_1.

Un démarche intéressante doit assurer : $\forall i, \alpha'_i \leq \alpha$

Détermination des scores locaux significatifs (5)

Première idée : la factorisation de la probabilité :

$$\mathbb{P}(H_n^{(1)} > s^{(1)}, ..., H_n^{(r)} > s^{(r)}) =$$

$$\mathbb{P}\left(H_n^{(1)} > s^{(1)}\right) \times \prod_{j=2}^{r} \mathbb{P}\left(H_n^{(j)} > s^{(j)} \mid \left\{H_n^{(k-1)} > s^{(j-1)}\right\}_{k=1,...,j-1}\right)$$

suggère de déterminer les seuils en fonction de la loi

$$\left[H_n^{(j)} \mid \left\{H_n^{(k-1)} > s^{(k-1)}\right\}_{k=1,...,j-1}\right]$$

- Sous H_0, $\mathbb{P}(H_n^{(1)} > s^{(1)}, ..., H_n^{(r)} > s^{(r)}) = \alpha^r \qquad (< \alpha)$
- puissance pour détecter k valeurs élevées correspond à un risque de première espèce α^k
- puissance pour détecter 1 valeur élevée est la même que pour la démarche classique

Détermination des scores locaux significatifs (6)

Problème : Difficulter de calcul des fonctions de répartition de

$$\left[H_n^{(j)} \mid \left\{ H_n^{(k-1)} > s^{(k-1)} \right\}_{k=1,\ldots,j-1} \right]$$

Possibilité de réutiliser les itérations de Monte-Carlo permettant de déterminer K^* et λ *(cf. rapport)*

Par contre, la loi conditionnée se calcule facilement :

$$[H_n^{(i)} \mid H_n^{(i-1)}, ..., H_n^{(1)}] = [H_n^{(i)} \mid H_n^{(i-1)}]$$

$$\mathbb{P}\left(H_n^{'(i)} \leq h^{'(i)} \mid H_n^{'(i-1)} = h^{'(i-1)} \right) = \exp\left(-e^{-h^{'(i)}} + e^{-h^{'(i-1)}} \right)$$

Détermination des scores locaux significatifs (7)

1. Standardisation des plus grandes valeurs de score local :

$$\forall j, H_n'^{(j)} = \lambda H_n^{(j)} - \log(K \times n)$$

2. On teste au risque α la plus grande valeur à l'aide du seuil $s^{(1)}$ tel que $\exp(-\exp(-s^{(1)})) = 1 - \alpha$.

3. Tant que l'hypothèse $H_0^{(i)}$ est rejetée ,
   Test de la valeur $H_n'^{(i+1)}$ en utilisant une des deux règles de décision suivantes : on calcule le seuil $s^{(i+1)}$ tel que
   (a) **Règle 1** : $\mathbb{P}(H_n'^{(i+1)} > s^{(i+1)} \mid H_n'^{(i)} = s^{(i)}) = \alpha$

   (b) **Règle 2** : $\mathbb{P}(H_n'^{(i+1)} > s^{(i+1)} \mid H_n'^{(i)} = h'^{(i)}) = \alpha$

Avantage de la règle 1 fondée sur le conditionnement par les seuils : les valeurs de $s^{(i)}$ peuvent être tabulées (ici $\alpha = 5\%$) :

1-5	2.970	2.277	1.872	1.584	1.361
6-10	1.178	1.024	0.891	0.773	0.668
11-15	0.572	0.485	0.405	0.331	0.262
16-20	0.198	0.137	0.080	0.026	-0.026
21-25	-0.074	-0.121	-0.165	-0.208	-0.249
26-30	-0.288	-0.326	-0.362	-0.397	-0.431
31-35	-0.464	-0.496	-0.526	-0.556	-0.585
36-40	-0.613	-0.641	-0.667	-0.693	-0.719
41-45	-0.743	-0.767	-0.791	-0.814	-0.836
46-50	-0.858	-0.880	-0.901	-0.922	-0.942

Détermination des scores locaux significatifs (8)

Etude de simulation des puissances respectives des deux règles

Modèle i.i.d. avec : $\mathbb{P}(X = 0) = 0.5$, $\mathbb{P}(X = 1) = 0.1$, $\mathbb{P}(X = 2) = 0.2$, $\mathbb{P}(X = 3) = 0.2$

1. La taille de la séquence : $n = 1000, 10000, 100000$.
2. La fonction de score :
   (a) $S_1(0) = -4, S_1(1) = -3, S_1(2) = 1, S_1(3) = 3, \mathbf{E}[\mathbf{S_1}(\mathbf{X})] = \mathbf{-1.5}$
   (b) $S_3(0) = -16, S_3(1) = -12, S_3(2) = 4, S_3(3) = 12, \mathbf{E}[\mathbf{S_3}(\mathbf{X})] = \mathbf{-6}$
3. Le nombre de plages sous H_1 : $r = 0, 5, 10, 20$
4. La longueur des plages sous H_1 : $l = 10, 20$

Sous H_1, on utilise **l'opposé** des scores sous H_0.

Nombre moyen de d'hypothèses $H_0^{(i)}$ rejetées.

		S_1					
n	r / l	5		10		20	
1000	10	2.20	1.83	5.48	4.00	12.94	8.49
	20	4.15	3.71	8.36	7.38	9.04	7.79
100000	10	0.19	0.18	0.42	0.34	0.96	0.72
	20	2.50	2.05	5.26	4.03	11.59	8.23

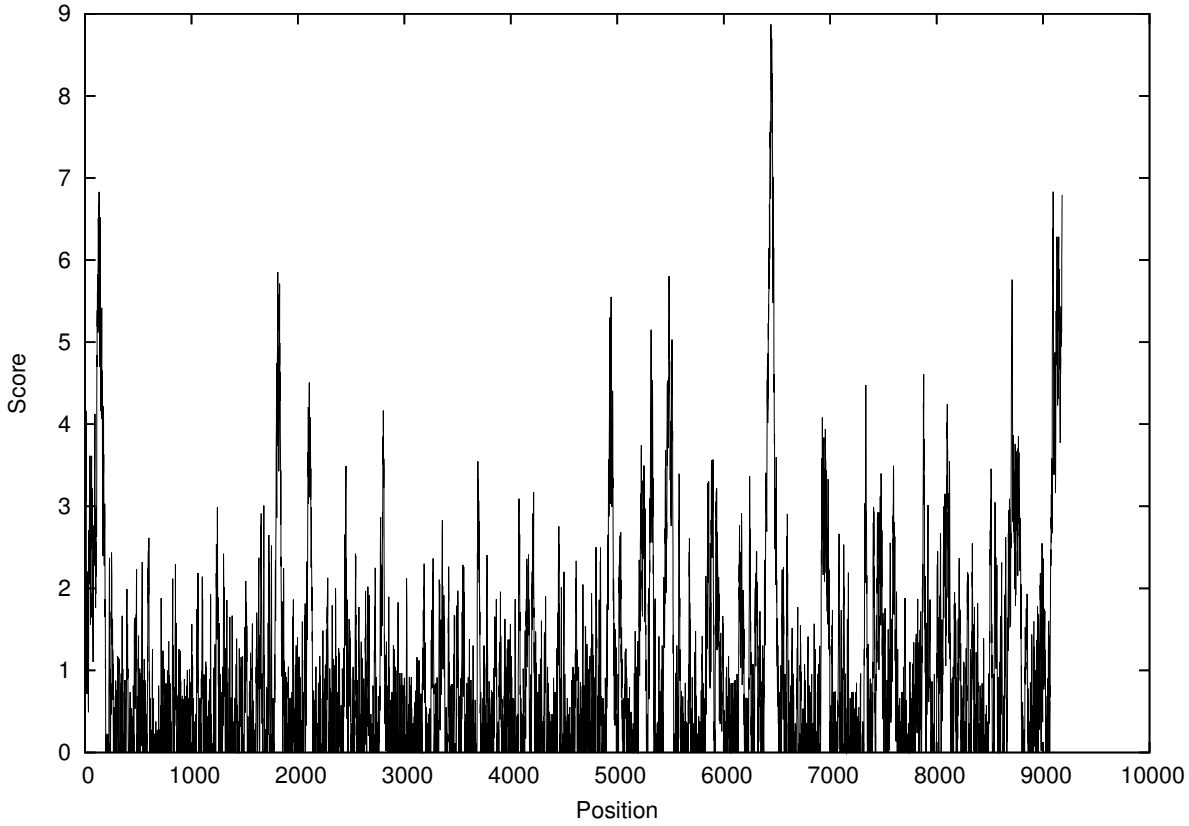
n : taille de la séquence ; r : nombre de segments sous H_1 ; l : taille des segments sous H_1. Chiffre de **gauche** : règle de décision numéro 1 (conditionnement par le seuil précédent), et chiffre de **droite** la régle de décision numéro 2 (conditionnement par la valeur observée précédente)

Applications et logiciels

Matériels et méthodes

- Séquence issues du serveur du NCBI
- Matrice de transition P^+ est **estimée** sur la séquence :
  Imprécision liée à l'estimation est négligée dans nos méthodes.
- Ordre de Markov : 1 à 4
- Organismes : VIH-1, SRAS, Bactériophage Lambda

Virus de l'Immunodéficience Humaine VIH1



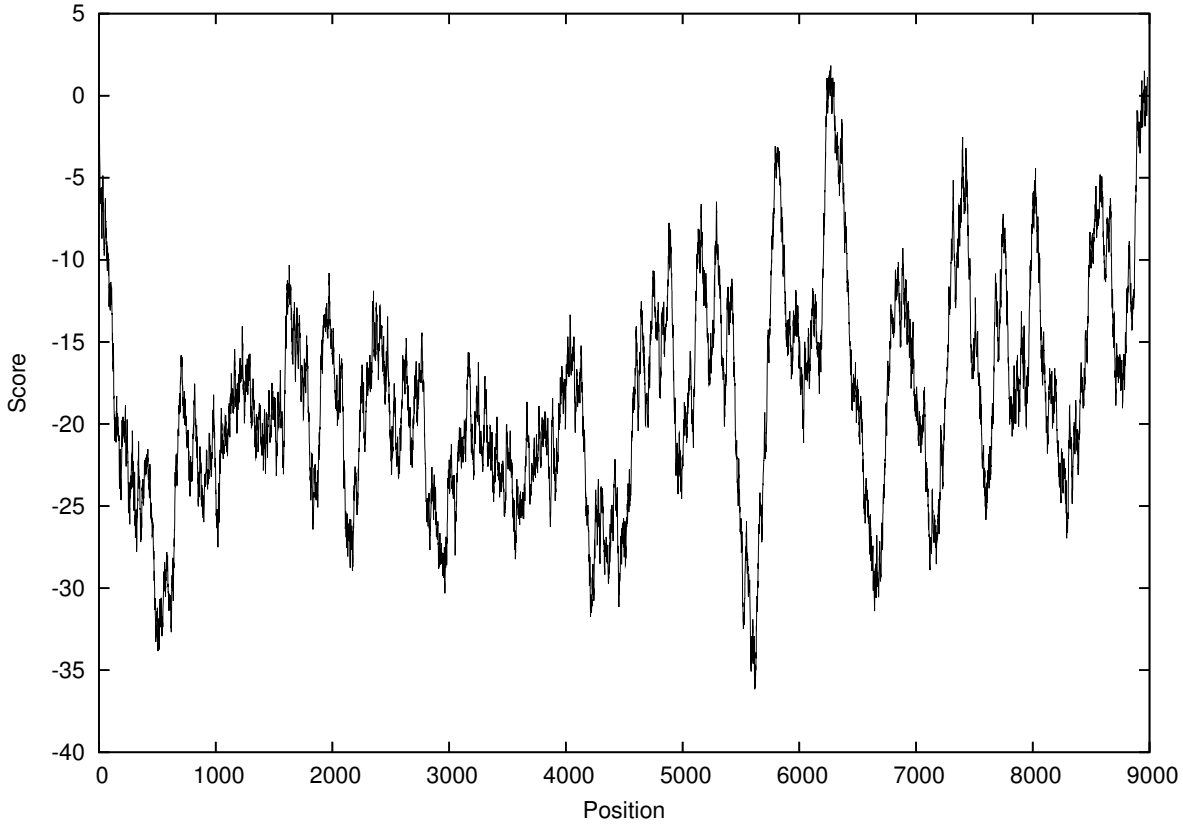
Score de retournement $S_k - \min_{i<k} S_k$

Virus de l'Immunodéficience Humaine VIH1

Ordre	Position	Value	p-value
1	1 - 174	4.64	0.01
	gtctctctggttagaccagatctgagcctgggagctctctggctaactagggaaccc		
	actgcttaagcctcaataaagcttgccttgagtgcttcaagtagtgtgtgcccgtct		
	gttgtgtgactctggtaactagagatccctcagacccttttagtcagtgtggaaaatctc		
1	9065 - 9179	2.90	0.05
	gtgctttttgcctgtactgggtctctctggttagaccagatctgagcctgggagctc		
	tctggctaactagggaacccactgcttaagcctcaataaagcttgccttgagtgctt		
2	6377 - 6444	2.60	0.07
	aggcctgtccaaaggtatcctttgagccaattcccatacattattgtgccccggctggttttgcgatt		

Retournements détectés par la méthode du score local.

Virus de l'Immunodéficience Humaine VIH1



Statistique de fenêtre glissante (ordre 2) et fenêtre de taille 200.

Syndrome Respiratoire Aigu Sévère

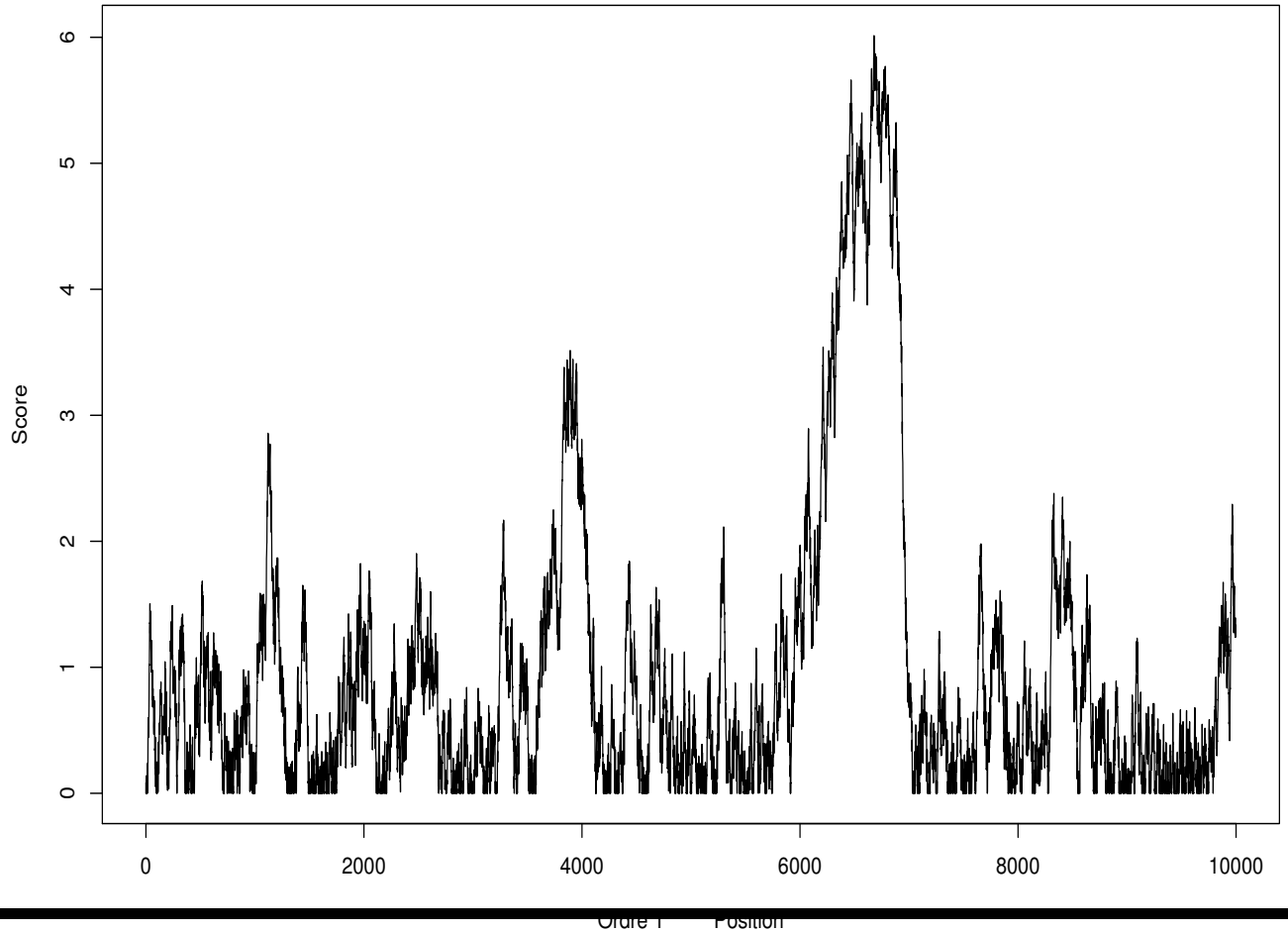
Le génome est découpé arbitrairement en trois parties de tailles 10000.

Retournements significatifs détectés :

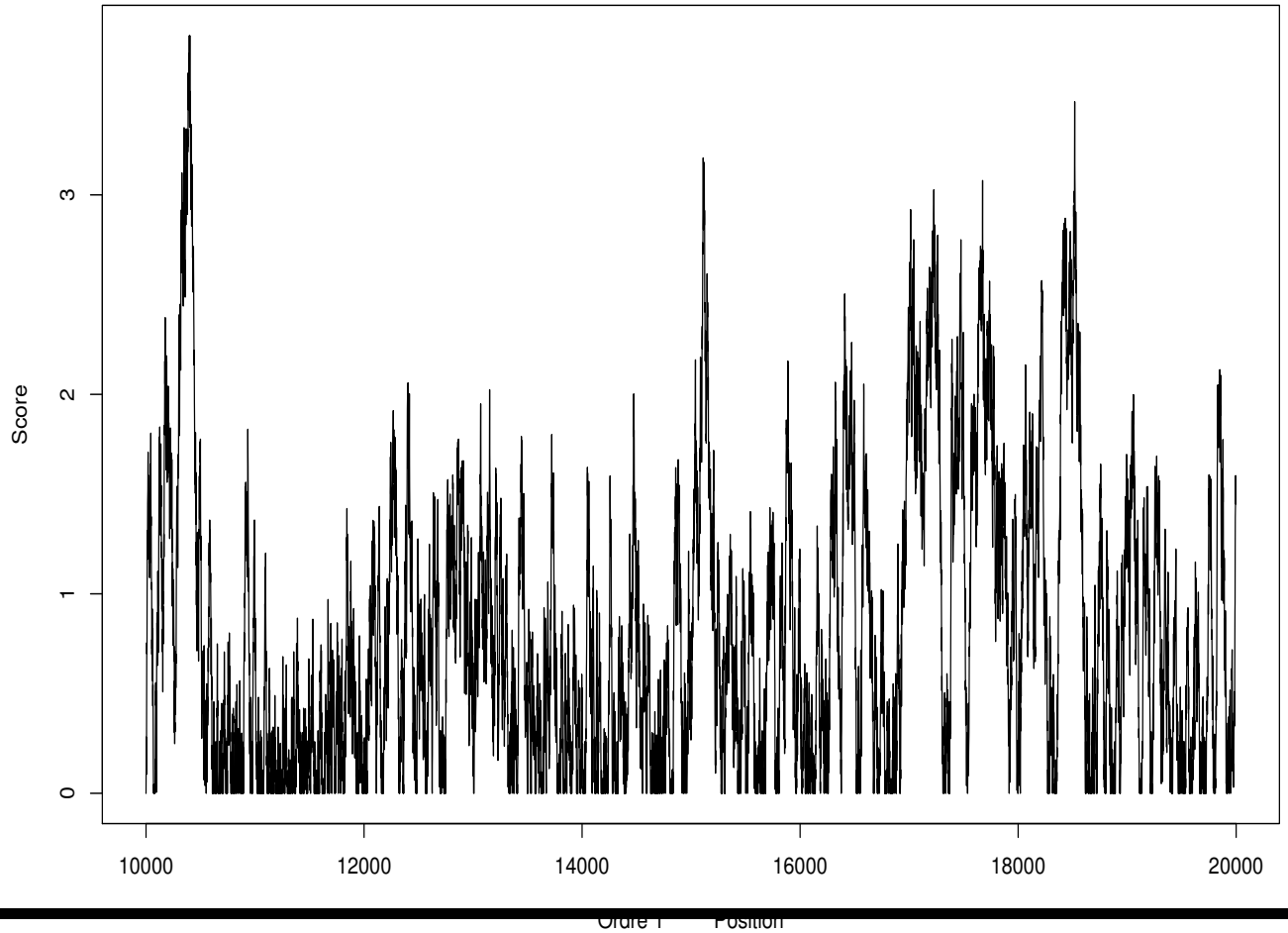
Position	Value	p-value
5915 - 6679	2.24	0.10
28134 - 28728	2.53	0.08

Le segment 28134 - 28728 correspond à la première moitié du gène codant la protéine nucléocapside

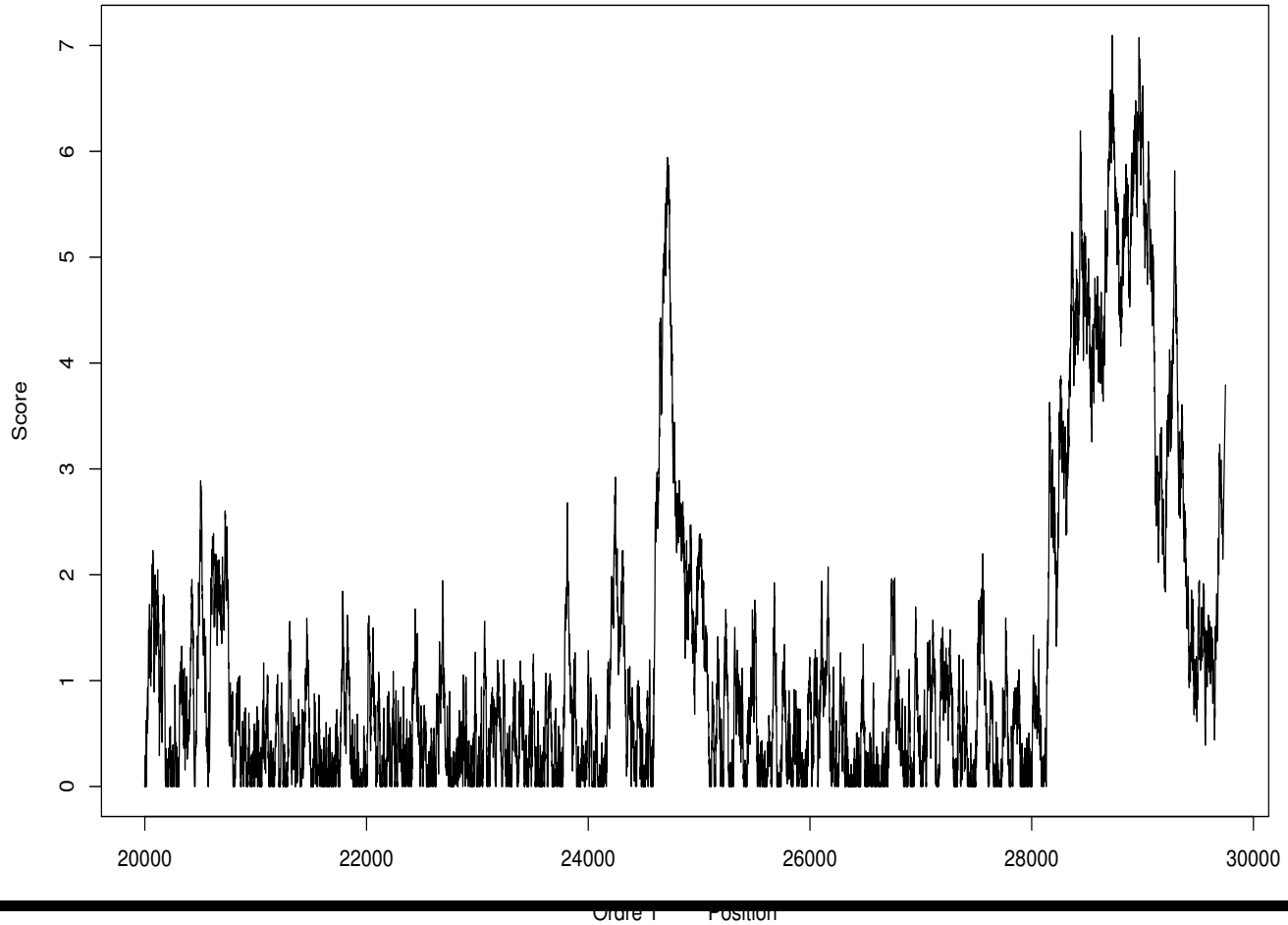
Syndrome Respiratoire Aigu Sévère (2)



Syndrome Respiratoire Aigu Sévère (3)

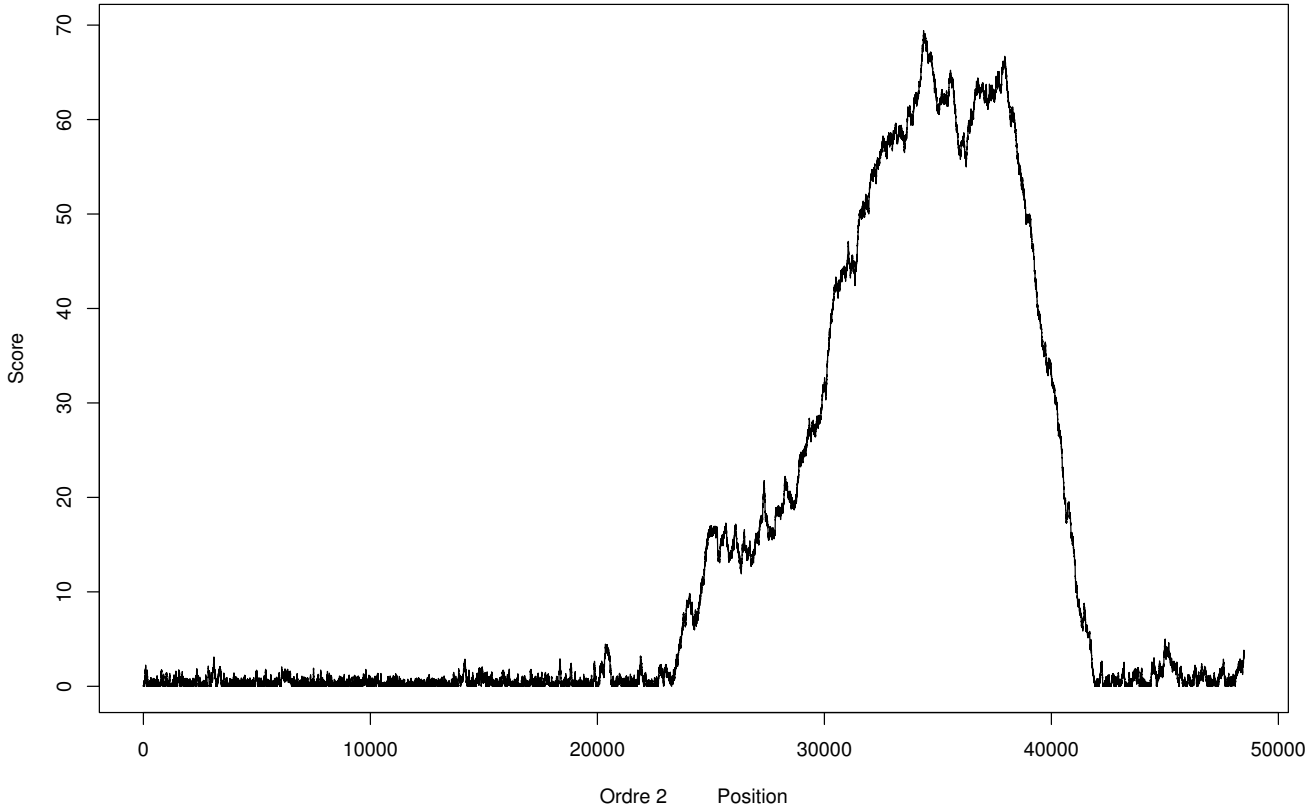


Syndrome Respiratoire Aigu Sévère (3)



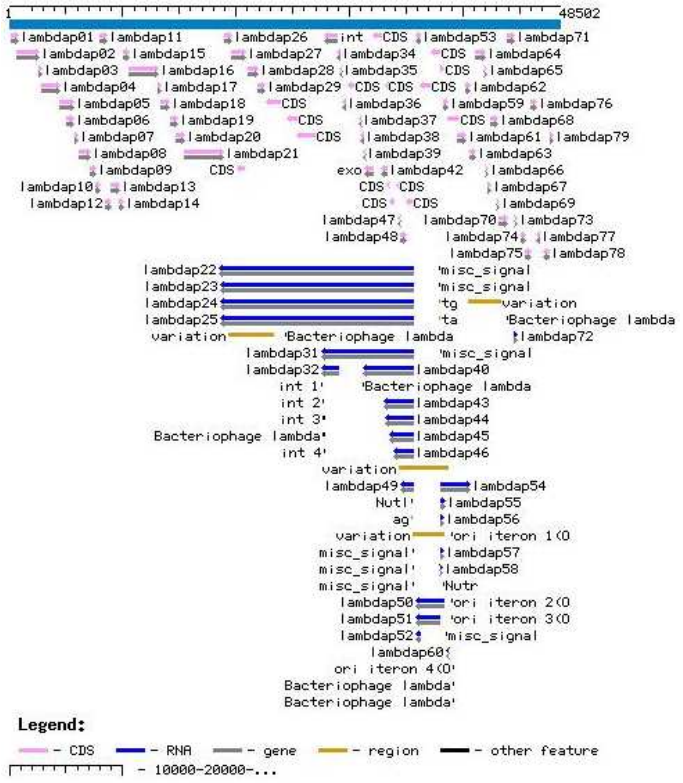
Bactériophage Lambda

Séquence complète - score local - ordre 2



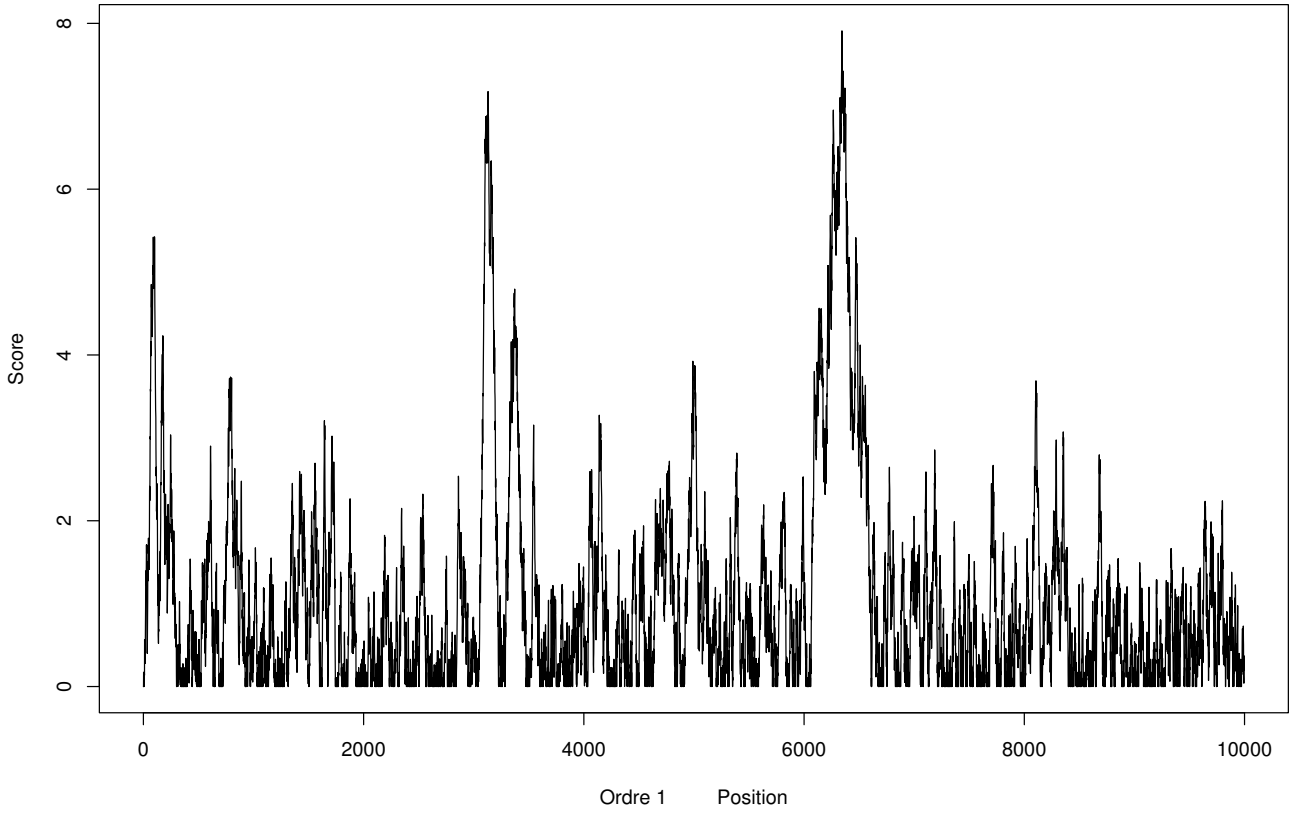
Bactériophage Lambda (2)

Large retournement entre les positions 20000 et 40000

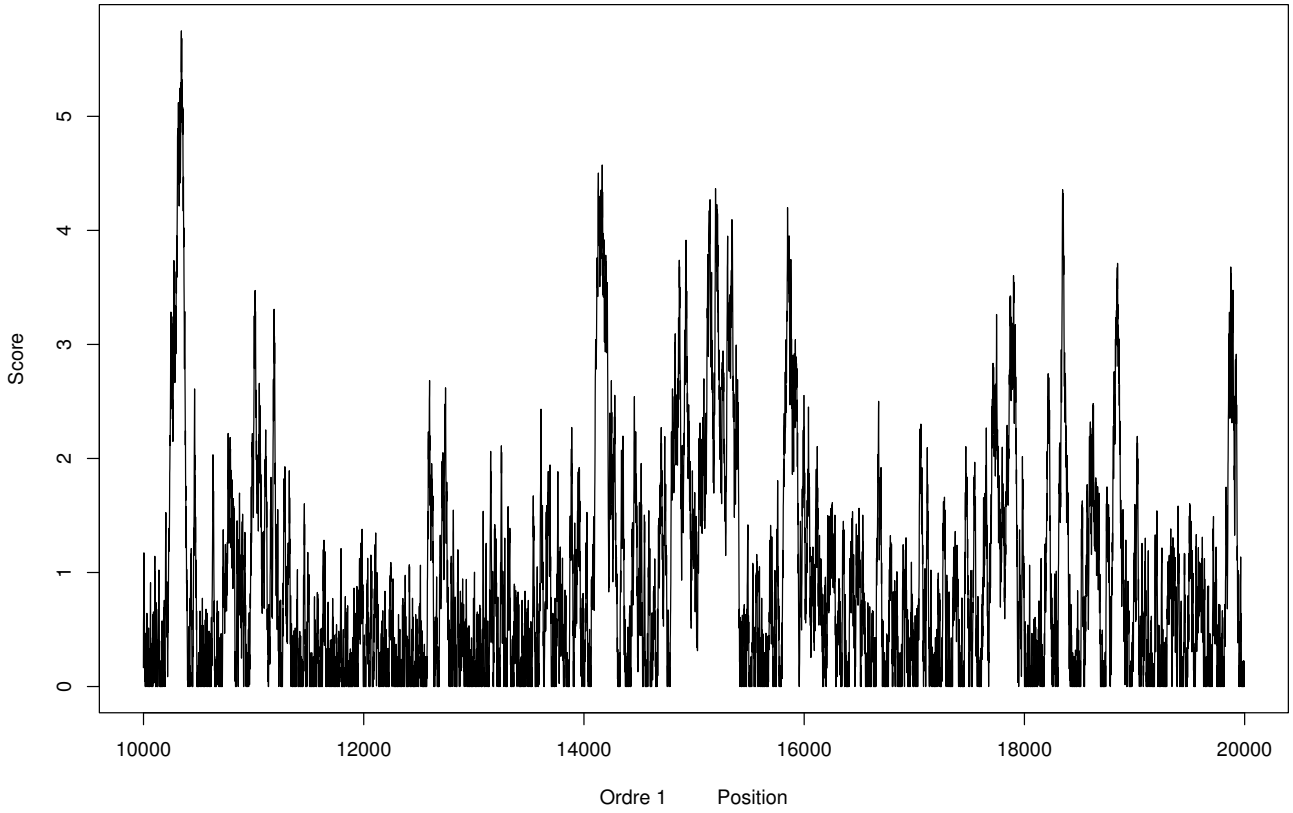


→ liés à la présence de gènes codant les ARNs

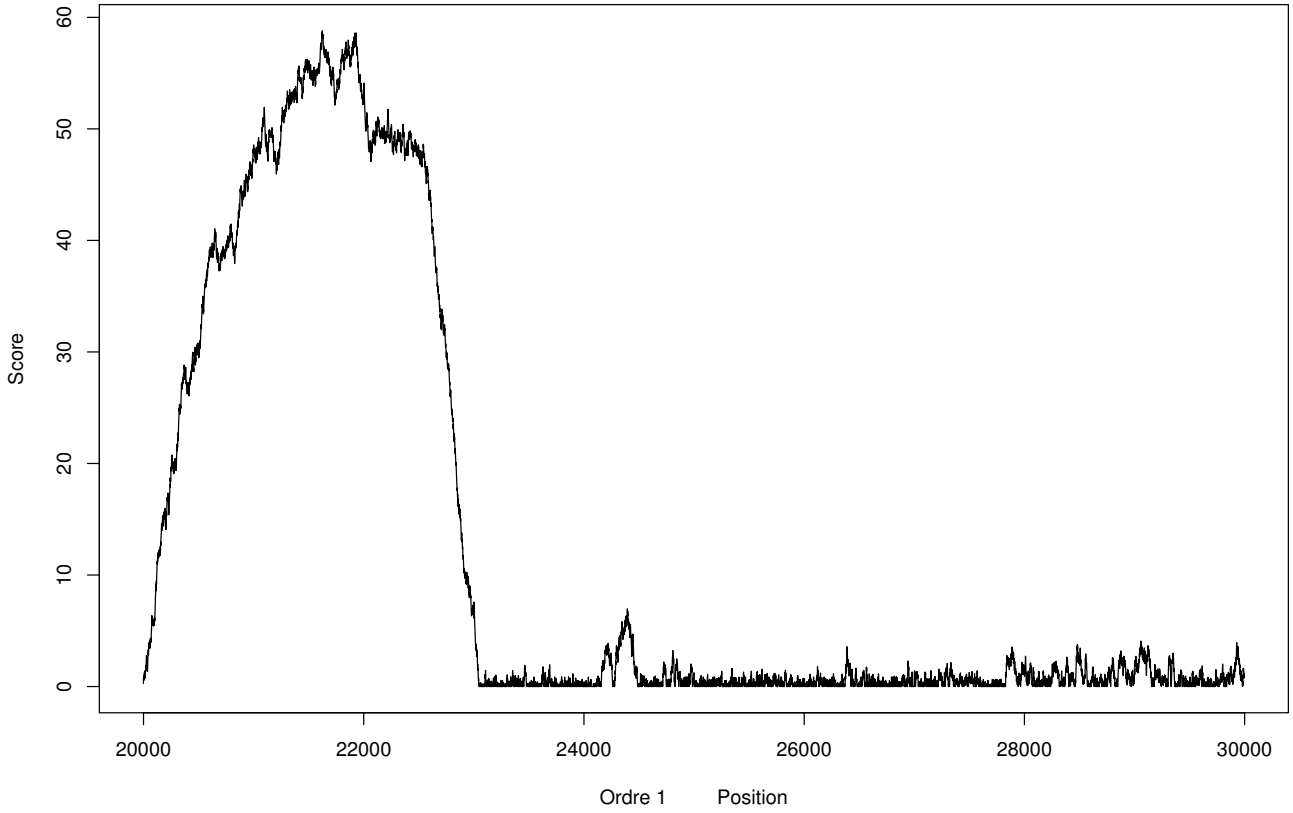
Bactériophage Lambda(3)



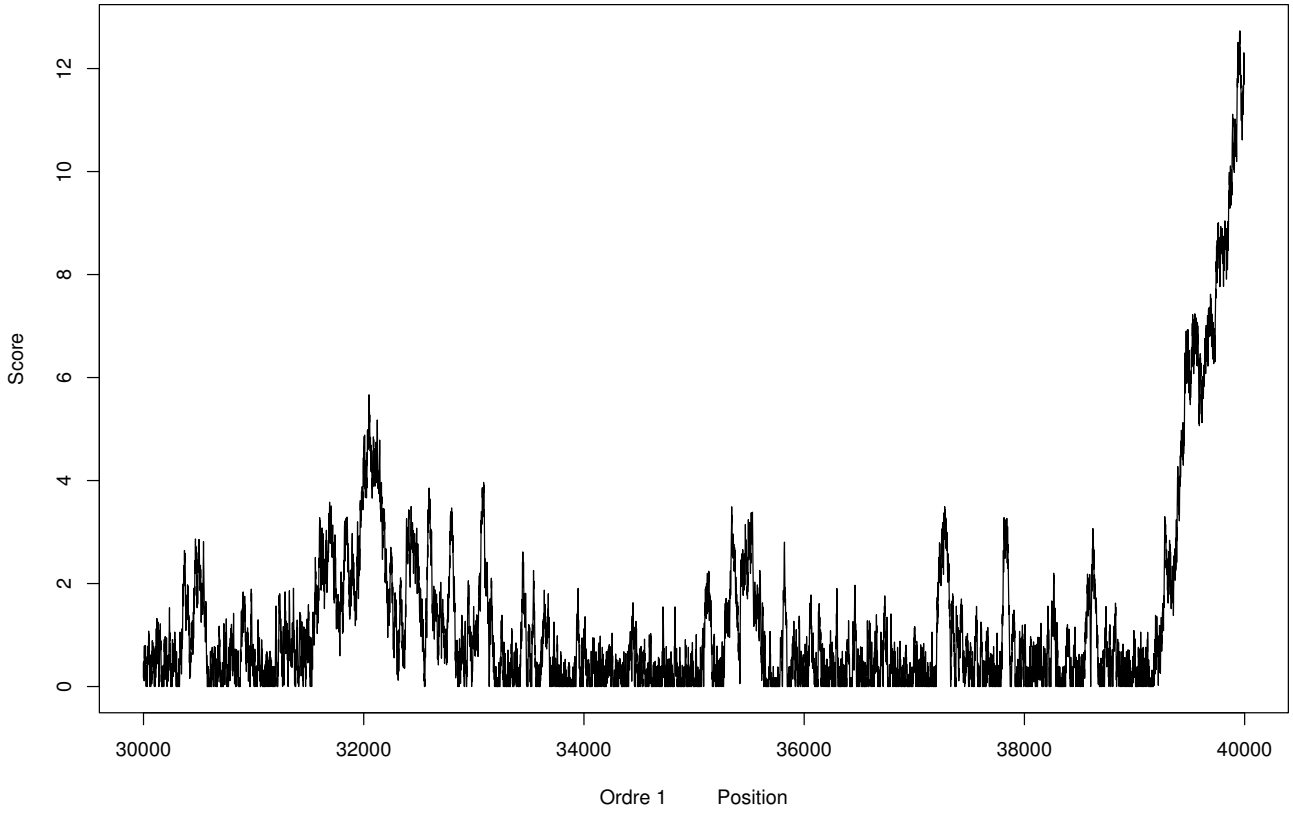
Bactériophage Lambda(4)



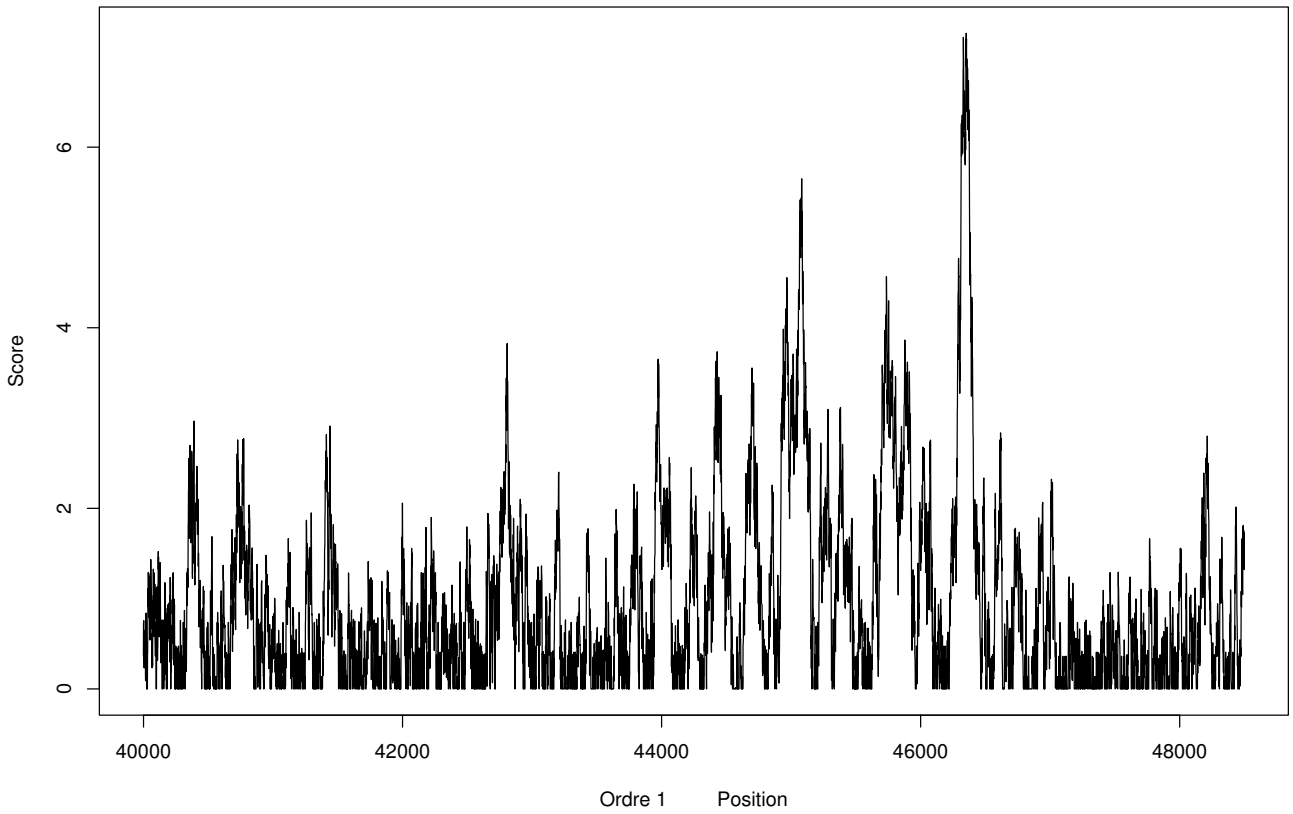
Bactériophage Lambda(5)



Bactériophage Lambda(6)



Bactériophage Lambda(7)



Bactériophage Lambda(7)

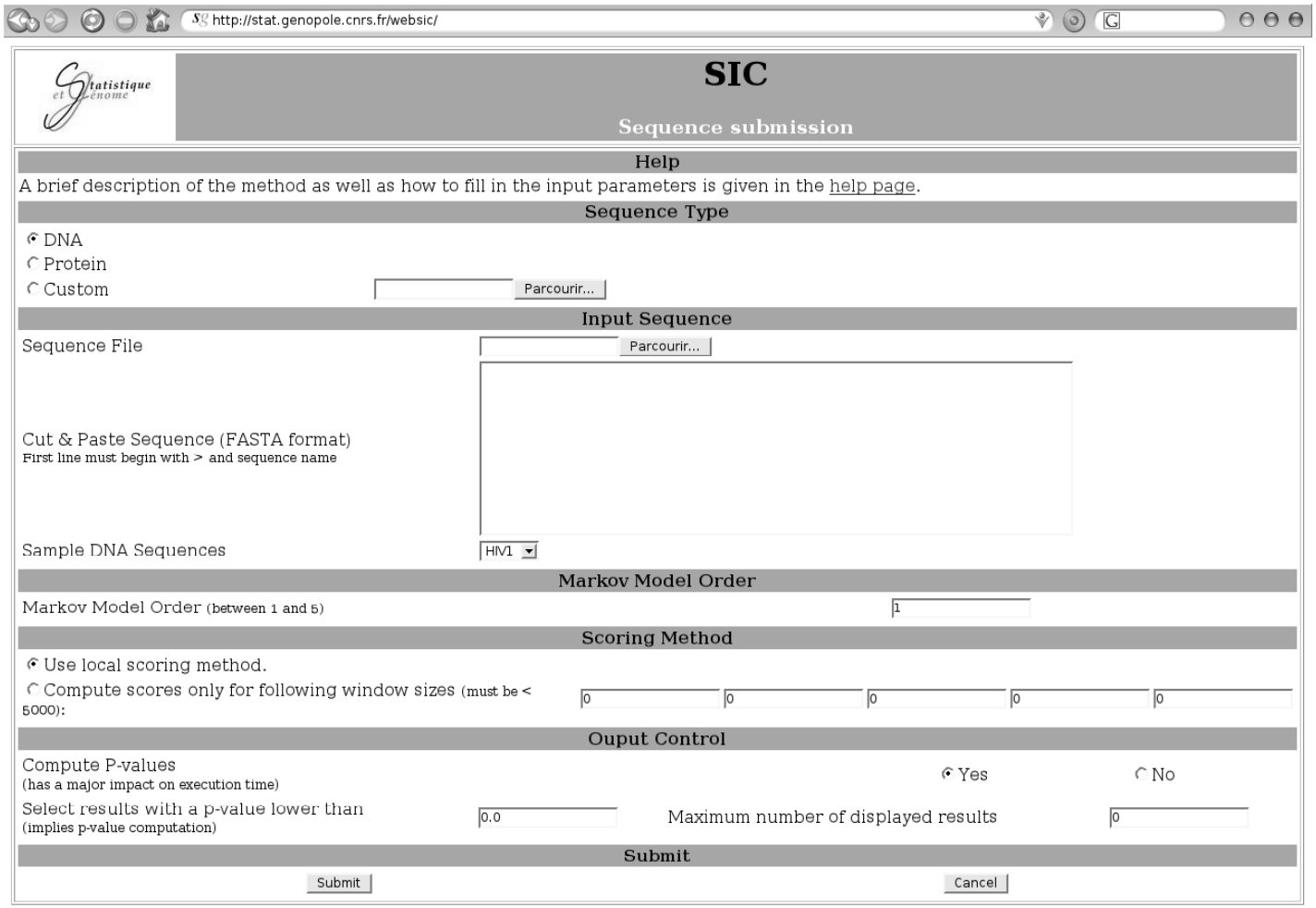
Retournements détectés

Position	Value	p-value	
3042 - 3130	2.10	0.11	capside (2836-4437)
6061 - 6344	2.87	0.06	capside (6135-7160)
20000 - 21625	55.67	$< 10^{-8}$	
39173 - 39958	7.65	5×10^{-4}	OR
46223 - 46353	1.95	0.13	Lyse de membrane de cellule hôte

Logiciel SIC (Scan Inverse Complementary)

Logiciel de recherche de segments inversés dans une séquence biologique

- licence GPL
- s'appuie sur la librairie Seq++ (Miele et al., 2005)
  $\rightarrow$ Modélisations Markoviennes de séquence biologique
- Interface Web (serveur d'application Tomcat)


http://stat.genopole.cnrs.fr/websic/
Statistique et Génome
SIC
Sequence submission
Help
A brief description of the method as well as how to fill in the input parameters is given in the help page.
Sequence Type
DNA
Protein
Custom
Parcourir...
Input Sequence
Sequence File
Parcourir...
Cut & Paste Sequence (FASTA format)
First line must begin with > and sequence name
Sample DNA Sequences
HIV1
Markov Model Order
Markov Model Order (between 1 and 5)
1
Scoring Method
Use local scoring method.
Compute scores only for following window sizes (must be < 5000):
0
0
0
0
0
Ouput Control
Compute P-values
(has a major impact on execution time)
Yes
No
Select results with a p-value lower than
(implies p-value computation)
0.0
Maximum number of displayed results
0
Submit
Submit
Cancel

http://stat.genopole.cnrs.fr/websic/results.do?jobId=2900563582903536638

SIC

Sequence Submission Results

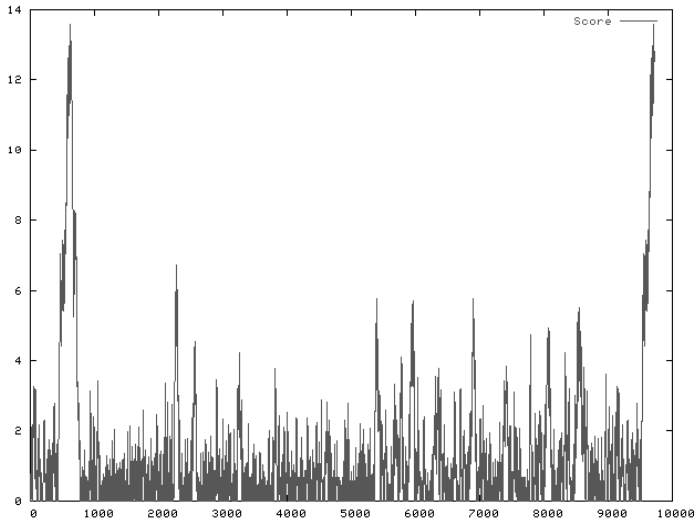
Job Details

- Sequence: HIV1.
- Alphabet: DNA.
- Order of Markov model: 1.
- Local score computation method used.
- P-Value calculation performed.

Total Variation Distance

Total Variation Distance	0.448601

Score Plot



Top Scoring Segments

Rank	Location	Score	P-value	Sequence
1	433 - 628	13.5956	5.87259E-4	agctgctttttgcctgtactgggtctctctggttagaccagatctgagcctgggagctc ctggctaactagggaacccactgcttaagcctcaataaagcttgccttgagtgcttcaa tagtgtgtgcccgtctgttgtgtgactctggtaactagagatccctcagacccttttag cagtgtggaaaatctc
2	9517 - 9712	13.5956	5.87259E-4	agctgctttttgcctgtactgggtctctctggttagaccagatctgagcctgggagctc ctggctaactagggaacccactgcttaagcctcaataaagcttgccttgagtgcttcaa tagtgtgtgcccgtctgttgtgtgactctggtaactagagatccctcagacccttttag cagtgtggaaaatctc

Conclusion

Problèmatique méthodologique :

- Loi jointe asymptotique des r plus grandes valeurs de score local - Procédure de détermination de r
- Résultats sur les CM à espace d'état discret fini.
- Faisabilité pratique des méthodes.

Problèmatique biologique :

- Méthode de détection ab initio
- Peu puissante pour détecter des segments très courts
- Fortement dépendante du "degré d'orientation" du génome
- Perspectives : génomique comparative