Test d'homogénéité non-paramétrique pour les données Hi-C [1]

Vincent Brault[1,2] & Céline Lévy-Leduc[1,3]& Sarah Ouadah[1,4]& Laure Sansonnet[1,5]

[1] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay*
[2] *vincent.brault@agroparistech.fr*
[3] *celine.levy-leduc@agroparistech.fr*
[4] *sarah.ouadah@agroparistech.fr*
[5] *laure.sansonnet@agroparistech.fr*

Résumé. Le but de cet exposé est de proposer une méthode statistique pour analyser les données Hi-C. Ces données fournissent des informations sur le degré d'interaction physique entre différentes positions du génome (voir par exemple Dixon et al. [1]). Elles se présentent sous forme de matrices dans lesquelles les zones de forte interaction correspondent à des blocs homogènes. Le but est de proposer une méthode automatique pour estimer les frontières de ces blocs homogènes.

Dans cet exposé, nous proposons un test d'homogénéité non-paramétrique pouvant être vu comme une généralisation du test de rang de Wilcoxon et nous en étudions les propriétés asymptotiques. Ces propriétés sont validées sur des données simulées.

Mots-clés. Tests, Statistique non-paramétrique, Données Hi-C.

Abstract. The goal of this talk is to propose a statistical method for analyzing Hi-C data. These data measure the interactions between various positions of the genome (see for example Dixon et al. [1]). They can be summarized as matrices in which zones of strong interactions correspond to homogeneous blocks. We propose in this talk an automatic approach for estimating the boundaries of these homogeneous blocks.

In this talk, we shall propose a nonparametric homogeneity test which can be seen as a generalization of the Wilcoxon rank test and we study its asymptotic properties. We shall also provide some numerical experiments in order to support our claims.

Keywords. Tests, Nonparametric statistic, Hi-C data.

1 Introduction

La technologie Hi-C (High Chromosome Contact map) permet de mesurer le degré d'interaction physique entre différents *loci* (positions) le long d'un chromosome. Les données Hi-C sont représentées sous forme d'une matrice symétrique ou non et peuvent

1. Les auteurs remercient le projet d'ANR ABS4NGS qui a financé en partie ces travaux.

être modélisées par des variables aléatoires ayant des lois identiques au sein de blocs homogènes formant un quadrillage. Le but de l'analyse des données Hi-C est de fournir une méthode automatique et efficace d'estimation des frontières de ces blocs. Il existe des méthodes pour détecter des blocs sur la diagonale (voir Dixon et al. [1] et Lévy-Leduc et al. [3]), notre but ici est d'estimer ces frontières en prenant en compte les interactions extra-diagonales dans un cadre non-paramétrique.

Pour cela, nous proposons d'étendre au cas multivarié la statistique de rang de Wilcoxon dans le prolongement de l'article de Lung-Yut-Fong et al. [4] dans lequel les observations appartiennent à un espace $\mathbb{R}^K$ avec K fixé tandis que, dans notre cas, la dimension de l'espace est égale au nombre d'observations. Nous présenons ici un test d'homogénéité non-paramétrique à deux échantillons qui peut s'étendre à plusieurs échantillons et donc à la détection de ruptures multiples.

2 Cadre statistique

2.1 Hypothèses de test

Les données sont représentées sous la forme d'une matrice symétrique $\mathbf{X} = (X_{i,j})_{1\leq i,j\leq n}$ de taille $n \times n$ où chaque case $X_{i,j}$ représente le degré d'interaction entre le *locus* i et le *locus* j. Les variables $(X_{i,j})_{i\geq j}$ sont supposées indépendantes.
Nous allons proposer un test d'homogénéité entre les deux échantillons : $(\mathbf{X}_1, \ldots, \mathbf{X}_{n_1})$ et $(\mathbf{X}_{n_1+1}, \ldots, \mathbf{X}_n)$ pour $1 \leq n_1 \leq n-1$ et $\mathbf{X}_j = (X_{1,j}, \ldots, X_{n,j})'$, où pour un vecteur U, U' représente la transposée de celui-ci.

Plus précisément, les hypothèses du test d'homogénéité sont les suivantes.

$(\mathcal{H}_0)$: $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ sont identiquement distribuées.
$(\mathcal{H}_1)$: les échantillons $(\mathbf{X}_1, \ldots, \mathbf{X}_{n_1})$ et $(\mathbf{X}_{n_1+1}, \ldots, \mathbf{X}_n)$ suivent des lois différentes.

Comme la matrice est symétrique, nous pouvons la décomposer en quatre blocs :

$$\mathbf{X} = \left(\begin{array}{c|c} R_{0,0} & R_{0,1} \\ \hline R_{1,0} & R_{1,1} \end{array}\right) \tag{1}$$

avec $R_{1,0}^T = R_{0,1}$. De plus, pour les deux hypothèses du test, les variables d'un même bloc suivent la même loi :

- $X_{i,j} \sim \mathbb{P}_{0,0}, \quad 1 \leq i, j \leq n_1$;
- $X_{i,j} \sim \mathbb{P}_{1,1}, \quad n_1 + 1 \leq i, j \leq n$;
- $X_{i,j} \sim \mathbb{P}_{0,1}, \quad 1 \leq i \leq n_1$ et $n_1 + 1 \leq j \leq n$ ou $n_1 + 1 \leq i \leq n$ et $1 \leq j \leq n_1$.

Remarque 1 *En particulier, nous pouvons réécrire les hypothèses du test de la façon suivante :*
$(\mathcal{H}_0)$ *:* $\mathbb{P}_{0,0} = \mathbb{P}_{1,1} = \mathbb{P}_{0,1}$.
$(\mathcal{H}_1)$ *:* $\mathbb{P}_{0,0} \neq \mathbb{P}_{0,1}$ *ou* $\mathbb{P}_{1,1} \neq \mathbb{P}_{0,1}$.

2.2 Statistique de test

Pour mettre en place le test d'homogénéité, nous avons choisi d'adapter la statistique proposée par Lung et al. [4] qui étend la statistique de rang de Wilcoxon–Mann–Whitney au cas multivarié. Pour cela, nous introduisons pour chaque ligne $i \in \{1, \ldots, n\}$, la U-statistique :

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

et nous définissons la statistique de test :

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

3 Propriétés de la statistique de test

Nous établissons le résultat suivant.

Théorème 1 *Sous l'hypothèse $\mathcal{H}_0$, sous la condition que la fonction de répartition des $X_{i,j}$ soit continue et sous la condition qu'il existe $\tau_1 \in]0;1[$ tel que*

$$\frac{n_1}{n} \underset{n \to +\infty}{\longrightarrow} \tau_1,$$

nous avons

$$\frac{S_n(n_1) - \mathbb{E}\left[S_n(n_1)\right]}{\sqrt{n}} = O_P(1)$$

lorsque n tend vers l'infini avec

$$\mathbb{E}\left[S_n(n_1)\right] = \frac{n+1}{3}.$$

Ce théorème montre que notre statistique de test recentrée et normalisée est bornée en probabilité. L'estimation des quantiles de notre statistique de test pourra être faite en utilisant des approches de type bootstrap (voir Efron et Tibshirani [4]).

4 Simulations

Estimation des quantiles

Pour estimer les quantiles d'ordre 95% de la loi de $T_n(n_1) = \frac{S_n(n_1) - \frac{n+1}{3}}{\sqrt{n}}$ sous l'hypothèse $\mathcal{H}_0$, nous avons simulé 1000 matrices symétriques de tailles $n \times n$ avec $n \in$

$\{50, 100, 500, 1000\}$ avec les cases $(X_{i,j})_{i \geq j}$ indépendantes de loi normale ($\mathcal{N}(0,1)$), de Cauchy ($\mathcal{C}au(0,1)$) ou exponentielle ($\mathcal{E}xp(2)$). Pour la statistique, nous avons pris $n_1 \in \{0.1n, 0.5n\}$. La table 1 recense les quantiles empiriques d'ordre 95% obtenus pour chaque configuration.
La valeur estimée semble être la même pour tous les cas et être aux alentours de 0.78.

	$n_1 = 0.1n$			$n_1 = 0.5n$		
	$\mathcal{N}(0,1)$	$\mathcal{C}au(0,1)$	$\mathcal{E}xp(2)$	$\mathcal{N}(0,1)$	$\mathcal{C}au(0,1)$	$\mathcal{E}xp(2)$
$n = 50$	0.87	0.82	0.87	0.76	0.81	0.85
$n = 100$	0.84	0.81	0.81	0.77	0.74	0.74
$n = 500$	0.79	0.75	0.77	0.78	0.80	0.80
$n = 1000$	0.78	0.81	0.81	0.83	0.79	0.74

TABLE 1 – Tableau des quantiles d'ordre 95% estimés de la loi de $T_n(n_1)$ obtenus par simulations suivant la valeur de n_1 (colonnes principales), la loi utilisée (sous-colonnes) et la taille des matrices (lignes)

Estimation en pratique

En pratique, nous pouvons choisir d'estimer les quantiles empiriques à partir des données. Pour cela, nous utilisons une procédure bootstrap considérant le fait que les matrices doivent être symétriques avec des variables indépendantes (pour la partie supérieure de la matrice) et de même loi. Admettant qu'il n'y a qu'une seule rupture n_1 nous serions dans la configuration du modèle (1) où les blocs $R_{0,0}$ et $R_{1,1}$ vérifient alors les conditions souhaitées. Nous choisissons de placer la rupture n_2 au sein de l'un de ces blocs de sorte que la proportion d'observations avant celle-ci soit égale à celle avant la rupture n_1 au sein de $\mathbf{X}$.
Pour la procédure sur ces blocs, nous expliquons la démarche pour le bloc symétrique $R_{0,0}$:

1. Création d'une matrice $\mathbf{Y}$ de même taille que $R_{0,0}$ (c'est-à-dire $n_1 \times n_1$).
2. Calcul de $n_2 = [n_1^2/n]$ où $[\cdot]$ représente la partie entière.
3. Pour $iter_{Boot}$ allant de 1 à N_{Boot} fois :
   (a) Remplissage de la partie supérieure droite incluant la diagonale de la matrice $\mathbf{Y}$ en tirant avec remise dans la partie supérieure droite de $R_{0,0}$.
   (b) Symétrisation de la matrice.
   (c) Calcul de la statistique $T_{n_1}^{(iter_{Boot})}(n_2)$.

4. Calcul du quantile d'ordre 95% de l'échantillon $\left(T_{n_1}^{(1)}(n_2), \ldots, T_{n_1}^{(N_{Boot})}(n_2)\right)$.

En ce qui concerne les variables du bloc $R_{1,0}$, nous appliquons la même procédure avec des matrices symétriques de taille $n_3 \times n_3$ avec $n_3 = \left[\sqrt{2n_1(n-n_1)}\right]$ et en remplissant la partie supérieure droite en tirant avec remise dans toutes les valeurs du bloc $R_{1,0}$ (qui sont par définition indépendantes).

Nous avons simulé 100 matrices pour chacune des configurations précédentes, lancé la procédure avec $N_{Boot} = 100$ et estimé empiriquement un quantile d'ordre 95% pour les procédures de chacun des 3 blocs ; nous avons également estimé le quantile en agrégeant les statistiques des trois procédures. Dans la table 2 sont recensés les moyennes et écart-types des quantiles obtenus. Nous pouvons voir que lorsqu'il y a peu d'informations (région $R_{0,0}$ pour $n_1 = 0.1n$), la valeur du quantile est beaucoup plus faible mais quand le nombre d'observations augmente (par exemple quand n augmente), les quantiles se rapprochent des valeurs obtenues précédemment.

5 Perspectives

Dans cet exposé, nous étendrons la statistique proposée aux cas de plusieurs échantillons et à la détection de ruptures multiples, ceci correspondant à l'estimation des frontières des blocs homogènes. Nous présenterons les résultats théoriques obtenus et conclurons par une utilisation sur des données simulées.

Bibliographie

[1] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, et B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398) : 376380, 2012.
[2] Efron, B., et Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.
[3] C. Lévy-Leduc, M. Delattre, T. Mary-Huard, et S. Robin. Two-dimensional segmentation for analyzing hic data. *Bioinformatics*, 30(17) :386392, 2014.
[4] A. Lung-Yut-Fong, C. Lévy-Leduc, and O. Cappé. Distributed detection/localization of change-points in high-dimensional network traffic data. *Statistics and Computing*, 22(2) :485-496, 2012.

		$n_1 = 0.1n$			$n_1 = 0.5n$		
		$\mathcal{N}(0,1)$	$\mathcal{C}au(0,1)$	$\mathcal{E}xp(2)$	$\mathcal{N}(0,1)$	$\mathcal{C}au(0,1)$	$\mathcal{E}xp(2)$
$n = 50$	$R_{0,0}$	-0.13	-0.14	-0.13	0.69	0.69	0.70
		(0.16)	(0.15)	(0.15)	(0.12)	(0.13)	(0.13)
	$R_{1,1}$	0.73	0.74	0.72	0.70	0.70	0.72
		(0.10)	(0.11)	(0.10)	(0.14)	(0.14)	(0.12)
	$R_{1,0}$	0.69	0.70	0.69	0.79	0.80	0.81
		(0.084)	(0.087)	(0.094)	(0.11)	(0.12)	(0.12)
	Tot	0.63	0.63	0.62	0.75	0.75	0.76
		(0.062)	(0.056)	(0.064)	(0.068)	(0.078)	(0.082)
$n = 100$	$R_{0,0}$	0.15	0.13	0.15	0.80	0.77	0.77
		(0.10)	(0.099)	(0.099)	(0.13)	(0.12)	(0.11)
	$R_{1,1}$	0.78	0.78	0.77	0.79	0.78	0.76
		(0.11)	(0.11)	(0.11)	(0.12)	(0.11)	(0.12)
	$R_{1,0}$	0.75	0.74	0.76	0.81	0.79	0.79
		(0.11)	(0.10)	(0.11)	(0.11)	(0.11)	(0.096)
	Tot	0.67	0.67	0.67	0.82	0.80	0.79
		(0.066)	(0.060)	(0.061)	(0.061)	(0.069)	(0.068)
$n = 500$	$R_{0,0}$	0.74	0.74	0.73	0.77	0.78	0.77
		(0.090)	(0.10)	(0.094)	(0.091)	(0.11)	(0.091)
	$R_{1,1}$	0.76	0.76	0.76	0.78	0.77	0.79
		(0.096)	(0.10)	(0.11)	(0.11)	(0.10)	(0.10)
	$R_{1,0}$	0.77	0.77	0.75	0.78	0.78	0.76
		(0.10)	(0.099)	(0.11)	(0.098)	(0.11)	(0.12)
	Tot	0.77	0.77	0.76	0.79	0.79	0.79
		(0.055)	(0.061)	(0.057)	(0.051)	(0.063)	(0.057)
$n = 1000$	$R_{0,0}$	0.76	0.78	0.77	0.75	0.76	0.78
		(0.10)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)
	$R_{1,1}$	0.77	0.75	0.77	0.78	0.76	0.78
		(0.091)	(0.10)	(0.096)	(0.11)	(0.11)	(0.097)
	$R_{1,0}$	0.76	0.74	0.74	0.77	0.76	0.74
		(0.099)	(0.098)	(0.099)	(0.093)	(0.10)	(0.10)
	Tot	0.78	0.77	0.77	0.78	0.78	0.78
		(0.059)	(0.062)	(0.060)	(0.058)	(0.058)	(0.060)

TABLE 2 – Tableau des moyennes et écart-types des quantiles d'ordre 95% estimés de la loi de $T_n(n_1)$ obtenus par méthode bootstrap suivant la valeur de n_1 (colonnes principales), la loi utilisée (sous-colonnes), la taille des matrices (lignes principales) et la région (sous-lignes)