Classification Ascendante Hiérarchique sous Contrainte de Contiguïté pour l'Analyse de données Hi-C

Nathanaël Randriamihamison

Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *27/10/2021* **par :**

Nathanaël RANDRIAMIHAMISON

Classification Ascendante Hiérarchique sous Contrainte de Contiguïté pour l'Analyse de Données Hi-C

JURY

AVNER BAR-HEN	Professeur, CNAM	Rapporteur
DAVID CAUSEUR	Professeur, Agrocampus Ouest	Examinateur
MARIE CHAVENT	Professeure, Université de Bordeaux	Co-directrice
SYLVAIN FOISSAC	Chargé de Recherche, INRAE	Co-directeur
GUILLEMETTE MAROT	Maîtresse de Conférences, Univ. de Lille	Examinatrice
CATHY MAUGIS-RABUSSEAU	Maîtresse de Conférences, INSA	Examinatrice
PIERRE NEUVIAL	Directeur de Recherche, CNRS	Co-directeur
FRANCK PICARD	Directeur de Recherche, CNRS	Examinateur
NATHALIE VIALANEIX	Directrice de Recherche, INRAE	Co-directrice

---

École doctorale et spécialité :

MITT : Domaine Mathématiques : Mathématiques appliquées

Unité de Recherche :

Mathématiques et Informatique Appliquées de Toulouse (UR875)

Directeur·rice·s de Thèse :

Marie Chavent, *Sylvain Foissac (invité)*, *Pierre Neuvial* et *Nathalie Vialaneix*

Rapporteur·e·s :

Avner Bar-Hen et *Susan Holmes* (Professor, Stanford University)

Remerciements

Tout d'abord, je tiens à remercier mes co-directeurs de thèse Nathalie, Marie, Pierre et Sylvain pour leur aide et leur soutien constant tout au long de mon doctorat. Vous m'avez toujours consacré du temps et soutenu, et c'est grâce à vous si je peux présenter cette thèse aujourd'hui.

Je suis également extrêmement reconnaissant envers l'ensemble des membres du jury : Mme Marot, Mme Maugis-Rabusseau, M. Bar-Hen, M. Causeur et M. Picard, d'avoir acceptés d'être présents pour ma soutenance de thèse. Je souhaite également remercier Mme Holmes et M. Bar-Hen d'avoir rapporté mon manuscrit. Leurs précieuses remarques et suggestions m'ont permis d'approfondir plusieurs points et je leur suis très reconnaissant d'avoir pris ce temps pour apprécier mon manuscrit.

J'ai énormément apprécié travailler au sein de l'unité Mathématiques et Informatique Appliquées de Toulouse pendant ces trois années et demie. J'ai eu beaucoup de chance de pouvoir réaliser cette thèse dans un endroit si accueillant et je tiens à remercier tous les membres de MIAT pour cela. Une mention spéciale pour Fabienne, Alain, Benjamine et Papa qui m'ont aidé dans bien des situations dont je n'aurais pas su me dépêtrer seul. Je remercie Sylvain J. pour les quelques discussions que l'on a pu avoir pendant ces trois ans et ses conseils avisés.

Je remercie également l'Institut de Mathématiques de Toulouse qui m'a permis de bénéficier de nombreuses ressources, ainsi que l'équipe CQFD de l'Inria Bordeaux Sud-Ouest de m'avoir accueilli.

Je souhaite également adresser un message tout particulier à Gaëlle avec qui j'ai partagé notre bureau pendant ces trois années : merci pour tous les gâteaux ! Tes talents culinaires n'ont d'égal que ta gentillesse et j'ai eu beaucoup de chance de pouvoir compter sur toi.

Je suis reconnaissant envers l'Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement et l'Inria de m'avoir permis de réaliser ce doctorat en finançant mon travail.

Enfin, je remercie ma famille de m'avoir soutenu pendant le doctorat : mes parents et ma petite sœur, ma grand-mère et ma belette (ainsi que les deux loupis).

Résumé

L'organisation spatiale du génome à l'intérieur du noyau des cellules a un impact majeur sur la régulation de l'expression des gènes. La technologie de séquençage Hi-C permet d'accéder à des mesures à haute résolution de la proximité spatiale dans le noyau entre paires de positions du génome, sous la forme de matrices de contact.

Dans ce travail, nous nous concentrons sur deux questions biologiques qui se prêtent naturellement à une modélisation statistique dédiée : d'une part, l'étude de l'organisation spatiale hiérarchique du chromosome à partir de données Hi-C, et d'autre part, la comparaison de cette organisation spatiale entre deux conditions biologiques différentes. La première partie est donc dédiée à l'étude des extensions d'un outil statistique naturel pour l'examen de structures hiérarchiques, la classification ascendante hiérarchique (CAH), dans l'optique de l'appliquer aux données Hi-C. En effet, le cadre standard de cette méthode est celui dans lequel les données d'entrée sont fournies sous la forme d'une matrice de distances euclidiennes entre paires d'objets et où les distances entre classes sont calculées grâce au lien de Ward. Cette partie est l'occasion d'uniformiser et de justifier l'application de la CAH à différents types de données (noyaux, similarités et dissimilarités). D'autre part, on s'intéresse à la CAH classique mais aussi à sa version sous contrainte de contiguïté (CAHCC) qui ne réunit que des objets adjacents au sens d'une relation définie a priori. On s'attarde particulièrement sur un cas simple de contrainte de contiguïté, la contrainte d'ordre (CAHCO) où la relation de contiguïté est déterminée par l'aspect ordonné des objets. Elle est adaptée pour représenter l'ordre linéaire induit par le génome et permet une meilleure interprétation biologique des classes intervenant dans la CAH. La question de la cohérence des représentations graphiques (arbre binaires) issues de la CAH, dépendante de divers facteurs (type de données, présence ou absence de contrainte, ...) est étudiée de façon systématique. Enfin, une comparaison de performances entre CAH et CAHCO met en évidence la supériorité de la version contrainte dans le cas où cette contrainte est en adéquation avec la structure des données.

La seconde question est l'occasion de développer une méthode d'analyse différentielle de structures prenant en compte l'aspect hiérarchique des données, basée sur la comparaison des résultats issus de la CAH. Ainsi, l'objectif est de développer un test de comparaison de deux ensembles d'arbres. Une étude bibliographique des approches de comparaisons d'arbres existantes (distances entre arbres essentiellement) est menée et permet de retenir une représentation vectorielle pertinente pour les arbres par leurs vecteurs de distances cophénétiques. Le problème se for-

malise alors comme une comparaison de moyennes multivariées dans un contexte de grande dimension et des approches de type « Hotelling » avec régularisation de variance sont envisagées. La structure de dépendance très particulière des vecteurs de distances cophénétiques nous amène à considérer une approche alternative basée sur une agrégation des p-valeurs associées aux tests individuels des coordonnées. C'est cette dernière approche qui est finalement retenue et validée empiriquement à l'aide d'une simulation basée sur des données GWAS (Genome Wide Association Study). La possibilité d'appliquer la méthode avec pertinence pour des données de phylogénie est également illustrée. Enfin, la méthode est appliquée sur données Hi-C.

Abstract

This work is motivated by a biological problem, namely the study of the three-dimensional structure of genome and of its variations in the cell nuclei. Spatial genome organization within the cell nuclei has a major impact on the regulation of gene expression, with important consequences in fetal development, cell differentiation or in the development of diseases such as cancer. High-resolution measurements of spatial proximity between pairs of genome positions can be obtained by a sequencing-based technology called Hi-C. This thesis focuses on statistical challenges raised by two biological questions in this context : on the one hand, the study of the hierarchical spatial organization of chromosomes using Hi-C data, and on the other hand, the comparison of this spatial organization between two different biological conditions.

The first question is addressed through the study of extensions of a natural statistical tool for the study of hierarchical structures, the Hierarchical Agglomerative Clustering (HAC), in order to apply it to Hi-C data. HAC standard framework handles inputs that are matrices of Euclidean distances between pairs of objects. Usually, the distances between classes are then calculated using Ward's linkage. This part of the work justifies the application of HAC to different types of data (kernels, similarities and dissimilarities). In this section, classical HAC is addressed but also in its contiguity-constrained version, which merges only adjacent objects. In particular, we focus on a simple contiguity constraint, the order constraint (OCHAC), where the contiguity relation is determined by the ordering of the objects. This case is suitable for representing the linear order induced by the genome and allows for a better biological interpretation of the clusters involved in HAC (TAD, meta-TAD, compartments, etc.). The issue of the consistency of graphical representations (binary trees) resulting from HAC, depending on various factors (data type, presence of a constraint, ...) is systematically studied. Finally, a performance comparison between HAC and OCHAC highlights the superiority of the constrained version in the case where this constraint is consistent with the data structure.

The second question is addressed by developing a differential analysis method for hierarchical structures, taking into account the hierarchical aspect of the data. This method is based on the comparison of the results of HAC. The aim is to develop a test for the comparison of two samples of trees. A bibliographic study of existing tree comparison approaches (mainly distances between trees) highlights the relevance of cophenetic distances to represent and compare trees. The question at hand is then formalized as a problem of comparison of means in a high-dimensional context. "Hotelling"-type approaches are considered with different variance regu-

larization options. Finally, the particular dependency structure of the vectors of cophenetic distances leads us to consider an alternative approach based on aggregating the p-values associated with the individual tests of the vector entries. The developed approach is validated by numerical experiments. A first application to GWAS (Genome Wide Association Study) data in a controlled framework allows empirical verification of the behavior of the statistic in the absence of biological difference between conditions. The application of the method is also successfully illustrated with phylogeny data. Finally, the method is applied to Hi-C data.

Table des matières

Notations

Échantillons et conditions :

— n : nombre d'individus (échantillons biologiques) : matrices Hi-C, dendrogrammes, ...
— $i \in \{1, \ldots, n\}$: indexation des individus
— $\mathcal{C}_1$ et $\mathcal{C}_2$: deux conditions biologiques telles que $\mathcal{C}_1 \cup \mathcal{C}_2 = \{1, \ldots, n\}$ et $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$
— $|\mathcal{C}_1| = n_1$ et $|\mathcal{C}_1| = n_2$: nombre d'individus dans chaque condition
— $r \in \{1, 2\}$: indexation des conditions

Matrices Hi-C et bins :

— H^i : matrice Hi-C correspondant à l'individu d'indice i
— B : nombre de bins
— L^i : profondeur de séquençage de la matrice Hi
— $(k, l) \in \{1, \ldots, B\}^2$: indexation des paires de bins
— h^i_{kl} : comptage des interactions physiques pour la matrice i et le couple de bins (k, l)
— $p = B(B-1)/2$: nombre de paires de bins
— $j \in \{1, \ldots, p\}$: indexation (vectorielle) des paires de bins
— c : indexation des chromosomes

Classification Ascendante Hiérarchique :

— D : matrice de dissimilarité
— S : matrice de similarité
— K : matrice de noyau
— G : classe intervenant dans le processus de CAH
— $\delta(G, G')$: lien de Ward entre les classes G et G'
— t : étape de la procédure de CAH
— $\mathcal{P}_t$: partition des feuilles obtenue à l'étape t de la CAH

Dendrogrammes, arbres et matrices de distances cophénétiques :
— τ^i : arbre (ou dendrogramme) correspondant à l'individu d'indice i
— $\mathcal{T}$: espace des arbres enracinés pondérés avec un nombre de feuilles fixées
— $\mathbf{X} = (x_{kl})_{1 \leq k,l \leq B}$: matrice de distances cophénétiques
— X : vecteur de $\mathbb{R}^p$ des coefficients de la partie triangulaire supérieure de $\mathbf{X}$

Estimation :
— $X^{(r)}$: vecteur aléatoire des distances cophénétiques pour la condition r
— $X_{i,j}$: j-ème composante de l'observation $X_i \in \mathbb{R}^p$ $(j = 1, \ldots, p)$
— Σ : matrice de variance-covariance de $X^{(r)}$
— σ_j^2 : variance individuelle associée à la j-ème coordonnée de $X^{(r)}$
— T^2 : Statistique de Hotelling
— χ^2_ν : loi du χ^2 à ν degrés de liberté
— $F(a, b)$: loi de Fisher à a et b degrés de liberté

Chapitre 1

Introduction

1.1 Expression génique et structure tridimensionnelle du matériel génétique

1.1.1 Contexte biologique

Information génétique et ADN. L'information génétique d'un être vivant est contenue dans les molécules d'Acide DésoxyriboNucléique (ADN), macromolécules copiées à l'identique dans l'ensemble de ses cellules. Une molécule d'ADN est composée de deux brins antiparallèles associés, formant un filament en double hélice de 2,5 nm de diamètre. C'est la structure linéaire des brins qui permet de coder l'information génétique sous la forme d'une séquence de bases : l'adénine (A), la thymine (T), la cytosine (C) et la guanine (G). La double hélice formée par les deux brins tire sa cohérence du fait que les bases sont associées en paires complémentaires (appariement adénine-thymine ou cytosine-guanine ; voir figure 1.1). L'ensemble du matériel génétique, commun pour chaque cellule d'un même être vivant, est constitué de plusieurs molécules d'ADN, chacune d'entre elles correspondant à un chromosome. La totalité des séquences associées aux molécules d'ADN d'un organisme constitue son génome.

Au sein de ce génome, des intervalles de plusieurs milliers de paires de bases successives en moyenne, sont utilisés d'un seul tenant par les cellules pour contribuer à leur fonctionnement et, plus généralement, à celui de l'organisme entier (Pearson, 2006). On appelle ces intervalles génomiques des gènes.

Expression génique et régulation. La grande majorité des cellules d'un même organisme ont un génome identique et par conséquent, partagent la même information génétique. Pourtant, elles diffèrent par de nombreux aspects (forme, fonction, métabolisme, ...), ce qui implique qu'elles n'utilisent pas cette information de la même façon. Les mécanismes biochimiques par lesquels l'information génétique stockée dans les gènes est lue et utilisée par les cellules constituent l'expression génique. Ce sont donc des variations dans l'expression des gènes qui permettent aux cellules, à partir d'un même génome, d'avoir la diversité de formes et de fonctionnements que l'on observe en pratique.

C'est l'environnement, au sens large, qui provoque des modifications d'expression génique. En effet, nos cellules reçoivent continûment des informations provenant de leur environnement qui leur permettent, par exemple, d'ajuster leur métabolisme à une situation donnée ou encore de se spécialiser au cours de leur développement. De nombreux mécanismes permettent de réguler l'expression génique. L'un d'entre eux est caractérisé par la conformation spatiale du génome dans le noyau de la cellule. C'est ce mécanisme qui est l'objet d'étude principal de cette thèse.

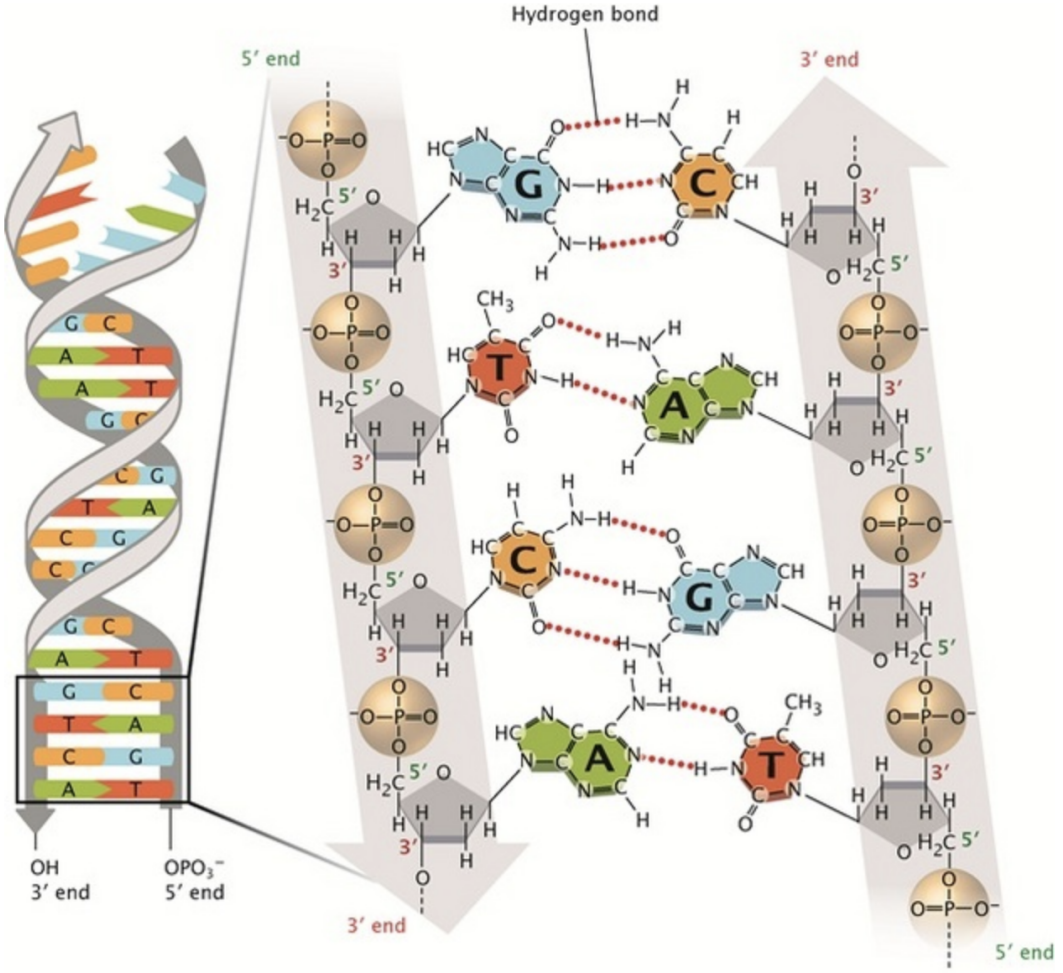


FIGURE 1.1: **Schéma d'une molécule d'ADN.** À gauche, la structure en double hélice est représentée avec les appariements A-T et C-G (bleu pour la guanine, jaune pour la cytosine, rouge pour la thymine et vert pour l'adénine). À droite, le détail moléculaire de cette structure est donnée ; la structure des brins est grisée et les bases sont en couleur. Figure extraite de Pray (2008).

1.1.2 Organisation tri-dimensionnelle du génome

Le nombre de bases constituant la séquence complète du génome est grand, souvent plusieurs milliards pour les mammifères par exemple. Ainsi, si on déroulait le génome d'une cellule d'un bout à l'autre, le résultat serait un filament de plusieurs mètres de long. Or, chez les eucaryotes (la plupart des organismes pluricellulaires ou non bactériens), cette information est contenue dans le noyau des cellules, dont le diamètre en moyenne est inférieur à 10 μm. Ainsi, l'organisation spatiale du gé-

nome à l'intérieur du noyau des cellules est fondamentale pour permettre à autant d'informations d'être contenues dans un si petit volume.

Compaction du génome. La première étape de la compaction du matériel génétique est rendue possible grâce à l'intervention de protéines, les histones. En se liant à l'ADN, elles permettent sa densification en une fibre plus épaisse appelée fibre chromatinienne qui s'enroule également plusieurs fois pour permettre une compaction du génome dans le noyau cellulaire (voir figure 1.2).

La structure tri-dimensionnelle du génome résultant de cette compaction n'est pas aléatoire. Elle fait apparaître des zones plus ou moins denses et ces différences de densités localisées déterminent des niveaux d'accessibilité divers jouant un rôle important dans l'utilisation du matériel génétique par la cellule (Bonev et Cavalli, 2016).

Une organisation multi-niveaux. L'organisation spatiale du génome est donc un sujet d'étude allant de la constitution de base de la fibre chromatinienne jusqu'à la formation des chromosomes et leur positionnement dans le noyau des cellules (Bonev et Cavalli, 2016). Dans cette opération de densification, la fibre chromatinienne met en proximité spatiale des régions du génome qui autrement, seraient distantes les unes des autres. Cette faible distance permet des interactions entre régions génomiques qui peuvent avoir des impacts sur la façon dont le matériel chromosomique est utilisé par la cellule. La conformation du génome peut influer sur le développement d'un organisme ou sur certaines maladies (Zheng et Xie, 2019). Comprendre les mécanismes de régulation associés à cette organisation est donc un enjeu important.

L'organisation spatiale du matériel génétique se met en place à différentes échelles, et ce de façon imbriquée, comme illustré dans la figure 1.3. Au niveau le plus grossier, les différents chromosomes ont tendance à occuper leurs régions propres dans le noyau et définissent ainsi des territoires chromosomiques. Au sein de ces territoires, la chromatine se répartit en zones plus ou moins denses, résultant en une compartimentation en deux classes (appelées A et B) du matériel génétique, qui rend compte de l'état de densité et d'accessibilité du matériel génétique. Si l'on réduit encore l'échelle, on trouve les Topologically Associating Domains (TADs), régions dont la taille est de l'ordre de la mégabase. Ce sont des domaines contigus le long du génome, composés de régions qui interagissent préférentiellement entre elles. Enfin, au niveau le plus fin généralement étudié, se trouvent les boucles de chromatine, structures de base de la fibre chromatinienne : elles sont contigües et permettent le contact physique de deux positions génomiques précises.

Dans cette thèse, on s'intéressera principalement au niveau d'organisation des TADs.

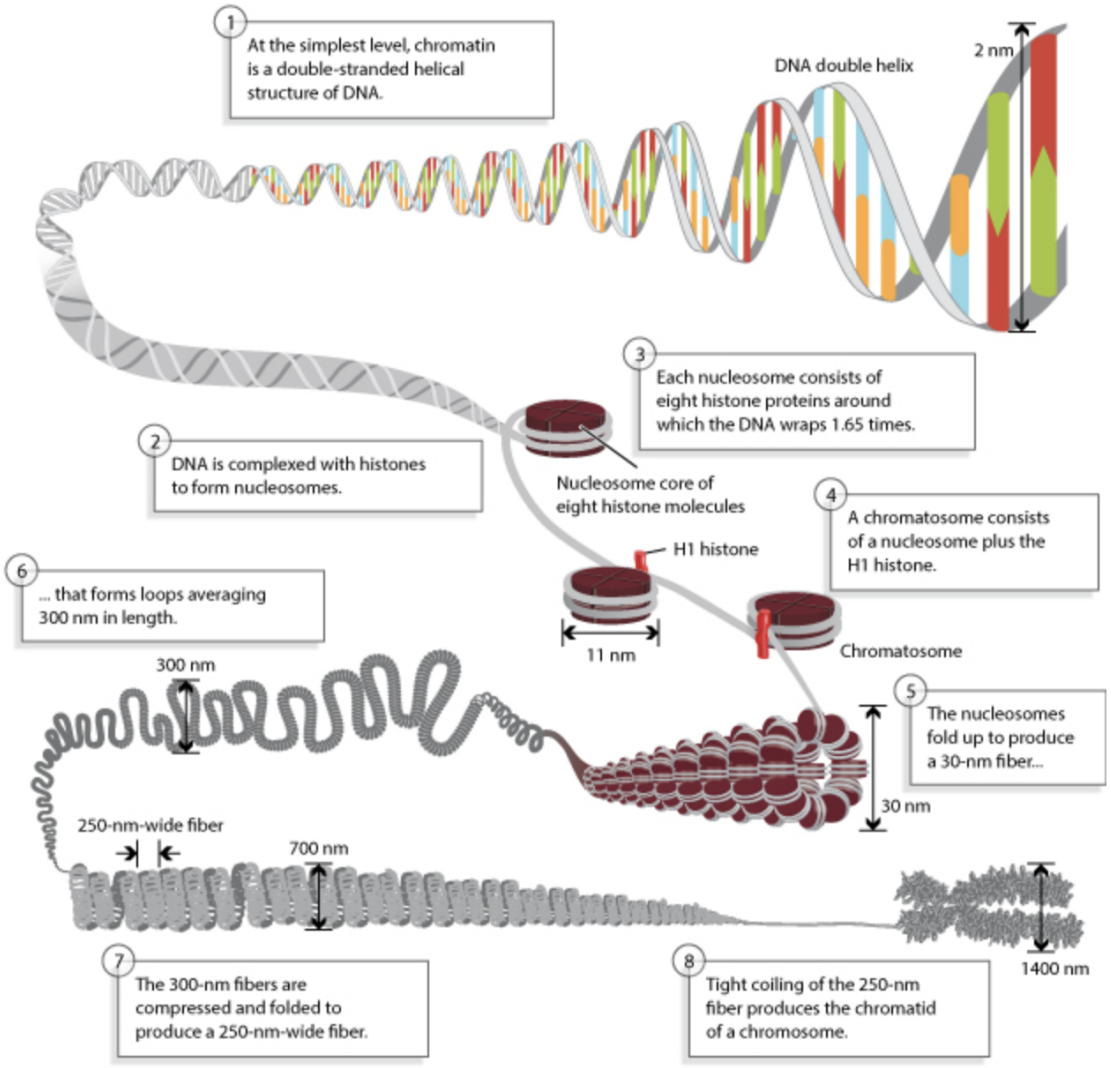


FIGURE 1.2: **Repliement et compaction de la molécule d'ADN.** La figure présente successivement la formation des nucléosomes, la formation de la fibre chromatinienne, puis le repliement et la compaction de la fibre chromatinienne jusqu'à la formation des chromosomes condensés. Figure extraite de Annunziato (2008).

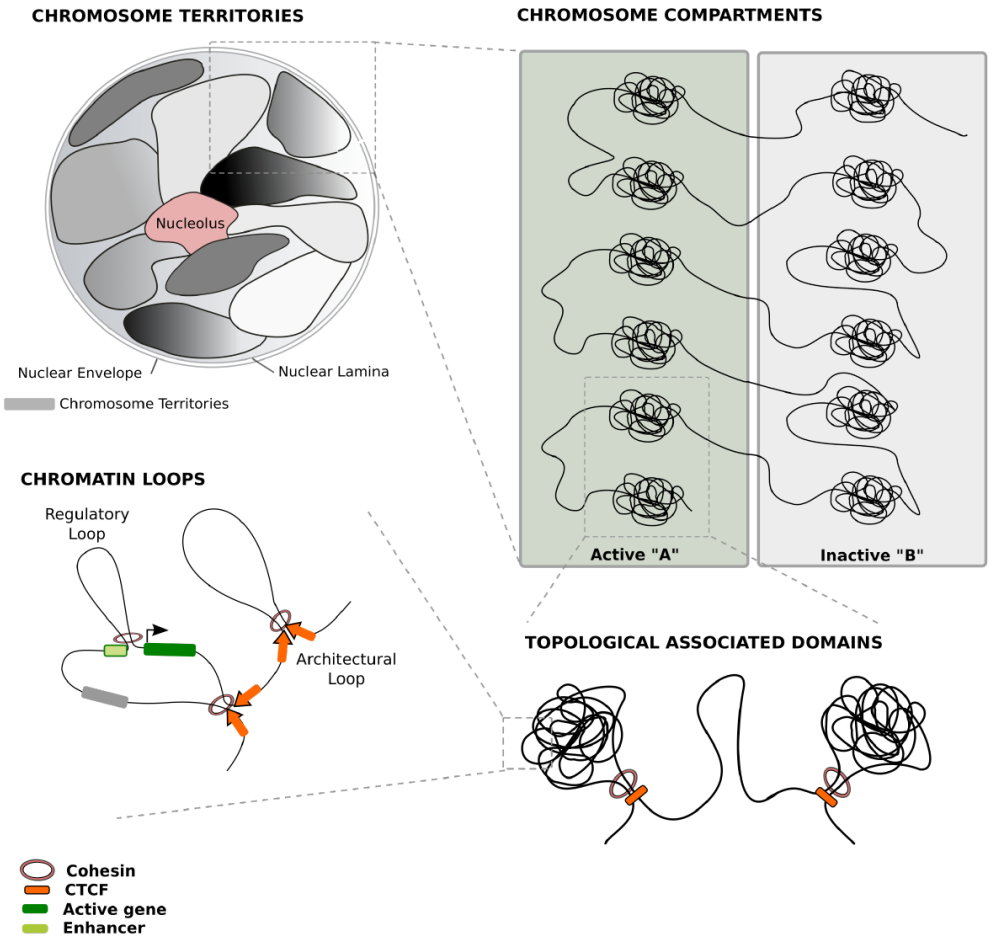


FIGURE 1.3: **Organisation multi-niveaux du génome.** Le génome est spatialement organisé de façon hiérarchique. Les chromosomes sont condensés dans des territoires propres dans le noyau. À l'intérieur des chromosomes, on distingue des compartiments déterminés par l'état de la chromatine : active (A) et inactive (B). À un niveau plus fin, la chromatine se divise en régions contigües du génome à l'intérieur desquelles les interactions sont fréquentes : ce sont les Topologically Associating Domains (TADs). Enfin, à la plus petite échelle considérée, les boucles de chromatine permettent la mise en contact de régions spécifiques du génome. Figure extraite de Servant (2017).

Les Topologically Associating Domains (TADs). Dans leurs travaux, Dixon et collab. (2012) ont mis en évidence l'existence de régions génomiques au sein desquelles les interactions entre positions génomiques sont plus intenses,

appelées Topologically Associating Domains (TADs). Il s'agit de structures ayant une grande importance dans l'organisation spatiale de la chromatine et qui jouent un rôle important dans la régulation génique.

En effet, deux positions génomiques au sein d'un même TAD ont tendance à intéragir plus souvent entre elles qu'avec des positions extérieures au TAD. Les TADs sont séparés par des frontières, caractérisées par des enrichissements en certains types de protéines comme la protéine CTCF ou la cohésine par exemple. Ces domaines ont la particularité d'être, en grande partie, conservés entre certaines espèces (la souris et l'homme par exemple) et entre lignées cellulaires également (Dixon et collab., 2012). Ils constituent les principales unités fonctionnelles intervenant dans l'organisation spatiale du génome. Ainsi, ils jouent un rôle par exemple dans les mécanismes de régulation de la transcription, dans la réplication de l'ADN, ou expliquent la co-régulation de certains gènes (voir Dixon et collab. (2016) pour une revue détaillée sur les TADs).

Une organisation imbriquée. Les différents niveaux d'organisation (territoires, compartiments, TADs, boucles) présentés précédemment sont une représentation nécessairement simplifiée de la structure tri-dimensionnelle et multi-échelles du génome. Certains auteurs ont défini d'autres structures, intermédiaires, comme les sous-TADs, zones de compaction très importante à l'intérieur des TADs (Berlivet et collab., 2013), ou les méga-TADs, regroupement de TADs (Fraser et collab., 2015), pour mieux décrire la continuité de l'imbrication existant au niveau même des TADs. Enfin, de plus en plus de travaux, comme ceux de Fraser et collab. (2015) ou Soler-Vila et collab. (2020), s'orientent vers une représentation hiérarchique des chromosomes pour rendre mieux compte des niveaux d'organisation imbriqués (essentiellement sous-TADs, TADs, méga-TADs et leurs interactions).

Modification de la structure tri-dimensionnelle du génome et conséquences. L'organisation spatiale du génome est donc un mécanisme complexe de régulation de l'expression des gènes. Elle influe sur l'accessibilité des gènes concernés à travers l'état de la chromatine mais aussi sur les interactions entre positions génomiques en rapprochant certaines régions génomiques parfois distantes de plusieurs centaines de milliers de bases.

Des changements dans l'organisation spatiale de la chromatine ont des conséquences parfois délétères sur la façon dont les gènes s'expriment (Kaiser et Semple, 2017; Lupiáñez et collab., 2016). Lupiáñez et collab. (2015) ont par exemple montré que la disparition d'une frontière entre deux TADs pouvait être responsable de différentes malformations de la main comme la polydactylie. Des modifications de frontières de TADs semblent aussi impliquées dans l'oncogenèse en favorisant des expressions aberrantes pour certains gènes (Northcott et collab., 2014; Hnisz

et collab., 2016).

1.2 Les données Hi-C

1.2.1 Différentes méthodes d'observation de la structure tridimensionnelle du génome

Il existe de nombreuses méthodes pour étudier l'organisation spatiale du génome (Kempfer et Pombo, 2019). Les premières approches utilisées ont été basées sur des méthodes de microscopie. En particulier, l'hybridation *in situ* en fluorescence (FISH) consiste à marquer des régions de la séquence d'ADN prédéterminées à l'aide de sondes fluorescentes complémentaires des séquences visées et d'observer les zones fluorescentes ainsi créées par microscopie. Ces approches sont limitées car elles ne permettent d'observer qu'un nombre restreint d'interactions, entre régions que l'on connaît suffisamment bien pour pouvoir créer une sonde associée.

Depuis une douzaine d'années, l'avènement des méthodes de séquençage haut débit (Next Generation Sequencing) a permis le développement de plusieurs types d'approches permettant d'obtenir des informations sur la structure 3D de manière plus massive. Le séquençage haut débit est une technologie qui permet la lecture rapide de la composition séquentielle en bases (A, C, T, G) de millions de petits fragments (morceaux) d'ADN simultanément. Les mots ainsi obtenus sont appelés lectures et il existe des approches bioinformatiques (« alignement ») pour déterminer leur emplacement d'origine sur le génome (à condition que ce génome soit connu au préalable) de l'espèce ou du type cellulaire dont sont issus les fragments (voir figure 1.4). En particulier, le séquençage est devenu la méthode standard pour identifier la séquence génomique d'une espèce ou pour mesurer le niveau d'expression des gènes dans un échantillon biologique d'intérêt (données « RNA-seq »).

Le séquençage à haut-débit est utilisé dans de nombreuses expériences de biologie moléculaire. Depuis une douzaine d'années il fait notamment partie d'une technique permettant d'étudier la structure tridimensionnelle du génome : le protocole Hi-C, pour HIgh-throughput Chromatin Conformation Capture (Lieberman-Aiden et collab. (2009)).

Dans le reste de cette thèse, je me suis plus particulièrement focalisé sur les approches Hi-C.

1.2.2 Méthode d'obtention des données Hi-C

Le principe de l'Hi-C est de mesurer la fréquence à laquelle deux positions génomiques sont observées en contact dans le noyau des cellules. Belton et collab. (2012) expliquent les différentes étapes de cette méthode (voir la figure 1.5) :

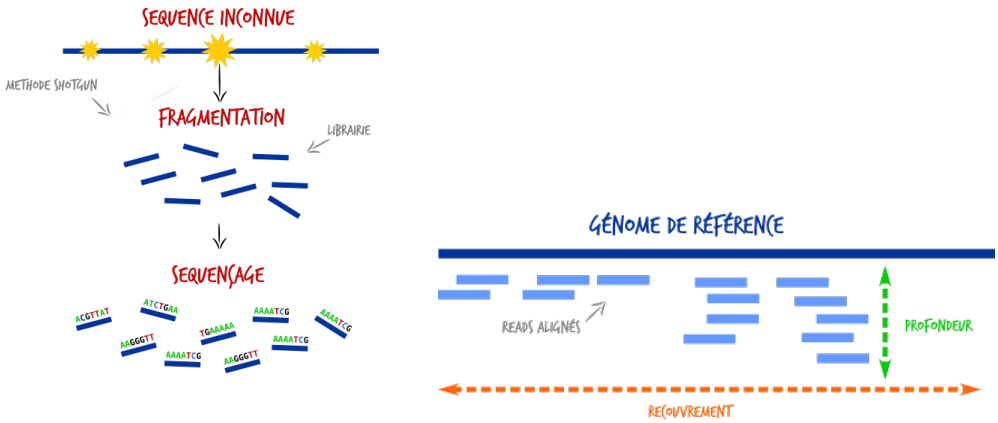


FIGURE 1.4: **Schéma simplifié du séquençage haut débit (gauche) et de l'alignement des lectures sur un génome connu (droite).** Cette dernière étape permet de connaître l'origine du fragment dans le génome et, inversement, de connaître la fréquence d'une position donnée du génome dans l'ensemble des fragments séquencés.
Figure extraite du site https://www.dridk.me/ngs.html.

à partir d'un échantillon de cellules, la première étape consiste à stabiliser les structures existantes avec de la formaldéhyde. L'étape suivante utilise une enzyme de digestion pour découper l'ADN en fragments (dont les extrémités sont marquées avec de la biotine). Vient ensuite l'étape dite de « ligation », qui consiste à rabouter certains fragments par leurs extrémités à l'aide d'une ligase. La particularité de cette ligation est qu'elle dépend principalement de la proximité spatiale entre les fragments : deux fragments ont d'autant plus de chances d'être associés que la distance les séparant est faible. Il en résulte des fragments hybrides marqués à la biotine au niveau du point de ligation. Après nettoyage et réduction, ceux-ci sont récupérés et séquencés par leurs extrémités. Ainsi, la technique produit des paires de lectures caractérisant des positions génomiques suffisamment proches pour avoir été associées par ligation. L'expérience étant réalisée sur des millions de cellules, le nombre d'associations observées entre deux fragments peut être considéré comme une mesure de leur proximité. C'est ainsi que la technologie Hi-C permet d'étudier la structure tridimensionnelle du génome.

1.2.3 Matrice de comptages

A la suite du séquençage, une liste de couples de séquences est obtenue, que le pré-traitement bioinformatique peut associer à des positions génomiques. Il suffit alors de compter ces couples pour obtenir une mesure de la fréquence d'interactions

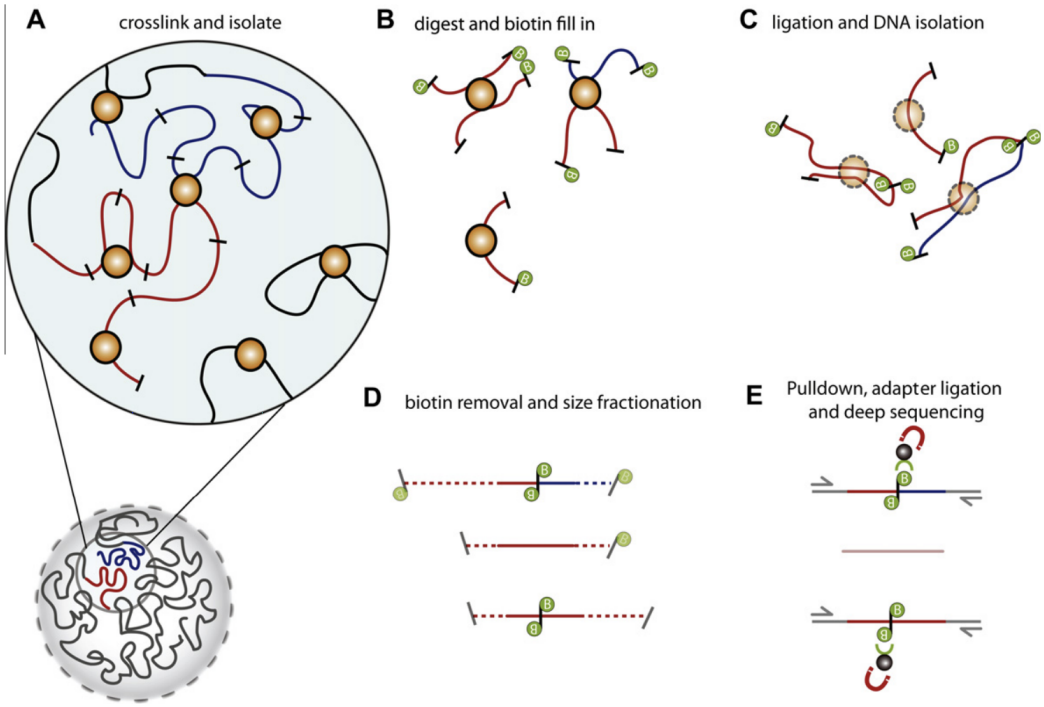


FIGURE 1.5: **Étapes de la méthode Hi-C.** (A) Fixation de l'ADN par le formaldéhyde. (B) Digestion enzymatique de l'ADN et marquage des extrémités des fragments par biotine. (C) Liaison des fragments proches par leurs extrémités. (D) Retrait de la biotine aux extrémités et raccourcissement des fragments. (E) Récupération des fragment hybrides contenant une biotine et séquençage haut-débit. Figure extraite de Belton et collab. (2012).

pour deux positions génomiques données.

De façon naturelle, la majorité des lectures correspondent à des couples de positions génomiques relativement proches du point de vue de la distance génomique (moins d'une mégabase). Pour avoir de l'information sur des contacts entre positions plus éloignées, il est donc nécessaire de séquencer beaucoup de fragments afin d'obtenir suffisamment d'information pour étudier les interactions distantes. Une solution complémentaire consiste à regrouper les observations que constituent les lectures en groupes correspondant à des segments génomiques plus grands, dits bins. Cette étape de *binning* est essentielle et est représentée dans la figure 1.6.

En effet, plus les bins seront grands, plus les nombres de lectures associées seront importants. Bien sûr, cette étape se fait aux dépens de la résolution qui diminue lorsque la taille du bin augmente. Le compromis fixant la taille des bins est donc à déterminer en fonction du nombre final de lectures obtenu après séquençage et du degré d'éloignement séparant les positions d'intérêts.

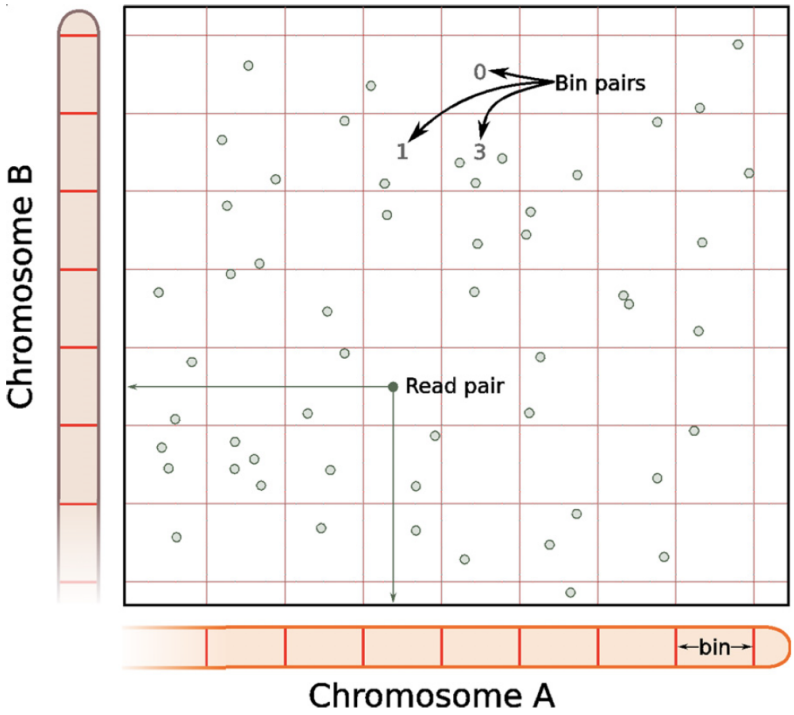


FIGURE 1.6: **Vue schématique de l'étape de *binning*.** Les paires de lectures, représentées par des points, sont positionnées à l'aide de leurs coordonnées génomiques. Elles sont ensuite attribuées à une paire de bins, correspondant à la case du quadrillage où elles se trouvent. Le coefficient de la matrice correspondant à une paire de bins donnée est alors le nombre de paires de lectures tombant dans la case associée à cette paire de bins. Figure extraite de Lun et Smyth (2015).

La représentation la plus classique des résultats groupés en bins est une matrice $H = (h_{kl})_{k,l=1,\dots,B}$ carrée symétrique de taille B égale au nombre de bins considérés, et dont le coefficient (k, l) correspond au nombre d'interactions associées au couple de régions génomiques définies par les bins k et l.

Pour visualiser une matrice Hi-C, on utilise généralement une heatmap, comme illustré dans la figure 1.7. La diagonale, correspondant à des interactions à courte distance, comporte toujours des comptages plus élevés que le reste de la matrice, ce qui se traduit par une forte intensité de couleur. Cela s'explique par le fait que plus des positions génomiques sont proches du point de vue de la distance linéaire du génome, plus elles ont de chances d'interagir en général, ce qui induit un gradient perpendiculaire à la diagonale de la matrice. La décroissance de l'intensité des comptages en fonction de l'éloignement à la diagonale est de tendance exponentielle (Lieberman-Aiden et collab., 2009).

Dans la figure 1.7, on peut également remarquer des carrés de plus forte intensité le long de la diagonale. Ces structures s'interprètent comme des régions du génome à l'intérieur desquelles les interactions entre bins sont plus fortes qu'avec l'extérieur. D'autre part, certains carrés sont eux-même contenus dans des carrés un peu moins intenses, et cette imbrication est à mettre en relation avec l'aspect multi-échelle de l'organisation tridimensionnelle du génome. Ainsi, les données Hi-C permettent la confirmation et l'étude du caractère multi-échelle de la conformation du matériel génétique.

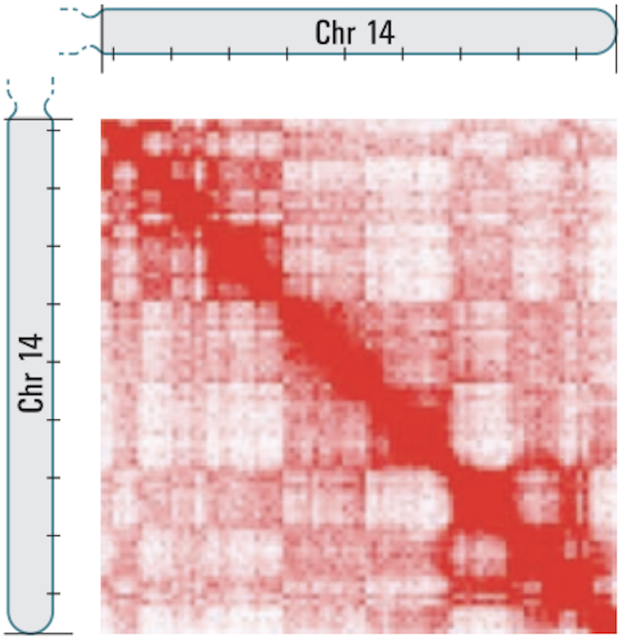


FIGURE 1.7: **Exemple de représentation graphique d'une matrice Hi-C.** Chaque coefficient (k, l) de la matrice correspond au nombre d'interactions entre les bins (régions génomiques) k et l déterminé expérimentalement. L'intensité du contact est représentée par niveaux de rouge. La matrice représentée ici correspond au chromosome 14 humain, avec une taille de bins fixée à 1 Mb. Figure extraite de Lieberman-Aiden et collab. (2009).

1.2.4 Biais

Différents biais liés aux différentes étapes du protocole expérimental Hi-C peuvent introduire du bruit dans les données. Parmi eux, on peut citer, par exemple, la composition en nucléotides de l'ADN : en effet, la proportion en guanine et cytosine (GC) au niveau de la fin des fragments influence les probabilités de

religation (Dohm et collab., 2008). Un autre biais est la variabilité de la longueur des fragments obtenus qui affecte aussi l'efficacité de la religation ainsi que le degré d'unicité de la séquence du fragment dans le génome, rendant l'alignement plus ou moins performant. Enfin, bien que les biais précédents soient essentiellement de nature biologique, on peut ajouter à cette liste des biais de nature technique, liés aux nombreuses et complexes étapes du protocole Hi-C. Yaffe et Tanay (2011) dressent une liste détaillée des biais rencontrés lors de l'obtention de données Hi-C.

Par conséquent, au vu des nombreux biais pouvant introduire du bruit dans la mesure des probabilités d'interactions spatiales entre positions génomiques, la valeur du comptage, h_{kl} d'une paire de bins, (k, l), donnée n'est pas toujours directement comparable à la valeur du comptage d'une autre paire de bins, (k', l') dans la même matrice. Des méthodes de normalisation ont été développées pour corriger ces divers biais appelées *méthodes de normalisation intra-matrice*. Elles se décomposent essentiellement entre méthodes de normalisation paramétriques (qui estiment tous les biais supposés de manière explicite et réalisent un modèle d'ajustement, comme HiCNorm (Hu et collab., 2012)) et méthodes de normalisation non paramétriques (qui font simplement l'hypothèse que les comptages entre tous les bins devraient être équilibrés et corrigent la matrice de manière à assurer cette propriété, comme ICE (Imakaev et collab., 2012)). Toutefois, dans le cadre de la comparaison non pas de paires de bins au sein d'une même matrice mais de matrices elles-mêmes, ces normalisations semblent moins pertinentes, à l'instar des normalisations par la longueur de gènes qui induisent une perte de puissance dans les analyses différentielles de données RNA-Seq (Dillies et collab., 2012). En effet, lors d'une comparaison entre matrices, il apparaît moins important de corriger ces biais internes car on peut supposer que les matrices sont affectées de manière similaire par ces biais. Nous ne développerons donc pas plus cette discussion sur les biais entre bins et leurs corrections, que l'on pourra considérer comme négligeables dans la suite de la thèse, pour le problème d'analyse différentielle auquel je me suis intéressé.

1.3 Analyse différentielle de données Hi-C

L'objet de cette partie est de présenter ce qu'est l'analyse différentielle de données Hi-C. Il s'agit du développement de méthodes comparatives pour détecter des différences significatives entre deux ensembles de matrices Hi-C obtenues dans deux conditions biologiques différentes. Les cas classiques traités dans la littérature incluent par exemple des stades de développement différents (Bonev et collab., 2017) ou encore des types cellulaires différents (Dixon et collab., 2015). L'intérêt de cette recherche est que ces différences de structures peuvent être à l'origine ou bien être la conséquence de la différence entre conditions biologiques et donner des

explications sur les mécanismes de régulation génique impliqués par ou expliquant celles-ci.

Dans la suite, on notera $\{H^i\}_{1\leq i\leq n}$ l'ensemble des matrices Hi-C considérées. Chacune de ces matrices sera associée à l'un des deux groupes $\mathcal{C}_1$ ou $\mathcal{C}_2$ (correspondant aux deux conditions biologiques) de sorte que :

$$\mathcal{C}_1 \cup \mathcal{C}_2 = \{1, 2, ..., n\} \quad \text{et} \quad \mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset.$$

Deux échantillons appartenant à une même condition biologique sont qualifiés de *réplicats biologiques*.

1.3.1 Normalisation entre échantillons

Utilité de la normalisation entre échantillons. Parmi les biais intervenant entre deux matrices Hi-C, l'un d'entre eux est particulièrement important : il s'agit du biais lié à une différence du nombre total de paires de lectures obtenu entre deux matrices Hi-C différentes. Ce nombre total de lectures est couramment appelé *profondeur de séquençage* et correspond à $L^i = \sum_{k\leq l} h^i_{kl}$. En effet, si les valeurs L^i et $L^{i'}$ sont très différentes, les comptages respectifs de la paire de bins (k, l) dans les deux matrices seront impactés en conséquence (une paire (k, l) a une probabilité d'autant plus grande de faire partie des lectures d'un échantillon que le nombre de fragments séquencés est grand) et les valeurs h^i_{kl} et $h^{i'}_{kl}$ ne sont alors pas comparables.

Une approche naïve de correction de ce type de biais consisterait à réaliser une simple transformation linéaire sur les coefficients, comme par exemple :

$$\tilde{h}^i_{kl} = \frac{h^i_{kl}}{L^i} \times L \quad \text{avec} \quad L^i = \sum_{k\leq l} h^i_{kl}$$

où L est un nombre arbitraire fixé, commun pour toutes les matrices. En pratique, on sait que les données issues du séquençage sont fréquemment très asymétriques et que ce type de normalisation naïve est alors influencée de manière négative par quelques paires de bins particulièrement fréquents dans l'échantillon (voir Robinson et Oshlack (2010) pour la description de ce phénomène sur des données de séquençage RNA-seq).

Bien que les méthodes de normalisation classiques de données RNA-Seq (RLE (Anders et Huber, 2010) ou TMM (Robinson et Oshlack, 2010) par exemple) soient applicables aux données Hi-C, Lun et Smyth (2015) estiment que le bruit présent dans ce type de données est trop important par rapport au niveau de bruit standard de données RNA-Seq pour que ces approches de normalisation donnent des résultats satisfaisants. Ils montrent que les biais entre deux matrices peuvent être

caractérisés par une tendance sur le *MA plot*, c'est à dire le nuage de points représentant les valeurs $M_{kl} = \log(h^i_{kl}) - \log(h^{i'}_{kl})$ (différences entre les logarithmes des comptages de deux matrices) en fonction des valeurs $A_{kl} = 1/2(\log(h^i_{kl}) + \log(h^{i'}_{kl}))$ (moyennes des logarithmes des comptages). La méthode estime alors la tendance par régression *loess* (LOcally weighted polynomial regrESSion) et les coefficients des deux matrices sont ensuite normalisés de façon à supprimer cette tendance. Cette méthode de normalisation est illustrée sur la figure 1.8.

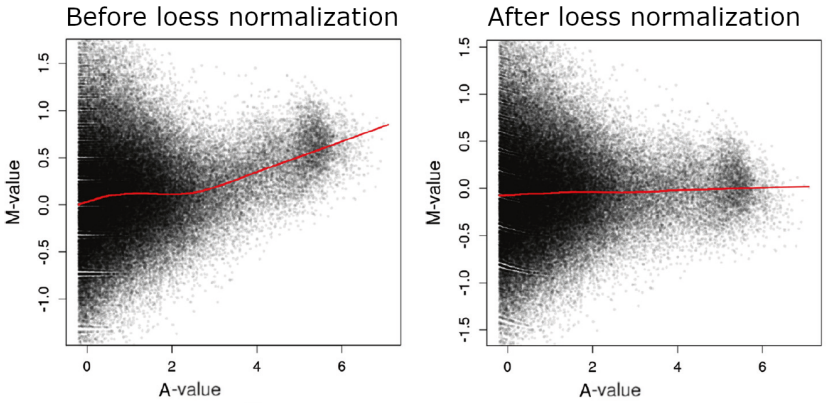


FIGURE 1.8: **Illustration de la méthode de normalisation basée sur MA plot.** Chaque point correspond à une paire de bins. En ordonnées, on représente les valeurs $M_{kl} = \log(h^i_{kl}) - \log(h^{i'}_{kl})$ et en abscisse $A_{kl} = 1/2(\log(h^i_{kl}) + \log(h^{i'}_{kl}))$. La tendance estimée par régression *loess* est représentée en rouge. Le graphique de gauche correspond au MA plot avant normalisation et celui de droite, après normalisation. Figure adaptée à partir de Lun et Smyth (2015).

Stansfield et collab. (2018) proposent une variante de cette approche dans le package R **HiCcompare**, plus adaptée à la structure particulière de la matrice Hi-C où le niveau des comptages h_{kl} est fortement influencé par la distance entre les deux bins $|k - l|$. Pour cela, les auteurs définissent un *MD plot* qui représente les valeurs M_{kl} en fonction de la distance génomique $|k - l|$. Une régression loess est alors utilisée pour estimer et corriger la tendance dans ce graphique, comme pour la normalisation précédente. Cette approche est illustrée dans la figure 1.9.

Pour normaliser plus d'une paire de matrices et donc être capable de prendre en compte des réplicats, ces deux méthodes peuvent être appliquées de manière cyclique (Stansfield et collab., 2019).

1.3.2 État de l'art

Comparaisons coefficients à coefficients. La plupart des méthodes existantes recherchent des différences au niveau des coefficients des matrices Hi-C,

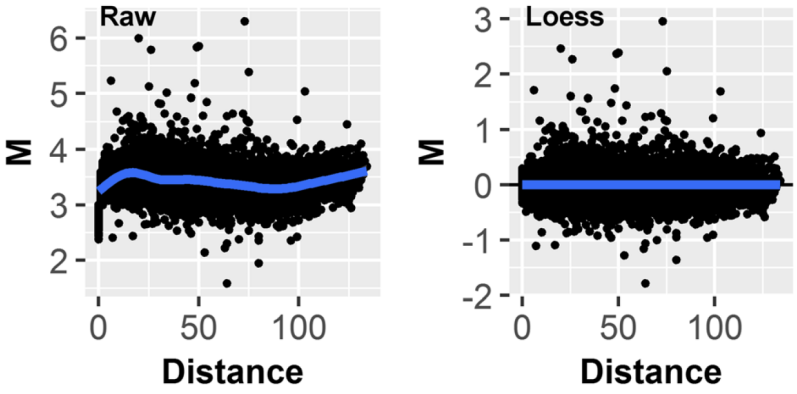


FIGURE 1.9: **Illustration de la méthode de normalisation basée sur MD plot.** Chaque point correspond à une paire de bins. En ordonnées, on représente les valeurs $M_{kl} = \log(h^i_{kl}) - \log(h^{i'}_{kl})$ et en abscisse, les distances génomiques $|k-l|$. La tendance estimée par régression *loess* est représentée en bleu. Le graphique de gauche correspond au MD plot avant normalisation et celui de droite, après normalisation. Figure extraite de Stansfield et collab. (2018).

c'est-à-dire au niveau des paires de bins. En procédant de cette façon, on cherche à obtenir un ensemble de couples de bins dont les interactions sont significativement différentes entre conditions biologiques. Cette approche amène donc à réaliser des millions de tests indépendamment, un test par coefficient (k,l), correspondant à l'hypothèse :

$$\mathcal{H}_0^{kl} : N_{kl}^1 = N_{kl}^2$$

où N_{kl}^r est la loi de la variable aléatoire décrivant le nombre de comptages du coefficient (k,l) pour la condition biologique $\mathcal{C}_r$. De nombreuses méthodes existent dans la littérature, faisant différentes hypothèses sur les lois des N_{kl}^r et l'indépendance de ces variables lorsqu'elles décrivent des paires de bins voisines.

Certaines méthodes de comparaison bins à bins sont basées sur la construction de Z-scores. Par exemple, Stansfield et collab. (2018) ont développé une méthode d'analyse différentielle entre deux matrices Hi-C basée sur le calcul d'un Z-score obtenu à partir de la matrice des différences des log-comptages. Après avoir normalisé les matrices par correction *loess* du MD plot comme décrit dans la section précédente, le Z-score associé au coefficient (k,l) est défini par :

$$z_{kl} = \frac{M_{kl} - \overline{M}}{\sigma_M}$$

où $M = \log(h^1_{kl}) - \log(h^2_{kl})$ est la matrice des log-différences des comptages, $\overline{M}$ est la moyenne des coefficients de cette matrice et σ_M l'écart type associé. L'hypothèse des auteurs consiste alors à considérer ce Z-score comme approximativement

normalement distribué, ce qui permet de le convertir en p-valeur en utilisant les quantiles de la loi normale centrée réduite. Enfin, étant donné le grand nombre de tests indépendamment réalisés, une correction de tests multiples est appliquée pour contrôler le FDR (False Discovery Rate), c'est-à-dire l'espérance de la proportion de faux positifs parmi les tests considérés comme significatifs (Benjamini et Hochberg, 1995).

D'autres méthodes se basent sur une modélisation des comptages par distribution binomiale négative. Lun et Smyth (2015) développent une telle approche pour la comparaison de matrices Hi-C nécessitant au moins 3 réplicats par conditions biologiques. Leur méthode repose sur des outils développés pour les données de RNA-seq (Robinson et collab., 2009), et les tests employés sont analogues à ceux utilisés en RNA-seq. Stansfield et collab. (2019) ont également développé une méthode basée sur une distribution binomiale négative des comptages. Le fonctionnement est similaire à la celui de la méthode de Lun et Smyth (2015), avec la particularité de s'appliquer de façon indépendante sur des groupes de coefficients à une même distance de la diagonale de la matrice, ce qui revient à fixer la distance entre les bins impliqués dans une paire. Ceci permet de prendre en compte la structure particulière des données Hi-C vis-à-vis de la distance à la diagonale (les comptages décroissent exponentiellement avec l'éloignement à la diagonale).

Enfin, certaines méthodes se basent sur la structure de proximité des données Hi-C. Djekidel et collab. (2018) développent une méthode où les comptages voisins d'un coefficient fixé sont modélisés par un processus spatial de Poisson homogène. En utilisant la répartition des comptages dans le voisinage du coefficient considéré, une estimation du paramètre λ d'intensité du processus de Poisson est obtenue par condition et par voisin. On peut alors réaliser plusieurs tests d'égalités des intensités entre conditions qui sont finalement agrégés en une seule p-valeur. Cette méthode nécessite au moins deux réplicats par condition et a l'avantage de permettre de prendre en compte la dépendance existant entre comptages proches du point de vue la distance génomique.

Ces méthodes répondent à la question de la recherche de différences significatives dans l'organisation spatiale du matériel génétique entre conditions biologiques par une liste d'interactions bin à bin. En procédant de cette façon, elle n'utilisent pas, ou peu, l'aspect intrinsèquement multi-échelle des données Hi-C. De plus, elles nécessitent également de tester de très nombreuses paires de bins dans les matrices Hi-C, de sorte qu'elles peuvent induire une perte de puissance à travers l'étape de correction de tests multiples.

Dans l'optique de résoudre ces problèmes, il est possible de modifier la stratégie de recherche de ces différences entre conditions. En effet, une idée peut consister à chercher à réaliser des comparaisons agrégées, basées sur des régions entières du génome, ou de façon équivalente, sur des ensembles de coefficients des matrices Hi-

C. Ces approches permettent ainsi de réduire potentiellement le nombre de tests à réaliser. Elles sont également susceptibles de prendre plus facilement en compte l'organisation multi-échelle du génome. Enfin, elles permettent des interprétations biologiques souvent facilitées en permettant de poser directement des questions complexes (modifications de frontières de TADs, état plus ou moins dense de la chromatine dans une région, etc).

Comparaison de structures. Une première famille d'approches pour les méthodes de comparaisons structurelles est la comparaison de partition. En effet, les différents niveaux d'organisation qui composent la structure spatiale du génome, et plus particulièrement, les TADs, font l'objet d'un intérêt particulier dans l'analyse différentielle de données Hi-C (Dixon et collab., 2016). Une façon de modéliser le groupement des régions génomiques est de considérer des partitions de l'ensemble des bins composant le chromosome (parfois, le génome en entier). La question de la détection de différences significatives entre conditions biologiques se transpose alors en la recherche d'une méthode de comparaison de partitions définies par un certain niveau d'organisation génomique.

La méthode TADcompare (Cresswell et Dozmorov, 2020) est basée sur une représentation sous forme de graphe des matrices Hi-C, ces dernières pouvant être interprétées comme des matrices d'adjacence. En utilisant le laplacien du graphe afin d'en extraire ses vecteurs propres, la méthode construit un vecteur de scores reflétant la probabilité d'une frontière de TAD pour chaque position génomique. L'analyse différentielle pour deux matrices Hi-C est ensuite menée par comparaison des vecteurs de score respectifs, en utilisant un Z-score construit à partir de leur log-différence.

La méthode TADpole (Soler-Vila et collab., 2020) est basée sur une modélisation hiérarchique de la structure tri-dimensionnelle induite par une matrice Hi-C. La matrice Hi-C est d'abord transformée en matrice de corrélation puis une analyse par composante principale est appliquée de façon à réduire la dimension. À partir de ces données, la classification ascendante hiérarchique avec contrainte de contiguïté est appliquée, et une partition du génome est sélectionnée pour décrire la segmentation en TADs du chromosome. Étant données deux partitions en TADs issues de deux matrices Hi-C différentes, les auteurs proposent un score le long du génome, noté DiffT, décrivant les différences cumulées entre les deux partitions depuis le début du chromosome. La première étape consiste à transformer respectivement chacune des deux partitions en une matrice P (resp. Q) dont le coefficient p_{kl} (resp. q_{kl}) vaut 1 si les bins k et l appartiennent à la même classe et 0 sinon. Le score, défini pour un bin b, est alors la proportion de différences depuis le début

de ce chromosome jusqu'à ce bin b :

$$\text{DiffT}(b) = \frac{\sum_{k=1}^{b}\sum_{l=1}^{B}|p_{kl} - q_{kl}|}{\sum_{k=1}^{B}\sum_{l=1}^{B}|p_{kl} - q_{kl}|}$$

où B désigne le nombre total de bins. L'idée est ensuite de considérer les sauts de la fonction comme des zones riches en différences. Les auteurs obtiennent alors des p-valeurs par comparaison avec une distribution empirique obtenue sous un modèle de partitions aléatoires. Cependant, cette méthode ne peut pas prendre en compte la variabilité biologique intra-condition puisqu'elle n'utilise pas de réplicats.

Certaines méthodes se basent sur la comparaison d'arbres modélisant la structure multi-échelle de l'organisation spatiale du génome. Fraser et collab. (2015) ont réalisé des comparaisons d'arbres modélisant l'organisation hiérarchiques des TADs. En effet, à partir de deux matrices Hi-C, les auteurs déterminent deux partitions en TADs puis définissent un ensemble de TADs communs. Une hiérarchie est ensuite construite sur ces TADs communs à partir de leur intensité de contact pour chacune des deux matrices. Les arbres de metaTADs (les feuilles de l'arbre étant des TADs, on se situe à l'échelle d'organisation supérieure) obtenus sont ensuite comparés en utilisant deux critères de comparaison d'arbres : la distance de Robinson-Foulds (Robinson et Foulds, 1981) et le coefficient de corrélation cophénétique (Sokal et Rohlf, 1962). Les auteurs ont ensuite essayé d'obtenir des garanties en utilisant des distributions aléatoires d'arbres développées par Paradis et collab. (2004), mais leur approche ne fournit pas la significativité des résultats obtenus. De plus, il n'est pas possible de comparer plusieurs réplicats pour chaque condition.

La méthode développée dans le cadre de cette thèse pose également la problématique de la recherche de différences significatives du point de vue de la comparaison de structures. Elle se base sur des comparaisons d'arbres modélisant l'organisation multi-échelle du génome, nécessite d'utiliser plusieurs réplicats par condition et permet d'obtenir des garanties statistiques concernant les différences d'organisation détectées.

1.4 Contributions de la thèse

Le travail réalisé au cours de cette thèse est donc motivé par la problématique biologique de la comparaison de la structure tridimensionnelle du génome entre conditions biologiques à partir de données Hi-C.

Le premier axe d'étude abordé dans le chapitre 2 de cette thèse a consisté en l'étude d'un outil statistique particulièrement adapté pour modéliser les structures hiérarchiques : la classification ascendante hiérarchique (CAH). L'idée

était d'utiliser cet outil pour obtenir une modélisation sous forme d'arbre binaire enraciné de la structure hiérarchique induite par les données Hi-C.

À partir des distances euclidiennes entre les objets étudiés, la CAH permet de construire une structure hiérarchique sur ces mêmes objets, représentée par un arbre binaire enraciné, ou dendrogramme. Les données Hi-C pouvant s'interpréter comme des données de similarité, un premier objectif, étudié dans la section 2.3, a été de déterminer et de justifier les possibilités d'extension de la CAH à divers types de données, dont en particulier les similarités. Nous avons ainsi pu justifier l'utilisation de la CAH pour modéliser la structure tridimensionnelle du génome à l'aide des dendrogrammes résultant de l'application de la CAH à une matrice Hi-C.
Ce chapitre a également été l'occasion d'introduire d'étudier la version sous contrainte de contiguïté de cette procédure, qui permet de restreindre les possibilités d'agrégation à chaque étape du processus de façon à obtenir une hiérarchie cohérente avec la problématique considérée (dans notre cas, l'ordre linéaire du génome).
Enfin, les divers cadres d'utilisation de la CAH (type de données, avec ou sans contrainte) ont amené à étudier les propriétés de l'algorithme de CAH et la cohérence de la représentation de ses résultats sous forme de dendrogrammes afin de compléter les résultats existant dans la littérature.
Enfin, nous avons comparé les performances entre la CAH et la CAH avec contrainte d'ordre. On montre de façon empirique que l'ajout d'une contrainte cohérente avec la structure des données améliore les résultats de la CAH en termes d'inertie intra-classes.

Le second axe, développé dans le chapitre 3, s'intéresse à l'élaboration d'une statistique de test dédiée à la comparaison de deux ensembles d'arbres. En effet, les résultats de la partie précédente nous ont permis de transposer de façon justifiée la question de la comparaison de deux ensembles de matrices Hi-C en celle de la comparaison de deux ensembles d'arbres.
Ce chapitre débute par la section 3.1 consistant en une revue de la littérature sur les différentes façons de quantifier les différences entre deux arbres. Cette étude a débouché sur le choix d'une représentation spécifique des arbres par le vecteur de distances cophénétiques. Celle-ci permet de bien quantifier divers types de différences entre arbres qui sont particulièrement intéressants dans notre contexte de comparaison.
La construction d'une statistique de test d'égalité de moyennes entre deux ensembles de vecteurs a ensuite été étudiée dans la section 3.2. On a considéré des approches de type Hotelling dans un contexte de grande dimension, cependant,

l'approche finalement retenue repose sur une agrégation de p-valeurs individuelles. La méthode statistique obtenue s'applique non seulement aux données Hi-C mais aussi à tout contexte impliquant des comparaisons de deux ensembles d'arbres.
Nous avons ensuite validé empiriquement la méthode (section 3.3) par des simulations issues de données GWAS (Genome Wide Association Study ou GWAS). Cela a permis de mettre en évidence les bonnes propriétés de la statistique de test et de confirmer en pratique les hypothèses de validité du test.
Enfin, deux applications ont été réalisées dans la section 3.4 afin de démontrer l'applicabilité de la méthode dans des contextes variés. La première sur des données issues de la phylogénie moléculaire, elles aussi représentées par des arbres, a permis de mettre en évidence la pertinence de la méthode dans un cadre différent de celui d'origine. Enfin, la méthode a été appliquée sur des données Hi-C et ses résultats ont montré une cohérence avec les études précédemment menées sur ce jeu de données.

Chapitre 2

Classification Ascendante Hiérarchique et données Hi-C

Nous avons vu dans l'introduction que les données Hi-C présentent un aspect multi-échelle marqué qui reflète l'imbrication des différents niveaux d'organisation du matériel génétique. Afin d'étudier les variations de cette organisation tridimensionnelle entre conditions biologiques, on peut chercher à modéliser l'aspect hiérarchique des données Hi-C.

La Classification Ascendante Hiérarchique (CAH) est un outil statistique adapté à la modélisation de structures hiérarchiques à partir de données de distances euclidiennes, introduit par Ward (1963). Elle permet, à partir d'une matrice de distance entre objets, de construire un arbre binaire enraciné, ou dendrogramme, qui rend compte de l'organisation mutuelle des objets telle qu'impliquée par la matrice de distance. Il existe différents critères d'agrégation pour la procédure de CAH et dans la thèse, nous utilisons le lien de Ward. En effet, à chaque étape de la CAH, le choix d'une agrégation selon le lien de Ward s'interprète simplement comme la décision minimisant la variation de l'inertie intra-classes.

Les données Hi-C, quant à elles, peuvent s'interpréter comme des similarités, c'est-à-dire comme des informations de proximité entre objets. La question est alors de savoir si on peut étendre la procédure de CAH avec lien de Ward à des données de similarités de façon justifiée. Cette approche permettrait alors de modéliser de façon naturelle l'information hiérarchique contenue dans les données Hi-C sous la forme de dendrogrammes.

D'autre part, la CAH peut également prendre en compte des contraintes de contiguïté entre objets dans le processus d'agrégation (Grimm, 1987; Chavent et collab., 2018). Cette possibilité est particulièrement intéressante dans notre contexte applicatif puisqu'elle permet de conserver l'information que constitue l'ordre linéaire du génome, afin d'obtenir des résultats cohérents d'un point de vue biologique. Dans l'optique d'intégrer une telle contrainte, il est donc nécessaire d'en étudier l'impact sur la procédure de CAH, notamment du point de vue des dendrogrammes.

L'article présenté dans la suite de ce chapitre a pour objet de justifier les extensions de la CAH à divers type de données et d'étudier les conséquences de ces extensions sur les propriétés des dendrogrammes résultant de la procédure. Il a aussi pour objectif de comparer deux versions de la CAH avec lien de Ward : celle avec contrainte de contiguïté (plus spécifiquement, avec contrainte d'ordre) et celle sans contrainte, d'un point de vue théorique et pratique.

Il permettra ainsi de justifier notre approche d'analyse différentielle Hi-C, basée sur une modélisation de l'organisation tridimensionnelle du matériel génétique par des dendrogrammes obtenus par Classification Ascendante Hiérarchique avec Contrainte d'Ordre (CAHCO) et de motiver la formulation du problème d'analyse différentielle de matrices Hi-C comme un problème de comparaison d'arbres binaires enracinés.

La section suivante est donc constitué de l'article en question, (Randriamihamison et collab., 2020), tel que publié dans le *Journal of Classification*. Les annexes de l'article correspondent aux annexes du chapitre 2, placées à la fin de ce manuscrit.

Les notations sont donc légèrement différentes du reste du manuscrit. En particulier, dans ce chapitre :

- n désigne le nombre d'objets à classer par la CAH : $(x_1, \ldots, x_n)$
- p désigne la dimension de l'espace $\mathbb{R}^p$ auquel appartiennent les objets à classer $(x_1, \ldots, x_n)$
- $(i, j) \in \{1, \ldots, n\}^2$ sont des indices réservés aux objets à classer
- h_{ij} désigne la première hauteur à laquelle les feuilles i et j apparaissent dans un même groupe lors du processus de CAH.

Applicability and interpretability of Ward's hierarchical agglomerative clustering with or without contiguity constraints

Nathanaël Randriamihamison[1,2], Nathalie Vialaneix[1] & Pierre Neuvial[2]

[1] INRAE, UR875 Mathématiques et Informatique Appliquées Toulouse, F-31326 Castanet-Tolosan, France
[2] Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS UPS, F-31062 Toulouse Cedex 9, France

Abstract

Hierarchical Agglomerative Clustering (HAC) with Ward's linkage has been widely used since its introduction by Ward (1963). This article reviews extensions of HAC to various input data and contiguity-constrained HAC, and provides applicability conditions. In addition, different versions of the graphical representation of the results as a dendrogram are also presented and their properties are clarified. We clarify and complete the results already available in an heterogeneous literature using a uniform background. In particular, this study reveals an important distinction between a consistency property of the dendrogram and the absence of crossover within it. Finally, a simulation study shows that the constrained version of HAC can sometimes provide more relevant results than its unconstrained version despite the fact that the constraint leads to optimize the objective criterion on a reduced set of solutions at each step. Overall, this article provides comprehensive recommendations, both for the use of HAC and constrained HAC depending on the input data, and for the representation of the results.

2.1 Introduction

Hierarchical Agglomerative Clustering (HAC) with Ward's linkage has been widely used since its introduction by Ward (1963). The method is appealing since it provides a simple approach to approximate, for any given number of clusters, the partition minimizing the within-cluster inertia or "error sum of squares". In addition to its simplicity and the fact that it is based on a natural quality criterion, HAC often comes with a popular graphical representation called a dendrogram, that is used as a support for model selection (choice of the number of clusters) and result interpretation. Originally described to cluster data in $\mathbb{R}^p$, the method has been applied more generally to data described by arbitrary distances (or dissimilarities). Constrained versions of HAC have also been proposed to incorporate a "contiguity" relation between objects into the clustering process (Lebart, 1978; Grimm, 1987; Gordon, 1996).

However, as already shown by Murtagh et Legendre (2014), confusions still exist between the different versions and how the results are represented with a dendrogram, which is also illustrated in (Grimm, 1987) that presents different alternatives for the representation. These have resulted in different implementations of the Ward's clustering algorithm, with notable differences in the results. More importantly, the applicability framework of the different versions is not always clear : Batagelj (1981) has given very general necessary and sufficient conditions on a linkage value to ensure that it is always increasing for any given dissimilarity. This property is important to ensure the consistency between the results of HAC and their graphical display as a dendrogram. Conditions on a general constraint are also provided in Ferligoj et Batagelj (1982) to ensure a similar property and Grimm (1987) proposes alternative solutions to the standard heights to address the fact that the linkage might sometimes fail to provide a consistent representation of the results of HAC. However, none of these articles fully cover the theoretical properties of these alternatives, for unconstrained and constrained versions of the method.

The goal of the present article is to clarify the conditions of applicability and interpretability of the different versions of HAC and contiguity-constrained HAC (CCHAC). We discuss the relevance of these methods to the analysis of different types of input data, and the interpretation of the corresponding results. We perform a systematic study of the monotonicity of the different versions of the dendrogram heights by reporting the results already available in the literature for standard HAC and its extensions and by completing the ones that were not available to our knowledge. In addition to providing a uniform presentation of a number of results partially present in the literature, this study reveals an important distinction between the consistency of representation and the absence of crossover within the dendrogram that was not discussed earlier to our knowledge.

Finally, we illustrate the respective behavior of HAC and CCHAC in a simulation study where different heights are used in order to represent the results. This simulation shows that, in addition to reducing the computational time needed to perform the method, the constrained version (CCHAC) can also provide better solutions than the standard one (HAC) when the constraint is consistent with the data, despite the fact that it optimizes the objective criterion on a reduced set of solutions at each step.

2.2 HAC and contiguity-constrained HAC

2.2.1 Hierarchical Agglomerative Clustering

HAC was initially described by Ward (1963) for data in $\mathbb{R}^p$. Let $\Omega := \{x_1, \cdots, x_n\}$ be the set of objects to be clustered, which are assumed to lie in $\mathbb{R}^p$. A cluster is a subset of Ω. The loss of information when grouping objects into a cluster $G \subset \Omega$ is quantified by the inertia (also known as *Error Sum of Squares*, ESS) :

$$I(G) = \sum_{i \in G} \|x_i - \bar{x}_G\|^2_{\mathbb{R}^p}, \tag{2.1}$$

where $\bar{x}_G = |G|^{-1} \sum_{x_i \in G} x_i$ is the center of gravity of G and $|G|$ denotes the cardinal of the set G. Starting from a partition $\mathcal{P} = \{G_1, \cdots, G_l\}$ of Ω, the loss of information when merging two clusters G_u and G_v of $\mathcal{P}$ is quantified by :

$$\delta(G_u, G_v) := I(G_u \cup G_v) - I(G_u) - I(G_v). \tag{2.2}$$

The quantity δ is known as Ward's linkage and it is equal to the variation of within-cluster inertia (also called *within-cluster sum of squares*) after merging two clusters. It also corresponds to the squared distance between centers of gravity :

$$\delta(G_u, G_v) = \frac{|G_u||G_v|}{|G_u| + |G_v|} \|\bar{x}_{G_u} - \bar{x}_{G_v}\|^2_{\mathbb{R}^p}. \tag{2.3}$$

The HAC algorithm is described in Algorithm 1. Starting from the trivial partition $\mathcal{P}_1 = \{\{x_1\}, \{x_2\}, \cdots, \{x_n\}\}$ with n singletons, the HAC algorithm creates a sequence of partitions by successively merging the two clusters whose linkage δ is the smallest [1], until all objects have been merged into a single cluster. Linkage values at step t can be efficiently updated using linkage values at step $t-1$ with a formula known as the Lance-Williams formula (Lance et Williams, 1967). In the case of Ward's linkage, this formula has first been demonstrated by Wishart (1969) :

$$\begin{aligned} \delta(G_u \cup G_v, G_w) \quad = \quad & \frac{|G_u| + |G_w|}{|G_u| + |G_v| + |G_w|} \delta(G_u, G_w) + \frac{|G_v| + |G_w|}{|G_u| + |G_v| + |G_w|} \delta(G_v, G_w) \\ & - \frac{|G_w|}{|G_u| + |G_v| + |G_w|} \delta(G_u, G_v). \end{aligned} \tag{2.4}$$

The framework of the current section can be extended straightforwardly to the case where the objects to cluster are weighted. However, this study focuses on uniform weights for the sake of simplicity.

1. In the rare situation when the minimal linkage is achieved by more than one merger, a choice between these mergers has to be made. Different choices are made by different implementations of HAC.

Algorithm 1 Standard Hierarchical Agglomerative Clustering (HAC)

1: **Initialization :** $\mathcal{P}_1 = \{\{x_1\}, \{x_2\}, \cdots, \{x_n\}\}$
2: **for** $t = 1$ to $n - 1$ **do**
3: Compute all pairwise linkage values between clusters of the current partition $\mathcal{P}_t$
4: Merge the two clusters with minimal linkage value to obtain the next partition $\mathcal{P}_{t+1}$
5: **end for**
6: **return** $\{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n\}$

2.2.2 HAC under contiguity constraint

A priori information about relations between objects can often be available in applications. For instance, it is the case for spatial statistics, where objects possess natural proximity relations, in genomics, where genomic loci are linearly ordered along the chromosome, or in neuroimaging, with the three-dimensional structure of the brain. According to this point of view, Contiguity-Constrained HAC (CCHAC) allows only mergers between contiguous objects. Considering this approach can have two benefits : (i) more interpretable results by taking into account the natural structure of the data ; (ii) a decreased computational time, because only a subset of all possible mergers are considered.

A very general framework for constrained HAC is described in Ferligoj et Batagelj (1982) : the contiguity is defined by an arbitrary symmetric relation $\mathcal{R} \subset \Omega \times \Omega$ that indicates which pairs of objects are said *contiguous*. Only these pairs are then allowed to be merged at the first step of the algorithm, using the same objective function than in the standard HAC algorithm. The next step iterates similarly, by using the following rule to extend the contiguity relation to merged clusters :

$$(G_u \cup G_v, G_w) \in \mathcal{R} \qquad \Leftrightarrow \qquad (G_u, G_w) \in \mathcal{R} \text{ or } (G_v, G_w) \in \mathcal{R}.$$

Algorithm 2 describes contiguity-constrained hierarchical agglomerative clustering (CCHAC). The only difference with standard HAC lies in the fact that only contiguous clusters are merged. From a computational viewpoint, only the linkage values for a subset of $\mathcal{P}_t \times \mathcal{P}_t$ have to be considered, which can drastically reduce the number of values to be computed with respect to the standard algorithm. This gain in computational time comes at the price of a (potential) loss in the objective function at a given step of the algorithm, especially if the constraint is not consistent with the dissimilarity or similarity values (see Section 2.5 for illustration and discussion). This also has a side effect on standard representations of the result of the algorithm, which is discussed in Section 2.4.

Algorithm 2 Contiguity-Constrained Hierarchical Agglomerative Clustering (CCHAC)

1: **Initialization :** $\mathcal{P}_1 = \{G_1^1, G_2^1, \cdots, G_n^1\}$ where $G_u^1 = \{x_u\}$. Contiguous singletons are defined by $\mathcal{R}_1 = \mathcal{R} \subset \Omega \times \Omega$.
2: **for** $t = 1$ to $n - 1$ **do**
3: Compute all pairwise linkage values between *contiguous* clusters of the current partition $\mathcal{P}_t$ with respect to $\mathcal{R}_t$
4: Merge the two *contiguous* clusters, $G_{v_1}^t$ and $G_{v_2}^t$ with minimal linkage value to obtain the next partition $\mathcal{P}_{t+1} = \{G_u^{t+1}\}_{u=1,\ldots,n-t}$
5: Extend the contiguity relation to the new cluster $G_{v_1}^t \cup G_{v_2}^t \in \mathcal{P}_{t+1}$ by setting

$$\left(G_{v_1}^t \cup G_{v_2}^t, G_w^t\right) \in \mathcal{R}_{t+1} \quad \Leftrightarrow \quad \left(G_{v_1}^t, G_w^t\right) \in \mathcal{R}_t \text{ or } \left(G_{v_2}^t, G_w^t\right) \in \mathcal{R}_t.$$

6: **end for**
7: **return** $\{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n\}$

Order-constrained HAC. A simple and useful case of contiguity constraint is the case when the symmetric relation is a contiguity relation defined along a line. This special case of constraint is frequently encountered in genomics (where the contiguity relation is deduced from genomic positions along a given chromosome) and will be called *order-constrained HAC* (OCHAC) in the sequel. In this context, every cluster has exactly two neighbors (except for the two positioned at the beginning and the end of the line) and at step t of the algorithm, only $n - t$ values of the linkage have to be computed (instead of $(n - t)(n - t - 1)/2$ for standard HAC). This case is the one implemented in the R package **adjclust** and an efficient algorithm is described in Ambroise et collab. (2019) for sparse datasets.

In this specific case, constrained HAC can be seen as a heuristic to approximate the search of an *optimal segmentation* (i.e., achieving minimal ESS among all possible segmentations) of the data into $K(= n - t)$ groups, for each possible K. This problem is also known as the "multiple changepoint problem". Strategies already exist to solve this problem both in Euclidean or non-Euclidean settings, and it is known that the sequence of optimal segmentations for each K can be found efficiently in a quadratic time and space complexity using dynamic programing (Steinley et Hubert (2008); Arlot et collab. (2019)). Nevertheless, those approaches are restrained to order constraints and cannot be applied to more general contiguity constraints, contrary to CCHAC. Moreover, the nestedness of the clustering sequences obtained from HAC allows useful graphical representations such as dendrograms (discussed in Section 2.4.1), contrary to the previously cited methods.

In the present paper, we demonstrate the good properties of the CCHAC for the case of a general contiguity relation and illustrate the opposite situation (where some good properties are not always satisfied for CCHAC) by providing counter-examples and illustrations in the specific case of OCHAC.

2.3 Validity of HAC in possibly non-Euclidean settings

In this section, we systematically justify the use of HAC algorithm (with or without contiguity constraints) for all kinds of proximity data, including dissimilarity and similarity data.

2.3.1 Extension to dissimilarity data

The HAC algorithm of Ward (1963) has been designed to cluster elements of $\mathbb{R}^p$. In practice however, the objects to be clustered are often only indirectly described by a matrix of pairwise dissimilarities, $D = (d_{ij})_{1\leq i,j\leq n}$. Formally, a dissimilarity is a generalization of a distance that is not necessarily embedded into a Euclidean space (*e.g.*, because the triangle inequality does not hold). Here, we only assume that D satisfies the following properties for all $i, j \in \{1, \ldots, n\}$:

$$d_{ij} \geq 0; \quad d_{ii} = 0; \quad d_{ij} = d_{ji}\,.$$

The HAC algorithm will be applicable to such a dissimilarity matrix D if D is Euclidean. Formally, D is Euclidean if there exists an Euclidean space $(E, \langle\cdot,\cdot\rangle)$ and n points $\{x_1, ..., x_n\} \subset E$ such that $d_{ij} = \|x_i - x_j\|$ for all $i, j \in \{1, ..., n\}$, with $\|\cdot\|$ the norm induced by the inner product, $\langle ., .\rangle$, on E. Under this assumption, the dissimilarity case is a simple extension of the original $\mathbb{R}^p$ framework described in Section 2.2. Different versions of necessary and sufficient conditions for which an observed dissimilarity matrix is Euclidean have been obtained in Schoenberg (1935); Young et Householder (1938); Krislock et Wolkowicz (2012).

When such conditions do not hold, D is simply called a dissimilarity dataset, which is a particular case of proximity or relational datasets. Schleif et Tino (2015) have proposed a typology of such datasets and described different approaches that can be used to extend statistical or learning methods defined for Euclidean data to such proximity data. In brief, the first main strategy consists in finding a way to turn a non-Euclidean dissimilarity into an Euclidean distance, that is the closest (in some sense) to the original dissimilarity. This can be performed using eigenvalue corrections (Chen et collab., 2009), embedding strategies (like multidimensional scaling, Kruskal (1964)) or solving a maximum alignment problem (Chen et Ye, 2008), for instance.

A general construction. Alternatively, by using an analogy between distance and dissimilarity, HAC can be directly extended to non-Euclidean data as in Chavent et collab. (2018). This extension stems from the fact that, in the Euclidean case of Section 2.2, the inertia of a cluster may be expressed only in function of sums of the entries of the pairwise distances $(\|x_i - x_j\|_{\mathbb{R}^p}, 1 \leq i, j \leq n)$:

$$I(G) = \frac{\Delta(G, G)}{2|G|}, \tag{2.5}$$

where Δ is defined by $\Delta(G_u, G_v) = \sum_{x_i \in G_u, x_j \in G_v} \|x_i - x_j\|^2_{\mathbb{R}^p}$ for any clusters G_u and G_v. As a consequence of (2.5), Ward's linkage between any two clusters G_u and G_v may itself be written in function of these pairwise distances, see, e.g., Murtagh et Legendre (2014, p. 279) :

$$\delta(G_u, G_v) = \frac{|G_u||G_v|}{|G_u| + |G_v|} \left(\frac{\Delta(G_u, G_v)}{|G_u||G_v|} - \frac{\Delta(G_u, G_u)}{2|G_u|^2} - \frac{\Delta(G_v, G_v)}{2|G_v|^2} \right). \tag{2.6}$$

Therefore, as proposed by Chavent et collab. (2018), an elegant way to extend Ward's HAC to dissimilarity data is to *define* the inertia of a cluster using (2.5), with (sums of) distances replaced by (sums of) dissimilarities, that is :

$$\tilde{I}(G) = \frac{\tilde{\Delta}(G, G)}{2|G|}, \tag{2.7}$$

where

$$\tilde{\Delta}(G_u, G_v) = \sum_{x_i \in G_u, x_j \in G_v} d^2_{ij}. \tag{2.8}$$

The corresponding HAC is then formally obtained as the output of Algorithm 1, as described in Section 2.2.1. In particular, Ward's linkage is still given by (2.6), with Δ formally replaced by $\tilde{\Delta}$, and, as a consequence, the Lance-Williams update formula is also still given by (2.4). When the elements of Ω do belong to an Euclidean space and the dissimilarities are the pairwise Euclidean distances $\|x_i - x_j\|_{\mathbb{R}^p}$, these two definitions of HAC coincide. Otherwise, HAC is still formally defined, and the linkage can still be seen as a measure of heterogeneity, but the interpretation of the inertia of a cluster as an average squared distance to the center of gravity of the cluster (as in Equation (2.1)) is lost. Since the two definitions, I and $\tilde{I}$ coincide for the Euclidean case, we will only use the notation I in the sequel for the sake of simplicity, even when the data are non-Euclidean dissimilarity data.

The above approach based on pairwise dissimilarities and pseudo-intertia may be used to recover generalizations of Ward-based HAC to non-Euclidean distances

already proposed in the literature. In particular, the Ward HAC algorithm associated to $d_{ij} = \|x_i - x_j\|_{\mathbb{R}^p}^{\alpha/2}$ for $0 < \alpha \leq 2$ and $d_{ij} = \|x_i - x_j\|_{1,\mathbb{R}^p}$ (the latter is also called the Manhattan distance) correspond to the methods proposed by Székely et Rizzo (2005) and Strauss et von Maltitz (2017), respectively.

Remark 1. *Székely et Rizzo (2005) and Strauss et von Maltitz (2017) take a different point of view : they* define *the linkage between two clusters by (2.6) (up to a scaling factor 1/2) ; their generalized HAC is then the HAC associated to this linkage. Then, they* prove *that the Lance-Williams Equation (2.4) is still valid for this linkage. We favor the above construction by Chavent et collab. (2018), which is simply based on* pairwise *dissimilarities, as it is more intrinsic. It provides a justification for the linkage formula, and the Lance-Williams formula is automatically valid with no proof required.*

Finally, there is an ambiguity in the definition of the pseudo-inertia as an extension of the Ward's case. If most authors consider that the dissimilarity is associated to a distance and therefore define the pseudo-inertia based on the squared values d_{ij}^2, some authors (as Strauss et von Maltitz (2017)) define a linkage equal to the one that would have been obtained with Ward's linkage and a pseudo-inertia described as $\frac{1}{2|G|}\sum_{x_i,x_j\in G} d_{ij}$. This ambiguity has long been enforced by popular implemented versions of the algorithm, as it was the case in the R function `hclust` before Murtagh et Legendre (2014) raised and corrected this problem.

2.3.2 Extension to kernel data

In some cases, proximity relations between objects are described by their resemblance instead of their dissimilarity. We start with the case when the data are described by a kernel matrix. A kernel matrix is a symmetric positive-definite matrix $K = (k_{ij})_{1\leq i,j\leq n}$ whose entry k_{ij} corresponds to a measure of resemblance between x_i and x_j. Here, contrary to the Euclidean setting, no specific structure is assumed for Ω, which can be an arbitrary set.

Aronszajn (1950) has proved that there exists a unique Hilbert space $\mathcal{H}$ equipped with the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and a unique map $\phi : \Omega \longrightarrow \mathcal{H}$, such that $k_{ij} = \langle\phi(x_i), \phi(x_j)\rangle_{\mathcal{H}}$. This allows to consider the associated distance in $\mathcal{H}$ between any two elements $\phi(x_i)$ and $\phi(x_j)$ for $x_i, x_j \in \Omega$, that implicitly defines a Euclidean distance in Ω by :

$$d_{ij} = d(x_i, x_j) := \|\phi(x_i) - \phi(x_j)\|_{\mathcal{H}}\,,$$

so that

$$d_{ij}^2 = k_{ii} + k_{jj} - 2k_{ij}\,. \tag{2.9}$$

Therefore, it is possible to use Algorithm 1 for kernel data, even when $\mathcal{H}$ is not known explicitly and/or when it is not finite-dimensional. This is an instance of the so-called "kernel trick" (Schölkopf et Smola, 2002). The associated Ward's linkage can itself be re-written directly using sums of elements of the kernel matrix, as shown, e.g., in Dehman (2015) :

$$\delta(G_u, G_v) = \frac{|G_u||G_v|}{|G_u| + |G_v|} \left(\frac{R_{G_u,G_u}}{|G_u|^2} + \frac{R_{G_v,G_v}}{|G_v|^2} - 2\frac{R_{G_u,G_v}}{|G_u||G_v|} \right) , \tag{2.10}$$

where $R_{G_u,G_v} = \sum_{(x_i,x_j)\in G_u\times G_v} k_{ij}$.

Contrary to the dissimilarity case described in Section 2.3.1, the kernel case is a truly interpretable generalization of Ward's original approach because Ward's linkage as calculated in (2.10) is the variation of within-cluster inertia in the associated Hilbert space $\mathcal{H}$. This case has been described previously in Qin et collab. (2003); Ah-Pine et Wang (2016), for instance.

2.3.3 Extension to similarity data

Similarity data also aim at describing pairwise resemblance relations between the objects of Ω through a matrix of similarity (or proximity) measures $S = (s_{ij})_{1\leq i,j\leq n}$. Even though the precise definition of a similarity matrix can differ within the literature (see *e.g.*, Hartigan (1967)), it is generally far less constrained than kernel matrices. In most cases, the only conditions required to define a similarity is the symmetry of the matrix S [2] and the positivity of its diagonal. Since both similarities and kernels describe resemblance relations, it seems natural to try to extend the background of Section 2.3.2 to similarity datasets by using Equation (2.10). This allows the definition of a linkage, δ_S, between clusters via sums of elements of S. However, this heuristic is not well justified since the quantity $s_{ii} + s_{jj} - 2s_{ij}$ is not necessarily non-negative when S is not a positive definite kernel. Thus, it can not be associated to a squared distance as in Equation (2.9).

The previous work of Miyamoto et collab. (2015) has explicitly linked similarity and kernel data in HAC results. More precisely, for any given similarity S, the matrix $S^\lambda = (s^\lambda_{ij})_{1\leq i,j\leq n}$ such that $s^\lambda_{ij} := s_{ij} + \mathbf{1}_{\{i=j\}}\lambda$ is definite positive for any λ larger than the absolute value of the smallest eigenvalue of S. Therefore, the kernel matrix S^λ induces a well-defined linkage δ_{S^λ} via Equation (2.10), which is linked to δ_S by :

$$\delta_{S^\lambda}(G_u, G_v) = \delta_S(G_u, G_v) + \lambda\,.$$

This proposition justifies the extension of Equation (2.10) to similarity data with $R_{G_u,G_v} = \sum_{(x_i,x_j)\in G_u\times G_v} s_{ij}$. Using this heuristic is indeed equivalent to using a

2. In some cases, similarity measures are also supposed to take non-negative values, but we will not make this assumption in the present article.

given kernel matrix obtained by translating the diagonal of the original similarity S : doing so, the clustering is unchanged and the linkage values are all translated from $+\lambda$ for the kernel matrix, which does not even change the global shape of the clustering representation when the heights in this representation are the values of the linkage (as discussed in Section 2.4). The invariance property to this type of correction is specific to Ward's linkage. Therefore, the choice of Ward's linkage is the only choice that provides a natural interpretation of similarity matrices as dot product matrices and that makes a direct link between general similarities and the standard case of Euclidean distances. However, as for general dissimilarity data in Section 2.3.1, the interpretation of the linkage as a variation of within-cluster inertia is lost.

Conclusion. In conclusion to this section, we are finally left with only two cases : the Euclidean case (in which objects are embedded in a direct or indirect manner in a Euclidean framework) and the non-Euclidean case. The first case includes the standard case, the case of Euclidean distance matrices and the case of kernels while the latter case includes general dissimilarity and similarity matrices. In the Euclidean case, the original description of the Ward's algorithm is valid as such while, in the second, the algorithm can still be formally applied in a very similar manner at the cost of a loss of the interpretability of the criterion.

2.4 Interpretability of dendrograms

2.4.1 Dendrograms

The results of HAC algorithms are usually displayed as dendrograms. A dendrogram is a binary tree in which each node corresponds to a cluster, and, in particular, the leaves are the original objects to be clustered. The edges connect the two clusters (nodes) merged at a given step of the algorithm. The height of the leaves is generally supposed to be $h_0 = 0$. In the case of OCHAC, these leaves are displayed as indicated by the natural ordering of the objects, while in the general case of unconstrained HAC they are ordered by a permutation of the class labels that ensures that the successive mergers are neighbors in the dendrogram. The height of the node corresponding to the cluster created at merger t, h_t, is often the value of the linkage. To distinguish the height of the dendrogram from the value of the linkage, we will denote by m_t the value of the linkage at step t. Alternative choices for the values of $(h_t)_t$ are discussed in Section 2.4.4.

Dendrograms are used to obtain clusterings by horizontal cuts of the tree structure at a chosen height. A desirable property of a dendrogram is thus that the

clusterings induced by such a cut corresponds to those defined by the HAC algorithm. This property is equivalent to the fact that the sequence of heights is non-decreasing. When this *monotonicity* property is not satisfied, a merging step t for which $h_t < h_{t-1}$, is called a *reversal*. Reversals can be of two types, depending on whether or not they correspond to a visible *crossover* between branches of the dendrogram. Mathematically, a crossover corresponds to the case when the height of a given merger $G_{v_1} \cup G_{v_2}$ is less than the height of G_{v_1} or the height of G_{v_2}. A toy example of reversal with crossover is shown in Figure 2.1, between nodes merged at steps 1 and 2, for the result of OCHAC.

The goal of this section is to study which settings and which definitions of height guarantee the absence of reversals – with and without crossovers.

2.4.2 Monotonicity, crossovers and ultrametricity

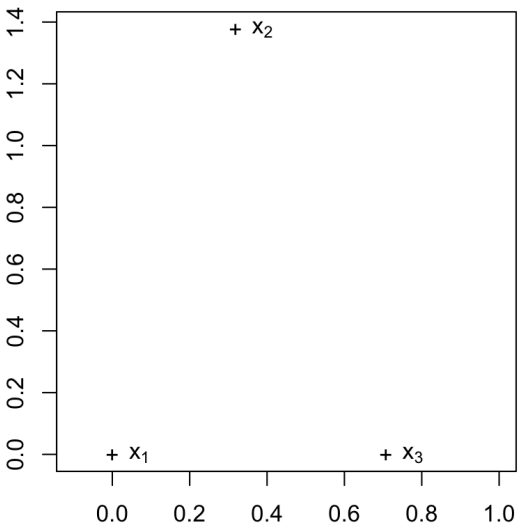
A crossover in a dendrogram automatically implies the non-monotonicity of the sequence of heights. The converse is true when the height of the dendrogram corresponds to the value of the linkage (or to a non-decreasing function of the linkage) for the corresponding merger, by virtue of Proposition 1 below.

Proposition 1. *Consider a dendrogram whose sequence of heights $(h_t)_t$ is a non-decreasing transformation of the linkage values $(m_t)_t$. Then the only reversals that can occur are crossovers.*

The proof of Proposition 1 is not specific to Ward's linkage and is a simple consequence of the fact that the linkage is the objective function of the clustering :

Proof of Proposition 1. Consider an arbitrary merger step of the HAC, characterized by the linkage value m_t. If the next merger does not involve the newly created cluster, then this merger was already a candidate at step t. Then, by optimality of the linkage value at step t, this merger can not be a reversal. Therefore, any reversal must involve the newly created cluster, and is thus a crossover. □

An important consequence of Proposition 1 is that when the height of the dendrogram is the corresponding linkage, the absence of crossovers is *equivalent* to the monotonicity of the sequence of heights.



$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0.45\sqrt{0.5} & \sqrt{1.89875} \\ \sqrt{0.5} & 0 \end{pmatrix}$$

$$D = \begin{pmatrix} 0 & \sqrt{2} & \sqrt{0.5} \\ \sqrt{2} & 0 & \sqrt{2.05} \\ \sqrt{0.5} & \sqrt{2.05} & 0 \end{pmatrix}$$

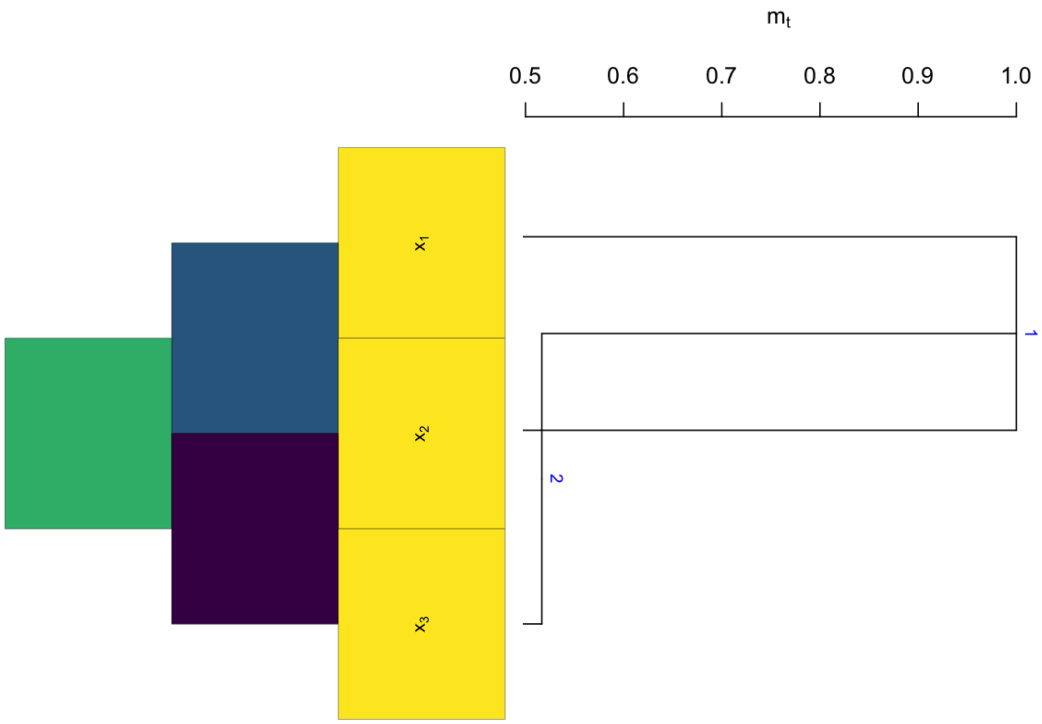


FIGURE 2.1: **A crossover for Euclidean OCHAC with height defined as the linkage m_t.** Top left : Configuration of the objects in $\mathbb{R}^2$. Top right : Coordinates of the objects and Euclidean distance matrix corresponding to this configuration. Bottom left : Representation of the values of the Euclidean distance (dark colors correspond to larger values, so to distant objects). Bottom right : Dendrogram obtained from OCHAC (the ordering is indicated by the indices of objects) and with the height corresponding to Ward's linkage.

We shall see in Section 2.4.4 that for an arbitrary height, the absence of crossover in the dendrogram is not necessarily equivalent to the monotonicity of the sequence of heights. The absence of crossover can be characterized by a mathematical property of the cophenetic distance associated to the heights of the dendrogram, called *ultrametricity* (see *e.g.*, Rammal et collab. (1986)). Formally, let us define, for all $i, j \in \{1, \ldots, n\}$, the *cophenetic distance* h_{ij} between i and j as the value of the height h_{t^*} such that t^* is the first step (or the smallest merge number) such that the i-th and j-th objects are in the same cluster. h is said to satisfy the ultrametric inequality if :

$$\forall i, j, k \in \{1, ..., n\}, \qquad h_{ij} \leq \max\{h_{ik}, h_{kj}\}.$$

As announced, this property is key to ensure the monotonicity of the sequence of heights. More precisely, Johnson (1967) has defined an explicit bijection between a hierarchy of clusterings with an associated sequence of non-decreasing "heights" (called "values" in the article) and matrix of values with a diagonal equal to zero and satisfying the ultrametric inequality. It turns out that this bijection explicitly defines the entries of the ultrametric matrix as the cophenetic distance of the dendrogram whose heights are the one of the associated hierarchy of clusterings. In other words, this means that a given sequence of heights defining a dendrogram is non-decreasing if and only if the cophenetic distance associated to this dendrogram (or equivalently to this sequence of heights) satisfies the ultrametric inequality.

2.4.3 Monotonicity of Ward's linkage

Ward's linkage corresponds to the variation of within-cluster inertia, so that the monotonicity of the linkage is ensured for Ward's standard HAC algorithm with Euclidean data. More generally, Batagelj (1981) gives necessary and sufficient conditions based only on the Lance-Williams coefficients that ensures monotonicity for a given linkage. These results apply to the extensions of HAC to non-Euclidean datasets and show that the monotonicity of the linkage values is always ensured for standard HAC with Ward's linkage. In addition, Ferligoj et Batagelj (1982) give necessary and sufficient conditions on the Lance-Williams coefficients to ensure the monotonicity of the linkage values in constrained HAC, for an arbitrary symmetric relational constraint. These conditions are not fulfilled for Ward's linkage. Therefore, monotonicity is not guaranteed for CCHAC with Ward's linkage, as also noted by Grimm (1987) for the specific case of OCHAC. It can be shown that even for Euclidean data, the contiguity constraint can induce non increasing linkage values for some steps of the algorithm, as illustrated by Figure 2.1.

More precisely, if we consider OCHAC, the following proposition establishes necessary and sufficient conditions on a dissimilarity d to observe a reversal at a given step of OCHAC when the height is defined by Ward's linkage :

Proposition 2. *Suppose that* $\Omega = \{x_i\}_{i=1,\dots,n}$ *is equipped with the symmetric contiguity relation* $x_i \mathcal{R} x_j \Leftrightarrow |i-j| = 1$ *(OCHAC). Denote by* l *and* r *the indices of the left and right clusters merged at a given step* t*, and by* $\bar{l}$ *and* $\bar{r}$ *their own left and right cluster, respectively. Then there is a reversal at step* $t+1$ *for the height defined by the linkage if and only if :*

$$\delta(G_l, G_r) \geq \min\left(\frac{g_{\bar{l}l}\delta(G_{\bar{l}}, G_l) + g_{\bar{l}r}\delta(G_{\bar{l}}, G_r)}{g_{\bar{l}l} + g_{\bar{l}r}}, \frac{g_{l\bar{r}}\delta(G_l, G_{\bar{r}}) + g_{r\bar{r}}\delta(G_r, G_{\bar{r}})}{g_{l\bar{r}} + g_{r\bar{r}}}\right) \tag{2.11}$$

where we have used the notation $g_{uv} := |G_u \cup G_v| = |G_u| + |G_v|$.

The fact that Condition (2.11) involves clusters contiguous to the last merger is a consequence of Proposition 1. The formulation of Condition (2.11) is quite intuitive : crossovers correspond to situations in which the Ward linkage between two newly merged clusters is larger than a (weighted) average Ward linkage between each of these two clusters and one of the contiguous clusters. The proof of Proposition 2 is given in Appendix A.1.

Let us apply Proposition 2 to the specific case of the first and second mergers in the algorithm. Assuming that the optimal merger at step 1 is between the l-th and r-th objects, and recalling that the Ward linkage between two singletons is simply $\delta(\{u\}, \{v\}) = d^2_{uv}/2$, Condition (2.11) reduces to :

$$2d^2_{l,r} > \min\left(d^2_{\bar{l},l} + d^2_{\bar{l},r}, d^2_{r,\bar{r}} + d^2_{l,\bar{r}}\right)$$

In particular, given the $p-1$ distances $(d^2_{i,i+1})_{1\leq i\leq p-1}$ that determine the first step of the OCHAC algorithm, it is always possible to find an adversarial dissimilarity yielding a reversal at the second step, *e.g.*, by choosing $d_{l,\bar{r}}$ such that $d^2_{l,\bar{r}} < 2d^2_{l,r} - d^2_{r,\bar{r}}$. This is the case in the counter-example of Figure 2.1.

An example of relevant reversal for OCHAC. Because of the possible presence of crossovers in OCHAC even in a simple Euclidean setting, CCHAC may appear as a deteriorated version of standard HAC, where the optimal merger is chosen within a reduced set of possible mergers compared to the unconstrained version. One may then expect that the total within-cluster inertia at a given step of the algorithm is larger than for the unconstrained version that chooses the "optimal" merger at this step (that is, the merger with the smallest increase of the total within-inertia). In addition, the algorithm does not necessarily exhibit a clear and understandable monotonic evolution of the objective criterion, $(m_t)_t$. However, it can be shown, even in a very simple example, that OCHAC can lead to better solutions in terms of within-cluser intertia, when the constraint is consistent to

the spatial structure of the data. This fact is illustrated in Figure 2.2 [3]. In this example, 7 data points are displayed in $\mathbb{R}^2$ with an order constraint illustrated by a line linking two points allowed to be merged. In this situation, $(m_t)_t$ is indeed non monotonic for OCHAC (bottom left figure) but leads to a better total within-cluster inertia for $k = 3$ clusters (vertical green line), which is also more relevant for the data configuration (top figures). This is a typical case where the constraint forces the algorithm to explore under-efficient configurations but that can be aggregated into a better solution, contrary to the unconstrained algorithm. This is explained by the fact that even the unconstrained algorithm is greedy, by construction, and thus not optimal compared to an exhaustive search of the best partition in k classes.

2.4.4 Monotonicity of alternative heights

Since reversals can occur in CCHAC dendrograms with Ward's linkage, alternative definitions of the height have been proposed to improve the interpretability of the result in this case. They are defined as quantities related to the heterogeneity of the partition. In this section, we study the monotonicity of such alternative heights.

Grimm (1987) presents three alternative heights to the standard *variation of within-cluster inertia* (m_t) :

— the *within-cluster (pseudo-)inertia* (or *Error Sum of Squares*) that corresponds to the value of the objective function. In this case, the height at step t is given by :

$$\mathrm{ESS}_t = \sum_{u=1}^{n-t} I(G_u^{t+1}),$$

where $\mathcal{P}^{t+1} = \{G_u^{t+1}\}_{u=1,\dots,n-t}$ is the partition obtained at step t of the algorithm. This alternative height is very natural (and the one implemented in the R package **rioja** for OCHAC) since it corresponds to the criterion whose minimization is approximated by HAC (and OCHAC) in a greedy way ;

— the *(pseudo-)inertia of the current merger*, which is defined as :

$$I_t = I(G_u^t \cup G_v^t)$$

where G_u^t and G_v^t are the two clusters merged at step t. Grimm (1987) remarks that this measure is very sensitive to the cluster size $|G_u^t| + |G_v^t|$.

3. The detailed analysis of all examples and counter-examples of this section is provided in Appendix A.2.

— the *average (pseudo-)inertia of the current merger*, that has been designed so as to avoid the bias related to the cluster size in I_t. It is defined as :

$$\overline{I}_t = \frac{I_t}{|G_u^t| + |G_v^t|}$$

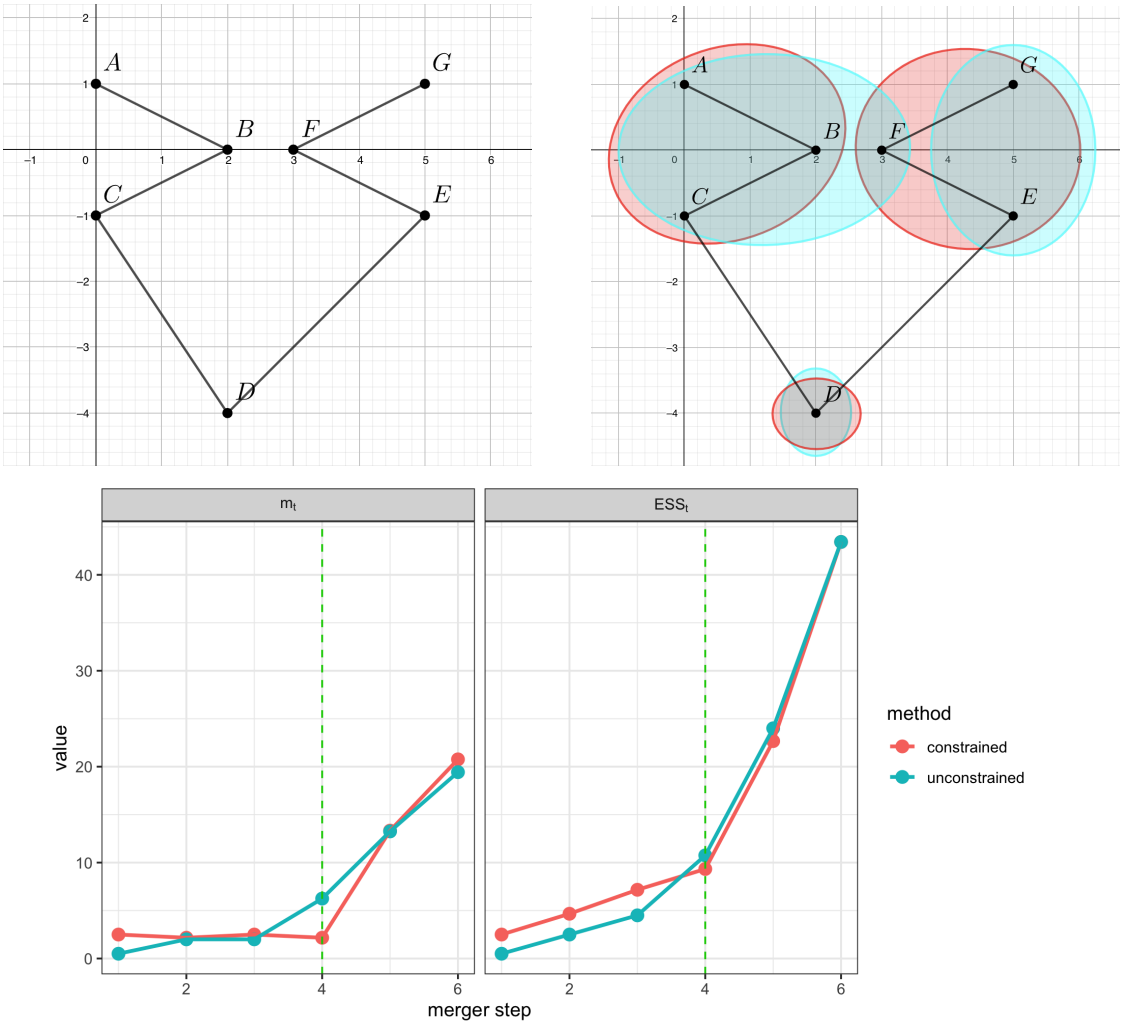


FIGURE 2.2: **Simple configuration in which OCHAC outperforms standard HAC.** Top left : Initial configuration with the order constraint represented by straight lines. Top right : Clustering with 3 clusters as produced by OCHAC (red) and standard HAC (blue). Bottom : Evolution of $(m_t)_t$ and of the total within-cluster inertia (also called, Error Sum of Squares : $(\text{ESS}_t)_t$) along the clustering processes, the green line correspond to the 3 components clustering.

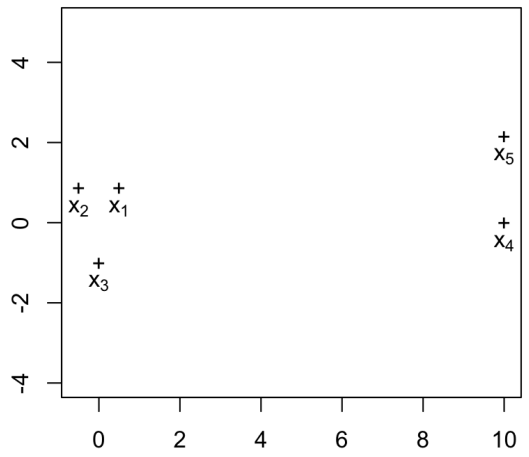
Standard HAC : Known properties of alternative heights. Note that $\text{ESS}_t = \sum_{t' \leq t} m_{t'}$. As explained in Section 2.4, $(m_t)_t$ is monotonic for standard HAC, both for Euclidean and non-Euclidean data. Since $m_0 = 0$ by definition, this ensures the monotonicity of $(\text{ESS}_t)_t$, for Euclidean and non-Euclidean data in the case of standard HAC.

On the contrary, I_t and $\overline{I}_t$ may induce reversals even for standard HAC and Euclidean data. More importantly, contrary to the case when the height of the dendrogram is m_t, even when the ultrametric property is satisfied, the monotonicity is not ensured for these criteria. This is illustrated in Figure 2.3 (and in Figure A.1 of the Appendix A.3), for I_t (and for $\bar{I}_t$, respectively) and data in $\mathbb{R}^2$.

In this case, the dendrogram has a conventional look but the mergers are not ordered by increasing heights. For instance, in Figure 2.3, the cluster merged at step 2 is above the one at step 3. Hence, cutting the dendrogram at height $h = 2.5$ leads to a clustering into $\{x_1, x_2\}, \{x_3\}, \{x_4, x_5\}$, but this clustering does not belong to the sequence of clusterings induced by the HAC (where the clustering in 3 clusters is the one obtained after the second merger, that is, $\{x_1, x_2, x_3\}, \{x_4\}, \{x_5\}$).

CCHAC : Known properties of alternative heights. Figures 2.3 and A.1 (the latter in Appendix A.3) provide counter-examples for the monotonicity of $(I_t)_t$ and $(\bar{I}_t)_t$ in the Euclidean case for HAC. If the objects are pre-ordered as the nodes in these figures, then OCHAC and standard HAC give identical hierarchical clusterings. Therefore, these examples also provide counter-examples for the monotonicity of $(I_t)_t$ and $(\bar{I}_t)_t$ in the Euclidean case for OCHAC, and show that there is no guarantee for monotonicity in the case of general CCHAC. The fact that $(\bar{I}_t)_t$ is not necessarily monotonous for OCHAC has already been mentioned by Grimm (1987).

CCHAC : Within-cluster pseudo-inertia for dissimilarity data. The only unanswered case is whether $(\text{ESS}_t)_t$ is monotonic or not for CCHAC and non-Euclidean data. We provide a counter-example that proves that the monotonicity is not ensured in this case : Figure 2.4 shows that the dendrogram obtained from OCHAC on a given non-Euclidean dissimilarity D contains a crossover ($m_4 < m_3$). In particular, the associated sequence of heights is not monotonic. However, Proposition 1 ensures that $(\text{ESS}_t)_t$ has the nice property that the absence of crossovers is equivalent to its monotonicity. Indeed, as $(\text{ESS}_t)_t$ corresponds to the cumulative sums of the linkage $(m_t)_t$, the mapping between m_t and ESS_t is equal to the addition of ESS_{t-1}. As, by definition, ESS_{t-1} is, as any $I(G_u^{t-1})$, positive, this ensures that this mapping is non-decreasing.



$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ 0 & -1 \\ 10 & 0 \\ 10 & 2.16 \end{pmatrix}$$

$$D \approx \begin{pmatrix} 0.00 & 1.00 & 1.93 & 9.54 & 9.59 \\ 1.00 & 0.00 & 1.93 & 10.54 & 10.58 \\ 1.93 & 1.93 & 0.00 & 10.05 & 10.49 \\ 9.54 & 10.54 & 10.05 & 0.00 & 2.16 \\ 9.59 & 10.58 & 10.49 & 2.16 & 0.00 \end{pmatrix}$$

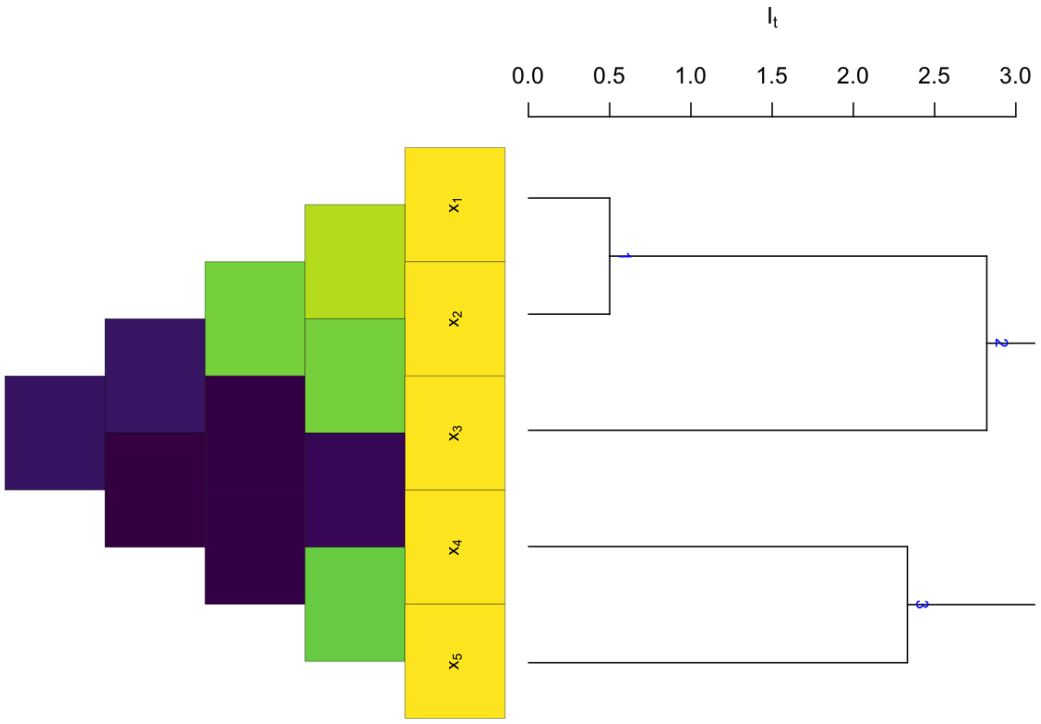


FIGURE 2.3: **A reversal for Euclidean standard HAC with height defined as I_t.** Top left : Configuration of the objects in $\mathbb{R}^2$. Top right : Coordinates of the objects and Euclidean distance matrix corresponding to this configuration. Bottom left : Representation of the values of the dissimilarity (dark colors correspond to larger values, so to distant objects). Bottom right : dendrogram obtained from standard HAC. Only the first 3 merges of the dendrogram is represented to ensure a comprehensive view of the sequence of heights.

$$D = \begin{pmatrix} 0 & \sqrt{1.99} & \sqrt{1.99} & \sqrt{1.99} & 0.1 & 1 \\ \sqrt{1.99} & 0 & \sqrt{2} & \sqrt{1.99} & 0.1 & 1 \\ \sqrt{1.99} & \sqrt{2} & 0 & \sqrt{2} & 0.1 & 1 \\ \sqrt{1.99} & \sqrt{1.99} & \sqrt{2} & 0 & \sqrt{2} & 1 \\ 0.1 & 0.1 & 0.1 & \sqrt{2} & 0 & \sqrt{2} \\ 1 & 1 & 1 & 1 & \sqrt{2} & 0 \end{pmatrix}$$

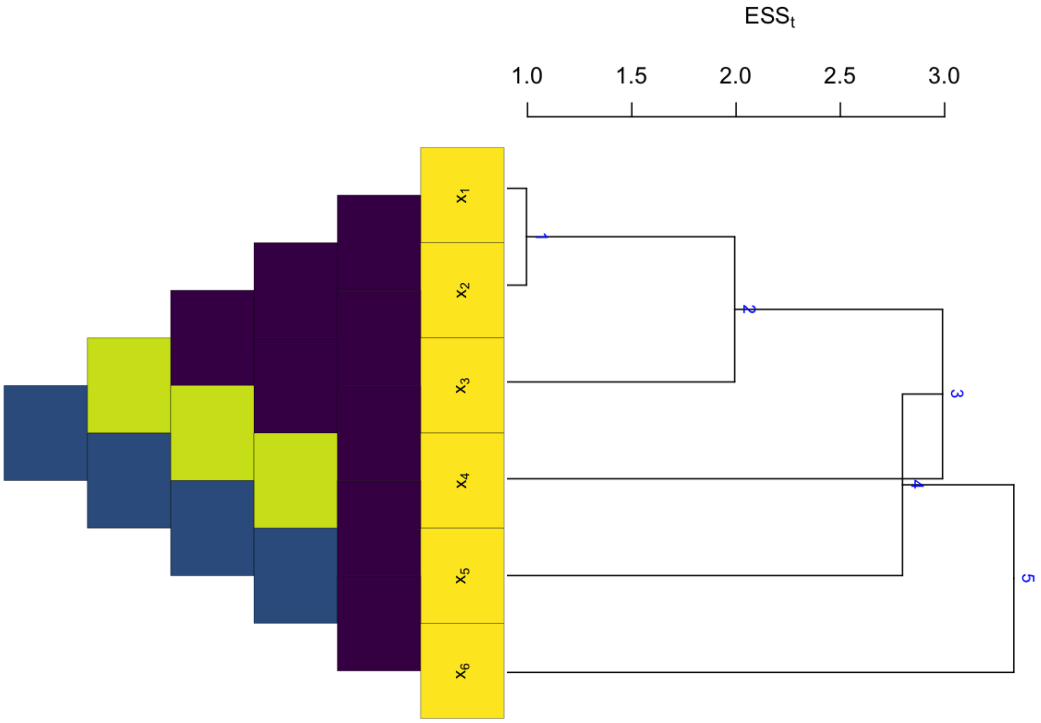


FIGURE 2.4: **A reversal for non-Euclidean OCHAC with height defined as ESS$_t$.** Top : Dissimilarity matrix. Bottom left : Representation of the values of the dissimilarity D (dark colors correspond to larger values, so to distant objects). Bottom right : Dendrogram obtained from OCHAC (the ordering is indicated by the indices of objects) and with the height corresponding to ESS$_t$.

Table 2.1 summarizes the properties of the different types of heights, respectively for standard HAC and CCHAC. Note that the monotonicity of ESS_t is a consequence of the positivity of m_t.

		m_t	ESS_t	I_t	$\overline{I}_t$
HAC	Euclidean	✓Ward (1963)	✓Ward (1963)	× [Fig. 2.3]	×[Fig. A.1]
	Non-Euclidean	✓Batagelj (1981)	✓Batagelj (1981)	× [Fig. 2.3]	×[Fig. A.1]
CCHAC	Euclidiean	×Grimm (1987)	✓Grimm (1987)	× [Fig. 2.3]	×Grimm (1987)
	Non-Euclidean	×Grimm (1987)	× [Fig. 2.4]	× [Fig. 2.3]	×Grimm (1987)

TABLE 2.1: Monotonicity of heights for standard HAC (top) and CCHAC (bottom).

2.5 Simulation

HAC can be seen as a greedy algorithm to solve the problem of finding the partition with minimal within-cluster inertia ESS_t of n objects into $n-t$ classes, for each $t = 1 \dots n-1$. It may be expected that the inertia of the partitions will be lower for HAC than OCHAC, since the possible mergers in OCHAC are chosen among a subset of the possible mergers in HAC. Can we quantify the impact of the order constraint on the quality of the partitions (as measured by ESS) obtained for HAC and OCHAC, depending on the strength of the actual order structure in the data ? In this section, we address this question by analyzing Hi-C data (Dixon et collab., 2012), which present a strong order structure, as illustrated by Figure 2.5. We use a perturbation process to progressively break the consistency between the data structure and the constraint imposed in OCHAC.

2.5.1 Data and method

Hi-C studies aim at characterizing proximity relationships in the 3D structure of a genome, by measuring the frequency of physical interaction between pairs of genomic locations via sequencing experiments. Formally, a Hi-C map is a symmetric matrix $S = (s_{ij})_{i,j}$ in which each entry s_{ij} is equal to the frequency of interaction between genomic loci i and j. Here, a locus is a fixed-size interval of genomic positions, also called a "bin". Hi-C maps are classically represented by the upper triangular part of the matrix, as shown in Figure 2.5. The matrix has a strong diagonal structure that reflects the linear order of DNA within chromosomes (loci that are close along the genome are more frequently interacting than distant loci). An important question in Hi-C studies is to identify Topologically

Associating Domains (TADs), which are self-interacting genomic regions appearing to be more compact than the rest of the genome. Indeed, TADs have been shown to play an important role in gene regulation (Dixon et collab., 2012). A number of TAD detection methods have been proposed (see *e.g.*, Zufferey et collab. (2018) for a review) and some are based on HAC or OCHAC (Fraser et collab., 2015; Haddad et collab., 2017; Ambroise et collab., 2019). This is both natural, since Hi-C maps can be seen as similarity matrices, and formally justified, as explained in Section 2.3.3. In practice, Hi-C maps are indeed non-positive, and as explained in Section 2.3.3, Ward's linkage is preferred in this situation since it is the only linkage that provides a natural interpretation of such matrices in terms of Euclidean dot products.

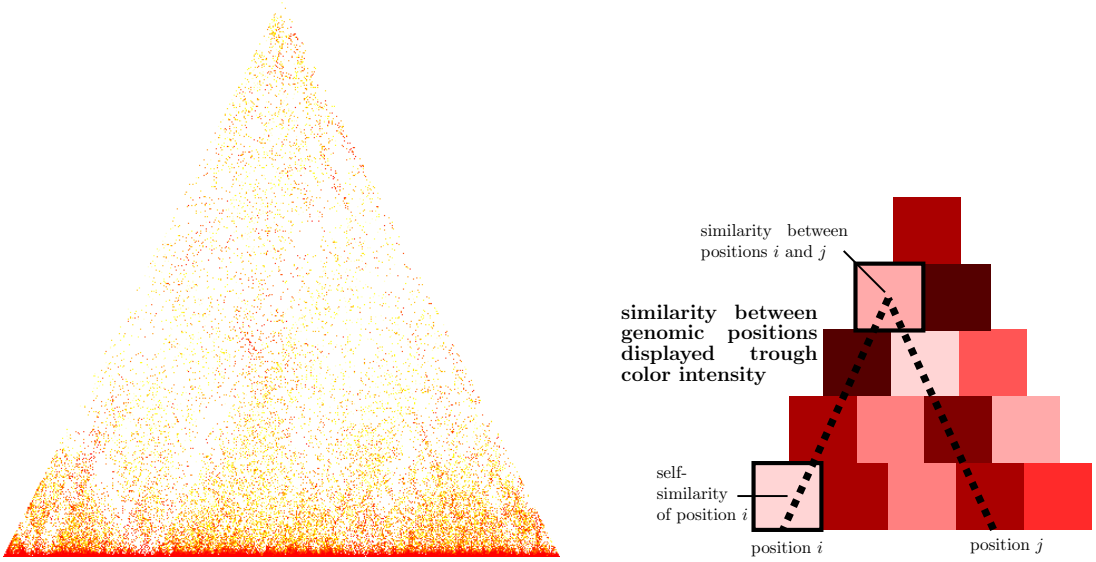


FIGURE 2.5: **Graphical representation of a Hi-C map**. Left : Classical representation of a Hi-C map as the upper half of an heatmap. Horizontal axis corresponds to the diagonal of the heatmap and horizontal position is defined by the indices of bins within a single chromosome. Intensities of the frequency of physical interaction between bins are represented by levels of red. Non-contiguous bin interactions corresponds to all interactions strictly above the horizontal axis (non-diagonal entries). Right : Schematic view of the graphical representation of a Hi-C map with detailed specific entries (self bin interactions and non-contiguous bin interactions).

The simulations in this section are based on a single chromosome (chromosome 3) from an experiment in human embryonic stem cells (hESC ; Dixon et collab.

(2012) [4]). The downloaded Hi-C matrix contains 4,864 bins. It has been obtained with a bin size of 40kb and normalized using ICE (Imakaev et collab., 2012). We further performed a log-transformation of the entries to reduce the distribution skewness prior clustering.

In order to assess the influence of the data structure on the quality of the partitions obtained by OCHAC and standard HAC algorithms, we have used a perturbation process to progressively remove the strong diagonal in the original Hi-C map. The perturbation consists in swapping two entries, s_{ij} and $s_{i'j'}$ of the matrix, in which (i, j) and (i', j') have been randomly sampled with uniform probability among the pairs $\{(u, v), (u', v')\}$ for which $u \leq v$, $u' \leq v'$ and $s_{uv} + s_{u'v'} > 0$, where the last condition avoids swapping entries that are both zero. The proportion of such swapped pairs, which we call perturbation level, varied from 0% up to 90% (Figure 2.6).

This process was repeated 50 times to allow assessing the variability. Since obtained matrices are not necessarily positive definite, we translate their diagonal by a small quantity that ensures the positivity of all $d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$ as described in Section 2.3.2.

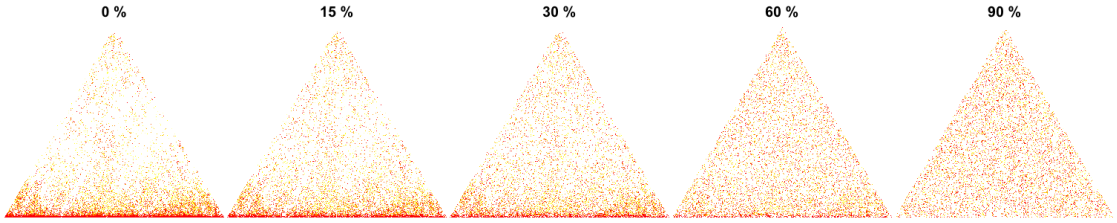


FIGURE 2.6: **Illustration of the perturbation process**. From left to right : example of Hi-C maps corresponding to increasing perturbation levels.

All simulations were performed with R. The results for standard HAC were computed with the function `hclust` (from the **stats** package) and those for OCHAC were computed with the function `adjClust` (from the **adjclust** package). Figures were obtained using **adjClust** or **ggplot2** (Wickham, 2016).

2.5.2 Comparison of standard HAC and OCHAC results

In this section, the results of standard HAC and OCHAC are compared through the corresponding height sequences of the dendrograms, through dendrograms themselves and through clusterings obtained by horizontal cuts of the dendrograms. While dendrograms and height sequences are a direct output of the HAC

4. The pre-processed and normalized data have been downloaded from the authors' website at `http://chromosome.sdsc.edu/mouse/hi-c/download.html` (raw sequence data are also published on the GEO website, accession number GSE35156).

process, clusterings are obtained using a model selection strategy. We have considered two such strategies : the broken stick (Bennett, 1996), as implemented in **adjclust**, and the slope heuristic (Arlot et collab., 2016), as implemented in **capushe**. The idea of the broken stick heuristic is to test the reduction of within-cluster inertia along the clusterings sequence considered backward (starting by the clustering consisting in the whole set of objects) against the reduction obtained for a model in which within-cluster inertia is divided with uniform probability in the corresponding number of components. On the other hand, the slope heuristic assumes the existence of a true clustering which is detected by a change in the slope of within-cluster inertia along the clustering sequence.

For both strategies, the "best" clustering is defined based on the within-cluster inertia of the sequence of clusterings obtained by the hierarchical process. As both strategies gave similar results, we chose to report only the results obtained for the broken stick heuristic here. For each Hi-C map of the simulation and for both methods of hierarchical clustering, clustering comparisons will be based on the clusterings selected by the broken stick heuristic.

Height sequences. Figure 2.7 shows the evolution of m_t (normalized by its maximal value among both methods at a given permutation level) and ESS_t (normalized by the total inertia of the set of bins) along the two clustering processes for increasing perturbation levels. For the original dataset, which presents an organization strongly consistent with the order constraint, the heights of standard HAC and OCHAC are very similar. However, interestingly, OCHAC improves the objective criteria (ESS_t and m_t) for low perturbation levels (15%-30%) across a wide range of merging levels.

More specifically, we compared the heights obtained for HAC and OCHAC at the merger number selected by the broken stick heuristic (Bennett (1996) ; vertical lines in Figure 2.7). At these numbers of clusters or in their close neighborhood, ESS_t is always smaller for OCHAC, which we interpret as more homogeneous clusterings for OCHAC than for HAC. The magnitude of the improvement achieved by OCHAC with respect to HAC depends on the perturbation level : for the original data, it is close to 5%, whereas it is much larger (25-30%) when the perturbation level is 15%-30%. It then decreases again ($< 20\%$) for larger perturbation levels (60%).

The fact that OCHAC can achieve lower values than HAC for ESS_t and m_t may be counter-intuitive, since –as explained at the beginning of Section 2.5– possible mergers in OCHAC are chosen among only a subset of the possible mergers in standard HAC. In fact, HAC itself is a heuristic for the minimization of ESS_t, because of its hierarchical agglomerative nature ; in contrast, the optimal clustering at step t in the sense of ESS_t may not necessary be obtained by merging two

clusters of the optimal clustering at step $t-1$. This result illustrates the robustness to noise of the constrained approach, which is very interesting in practice : in Hi-C experiments, for instance, many biases (genomic, experimental, etc.) are encountered. Thus, OCHAC has to be preferred in such contexts and will additionally result in a lower computational cost. The benefit of using a relevant constraint had already been observed by Steinley et Hubert (2008) : their simulations proved that a relevant order constraint (in their case, obtained from the data) could improve the recovery of the true cluster structure (although possibly at the cost of a slight decrease in ESS_t compared to the unconstrained version).

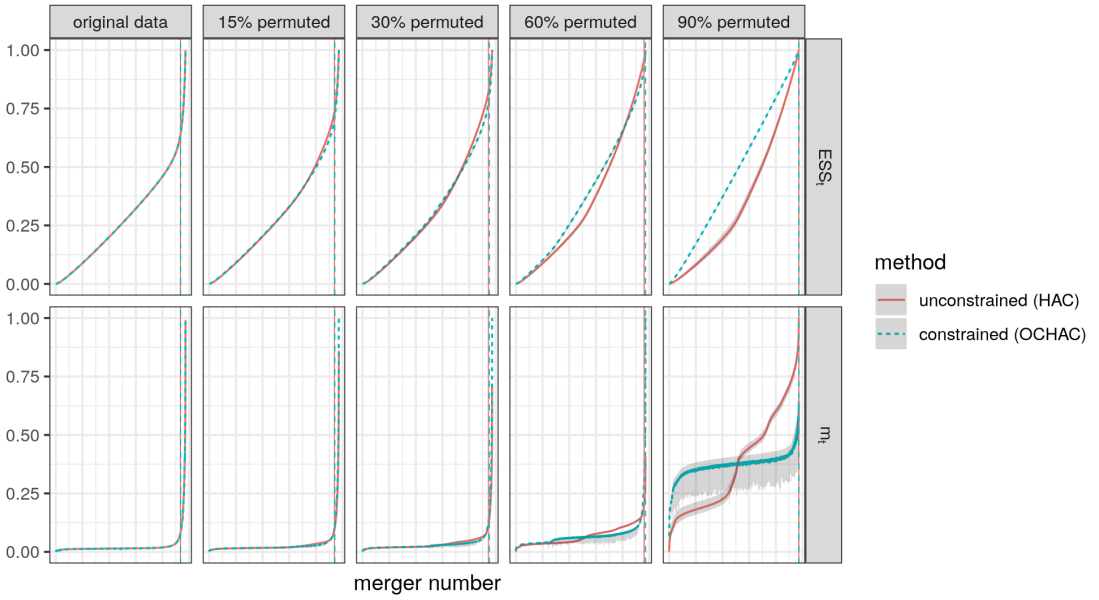


FIGURE 2.7: **Comparison of the height sequences for standard HAC (red, solid) and OCHAC (blue, dashed)** for ESS_t (top) and m_t (bottom) with increasing levels of perturbation of the original Hi-C matrix. The curves correspond to the average criteria over 50 simulations and the grey shadows correspond to the minimum and maximum of the criteria over 50 simulations. The vertical lines correspond to the average number of clusters chosen by the broken stick heuristic, respectively for standard HAC and OCHAC (red, solid and blue, dashed).

For perturbation levels larger than 60%, the data structure is no more compatible with the constraint (see Figure 2.6) and standard HAC seems to perform globally better than OCHAC, as expected. In addition, in this extreme situation, OCHAC exibits very large reversals for m_t (seen with the grey shadow in Figure 2.7), that are due to sudden breaks in the quality of the clusterings, induced by the constraint. The presence of such large reversals is a practical and visible

indication that the constraint is not relevant for the data and that OCHAC should not be used.

Dendrograms and clusterings. The same type of conclusion can be drawn when comparing not just the heights of the dendrograms but the dendrograms themselves or the clusterings induced by these dendrograms. Figure 2.8 shows the distribution of a measure of similarity between the order of fusion in the dendrogram. More precisely, the cophenetic distances have been computed for all pairs of objects in the dendrograms induced by standard HAC and OCHAC at different levels of perturbation and the Spearman correlation betwen these two vectors of cophenetic distances (coming from the constrained and the unconstrained version of the algorithm) has been obtained. As the perturbation level increases, the Spear-

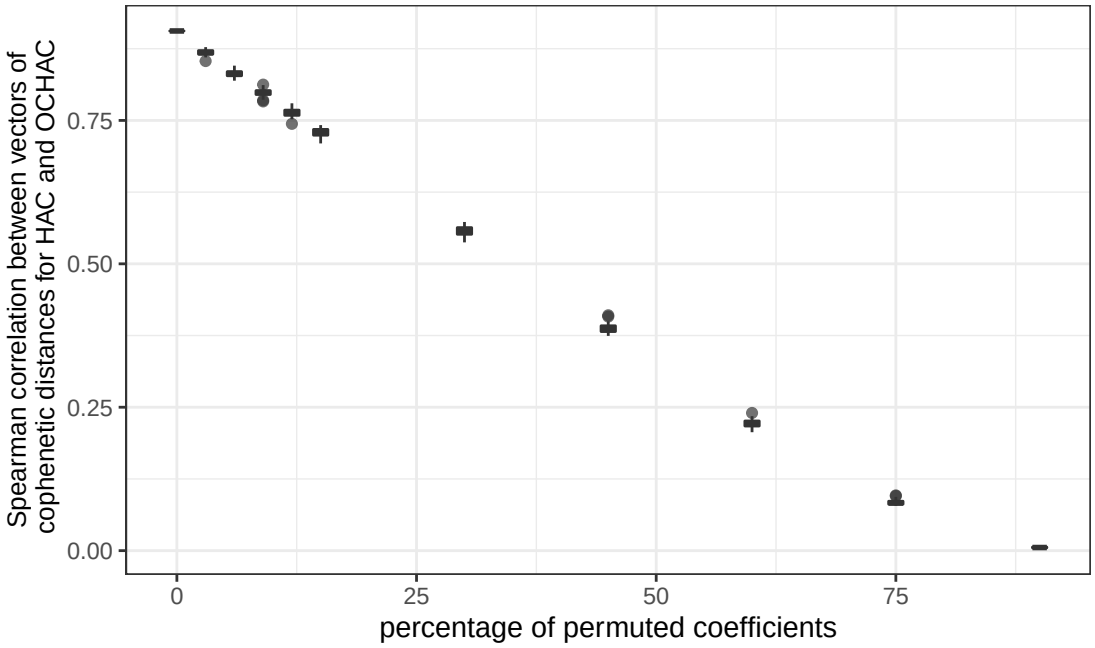


FIGURE 2.8: **Spearman correlation between vectors of cophenetic distances for HAC and OCHAC.**

man correlation linearly decreases from a value close to 1 (implying very similar dendrograms) to a value close to 0 (implying completely different dendrograms).

Finally, we compared the clusterings obtained by the broken stick heuristic (Bennett, 1996) as follows. For larger perturbation levels (more than 60%) of permuted coefficients, we obtained a trivial clustering with only one cluster, a strong indication that the cluster structure had disappeared at these levels. For lower perturbation levels, the obtained clusterings were compared using the Normalized

Mutual Information (NMI, Danon et collab. (2005)). As for the Spearman correlation, the NMI values obtained for the original data and low levels of perturbations (up to 30%) are very close to 1, which shows a strong similarity of the induced clusterings. As the perturbation level increases, the obtained partitions became more and more different, with NMI values below 0.6 (results not shown).

2.5.3 Reversals for the different heights

In this section, we investigate the reversals obtained for different heights and for standard HAC and OCHAC. Figure 2.9 gives the evolution of the percentage of reversals (relative to the total number of simulations, 50), for standard HAC and OCHAC and for the different types of heights, along the hierarchical clustering process.

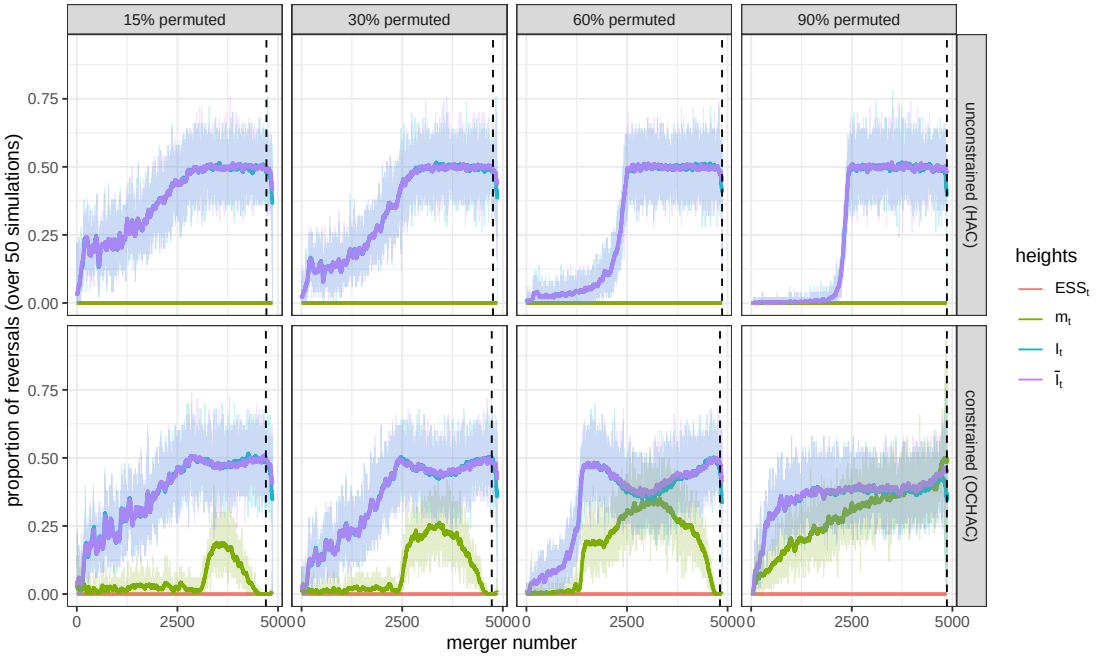


FIGURE 2.9: **Evolution of the number of reversals** for ESS_t, m_t, I_t and $\bar{I}_t$ for standard HAC (top) and OCHAC (bottom) for increasing levels of perturbation of the original Hi-C matrix. The background shadow is the actual value and the strong line is a smoothed value (box kernel, bandwith equal to 50). The dotted vertical line corresponds to the average number of clusters chosen by the broken stick heuristic.

As expected from Section 2.3 (Table 2.1), $(\mathrm{ESS}_t)_t$ does not have reversals and $(m_t)_t$ only has reversals for OCHAC. When the perturbation level increases, the

evolution of the number of reversals in $(I_t)_t$ and $(\bar{I}_t)_t$ is markedly different from that of $(m_t)_t$. For the smallest perturbation levels (up to 30%), the number of reversals of $(m_t)_t$ is close to 0, while it ranges from 10 to 50% for $(I_t)_t$ and $(\bar{I}_t)_t$. At these perturbation levels, $(m_t)_t$ almost never has a reversal at a merger number that corresponds to the number of clusters chosen by the broken stick heuristic : most reversals are concentrated at a merger number smaller than the merger chosen by the broken stick heuristic. Actually, for small perturbation levels, these reversals in m_t values help improve the quality of further clusterings by choosing a solution that is less efficient than that of standard HAC but more consistent with the data (as already discussed in the example of Figure 2.2). Hence, when the data structure is consistent with the constraint, $(m_t)_t$ typically provides an interpretable dendrogram. This nice property is, of course, lost when the constraint is no more consistent with the data structure (above a perturbation level of 60%), which is explained by the fact that the OCHAC has a poor performance in that context, as already discussed in the previous section.

On the contrary, $(I_t)_t$ and $(\bar{I}_t)_t$ exhibit larger numbers of reversals. This is particularly the case for the last mergers, even for small levels of perturbation and even in the unconstrained case : 40-60% of the simulations have reversals for both OCHAC and standard HAC at a number of clusters corresponding to the selected clustering. We also observe that the percentage of simulations showing a reversal for standard HAC tends to decrease when the perturbation level in the data increases for the first steps of the hierarchical process (the same can be observed, to a much lesser extent, for OCHAC). This phenomenon is explained below.

Figure 2.10 displays the evolution of the merged cluster size thorough the hierarchical clustering and provides an explanation for this fact. For standard HAC, the number of clusters with a size equal to 2 during the first steps of the algorithm is strongly increasing when the perturbation level increases. For a permutation level of 90%, most of the mergers have a size equal to 2 during half of the clustering process (for fusion numbers ranging from 1 to at least 2,000). However, for clusters with a size equal to 2, I_t is equal to m_t which explains the similiarities between m_t and I_t curves during the first steps of the clustering process, as the perturbation level increases. Since $(m_t)_t$ is increasing for standard HAC, this explains why I_t has less reversals in standard HAC for the first merger numbers when the perturbation level is higher. The same holds for $\bar{I}_t$ up to a fixed size factor of 2.

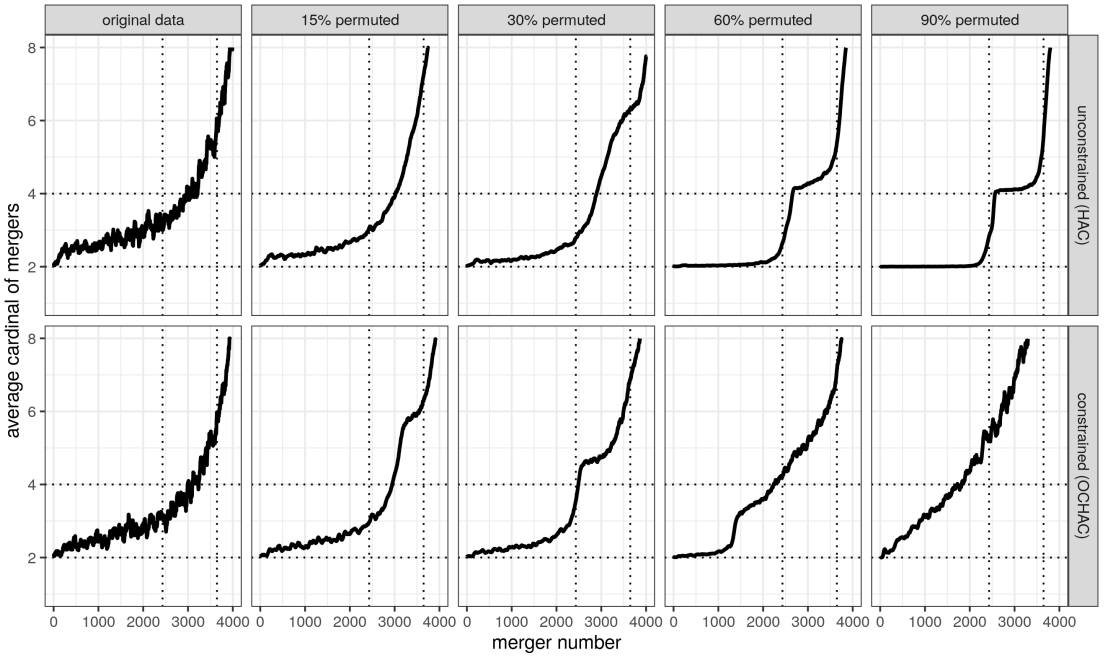


FIGURE 2.10: **Evolution of the average cardinal of mergers along the hierarchical clustering process** for standard HAC (top) and OCHAC (bottom), and different levels of perturbation. Note that the average is computed over the 50 simulations, whereas the original data correspond to a unique value. Data are shown only for the first 4,000 mergers and for a cardinal smaller than 8 for the sake of readability. The two dotted vertical lines correspond, respectively to $n/2$ and $3n/4$.

2.6 Conclusion

In this article, we have studied the applicability of HAC and its constrained version to a wide range of input data. In particular, we have shown that these applications are justified beyond the Euclidean framework. We have also shown that the monotonicity of the sequence of heights is not always ensured, although this property is necessary for the sequence of clusterings obtained by cutting dendrograms to be consistent with the sequence of clusterings of the algorithm. We have clarified which heights have this property depending on the input data types and for the constrained and unconstrained HAC. We have also pinpointed an important distinction between this monotonicity and the existence of crossovers.

These results imply that the variance of the merged cluster, I_t, or the average variance of the merged cluster, $\bar{I}_t$, are never ensured to be monotonic, and should thus not be chosen to represent the dendrogram heights. Strikingly, we have also

shown that the constrained version of the HAC can provide more relevant and efficient solutions than its unconstrained versions, not only in terms of algorithmic complexity, but also in terms of the values of the objective function ESS_t. In such cases, a small number of reversals can actually be beneficial to explore intermediate solutions closer to the data and that lead to more relevant clusters.

Acknowledgements

The authors would like to thank Marie Chavent for numerous instructive discussions on this paper.

The authors are grateful to the GenoToul bioinformatics platform (INRAE Toulouse, http ://bioinfo.genotoul.fr/) and its staff for providing computing facilities.

Funding

The PhD thesis of N.R. is funded by the INRAE/Inria doctoral program 2018. This work was also supported by the SCALES project funded by CNRS (Mission “Osez l’interdisciplinarité”).

Chapitre 3

Comparaisons d'arbres

Le chapitre précédent nous permet de motiver l'utilisation de la CAHCO (appelée OCHAC au chapitre précédent) sur les données Hi-C afin de modéliser la structure tridimensionelle du génome par des arbres binaires enracinés ou dendrogrammes. L'idée de notre méthode d'analyse différentielle repose sur le fait de traduire le problème d'analyse différentielle initialement posé pour n matrices Hi-C $\{H^i\}_{1\leq i\leq n}$ réparties en deux conditions $\mathcal{C}_1$ et $\mathcal{C}_2$ (telles que $\mathcal{C}_1 \cup \mathcal{C}_2 = \{1, \ldots, n\}$ et $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$), en un problème de comparaison de deux familles d'arbres, $\{\tau^i\}_{i\in\mathcal{C}_1}$ et $\{\tau^i\}_{i\in\mathcal{C}_2}$, construits sur les mêmes feuilles. Il est donc nécessaire de disposer d'une méthode statistique de comparaison de deux familles d'arbres.

Dans ce chapitre, on se sépare donc momentanément du contexte d'analyse différentielle de données Hi-C pour se focaliser sur celui de la comparaison de deux familles d'arbres d'un point de vue statistique. Ainsi, les points développés ici peuvent s'appliquer aux données Hi-C mais aussi à tout autre type de données faisant intervenir des arbres enracinés (données de Genome Wide Association Studies [GWAS], phylogénies, etc).

La littérature sur le sujet de la comparaison d'arbres s'oriente essentiellement vers la comparaison d'un couple d'arbres à travers l'utilisation de distances. En effet, les distances entre arbres sont nombreuses et possèdent des propriétés très variées. Certaines dites *topologiques* ne s'intéressent qu'à l'ordre des associations entre feuilles tandis que les distances pondérées, à l'inverse, prennent en compte la longueur des branches. Une étude de ces distances et de leur propriétés est menée dans la section 3.1.

Néanmoins, l'utilisation d'une distance seule est insuffisante pour répondre à notre problématique et la recherche d'une statistique de comparaison entre deux familles est développée dans la section 3.2.

On valide ensuite l'approche développée sur données réelles (GWAS) dans la section 3.3. Enfin, nous illustrons ses possibilités d'application dans différents contextes faisant intervenir des arbres, tels que les données Hi-C ou de phylogénies dans la section 3.4.

3.1 Distances entre arbres

En théorie des graphes, un arbre est un graphe non orienté, acyclique et connexe. On peut distinguer deux types de sommets dans un arbre : les feuilles, qui correspondent aux sommets de degré 1 et les nœuds internes qui correspondent aux sommets dont le degré est strictement supérieur à 1. Il est possible d'enraciner un arbre, c'est-à-dire de définir un nœud racine, et de ce fait, d'attribuer une orientation aux arêtes. Enfin, les arbres obtenus par CAH sont particuliers : il s'agit d'arbres enracinés *binaires*, c'est-à-dire des arbres dont tous les nœuds internes sont de degré 3, excepté la racine qui est de degré 2. Un exemple d'un tel arbre

est représenté dans la figure 3.1.

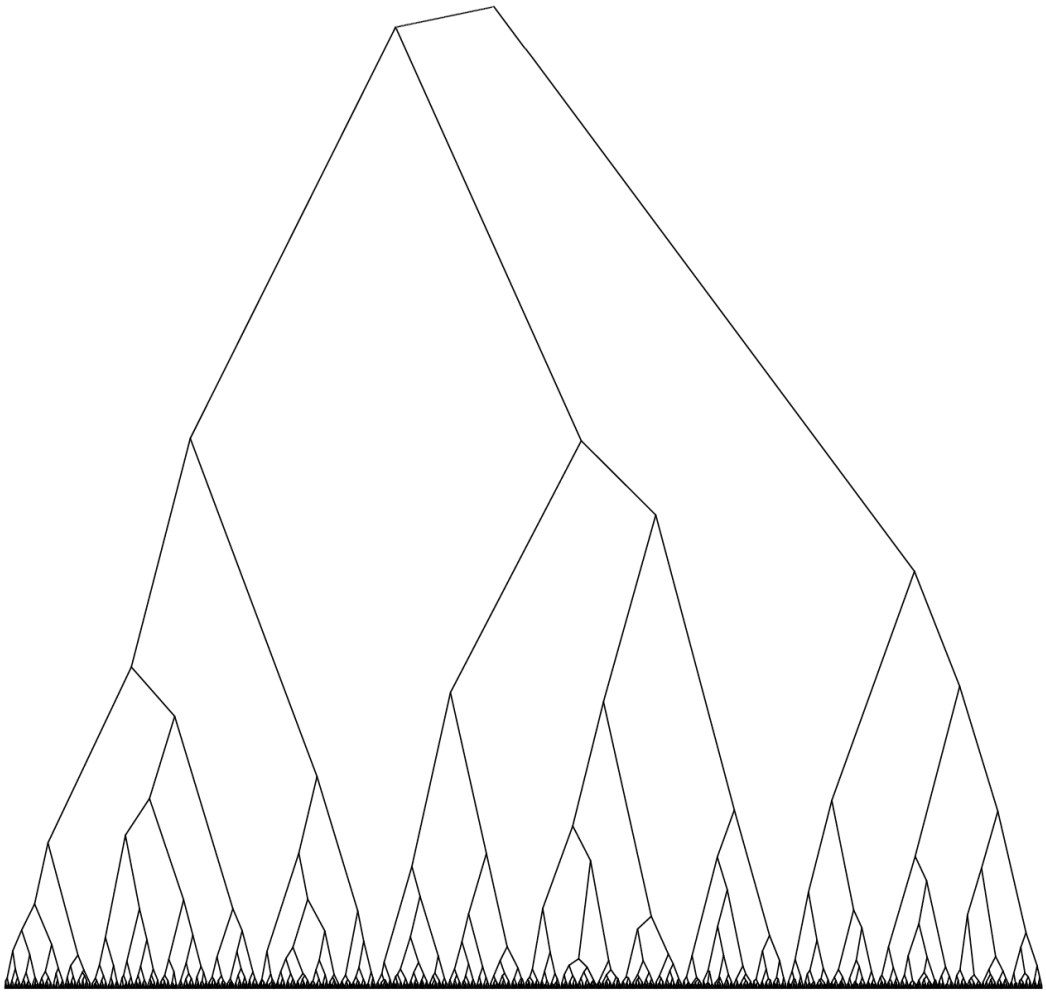


FIGURE 3.1: **Exemple d'arbre binaire enraciné obtenu par CAH.** La racine correspond ici au nœud le plus haut tandis que les feuilles sont représentées horizontalement en bas de l'arbre.

Dans la littérature, la majorité des approches de comparaisons entre arbres se font deux à deux et exploitent des distances. La plupart de ces distances ont été développées dans le cadre des analyses de phylogénies, domaine où les représentations sous forme d'arbres sont très présentes. On peut également citer les distances entre graphes, qui s'appliquent naturellement aux arbres en tant que cas particuliers, mais on se restreindra dans cette partie aux distances spécifiques aux arbres.

Dans la suite, on désignera par τ un arbre enraciné quelconque.

3.1.1 Généralités sur les distances entre arbres

Il y a essentiellement deux approches différentes pour comparer deux arbres à l'aide d'une distance. La première consiste à considérer uniquement l'ordre des branchements entre les différentes feuilles constituant l'arbre, en laissant de côté les longueurs des branches. On ne s'intéresse alors qu'à la structure ou topologie de l'arbre, c'est pourquoi on qualifiera ces distances de *topologiques*. L'autre point de vue possible est de prendre en compte la longueur des branches lors du calcul de la métrique. On parle, dans ce dernier cas, de distances *pondérées*.

Il est fréquent, lorsqu'on cherche à évaluer la dissimilarité entre deux arbres, d'avoir recours à des pseudo-distances. Cependant, nous nous limiterons ici à des distances au sens mathématique du terme, c'est-à-dire vérifiant les axiomes de symétrie, séparabilité ainsi que l'inégalité triangulaire.

Distances topologiques. Une façon générique d'obtenir une distance topologique est de choisir un type de modification élémentaire de la structure topologique d'un arbre et de définir la distance par le nombre minimal d'opérations de ce type nécessaires pour passer d'un arbre à l'autre. Dans la suite, on qualifiera ce type de distance de *distances d'édition* (DasGupta et collab., 1998). Elles vérifient par construction, les axiomes de symétrie et de séparation. De plus, elles vérifient également l'inégalité triangulaire,

$$\forall(\tau,\tau',\tau''),\quad d(\tau,\tau')\leq d(\tau,\tau'')+d(\tau'',\tau')$$

puisque passer de τ à τ'' puis de τ'' à τ' donne un chemin de τ à τ' dont le nombre d'étapes n'est pas minimal en général. Il s'agit donc de distances sur l'espace des arbres enracinés topologiques (c'est-à-dire, sans information de longueurs sur les branches).

Une des distances d'édition les plus connues est la distance de Robinson-Foulds (Robinson et Foulds, 1981) qui est le nombre minimal de contractions/décontractions nécessaires pour passer d'un arbre à l'autre. Une contraction consiste à fusionner deux nœuds séparés par une arête, en créant ainsi un nœud dont le degré est la somme des deux précédents diminuée de 2. Une décontraction consiste en l'opération inverse, où l'on peut alors choisir comment répartir les arêtes incidentes lors du dédoublement du nœud considéré. Une suite d'opérations de contraction/décontraction est représentée dans la figure 3.2.

Ainsi, la distance de Robinson-Foulds n'est pas restreinte aux arbres binaires. Elle est aussi qualifiée de *symmetric difference metric* en référence au fait qu'une façon de la calculer est de déterminer le nombre de bipartitions des feuilles (partition des feuilles en deux classes induite par la suppression d'une branche interne) spécifiques de chacun des deux arbres. Plus formellement :

$$d_{RF}(\tau,\tau')=|b(\tau)|+|b(\tau')|-|b(\tau)\cap b(\tau')|=|\Delta(b(\tau),b(\tau'))|$$

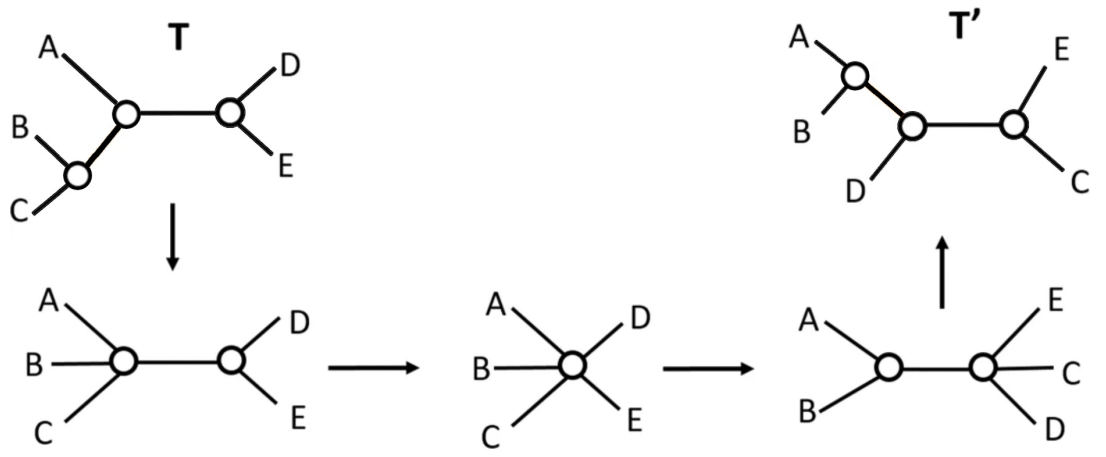


FIGURE 3.2: **Exemple de chemin de contraction/décontraction entre les arbres T et T'.** Figure adaptée de Briand et collab. (2020).

où $b(\tau)$ (resp. $b(\tau')$) désigne l'ensemble des bipartitions de l'arbre τ (resp. τ') et Δ, la différence symétrique ensembliste.

Il existe d'autres distances basées sur le principe de compter le nombre minimal d'un certain type de modification pour passer d'un arbre à un autre. Parmi les plus connues, on peut citer la *SPR distance* (Subtree Pruning and Regrafting) basée sur le débranchement et le rebranchement de sous-arbres, ou encore la *NNI distance* (Nearest Neighbor Interchange) introduite de façon indépendante par Robinson (1971) et Moore et collab. (1973), dont la modification élémentaire consiste à échanger les positions de deux sous-arbres voisins, c'est-à-dire, séparés par une seule branche interne.

La distance quartet (ou *Quartet metric*) est un autre type de distance topologique (Bandelt et Dress, 1986). Elle s'obtient comme la différence symétrique des ensembles de sous-arbres à 4 feuilles induits respectivement par chacun des deux dendrogrammes :

$$d_{quartet}(\tau, \tau') = |b'(\tau)| + |b'(\tau')| - |b'(\tau) \cap b'(\tau')| = |\Delta(b'(\tau, \tau')|$$

où $b'(\tau)$ (resp. $b'(\tau')$) désigne l'ensemble des sous-arbres à 4 feuilles de l'arbre τ (resp. τ') et Δ, la différence symétrique ensembliste.

Enfin, certains travaux utilisent des distances initialement prévues pour d'autres objets en se basant sur des représentations spécifiques des arbres. Par exemple, Diaconis et Holmes (1998) montrent l'existence d'une bijection entre les arbres binaires enracinés à B feuilles et les groupements des $2B-2$ premiers entiers en paires, qui peuvent être interprétés comme des produits de transpositions. On a alors une application injective de l'espace des arbres dans l'ensemble des

permutations, qui permet de transporter une distance initialement conçue pour les permutations sur l'espace des arbres.

Distances pondérées. Les distances pondérées diffèrent des distances évoquées précédemment par le fait qu'elles prennent en compte la longueur des branches lors des calculs.

Un premier point intéressant à noter est que les distances d'édition peuvent souvent s'étendre au cas pondéré de la façon suivante : au lieu d'associer la valeur unité à une modification de base dans le calcul de la distance, on associe la longueur de l'arête en jeu dans cette modification. Étendues de cette façon, elles vérifient encore les axiomes de la définition d'une distance mais cette fois sur l'espace des arbres enracinés pondérés. Ainsi, la distance de Robinson-Foulds par exemple s'adapte naturellement pour prendre en compte la longueur des branches des arbres (Robinson et Foulds, 1979). De même, la distance NNI se généralise facilement au cas pondéré en pondérant chaque échange de sous-arbres voisins par la longueur de la branche interne concernée.

La distance BHV (*Billera-Holmes-Vogtmann distance*) est introduite par l'article de Billera et collab. (2001) dans lequel est étudié l'espace géométrique des arbres pondérés enracinés basés sur un même ensemble de feuilles. Chaque topologie d'un arbre binaire y est représentée par un orthant (on appelle orthant de dimension p l'ensemble $\mathbb{R}_+^p$) dont la dimension correspond au nombre de branches internes de l'arbre. Ainsi, pour un arbre à B feuilles, la dimension de l'orthant sera $B-2$. Tout point de l'orthant correspond alors à un arbre et les coordonnées de ce point dans l'orthant correspondent aux longueurs des branches internes de cet arbre. L'espace des arbres est composé d'une juxtaposition d'orthants, dont le nombre est égal au nombre de topologies possibles parmi les arbres binaires enracinés à B feuilles. Dans ce contexte, les frontières d'un orthant correspondent donc à des topologies dégénérées, dont certaines des branches internes ont une longueur nulle, et permettent donc la transition d'une topologie vers une autre, comme illustré dans la figure 3.3.

À l'intérieur de chaque orthant, on dispose de la distance euclidienne classique $\|\cdot\|_2$ de $\mathbb{R}^{B-2}$, correspondant à la longueur du segment de ligne droite joignant les deux arbres. Pour définir une distance entre deux arbres quelconques de cet espace des arbres pondérés enracinés, on va chercher à prolonger la distance euclidienne classique disponible dans chaque orthant, à une paire d'arbres n'appartenant pas nécessairement au même orthant. En effet, il est possible de construire un plus court chemin entre deux arbres, ou *géodésique*, composé de segments de droites correspondant à chaque orthant traversé. La longueur de ce plus court chemin permet de définir une distance géodésique : la distance BHV.

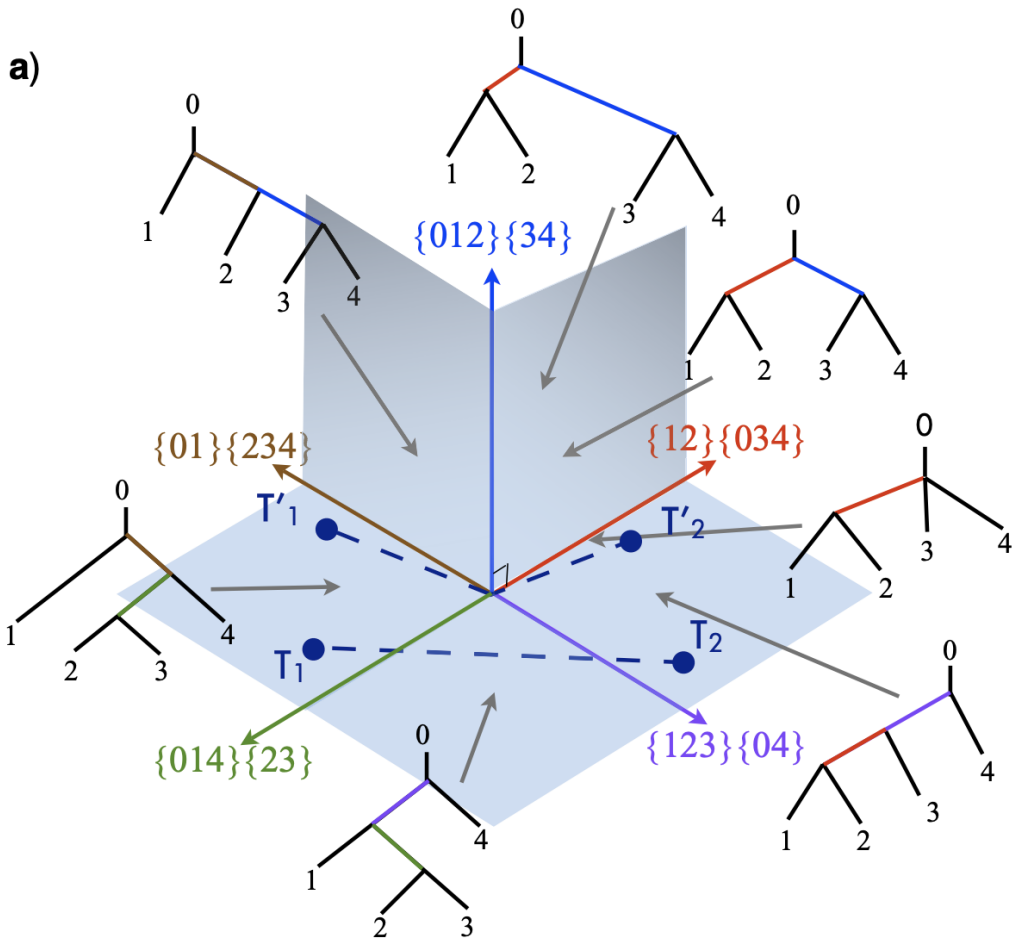


FIGURE 3.3: **Illustration de cinq orthants parmi les quinze de l'espace géométrique des arbres binaires enracinés à 4 feuilles.** Dans cette illustration, on peut voir cinq orthants (ici, des quadrants $\mathbb{R}^2_+$) correspondant à cinq topologies binaires distinctes. Ces quadrants partagent des frontières qui correspondent à des topologies binaires dégénérées, c'est-à-dire obtenues à partir des différentes topologies en annulant la longueur de l'une des deux branches internes. Chaque axe-frontière quantifie la longueur d'une branche interne qui est représentée par la bipartition associée (la couleur d'un axe est la même que celle de l'arête interne qu'il décrit). Enfin, deux géodésiques, plus courts chemins entre deux arbres, sont réprésentées en pointillés bleus. Sur ces arbres, la racine est distinguée par la présence d'une feuille 0. Figure extraite de Brown et Owen (2019).

Une autre stratégie de construction d'une distance pondérée est d'utiliser une représentation vectorielle des arbres associée à une distance standard dans

l'espace vectoriel. Si la représentation vectorielle est injective, alors la distance induite sur l'espace des arbres vérifiera bien les axiomes de la définition d'une distance. C'est le cas des représentations vectorielles évoquées dans la suite.
Ainsi par exemple, la *Branch Score Distance* (Kuhner et Felsenstein, 1994) est basée sur une représentation des arbres par leurs longueurs de branches. Les deux arbres ayant les mêmes ensembles de feuilles, on construit pour chaque arbre le vecteur de l'ensemble des longueurs de branches pour toutes les branches possibles, même celles n'existant pas forcément dans les arbres considérés. En effet, une branche possible est déterminée par une bipartition, c'est-à-dire une séparation en deux groupes de l'ensemble des feuilles (séparation que l'on obtiendrait en supprimant cette même branche et en considérant les feuilles des deux parties connexes obtenues). Étant donné un arbre, le vecteur est construit en pratique de la façon suivante : si la branche n'est pas présente dans l'arbre, alors la valeur associée est 0 sinon, il s'agit de sa longueur dans l'arbre. La *Branch Score Distance* est alors la distance euclidienne canonique entre ces deux vecteurs. Cependant, la dimension des vecteurs obtenus, $2^{B-1}-1$, croît fortement avec le nombre de feuilles commun B des arbres considérés.

La *weighted Path Difference Metric* (Steel et Penny, 1993) (wPD) est elle aussi basée sur une représentation vectorielle des arbres mais dans un espace de dimension $B(B-1)/2$ plus faible que pour la distance précédente. Dans ce cas, les arbres sont représentés par des vecteurs composés de l'ensemble des longueurs de chemins entre paires de feuilles. La *weighted Path Difference Metric* est alors la distance euclidienne canonique entre ces deux vecteurs comme illustré sur la figure 3.4.

Il existe un lien entre la distance **wPD** et la matrice de distance cophénétique. La matrice de distance cophénétique est la matrice $\mathbf{X} = (x_{kl})_{1\leq k,l\leq B}$ dont le coefficient x_{kl} est la hauteur à laquelle les feuilles k et l apparaissent pour la première fois sous le même nœud. La longueur du chemin joignant la feuille k et la feuille l est alors le double de x_{kl} (en effet, lors du calcul de la longueur du chemin les arêtes horizontales ne comptent pas car elles dépendent de la représentation graphique choisie).
On notera $X \in \mathbb{R}^p$ (resp. $X' \in \mathbb{R}^p$), avec $p = B(B-1)/2$, le vecteur des coefficients de la partie triangulaire supérieure de la matrice des distances cophénétiques $\mathbf{X}$ (resp. $\mathbf{X}'$). On a alors :

$$d_{\text{wPD}}(\tau, \tau') = 2\|X - X'\|_2 \tag{3.1}$$

Enfin, il existe d'autres mesures basées sur les matrices de distances cophénétiques. Elles se basent sur des mesures de la similarité entre deux arbres comme par exemple, le coefficient de corrélation cophénétique de Sokal et Rohlf (1962). Le principe est de calculer la corrélation de Pearson entre les matrices de distances

cophénétiques de deux dendrogrammes. Dans le cas d'une comparaison topologique, au lieu d'utiliser la hauteur du premier nœud regroupant deux feuilles pour construire la matrice de distance cophénétique, on remplace cette valeur par l'indice de ce nœud dans l'ordre d'agrégation induit par les hauteurs (prises dans l'ordre croissant). Dans ce cas, Baker (1974) conseille pour calculer la corrélation l'utilisation du gamma de Kruskal et Goodman (Goodman et Kruskal, 1959), mieux adapté à la gestion des rangs et des ex-æquo.

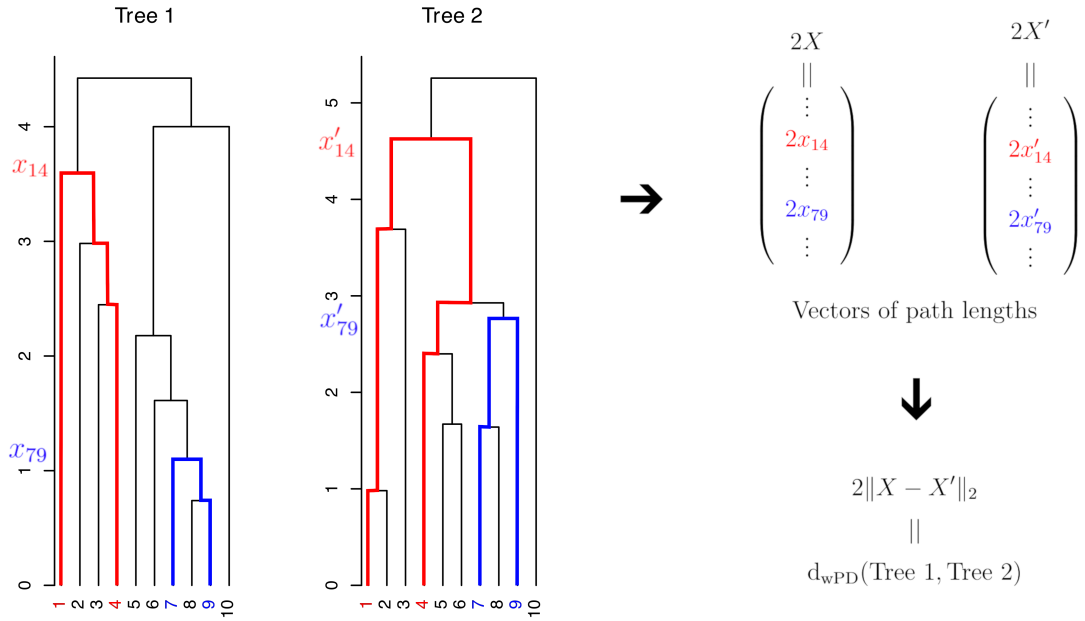


FIGURE 3.4: **Illustration du calcul de la *weighted Path Difference Metric*.** On commence par calculer l'ensemble des longueurs de chemins entre paires de feuilles. Pour l'exemple, on représente en couleur (rouge et bleu respectivement) deux chemins se correspondant pour les deux arbres. On stocke ces quantités dans des vecteurs représentant les arbres à comparer (les vecteurs des longueurs de chemin correspondent au double des vecteurs de distances cophénétiques). La valeur de la *weighted Path Difference Metric* est alors la distance euclidienne canonique entre ces deux vecteurs.

3.1.2 Propriétés des distances

Devant le grand nombre de distances disponibles, il peut être difficile de réaliser un choix pertinent. Cette sous-section s'intéresse aux critères, théoriques ou pratiques, qui peuvent guider le choix d'une distance plutôt qu'une autre.

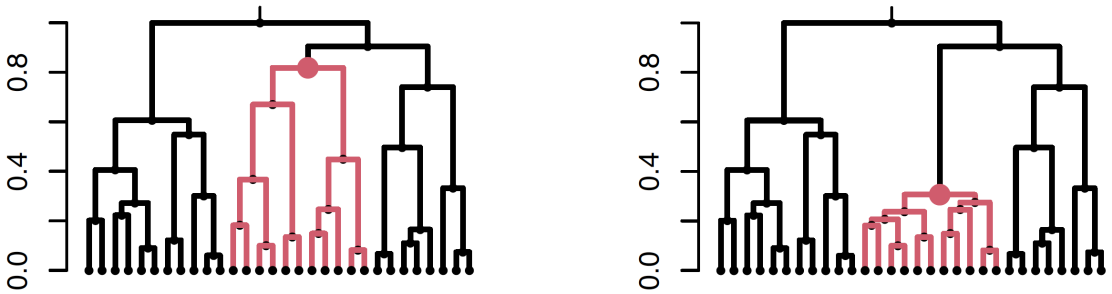
Le premier élément à considérer au moment de sélectionner une distance est son aspect pondéré ou topologique. Une distance topologique ne tiendra compte que

de l'ordre des branchements des arbres. Ce type de distance est adapté quand les arbres ont toutes leurs branches internes de même longueur ou quand l'information portée par la longueur des branches est jugée non pertinente.
En revanche, si l'on considère que la longueur des branches est informative, on aura tout intérêt à privilégier une distance pondérée. On pourra alors distinguer deux arbres de même topologie mais dont les longueurs des branches varient, ce qui n'est pas possible avec une distance topologique.
Dans ce travail, on se placera dans le cas où les longueurs des branches portent de l'information et on privilégiera donc les distances pondérées comme par exemple la distance de Robinson-Foulds pondérée, la distance BHV, ou encore la *weighted Path Difference Metric*.

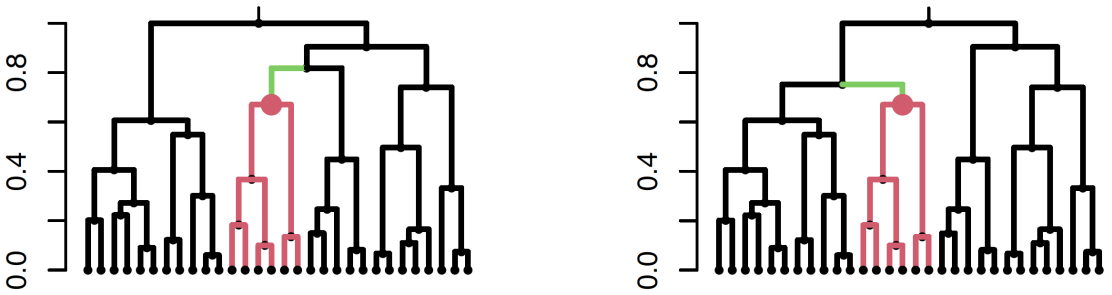
Un autre point très important motivant le choix d'une distance entre arbres est le suivant : la bonne quantification par la distance de certains types de différences dans la structure des arbres. En effet, les arbres binaires enracinés pondérés étant des objets relativement complexes, les différences observées peuvent être de natures distinctes. Ces types de différences correspondent à la traduction en termes d'arbres des modifications évoquées pour les matrices Hi-C par Cresswell et Dozmorov (2020) (différences illustrées dans la figure 3.5). Il est souhaitable d'une part qu'une distance soit suffisamment sensible pour percevoir ces différences, mais d'autre part, qu'elle ait une forme de continuité vis-à-vis de ces différences, afin de pouvoir quantifier avec pertinence leur amplitude. La figure 3.5 présente les types de différences qui nous intéresseront majoritairement par la suite.

Certaines différences entre arbres ne reposent pas sur des modifications de la topologie des arbres, d'où la nécessité de disposer d'une distance pondérée. C'est le cas de l'apparition ou de la disparition d'une frontière entre deux groupes de feuilles, qui va se traduire respectivement par une élévation ou un abaissement marqué de la hauteur à laquelle ces deux groupes sont agrégés (cas **(A)** de la figure 3.5). D'autres modifications, en revanche, joueront plus sur l'ordre des branchements : c'est le cas des décalages de frontières, par exemple, qui étant donnée une frontière, consistent à faire passer un groupement de feuilles d'un côté de la frontière à l'autre, décalant par conséquent cette dernière (cas **(B)** de la figure 3.5). Enfin, des modifications un peu plus complexes, comme des différences de subdivisions pour un groupe de feuilles donné, qui ne se décomposerait pas en le même nombre de sous-groupes pour les deux arbres, sont également intéressantes à détecter (cas **(C)** de la figure 3.5).

(A). Différence de hauteurs



(B). Différence de branchement



(C). Différence mixte

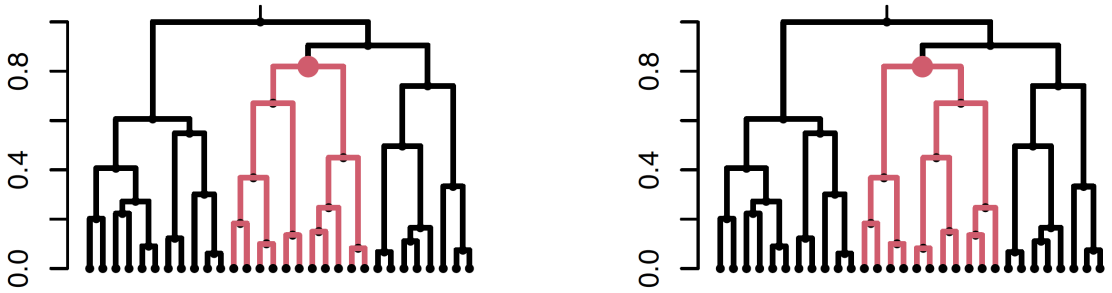


FIGURE 3.5: **Illustrations des types de différences entre arbres.** Les sous-arbres sur lesquels les différences sont appliquées sont représentés en rouge. À gauche : pour chaque ligne, le même arbre témoin est représenté. À droite : pour chaque ligne, un arbre obtenu par modification de l'arbre témoin est représenté. La ligne **(A)** présente une différence de hauteurs : le sous-arbre central est plus dense à droite qu'à gauche (aucune modification de branchements en revanche). La ligne **(B)** présente une différence de branchement (matérialisée en vert) : le sous-arbre rouge ne se rebranche pas de la même façon entre l'arbre témoin et l'arbre modifié. La ligne **(C)** présente une différence dite « mixte » portant sur le sous-arbre rouge car elle se compose à la fois de différences de hauteurs et de branchements.

Toutes les distances entre arbres ne remplissent pas ces conditions. Il est, par exemple, connu que la distance topologique de Robinson-Foulds peut prendre sa valeur maximale pour une différence consistant en un unique échange de feuilles (Böcker et collab., 2013).

Les aspects pratiques computationnels sont également à prendre en compte en fonction de la taille (nombre de feuilles) des arbres considérés. Ainsi, certaines distances sont compliquées à évaluer comme la distance NNI ou la distance SPR qui sont NP-difficiles à calculer (Dasgupta et collab., 1997; Bordewich et Semple, 2005). Dans ces cas, on utilise la plupart du temps des approximations, plus simples à obtenir, des distances en question. D'autres en revanche sont rapides à évaluer. Pour la distance de Robinson-Foulds par exemple, Day (1985) a introduit un algorithme de calcul en temps linéaire. La distance BHV, quand à elle, peut être calculée en temps polynomial (Owen et Provan, 2011).

Enfin, la majorité des méthodes évoquées ci-dessus possèdent des implémentations disponibles librement. Ainsi, le package R **phangorn** développé par Schliep (2010), permet d'avoir accès à la distance de Robinson-Foulds et à sa version pondérée, à la Branch Score Distance, à la weighted Path Difference Metric et à une version approchée de la SPR distance (de Oliveira Martins et collab., 2008). Le package R **distory**, développé par Chakerian et Holmes (2012) permet quant à lui d'avoir accès à la distance BHV. On peut également citer le package R **TreeDist** (Smith, 2020) qui implémente un certain nombre d'autres distances parmi lesquelles une approximation en temps polynomial de la distance NNI (Li et collab., 1996).

Pour la suite de ce travail, nous nous sommes focalisés sur la weighted Path Difference Metric. En effet, il s'agit d'une distance pondérée, relativement simple à calculer et dont une implémentation est disponible dans le package R **phangorn**. D'autre part, nous avons pu constater empiriquement ses bonnes propriétés en termes de détection des différences puisqu'elle est sensible aux trois types illustrées dans la figure 3.5, et permet de quantifier l'intensité de ces différences. Nous retenons donc pour la suite de la thèse la représentation des arbres par leurs vecteurs de longueurs des chemins entre paires de feuilles (ce qui est équivalent à la représentation des arbres par leurs vecteurs de distances cophénétiques ; voir la définition 3.1 de la weighted Path Difference Metric).

Cependant, une distance seule ne permet pas de comparer statistiquement deux ensembles d'arbres. Des travaux (Steel, 1988; Steel et Penny, 1993; Critchlow et collab., 1996) ont été menés dans l'idée d'attribuer une signification statistique à la valeur d'une distance entre un couple d'arbres en étudiant les distributions de distances obtenues après avoir fixé au préalable un modèle de distribution sur l'espace des arbres lui-même. Une limitation de cette approche est qu'il est généralement difficile de vérifier l'adéquation d'un modèle de distribution d'arbres avec

un contexte applicatif donné. D'autre part, nous nous intéressons à la comparaison de deux groupes d'arbres et non uniquement d'un couple, et il est donc nécessaire d'utiliser une statistique de comparaison capable de prendre en compte plusieurs individus par conditions. La partie suivante est donc dédiée à la recherche et au développement d'une telle statistique.

3.1.3 D'une distance entre arbres vers une statistique de comparaison de deux ensembles d'arbres

Vouloir comparer statistiquement deux ensembles d'arbres suppose d'être capable de définir une distribution de probabilité sur l'espace des arbres enracinés à B feuilles, noté $\mathcal{T}$. Billera et collab. (2001) proposent un modèle de distribution de probabilité basé sur l'utilisation d'une distance entre arbres : il s'agit du modèle de Mallows (Mallows, 1957). Initialement conçu pour des classements d'objets, ce modèle s'étend à tout espace métrique muni d'une distance d. La densité de probabilité pour un arbre τ est définie par :

$$f(\tau) = K \exp^{-\lambda d(\tau,\tau_0)}, \quad \tau \in \mathcal{T} \tag{3.2}$$

où τ_0 désigne un arbre « central » pour la distribution, λ est un paramètre de dispersion, et K une constante de normalisation. Ce cadre permet alors de faire de l'inférence statistique sur l'espace des arbres (Holmes, 2003; Chakerian et Holmes, 2012), et donc de réaliser des tests d'hypothèses sur les paramètres des distributions respectives des deux ensembles d'arbres.

Dans notre cas, nous raisonnons par analogie et remplaçons d dans l'expression de la densité dans l'équation (3.2) par le carré de la distance wPD pour les bonnes propriétés de cette distance, évoquées dans la section précédente. L'estimation des paramètres τ_0 et λ de la distribution devrait ensuite se faire sur l'espace des arbres. Cependant, étant donné la complexité de cet espace, il peut être intéressant de simplifier ce problème d'estimation en exploitant la définition 3.1 de la distance wPD. En effet, $d^2_{\text{wPD}}(\tau, \tau_0) = 4\|X - X_0\|_2^2$ où $X \in \mathbb{R}^p$ (resp. $X_0 \in \mathbb{R}^p$) est le vecteur des distances cophénétiques de τ (resp. τ_0). L'idée est alors de considérer l'approximation consistant à remplacer les arbres par leurs vecteurs de longueurs de chemins et à estimer X_0 par la moyenne de ces vecteurs, en s'affranchissant du problème d'estimation de l'arbre central τ_0 dans $\mathcal{T}$. Il est important de noter que cette approximation est abusive dans la mesure où le plongement de l'espace des arbres dans $\mathbb{R}^p$ est injectif mais pas surjectif. Une conséquence est donc qu'on ne peut pas faire correspondre en général la moyenne de vecteurs de longueurs de chemins dans $\mathbb{R}^p$ à un arbre τ_0 de $\mathcal{T}$. Cependant, cette approximation permet de reformuler la densité sur $\mathbb{R}^p$ de la façon suivante :

$$f(X) = K \exp^{-\lambda' \|X - X_0\|_2^2}, \quad X \in \mathbb{R}^p \qquad (\lambda' = 4\lambda)$$

On reconnait alors la densité d'une loi gaussienne multivariée $\mathcal{N}_p(X_0, \Sigma)$ avec $\Sigma = 1/(2\lambda)\mathbb{I}_p$. L'hypothèse de relaxation consistant à plonger les arbres dans $\mathbb{R}^p$ selon la modalité suggérée par la distance wPD nous oriente donc vers une modélisation des vecteurs de longueurs de chemins par une loi gaussienne multivariée (ici avec une matrice de variance-covariance diagonale homoscédastique).

Faire l'hypothèse d'une distribution gaussienne multivariée pour les vecteurs de longueurs de chemins permet l'utilisation d'une statistique naturelle de test pour la comparaison des moyennes des deux ensembles d'arbres, la statistique de Hotelling. Dans la suite, nous allons donc explorer cette piste.

3.2 Construction d'une statistique de comparaison

3.2.1 Formalisation du problème

Dans cette section, on va formaliser notre problème de comparaison afin de pouvoir y répondre statistiquement. On dispose de n arbres enracinés, $\{\tau^1, \dots, \tau^n\}$, avec un même nombre de feuilles B. De plus ces arbres sont séparés en deux conditions $\mathcal{C}_1$ et $\mathcal{C}_2$ telles que :

$$\mathcal{C}_1 \cup \mathcal{C}_2 = \{1, \dots, n\} \quad \text{et} \quad \mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$$

On pose $n_1 = |\mathcal{C}_1|$ et $n_2 = |\mathcal{C}_2|$.

Les arbres τ^i sont donc représentés par leurs vecteurs de distances cophénétiques (définis dans la sous-section 3.1 dans le paragraphe consacré à la *weighted Path Difference Metric*), $X_i \in \mathbb{R}^p$, avec $p = B(B-1)/2$. Le problème de comparaison des deux familles d'arbres est alors transposé en un problème de comparaison de deux ensembles de vecteurs de $\mathbb{R}^p$, $\{X_i\}_{i\in\mathcal{C}_1}$ et $\{X_i\}_{i\in\mathcal{C}_2}$. Les $\{X_i\}_{i\in\mathcal{C}_1}$ sont n_1 observations du vecteur aléatoire $X^{(1)} \in \mathbb{R}^p$ décrivant la loi propre à la condition 1, et les $\{X_i\}_{i\in\mathcal{C}_2}$, n_2 observations de $X^{(2)} \in \mathbb{R}^p$ décrivant la loi propre à la condition 2. Notre objectif est de tester l'égalité des espérances des deux vecteurs aléatoires $X^{(1)}$ et $X^{(2)}$, notées respectivement $\mu^{(1)}$ et $\mu^{(2)}$. Formellement, on souhaite tester sur $\mu^{(1)} \in \mathbb{R}^p$ et $\mu^{(2)} \in \mathbb{R}^p$ l'hypothèse nulle :

$$\mathcal{H}^0 : \mu^{(1)} = \mu^{(2)} = \mu \tag{3.3}$$

contre l'hypothèse alternative :

$$\mathcal{H}^1 : \mu^{(1)} \neq \mu^{(2)} \tag{3.4}$$

De plus, en notant $\mu^{(1)} = (\mu_j^{(1)})_{1\leq j\leq p}$ et $\mu^{(2)} = (\mu_j^{(2)})_{1\leq j\leq p}$, on peut décomposer l'hypothèse nulle (3.3) de la façon suivante :

$$\mathcal{H}^0 = \cap_{j=1}^p \mathcal{H}_j^0 \quad \text{avec} \quad \mathcal{H}_j^0 : \mu_j^{(1)} = \mu_j^{(2)} \tag{3.5}$$

Comme expliqué dans la sous-section 3.1.3, on suppose que les lois de $X^{(1)}$ et $X^{(2)}$ sont des gaussiennes multivariées. On fait en outre l'hypothèse que la matrice de variance-covariance est commune à $X^{(1)}$ et $X^{(2)}$. On a donc $X^{(1)} \sim \mathcal{N}_p(\mu^{(1)}, \Sigma)$ et $X^{(2)} \sim \mathcal{N}_p(\mu^{(2)}, \Sigma)$.

On notera par la suite $X_{i,j}$ la j-ème coordonnée de l'observation $X_i \in \mathbb{R}^p$ et $\bar{X}_j^{(r)} = (n_r)^{-1} \sum_{i \in \mathcal{C}_r} X_{i,j}$ (et donc $\bar{X}^{(r)} = (\bar{X}_1^{(r)}, \ldots, \bar{X}_p^{(r)})$) avec $r = 1$ ou 2 selon la condition.

Une statistique naturelle pour tester $\mathcal{H}_0$ est la statistique du T^2 de Hotelling (Hotelling, 1931). Cette dernière est définie par :

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{X}^{(1)} - \bar{X}^{(2)})^\top \hat{\Sigma}^{-1} (\bar{X}^{(1)} - \bar{X}^{(2)}) \tag{3.6}$$

où $\hat{\Sigma}$ est l'estimateur empirique de Σ obtenu à partir des observations de $X^{(1)}$ et $X^{(2)}$ de la façon suivante :

$$\hat{\Sigma} = \frac{(n_1 - 1)\hat{\Sigma}^{(1)} + (n_2 - 1)\hat{\Sigma}^{(2)}}{n_1 + n_2 - 2} \tag{3.7}$$

où $\hat{\Sigma}^{(r)} = (n_r - 1)^{-1} \sum_{i \in \mathcal{C}_r} (X_i - \bar{X}^{(r)})(X_i - \bar{X}^{(r)})^\top$

Toutefois, le calcul de la statistique du T^2 de Hotelling définie par l'équation (3.6) n'est possible que si l'estimateur de la matrice de variance-covariance, $\hat{\Sigma}$, est inversible. Or, $\hat{\Sigma}$ est par construction de rang inférieur ou égal à $n - 1$, donc non inversible dès que $n \leq p$. Cette statistique n'est donc pas directement utilisable dans de nombreux contextes applicatifs où les observations sont trop peu nombreuses par rapport à la dimension du problème, comme par exemple l'analyse de données Hi-C.

Dans la suite de cette section, nous faisons une synthèse de différentes approches permettant de résoudre ce problème d'estimation, et nous proposons une méthode adaptée au cas où $n \ll p$.

3.2.2 Régularisation de l'estimateur de la matrice de variance-covariance $\hat{\Sigma}$

On peut comprendre la statistique de Hotelling comme une méthode particulière d'agrégation de statistiques univariées prenant en compte la dépendance entre les coordonnées des vecteurs $\bar{X}^{(1)}$ et $\bar{X}^{(2)}$ à travers une étape de décorrélation (multiplication de la matrice des données par $\hat{\Sigma}^{-1/2}$). Si prendre en compte la dépendance à travers la décorrélation des coordonnées constitue un point commun des méthodes optimales en termes de puissance statistique en petite dimension, le bénéfice d'une telle étape dans un contexte de grande dimension et/ou de forte dépendance est moins immédiat. En effet, pour la statistique de Hotelling par

exemple, c'est le calcul de $\hat{\Sigma}^{-1}$ en grande dimension, et donc l'étape de décorrélation, qui pose problème en rendant la statistique soit impossible à calculer, soit particulièrement instable.

Dans ce contexte de la grande dimension, deux points de vue existent dans la littérature : l'un préconisant la prise en compte de la dépendance (Liu et Xie, 2018) et l'autre suggérant qu'ignorer la dépendance puisse améliorer les résultats (Dudoit et collab., 2002; Wu et collab., 2014; Barnett et collab., 2017). Hébert (2019); Hébert et collab. (2021) proposent une approche adaptative pour réaliser un compromis adaptatif entre ces deux points de vue et montrent que le bon degré de prise en compte de la dépendance varie en fonction de la combinaison du vecteur $\mu^{(1)} - \mu^{(2)}$ et de Σ. De manière similaire, dans le contexte de la comparaison de moyennes multivariées en grande dimension, Cai et collab. (2013) insistent sur l'influence de la distribution des entrées du vecteur $\mu^{(1)} - \mu^{(2)}$ sur la puissance statistique des procédures de test.

Dans la suite de cette section, nous explorons ces deux pistes pour choisir la méthodologie la plus adaptée à notre contexte.

Régularisation globale de $\hat{\Sigma}$

Dans leur article, Dong et collab. (2016) évoquent une piste basée sur la régularisation globale de $\hat{\Sigma}$. Cette approche est développée par plusieurs auteurs : Chen et collab. (2011) proposent de remplacer $\hat{\Sigma}$ par $\hat{\Sigma} + \lambda \mathbb{I}_p$ ($\lambda \in \mathbb{R}_+$) tandis que Shen et collab. (2011) utilisent une combinaison convexe $\lambda\hat{\Sigma} + (1-\lambda)\mathbb{I}_p$ ($\lambda \in [0,1]$), où λ est un paramètre de régularisation.

Dans Chen et collab. (2011), des résultats théoriques sont obtenus pour le cas $n < p$. En particulier, sous certaines hypothèses (notamment de contrôle sur la croissance du ratio p/n), les auteurs montrent que la statistique de Hotelling régularisée suit une loi asymptotiquement gaussienne. Cependant, cette hypothèse de contrôle n'est pas réaliste en pratique dans le type d'application qui nous intéresse. Dans Shen et collab. (2011), les auteurs proposent l'utilisation d'un test par permutations. Mais dans les cas où n est très petit, par exemple $n = 6$ et $n_1 = n_2 = 3$ (cas d'étude de la sous-section 3.4.2), le nombre de permutations disponibles est de $\binom{n}{n_1} = 20$, ce qui est trop faible pour approcher la distribution empirique dans l'optique de réaliser un test.

T^2 sans estimation de la matrice de variance-covariance

Une autre idée, relativement simple et exposée dans Dong et collab. (2016), correspond au cas extrême $\lambda = 1$ de la régularisation de Shen et collab. (2011). Elle consiste à remplacer $\hat{\Sigma}$ par la matrice identité, $\mathbb{I}_p$, dans le calcul de la statistique

de Hotelling. Dans ce cas, la statistique de test devient :

$$T^2_{\text{alt}} = \frac{n_1 n_2}{n_1 + n_2}(\bar{X}^{(1)} - \bar{X}^{(2)})^\top(\bar{X}^{(1)} - \bar{X}^{(2)}) = \frac{n_1 n_2}{n_1 + n_2}\|\bar{X}^{(1)} - \bar{X}^{(2)}\|_2$$

Cette approche a été proposée initialement par Bai et Saranadasa (1996) dans le cas où p et n sont du même ordre de grandeur. Elle a été étendue au cas de la grande dimension ($n \ll p$) par Zhang et Xu (2009) et Chen et Qin (2010). Cette hypothèse sur la structure de la matrice de variance-covariance revient donc à considérer une statistique basée sur la norme L^2 de $(\bar{X}^{(1)} - \bar{X}^{(2)})$. Le résultat principal obtenu dans ces articles est une distribution asymptotique de la statistique qui ne requiert pas que le ratio p/n soit contrôlé et autorise donc la grande dimension.

Cependant, en pratique Chen et Qin (2010) conseillent un nombre d'observations $n \geq 20\log(p)$ afin de garantir une puissance raisonnable. Dans notre cas, n reste relativement petit du fait de la disponibilité limitée en observations et donc ces résultats asymptotiques semblent à nouveau inadaptés.

3.2.3 Régularisation des variances individuelles

Les deux types de solutions évoquées dans la sous-section précédente semblent finalement peu adaptées au cas où $n \ll p$. Ce problème a mené à une troisième famille de méthodes qui consiste à faire l'hypothèse que la matrice de variance-covariance commune a une structure diagonale :

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \sigma_p^2 \end{pmatrix}$$

Cette hypothèse a été faite dans ce contexte de grande dimension ($p \gg n$) dans différents travaux (Dudoit et collab., 2002; Bickel et Levina, 2004; Tong et Wang, 2012), avec de meilleurs résultats que les approches citées dans la section précédente (Dong et collab., 2016). Elle permet de réduire de façon importante le nombre de paramètres à estimer, notamment lorsque p est grand.

Néanmoins, le faible nombre d'observations fait que l'estimation des variances σ_j^2 est malgré tout difficile et même dans ce contexte simplifié, on a encore besoin de méthodes de régularisation pour les σ_j^2.

Régularisation des variances σ_j^2

Tusher et collab. (2001) proposent une régularisation relativement simple, au niveau de l'estimation des variances individuelles. En effet, l'estimateur de l'écart

type commun de la j-ème composante de $X^{(1)}$ et de $X^{(2)}$ défini par :

$$\hat{\sigma}_j = \sqrt{\frac{1}{n_1+n_2-2}\left(\sum_{i\in\mathcal{C}_1}\left(X_{i,j}-\bar{X}_j^{(1)}\right)^2+\sum_{i\in\mathcal{C}_2}\left(X_{i,j}-\bar{X}_j^{(2)}\right)^2\right)} \tag{3.8}$$

est remplacé par $\hat{\sigma}_j + \lambda$ où $\lambda \in \mathbb{R}^+$ est choisi pour minimiser un certain critère sur l'ensemble des j. Les auteurs utilisent ensuite une distribution obtenue par permutation des étiquettes des groupes. Tong et Wang (2012) apportent une autre réponse à ce problème en proposant un estimateur régularisé de σ_j^{-2} directement (au lieu d'utiliser l'inverse d'un estimateur de σ_j^2) en utilisant des informations provenant de tous les $(\sigma_j^{-2})_{j=1,\dots,p}$. L'idée est d'utiliser comme estimateur de σ_j^{-2} une moyenne géométrique pondérée de l'estimateur classique de σ_j^{-2} et d'un estimateur poolé obtenu à partir des estimateurs respectifs de tous les σ_j^2. Les auteurs montrent alors l'existence de paramètres de pondération de la moyenne géométrique optimaux pour la perte de Stein (James et Stein, 1992) ou celles des moindres carrés.

Une autre approche, cette fois-ci justifiée par un cadre de modélisation bayésienne, a également été développée pour l'analyse différentielle de données de puces à ADN et est implémentée dans le package R **limma**. Les fondements de cette approche sont décrits dans la section 3 de l'article de Smyth (2004). Cette approche consiste à supposer une loi *a priori* sur σ_j^{-2} :

$$\frac{1}{\sigma_j^2} \sim \frac{1}{\nu_0\sigma_0^2}\chi^2_{\nu_0}$$

où $\chi^2_{\nu_0}$ désigne une loi du χ^2 à ν_0 degrés de liberté et, ν_0 et σ_0 sont des hyperparamètres à estimer.

D'autre part, puisque l'on a supposé que $X^{(1)}$ et $X^{(2)}$ sont deux variables gaussiennes multivariées indépendantes avec même matrice de variance-covariance diagonale, on a :

$$(\bar{X}_j^{(1)}-\bar{X}_j^{(2)})|(\mu^{(1)}-\mu^{(2)},\sigma_j^2) \sim \mathcal{N}(\mu_j^{(1)}-\mu_j^{(2)},(n_1+n_2)/(n_1n_2)\sigma_j^2)$$
$$\text{et} \quad \hat{\sigma}_j^2|\sigma_j^2 \sim \frac{\sigma_j^2}{\nu}\chi^2_\nu \quad \text{avec} \quad \nu = n_1+n_2-2$$

où $\hat{\sigma}_j^2$ désigne la variance empirique (définie dans l'équation 3.8).

Dans ces conditions, en posant comme estimateur régularisé de la variance son espérance *a posteriori*, on a alors :

$$\tilde{\sigma}_j^2 := \mathbb{E}(\sigma_j^2|\hat{\sigma}_j^2) = \frac{\nu_0\sigma_0^2+\nu\hat{\sigma}_j^2}{\nu_0+\nu} \tag{3.9}$$

On en déduit que le quotient de la différence des moyennes empiriques par condition de la j-ème composante par la variance régularisée :

$$\tilde{t}_j = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \times \frac{\bar{X}_j^{(1)} - \bar{X}_j^{(2)}}{\tilde{\sigma}_j} \tag{3.10}$$

suit une loi de Student à $\nu + \nu_0$ degrés de liberté sous l'hypothèse nulle (3.3).

En pratique, il reste à déterminer les hyperparamètres ν_0 et σ_0^2. Ils sont obtenus par estimation à partir des deux premiers moments des $\ln(\hat{\sigma}_j^2)$:

$$\mathbb{E}(\ln \hat{\sigma}_j^2) = \ln(\sigma_0^2) + \psi(\nu/2) - \psi(\nu_0/2) + \ln(\nu_0/\nu) \quad \text{et} \quad \text{var}(\ln \hat{\sigma}_j^2) = \phi(\nu/2) + \phi(\nu_0/2)$$

où ψ et ϕ désignent les fonctions digamma et trigamma respectivement.

3.2.4 Approches proposées

Statistique de Hotelling avec matrice de variance/covariance diagonale

Dans le contexte de la grande dimension, la statistique de Hotelling dite *diagonale* est une approche privilégiée (Dong et collab., 2016) car elle permet de simplifier l'estimation de Σ. Dans notre cas, elle consiste à supposer des covariances nulles entre les différentes composantes des vecteurs de longueurs de chemins. Avec cette hypothèse de structure diagonale, la statistique de Hotelling s'écrit de la façon suivante :

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} \sum_{j=1}^{p} \frac{(\bar{X}_j^{(1)} - \bar{X}_j^{(2)})^2}{\hat{\sigma}_j^2}$$

Cependant, comme expliqué dans les sections précédentes, le fait que $n \ll p$ peut entraîner des difficultés lors de l'estimation des σ_j^2 avec des variances trop petites qui engendrent une instabilité de la statistique de Hotelling. Une idée pour lever ces difficultés est donc d'utiliser une version régularisée des estimateurs, comme cela est fait dans Dong et collab. (2016). Nous définissons donc une statistique de Hotelling diagonale régularisée de la façon suivante :

$$\tilde{T}^2 = \frac{n_1 n_2}{n_1 + n_2} \sum_{j=1}^{p} \frac{(\bar{X}_j^{(1)} - \bar{X}_j^{(2)})^2}{\tilde{\sigma}_j^2} = \sum_{j=1}^{p} \tilde{t}_j{}^2 \tag{3.11}$$

où $\tilde{\sigma}_j^2$ est la variance régularisée définie par l'équation (3.9).

On voit alors que $\tilde{T}^2$ est la somme des p carrés des variables $\tilde{t}_j$, qui suivent une loi de Student de paramètre $\nu_0 + \nu$ d'après la section 3.2.3. Autrement dit, $\tilde{T}^2$ est une somme de variables de Fisher $F(1, \nu_0 + \nu)$.

Cependant, on rencontre plusieurs difficultés avec cette approche en pratique. D'une part, l'hypothèse des covariances nulles n'est pas vérifiée pour des objets tels que les arbres car les vecteurs de longueurs de chemin ont des structures très particulières. Plus précisément, étant donné un vecteur de longueurs de chemins d'un arbre à B feuilles, ses $p = B(B-1)/2$ composantes ne peuvent prendre au plus que $B-1$ valeurs distinctes, avec des plages de valeurs constantes situées en des endroits similaires entre vecteurs. En effet, la longueur d'un chemin entre deux feuilles est déterminée par la hauteur du plus haut nœud interne traversé par ce chemin, et il n'y a que $B-1$ nœuds internes dans un arbre binaire à B feuilles. La figure 3.9 permet de mettre en évidence les structures de dépendance existantes à travers la visualisation d'un extrait de la matrice de corrélation pour un ensemble de vecteurs de distances cophénétiques obtenus sous hypothèse nulle simulée (voir la section 3.3).

D'autre part, cet effet de dépendance entre les différents termes de la somme est potentiellement renforcé par le fait que les $\tilde{\sigma}_j^2$ ne sont pas indépendants par construction. Cette propriété rend cette statistique régularisée difficilement exploitable en pratique : d'une part, on ne peut pas utiliser de loi théorique pour déterminer les p-valeurs ; d'autre part, obtenir une distribution empirique par permutation n'est pas envisageable lorsque les effectifs des conditions sont faibles.

Pour ces raisons, il est nécessaire de trouver une alternative plus permissive en termes de dépendance entre les composantes des vecteurs de longueurs de chemins.

Agrégation de Simes des p-valeurs individuelles

On introduit ici une méthode capable de lever la difficulté liée au caractère dépendant des différents termes de la somme de $\tilde{T}^2$. En effet, les statistiques individuelles $\tilde{t}_j$ (définie par l'équation 3.10) qui sont les termes composant $\tilde{T}^2$ dans l'équation (3.11), suivent une loi de Student comme expliqué dans la section 3.2.3. On peut donc construire un vecteur de p-valeurs individuelles $\{p_j\}_{1\leq j\leq p}$ correspondant aux tests des hypothèses $\{\mathcal{H}_j^0\}_{1\leq j\leq p}$ par ces statistiques individuelles. L'idée est ensuite d'agréger ces p-valeurs pour construire une p-valeur correspondant au test de l'hypothèse nulle d'intersection $\mathcal{H}_0 = \cap_{j=1}^p \mathcal{H}_j^0$, associée à la comparaison globale des vecteurs (équation (3.5)). Nous proposons pour ce faire d'utiliser la méthode de Simes (1986), qui définit une p-valeur agrégée comme suit :

$$p_{\text{Simes}} = \min\left\{p\frac{p_{(j)}}{j}; j = 1, \ldots, p\right\} \tag{3.12}$$

où $\{p_{(j)}\}_{1\leq j\leq p}$ est le vecteur des p-valeurs rangées dans l'ordre croissant. Cette méthode a par exemple été utilisée par Lun et Smyth (2014) pour l'analyse de données ChIP-seq. Le test associé est valable non seulement lorsque les p-valeurs individuelles sont indépendantes (Simes, 1986), mais également sous une hypothèse

moins restrictive de dépendance positive appelée Positive Regression Dependency on a Subset (PRDS) (Benjamini et Yekutieli, 2001). Bien que cette hypothèse soit difficilement vérifiable formellement en pratique, elle est communément admise en génomique ; en particulier, c'est la condition de dépendance la plus faible sous laquelle la procédure de Benjamini et Hochberg (1995) contrôle le False Discovery Rate (Benjamini et Yekutieli, 2001).

3.3 Validation

Pour pouvoir évaluer la pertinence d'une méthode de test, et notamment si les hypothèses sous-jacentes sont réalistes et si la méthode contrôle correctement le taux de fausses découvertes dans la pratique, l'approche standard consiste à simuler des données sous $\mathcal{H}_0$ ou $\mathcal{H}_1$ et à étudier le comportement de la méthode proposée sur ces simulations. Cependant, il est très difficile de procéder de cette façon dans notre cas car il n'y a pas de cadre clair et naturel de simulations pour les arbres aléatoires. De nombreuses distributions d'arbres ont été étudiées dans la littérature (voir Steel et Penny (1993) et Paradis (2011) par exemple), avec des propriétés très différentes en fonction du modèle de génération aléatoire choisi. Le problème pour exploiter ces travaux réside dans le fait qu'étant donné un contexte applicatif, il est extrêmement difficile d'avoir des arguments rigoureux pour justifier la sélection de l'une de ces distributions comme modèle. Comme les résultats de l'analyse dépendent fortement de ce choix, cela rend cette approche problématique.

Nous choisissons donc de baser notre simulation sur l'utilisation de données GWAS (Genome Wide Association Study) pour se placer dans un cadre d'hypothèse nulle réaliste tout en évitant le problème de la sélection d'un modèle aléatoire d'arbres.

3.3.1 Données GWAS et procédure de simulation

Les données GWAS permettent de cartographier les variations de séquence pour un individu (correspondant à un organisme vivant) par rapport à un génome de référence en collectant des occurrences de versions d'allèles minoritaires le long du génome (un allèle correspond à une certaine version d'un gène). À chaque individu, est associée une suite dans $\{0, 1, 2\}$. Le j-ème terme de cette suite indique si la position j comporte 0, 1 ou 2 occurrences de l'allèle minoritaire chez l'individu considéré, par rapport au génome de référence de la population. Les positions génomiques sur lesquelles ont lieu ces variations de séquence sont appelées Single Nucleotide Polymorphisms (SNP) car elles ne concernent le changement que d'une seule base. Ces données sont fréquemment utilisées pour calculer le déséquilibre de liaison (LD), qui correspond à une mesure de corrélation entre ces paires de

position (voir Ambroise et collab. (2019) pour une définition précise). Un intervalle du génome qui présente un grand déséquilibre de liaison par rapport au reste du génome est l'indication que les SNPs correspondants ont une tendance plus forte à être transmis conjointement par les parents que ce qui serait attendu si leur transmission était indépendante. Ces intervalles sont généralement stables pour l'ensemble de la population et organisés hiérarchiquement (voir la figure 3.6). Dans cette section, nous utilisons des variations de ces données obtenues par sous-échantillonnage d'une population pour évaluer la pertinence de notre méthode.

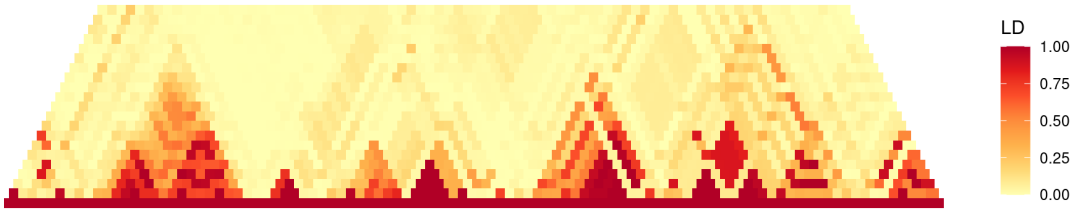


FIGURE 3.6: Heatmap d'une matrice de déséquilibre de liaison. Seule la partie triangulaire supérieure est représentée pour des raisons de symétrie. L'axe horizontal correspond aux SNPs dans l'ordre induit par le génome. La matrice est tronquée pour ne représenter que les interactions entre positions possédant une distance inférieure à un certain seuil.

Plus précisément, nous avons utilisé un fichier de données du projet international HapMap (The International HapMap Consortium, 2003) portant sur 603 SNPs contigus couvrant une région d'une mégabase sur le chromosome 22, pour un ensemble de 90 Européens. Cet ensemble de données a été obtenu à partir du package R **snpStats**. Nous avons échantillonné uniformément 60% des individus de l'échantillon initial n fois pour obtenir n sous-populations de la population d'origine. Nous avons également sélectionné une région de B SNPs contigus à partir desquels nous avons construit n matrices de déséquilibre de liaison. Ces matrices, correspondant à des similarités, ont ensuite été utilisées comme entrée pour la CAH contrainte avec lien de Ward, à l'aide du package R **adjClust**, afin de représenter leur structure hiérarchique par des arbres (Ambroise et collab., 2019). Nous avons ainsi obtenu n arbres à B feuilles (correspondant aux SNPs), que nous avons séparés arbitrairement en deux groupes de même cardinal, sur lesquels notre test a été effectué. Ce processus de simulation a été répété 1000 fois, afin de pouvoir évaluer les distributions de la statistique de test et des p-valeurs, et différentes variations des paramètres de simulation (n et B) ont été étudiées. Par construction, il n'y a pas de vraies différences entre les deux groupes, c'est pourquoi les résultats de ces expériences numériques peuvent être utilisés pour vérifier le comportement de la méthode proposée sous l'hypothèse nulle $\mathcal{H}_0$.

3.3.2 Résultats

Les figures 3.7 et 3.8 montrent respectivement la distribution des p-valeurs combinées par agrégation de Simes (définies par l'équation (3.12)) et de la statistique individuelle de comparaison d'une paire de feuilles (définie par l'équation (3.10)) pour 1000 simulations, et différentes valeurs de n et B. Dans la figure 3.8, la distribution théorique a été évaluée par tirage aléatoire selon la loi théorique : loi de Student à $n+\nu_0-2$ degrés de liberté (voir 3.9 pour la définition de ν_0). Puisque ν_0 dépend de la simulation, pour chacune des 1000 simulations, on a tiré 10 variables aléatoires selon la loi théorique de la simulation considérée. C'est la densité empirique de l'ensemble de ces simulations qui est utilisée pour représenter la densité « théorique » attendue, en bleu sur la figure 3.8.

La première conclusion de ces simulations visible dans la figure 3.7, est qu'à l'exception du cas le plus défavorable de simulation (n très petit et B petit), la méthode contrôle correctement le risque de première espèce pour un niveau de risque de 5% et la distribution des p-valeurs ne présente pas de grande divergence avec la distribution uniforme, comme cela est attendu. Un n grand améliore ce contrôle du risque, avec un meilleur contrôle pour les simulations avec un nombre de feuilles B plus grand (le contrôle du risque est similaire pour les simulations avec $n = 40$ et $B = 100$ et les simulations avec $n = 100$ et $B = 20$). Comme on peut le voir dans la figure 3.8, l'adéquation entre la distribution des statistiques individuelles des paires de feuilles et leur distribution théorique sous l'hypothèse nulle est également bonne, avec la meilleure correspondance pour la simulation avec la plus grande valeur de B. Ces résultats montrent donc le bon comportement de notre test.

Pour justifier plus avant notre choix concernant cette approche, notamment le choix de l'agrégation de Simes pour définir la p-valeur combinée, on extrait une partie de la matrice de covariance empirique (ou plutôt de la matrice de corrélation pour en simplifier l'interprétation), associée, pour une feuille fixée (c'est-à-dire un SNP fixé), ℓ, à toutes les entrées de $\widehat{\Sigma}$ correspondant aux paires de feuilles (ℓ, k) et (ℓ, k') pour $k, k' \in \{1, \ldots, B\}$ avec $(k, k' \neq \ell)$. Cela nous a permis d'obtenir une représentation de cette partie de la matrice de corrélation qui est en accord avec l'ordre naturel des SNPs, lui-même induit par celui du génome (l'ordre des paires (ℓ, k) est le même que celui des SNPs k correspondants le long génome). Un exemple de ces extraits de matrices de corrélation est donné dans la figure 3.9, pour la simulation de paramètres $n = 40$ et $B = 100$ et correspondant à la p-valeur médiane parmi les 1000 simulations de mêmes paramètres. La feuille ℓ a été tirée aléatoirement de façon uniforme parmi toutes les feuilles. Comme attendu, cet extrait de la matrice de covariance montre une structure de corrélation très marquée, qui est en accord avec la structure de l'arbre, avec des blocs de corrélations identiques.

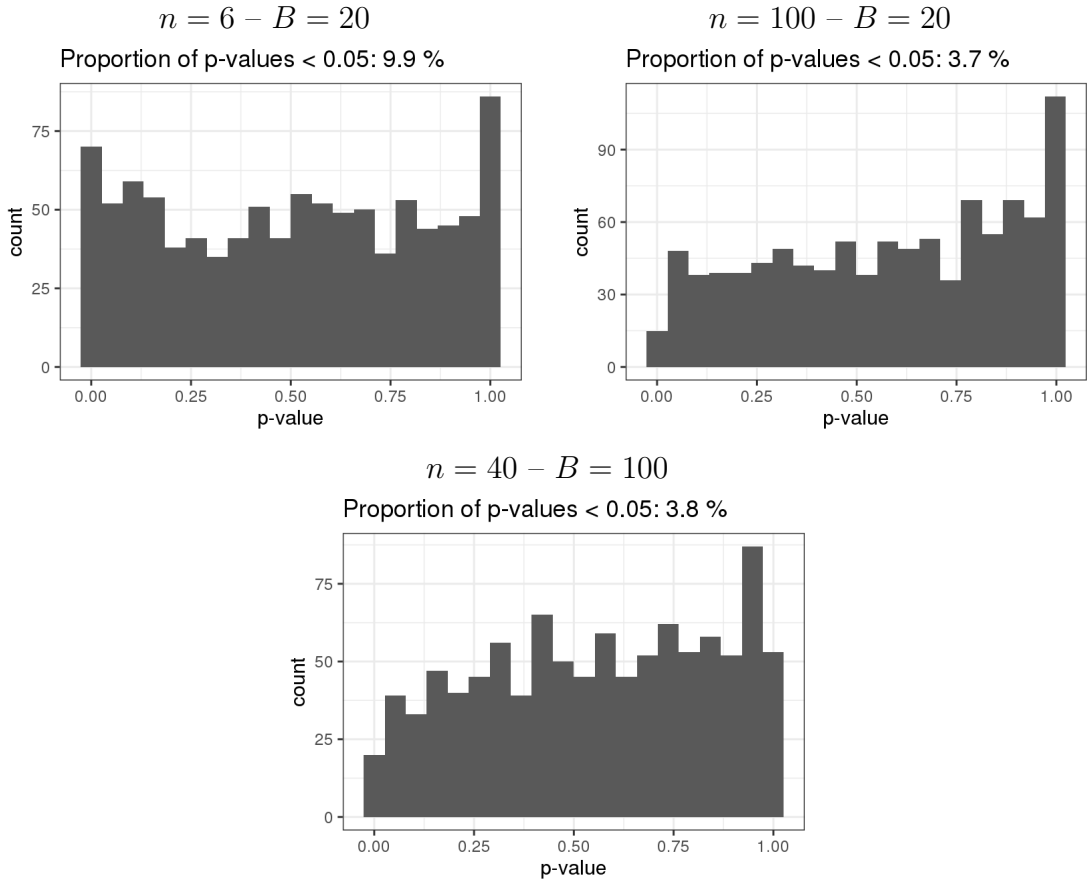


FIGURE 3.7: Distribution des p-valeurs p_{Simes} (définies dans l'équation 3.12) pour l'analyse différentielle d'arbres à partir de la simulation sur données GWAS, pour différentes valeurs de n (nombre total d'arbres dans les deux conditions) et B (nombre de feuilles dans les arbres).

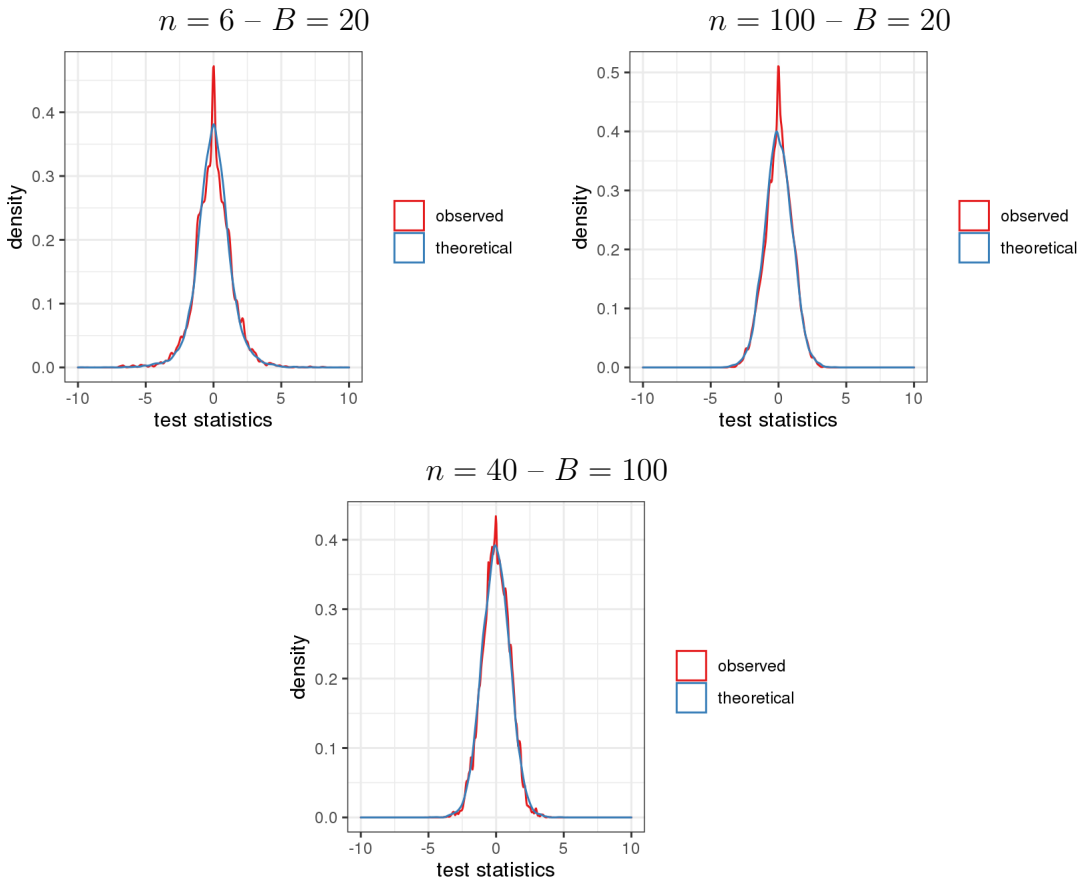


FIGURE 3.8: Distribution empirique des statistiques $\tilde{t}_j$ des paires de SNPs (comme définies dans l'équation (3.10), en rouge) comparée à la distribution théorique sous l'hypothèse nulle (en bleu).

3.3.3 Conclusion

Les données GWAS nous ont permis de tester notre statistique de comparaison de deux ensembles d'arbres sous un cadre d'hypothèse nulle simulé réaliste.
Cette simulation a permis de mettre en évidence le bon comportement de la statistique en termes de distributions empiriques, qu'il s'agisse des p-valeurs associées à la statistique finale agrégée (figure 3.7) ou des statistiques individuelles de comparaison de paires de feuilles (figure 3.8) qui sont cohérentes avec les distributions théoriques attendues. Plusieurs valeurs des paramètres que sont le nombre total d'arbres n et le nombre de feuilles B ont été étudiées et, même dans des cas défavorables ($n = 6$ et $B = 20$ par exemple), les résultats en termes de distributions et de contrôle du risque de première espèce sont satisfaisants.

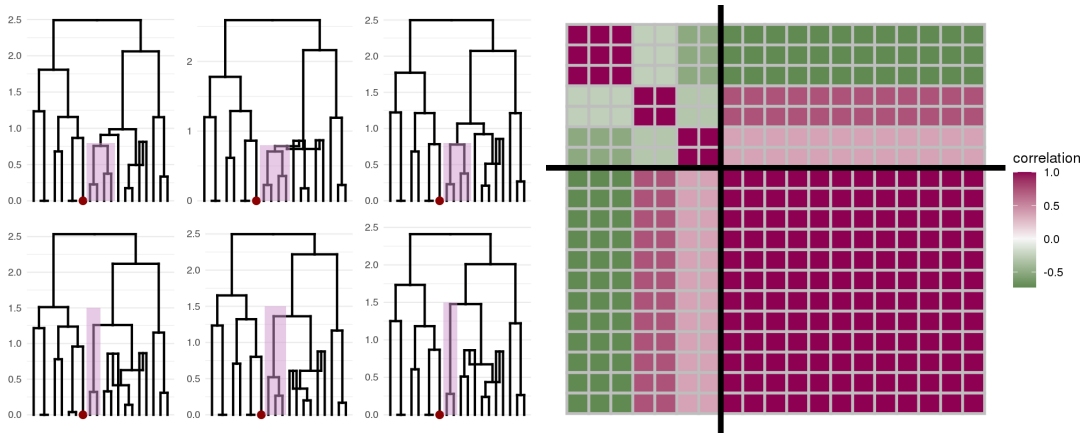


FIGURE 3.9: Gauche : Dendrogrammes obtenus pour une simulation de paramètres $n = 6$ et $B = 20$. Ces dendrogrammes sont ceux amenant à la p-valeur la plus grande parmi celles inférieures à 0,05. Les trois dendrogrammes de la ligne supérieure correspondent à la condition 1 et les trois de la ligne inférieure à la condition 2. Le fond rose met en valeur la partie des arbres contribuant le plus à la significativité de la statistique de test. Droite : Extrait de la matrice de corrélation pour la simulation correspondante. Le point rouge (à gauche) et les lignes noires (à droite) indique la position de la feuille fixée ℓ (voir le texte).

Enfin, cette simulation a permis de souligner les structures particulières de corrélation existant pour les vecteurs de distances cophénétiques (figure 3.9). Ces corrélations marquées confortent le choix d'une p-valeur associée à la comparaison globale obtenue par agrégation de Simes. En effet, le niveau de contrôle de la statistique est garanti par des hypothèses de dépendance positive entre les statistiques individuelles (PRDS, Benjamini et Yekutieli (2001)) relativement permissives et donc, mieux adaptées au cas des vecteurs de distances cophénétiques.

3.4 Applications

3.4.1 Application aux arbres phylogénétiques

Cette sous-section a pour but de démontrer la possibilité d'appliquer avec pertinence la méthode développée dans la section 3.2 à des arbres phylogénétiques.

La phylogénie est une sous-branche de la taxonomie, dont l'objet est l'étude des liens de parentés entre différentes espèces dans le but de les classer de façon hiérarchisée (genres, familles, ordres, ...). On peut comprendre de façon synthétique cet objectif comme la recherche d'un « arbre généalogique des espèces », comme illustré dans la figure 3.10. À l'origine, ces classifications étaient établies sur la base de comparaisons de traits physiques entre organismes (comparaisons de *phénotypes*). Ainsi, plus des espèces présentent de différences l'une avec l'autre, plus l'événement dit de spéciation (mécanisme de différentiation des espèces) à l'origine de la divergence entre ces deux espèces est considéré comme ancien. La représentation sous forme d'arbre permet de représenter cette information à travers la hauteur du premier nœud regroupant les deux espèces en question (correspondant au premier ancêtre commun), et de faire cela pour tout un ensemble d'espèces simultanément.

La phylogénie moléculaire permet d'aborder cette question en utilisant l'information que constituent les séquences génétiques des organismes. Plus précisément, pour construire l'arbre évolutif d'un ensemble d'organismes, on remplace la comparaison de traits physiques par la comparaison de séquences biologiques, comme des gènes par exemple, que l'on retrouve chez toutes les espèces considérées, mais avec des variations propres aux espèces. Ces gènes « communs » permettent de comparer, du point de vue des séquences, les organismes qui les possèdent. La valeur de la distance entre deux séquences renseigne sur l'événement de spéciation entre les deux espèces associées à chacune des séquences. On peut ainsi essayer de déterminer les liens évolutifs existant entre l'ensemble des organismes considérés, et la représentation synthétique du résultat pour toutes les espèces se fait sous la forme d'un *arbre phylogénétique* comme illustré dans la figure 3.11, dont les feuilles sont les organismes et la topologie représente l'histoire évolutive induite par le gène commun pour ces organismes.

Avec l'essor des études de phylogénie moléculaire, des divergences ont été mises en évidence entre les arbres phylogénétiques établis à partir de gènes différents. Nous proposons d'appliquer notre méthode de comparaison à des dendrogrammes issus de la phylogénie moléculaire, afin d'essayer de détecter des différences dans les arbres impliqués par différentes familles de gènes pour un même ensemble d'organismes.

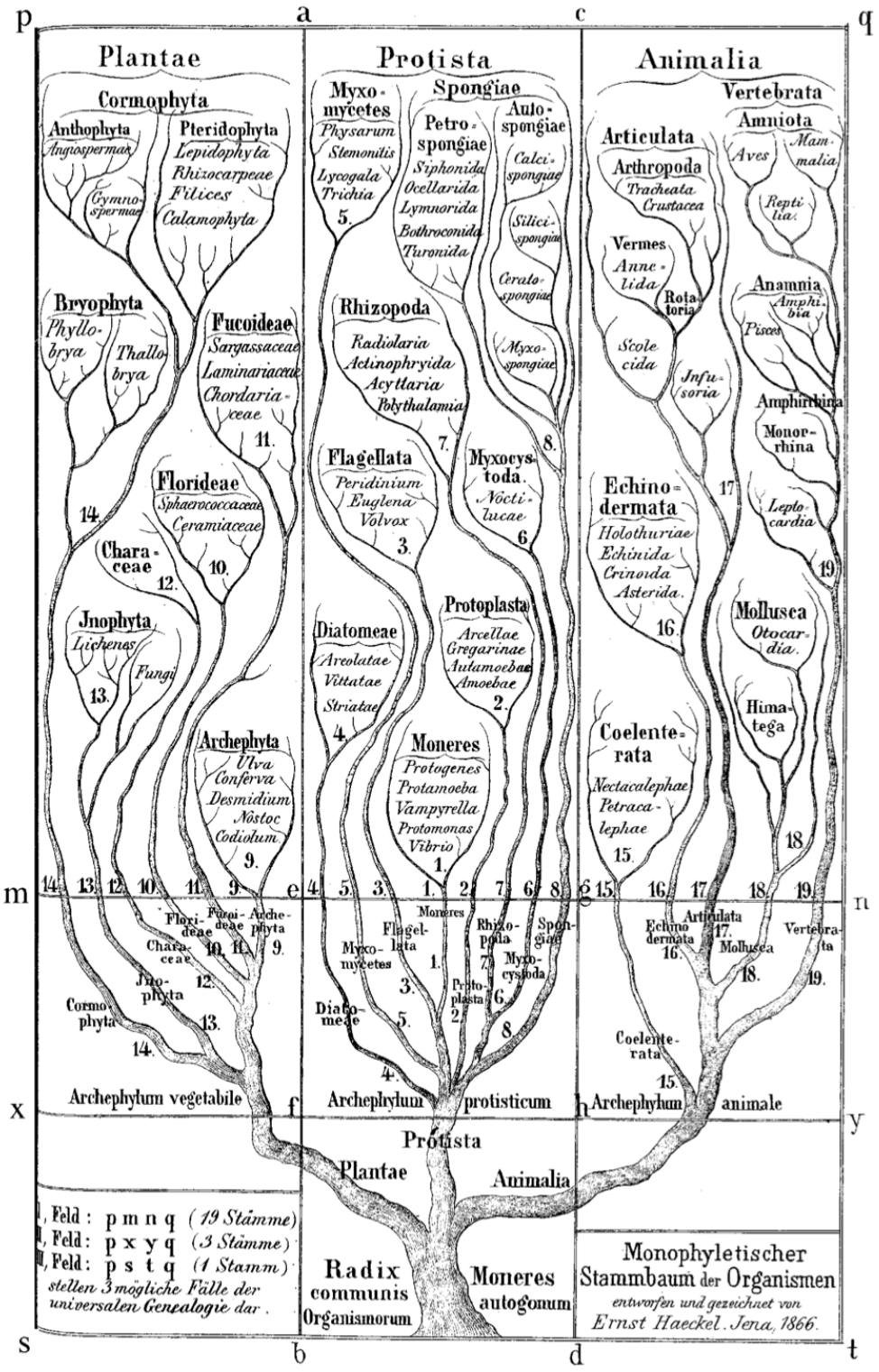


FIGURE 3.10: **Arbre généalogique monophylétique des organismes de Ernst Haeckel (1866).** Représentation sous forme d'arbre d'une tentative de reconstitution de la filiation des espèces. On note par exemple, la première division entre plantes, animaux et protistes (classe obsolète regroupant certains organismes unicellulaires). L'arbre est dit *monophylétique* car tous les organismes sont ici issus d'un même ancêtre commun représenté par la racine de l'arbre.

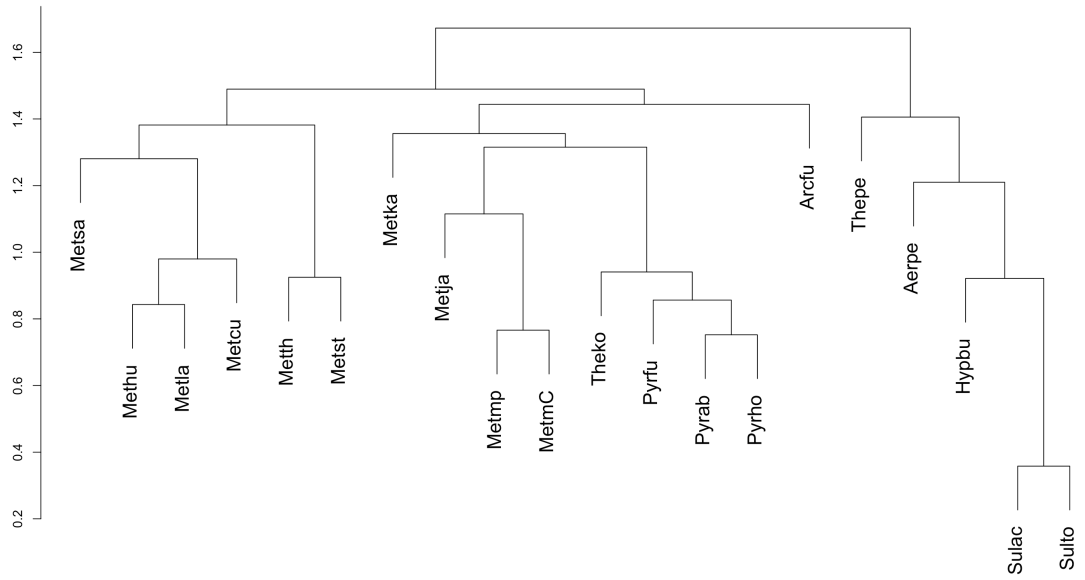


FIGURE 3.11: **Exemple d'arbre phylogénétique.** L'arbre porte sur 20 espèces d'archées. L'ordonnée des nœuds est calculée par la méthode de construction de l'arbre et s'interprète comme une distance entre les espèces que le nœud réunit. Arbre obtenu à partir des données de l'article de Puigbò et collab. (2009).

Présentation des données

Les arbres phylogénétiques considérés ici proviennent de l'article de Puigbò et collab. (2009) dans lequel les auteurs s'intéressent à la possibilité de retrouver l'histoire évolutive commune de façon cohérente à partir d'arbres phylogénétiques provenant de gènes différents. Dans leur article, Puigbò et collab. (2009) étudient un ensemble de 100 espèces d'organismes unicellulaires (plus précisément deux types de *procaryotes* : 41 archées et 59 bactéries), pour lesquels ils disposent de 6901 arbres phylogénétiques partiels (chaque arbre ne porte que sur un sous-ensemble des 100 procaryotes considérés). En particulier, ils disposent de 102 arbres dont les feuilles sont constituées de plus de 90 des 100 organismes considérés. C'est sur ces derniers que se concentre notre application.

Ces arbres sont classés en groupes appelés *classes fonctionnelles* (indiqués par une lettre) correspondant à la fonction des gènes ayant servi à leur élaboration. La correspondance entre chaque arbre et la classe fonctionnelle associée est donnée dans l'article de Puigbò et collab. (2009). Ainsi, à chaque classe fonctionnelle de gènes correspond un groupe d'arbres phylogénétiques. Ici, nous allons considérer trois classes fonctionnelles de gènes : les classes F, J et L.

— La classe F est associée à la classe fonctionnelle « Nucleotide transport and metabolism » et contient **10** arbres phylogénétiques.
— La classe J est associée à la classe fonctionnelle « Translation, ribosomal structure and biogenesis » et contient **54** arbres phylogénétiques.
— La classe L est associée à la classe fonctionnelle « Combination, Recombination and Repair » et contient **5** arbres phylogénétiques.

Les feuilles des arbres phylogénétiques étudiés correspondent donc soit à des archées, soit à des bactéries. D'après Puigbò et collab. (2009), il est attendu que les arbres obtenus à partir de certaines familles fonctionnelles de gènes séparent bien les deux groupes « bactéries » et « archées » : c'est le cas des classes classes J et L par exemple. Tandis que pour d'autres, comme la classe F, bactéries et archées seraient moins bien séparées. Du point de vue de l'évolution, la séparation des bactéries et des archées est attestée par la littérature (Gribaldo et Brochier-Armanet, 2006).

Les arbres phylogénétiques associés à ces classes sont représentés dans la figure 3.12. On peut constater la bonne séparation des bactéries et des archées pour les classes J et L tandis qu'elles sont relativement mélangées dans la classe F.

On peut également constater cette propriété de séparation des groupes bactéries et archées dans les arbres phylogénétiques en comparant les matrices de distances cophénétiques moyennes par classes fonctionnelles, présentées dans la figure 3.13. On voit clairement une séparation pour les classes J et L tandis que pour la classe F, la séparation est beaucoup moins nette.

L'idée générale de cette application est donc de comparer la valeur de la statistique de test pour une comparaison des classes F et J, et une comparaison des classes L et J. D'un point de vue biologique, on s'attend à ce que les classes J et L aient des arbres phylogénétiques relativement semblables. À l'inverse, l'histoire évolutive induite par les gènes de la classe F devrait différer de celle de la classe J. En particulier, on sait que la séparation entre bactéries et archées y est moins nette, et on souhaite mettre en évidence la capacité de notre statistique à détecter ce type de différence.

Comparaisons des histoires génétiques induites par les classes F, J et L

Prétraitement des arbres. Les arbres considérés sont partiels dans la mesure où ils ne portent que sur des sous-ensembles des 100 espèces de procaryotes considérées. La méthode développée dans la partie 3.2 ne s'applique qu'à des arbres établis sur le même ensemble de feuilles. Une solution pour rendre possibles les comparaisons entre des arbres n'ayant pas exactement les mêmes feuilles mais qui en partagent un certain nombre, peut consister à élaguer les arbres à un ensemble de feuilles communes à tous les arbres intervenant dans la comparaison. Un choix naturel est donc de considérer l'intersection des feuilles sur l'ensemble des arbres

concernés. Ainsi, bien que les arbres des classes F, J et L retenus soient des arbres portant sur plus de 90 espèces parmi les 100 considérées, l'intersection des feuilles de ces 69 arbres mène à un ensemble de feuilles communes de 55 procaryotes (22 archées et 33 bactéries). Dans la suite, pour chaque arbre phylogénétique, on considèrera donc sa version restreinte à ces 55 feuilles.

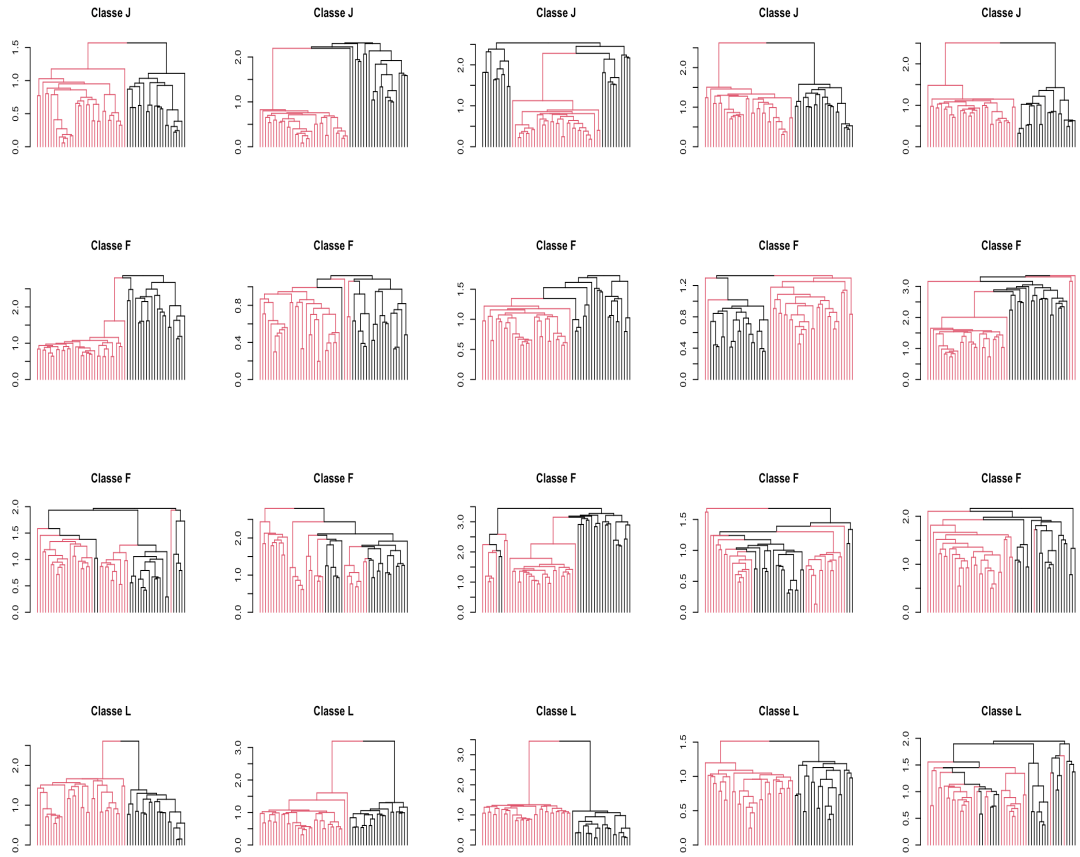


FIGURE 3.12: **Arbres phylogénétiques associés aux différentes classes.** Les arbres phylogénétiques sont représentés avec les branches associées aux bactéries en rouge et celles menant aux archées en noir. La première ligne correspond à 5 arbres choisis aléatoirement parmi la classe fonctionnelle J. La deuxième et la troisième ligne correspondent aux arbres de la classe fonctionnelle F, et la quatrième à ceux de la classe fonctionnelle L. Une bonne séparation entre bactéries et archées est notamment caractérisée par une absence de mélange entre les branches rouges et noires. Ainsi, on peut voir que les classes J et L séparent globalement bien bactéries et archées tandis que la classe F possède une majorité d'arbres qui comportent des mélanges.

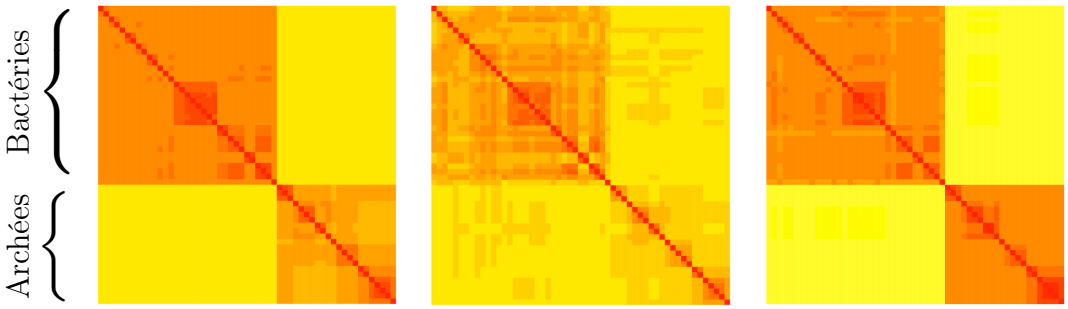


FIGURE 3.13: **Heatmaps des matrices de distances cophénétiques moyennes par classe fonctionnelle.** L'échelle de couleur est identique pour les 3 matrices. Les deux blocs diagonaux foncés observables sur la matrice de gauche (classe J) et de droite (classe L) correspondent aux sous-matrices associées respectivement aux bactéries et aux archées. Pour la matrice centrale (classe F), la séparation est moins marquée : on voit des mélanges (motifs en damiers) et les deux classes apparaissent beaucoup moins bien séparées, ce qui se traduit par un contraste moins fort entre les deux blocs diagonaux et anti-diagonaux.

Un deuxième aspect du prétraitement concerne la normalisation des arbres phylogénétiques. Dans le cas des arbres considérés ici, les racines et les feuilles présentent souvent de grandes différences en termes de hauteurs, parfois au sein d'une même classe fonctionnelle. On peut voir ces différences sur la figure 3.12 par exemple, où le second arbre de la seconde ligne (classe F) et le cinquième présentent un grand écart de hauteur de racine relativement à leur taille. Ces différences ont tendance à en écraser de plus fines, comme des différences de topologie interne par exemple. Ici, il est donc nécessaire d'appliquer une normalisation sur les hauteurs des nœuds internes (c'est-à-dire différents d'une feuille) des arbres, afin de rendre les arbres comparables dans leur ensemble. Cette normalisation consiste en une simple transformation affine des hauteurs afin de rendre le premier nœud de hauteur nulle et la racine de hauteur standard commune à tous les arbres de la comparaison. Plus précisément, en notant $(h_t)_t$ la suite des hauteurs des nœuds internes d'un arbre, la normalisation consiste à définir les $(\tilde{h}_t)_t$ de la façon suivante :

$$\tilde{h}_t = \frac{h_t - \min_r(h_r)}{\max_r(h_r) - \min_r(h_r)} \quad \in \quad [0,1]$$

Cette normalisation est ensuite appliquée globalement à l'ensemble des arbres intervenant dans la comparaison.

Comparaison des arbres phylogénétiques des classes fonctionnelles F, J et L. La tableau 3.1 présente les résultats obtenus à l'aide de la statistique développée dans la section 3.2 en comparant les 10 arbres phylogénétiques de la classe F avec les 54 de la classe J et en comparant les 5 de la classe L avec les 54 de la classe J.

	p-valeur
Comparaison de la classe F avec la classe J	$6,7 \times 10^{-5}$
Comparaison de la classe L avec la classe J	0,85

TABLE 3.1: **Tableau des p-valeurs obtenues pour les comparaisons des arbres phylogénétiques des classes fonctionnelles F, J et L.**

Les résultats du tableau 3.1 montrent une présence de signal dans la comparaison de la classe F avec la classe J tandis la statistique n'en détecte pas dans la comparaison de la classe L avec la classe J, comme cela est attendu d'un point de vue biologique. La comparaison de la classe F et de la classe L a également été réalisée et donne une p-valeur de 0,13.

On peut se poser la question de l'effet de la forte différence d'effectifs dans la première condition (10 arbres pour la classe F et 5 arbres pour la classe J) sur la significativité du test. Pour répondre à cette question, on a étudié la distribution des p-valeurs obtenues en choisissant 5 arbres parmi les 10 de la classe F et en réalisant la comparaison entre ces 5 arbres avec les 54 de la classe J. Il y a $\binom{10}{5} = 252$ façons de choisir ces 5 arbres dans la classe F. Dans la figure 3.14, on représente la distribution de p-valeurs obtenue pour ces 252 façons de choisir ces 5 arbres lorsqu'on réalise la comparaison avec les 54 arbres de la classe J. Cette distribution suggère que la p-valeur obtenue dans la comparaison des 10 arbres de la classe F avec les 54 de la classe J, donnée dans le tableau 3.1, n'est pas due à la différence d'effectifs de la condition 1 mais à une réelle présence de signal dans les données.

Conclusion

L'application de notre méthode aux classes considérées permet de conclure à la présence d'une différence significative entre les arbres phylogénétiques obtenues à l'aide des gènes de la classe fonctionnelle F et ceux obtenus à l'aide des gènes de la classe J. Cette différence peut s'expliquer par la mauvaise séparation des bactéries et des archées dans les arbres de la classe F (illustrée dans les figures 3.12 et 3.13), comparée à la classe J.

Cependant, l'article de Puigbò et collab. (2009) nous permet d'interpréter cette différence plus précisément. En effet, les différences de topologie entre les arbres de

la classe F et ceux de la classe J s'expliquent notamment par les *transferts horizontaux de gènes*, mécanismes qui permettent des partages de matériel génétique entre individus non apparentés et se produisant particulièrement fréquemment pour les procaryotes. Ainsi, il est attendu que ces transferts soient nombreux pour la classe F contrairement à la classe J ou L, ce qui rend les histoires retrouvées à partir de gènes de la classe F perméables entre bactéries et archées, expliquant ainsi les mélanges entre bactéries et archées présents dans les arbres phylogénétiques de la classe F.

Cette application permet donc de dire que notre méthode est donc d'une part, applicable avec pertinence à des données phylogénétiques, et d'autre part, capable de retrouver des différences attendues d'un point de vue biologique.

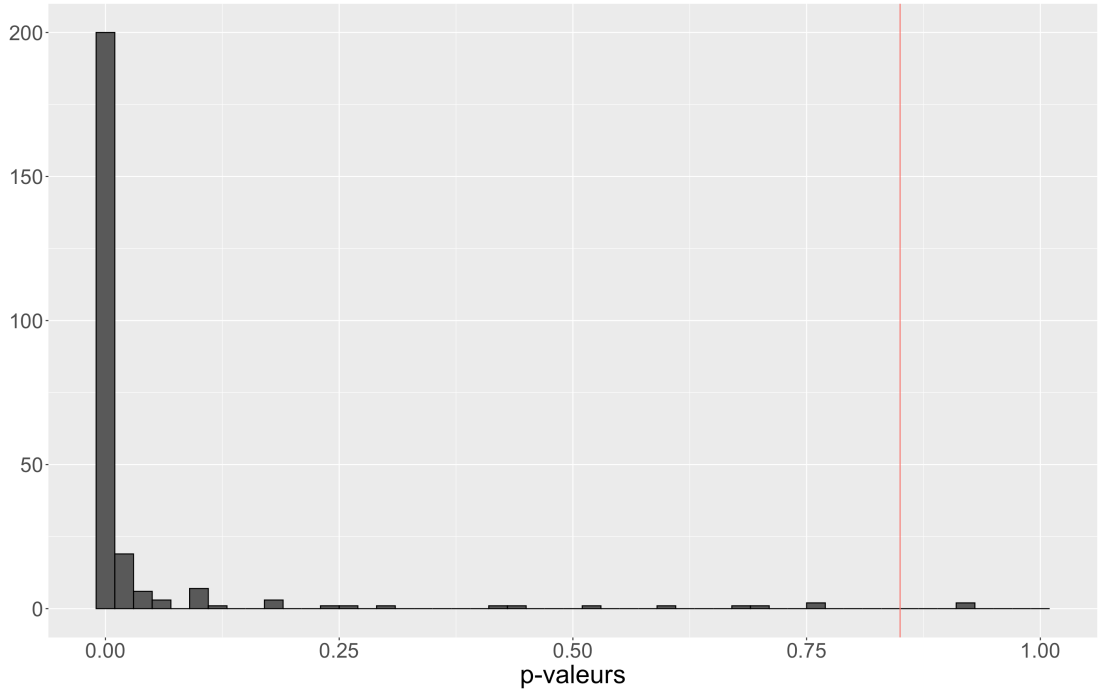


FIGURE 3.14: **Histogramme de la distribution des p-valeurs dans la comparaison d'un ensemble de 5 arbres de la classe F avec les 54 arbres de la classe J.** L'histogramme représente les 252 p-valeurs correspondant à chacune des combinaisons de 5 arbres parmi les 10 de la classe F. La ligne rouge correspond à la p-valeur de la comparaison de la classe L avec la classe J.

3.4.2 Application sur données Hi-C

Cette sous-section a pour but de démontrer la possibilité d'appliquer avec pertinence la méthode développée dans la section 3.2 sur des données Hi-C.

Présentation des données

Pour cette application, nous avons utilisé les données décrites dans Marti-Marimon et collab. (2021). Elles sont composées de matrices Hi-C provenant de 6 expériences sur les tissus musculaires de 6 fœtus de cochons, chacune de ces expériences ayant conduit à 18 matrices Hi-C intra-chromosomique [1], $(H^{c,i})_{c=1,\dots,18,\, i=1,\dots,6}$. Les coefficients de ces matrices $H^{c,i}$ correspondent au nombre de comptages d'interactions physiques entre bins d'une longueur de $200,000$ paires de bases (200 kb). Les comptages représentent le nombre de fois que les bins d'une paire donnée ont été observés en contact dans l'expérience (voir la section 1.2 pour plus de détails). De plus, les fœtus de cochons considérés sont répartis en deux conditions biologiques en fonction de leur stade de développement : 3 échantillons correspondent à 90 jours de gestation et les 3 autres à 110 jours de gestation (soit approximativement à la fin de la gestation). Il est déjà connu que des changements métaboliques majeurs se produisent entre ces deux stades pour permettre la pleine maturité des porcelets à la naissance et notamment leur survie (Voillet et collab., 2014; Lefort et collab., 2020).

Jusqu'à présent, les différences de conformation de la chromatine pour ces données n'ont été étudiées qu'au travers de comparaisons basées sur les paires de bins, les TADs ou les compartiments A/B (voir la section 1.1.2 de l'introduction pour la définition de ces différents niveaux d'organisation) et des analyses différentielles indépendantes ont été réalisées pour chacun de ces niveaux d'organisation (Lun et Smyth, 2015). Bien qu'intéressante, cette approche ne répond pas entièrement à la question d'identifier des modifications dans la structure tridimensionnelle du génome.

Procédure d'analyse différentielle

Les étapes de l'analyse différentielle que nous avons menée sur ces données sont les suivantes :

1. Pour $c = 1, \dots, 18$, les 6 matrices ont été normalisées pour corriger les différences de profondeur de séquençage entre les différentes expériences, i, en utilisant une normalisation MA (décrite dans la section 1.3.1). On a ainsi obtenu des matrices corrigées, $(\tilde{H}^{c,i})_{c=1,\dots,18,\, i=1,\dots,6}$, dont les coefficients sont comparables entre deux expériences i et i' ;
2. Pour $c = 1, \dots, 18$, on calcule la log-transformation de la matrice totale $\tilde{H}^c = \sum_{i=1}^{6} \tilde{H}^{c,i}$ et on l'utilise comme entrée de classification ascendante hiérarchique avec contrainte d'ordre (et avec lien de Ward) afin d'obtenir un dendrogramme consensus à l'aide du package R **adjClust**. Ce dendrogramme

1. Le cochon a 2×19 chromosomes mais ceux de cette expérience n'étant pas du même sexe, on a omis les chromosomes sexuels.

est utilisé pour déterminer une partition du chromosome c en K_c classes de bins contigus notées $(G_k^c)_{k=1,\dots,K_c}$ (à partir de l'heuristique Broken Stick (Bennett, 1996)) ;

3. Pour $c = 1,\dots,18$, $k = 1,\dots,K_c$, et $i = 1,\dots,6$, les sous-matrices $\tilde{H}_k^{c,i} = (\tilde{h}_{j,l}^{c,i})_{j,l\in G_k^c}$ (avec $\tilde{H}^{c,i} = (\tilde{h}_{j,l}^{c,i})_{j,l}$) sont log-transformées puis utilisées comme entrée de la CAHCO. Notre méthode d'analyse différentielle est ensuite appliquée aux 6 dendrogrammes obtenus par chromosome c et par classe G_k^c. Une régularisation des variances est ensuite réalisée à l'échelle du génome et on construit une p-valeur par agrégation des p-valeurs individuelles comme décrit dans les équations (3.10) et (3.12), pour chaque classe G_k^c. On obtient ainsi $\sum_{c=1}^{18} K_c = 743$ p-valeurs. Les résultats présentés ci-dessous sont obtenues en appliquant la correction de tests multiples de Benjamini et Hochberg (1995) aux p-valeurs précédentes afin de contrôler le FDR.

Le code utilisé est disponible à https://forgemia.inra.fr/scales/differential-analysis-of-trees.

Résultats

Pour mieux évaluer le comportement de notre procédure de test, nous avons également réalisé notre test en permutant les étiquettes des conditions biologiques.Les distributions des p-valeurs brutes et ajustées pour les données d'origines et les données permutées sont présentées dans la figure 3.15.

Ces résultats montrent que notre procédure contrôle correctement le FDR avec approximativement 1,7% de tests déclarés positifs à partir des p-valeurs ajustées pour les données permutées au niveau de risque 5%. De plus, les résultats positifs identifiés sur les données d'origine correspondent approximativement au tiers (35,4%) du génome, ce qui est en accord avec les découvertes précédentes concernant la conservation des TADs (Dixon et collab., 2015). La même analyse, réalisée sur des données de résolutions différentes (longueur de bins de 40kb et 500kb), a montré des conclusions similaires, avec une plus petite fréquence de détection pour la plus grande résolution. Celle-ci est dûe à des classes plus petites et donc moins fiables (10% de classes déclarées différentielles sur l'ensemble du génome). Dans tous les cas, le FDR est proprement contrôlé dans le contexte des données permutées (résultats non montrés).

Certains des arbres détectés comme différentiels par notre analyse sont donnés dans la figure 3.16 en utilisant la représentation sous forme de dendrogrammes des sous-matrices Hi-C correspondantes.

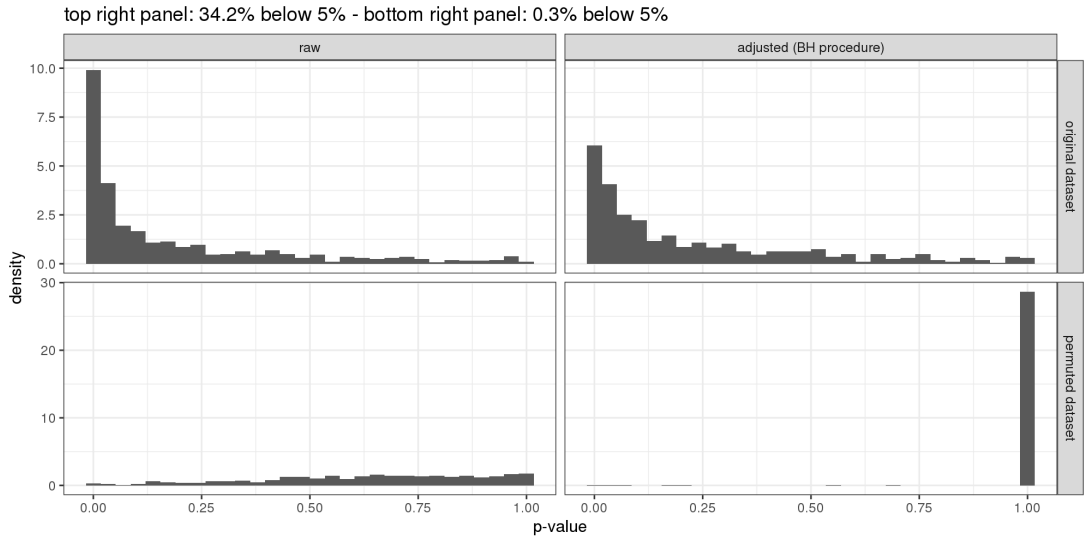


FIGURE 3.15: **Distribution des p-valeurs pour l'analyse différentielle d'arbres issus de données Hi-C** : données d'origine (ligne du haut) et données permutées (ligne du bas), p-valeurs brutes (gauche) et p-valeurs ajustées (droite). 34,5% et 0,3% des ensembles d'arbres testés ont été détectés comme différentiels, respectivement pour les données d'origine et les données permutées avec un niveau de risque de 5%

Ces résultats montrent que le test est en effet capable d'identifier des différences de structures mais peut aussi donner une information supplémentaire en détectant les paires de feuilles contribuant le plus à cette différence. En effet, il attribue plus d'importance d'un point de vue statistique, c'est-à-dire des p-valeurs individuelles p_j plus petites (associées aux statistiques individuelles définies par l'équation (3.10)), aux changements qui concernent une grande partie des arbres (par exemple, se produisant à des niveaux proches de la racine) contrairement à des changements qui affectent seulement une petite partie des arbres (se produisant à des niveaux bas, proches des feuilles). Ainsi, dans la partie **(A)** de la figure 3.16, on voit que la différence entre les deux ensembles d'arbres est principalement dûe à un décalage de la frontière située entre la feuille 5 et la feuille 6 pour les arbres correspondant à 90 jours de gestation vers une frontière entre les feuilles 7 et 8 pour les arbres à 110 jours de gestation. Dans la partie **(B)** de la figure 3.16, la différence est majoritairement dûe à un décalage de frontière également : le groupe constitué des feuilles 17, 18 et 19 change de côté par rapport à la frontière entre les deux conditions.

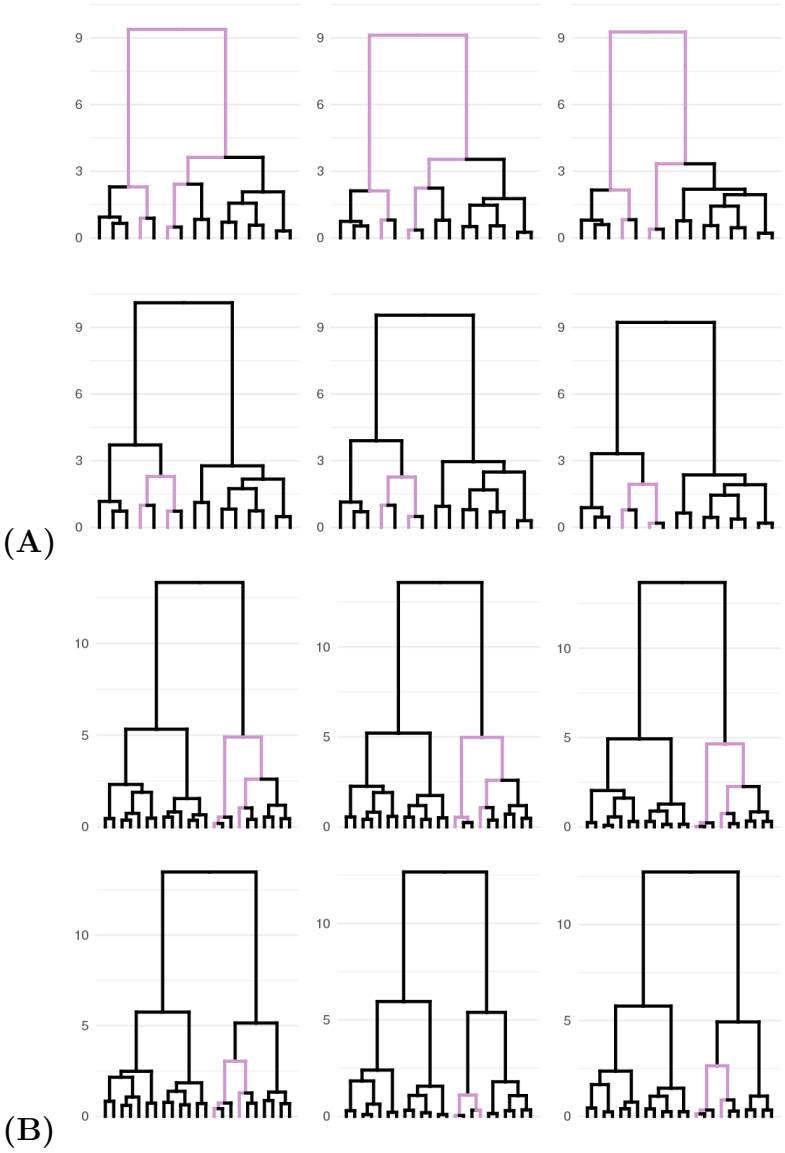


FIGURE 3.16: Deux groupes ((A) et (B)) de deux ensembles d'arbres significativement différents selon notre méthode. **(A)** : Ensemble d'arbres correspondant à la p-valeur ajustée minimale sur l'ensemble du génome ($2,01 \times 10^{-5}$) et **(B)** : au troisième quartile des p-valeurs ajustées ($1,35 \times 10^{-2}$). Dans les deux figures ((A) et (B)), la ligne du haut est composée des arbres correspondant à 90 jours de gestation et celle du bas, aux arbres correspondant à 110 jours de gestation. La branche surlignée correspond à la paire de bins avec la plus grande statistique individuelle (comme définie dans l'équation (3.10)).

Enfin, pour confirmer la validité de notre test dans ce contexte, on compare la distribution empirique des statistiques individuelles de comparaison des paires de feuilles (définies dans l'équation (3.10)) avec la distribution théorique sous l'hypothèse nulle. Puisque la régularisation des variances a consisté à augmenter de $\nu_0 = 1,17$ et $1,20$ le degré de liberté pour les données d'origine et permutées respectivement, la distribution sous l'hypothèse nulle suit une loi de Student avec un $\nu = 6 + \nu_0 = 7,17$ et $7,20$ respectivement. Par conséquent, les distributions empiriques sont donc comparées avec la distribution de Student de degré de liberté $\nu = 7,17$ (la différence avec l'autre distribution théorique étant négligeable). La figure 3.17 présente cette comparaison.

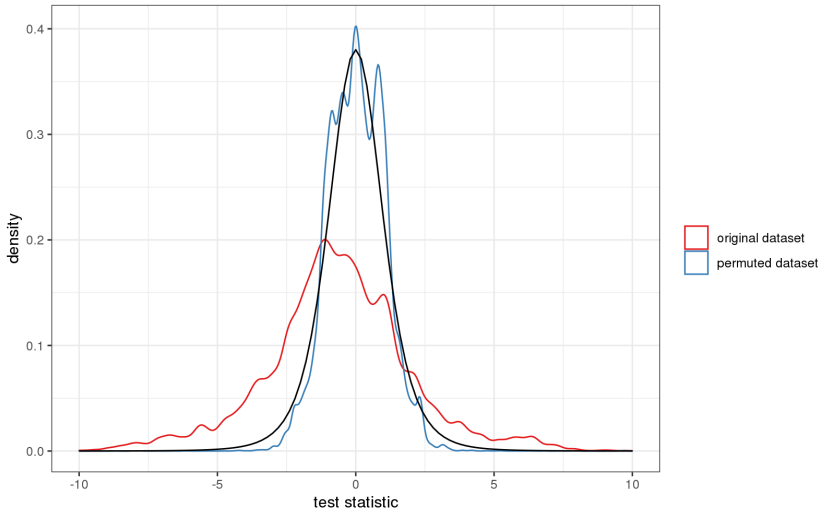


FIGURE 3.17: Distribution empirique des statistiques individuelles $\tilde{t}_j$ associées aux paires de bins (comme définies dans l'équation (3.10)), pour les données d'origine (en rouge) et les données permutées (en bleu), comparée à la distribution théorique sous l'hypothèse nulle (en noir).

La figure 3.17 permet donc de vérifier la bonne adéquation entre la distribution issue des données permutées et la théorie, et donc de donner un argument supplémentaire en faveur de la validité des résultats positifs détectés au moyen de ce test. On note un léger décalage à gauche de la distribution empirique (en rouge sur la figure) par rapport à la distribution théorique. Ceci indique que les distances cophénétiques ont tendance à être plus grandes à 110 jours qu'à 90 jours de gestation, ce qui suggère un plus grand niveau de compaction de la chromatine à 110 jours qu'à 90 jours de gestation dans les régions génomiques correspondantes.

Enfin, les résultats ont été comparés à ceux précédemment obtenus par comparaisons au niveau des paires des bins (Marti-Marimon et collab., 2021). Pour réaliser cette comparaison, une paire de bins a été désignée comme « différentielle » par

notre méthode si elle appartenait à un arbre identifié par notre statistique comme différentiel entre conditions. Plus précisément, les résultats de la précédente analyse différentielle (Marti-Marimon et collab., 2021) sur les paires de bins ont été comparés aux p-valeurs individuelles p_j (telles que définies par l'équation (3.10) et après correction de tests multiples) que nous avons obtenues.

Bien que l'article de Marti-Marimon et collab. (2021) se soit plutôt intéressé à des matrices dont la taille des bins était de 500kb, on a refait les mêmes analyses à 200kb pour être dans la plage de résolution où notre méthode est la plus pertinente. Les deux ensembles de p-valeurs ajustées ont ensuite été classés en résultats positifs (p-valeur ajustée $< 0,05$) et en résultats négatifs. On peut noter qu'à cause de fortes différences entre les deux méthodes, le nombre de résultats positifs pour l'analyse standard est bien plus faible que pour notre méthode (675 contre 12 962 respectivement). Cependant, parmi les résultats négatifs de notre analyse, approximativement 0,5% sont positifs pour l'analyse standard et parmi les résultats positifs de notre analyse, approximativement 2,0% sont positifs pour l'analyse standard. Un test exact de Fisher réalisé sur la table de contingence des résultats positifs/négatifs pour le test standard et celui basé sur les arbres, donne un résultat très significatif (p-valeur $< 2,2 \times 10^{-16}$) qui permet de rejeter l'hypothèse selon laquelle les résultats des deux approches sont indépendants.

Conclusion

Cette application nous permet de vérifier à nouveau mais dans un cadre différent le bon comportement de notre statistique de test. En effet, le fait d'avoir, pour les données permutées, un bon contrôle du FDR d'une part et une forte adéquation de la distribution des statistiques individuelles avec leur distribution théorique sous l'hypothèse nulle d'autre part, nous conforte quant à la validité des résultats positifs obtenus.

De plus, la vérification visuelle de certains résultats positifs confirme la pertinence des différences détectées par la méthode en termes de modifications de structure entre conditions. La représentation graphique met en évidence une caractéristique intéressante de notre approche qui permet d'isoler les paires de feuilles contribuant le plus à la significativité du résultat, conduisant ainsi à une meilleure interprétation de la différence détectée.

Enfin, nos résultats semblent en accord avec les études précédentes. En effet, la proportion de résultats positifs en termes de régions génomiques est en adéquation avec les résultats existant dans la littérature (Dixon et collab., 2015). De plus, les précédentes analyses de ces mêmes données menées par (Marti-Marimon et collab., 2021) montrent aussi une certaine cohérence avec nos résultats, bien que la question motivant leur travail soit légèrement différente de la nôtre, ce qui implique une différence dans le nombre total de différences détectées.

Chapitre 4

Conclusion et perspectives

Cette thèse a été motivée par une problématique biologique : celle de la comparaison de la structure tridimensionnelle du génome entre deux conditions biologiques différentes. Cet objectif a été abordé par l'étude de données Hi-C, mesures de similarités entre positions génomiques, qui renseignent sur la probabilité d'une interaction physique entre deux positions. Ces données présentent une forte structure hiérarchique, expliquée par l'organisation multi-échelle du génome. L'objectif de cette thèse a été de développer une méthodologie permettant de réaliser des comparaisons de conditions biologiques basées sur ces données, en cohérence avec leur structure hiérarchique.

La question de la modélisation de la structure 3D du matériel génétique à partir de matrices Hi-C a été résolue par l'utilisation de la procédure de Classification Ascendante Hiérarchique (CAH). Ce point a donné lieu à une justification des extensions possibles de la CAH (en particulier, aux similarités, qui correspondent au cas des données Hi-C). La version sous contrainte de contiguïté (CAHCC) a été décrite, avec notamment le cas particulier de la contrainte d'ordre (CAHCO). Ces contraintes permettent de restreindre les agrégations possibles à chaque étape de la procédure et d'obtenir des résultats plus cohérents du point de vue d'une réalité terrain connue *a priori* (l'ordre linéaire du génome par exemple). Les différents contextes d'application de la CAH (types de données, avec ou sans contrainte) ont mené à une étude systématique des propriétés des résultats et des représentations graphiques de la CAH (dendrogrammes). Enfin, bien que dans le cas de la version sous contrainte de contiguïté, l'agrégation soit réalisée de manière sous-optimale à chaque étape de la procédure, on a montré qu'une contrainte en adéquation avec la structure intrinsèque des données pouvait améliorer le résultat global de la classification en termes d'inertie intra-classes.

Ce premier point a donc permis de justifier l'utilisation de la CAH sur les données Hi-C afin de modéliser la structure hiérarchique du matériel génétique associée à une matrice Hi-C par un arbre binaire résultant de la CAH. L'étape

suivante a été d'établir une méthode statistique de comparaison de deux ensembles d'arbres. La méthode développée va au-delà du contexte biologique motivant la thèse (données Hi-C) et s'applique à tout domaine faisant intervenir des données sous forme d'arbres.

Dans la seconde partie de la thèse, on a d'abord étudié les manières de quantifier les différences entre arbres à travers une étude bibliographique des distances entre arbres. Ce travail a débouché sur la sélection d'une représentation vectorielle des arbres par leur vecteurs de distances cophénétiques. Cette transposition sous forme vectorielle a amené à formaliser le problème statistique comme une comparaison de moyennes multivariées en grande dimension. Dans ce contexte, on a considéré des variantes de la statistique de Hotelling, adaptées au cas de la grande dimension par une hypothèse d'indépendance des coordonnées et des approches de régularisation des variances. Cependant, on a montré que les vecteurs de distances cophénétiques présentent une structure de dépendance marquée, pouvant remettre en cause la stratégie d'agrégation des statistiques univariées par somme de carrés, comme cela est fait dans les statistiques de type « Hotelling » diagonales. C'est pourquoi on a finalement opté pour une méthode d'agrégation plus permissive en termes de dépendances, l'agrégation de p-valeurs de Simes. On a ensuite réalisé une simulation basée sur données GWAS qui a permis de valider le bon comportement de la méthode, en termes de distributions attendues des statistiques de test sous l'hypothèse nulle. Dans le but d'illustrer les applications potentiellement variées de la méthode développée, elle a aussi été utilisée sur des arbres phylogénétiques et a permis, dans ce contexte, de retrouver une différence attendue d'un point de vue biologique. Enfin, la méthode a été appliquée sur des données Hi-C et a permis de détecter des différences en termes de structure du matériel génétique entre deux stades de développement fœtal porcin. Le code utilisé pour les applications aux données GWAS et Hi-C est disponible à https://forgemia.inra.fr/scales/differential-analysis-of-trees.

La question de la dépendance permet d'ouvrir des perspectives à deux niveaux :

Décorrélation. La structure de dépendance des vecteurs de distances cophénétiques est très particulière et dûe au fait que ces vecteurs dérivent d'arbres. L'approche retenue à l'issue de cette thèse permet d'obtenir une p-valeur pour la comparaison multivariée à l'aide d'une agrégation de Simes des p-valeurs individuelles. L'avantage de l'agrégation de Simes est que les hypothèses requises sur les données (PRDS) sont relativement robustes à la dépendance, et leur validité est couramment admise dans le contexte de la génomique.

Cependant, cette approche est moins fine qu'une approche de décorrélation, dans la mesure où on se limite à faire une hypothèse sur la structure de dépen-

dance des données au lieu de prendre en compte explicitement cette structure de dépendance. Bien que la validité de l'approche ait été vérifiée empiriquement, il est possible qu'elle soit trop conservatrice et qu'il en résulte une perte de puissance du test. Dans cette thèse, les approches de décorrélation déjà envisagées (variantes de la statistique de Hotelling basées sur des régularisation de $\hat{\Sigma}$) n'ont soit pas pu être appliquées, soit n'ont pas donné de résultats pertinents, notamment en raison du problème de la grande dimension. Cependant, il existe des approches (Cai et collab., 2013) dédiées à la comparaison de moyennes multivariées en grande dimension et sous hypothèses de dépendance, adaptées à certains types de distribution du signal. Il serait donc utile d'étudier de façon plus précise le lien entre la forme du signal (différence de moyennes de vecteurs de distances cophénétiques) et la structure de corrélation des vecteurs de distances cophénétiques afin d'envisager une prise en compte adaptative de cette structure de corrélation, comme cela est suggéré dans les travaux de Hébert (2019); Hébert et collab. (2021).

Tests hiérarchiques. Dans certains contextes applicatifs, il peut arriver que les arbres à comparer aient un trop grand nombre de feuilles pour qu'une comparaison directe soit pertinente. C'est par exemple le cas dans l'application présentée dans la sous-section 3.4.2, où un arbre associée à une matrice Hi-C intra-chromosomique peut avoir jusqu'à plusieurs milliers de feuilles. Il est alors préférable, tant du point de vue computationnel que de celui de l'interprétation des résultats, de trouver un moyen de définir des arbres plus petits pour appliquer la méthode de comparaison.

Dans la sous-section 3.4.2, une partition du chromosome est obtenue et la méthode est appliquée indépendamment sur les « sous-arbres » correspondant aux sous-matrices Hi-C induites par la partition. Cependant, il est naturel de se demander quel aurait été le résultat de l'analyse si elle avait avait été menée sur une autre partition de ce chromosome, ou plus généralement, comment trouver une partition ou des sous-arbres pertinents à tester.

En effet, réaliser l'analyse différentielle à différents niveaux du dendrogramme global peut présenter divers intérêts. Dans le cas des données Hi-C, cela revient, par exemple, à tester des sous-arbres plus ou moins grands, correspondant au choix d'une partition plus ou moins fine du chromosome. On a vu que plus une différence intervient proche de la racine, plus elle a d'impact dans le calcul de la p-valeur associée, ce qui implique que le niveau auquel on réalise le test est déterminant pour détecter un signal présent dans les données. Etant donné que le signal peut concerner différents niveaux d'organisation du matériel génétique (voir la sous-section 1.1.2), appliquer la méthode sur des sous-arbres de tailles variables, éventuellement imbriqués les uns dans les autres, semble être une bonne direction pour ne pas négliger de différences de structure dans l'analyse. Une illustration de résultats préliminaires pour une comparaison basée sur des partitions d'un

chromosome à différents niveaux est donnée dans la figure 4.1.

Du fait de la structure hiérarchique des arbres, il est donc nécessaire de définir une stratégie de parcours des sous-arbres que l'on souhaite tester afin de considérer les différents niveaux présents dans les données. Ces sous-arbres peuvent alors être imbriqués, ou même se recouvrir partiellement et les tests réalisés ne peuvent plus être considérés comme indépendants. Une nouvelle problématique est alors le contrôle du risque de première espèce le long de ce parcours de sous-arbres. Trouver une procédure qui permettent de contrôler efficacement le FDR dans ce contexte de dépendance induit par la structure hiérarchique des arbres constitue une perspective importante de ce travail. Le travail de Yekutieli (2008) sur les tests hiérarchiques, qui donne une procédure de contrôle du FDR global grâce à un contrôle du FDR de chaque niveau de la hiérarchie, constitue un bon point de départ pour aborder cette question.

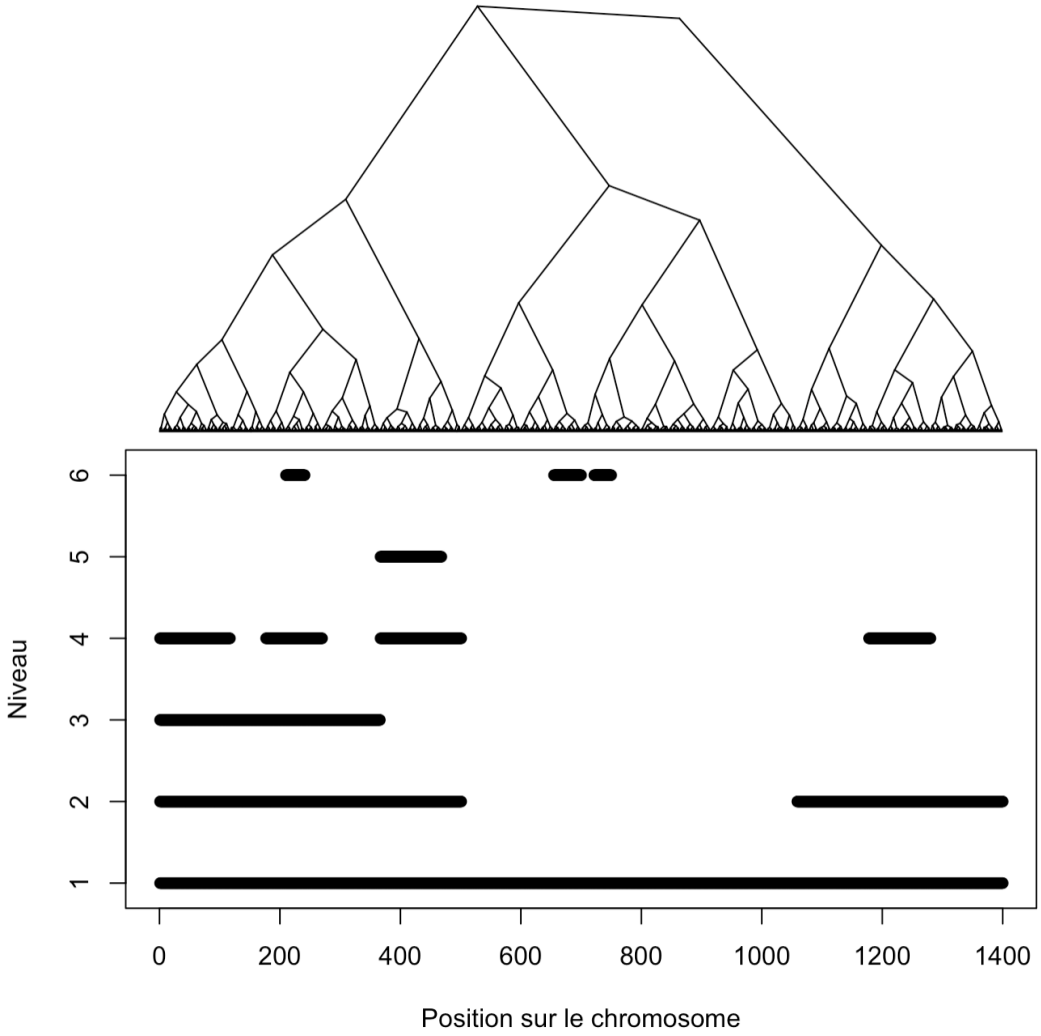


FIGURE 4.1: **Illustration des résultats de la méthode d'analyse différentielle en fonction du niveau de la partition d'un chromosome.** En haut, dendrogramme obtenue en appliquant la CAH sur la moyenne des 3×2 matrices Hi-C correspondant au 18ème chromosome porcin pour la comparaison décrite dans la section 3.4.2. En bas, graphique représentant les régions différentielles obtenues avec la statistique en fonction du niveau considéré dans l'arbre. En abscisse, on représente les 1400 positions génomiques (ou bins) du chromosome. En ordonnées, on représente les niveaux des partitions (1 correspond à une partition en 2 classes du chromosome et au niveau le plus grossier, et 6 à une partition en 64 classes et au niveau le plus fin). Les segments noirs correspondent aux régions génomiques détectées comme différentielles entre conditions par la méthode au niveau de risque 5%. La procédure de correction de tests multiples de Benjamini et Hochberg (1995) a été appliquée pour chaque niveau de façon indépendante.

Annexe A

Annexes du chapitre 2

A.1 Proof of Proposition 2

Proof of Proposition 2. We begin by noting that by Proposition 1, the only reversals that may occur are crossovers. With the notation of Proposition 2, a crossover at step $t+1$ corresponds to the situation where

$$\delta(G_l, G_r) \geq \delta(G_l \cup G_r, G_{\bar{r}}) \text{ or } \delta(G_l, G_r) \geq \delta(G_l \cup G_r, G_{\bar{l}}).$$

By symmetry we focus on the first case. With the notation of Proposition 2, and using the Lance-Willams formula (2.4), the first condition is equivalent to

$$\delta(G_l, G_r) \geq \frac{g_{lr'}\delta(G_l, G_{\bar{r}}) + g_{rr'}\delta(G_r, G_{\bar{r}})}{g_{lr'} + g_{rr'}}$$

while the second one is equivalent to

$$\delta(G_l, G_r) \geq \frac{g_{\bar{l}l}\delta(G_{\bar{l}}, G_l) + g_{\bar{l}r}\delta(G_{\bar{l}}, G_r)}{g_{\bar{l}l} + g_{\bar{l}r}}$$

hence the result. □

A.2 Step-by-step description of the counter-examples

In the following tables, red color is used to signal reversals. Green color in details of Figure 2.2 is used to highlight the value of the objective function (ESS_t) for the clustering with 3 clusters.

Merger	cluster 1	cluster 2	m_t	ESS_t	I_t	$\overline{I}_t$
1	$\{x_1\}$	$\{x_2\}$	1.000	1.000	1.000	0.500
2	$\{x_1, x_2\}$	$\{x_3\}$	0.517	1.517	1.517	0.506

TABLE A.1: Details of Figure 2.1

OCHAC

Merger	cluster 1	cluster 2	m_t	ESS_t	I_t	$\overline{I}_t$
1	$\{x_1\}$	$\{x_2\}$	2.500	2.500	2.500	1.250
2	$\{x_1, x_2\}$	$\{x_3\}$	2.167	4.667	4.667	1.556
3	$\{x_6\}$	$\{x_7\}$	2.500	7.167	2.500	1.250
4	$\{x_5\}$	$\{x_6, x_7\}$	2.167	9.333	4.667	1.556
5	$\{x_1, x_2, x_3\}$	$\{x_4\}$	13.333	22.667	18.000	4.500
6	$\{x_1, x_2, x_3, x_4\}$	$\{x_5, x_6, x_7\}$	20.762	43.429	43.429	6.204

HAC

Merger	cluster 1	cluster 2	m_t	ESS_t	I_t	$\overline{I}_t$
1	$\{x_2\}$	$\{x_6\}$	0.500	0.500	0.500	0.250
2	$\{x_1\}$	$\{x_3\}$	2.000	2.500	2.000	1.000
3	$\{x_5\}$	$\{x_7\}$	2.000	4.500	2.000	1.000
4	$\{x_2, x_6\}$	$\{x_1, x_3\}$	6.250	10.750	8.750	2.188
5	$\{x_4\}$	$\{x_1, x_2, x_3, x_6\}$	13.250	24.000	22.000	4.400
6	$\{x_5, x_7\}$	$\{x_1, x_2, x_3, x_4, x_6\}$	19.429	43.429	43.429	6.204

TABLE A.2: Details of Figure 2.2

Merger	cluster 1	cluster 2	m_t	ESS_t	I_t	$\overline{I}_t$
1	$\{x_1\}$	$\{x_2\}$	0.50	0.50	0.50	0.25
2	$\{x_1, x_2\}$	$\{x_3\}$	2.32	2.82	2.82	0.94
3	$\{x_4\}$	$\{x_5\}$	2.33	5.15	2.33	1.17
4	$\{x_1, x_2, x_3\}$	$\{x_4, x_5\}$	120.84	125.99	125.99	25.20

TABLE A.3: Details of Figure 2.3

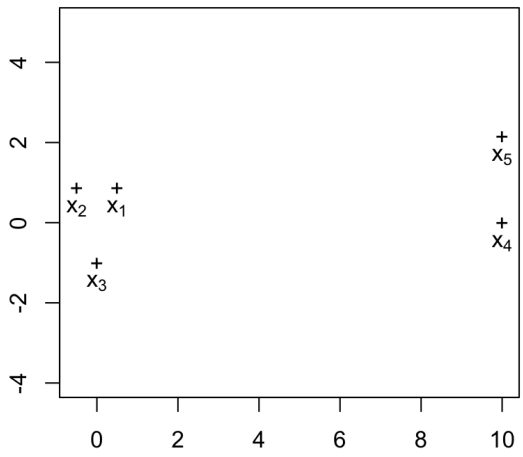
Merger	cluster 1	cluster 2	m_t	ESS_t	I_t	$\overline{I}_t$
1	$\{x_1\}$	$\{x_2\}$	0.995	0.995	0.995	0.498
2	$\{x_1, x_2\}$	$\{x_3\}$	0.998	1.993	1.993	0.664
3	$\{x_1, x_2, x_3\}$	$\{x_4\}$	0.997	2.990	2.990	0.748
4	$\{x_1, x_2, x_3, x_4\}$	$\{x_5\}$	-0.192	2.798	2.798	0.560
5	$\{x_1, x_2, x_3, x_4, x_5\}$	$\{x_6\}$	0.534	3.332	3.332	0.555

TABLE A.4: Details of Figure 2.4

Merger	cluster 1	cluster 2	m_t	ESS_t	I_t	$\overline{I}_t$
1	$\{x_1\}$	$\{x_2\}$	0.50	0.50	0.50	0.25
2	$\{x_4\}$	$\{x_5\}$	2.31	2.81	2.31	1.16
3	$\{x_1, x_2\}$	$\{x_3\}$	2.32	5.13	2.82	0.94
4	$\{x_1, x_2, x_3\}$	$\{x_4, x_5\}$	120.83	125.96	125.96	25.19

TABLE A.5: Details of Figure A.1

A.3 Counter-example of the monotonicity of $\bar{I}_t$ for standard HAC in the Euclidean case



$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ 0 & -1 \\ 10 & 0 \\ 10 & 2.15 \end{pmatrix}$$

$$D \approx \begin{pmatrix} 0.00 & 1.00 & 1.93 & 9.54 & 9.59 \\ 1.00 & 0.00 & 1.93 & 10.54 & 10.58 \\ 1.93 & 1.93 & 0.00 & 10.05 & 10.48 \\ 9.54 & 10.54 & 10.05 & 0.00 & 2.15 \\ 9.59 & 10.58 & 10.48 & 2.15 & 0.00 \end{pmatrix}$$

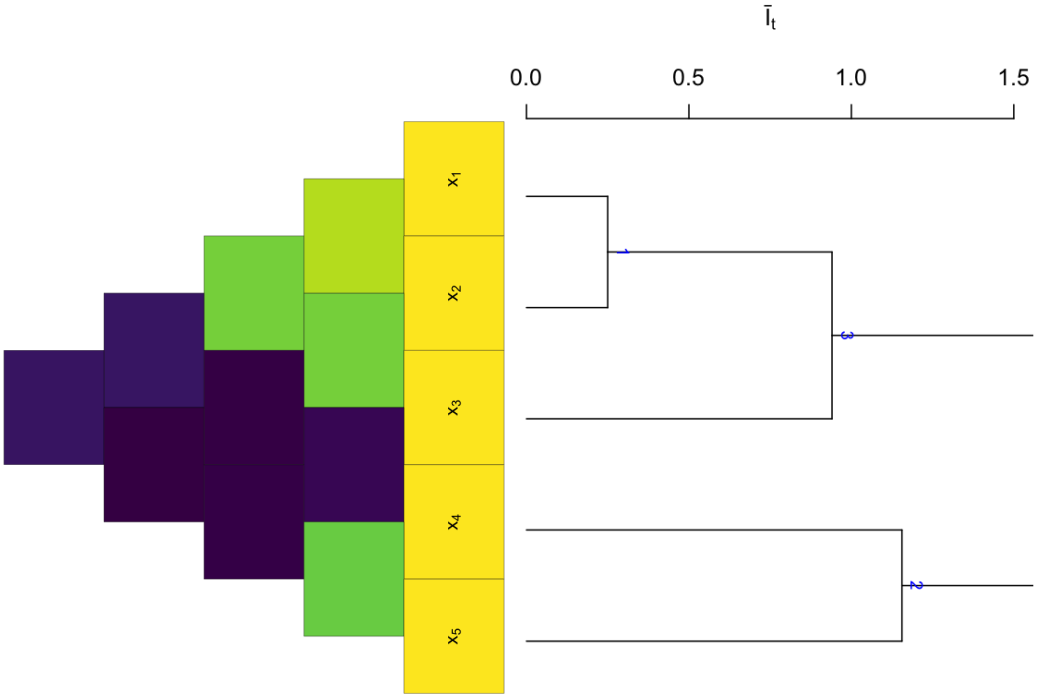


FIGURE A.1: **A reversal for Euclidean standard HAC with height defined as $\bar{I}_t$.** Top left : Configuration of the objects in $\mathbb{R}^2$. Top right : Coordinates of the objects and Euclidean distance matrix corresponding to this configuration. Bottom left : Representation of the values of the dissimilarity (dark colors correspond to larger values, so distant objects). Bottom right : dendrogram obtained from standard HAC. Only the first 3 merges of the dendrogram is represented to ensure a comprehensive view of the sequence of heights.

Bibliography

Ah-Pine, J. et X. Wang. 2016, «Similarity based hierarchical clustering with an application to text collections», dans *Proceedings of the 15th International Symposium on Intelligent Data Analysis (IDA 2016)*, édité par H. Boström, A. Knobbe, C. Soares et P. Papapetrou, Lecture Notes in Computer Sciences, Stockholm, Sweden, p. 320–331, doi:10.1007/978-3-319-46349-0. URL https://hal.archives-ouvertes.fr/hal-01437124.

Ambroise, C., A. Dehman, P. Neuvial, G. Rigaill et N. Vialaneix. 2019, «Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics», *Algorithms for Molecular Biology*, vol. 14, doi: 10.1186/s13015-019-0157-4, p. 22.

Anders, S. et W. Huber. 2010, «Differential expression analysis for sequence count data», *Genome Biology*, vol. 11, n° 10, doi:10.1186/gb-2010-11-10-r106.

Annunziato, A. T. 2008, «DNA packaging: Nucleosomes and chromatin», *Nature Education*, vol. 1, n° 1, p. 26.

Arlot, S., V. Brault, J.-P. Baudry, C. Maugis et B. Michel. 2016, *capushe: CAlibrating Penalities Using Slope HEuristics*. URL https://CRAN.R-project.org/package=capushe, r package version 1.1.1.

Arlot, S., A. Celisse et Z. Harchaoui. 2019, «A kernel multiple change-point algorithm via model selection», URL https://arxiv.org/abs/1202.3878, preprint arXiv: 1202.3878.

Aronszajn, N. 1950, «Theory of reproducing kernels», *Transactions of the American Mathematical Society*, vol. 68, n° 3, doi: 10.1090/s0002-9947-1950-0051437-7, p. 337–337.

Bai, Z. et H. Saranadasa. 1996, «Effect of high dimension: by an example of a two sample problem», *Statistica Sinica*, p. 311–329.

Baker, F. B. 1974, «Stability of two hierarchical grouping techniques case I: sensitivity to data errors», *Journal of the American Statistical Association*, vol. 69, nº 346, doi:10.1080/01621459.1974.10482971, p. 440–445.

Bandelt, H.-J. et A. Dress. 1986, «Reconstructing the shape of a tree from observed dissimilarity data», *Advances in Applied Mathematics*, vol. 7, nº 3, doi:10.1016/0196-8858(86)90038-2, p. 309–343.

Barnett, I., R. Mukherjee et X. Lin. 2017, «The generalized higher criticism for testing SNP-set effects in genetic association studies», *Journal of the American Statistical Association*, vol. 112, nº 517, doi:10.1080/01621459.2016.1192039, p. 64–76.

Batagelj, V. 1981, «Note on ultrametric hierarchical clustering algorithms», *Psychometrika*, vol. 46, nº 3, doi:10.1007/bf02293743, p. 351–352.

Belton, J.-M., R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan et J. Dekker. 2012, «Hi-c: a comprehensive technique to capture the conformation of genomes.», *Methods (San Diego, Calif.)*, vol. 58, doi:10.1016/j.ymeth.2012.05.001, p. 268–276, ISSN 1095-9130.

Benjamini, Y. et Y. Hochberg. 1995, «Controlling the false discovery rate: a practical and powerful approach to multiple testing», *Journal of the Royal Statistical Society Series B*, vol. 57, nº 1, p. 289–300. URL https://www.jstor.org/stable/2346101.

Benjamini, Y. et D. Yekutieli. 2001, «The control of the false discovery rate in multiple testing under dependency», *The Annals of Statistics*, vol. 29, nº 4, doi:10.1214/aos/1013699998.

Bennett, K. D. 1996, «Determination of the number of zones in a biostratigraphical sequence», *New Phytologist*, vol. 132, nº 1, doi:10.1111/j.1469-8137.1996.tb04521.x, p. 155–170.

Berlivet, S., D. Paquette, A. Dumouchel, D. Langlais, J. Dostie et M. Kmita. 2013, «Clustering of tissue-specific sub-TADs accompanies the regulation of HoxA genes in developing limbs», *PLoS Genetics*, vol. 9, nº 12, doi:10.1371/journal.pgen.1004018, p. e1004 018.

Bickel, P. J. et E. Levina. 2004, «Some theory for fisher's linear discriminant function, naive bayes', and some alternatives when there are many more variables than observations», *Bernoulli*, vol. 10, nº 6, doi:10.3150/bj/1106314847, p. 989–1010.

Billera, L. J., S. P. Holmes et K. Vogtmann. 2001, «Geometry of the space of phylogenetic trees», *Advances in Applied Mathematics*, vol. 27, n° 4, doi:10.1006/aama.2001.0759, p. 733–767.

Böcker, S., S. Canzar et G. W. Klau. 2013, «The generalized robinson-foulds metric», dans *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, p. 156–169, doi:10.1007/978-3-642-40453-5_13.

Bonev, B. et G. Cavalli. 2016, «Organization and function of the 3d genome», *Nature Reviews Genetics*, vol. 17, n° 11, doi:10.1038/nrg.2016.112, p. 661–678.

Bonev, B., N. M. Cohen, Q. Szabo, L. Fritsch, G. L. Papadopoulos, Y. Lubling, X. Xu, X. Lv, J.-P. Hugnot, A. Tanay et G. Cavalli. 2017, «Multiscale 3d genome rewiring during mouse neural development», *Cell*, vol. 171, n° 3, doi:10.1016/j.cell.2017.09.043, p. 557–572.e24.

Bordewich, M. et C. Semple. 2005, «On the computational complexity of the rooted subtree prune and regraft distance», *Annals of Combinatorics*, vol. 8, n° 4, doi:10.1007/s00026-004-0229-z, p. 409–423.

Briand, S., C. Dessimoz, N. El-Mabrouk, M. Lafond et G. Lobinska. 2020, «A generalized robinson-foulds distance for labeled trees», *BMC Genomics*, vol. 21, n° S10, doi:10.1186/s12864-020-07011-0.

Brown, D. G. et M. Owen. 2019, «Mean and variance of phylogenetic trees», *Systematic Biology*, vol. 69, n° 1, doi:10.1093/sysbio/syz041, p. 139–154.

Cai, T. T., W. Liu et Y. Xia. 2013, «Two-sample test of high dimensional means under dependence», *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, n° 2, doi:10.1111/rssb.12034, p. 349–372.

Chakerian, J. et S. Holmes. 2012, «Computational tools for evaluating phylogenetic and hierarchical clustering trees», *Journal of Computational and Graphical Statistics*, vol. 21, n° 3, doi:10.1080/10618600.2012.640901, p. 581–599.

Chavent, M., V. Kuentz-Simonet, A. Labenne et J. Saracco. 2018, «ClustGeo2: an R package for hierarchical clustering with spatial constraints», *Computational Statistics*, vol. 33, n° 4, doi:10.1007/s00180-018-0791-1, p. 1799–1822.

Chen, J. et J. Ye. 2008, «Training SVM with indefinite kernels», dans *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, édité par W. Cohen, A. McCallum et S. Roweis, ACM, New York, NY, USA, Helsinki, Finland, p. 136–146, doi:10.1145/1390156.1390174.

Chen, L. S., D. Paul, R. L. Prentice et P. Wang. 2011, «A regularized Hotelling's T^2 test for pathway analysis in proteomic studies», *Journal of the American Statistical Association*, vol. 106, n° 496, doi:10.1198/jasa.2011.ap10599, p. 1345–1360.

Chen, S. X. et Y.-L. Qin. 2010, «A two-sample test for high-dimensional data with applications to gene-set testing», *The Annals of Statistics*, vol. 38, n° 2, doi:10.1214/09-aos716, p. 808–835.

Chen, Y., E. Garcia, M. Gupta, A. Rahimi et L. Cazzanti. 2009, «Similarity-based classification: concepts and algorithm», *Journal of Machine Learning Research*, vol. 10, p. 747–776.

Cresswell, K. G. et M. G. Dozmorov. 2020, «TADCompare: An R package for differential and temporal analysis of topologically associated domains», *Frontiers in Genetics*, vol. 11, doi:10.3389/fgene.2020.00158.

Critchlow, D. E., D. K. Pearl et C. Qian. 1996, «The triples distance for rooted bifurcating phylogenetic trees», *Systematic Biology*, vol. 45, n° 3, doi:10.1093/sysbio/45.3.323, p. 323–334.

Danon, L., A. Diaz-Guilera, J. Duch et A. Arenas. 2005, «Comparing community structure identification», *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, doi:10.1088/1742-5468/2005/09/P09008, p. P09 008.

DasGupta, B., X. He, T. Jiang, M. Li, J. Tromp, L. Wang et L. Zhang. 1998, «Computing distances between evolutionary trees», dans *Handbook of Combinatorial Optimization*, Springer US, p. 781–822, doi:10.1007/978-1-4613-0303-9_11.

Dasgupta, B., X. He, T. Jiang, M. Li, J. Tromp et L. Zhang. 1997, «On distances between phylogenetic trees», *SODA '97: Proceedings of the eighth annual ACM-SIAM symposium on Discrete algorithms*.

Day, W. H. E. 1985, «Optimal algorithms for comparing trees with labeled leaves», *Journal of Classification*, vol. 2, n° 1, doi:10.1007/bf01908061, p. 7–28.

Dehman, A. 2015, *Spatial Clustering of Linkage Disequilibrium Blocks for Genome-Wide Association Studies*, thèse de doctorat, Université Paris Saclay.

Diaconis, P. W. et S. P. Holmes. 1998, «Matchings and phylogenetic trees.», *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, p. 14 600–14 602, ISSN 0027-8424.

Dillies, M.-A., A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloe, C. L. Gall, B. Schaeffer, S. L. Crom, M. Guedj et F. J. and. 2012, «A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis», *Briefings in Bioinformatics*, vol. 14, nº 6, doi:10.1093/bib/bbs046, p. 671–683.

Dixon, J., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. Liu et B. Ren. 2012, «Topological domains in mammalian genomes identified by analysis of chromatin interactions», *Nature*, vol. 485, doi:10.1038/nature11082, p. 376–380.

Dixon, J. R., D. U. Gorkin et B. Ren. 2016, «Chromatin domains: the unit of chromosome organization», *Molecular Cell*, vol. 62, nº 5, doi:10.1016/j.molcel.2016.05.018, p. 668–680.

Dixon, J. R., I. Jung, S. Selvaraj, Y. Shen, J. E. Antosiewicz-Bourget, A. Y. Lee, Z. Ye, A. Kim, N. Rajagopal, W. Xie, Y. Diao, J. Liang, H. Zhao, V. V. Lobanenkov, J. R. Ecker, J. A. Thomson et B. Ren. 2015, «Chromatin architecture reorganization during stem cell differentiation», *Nature*, vol. 518, nº 7539, doi:10.1038/nature14222, p. 331–336.

Djekidel, M. N., Y. Chen et M. Q. Zhang. 2018, «FIND: difFerential chromatin INteractions Detection using a spatial Poisson process», *Genome Research*, vol. 28, nº 3, doi:10.1101/gr.212241.116, p. 412–422.

Dohm, J. C., C. Lottaz, T. Borodina et H. Himmelbauer. 2008, «Substantial biases in ultra-short read data sets from high-throughput DNA sequencing», *Nucleic Acids Research*, vol. 36, nº 16, doi:10.1093/nar/gkn425, p. e105–e105.

Dong, K., H. Pang, T. Tong et M. G. Genton. 2016, «Shrinkage-based diagonal Hotelling's tests for high-dimensional small sample size data», *Journal of Multivariate Analysis*, vol. 143, doi:10.1016/j.jmva.2015.08.022, p. 127–142.

Dudoit, S., J. Fridlyand et T. P. Speed. 2002, «Comparison of discrimination methods for the classification of tumors using gene expression data», *Journal of the American Statistical Association*, vol. 97, nº 457, doi:10.1198/016214502753479248, p. 77–87.

Ferligoj, A. et V. Batagelj. 1982, «Clustering with relational constraint», *Psychometrika*, vol. 47, nº 4, doi:10.1007/bf02293706, p. 413–426.

Fraser, J., C. Ferrai, A. M. Chiariello, M. Schueler, T. Rito, G. Laudanno, M. Barbieri, B. L. Moore, D. C. Kraemer, S. Aitken, S. Q. Xie, K. J. Morris, M. Itoh, H. Kawaji, I. Jaeger, Y. Hayashizaki, P. Carninci, A. R. Forrest, The FANTOM

Consortium, C. A. Semple, J. Dostie, A. Pombo et M. Nicodemi. 2015, «Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation», *Molecular Systems Biology*, vol. 11, doi:10.15252/msb.20156492, p. 852.

Goodman, L. A. et W. H. Kruskal. 1959, «Measures of association for cross classifications. II: Further discussion and references», *Journal of the American Statistical Association*, vol. 54, n° 285, doi:10.1080/01621459.1959.10501503, p. 123–163.

Gordon, A. 1996, «A survey of constrained classification», *Computational Statistics & Data Analysis*, vol. 21, n° 1, doi:10.1016/0167-9473(95)00005-4, p. 17–29.

Gribaldo, S. et C. Brochier-Armanet. 2006, «The origin and evolution of archaea: a state of the art», *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 361, n° 1470, doi:10.1098/rstb.2006.1841, p. 1007–1022.

Grimm, E. C. 1987, «CONISS: a FORTRAN 77 program for stratigraphically constrained analysis by the method of incremental sum of squares», *Computers & Geosciences*, vol. 13, n° 1, doi:10.1016/0098-3004(87)90022-7, p. 13–35.

Haddad, N., C. Vaillant et D. Jost. 2017, «IC-Finder: inferring robustly the hierarchical organization of chromatin folding», *Nucleic Acids Research*, vol. 45, n° 10, doi:10.1093/nar/gkx036, p. e81–e81.

Hartigan, J. A. 1967, «Representation of similarity matrices by trees», *Journal of the American Statistical Association*, vol. 62, n° 320, doi:10.2307/2283766, p. 1140–1158.

Hébert, F. 2019, *Prise en compte de la dépendance pour des problèmes de test global et de prédiction*, thèse de doctorat, IRMAR / Agrocampus Ouest.

Hébert, F., D. Causeur et M. Emily. 2021, «An adaptive decorrelation procedure for signal detection», *Computational Statistics & Data Analysis*, vol. 153, doi:10.1016/j.csda.2020.107082, p. 107 082.

Hnisz, D., A. S. Weintraub, D. S. Day, A.-L. Valton, R. O. Bak, C. H. Li, J. Goldmann, B. R. Lajoie, Z. P. Fan, A. A. Sigova, J. Reddy, D. Borges-Rivera, T. I. Lee, R. Jaenisch, M. H. Porteus, J. Dekker et R. A. Young. 2016, «Activation of proto-oncogenes by disruption of chromosome neighborhoods», *Science*, vol. 351, n° 6280, doi:10.1126/science.aad9024, p. 1454–1458.

Holmes, S. 2003, «Statistics for phylogenetic trees», *Theoretical Population Biology*, vol. 63, n° 1, doi:10.1016/s0040-5809(02)00005-9, p. 17–32.

Hotelling, H. 1931, «The generalization of student's ratio», *The Annals of Mathematical Statistics*, vol. 2, nº 3, doi:10.1214/aoms/1177732979, p. 360–378.

Hu, M., K. Deng, S. Selvaraj, Z. Qin, B. Ren et J. S. Liu. 2012, «HiCNorm: removing biases in Hi-C data via poisson regression», *Bioinformatics*, vol. 28, nº 23, doi:10.1093/bioinformatics/bts570, p. 3131–3133.

Imakaev, M., G. Fudenberg, R. McCord, N. Naumova, A. Goloborodko, B. Lajoie, J. Dekker et L. Mirny. 2012, «Iterative correction of Hi-C data reveals hallmarks of chromosome organization», *Nature Methods*, vol. 9, nº 10, doi:10.1038/nmeth.2148, p. 999–1003.

James, W. et C. Stein. 1992, «Estimation with quadratic loss», dans *Springer Series in Statistics*, Springer New York, p. 443–460, doi:10.1007/978-1-4612-0919-5_30.

Johnson, S. C. 1967, «Hierarchical clustering schemes», *Psychometrika*, vol. 32, nº 3, doi:10.1007/bf02289588, p. 241–254.

Kaiser, V. B. et C. A. Semple. 2017, «When TADs go bad: chromatin structure and nuclear organisation in human disease», *F1000Research*, vol. 6, doi:10.12688/f1000research.10792.1, p. 314.

Kempfer, R. et A. Pombo. 2019, «Methods for mapping 3D chromosome architecture», *Nature Reviews Genetics*, vol. 21, nº 4, doi:10.1038/s41576-019-0195-2, p. 207–226.

Krislock, N. et H. Wolkowicz. 2012, *Handbook on Semidefinite, Conic and Polynomial Optimization*, *International Series in Operations Research & Management Science*, vol. 166, chap. Euclidean distance matrices and applications, Springer, New York, Dordrecht, Heidelberg, London, p. 879–914.

Kruskal, B., Joseph. 1964, «Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis», *Psychometrika*, vol. 29, nº 1, doi:10.1007/bf02289565, p. 1–27.

Kuhner, M. K. et J. Felsenstein. 1994, «A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates.», *Molecular Biology and Evolution*, doi:10.1093/oxfordjournals.molbev.a040126.

Lance, G. et W. Williams. 1967, «A general theory of classificatory sorting strategies: 1. Hierarchical systems», *The Computer Journal*, vol. 9, nº 4, doi:10.1093/comjnl/9.4.373, p. 373–380.

Lebart, L. 1978, «Programme d'agrégation avec contraintes», *Les Cahiers de l'Analyse des Données*, vol. 3, n° 3, p. 275–287. URL http://www.numdam.org/item?id=CAD_1978__3_3_275_0.

Lefort, G., R. Servien, H. Quesnel, Y. Billon, L. Canario, N. Iannucelli, C. Canlet, A. Paris, N. Vialaneix et L. Liaubet. 2020, «The maturity in fetal pigs using a multi-fluid metabolomic approach», *Scientific Report*, vol. 10, doi:10.1038/s41598-020-76709-8, p. 19 912.

Li, M., J. Tromp et L. Zhang. 1996, «Some notes on the nearest neighbour interchange distance», dans *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, p. 343–351, doi:10.1007/3-540-61332-3_168.

Lieberman-Aiden, E., N. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. Lajoie, P. Sabo, M. Dorschner, R. Sandstrom, B. Bernstein, M. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. Mirny, E. Lander et J. Dekker. 2009, «Comprehensive mapping of long-range interactions reveals folding principles of the human genome», *Science*, vol. 326, n° 5950, doi:10.1126/science.1181369, p. 289–293.

Liu, Y. et J. Xie. 2018, «Powerful test based on conditional effects for genome-wide screening», *The Annals of Applied Statistics*, vol. 12, n° 1, doi:10.1214/17-aoas1103.

Lun, A. T. et G. K. Smyth. 2014, «De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly», *Nucleic Acids Research*, vol. 42, n° 11, doi:10.1093/nar/gku351, p. e95–e95.

Lun, A. T. et G. K. Smyth. 2015, «diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data», *BMC Bioinformatics*, vol. 16, doi:10.1186/s12859-015-0683-0, p. 258.

Lupiáñez, D. G., K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel et S. Mundlos. 2015, «Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions», *Cell*, vol. 161, n° 5, doi:10.1016/j.cell.2015.04.004, p. 1012–1025.

Lupiáñez, D. G., M. Spielmann et S. Mundlos. 2016, «Breaking TADs: how alterations of chromatin domains result in disease», *Trends in Genetics*, vol. 32, n° 4, doi:10.1016/j.tig.2016.01.003, p. 225–237.

Mallows, C. 1957, «Non-null ranking models. I», *Biometrika*, vol. 44, n° 1-2, doi: 10.1093/biomet/44.1-2.114, p. 114–130.

Marti-Marimon, M., N. Vialaneix, Y. Lahbib-Mansais, M. Zytnicki, S. Camut, D. Robelin, M. Bouissou-Matet et S. Foissac. 2021, «Major reorganization of chromosome conformation during muscle development in pig», cahier de recherche, INRAe, France. Preprint.

Miyamoto, S., R. Abe, Y. Endo et J.-I. Takeshita. 2015, «Ward method of hierarchical clustering for non-Euclidean similarity measures», dans *Proceedings of the VIIth International Conference of Soft Computing and Pattern Recognition (SoCPaR 2015)*, IEEE, Fukuoka, Japan, doi:10.1109/socpar.2015.7492784.

Moore, G., M. Goodman et J. Barnabas. 1973, «An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets», *Journal of Theoretical Biology*, vol. 38, n° 3, doi: 10.1016/0022-5193(73)90251-8, p. 423–457.

Murtagh, F. et P. Legendre. 2014, «Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion», *Journal of Classification*, vol. 31, n° 3, doi:10.1007/s00357-014-9161-z, p. 274–295.

Northcott, P. A., C. Lee, T. Zichner, A. M. Stütz, S. Erkek, D. Kawauchi, D. J. H. Shih, V. Hovestadt, M. Zapatka, D. Sturm, D. T. W. Jones, M. Kool, M. Remke, F. M. G. Cavalli, S. Zuyderduyn, G. D. Bader, S. VandenBerg, L. A. Esparza, M. Ryzhova, W. Wang, A. Wittmann, S. Stark, L. Sieber, H. Seker-Cin, L. Linke, F. Kratochwil, N. Jäger, I. Buchhalter, C. D. Imbusch, G. Zipprich, B. Raeder, S. Schmidt, N. Diessl, S. Wolf, S. Wiemann, B. Brors, C. Lawerenz, J. Eils, H.-J. Warnatz, T. Risch, M.-L. Yaspo, U. D. Weber, C. C. Bartholomae, C. von Kalle, E. Turányi, P. Hauser, E. Sanden, A. Darabi, P. Siesjö, J. Sterba, K. Zitterbart, D. Sumerauer, P. van Sluis, R. Versteeg, R. Volckmann, J. Koster, M. U. Schuhmann, M. Ebinger, H. L. Grimes, G. W. Robinson, A. Gajjar, M. Mynarek, K. von Hoff, S. Rutkowski, T. Pietsch, W. Scheurlen, J. Felsberg, G. Reifenberger, A. E. Kulozik, A. von Deimling, O. Witt, R. Eils, R. J. Gilbertson, A. Korshunov, M. D. Taylor, P. Lichter, J. O. Korbel, R. J. Wechsler-Reya et S. M. Pfister. 2014, «Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma», *Nature*, vol. 511, n° 7510, doi:10.1038/nature13379, p. 428–434.

de Oliveira Martins, L., É. Leal et H. Kishino. 2008, «Phylogenetic detection of recombination with a bayesian prior on the distance between trees», *PLoS ONE*, vol. 3, n° 7, doi:10.1371/journal.pone.0002651, p. e2651.

Owen, M. et J. S. Provan. 2011, «A fast algorithm for computing geodesic distances in tree space», *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, nº 1, doi:10.1109/tcbb.2010.3, p. 2–13.

Paradis, E. 2011, «Simulating phylogenies and evolutionary data», dans *Analysis of Phylogenetics and Evolution with R*, Springer New York, p. 313–330, doi:10.1007/978-1-4614-1743-9_7.

Paradis, E., J. Claude et K. Strimmer. 2004, «APE: Analyses of phylogenetics and evolution in R language», *Bioinformatics*, vol. 20, nº 2, doi:10.1093/bioinformatics/btg412, p. 289–290.

Pearson, H. 2006, «What is a gene?», *Nature*, vol. 441, nº 7092, doi:10.1038/441398a, p. 398–401.

Pray, L. A. 2008, «Discovery of DNA structure and function: Watson and Crick», *Nature Education*, vol. 1, nº 1, p. 100.

Puigbò, P., Y. I. Wolf et E. V. Koonin. 2009, «Search for a tree of life in the thicket of the phylogenetic forest», *Journal of Biology*, vol. 8, nº 6, doi:10.1186/jbiol159, p. 59.

Qin, J., D. P. Lewis et W. S. Noble. 2003, «Kernel hierarchical gene clustering from microarray expression data», *Bioinformatics*, vol. 19, nº 16, doi:10.1093/bioinformatics/btg288, p. 2097–2104.

Rammal, R., G. Toulouse et M. A. Virasoro. 1986, «Ultrametricity for physicists», *Reviews of Modern Physics*, vol. 58, nº 3, doi:10.1103/revmodphys.58.765, p. 765–788.

Randriamihamison, N., N. Vialaneix et P. Neuvial. 2020, «Applicability and interpretability of ward's hierarchical agglomerative clustering with or without contiguity constraints», *Journal of Classification*, doi:10.1007/s00357-020-09377-y.

Robinson, D. 1971, «Comparison of labeled trees with valency three», *Journal of Combinatorial Theory, Series B*, vol. 11, nº 2, doi:10.1016/0095-8956(71)90020-7, p. 105–119.

Robinson, D. et L. Foulds. 1981, «Comparison of phylogenetic trees», *Mathematical Biosciences*, vol. 53, nº 1-2, doi:10.1016/0025-5564(81)90043-2, p. 131–147.

Robinson, D. F. et L. R. Foulds. 1979, «Comparison of weighted labelled trees», dans *Lecture Notes in Mathematics*, Springer Berlin Heidelberg, p. 119–126, doi:10.1007/bfb0102690.

Robinson, M. D., D. J. McCarthy et G. K. Smyth. 2009, «edgeR: a bioconductor package for differential expression analysis of digital gene expression data», *Bioinformatics*, vol. 26, n° 1, doi:10.1093/bioinformatics/btp616, p. 139–140.

Robinson, M. D. et A. Oshlack. 2010, «A scaling normalization method for differential expression analysis of RNA-seq data», *Genome Biology*, vol. 11, n° 3, doi:10.1186/gb-2010-11-3-r25, p. R25.

Schleif, F.-M. et P. Tino. 2015, «Indefinite proximity learning: a review», *Neural Computation*, vol. 27, n° 10, doi:10.1162/neco_a_00770, p. 2039–2096.

Schliep, K. P. 2010, «phangorn: phylogenetic analysis in r», *Bioinformatics*, vol. 27, n° 4, doi:10.1093/bioinformatics/btq706, p. 592–593.

Schoenberg, I. 1935, «Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de Hilbert"», *Annals of Mathematics*, vol. 36, p. 724–732.

Schölkopf, B. et A. J. Smola. 2002, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, ISBN 0262536579. URL https://www.ebook.de/de/product/33404701/bernhard_scholkopf_alexander_j_smola_learning_with_kernels_support_vector_machines_regularization_optimization_and_beyond.html.

Servant, N. 2017, *Analysis of Chromosome Conformation Data and Application to Cancer*, thèse de doctorat, Ecole doctorale Complexité du vivant (Paris).

Shen, Y., Z. Lin et J. Zhu. 2011, «Shrinkage-based regularization tests for high-dimensional data with application to gene set analysis», *Computational Statistics & Data Analysis*, vol. 55, n° 7, doi:10.1016/j.csda.2010.12.013, p. 2221–2233.

Simes, R. J. 1986, «An improved bonferroni procedure for multiple tests of significance», *Biometrika*, vol. 73, n° 3, doi:10.1093/biomet/73.3.751, p. 751–754.

Smith, M. R. 2020, *TreeDist: Distances between Phylogenetic Trees. R package version 2.0.3*, doi:10.5281/zenodo.3528124.

Smyth, G. K. 2004, «Linear models and empirical bayes methods for assessing differential expression in microarray experiments», *Statistical Applications in Genetics and Molecular Biology*, vol. 3, n° 1, doi:10.2202/1544-6115.1027, p. 1–25.

Sokal, R. R. et F. J. Rohlf. 1962, «The comparison of dendrograms by objective methods», *Taxon*, vol. 11, n° 2, doi:10.2307/1217208, p. 33.

Soler-Vila, P., P. Cuscó, I. Farabella, M. D. Stefano et M. A. Marti-Renom. 2020, «Hierarchical chromatin organization detected by TADpole», *Nucleic Acids Research*, vol. 48, n° 7, doi:10.1093/nar/gkaa087, p. e39–e39.

Stansfield, J. C., K. G. Cresswell et M. G. Dozmorov. 2019, «multiHiCcompare: joint normalization and comparative analysis of complex hi-c experiments», *Bioinformatics*, vol. 35, n° 17, doi:10.1093/bioinformatics/btz048, p. 2916–2923.

Stansfield, J. C., K. G. Cresswell, V. I. Vladimirov et M. G. Dozmorov. 2018, «HiCcompare: an R-package for joint normalization and comparison of Hi-C datasets», *BMC Bioinformatics*, vol. 19, doi:10.1186/s12859-018-2288-x, p. 279.

Steel, M. A. 1988, «Distribution of the symmetric difference metric on phylogenetic trees», *SIAM Journal on Discrete Mathematics*, vol. 1, n° 4, doi:10.1137/0401050, p. 541–551.

Steel, M. A. et D. Penny. 1993, «Distributions of tree comparison metrics–some new results», *Systematic Biology*, vol. 42, n° 2, doi:10.1093/sysbio/42.2.126, p. 126–141.

Steinley, D. et L. Hubert. 2008, «Order-constrained solutions in K-means clustering: even better than being globally optimal», *Psychometrika*, vol. 73, n° 4, doi:10.1007/s11336-008-9058-z, p. 647–664.

Strauss, T. et M. J. von Maltitz. 2017, «Generalising Ward's method for use with Manhattan distances», *PLoS ONE*, vol. 12, doi:10.1371/journal.pone.0168288, p. e0168 288.

Székely, G. J. et M. L. Rizzo. 2005, «Hierarchical clustering via joint between-within distances: extending Ward's minimum variance method», *Journal of Classification*, vol. 22, n° 2, doi:10.1007/s00357-005-0012-9, p. 151–183.

The International HapMap Consortium. 2003, «The international HapMap project», *Nature*, vol. 426, doi:10.1038/nature02168, p. 789–796.

Tong, T. et Y. Wang. 2012, «Optimal shrinkage estimation of variances with applications to microarray data analysis», *Journal of the American Statistical Association*, vol. 102, n° 477, doi:10.1198/016214506000001266, p. 113–122.

Tusher, V. G., R. Tibshirani et G. Chu. 2001, «Significance analysis of microarrays applied to the ionizing radiation response», *Proceedings of the National Academy of Sciences*, vol. 98, n° 9, doi:10.1073/pnas.091062498, p. 5116–5121.

Voillet, V., M. SanCristobal, Y. Lippi, P. G. Martin, N. Iannuccelli, C. Lascor, F. Vignoles, Y. Billon, L. Canario et L. Liaubet. 2014, «Muscle transcriptomic investigation of late fetal development identifies candidate genes for piglet maturity», *BMC Genomics*, vol. 15, doi:10.1186/1471-2164-15-797, p. 797.

Ward, J. H. 1963, «Hierarchical grouping to optimize an objective function», *Journal of the American Statistical Association*, vol. 58, nº 301, doi:10.1080/01621459.1963.10500845, p. 236–244.

Wickham, H. 2016, *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag, New York, USA, ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Wishart, D. 1969, «An algorithm for hierarchical classifications», *Biometrics*, vol. 25, nº 1, doi:10.2307/2528688, p. 165–170.

Wu, Z., Y. Sun, S. He, J. Cho, H. Zhao et J. Jin. 2014, «Detection boundary and higher criticism approach for rare and weak genetic effects», *The Annals of Applied Statistics*, vol. 8, nº 2, doi:10.1214/14-aoas724.

Yaffe, E. et A. Tanay. 2011, «Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture», *Nature Genetics*, vol. 43, nº 11, doi:10.1038/ng.947, p. 1059–1065.

Yekutieli, D. 2008, «Hierarchical false discovery rate–controlling methodology», *Journal of the American Statistical Association*, vol. 103, nº 481, doi:10.1198/016214507000001373, p. 309–316.

Young, G. et A. Householder. 1938, «Discussion of a set of points in terms of their mutual distances», *Psychometrika*, vol. 3, doi:10.1007/bf02287916, p. 19–22.

Zhang, J. et J. Xu. 2009, «On the k-sample Behrens-Fisher problem for high-dimensional data», *Science in China Series A: Mathematics*, vol. 52, nº 6, doi:10.1007/s11425-009-0091-x, p. 1285–1304.

Zheng, H. et W. Xie. 2019, «The role of 3D genome organization in development and cell differentiation», *Nature Reviews Molecular Cell Biology*, vol. 20, nº 9, doi:10.1038/s41580-019-0132-4, p. 535–550.

Zufferey, M., D. Tavernari, E. Oricchio et G. Ciriello. 2018, «Comparison of computational methods for the identification of topologically associating domains», *Genome biology*, vol. 19, nº 1, p. 217.