Segmentation bidimensionnelle pour la détection de domaines dans les données HiC

Maud Delattre [1,2] & Céline Lévy-Leduc [1,2] & Tristan Mary-Huard [1,2,3] & Stéphane Robin [1,2]

maud.delattre@agroparistech.fr
[1] *INRA, UMR 518 MIA, F-75005 Paris, France*
[2] *AgroParisTech, UMR 518 MIA, F-75005 Paris, France*
[3] *UMR de Génétique Végétale, INRA, Université Paris-Sud, CNRS, Gif-sur-Yvette, France*

Résumé. Les données HiC mesurent le degré d'interaction physique entre différentes positions chromosomiques. Ces données se présentent sous la forme de matrices où les zones de forte interaction dans le génome forment des blocs diagonaux de valeurs homogènes. Notre objectif est de retrouver les positions (ruptures) délimitant ces blocs. Nous formulons cet objectif sous la forme d'un problème d'estimation de ruptures (segmentation) bidimensionnel. Nous proposons d'estimer les ruptures par maximum de vraisemblance et de réaliser l'inférence au moyen d'un algorithme de programmation dynamique. Nous validons cette méthodologie sur des données simulées et proposons une application sur données réelles.

Mots-clés. HiC, Programmation dynamique, Segmentation.

Abstract. HiC data measure the degree of interaction between various positions within a chromosome. The analyzed data are matrices where zones of strong chromosomic interactions consist in homogeneous diagonal blocks. Our objective is to find the positions (change-points) of these blocks. This can be cast into a two-dimensional change-points estimation (segmentation) problem. We propose to estimate these change-points by using a maximum likelihood approach where the maximization is achieved with a dynamic programming algorithm. This methodology is assessed on synthetic and real data.

Keywords. HiC, Dynamic Programming, Segmentation.

1 Résumé long

La co-expression des gènes est favorisée par la proximité spatiale au sein du noyau de la cellule de régions codantes correspondantes. Les technologies HiC ont été développées dans le but de mieux connaître la conformation spatiale de l'ADN et ainsi mieux comprendre les mécanismes de régulation de l'expression génique. Les données recueillies par ces techniques s'organisent sous forme de matrices (dites matrices d'interactions) dont les

éléments fournissent une mesure du degré d'interaction entre différentes positions chromosomiques. L'objectif de notre contribution est d'identifier les zones d'interaction à partir de ces matrices.

1.1 Modèle

Nous définissons la matrice d'interaction des données HiC comme une matrice symétrique $(Y_{i,j})_{1\leq i,j\leq n}$ dont les éléments $Y_{i,j}$ sont des variables aléatoires indépendantes telles que :

$$Y_{i,j} = Y_{j,i} \sim \mathcal{L}(\mu_{i,j}, \eta),\ 1 \leq i \leq j \leq n, \tag{1}$$

où $\mu_{i,j}$ est le paramètre de moyenne et η est un paramètre de nuisance. Ci-après, nous supposerons que les $Y_{i,j}$ suivent une loi Gaussienne, Poisson ou binomiale négative. Plus précisément, nous supposerons que les $Y_{i,j}$ sont des variables aléatoires indépendantes vérifiant les hypothèses (G), (P) ou (B) suivantes:

$$\begin{aligned} &\text{(G)}: \quad Y_{i,j} \sim \mathcal{N}(\mu_{i,j}, \sigma^2), \\ &\text{(P)}: \quad Y_{i,j} \sim \mathcal{P}(\mu_{i,j}), \\ &\text{(B)}: \quad Y_{i,j} \sim \mathcal{NB}(\mu_{i,j}, \phi). \end{aligned} \tag{2}$$

Une modélisation Gaussienne (G) sera adaptée à des données HiC normalisées alors que les deux autres distributions ((P) et (B)) conviendront aux données HiC brutes qui sont des données de comptage. Dans les modèles (G) et (B), nous supposerons les paramètres σ et ϕ constants et indépendants de i et j. Par ailleurs $(\mu_{i,j})$ est une matrice symétrique vérifiant pour $1 \leq i \leq j \leq n$

$$\mu_{i,j} = \sum_{k=1}^{K^\star+1} \mu_k \mathbf{1}_{\left\{(i,j)\in D_k^\star\right\}} + \mu_0 \mathbf{1}_{\left\{(i,j):\, i\leq j \text{ et } (i,j)\notin \left(\bigcup_{\ell=1}^{K^\star+1} D_\ell^\star\right)\right\}}, \tag{3}$$

avec $\mu_k \neq \mu_\ell$ si $k \neq \ell$, et $D_k^\star$ l'ensemble $D_k^\star = \{(i,j) : t_{k-1}^\star \leq i \leq j \leq t_k^\star - 1\}$ où $1 = t_0^\star < t_1^\star < \cdots < t_{K^\star+1}^\star = n+1$.

Dans le modèle (3), les moyennes des observations forment une matrice diagonale par blocs. Les moyennes sont supposées constantes à l'intérieur des blocs et les $(t_k^\star)$ sont les coordonnées des extrémités de ces blocs. La moyenne μ_0 est supposée constante à l'extérieur des blocs diagonaux. Un exemple est donné en Figure 1.

Dans ce travail, nous proposons une méthodologie efficace pour estimer les ruptures $(t_k^\star)_{1\leq k\leq K^\star}$ et le nombre de ruptures $K^\star$.

1.2 Inférence

Lorsque $K^\star$ est connu, nous estimons les $(t_k^\star)_{1\leq k\leq K^\star}$ par maximum de vraisemblance. Les estimateurs $(\hat{t}_k)_{1\leq k\leq K^\star}$ de $(t_k^\star)_{1\leq k\leq K^\star}$ sont donc obtenus par maximisation par rapport à

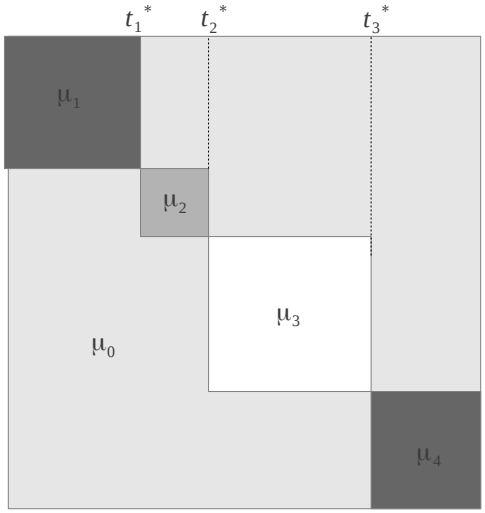

Figure 1: Exemple de matrice diagonale par blocs définie par (3).

$t_1, \ldots, t_{K^\star}$ du critère :

$$\mathcal{L}(K^\star) = \max_{t_0=1<t_1<t_2<\cdots<t_{K^\star}} \sum_{k=1}^{K^\star+1} \sum_{(i,j)\in D_k} \ell_k(Y_{i,j}) + \sum_{\substack{(i,j)\notin(\bigcup_{\ell=1}^{K^\star+1} D_\ell) \\ \text{t.q } i\leq j}} \ell_0(Y_{i,j}) \ , \tag{4}$$

où

$$D_k = \{(i,j) : t_{k-1} \leq i \leq j \leq t_k - 1\}, \tag{5}$$

et où $\ell_.(Y_{i,j})$ désigne la contribution de $Y_{i,j}$ à la log-vraisemblance, *i.e.* $\ell_k(Y_{ij}) = \log p(Y_{ij}; \mu_k)$.

Lorsque la moyenne à l'extérieur des blocs diagonaux μ_0 et le paramètre de nuisance η sont connus, la maximisation de la log-vraisemblance (4) peut être réalisée de manière exacte au moyen d'un algorithme de programmation dynamique. Lorsque μ_0 et η sont inconnus, nous proposons un algorithme approché pour estimer les $(t_k^\star)_{1\leq k\leq K^\star}$.

Lorsque $K^\star$ est inconnu, nous proposons de l'estimer par

$$\hat{K} = \text{argmax}_{K=1,\ldots,K_{max}} \ \mathcal{L}(K),$$

où K_{max} correspond au nombre maximal d'instants de rupture recherchés.

On peut remarquer que les segmentations en un nombre croissant de segments ne sont pas emboîtées les unes dans les autres. $\mathcal{L}(K)$ ne croit donc pas nécessairement avec K.

1.3 Résultats

Cette méthode a été implémentée dans un package R.

Nous avons validé cette méthodologie sur des jeux de données simulées. La Figure 2 illustre les résultats obtenus pour le modèle (B) décrit dans (2).

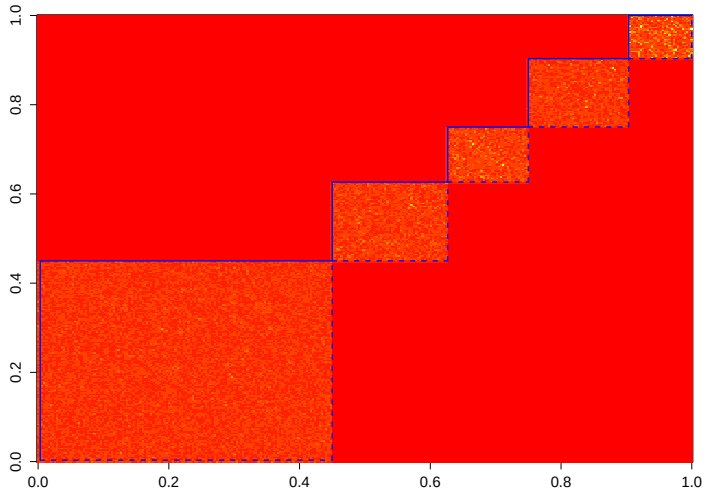

Figure 2: Vrais blocs (partie triangulaire inférieure de la matrice) et blocs détectés par notre méthode (partie triangulaire supérieure de la matrice) sur un jeu de données simulées selon le modèle (B).

Nous avons également appliqué cette méthodologie à des données publiques pour lesquelles Dixon & al. (2012) ont proposé une liste de domaines topologiques, et nous identifions des domaines proches des leurs (voir Figure 3).

Bibliographie

[1] Dixon, J.R. & al. (2012). *Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature*, **485**, 376 – 380

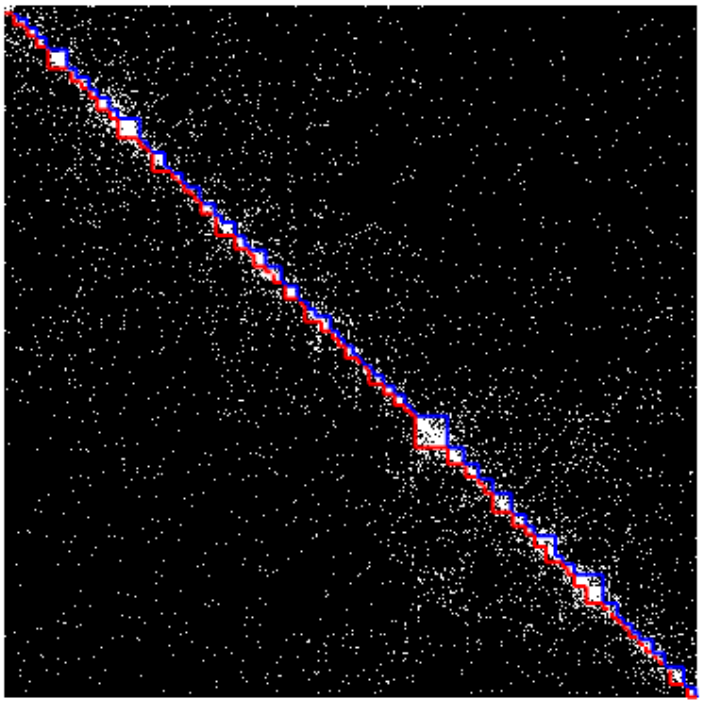

Figure 3: Domaines détectés par Dixon & al. (2012) (partie triangulaire inférieure de la matrice) et par notre méthode (partie triangulaire supérieure de la matrice) sur les données HiC du chromosome 17 de cellules souches embryonnaires humaines.