Détection de ruptures en 2D – Cadre non paramétrique – Données Hi-C

Sarah Ouadah

en collaboration avec

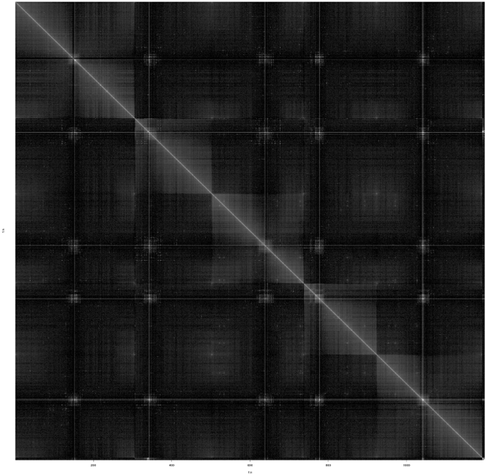
Vincent Brault, Laure Sansonnet et Céline Lévy-Leduc

UMR MIA-Paris, AgroParisTech, INRA

Séminaire MathForGenomics

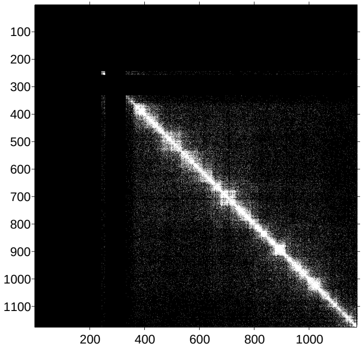
11 avril 2018, Evry

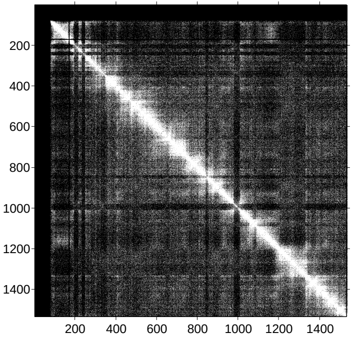
Données Hi-C – Arabidopsis Thaliana



5 chromosomes (UMR 8618 IBP - Institut de Biologie des Plantes, Université Paris Sud, CNRS)

Données Hi-C – Bing Ren Lab





Chromosome 21 (hESC) et Chromosome 19 (cortex de souris)

Questions

Modèle – Test – Estimation – Consistance – Application

- Quel modèle est adapté à la structure des données Hi-C ?
- Y-a-t'il une décomposition en blocs de la matrice des données Hi-C ?
- Si oui où se trouvent les frontières de ces blocs, i.e les ruptures ?
- Quelles sont les performances statistiques de la procédure d'estimation des frontières ?
- Comment obtenir des résultats en pratique ?

Sommaire

Quel modèle est adapté à la structure des données Hi-C ?

Y-a-t'il une décomposition en blocs de la matrice des données Hi-C?

Où se trouvent les frontières des blocs, i.e les ruptures ?

Quelles sont les performances statistiques de la procédure d'estimation des frontières ?

Comment obtenir des résultats en pratique ?

Structure des données (1)

Données : $\mathbf{X} = (X_{i,j})_{1 \leq i,j \leq n}$ matrice d'interaction

- $X_{i,j}$: intensité d'interaction entre les loci i et j (nombre total de paires de read communes)

Caractéristiques de $\mathbf{X}$

- $\mathbf{X}$ est symétrique
- $\mathbf{X}$ possède L potentielles ruptures $n_1, \ldots, n_L \in \{1, \ldots, n\}$, i.e $(L+1)^2$ blocs

Caractéristiques des $X_{i,j}$s

- les $X_{i,j}$s sont des variables aléatoires indépendantes pour $i \geq j$
- la distribution des $X_{i,j}$s est continue

Structure des données (2)

Soient

- $\mathbf{X}^{(j)} = (X_{1,j}, \ldots, X_{n,j})'$ la j-ème colonne de $\mathbf{X}$ *i.e, le vecteur des intensités d'interaction du locus j avec les autres loci*
- $\mathbf{X} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)})$

Particularité de $\mathbf{X}$

- n le nombre de vecteurs et K le nombre d'observations (taille du vecteur) sont <u>égaux et grands</u>, $\mathbf{X}$ <u>symétrique</u> avec de potentiels <u>blocs non chevauchants</u>

Bibliographie détection de ruptures en 2D (matrice)

- Cas où $K \neq n$ potentiellement grands, paramétrique/séries temporelles : Bai, 2010; Horvath and Huskova, 2012; Jirak, 2015
- Cas où n grand et K fixe, non paramétrique : Matteson and James, 2014; Lung-Yut-Fong et al., 2015

Sommaire

Quel modèle est adapté à la structure des données Hi-C ?

Y-a-t'il une décomposition en blocs de la matrice des données Hi-C?

Où se trouvent les frontières des blocs, i.e les ruptures ?

Quelles sont les performances statistiques de la procédure d'estimation des frontières ?

Comment obtenir des résultats en pratique ?

Y-a-t'il une décomposition en blocs de la matrice des données Hi-C ? Cas d'une seule rupture $0 < n_1 < n$

Test

$\mathcal{H}_0$: "les intensités d'interaction entre loci ont toutes la même distribution"

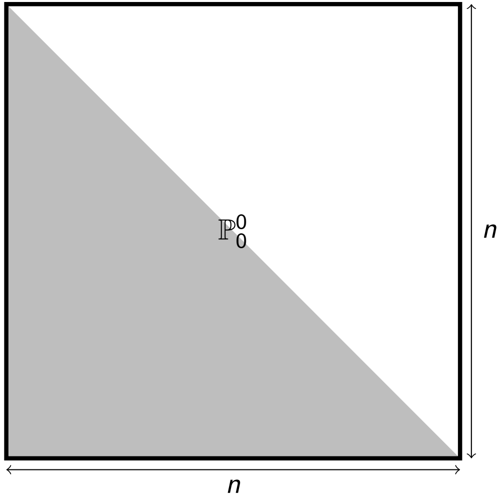
$\mathcal{H}_1$: la distribution des intensités d'interaction entre loci change d'une région à l'autre (avant et après la rupture)"

$\mathcal{H}_0$: "$(\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(n_1)}) \sim \mathbb{P}$ et $(\mathbf{X}^{(n_1+1)}, \dots, \mathbf{X}^{(n)}) \sim \mathbb{P}$"

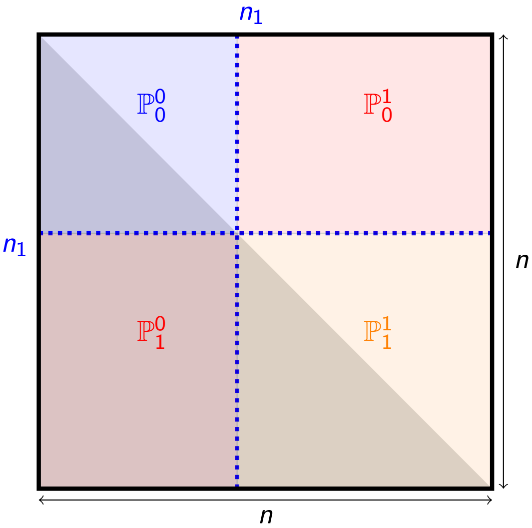
$\mathcal{H}_1$: "$(\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(n_1)}) \sim \mathbb{P}_1$ et $(\mathbf{X}^{(n_1+1)}, \dots, \mathbf{X}^{(n)}) \sim \mathbb{P}_2$ avec $\mathbb{P}_1 \neq \mathbb{P}_2$"

où $\mathbf{X}^{(j)} = (X_{1,j}, \dots, X_{n,j})'$ la j-ème colonne de $\mathbf{X}$

Hypothèse nulle $\mathcal{H}_0$



Hypothèse alternative $\mathcal{H}_1$



$\mathcal{H}_0$: "$\mathbb{P}_0^0 = \mathbb{P}_0^1 = \mathbb{P}_1^1$" et $\mathcal{H}_1$: "$\mathbb{P}_0^0 \neq \mathbb{P}_1^0$ ou "$\mathbb{P}_1^0 \neq \mathbb{P}_1^1$"

Statistique de test (1)

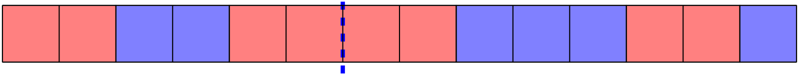
Extension de la statistique de *Lung-Yut-Fong et al., 2015*

$$S_n(n_1) = \sum_{i=1}^{n} U_{n,i}^2(n_1)$$

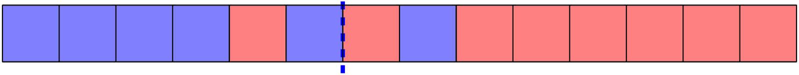
où

- $U_{n,i}(n_1) = \frac{1}{\sqrt{nn_1(n-n_1)}} \sum_{j_0=1}^{n_1} \sum_{j_1=n_1+1}^{n} h(X_{i,j_0}, X_{i,j_1})$
- $h(x,y) = \mathbb{1}_{\{x \leq y\}} - \mathbb{1}_{\{y \leq x\}}$

Sous $\mathcal{H}_0$, $U_{n,i}(n_1) \approx 0$



Sous $\mathcal{H}_1$, $|U_{n,i}(n_1)| >> 0$



Statistique de test (2)

Extension de la statistique de rang de Wilcoxon

$$U_{n,i}(n_1) = \frac{2}{\sqrt{nn_1(n-n_1)}} \sum_{j_0=1}^{n_1} \left(\frac{n+1}{2} - R_{j_0}^{(i)} \right)$$
$$= \frac{2}{\sqrt{nn_1(n-n_1)}} \sum_{j_1=n_1+1}^{n} \left(R_{j_1}^{(i)} - \frac{n+1}{2} \right),$$

où

$$R_j^{(i)} = \sum_{k=1}^{n} \mathbb{1}_{\{X_{i,k} \le X_{i,j}\}}$$

est le rang de $X_{i,j}$ au sein de $(X_{i,1}, \ldots, X_{i,n})$.

Y-a-t'il une décomposition en blocs de la matrice des données Hi-C ? Cas de L ruptures $0 < n_1 < \ldots < n_L < n$

Test

$\mathcal{H}_0$: "les intensités d'interaction entre loci ont tous la même distribution"

$\mathcal{H}_1$: "la distribution des intensités d'interaction entre loci change selon les régions"

$\mathcal{H}_0$: "$(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n_1)}) \sim \mathbb{P}$, $(\mathbf{X}^{(n_1+1)}, \ldots, \mathbf{X}^{(n_2)}) \sim \mathbb{P}$, …, $(\mathbf{X}^{(n_L+1)}, \ldots, \mathbf{X}^{(n)}) \sim \mathbb{P}$"

$\mathcal{H}_1$: "il existe un $\ell \in \{1, \ldots, L\}$ tel que $(\mathbf{X}^{(n_{\ell-1}+1)}, \ldots, \mathbf{X}^{(n_\ell)}) \sim \mathbb{P}_\ell$ et $(\mathbf{X}^{(n_\ell+1)}, \ldots, \mathbf{X}^{(n_{\ell+1})}) \sim \mathbb{P}_{\ell+1}$ avec $\mathbb{P}_\ell \neq \mathbb{P}_{\ell+1}$"

Statistique de test

Extension de la statistique de *Lung-Yut-Fong et al., 2015* / Kruskal-Wallis

$$S_n(n_1, \ldots, n_L) = \frac{4}{n^2} \sum_{\ell=0}^{L} (n_{\ell+1} - n_\ell) \sum_{i=1}^{n} \left(\overline{R}_\ell^{(i)} - \frac{n+1}{2} \right)^2$$

avec

$$\overline{R}_\ell^{(i)} = \frac{1}{n_{\ell+1} - n_\ell} \sum_{j=n_\ell+1}^{n_{\ell+1}} R_j^{(i)}$$

où $\overline{R}_\ell^{(i)}$ est la moyenne des rangs $R_j^{(i)}$ du groupe ℓ.

Test d'homogénéité

Théorème 1

Soit $\mathbf{X} = (X_{i,j})_{1 \leq i,j \leq n}$ une matrice symétrique telle que les variables aléatoires $X_{i,j}$s sont i.i.d. pour $i \geq j$ et de fonction de répartition continue. On suppose qu'il existe $0 < \tau_1 < \tau_2 < \ldots < \tau_L < 1$ tels que pour tout $\ell \in \{1, \ldots, L\}$, $n_\ell / n \to \tau_\ell$ quand $n \to \infty$.

Alors, lorsque $n \to \infty$

$$T_n := n^{-1/2} \big(S_n (n_1, \ldots, n_L) - \mathbb{E} [S_n (n_1, \ldots, n_L)] \big) = O_P(1)$$

et

$$\mathbb{E} [S_n (n_1, \ldots, n_L)] = \frac{L(n+1)}{3}$$

La statistique de test prend des valeurs finies.

On rejette $\mathcal{H}_0$ lorsque T_n dépasse un certain seuil

Calibration de la région de rejet

10 000 matrices $\mathbf{X}$ symétriques $n \times n$

- $n \in \{50, 100, 500, 1000\}$
- distributions des $(X_{i,j})_{i \geq j}$:
  - $\mathcal{N}(0, 1)$
  - $\mathcal{C}au(0, 1)$
  - $\mathcal{E}xp(2)$
- $n_1 = \lfloor 0.1n \rfloor$ or $n_1 = \lfloor 0.5n \rfloor$

Quantiles empiriques de $T_n(n_1)$ d'ordre 0.95

	$n_1 = \lfloor 0.1n \rfloor$			$n_1 = \lfloor 0.5n \rfloor$		
	$\mathcal{N}(0,1)$	$\mathcal{C}au(0,1)$	$\mathcal{E}xp(2)$	$\mathcal{N}(0,1)$	$\mathcal{C}au(0,1)$	$\mathcal{E}xp(2)$
$n = 50$	0.83	0.83	0.82	0.78	0.79	0.76
$n = 100$	0.81	0.8	0.82	0.78	0.8	0.78
$n = 500$	0.78	0.8	0.81	0.8	0.78	0.77
$n = 1000$	0.79	0.78	0.79	0.78	0.77	0.79

Puissance du test (1)

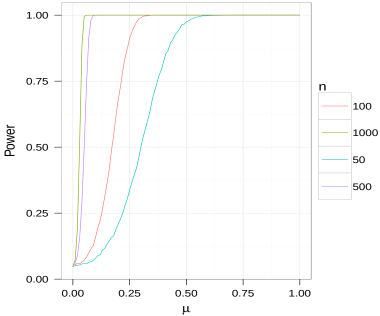
10 000 matrices $\mathbf{X}$ symétriques $n \times n$

- $n \in \{50, 100, 500, 1000\}$
- Configuration :

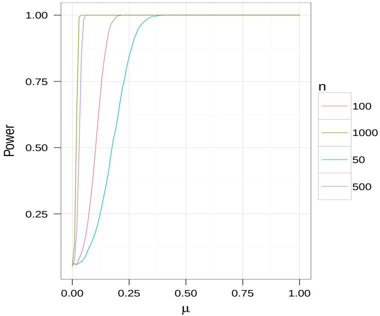
$$\left(\begin{array}{c|c} \mathcal{N}(0,1) & \mathcal{N}(\mu,1) \\ \hline \mathcal{N}(\mu,1) & \mathcal{N}(0,1) \end{array}\right) \text{ avec } \mu \in [0,1]$$

- $n_1 = \lfloor 0.1n \rfloor$ ou $n_1 = \lfloor 0.5n \rfloor$
- Rejet si $T_n(n_1) > 0.78$

Puissance du test (2)



$$n_1 = \lfloor 0.1n \rfloor$$



$$n_1 = \lfloor 0.5n \rfloor$$

Sommaire

Estimation des frontières des blocs

Estimation de $(n_1^*, n_2^*, \cdots, n_L^*)$

$$(\widehat{n}_1, \ldots, \widehat{n}_L) \in \underset{0<n_1<\cdots<n_L<n}{\arg\max}\; S_n(n_1, \ldots, n_L)$$

où

$$S_n(n_1, \ldots, n_L) = \frac{4}{n^2} \sum_{\ell=0}^{L} (n_{\ell+1} - n_\ell) \sum_{i=1}^{n} \left(\overline{R}_\ell^{(i)} - \frac{n+1}{2} \right)^2$$

est la statistique de test du test d'homogénéité.

Maximisation

- Stratégie : programmation dynamique (Bellman, 1961; Kay, 1993)
- Complexité de l'algorithme : $\mathcal{O}(n^3)$

Consistance (1)

Soit $\mathbf{X} = (X_{i,j})_{1 \leq i,j \leq n}$ une matrice symétrique telle que les variables aléatoires $X_{i,j}$s sont i.i.d. pour $i \geq j$ et de fonction de répartition continue. Soient $\mathbb{P}_{\ell_1}^{\ell_2}$ la distribution de $X_{i,j}$ pour $i \in D_{\ell_2}^*$, $j \in D_{\ell_1}^*$ et F_{ℓ_2,ℓ_1} la fonction de répartition associée, où
$D_\ell^* = \left\{i \in \{1, \ldots, n\} : n_\ell^* + 1 \leq i \leq n_{\ell+1}^*\right\}$.

Hypothèses

(A1) pour tout ℓ in $\{1, \ldots, L\}$, il existe $\tau_\ell^* \in (0, 1)$ tel que $n_\ell^*/n \to \tau_\ell^*$, lorsque $n \to \infty$ et tel que $\Delta_\tau^\star = \min_{0 \leq \ell \leq L} |\tau_{\ell+1}^* - \tau_\ell^*| > 0$

(A2) pour tout $\ell_1 \in \{0, \ldots, L-1\}$, il existe $\ell_4 \in \{0, \ldots, L\}$ tel que

$$\sum_{\ell_3=0}^{L} \left(\tau_{\ell_3+1}^\star - \tau_{\ell_3}^\star\right) \mathbb{E}\left[F_{\ell_4,\ell_1}(X) - F_{\ell_4,\ell_1+1}(X)\right] \neq 0, \quad X \sim \mathbb{P}_{\ell_3}^{\ell_4}$$

Consistance (2)

Théorème 2

Soit $\mathbf{X} = (X_{i,j})_{1 \leq i,j \leq n}$ une matrice symétrique telle que les variables aléatoires $X_{i,j}$s sont i.i.d. pour $i \geq j$ et de fonction de répartition continue.
Sous les hypothèses $(A1)$ et $(A2)$, pour tout $\delta > 0$, on a:

$$\mathbb{P}\big(\|\widehat{\mathbf{n}} - \mathbf{n}^*\|_\infty \geq n\delta\big) \to 0, \text{ lorsque } n \to \infty$$

où $\|\widehat{\mathbf{n}} - \mathbf{n}^*\|_\infty = \max_{0 \leq l \leq L} |\hat{n}_\ell - {n_\ell}^*|$.

Sommaire

Quel modèle est adapté à la structure des données Hi-C ?

Y-a-t'il une décomposition en blocs de la matrice des données Hi-C?

Où se trouvent les frontières des blocs, i.e les ruptures ?

Quelles sont les performances statistiques de la procédure d'estimation des frontières ?

Comment obtenir des résultats en pratique ?

Données simulées

10 000 matrices $\mathbf{X}$ symétriques $n \times n$

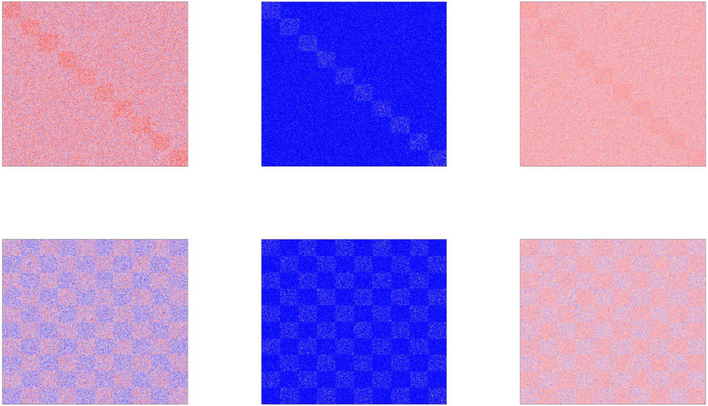
- $n \in \{50, 100, 200, 300, 400\}$
- $L = 10$
- Configurations :

Blocs diagonaux	Échiquier
$\begin{pmatrix} \mathcal{L}_1 & \mathcal{L}_2 & \cdots & \cdots & \mathcal{L}_2 \\ \mathcal{L}_2 & \mathcal{L}_1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \mathcal{L}_1 & \mathcal{L}_2 \\ \mathcal{L}_2 & \cdots & \cdots & \mathcal{L}_2 & \mathcal{L}_1 \end{pmatrix}$	$\begin{pmatrix} \mathcal{L}_1 & \mathcal{L}_2 & \mathcal{L}_1 & \cdots & \mathcal{L}_1 & \mathcal{L}_2 \\ \mathcal{L}_2 & \mathcal{L}_1 & \mathcal{L}_2 & \cdots & \mathcal{L}_2 & \mathcal{L}_1 \\ \mathcal{L}_1 & \mathcal{L}_2 & \mathcal{L}_1 & \cdots & \mathcal{L}_1 & \mathcal{L}_2 \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ \mathcal{L}_2 & \mathcal{L}_1 & \mathcal{L}_2 & \cdots & \mathcal{L}_2 & \mathcal{L}_1 \end{pmatrix}$

- Distributions:

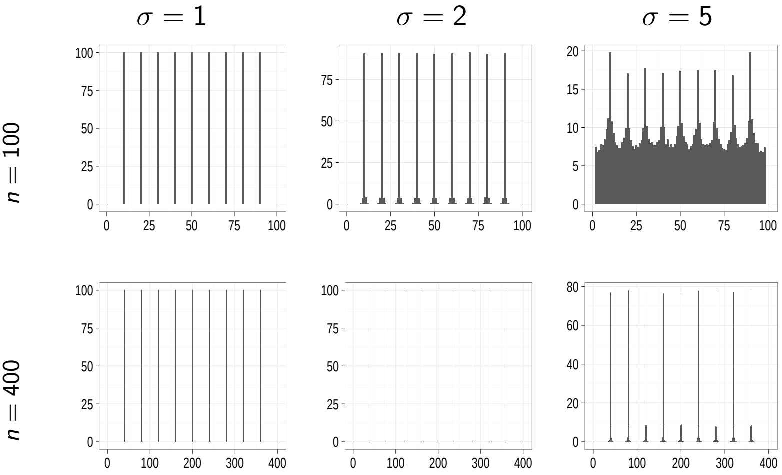
	$\mathcal{L}_1$	$\mathcal{L}_2$	Difficulté
Cas 1	$\mathcal{N}(1, \sigma^2)$	$\mathcal{N}(0, \sigma^2)$	$\sigma \in \{1, 2, 5\}$
Cas 2	$\mathcal{E}xp(2)$	$\mathcal{E}xp(\lambda)$	$\lambda \in \{1, 0.5, 4\}$
Cas 3	$\mathcal{C}au(1, a)$	$\mathcal{C}au(0, a)$	$a \in \{1, 2, 5\}$

Exemples de structures



Exemples de matrices $\mathbf{X}$ 400×400 . Haut : configuration blocs diagonaux, bas : configuration échiquier. Gauche : $\mathcal{L}_1 = \mathcal{N}(1,4)$, $\mathcal{L}_2 = \mathcal{N}(0,4)$, milieu : $\mathcal{L}_1 = \mathcal{E}xp(2)$, $\mathcal{L}_2 = \mathcal{E}xp(1)$, droite : $\mathcal{L}_1 = \mathcal{C}au(1,1)$, $\mathcal{L}_2 = \mathcal{C}au(0,1)$.

Configuration échiquier cas gaussien



Comparaison avec *Matteson and James, 2014* (ecp)

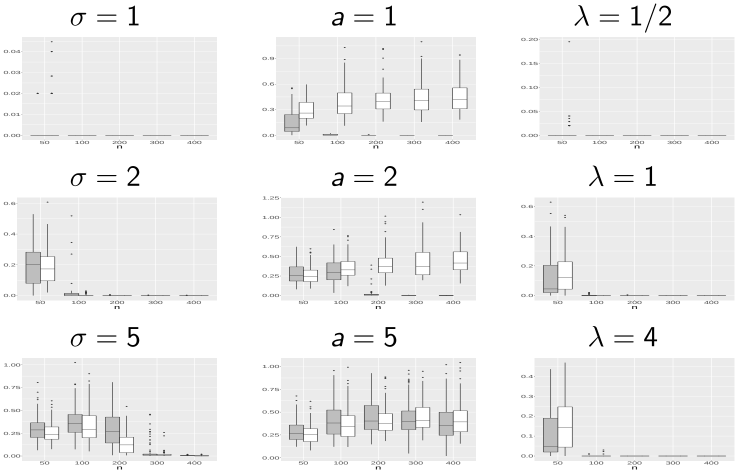
- 100 réplicats de $n \times n$ matrices symétriques
- $n \in \{50, 100, 200, 300, 400\}$
- configurations blocs diagonaux et échiquier

Distance

$$D(\widehat{\mathbf{n}}, \mathbf{n}^\star) = \frac{1}{n}\sqrt{\sum_{\ell=1}^{L^\star}(\widehat{n}_\ell - n_\ell^\star)^2}$$

avec $\mathbf{n}^\star = (n_1^\star, \ldots, n_{L^\star}^\star)$ le vecteur des vraies $L^\star$ ruptures et $\widehat{\mathbf{n}} = (\widehat{n}_1, \ldots, \widehat{n}_{L^\star})$ son estimation.

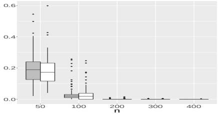
Échiquier



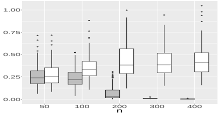
Gris: MuChPoint Blanc: ecp

Blocs diagonaux

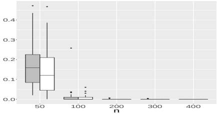
$\sigma = 1$



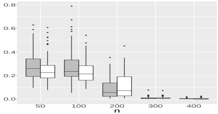
$a = 1$



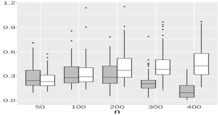
$\lambda = 1/2$



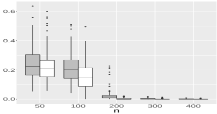
$\sigma = 2$



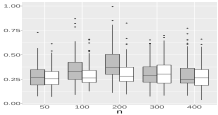
$a = 2$



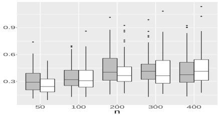
$\lambda = 1$



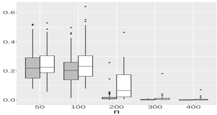
$\sigma = 5$



$a = 5$



$\lambda = 4$



Gris: MuChPoint Blanc: ecp

Sommaire

Quel modèle est adapté à la structure des données Hi-C ?

Y-a-t'il une décomposition en blocs de la matrice des données Hi-C?

Où se trouvent les frontières des blocs, i.e les ruptures ?

Quelles sont les performances statistiques de la procédure d'estimation des frontières ?

Comment obtenir des résultats en pratique ?

Données Hi-C - cortex de souris – Package R MuchPoint

Données Hi-C – Cortex de souris

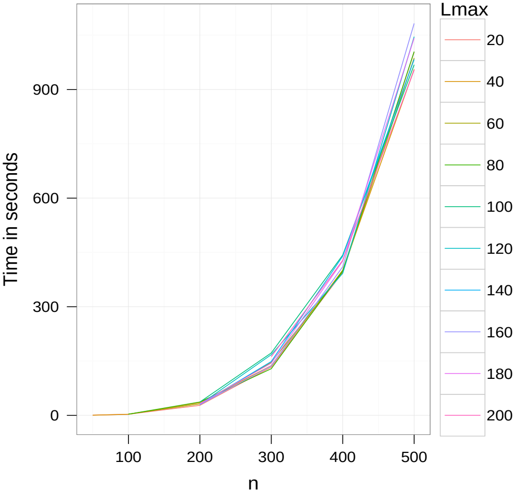
- données publiques :
  `http://chromosome.sdsc.edu/mouse/hi-c/download.html`
  (Dixon et al., 2012)
- Matrice des intensités d'interaction pour le Chromosome 19 du cortex de la souris avec une résolution 40 kb
- $n = 1534$

Package R MuchPoint

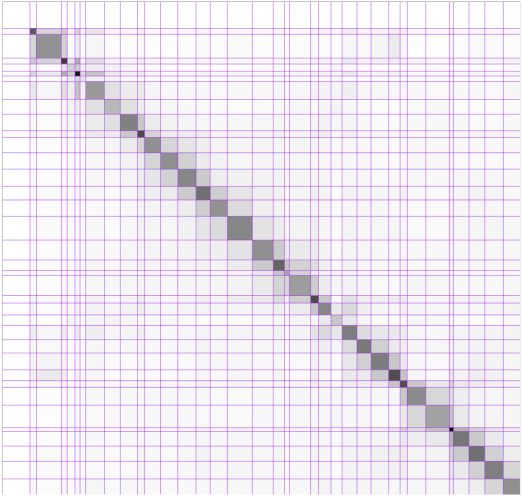
MuChPoint($\mathbf{X}$,Lmax=nrow($\mathbf{X}$/2),N=NULL,selection=T,cores=4,verbose=T)

avec $\mathbf{X}$ la matrice symétrique des interactions entre loci.

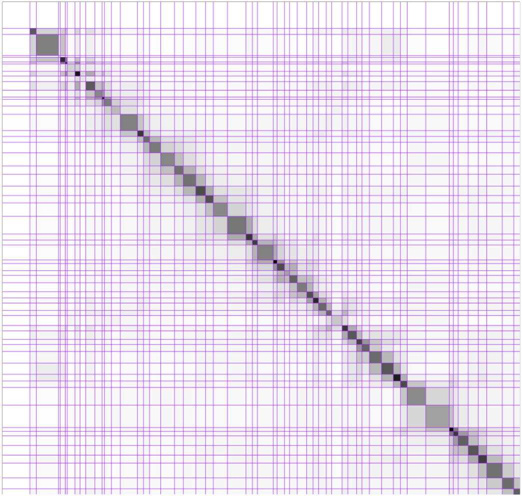
Temps de calcul de l'algorithme de programmation dynamique



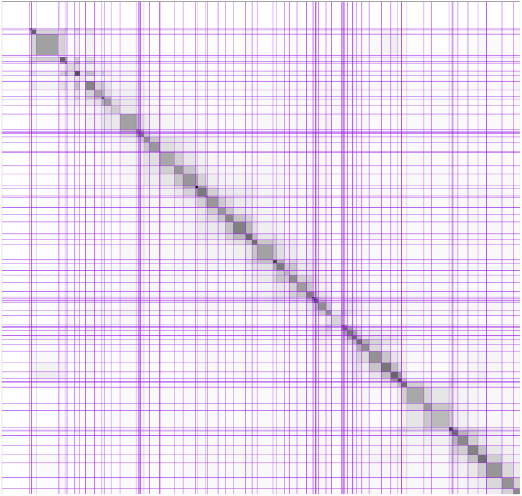
$\mathbf{X}$ estimée pour 35 ruptures



X estimée pour 55 ruptures



X estimée pour 75 ruptures



Comparaison avec *Dixon et al., 2012*

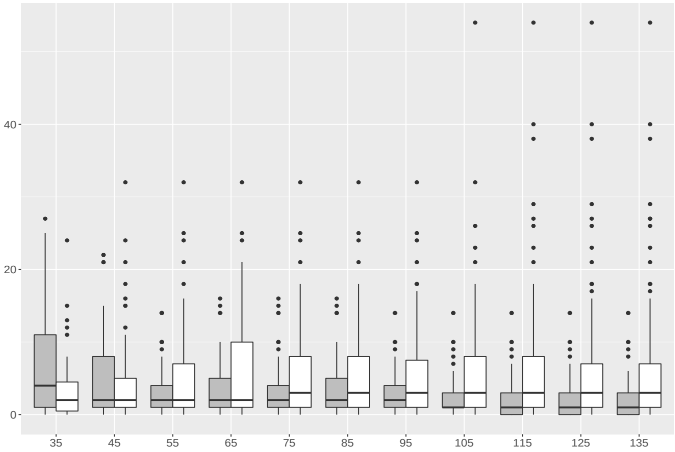
Distance de Hausdorff

$$d\left(\widehat{\boldsymbol{n}}_M, \widehat{\boldsymbol{n}}_D\right) = \max\left(d_1\left(\widehat{\boldsymbol{n}}_M, \widehat{\boldsymbol{n}}_D\right), d_2\left(\widehat{\boldsymbol{n}}_M, \widehat{\boldsymbol{n}}_D\right)\right)$$

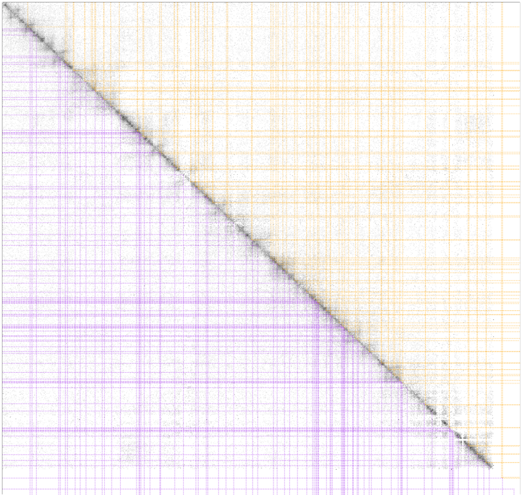
avec $\widehat{\boldsymbol{n}}_M$ et $\widehat{\boldsymbol{n}}_D$ les vecteurs des ruptures estimés avec MuchPoint et celle de *Dixon et al., 2012*, respectivement, et où

$$\begin{aligned} d_1\left(\mathbf{a}, \mathbf{b}\right) &= \sup_{b \in \mathbf{b}} \inf_{a \in \mathbf{a}} |a - b| \\ d_2\left(\mathbf{a}, \mathbf{b}\right) &= d_1\left(\mathbf{b}, \mathbf{a}\right). \end{aligned}$$

Boxplots de d_1 et d_2 selon L



Comparaison avec 85 ruptures



Références

V. Brault, S. Ouadah, L. Sansonnet, C. Lévy-Leduc. Nonparametric homogeneity tests and multiple change-point estimation for analyzing large Hi-C data matrices. Journal of Multivariate Analysis, 2017.

J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485(7398):376-380, 2012.

A. Lung-Yut-Fong, C. Lévy-Leduc, O. Cappé. Homogeneity and change-point detection tests for multivariate data using rank statistics, J. Soc. Fr. Statist. 156:133-162, 2015.

D. S. Matteson and N. A. James. A nonparametric approach for multiple change point analysis of multivariate data. Journal of the American Statistical Association, 109 (505):334-345, 2014.