Rapport de stage

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

Etudiant :

Randriamihamison
Nathanaël
Master 2 MApI^3

Encadrants :

Mme Vialaneix Nathalie
M. Neuvial Pierre

Stage effectué du 5 mars au 5 septembre 2018

Remerciements

Je tiens à remercier ma formation, le master II Mathématiques Appliquées à l'Industrie, l'Ingénierie et l'Innovation de l'université Toulouse III - Paul Sabatier, et l'ensemble de ses intervenants qui m'ont permis, par leurs enseignements, de faire ce stage.

Je suis reconnaissant envers l'ensemble du personnel du centre INRA - Occitanie dont l'accueil chaleureux m'a fourni un cadre de travail précieux et privilégié.

Je remercie également très sincèrement mes encadrants, Nathalie Vialaneix, Pierre Neuvial et Sylvain Foissac pour leur aide et leur implication tout au long de ce stage. Grâce à leur soutien, j'ai la chance de pouvoir poursuivre en thèse les recherches initiées à l'occasion de ce stage.

Table des matières

Introduction

Dans les élevages porcins, le progrès génétique des dernières années s'est accompagné d'une augmentation substantielle de la mortalité des porcelets. La maturité du porcelet, définie comme l'état de plein développement permettant la survie à la naissance, est un déterminant important de la mortalité précoce mais ces mécanismes sont encore largement méconnus. La motivation initiale de ce stage était l'étude de données produites à l'INRA de Toulouse dans le cadre du projet SCALES. En effet, ces données sont susceptibles d'améliorer la compréhension des mécanismes biologiques impliqués dans la maturité à la naissance et donc la survie périnatale.

De manière plus précise, les données collectées sont des données sur la conformation spatiale (l'enroulement) des chromosomes dans la cellule et sont appelées données Hi-C (High Chromosome Contact Map). Elles ont été obtenues à deux stades de développements fœtaux différents, avant la naissance, et permettent donc de voir l'évolution de cette conformation durant la phase finale de la gestation. Comparer les données pour ces deux conditions biologiques pourrait permettre une meilleure compréhension des mécanismes biologiques impliqués dans la survie du porcelet après la naissance.

La méthode statistique choisie pour l'analyse de ces données Hi-C est la classification ascendante hiérarchique (CAH). En effet, l'aspect hiérarchique de cette méthode de classification permet de modéliser la structure spatiale, elle aussi supposée hiérarchique, du chromosome. Pour prendre en compte l'aspect linéaire du génome, une version contrainte (sous contrainte de contiguïté, le long du chromosome) est utilisée.

Dans le cadre de ce stage, j'ai étudié les propriétés de ces deux versions de la CAH (contrainte et non contrainte) et également leurs extensions à différents types de données. En particulier, je me suis intéressé aux propriétés des représentations graphiques des résultats de ce type de méthode, les dendrogrammes. Par ailleurs, j'ai réalisé une étude bibliographique concernant l'agrégation et la comparaison des classifications ascendantes hiérarchiques en vue de futurs travaux permettant de mettre en valeur les différences de structures hiérarchiques entre les deux conditions biologiques d'intérêt décrites ci-dessus. Enfin, d'un point de vue pratique, j'ai participé au développement du package R **adjclust** implémentant diverses versions (pour divers types de données) de la classification hiérarchique sous contrainte de contiguïté.

Ce rapport de stage est organisé comme suit. Après une description des organismes d'accueil et du projet dans lequel s'inscrit ce stage, le contexte et le sujet du stage sont présentés. Ensuite, en tant qu'outil statistique d'étude du génome, la CAH est introduite, et ses extensions à différents types de données sont étudiées. Puis les représentations graphiques de CAH sont définies et nous mettons en évidence certaines de leurs propriétés dans des contextes variés. Les méthodes de comparaisons de CAH sont ensuite abordées d'un point de vue bibliographique.

1 Organisme d'accueil et contexte du projet

J'ai réalisé mon stage dans deux laboratoires de recherche :
- l'Institut de Mathématiques de Toulouse (IMT), dans l'équipe Statistique et Probabilités ;
- le département Mathématiques, Informatique et Applications de Toulouse (MIAT) du centre INRA-Occitanie-Toulouse.

Ce stage a fait l'objet d'un co-encadrement par un chercheur CNRS, M. Pierre Neuvial (IMT), et une chercheuse INRA, Mme Nathalie Vialaneix (MIAT).

1.1 Institut de Mathématiques de Toulouse

L'Institut de Mathématiques de Toulouse rassemble 240 enseignants-chercheurs et chercheurs permanents, ingénieurs, techniciens et administratifs ainsi que 120 doctorants et environ 30 post-doctorants en moyenne. Il s'agit d'une Unité Mixte de Recherche (UMR 5219) dont les tutelles principales sont l'Université Paul Sabatier (UPS), le Centre National de la Recherche Scientifique (CNRS), et l'Institut National des Sciences Appliquées (INSA) de Toulouse.

Les thèmes de recherche couvrent l'ensemble des domaines mathématiques depuis les aspects les plus théoriques jusqu'aux plus appliqués et s'organisent autour de 3 équipes :
- Mathématiques pour l'Industrie et la Physique ;
- Mathématiques Fondamentales Emile Picard ;
- Statistique et Probabilités.

L'équipe Mathématiques pour l'Industrie et la Physique intervient dans le domaine des mathématiques appliquées. Ses axes de recherche couvrent un large spectre incluant la théorie des équations aux dérivées partielles, le calcul scientifique intensif et la visualisation, en passant par la modélisation, l'algorithmique et l'optimisation. L'équipe est structurée autour des thèmes suivants :
- Modélisation - Calcul scientifique ;
- Équations aux dérivées partielles - Systèmes dynamiques ;
- Contrôle - Optimisation - Image.

L'activité scientifique de l'équipe Emile Picard de l'IMT regroupe pour l'essentiel la partie des mathématiques de l'IMT qui relève de la section 25 du CNU et qui est traditionnellement appelée "Mathématiques Pures" :
- Algèbre, Arithmétique, et Géométrie Algébrique ;
- Géométrie et Topologie ;
- Analyse et systèmes dynamiques.

L'Équipe de Statistique et Probabilités couvre tous les domaines de l'aléatoire depuis les plus théoriques comme les applications de la théorie des probabilités à l'algèbre, l'analyse et la géométrie, jusqu'aux plus appliqués comme l'épidémiologie, la biométrie, les mathématiques financières, le traitement du signal et de l'image, la statistique industrielle, le Big Data. L'équipe est structurée autour des thèmes suivants :
- Probabilités et Analyse ;
- Matrices et graphes aléatoires ;
- Calcul stochastique et processus fractionnaires ;
- Modèles markoviens, physique statistique et quantique, théorie ergodique ;
- Statistique fonctionnelle et opératorielle ;

— Statistique en grande dimension et apprentissage ;
— Méthodes de l'aléatoire en interactions.

Certains thèmes, transverses, sont en commun entre l'équipe Mathématiques pour l'Industrie et la Physique et l'équipe Statistique et Probabilités :
— Équipe projet AOC - Apprentissage, Optimisation, Complexité ;
— Mathématiques, Biologie et Santé.

1.2 Institut National de la Recherche Agronomique

Mon stage étant partagé entre IMT et INRA, j'ai également passé du temps au centre INRA Occitanie - Toulouse dans l'unité MIAT et plus précisément l'équipe SaAB. Je vais donc situer l'organisme, le département, l'unité et l'équipe.

1.2.1 L'INRA

L'Institut National de la Recherche Agronomique (INRA) est un organisme de recherche scientifique public, placé sous la double tutelle du Ministère de l'Enseignement Supérieur et de la Recherche et du ministère de l'Agriculture et de l'Alimentation. Il a été créé en 1946 et est constitué aujourd'hui de 13 départements scientifiques, répartis sur 17 centres régionaux. Ses recherches se concentrent sur les questions liées à l'agriculture, à l'alimentation et à la sécurité des aliments, à l'environnement et à la gestion des territoires, avec une perspective de développement durable.

Il a pour missions de :

1. produire et diffuser des connaissances scientifiques ;
2. concevoir des innovations et des savoir-faire pour la société ;
3. éclairer par son expertise, les décisions des acteurs publics et privés ;
4. développer la culture scientifique et technique et participer au débat science/société ;
5. élaborer des stratégies de recherche ;
6. former à la recherche et par la recherche ;
7. promouvoir éthique et déontologie.

Pour cela, l'INRA est présent au niveau mondial et est en permanence au contact des acteurs académiques, économiques, associatifs et territoriaux. Ces différents acteurs agissent au travers de branches scientifiques très diverses : les sciences de la vie en majorité, les sciences des milieux et des procédés, l'ingénierie écologique, les écotechnologies et les biotechnologies, de même que les sciences économiques et sociales et les sciences du numérique.

1.2.2 Département de Mathématiques et Informatique Appliquées

Le de Mathématiques et Informatique Appliquées (MIA) a vocation à mener des recherches en mathématiques et en informatique sur des verrous méthodologiques qui émergent des enjeux prioritaires de la recherche agronomique, et à mettre en œuvre ces recherches via des partenariats (projets, thèses, etc.). Le département a également pour mission de conduire, dans un cadre inter-disciplinaire, des recherches à l'interface sur des

enjeux prioritaires de l'INRA pour lesquels le rôle des mathématiques et de l'informatique, nouveau ou générique, est incontournable. De façon pratique, il vise la production de connaissances génériques et finalisées, la mise au point de méthodes, d'outils et de savoir-faire, applicables aux domaines de l'alimentation, l'agriculture et l'environnement. Enfin, un de ces buts est d'accompagner le développement des mathématiques et de l'informatique à l'INRA, concernant en particulier :

1. l'ingénierie du dispositif INRA en matière de traitement, gestion et analyse de données, de calcul et de simulation, en particulier dans le cadre de plates-formes ;
2. l'expertise en méthodologie mathématiques-informatique et en ingénierie informatique et calcul intensif en direction des départements et des programmes ;
3. la formation, l'entretien de la compétence métier, la diffusion et la promotion de la culture mathématiques-informatique ;
4. le suivi des partenariats entre l'INRA et les autres organismes concernant les mathématiques et l'informatique.

Pour détailler plus avant le contexte de recherche dans lequel s'inscrit mon stage, je vais citer l'axe méthodologique et le champ thématique du département dont mon stage relève :

> Axe Méthodologique 1
> **Extraction de connaissances à partir de données :**
> La science des données se développe actuellement en mixant les apports des différentes communautés scientifiques du numérique. MIA s'y investit pour répondre à la croissance des besoins et à la diversité des questions posées dans les sciences du vivant, en génétique et bioinformatique notamment, mais aussi dans les sciences travaillant à de plus grandes échelles.
>
> Les priorités méthodologiques pour MIA portent sur :
> — les représentations formelles et les traitements permettant d'extraire des connaissances de données issues de sources d'information multiples et hétérogènes (résultats expérimentaux, enquêtes, images, données numériques ou textuelles recueillies via des démarches de science participative ou sur le Web) ;
> — les méthodes d'apprentissage en grande dimension, supervisé ou non, soulevant des questions de réconciliation de données, de réduction de dimension et de visualisation, sur des données ayant souvent une structure spatiale, dynamique ou hiérarchique ;
> — la modélisation, l'inférence statistique et l'échantillonnage de processus spatio-temporels ou de processus ayant pour support des réseaux dynamiques, dans des contextes où les données sont fréquemment dispersées et hétérogènes.

> Champ Thématique 1
> **Bioinformatique et modélisation pour la biologie des systèmes et de synthèse :**
> La bioinformatique, la génomique et la biologie des systèmes sont au cœur de ce champ thématique. Initialement focalisées sur le génome et la compréhension du vivant à des échelles allant de la cellule à l'individu, les recherches

menées à MIA se développent aujourd'hui de l'échelle moléculaire à celle de la population dans des écosystèmes microbiens, ou encore à celle de l'individu eucaryote en interaction avec son environnement. Les enjeux sont d'améliorer la compréhension des mécanismes en jeu et la capacité de les simuler, offrant ainsi la possibilité de maîtriser leur impact sur des fonctions d'intérêt. Pour cela, MIA développe :

— des méthodes de représentation et d'analyse des données issues de la génomique et de la métagénomique, pour détecter et annoter des éléments fonctionnels. Cela correspond aux orientations [#OpenScience-2] [#BioRes-2] du schéma directeur #INRA2025 ;
— des méthodes d'extraction de connaissances, d'apprentissage et d'inférence sur les relations entre gènes, sur les réseaux de régulation et les réseaux métaboliques, à partir de sources d'information diverses et hétérogènes [#OpenScience-3] ;
— des approches de modélisation basées sur la biologie des systèmes et s'appuyant sur des données de génotypage et de phénotypage haut-débit, permettant de simuler des phénotypes microbiens, végétaux et animaux dans des environnements variés [#3Perf-2] ;
— des techniques d'inversion de modèles et d'optimisation pour des objectifs liés à la sélection génétique ou à la biologie de synthèse [#BioRes-1].

1.2.3 Unité Mathématiques et Informatique Appliquées de Toulouse

Mon stage s'est déroulé dans une unité de recherche propre du département MIA : l'unité Mathématiques et Informatique appliquées de Toulouse (MIAT) situé dans le centre INRA Occitanie - Toulouse.

En accord avec les missions de MIA, l'unité MIAT a pour mission de développer et de mettre à jour des méthodes et des compétences en mathématiques et/ou en informatique pour la résolution des problèmes que peuvent rencontrer les chercheurs des autres départements de l'INRA. L'unité est composée de deux grandes équipes de recherche :

— SaAB : Statistique et Algorithmique pour la Biologie
— MAD : Modélisation des Agro-écosystèmes et Décision

Mon stage était rattaché à l'équipe SaAB qui a pour objectif de développer et de mettre à disposition des biologistes des méthodes mathématiques, statistiques et informatiques permettant de contribuer à la compréhension du vivant. En ce sens, l'équipe mobilise et développe ces méthodes en portant une attention particulière à la valorisation de celles-ci dans des outils logiciels directement utilisables par les biologistes. L'équipe s'intéresse à la localisation et l'identification d'éléments fonctionnels dans les génomes des bactéries, plantes et animaux, et de façon croissante aux interactions qui existent entre ces différents éléments :

— au niveau génétique ;
— au niveau ARN/expression des gènes ;
— au niveau protéine.

1.3 Projet CNRS « SCALES »

Enfin, mon stage s'inscrit dans le cadre du projet CNRS « SCALES » dont l'objet est l'étude de l'Inférence statistique multi-échelle et données-dépendante pour la génomique. Je décris ici un peu plus en détails ce projet.

Le projet « SCALES » repose sur le constat qu'il existe un fossé entre les enjeux de l'analyse des données génomiques et les méthodes statistiques existantes. D'un côté, les méthodes statistiques reposent généralement sur une définition a priori et généralement figée d'objets d'intérêt sur lesquels porteront l'inférence, la prédiction, les garanties probabilistes. D'un autre côté, les processus biologiques sont par nature souvent multi-échelles (l'expression génique peut ainsi être étudiée à différents niveaux), et/ou les échelles pertinentes peuvent être dépendantes des données expérimentales (déséquilibre de liaison dans les études GWAS ; domaines topologiquement associés (TAD) dans les études Hi-C). L'idée directrice de ce projet est de combler ce fossé en construisant de nouvelles méthodes statistiques adaptées à cette notion d'échelle, en appliquant ces méthodes à des problématiques biologiques concrètes, et en diffusant ces méthodes dans la communauté biomédicale grâce à des outils d'analyse et de visualisation dédiés, ainsi qu'à des formations.

Ce projet est par nature à l'interface entre le domaine bio-médical, d'une part, et les mathématiques et l'informatique, d'autre part. En effet, il nécessite à la fois la maîtrise de problématiques biologiques et/ou cliniques pointues, ainsi que celle du développement d'outils mathématiques et informatiques pertinents pour répondre à ces problématiques. Il est construit autour de trois applications pilotes qui ont été identifiées comme nécessitant de nouveaux développements statistiques intégrant de façon fondamentale la notion d'échelle. Parmi elles, on note celle dans laquelle s'inscrit mon stage :

> **diff-HiC** : Recherche de TAD (topologically associated domains) différentiels entre deux phases du développement embryonnaire de mammifères, à partir de données de capture de conformation chromosomique de la chromatine (Hi-C). Cette question fait l'objet d'une collaboration entre l'IMT et deux unités de l'INRA de Toulouse, l'unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT) et l'unité de Génétique Physiologie et Systèmes d'Elevage (GenPhySE).

Les technologies de la biologie moléculaire produisent des mesures quantitatives à des échelles de plus en plus fines, atteignant le nucléotide pour le séquençage. Les approches standard font des tests au niveau d'échelle le plus détaillé, négligeant d'une part le fait que les résultats des tests sont liés par des proximités génomiques, et d'autre part que la question d'intérêt ne se pose pas nécessairement au niveau le plus fin mais que l'échelle optimale n'est souvent pas connue a priori. Une innovation majeure de ce projet est de prendre en compte explicitement cet aspect « échelle ». Ce projet passe par une synthèse entre deux branches de la statistique traditionnellement bien distinctes : statistique exploratoire et statistique inférentielle. Cette synthèse est rendue nécessaire par le fait que d'une part, la pratique quotidienne dans les applications génomiques consiste fréquemment à utiliser des méthodes d'inférence statistique (en particulier les test d'hypothèses et les méthodes de prédiction ou de classification supervisée) pour satisfaire un objectif de « génération d'hypothèse », par nature exploratoire. À l'heure actuelle, la théorie statistique

ne fournit pas ou peu de garanties statistiques/probabilistes rigoureuses dans ce type de cas. Par ailleurs, ce projet se situe dans le contexte de « crise de la reproductibilité » que traversent aujourd'hui la science en général et le domaine biomédical en particulier. L'approche développée ici propose des outils d'inférence statistique adaptés à la nature exploratoire des recherches biomédicales. Son application contribuera à limiter les « biais d'optimisme » dans la littérature scientifique. Enfin, une originalité de ce projet est qu'il propose le développement d'outils de visualisation interactive des données génomiques. L'objectif est de permettre une exploration dynamique multi-échelle des données, tout en fournissant à l'utilisateur des garanties statistiques sur les hypothèses testées.

2 Contexte et sujet

Dans cette partie, nous commençons par décrire les éléments de biologie qui contextualisent le sujet. Puis, nous présentons les données Hi-C, qui sont les données qui ont motivé le travail réalisé durant ce stage. Enfin, le cas pratique d'intérêt pour l'application de ce que nous avons étudié pendant le stage est introduit et nous finissons cette partie en exposant le sujet du stage lui-même.

2.1 Contexte biologique

L'acide désoxyribonucléique (ADN) L'acide désoxyribonucléique, ou ADN, est une macromolécule présente dans toutes les cellules vivantes. Elle contient l'ensemble des informations nécessaires pour le développement, le fonctionnement et la reproduction des êtres vivants. Il s'agit d'un code génétique utilisé, entre autres, pour synthétiser les protéines qui sont nécessaires au fonctionnement de la cellule. Cette molécule est formée de deux brins, chacun d'entre eux constitués par une séquence de nucléotides. Chaque nucléotide est composé de trois parties :

— une base nucléique parmi adénine (A), cytozine (C), guanine (G) ou thymine (T)
— un ose, ici le désoxyribose
— un groupe phosphate

Les nucléotides présentent une complémentarité qui leur permet de se lier deux à deux. Ceci permet la structure en double hélice de l'ADN comme illustré dans la figure 1.

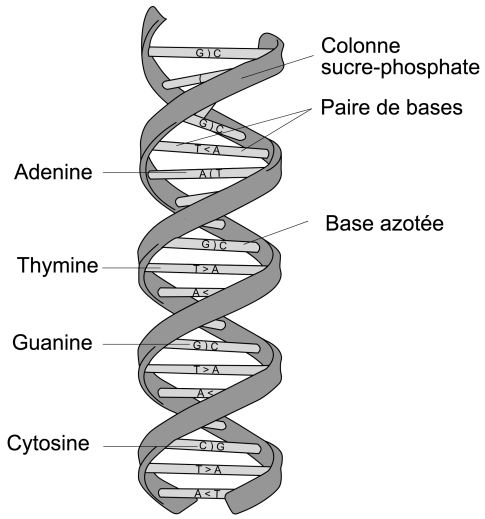

FIGURE 1 – Schéma d'une molécule d'acide désoxyribonucléique

Source: Wikimedia Commons, attribuable à MesserWoland

L'ensemble de l'information génétique codée dans la molécule d'ADN constitue le *génome*.

L'acide ribonucléique messager (ARNm) L'information génétique contenue dans l'ADN n'est pas directement utilisée par la cellule pour la synthèse de protéines. Une étape intermédiaire dite de *transcription* implique la création des acides ribonucléiques messagers. Un ARNm est une copie d'un morceau de l'un des deux brins d'ADN, où la thymine est remplacée par l'uracile. C'est cette molécule d'ARNm qui va ensuite être lue par certains organites de la cellule pour permettre la fabrication d'une ou plusieurs protéines. La fabrication d'une protéine à partir d'un ARNm est l'étape de *traduction*.

L'ensemble des ARN issus de la transcription du génome dans un tissu et des conditions donnés constitue le *transcriptome*.

La figure 2 ci-dessous illustre les mécanismes de transcription et de traduction.

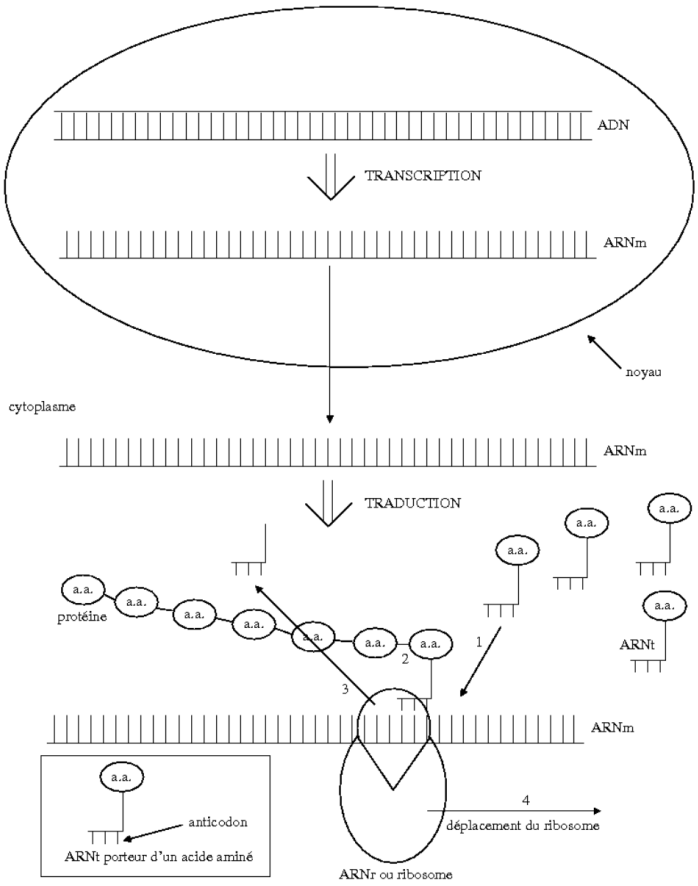

FIGURE 2 – Schéma de la fabrication d'une protéine à partir de la molécule d'ADN

Source: Wikimedia Commons, attribuable à Bionet

Expression génique et régulation Pour simplifier, on appellera *gène* une séquence d'ADN susceptible d'être transcrite en ARN. Il existe, au sein de la cellule, des mécanismes complexes d'intéractions entre gènes permettant de réguler (activer ou inhiber) leur expression et donc augmenter ou diminuer les quantités des produits de l'expression des gènes (ARN et protéines). Ces mécanismes permettent, entre autre, une adaptation des organismes vivants à leur environnement. Il existe plusieurs mécanismes connus ou supposés de régulation parmi lesquels on peut citer la *méthylation* de l'ADN, l'expression d'*ARN non codants* (ou de petits ARN), ou encore la conformation spatiale des chromosomes dans la cellule.

L'organisation spatiale du génome Lors de la mitose, le matériel génétique des cellules eucaryotes s'organise en chromosomes condensés et bien délimités. Cependant, le reste du temps, l'information génétique est stockée dans une structure complexe très compacte composée d'ADN, d'ARN et de protéines : la chromatine. On peut avoir un aperçu de l'organisation spatiale du matériel génétique grâce à la figure 3.

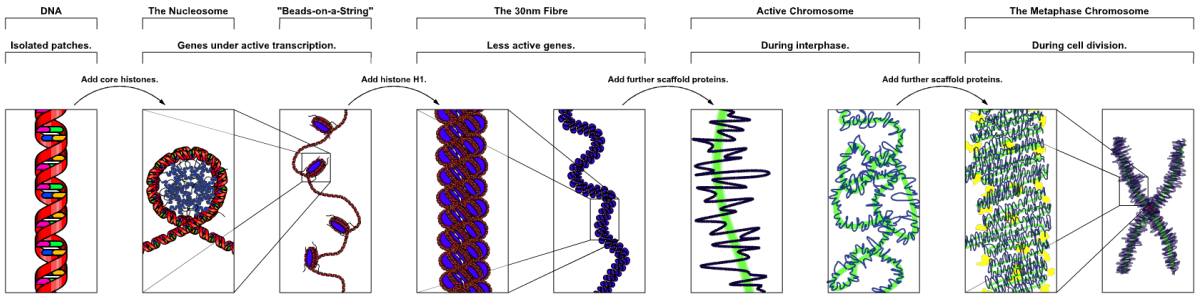

FIGURE 3 – Schéma de l'organisation de la chromatine

Source: Wikimedia Commons, attribuable à Richard Wheeler

Les techniques de Capture de la Conformation des Chromosomes (3C), dont fait partie le Hi-C, nous permettent d'obtenir des informations sur l'organisation en trois dimensions de cette chromatine. En effet, les techniques 3C mettent en évidence les régions du génome en contact physique et font apparaître la structure hiérarchique de la chromatine.

Il est désormais reconnu que l'organisation spatiale du génome dans la cellule a un impact important sur les phénomènes de régulation entre gènes, avec des implications dans la différenciation cellulaire [Dixon et al., 2015] ou encore le développement de certaines maladies [Lupiáñez et al., 2015]. C'est pourquoi obtenir des informations quant à l'organisation tri-dimensionnelle du génome dans la cellule est un enjeu de la recherche en génomique et en épigénétique.

2.2 Les données Hi-C et les domaines topologiques associés

Présentation Les données Hi-C sont des données issues du séquençage haut-débit de nouvelle génération. Leur particularité est qu'elles nous permettent d'avoir accès à une mesure de la proximité spatiale entre paires de positions à travers l'ensemble du génome contrairement aux autres méthodes de capture de la conformation des chromosomes qui s'attachent à des régions spécifiques ciblées.

Acquisition des données Pour obtenir ces données, [Lieberman-Aiden et al., 2009] décrivent un protocole expérimental qui est illustré dans la figure 4.

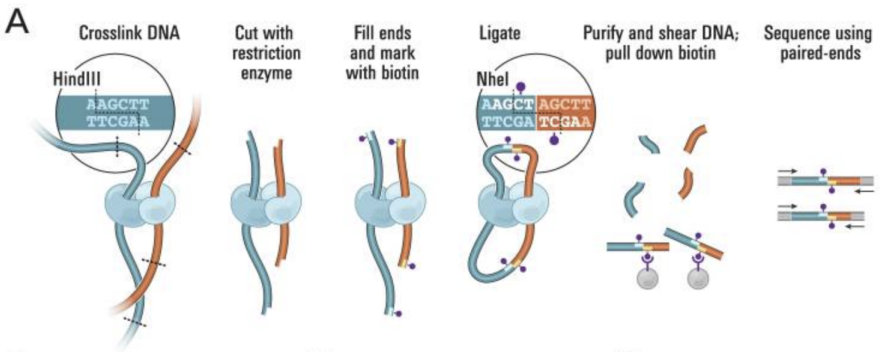

FIGURE 4 – Procédure pour l'acquisition de données Hi-C

Source: [Lieberman-Aiden et al., 2009]

On commence à partir d'une culture cellulaire et on procède à une *réticulation* qui consiste à enchaîner les fragments d'ADN spatialement proches les uns des autres, c'est-à-dire, deux *loci* (positions dans le génome) spatialement proches dans la chromatine. Ensuite, on découpe ces morceaux d'ADN enchaînés à l'aide d'une enzyme de restriction puis on en lie les extrémités afin d'obtenir un segment d'ADN dans lequel les deux loci se suivent. On obtient alors un ensemble de fragments d'ADN, chacun étant issus de deux loci à l'origine spatialement proches dans le noyau : il s'agit d'une *librairie de séquençage*. Cette première étape constitue un protocole 3C standard. Pour passer au Hi-C, il faut une dernière étape de séquençage haut-débit. Chaque fragment est lu, ce qui donne un ensemble de *reads*, c'est-à-dire, un ensemble de petites séquences de quelques centaines d'acides nucléiques qui correspondent aux séquences proches des deux loci initialement enchaînés. on obtient ce que l'on appelle une paire de *reads*. Ces reads sont ensuite alignés sur le génome de référence afin de déterminer les coordonnées génomiques des deux séquences correspondant aux deux loci initialement enchaînés : une paire de reads lue correspond donc à deux régions du génome spatialement proches dans la cellule. L'ensemble des traitements biologiques et bioinformatiques permettant de passer de l'échantillon biologique aux données Hi-C telles que décrites dans la section suivante est décrit dans la revue de [Ay and Noble, 2015].

Description et visualisation des données Les données Hi-C se présentent sous la forme d'une matrice S de comptages du nombre d'interactions ou « matrice de contact ». L'entrée $s(i, j)$ de la matrice correspond au nombre d'interactions lues entre les positions génomiques i et j. Le protocole Hi-C ne permettant pas d'obtenir des résultats à l'échelle des bases individuellement, chaque position correspond à un segment de longueur (en paire de bases de nucléotides) prédéfinie appelé *bin*. Typiquement, la résolution de ces segments est comprise entre 40 000 et 500 000 paires de nucléotides. Une des particularités de ces données est que la plupart des entrées de la matrice de données sont nulles car il y a généralement peu d'interactions entre zones éloignées sur le long du chromosome.

Les données Hi-C sont donc des matrices symétriques, à entrées positives et creuses. Elles peuvent être représentées par des *heatmaps* triangulaires correspondant à des demi-matrices de contact. La base horizontale du triangle correspond aux positions génomiques (diagonale de la matrice de contact) et l'intensité en couleur représente à la valeur du coefficient de la matrice. La figure 5 est un exemple de représentation de carte Hi-C.

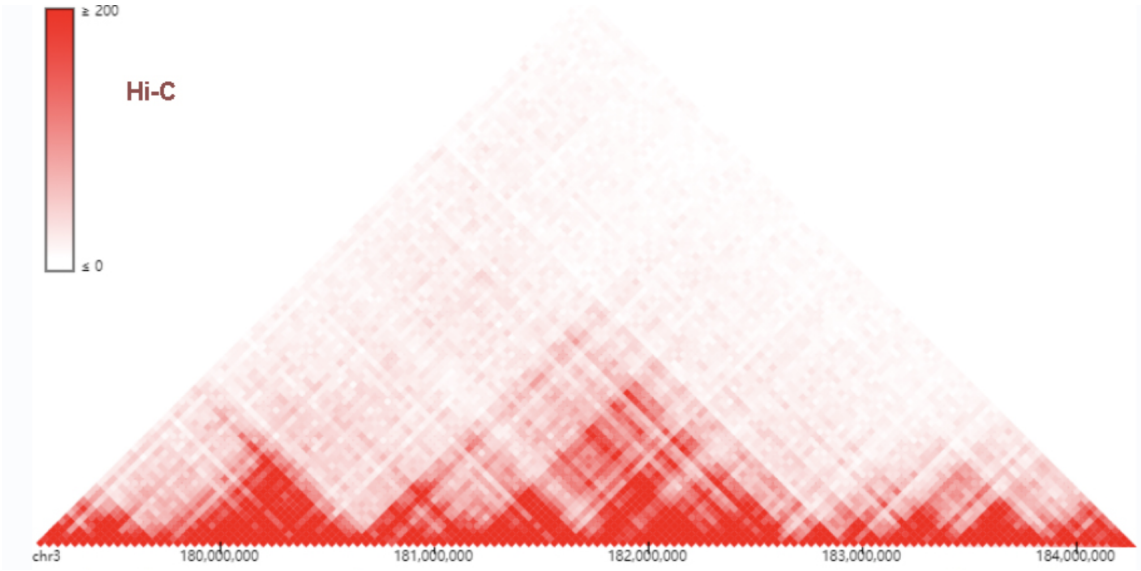

FIGURE 5 – Représentation graphique d'une carte Hi-C

Source: Feng Yue Lab (Penn State)

Application à l'étude des domaines topologiques associés Ces données ont permis de mieux comprendre les mécanismes de l'organisation spatiale du génome dans la cellule et de la structure hiérarchique de la chromatine. En particulier, elles ont permis de mettre en évidence l'existence de régions du génome spatialement proches dans le noyau de la cellule appelés *domaines topologiques associés (TAD)* [Dixon et al., 2012, Sexton et al., 2012]. Il s'agit de zones du chromosome particulièrement compactées, comme illustré sur la figure 6.

FIGURE 6 – Schéma de trois domaines topologiques associés

Source: [Gómez-Díaz and Corces, 2014]

Fonction des domaines topologiques associés

Des études, comme par exemple celle de [Bonev and Cavalli, 2016], ont montré que ces régions ont un rôle lors de la régulation de l'expression des gènes. La proximité spatiale semble impliquée dans l'activation de la transcription [Dixon et al., 2016] et il a été montré que les gènes du même domaine topologique associé présentent des profils d'expression similaires [Nora et al., 2012] et pourraient donc être co-régulés. Cette organisation spatiale est fortement lié au fonctionnement de la cellule : plusieurs études, comme celles de [Giorgio et al., 2015] et [Lupiáñez et al., 2015], ont montré que la fusion de deux domaines topologiques associés, initialement séparés, pouvait entraîner une sur-expression génique délétère pour l'organisme.

Détection des domaines topologiques associés Plusieurs méthodes [Dixon et al., 2012, Levy-Leduc et al., 2014, Weinreb and Raphael, 2015,

Serra et al., 2016] basées sur des heuristiques existent pour la recherche de TADs (voir la revue de [Forcato et al., 2017] pour une comparaison de ces méthodes). La détection des domaines topologiques associés peut être vue comme un problème de segmentation du génome dont le but est d'obtenir une partition du génome en segments contigus (chacun d'entre eux correspondant à un TAD).

Dans le cadre de ce stage, la classification ascendante hiérarchique (CAH), plus spécifiquement sa version sous contrainte de contiguïté le long du chromosome, est la méthode qui a été retenue.

Contrairement aux approches standard, la version utilisée pour traiter des données Hi-C doit prendre en entrée non pas une matrice de distance mais une matrice de similarité, donnée directement par les comptages de la matrice Hi-C, $S = (s_{ij})_{i,j=1,\dots,n}$. L'avantage de cette approche, contrairement à la plupart des méthodes existantes de recherche de TADs, est qu'en plus de la segmentation du chromosome en classes, elle fournit un modèle hiérarchique de la conformation spatiale du chromosome dans la cellule, modèle qui pourra être utilisé dans des études de comparaison.

2.3 Cas pratique

Le constat motivant le cas pratique de ce stage est le suivant : dans les élevages porcins, le progrès génétique des dernières années s'est accompagné d'une augmentation significative de la mortalité des porcelets. La maturité du porcelet, définie comme l'état de plein développement permettant la survie à la naissance, est un déterminant important de la mortalité précoce mais ces mécanismes sont encore largement méconnus.

Dans le cadre du projet SCALES, 2 jeux de 3 cartes Hi-C ont été produits par l'INRA de Toulouse. Ils correspondent à des échantillons de fœtus de porcs obtenus à deux stades de développement embryonnaire (90 et 110 jours) pour 3 individus. Des études préliminaires ont montré une évolution de l'organisation spatiale de l'ADN dans les cellules musculaires entre ces deux stades et l'objectif à terme sera d'analyser plus finement cette réorganisation.

Ainsi, le but de ce travail est de fournir des résultats et des méthodes applicables à ces données et de permettre d'améliorer la compréhension des mécanismes physiologiques permettant la survie à la naissance en identifiant les loci concernés par la réorganisation spatiale de la chromatine en fin de gestation.

2.4 Sujet

Ce projet est donc motivé par deux questions biologiques qui se prêtent naturellement à une modélisation statistique dédiée : d'une part l'identification de domaines topologiques associés à partir de données Hi-C avec la CAH contrainte, et d'autre part la comparaison de la conformation spatiale entre deux conditions biologiques différentes.

Ce stage s'est plus précisément focalisé sur la classification ascendante hiérarchique en étudiant ses extensions à des données de similarité et au cas contraint. En particulier, j'ai étudié le cadre formel de ces extensions ainsi que les propriétés théoriques des résultats de la méthode dans ces diverses extensions. Enfin, j'ai réalisé en fin de stage une étude bibliographique sur les méthodes de comparaison des résultats de CAH pour répondre à la question biologique d'intérêt décrite ci-dessus dans de futurs travaux.

3 Classification Ascendante Hiérarchique

La Classification Ascendante Hiérarchique (CAH) est une méthode d'apprentissage non supervisée dont le but est la partition automatique d'objets (ou individus) en sous-groupes d'individus similaires pour une certaine mesure de ressemblance. Elle fait partie des méthodes dites de *regroupement hiérarchique*. Dans la suite, on notera Ω un ensemble de n individus, $\{x_1, \ldots, x_n\}$, à partitionner.

Dans la suite, je présenterai, dans la section 3.1 le cadre formel standard de la CAH puis je m'intéresserai aux extensions de cette approche, d'une part à des données décrites par des dissimilarités non euclidiennes ou par des similarités dans la section 3.2 et, d'autre part, à une version permettant d'incorporer des contraintes de contiguïté dans les classes dans la section 3.3.

3.1 Classification ascendante hiérarchique : version standard

Le cadre standard de la CAH considère que les individus sont décrits par une relation de dissimilarité, d. Nous noterons aussi $D = (d_{ij})_{1 \leq i,j \leq n}$, avec $d_{ij} = d(x_i, x_j)$. Dans cette section, on suppose que cette dissimilarité est une distance euclidienne. Dans ce cas, on supposera (sans perte de généralité) que $(x_i)_i \in \mathbb{R}^p$ et que $d(x_i, x_j) = \|x_i - x_j\|$ où $\|.\|$ désigne la norme usuelle de $\mathbb{R}^p$, associée au produit scalaire $\langle \cdot, \cdot \rangle$.

L'algorithme standard prend la matrice de dissimilarité $D = (d_{ij})_{1 \leq i,j \leq n}$ comme entrée. Le résultat final de la classification ascendante hiérarchique consistera en une suite de partitions imbriquées, ou *hiérarchie* :

Définition 1 (Hiérarchie et hiérarchie indicée)**.**
Une hiérarchie $(\mathcal{P}_t)_{t=1,\ldots,T}$ d'un ensemble Ω est un sous-ensemble de $\mathcal{P}(\Omega)$ vérifiant :
— $\Omega \in (\mathcal{P}_t)_t$
— $\forall x \in \Omega, \quad \{x\} \in (\mathcal{P}_t)_t$
— $\forall (\mathcal{P}, \mathcal{P}') \in (\mathcal{P}_t)_t^2, \quad \mathcal{P} \cap \mathcal{P}' \in \{\emptyset, \mathcal{P}, \mathcal{P}'\}$

On appelle indice sur une hiérarchie $(\mathcal{P}_t)_{t=1,\ldots,T}$ de Ω une fonction j de $(\mathcal{P}_t)_{t=1,\ldots,T}$ dans $\mathbb{R}^+$ vérifiant les propriétés :
— *si $\mathcal{P} \subset \mathcal{P}'$ avec $\mathcal{P} \neq \mathcal{P}'$ alors $j(\mathcal{P}) < j(\mathcal{P}')$*
— $\forall x \in \Omega, \quad j(\{x\}) = 0$

Le couple $((\mathcal{P}_t)_t, j)$ est alors appelée une hiérarchie indicée.

Le principe de l'algorithme CAH consiste à commencer par une partition triviale, $\mathcal{P}_1$, où chaque classe de la partition est un singleton, puis par fusions successives, l'algorithme se termine avec une autre partition triviale, $\mathcal{P}_n$, composée d'une unique classe dans laquelle sont regroupés tous les objets. A chaque étape t, permettant de passer de la partition $\mathcal{P}_t$ à la partition $\mathcal{P}_{t+1}$, composée d'une classe de moins, deux des classes de $\mathcal{P}_t$ sont fusionnées en une unique classe dans $\mathcal{P}_{t+1}$. Les fusions sont choisies de sorte à minimiser une certaine quantité définie par le choix d'un critère de lien. Le choix d'un lien définit une *dissimilarité entre sous-ensembles disjoints* de Ω, notée δ, à partir de la donnée de D : à chaque étape, les deux classes de la partition courante les plus proches selon ce lien δ sont fusionnées.

Dans la suite, je me focaliserai sur le *lien de Ward* (introduit par [Ward, 1963]) qui est le plus couramment utilisé en statistique, car il a une interprétation simple. Celui-ci est basé sur la notion d'*inertie intra-classes*.

Définition 2 (Inertie et inertie intra-classes)**.**

Soit $\mathcal{P}$, $\mathcal{P} = (\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_K)$, une partition en K classes de Ω, l'inertie d'une classe $\mathcal{C}_k$ de cardinal μ_k est la quantité :

$$\mathcal{I}(\mathcal{C}_k) = \sum_{x_i \in \mathcal{C}_k} \|x_i - g_k\|^2$$

avec $g_k = \frac{1}{\mu_k} \sum_{x_i \in \mathcal{C}_k} x_i$, le centre de gravité de la classe $\mathcal{C}_k$.

De manière similaire, l'inertie intra-classes est la somme des inerties de chacune des classes :

$$\mathcal{I}_{intra}(\mathcal{P}) = \sum_{k=1}^{K} \mathcal{I}(\mathcal{C}_k)$$

Le lien de Ward entre deux sous-ensembles disjoints est défini comme l'augmentation de l'inertie intra-classe suite à la fusion de ces deux classes, par rapport à l'état précédent :

Définition 3 (Lien de Ward)**.** *Soit $(A, B) \subset \Omega^2$ disjoints, la valeur du lien de Ward entre A et B, notée $\delta(A, B)$ est définie par :*

$$\delta(A, B) = \mathcal{I}(A \cup B) - \mathcal{I}(A) - \mathcal{I}(B) \tag{1}$$

Remarque. *Une formulation équivalente du lien de Ward permet de l'exprimer comme une distance entre centres de gravité des classes [Kaufman and Rousseeuw, 1990] :*

$$\forall (A, B) \subset \Omega^2 \text{ disjoints}, \quad \delta(A, B) = \frac{\mu_A \mu_B}{\mu_A + \mu_B} \|g_A - g_B\|^2.$$

Démonstration.

On a la formulation équivalente suivante en termes de produit scalaire pour l'inertie :

$$\mathcal{I}(\mathcal{C}_i) = \sum_{x_j \in \mathcal{C}_i} \langle x_j - g_i, x_j - g_i \rangle = \frac{1}{\mu_i^2} \sum_{(x_j, x_k, x_l) \in \mathcal{C}_i^3} \langle x_j - x_k, x_j - x_l \rangle = \sum_{x_j \in \mathcal{C}_i} \|x_j\|^2 - \frac{1}{\mu_i} \sum_{(x_k, x_l) \in \mathcal{C}_i^2} \langle x_k, x_l \rangle$$

En utilisant l'équation précédente, on peut réécrire l'équation (1) :

$$\begin{aligned}
\delta(A, B) &= \frac{1}{\mu_A} \sum_{(x_i, x_j) \in A^2} \langle x_i, x_j \rangle + \frac{1}{\mu_B} \sum_{(x_i, x_j) \in B^2} \langle x_i, x_j \rangle - \frac{1}{\mu_A + \mu_B} \sum_{(x_i, x_j) \in A \cup B^2} \langle x_i, x_j \rangle \\
&= \frac{\mu_B}{\mu_A(\mu_A + \mu_B)} \sum_{(x_i, x_j) \in A^2} \langle x_i, x_j \rangle + \frac{\mu_A}{\mu_B(\mu_A + \mu_B)} \sum_{(x_i, x_j) \in B^2} \langle x_i, x_j \rangle - \frac{2}{\mu_A + \mu_B} \sum_{(x_i, x_j) \in A \times B} \langle x_i, x_j \rangle
\end{aligned}$$

et donc,

$$\begin{aligned}
\frac{\mu_A \mu_B}{\mu_A + \mu_B} \|g_A - g_B\|^2 &= \frac{\mu_A \mu_B}{\mu_A + \mu_B} \frac{1}{\mu_A^2 \mu_B^2} \sum_{(x_i, x_j) \in A^2} \sum_{(x_k, x_l) \in B^2} \langle x_i - x_k, x_j - x_l \rangle \\
&= \frac{1}{\mu_A + \mu_B} \left(\frac{\mu_B}{\mu_A} \sum_{(x_i, x_j) \in A^2} \langle x_i, x_j \rangle + \frac{\mu_A}{\mu_B} \sum_{(x_i, x_j) \in B^2} \langle x_i, x_j \rangle - 2 \sum_{(x_i, x_j) \in A \times B} \langle x_i, x_j \rangle \right) \\
&= \delta(A, B)
\end{aligned}$$

□

Pour éviter d'avoir à recalculer l'intégralité des valeurs des liens entre classes à chaque étape de l'algorithme, il existe une formule permettant une mise à jour rapide de celles-ci, la relation de Lance-Williams [Lance and Williams, 1967] :

$$\forall\, A,\, B,\, G \in \mathcal{P}_t, \qquad \delta(G, A \cup B) = \alpha_A \delta(G, A) + \alpha_B \delta(G, B) + \beta \delta(A, B).$$

Cette formule est générale pour la CAH et les valeurs précises dans le cadre du lien de Ward sont [Cormack, 1971] :

$$\alpha_A = \frac{\mu_A + \mu_G}{\mu_A + \mu_B + \mu_G}, \qquad \alpha_B = \frac{\mu_B + \mu_G}{\mu_A + \mu_B + \mu_G}, \qquad \beta = -\frac{\mu_G}{\mu_A + \mu_B + \mu_G}.$$

3.2 Extensions aux cas non euclidiens

Dans cette partie, nous étendons le cadre de la CAH à des données qui ne sont pas seulement décrites par une distance euclidienne. De manière plus précise, je décrirai trois extensions possibles, le cas des données décrites par des dissimilarités, le cas des données décrites par des noyaux et le cas des données décrites par des similarités quelconques.

3.2.1 Dissimilarité arbitraire

Dans cette section, on montre comment étendre (et justifier) le cadre de la CAH avec lien de Ward à des données décrites par des dissimilarités quelconques. Dans cette partie, on suppose que les $(x_i)_i$ prennent leurs valeurs dans un espace, $\mathcal{X}$, quelconque (non nécessairement euclidien) et qu'ils sont décrits par une relation de dissimilarité qui vérifient les propriétés suivantes :

— d est à valeurs positives : $d_{ij} \geq 0$;
— d est à diagonale nulle : $d_{ii} = 0$;
— d est symétrique : $d_{ij} = d_{ji}$.

Lorsqu'en outre, la dissimilarité vérifie l'inégalité triangulaire ($d_{ij} \leq d_{ik} + d_{kj}$), elle est dit métrique. En particulier, une dissimilarité issue du cadre euclidien est toujours métrique même si la réciproque est fausse en général.

Une des approches pour étendre une méthode basée sur une distance euclidienne à des données décrites par des dissimilarités quelconques consiste à transformer la dissimilarité en distance euclidienne. Par exemple, [Lingoes, 1971] montre qu'il existe $c_1 \in \mathbb{R}$ tel que $\tilde{d}$ définie par $\tilde{d}_{ij} = \sqrt{d_{ij}^2 + c_1}$ est euclidienne et [Cailliez, 1983] montre qu'il existe $c_2 \in \mathbb{R}$ tel que $\tilde{d}$ définie par $\tilde{d}_{ij} = d_{ij} + c_2$ est euclidienne.

Dans le cas d'une dissimilarité quelconque (non euclidienne), [Chavent et al., 2017] définit le lien de Ward, δ, de façon analogue à l'aide de la *pseudo-inertie intra-classes* qui est l'extension directe de la notion d'inertie pour $d_{ij}^2 = \|x_i - x_j\|^2$ (ce qui assure donc l'équivalence entre les deux notions dans le cas euclidien) :

Définition 4 (Pseudo-inertie et pseudo-inertie intra-classes)**.** *Soit $\mathcal{P}$, $\mathcal{P} = (\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_K)$, une partition en K classes de Ω, on appelle pseudo-inertie d'une classe $\mathcal{C}_k$ la quantité*

$$\tilde{\mathcal{I}}(\mathcal{C}_k) = \frac{1}{2\mu_k} \sum_{(x_i, x_j) \in \mathcal{C}_k^2} d_{ij}^2,$$

avec μ_k le cardinal de la classe $\mathcal{C}_k$, et pseudo-inertie intra-classes est la somme des pseudo-inerties de chacune des classes :

$$\tilde{\mathcal{I}}_{intra}(\mathcal{P}) = \sum_{k=1}^{K} \tilde{\mathcal{I}}(\mathcal{C}_k).$$

Le lien de Ward se généralise donc au cadre de dissimilarités quelconques par utilisation de la pseudo-inertie au lieu de l'inertie. Aussi, dans la suite, on omettra la distinction entre inertie et pseudo-inertie.

Dans ce contexte généralisé, le lien défini à l'aide de la pseudo-inertie intra-classes ne peut plus s'interpréter comme dans le cadre euclidien. On peut le voir comme un critère d'homogénéité qu'on cherche à optimiser à chaque étape mais le contexte n'est plus aussi bien défini. Ceci sera étudié plus en détail dans la section 4.3.

3.2.2 Noyau

Dans un certain nombre de cas pratiques (dont les données Hi-C font partie), les objets sont décrits par leurs ressemblances et non leurs dissemblances. C'est le cas, en particulier, lorsque les données sont décrites par un noyau [Schölkopf et al., 2004] ou par une mesure de similarité. Il existe, dans certains cas, des façons naturelles de passer d'une similarité à une dissimilarité et réciproquement. Toutefois, il n'y a pas toujours d'équivalence stricte entre ces deux notions, et on peut donc souhaiter réaliser une classification hiérarchique directement à partir d'une similarité. Dans cette section, nous décrivons comment la CAH peut être ré-écrite pour des données décrites par un noyau puis étendrons ce cadre, dans la section suivante, à des données décrites par des similarités quelconques.

On suppose donc ici que les données $(x_i)_i$ sont décrites par un noyau c'est-à-dire par une fonction $k : \mathcal{X}^2 \to \mathbb{R}$ qui, à chaque paire d'objet (x_i, x_j) associe une valeur $k_{ij} = k(x_i, x_j)$ et qui vérifie les propriétés suivantes :

— k est symétrique : $k(x_i, x_j) = k(x_j, x_i)$;

— k est positive : $\forall N \in \mathbb{N}$, $\forall (x_i)_{i=1,\dots,N} \subset \mathbb{N}$ and $\forall (\alpha_i)_{i=1,\dots,N} \subset \mathcal{X}$, $\sum_{i,j=1}^{N} \alpha_i \alpha_j k(x_i, x_j) \geq 0$.

Ces propriétés sont équivalentes au fait que la matrice $K = (k(x_i, x_j))_{i,j=1,\dots,n}$ est symétrique définie positive.

Dans ce cadre, [Aronszajn, 1950] prouve qu'il existe un unique espace de Hilbert, $(\mathcal{H}, \langle ., . \rangle_{\mathcal{H}})$, appelé espace de représentation, et une unique application $\phi : \mathcal{X} \to \mathcal{H}$, telle que le noyau k correspond au produit scalaire dans $\mathcal{H}$:

$$\forall\, x, x' \in \mathcal{X}, \qquad k(x, x') = \langle \phi(x), \phi(x') \rangle.$$

La CAH s'étend donc de manière évidente au cas de données décrites par un noyau en utilisant l'espace de Hilbert sous-jacent et la distance euclidienne définie dans cet espace de Hilbert :

$$d(x_i, x_j) := \sqrt{k_{ii} + k_{jj} - 2k_{ij}}. \tag{2}$$

Le critère de Ward s'interprète donc à nouveau comme la minimisation de l'augmentation de l'inertie intra-classe, calculée dans l'espace de représentation $\mathcal{H}$.

Dans ce cas particulier, on dispose alors d'une réécriture du critère de Ward directement en fonction du noyau k [Dehman, 2015] :

$$\delta(A, B) = \frac{\mu_A \mu_B}{\mu_A + \mu_B} \left(\frac{1}{\mu_A^2} K_{A,A} + \frac{1}{\mu_B^2} K_{B,B} - \frac{2}{\mu_A \mu_B} K_{A,B} \right) \tag{3}$$

avec $K_{A,B} = \sum_{x_i \in A, x_j \in B} k(x_i, x_j)$.

Démonstration. Dans l'espace de représentation, le lien de Ward s'écrit de la façon suivante :

$$\delta(A, B) = \frac{\mu_A \mu_B}{\mu_A + \mu_B} \|g_A - g_B\|_{\mathcal{H}}^2,$$

avec $g_C := \frac{1}{\mu_C} \sum_{x_i \in C} \phi(x_i)$ le centre de gravité d'une classe C, dans l'espace de représentation. On a

$$\begin{aligned} \|g_A - g_B\|^2_{\mathcal{H}} &= \langle g_A - g_B, g_A - g_B \rangle_{\mathcal{H}} \\ &= \langle g_A, g_A \rangle_{\mathcal{H}} + \langle g_B, g_B \rangle_{\mathcal{H}} - 2\langle g_A, g_B \rangle_{\mathcal{H}} \end{aligned}$$

Et donc, en remarquant que

$$\langle g_A, g_B \rangle_{\mathcal{H}} = \frac{1}{\mu_A \mu_B} \sum_{x_i \in A, x_j \in B} k(x_i, x_j),$$

on obtient le résultat attendu. □

Ainsi, le cas des données sous forme de noyaux est strictement équivalent au cas classique (c'est-à-dire sous formes de distances euclidiennes). En revanche, on dispose de la formule (3) pour le traiter directement sans avoir à utiliser de dissimilarités. Ce cas va servir de base pour définir l'extension de la CAH aux similarités quelconques.

3.2.3 Similarités

Les similarités constituent un autre type standard de données où les n objets $(x_i)_i$ sont décrits par une relation de ressemblance, $s_{ij} = s(x_i, x_j)$ pour $i, j = 1, \ldots, n$. Bien qu'il n'y ait pas de consensus sur la définition exacte d'une similarité, habituellement, ces mesures prennent de grandes valeurs positives pour des objets semblables et de petites valeurs positives pour des objets dissemblables. Dans certains cas, il est pertinent de considérer des similarités avec des coefficients négatifs. Elles peuvent être perçues comme une généralisation du produit scalaire ou une extension des noyaux. Pour la suite, on supposera que la similarité s considérée est symétrique : $s_{ij} = s_{ji}$. On notera également $S = (s_{ij})_{i,j=1,\ldots,n}$ la matrice des similarités paire à paire sur le jeu de données.

Dans le cas de données décrites par une similarité, et par opposition au cas des noyaux, il n'y a pas de façon « naturelle » pour passer d'une similarité à une dissimilarité. Par exemple, il est fréquent, si s vérifie $\forall\, i \in [\![1, n]\!],\ s_{ii} = M$, de considérer la dissimilarité $d_M(x_i, x_j) := M - s_{ij}$ qui est une dissimilarité arbitraire pour laquelle la notion de pseudo-inertie (section 3.2.1) s'applique. Si, en outre, la matrice S est définie positive (s est donc un noyau), l'équation (2) permet de calculer la distance euclidienne induite par le noyau s :

$$d_s^2(x_i, x_j) := s_{ii} + s_{jj} - 2s_{ij}.$$

d_M et d_s sont liées par la relation suivante :

$$2d_M(x_i, x_j) = d_s^2(x_i, x_j).$$

mais ne donneront pas les mêmes résultats lorsque fournies en entrée d'une CAH.

Pour fournir un cadre plus objectif pour ce type de données, nous avons choisi de nous appuyer sur l'analogie similarité / noyau et d'utiliser un analogue de l'équation (3) pour les similarités :

$$\delta(A, B) := \frac{\mu_A \mu_B}{\mu_A + \mu_B} \left(\frac{1}{\mu_A^2} S_{A,A} + \frac{1}{\mu_B^2} S_{B,B} - \frac{2}{\mu_A \mu_B} S_{A,B} \right) \tag{4}$$

avec $S_{A,B} = \sum_{x_i \in A, x_j \in B} s(x_i, x_j)$.

Cette approche est justifiée dans le cas particulier où la similarité considérée est un noyau. Dans les autres cas, il n'existe pas de distance euclidienne sous-jacente dont s serait le produit scalaire et la quantité $s_{ii} + s_{jj} - 2s_{ij}$, analogue à la distance sous-jacente à s au carré, peut même être négative, ce qui rend difficile l'interprétation des notions d'inertie et pseudo-inertie dans ce cadre.

Toutefois, il existe une relation simple permettant de passer d'une similarité quelconque à un noyau :

Proposition 1. *Soit S une similarité quelconque*

1. *Il existe $\lambda > 0$ tel que la similarité k_λ définie comme suit est un noyau :*

$$k_\lambda(x_i, x_j) := s_{ij} + \mathbf{1}_{\{i=j\}}\lambda$$

2. *On note δ_λ le lien de Ward pour le noyau k_λ. Si A_t et B_t sont fusionnés à l'étape t, alors*

$$\delta_\lambda(A_t, B_t) = \delta(A_t, B_t) + \lambda.$$

Démonstration de 2. Notons, pour A, $B \subset \Omega$, $K_{A,B} = \sum_{x_i \in A, x_j \in B} k_\lambda(x_i, x_j)$. On a

$$\begin{aligned}
\delta_\lambda(A, B) &= \frac{\mu_A \mu_B}{\mu_A + \mu_B}\left(\frac{1}{\mu_A^2}K_{A,A} + \frac{1}{\mu_B^2}S_{B,B} - \frac{2}{\mu_A\mu_B}S_{A,B}\right) \\
&= \frac{\mu_A \mu_B}{\mu_A + \mu_B}\left(\frac{1}{\mu_A^2}S_{A,A} + \frac{\mu_A}{\mu_A^2}\lambda + \frac{1}{\mu_B^2}S_{B,B} + \frac{\mu_B}{\mu_B^2}\lambda - \frac{2}{\mu_A\mu_B}S_{A,B}\right) \\
&\qquad \text{par définition de } k_\lambda \\
&= \delta(A, B) + \lambda
\end{aligned}$$

□

Ainsi, étant donnée une similarité quelconque, la CAH induite par le lien δ_λ et la CAH induite par le lien δ ont des suites de partitions identiques et les niveaux de fusion sont translatés de $(+\lambda)$ dans le second cas par rapport au premier. Ce résultat, qui justifie l'utilisation de l'équation (4) dans le cas d'une similarité quelconque, est l'analogue de celui donné dans [Miyamoto et al., 2015] sur les dissimilarités (ou distances) sous-jacentes à s et k_λ.

3.3 Classification ascendante hiérarchique sous contrainte de contiguïté

Dans le cas de la classification ascendante hiérarchique sous contrainte de contiguïté (CAHCC), on suppose que seulement deux classes adjacentes peuvent être fusionnées. Ce point de vue est un cas particulier de celui décrit dans [Ferligoj and Batagelj, 1982] pour lequel la classification est réalisée sous des contraintes définies par une relation symétrique arbitraire. Pour $t = 1, \ldots n-1$, décrivons formellement le passage de la partition $\mathcal{P}_t$ à $\mathcal{P}_{t+1}$ dan s le cas contraint. Les $n - t + 1$ classes de la partition $\mathcal{P}_t$ sont désignées par G_1^t, G_2^t, …, G_{n-t+1}^t. On note

$$u^* = \text{argmin}_{u=1,\ldots,n-t+1}\, \delta(G_u^t, G_{u+1}^t)$$

Les classes de la partition $\mathcal{P}_{t+1}$ sont donc :

$$\forall\, u = 1, \ldots, n-t, \qquad G_u^{t+1} = \begin{cases} G_u^t & \text{si } u < u^* \\ G_u^t \cup G_{u+1}^t & \text{si } u = u^* \\ G_{u+1}^t & \text{si } u > u^* \end{cases}$$

Les valeurs de la dissimilarité δ entre classes $(G_u^{t+1})_{u=1,\ldots,n-t}$ peuvent être obtenues de façon standard, en utilisant la formule de Lance-Williams. Du point de vue algorithmique, l'introduction de cette contrainte de contiguïté permet de passer d'une complexité (en nombre d'opérations) cubique ($O(n^3)$) à une complexité quadratique ($O(n^2)$).

4 Propriétés des dendrogrammes

Dans cette partie nous introduisons la notion de dendrogramme qui est un outil classique de représentation graphique des résultats de classification ascendante hiérarchique. Un dendrogramme est une représentation sous forme d'arbre de la hiérarchie induite par la CAH. Cet arbre peut-être pondéré ou non, ce qui induit deux points de vue, topologique ou pondéré.

La section 4.1 a pour but de définir la notion de dendrogramme et de hauteur et la section 4.2 introduit des hauteurs courantes dans la littérature. Ensuite, nous étudierons les propriétés de ces hauteurs dans les cas non contraint (section 4.3) et contraint (section 4.4). Dans cette partie figurent certains résultats que j'ai démontrés.

4.1 Les dendrogrammes

On a besoin de la notion d'arbre pour définir les dendrogrammes. Un arbre binaire est un graphe connexe acyclique tel que le degré de chaque noeud est au plus 3. Il est dit enraciné si un des noeuds, de degré au plus 2, a été défini comme racine. Considérons la hiérarchie (au sens de la définition 1) induite par la CAH de n objets. On appelle dendrogramme associé à cette CAH un arbre binaire dont les noeuds sont les éléments de la hiérarchie, et dont les arêtes représentent les fusions successives. Un tel dendrogramme possède donc n feuilles, chacune associée à un des n objets à classer, et $n-1$ noeuds internes correspondant aux $n-1$ fusions de la classification. La racine de l'arbre est la partition triviale contenant tous les éléments à classer.

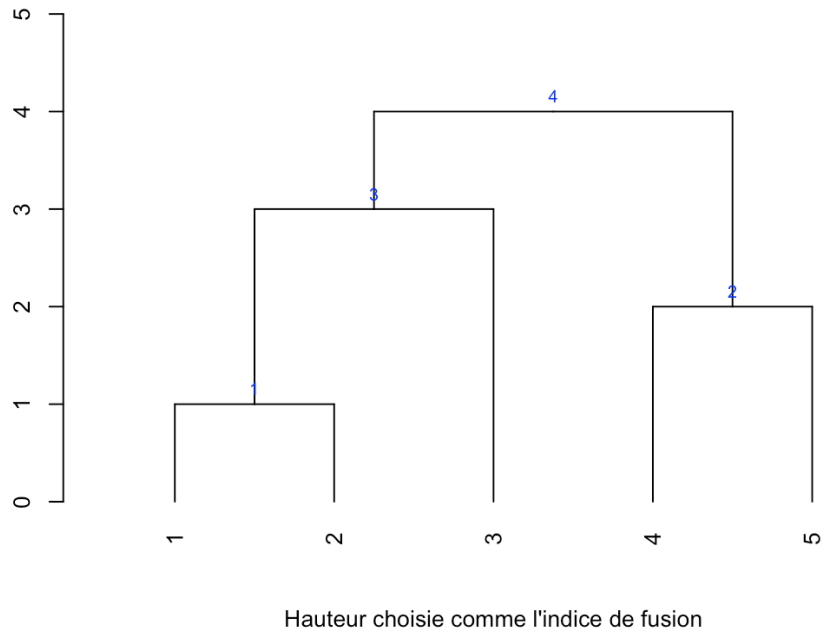

FIGURE 7 – Illustration du point de vue topologique

Dans la figure 7 ci-dessus, le placement des noeuds reflète l'ordre des fusions successives, et leur hauteur représente l'indice de fusion. Il s'agit là d'une représentation *topologique*. Cependant, la CAH fournit une information plus détaillée que le simple indice de fusion car on dispose d'informations quantitatives sur les fusions successives (comme la

valeur du lien de Ward entre classes fusionnées). Il paraît donc naturel de proposer une représentation *pondérée*, qui exploite ces informations pour définir la hauteur des fusions et rendre compte de la hiérarchie indexée que produit la CAH, comme dans la Figure 8.

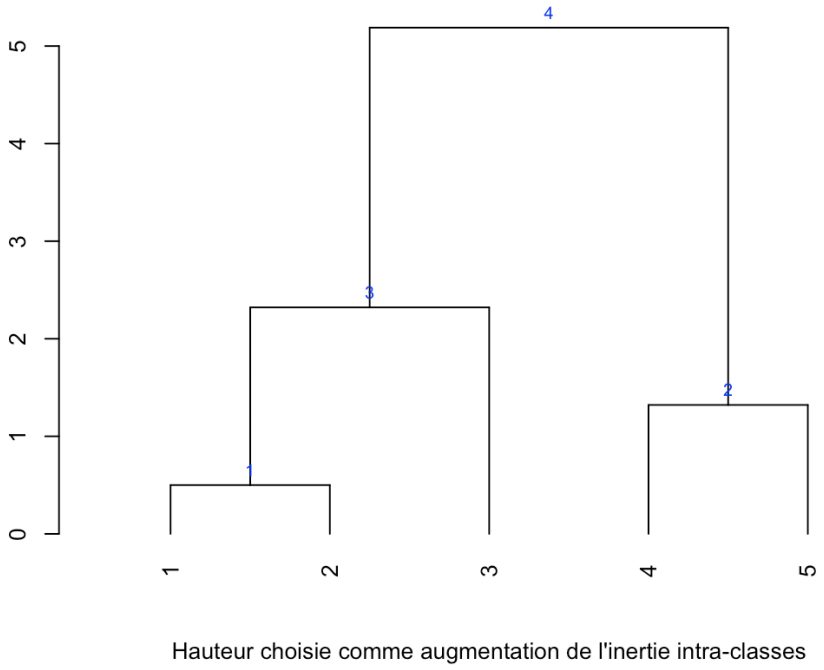

FIGURE 8 – Illustration du point de vue pondéré

Il existe différents choix de hauteur possibles, chacun d'entre eux menant à des propriétés spécifiques en fonction des données. Généralement, on souhaite que l'ordre induit par la hauteur coincide avec celui des fusions mais nous allons voir que cette propriété n'est pas toujours vérifiée, suivant le choix de la hauteur et le cadre considéré (CAH contrainte ou non).

4.2 Hauteurs d'un dendrogramme

[Grimm, 1987] propose quatre définitions de la notion de hauteur dans un dendrogramme. Dans la suite, $(G_u^t)_{u\in[\![1,n-t+1]\!]}$, désignera les classes de la partition $\mathcal{P}_t$ issue de l'étape $t-1$ de l'algorithme de classification ascendante hiérarchique.

— **Augmentation de l'inertie intra-classes** : il s'agit de l'approche standard. La hauteur est définie comme l'augmentation d'inertie intra-classes associée à la fusion des classes (valeur du lien de Ward). Supposons que G_i^t et G_j^t aient été fusionnés à l'étape t, on note $m_t = \delta(G_i^t \cup G_j^t) = \mathcal{I}(G_i^t \cup G_j^t) - \mathcal{I}(G_i^t) - \mathcal{I}(G_j^t)$ cette augmentation, appelée niveau de fusion à l'étape t.

— **Inertie intra-classes** : La hauteur est l'inertie intra-classes :

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

C'est la mesure suggérée à l'origine par [Ward, 1963]. Elle est également appelée *Error Sum of Squares* car dans le cas où les objets sont représentables dans un

espace euclidien, elle peut être décomposée en :

$$\mathrm{ESS}_t = \sum_{u=1}^{n-t} \sum_{x_i \in G_u^{t+1}} d^2(x_i, g_u^{t+1}),$$

où g_u^{t+1} est le centre de gravité de G_u^{t+1}.

— **Inertie** : La hauteur est l'inertie de la classe créée suite à la fusion l'inertie $\mathcal{I}(G_i^t \cup G_j^t)$. Cette mesure « locale » est très sensible à la taille des sous-ensembles impliqués et peut produire des dendrogrammes très aplatis.

— **Inertie moyenne** : La hauteur est l'*inertie moyenne*, notée $\overline{\mathcal{I}(G_i^t \cup G_j^t)}$, où $\overline{\mathcal{I}(A \cup B)} = \mathcal{I}(A \cup B)/(\mu_A + \mu_B)$. Cette définition permet d'éviter le biais lié de l'intertie à la taille des classes. Elle correspond à la variance de la classe lorsque les objets appartiennent à un espace euclidien.

La figure 9 illustre ces quatre possibilités de choix de hauteurs.

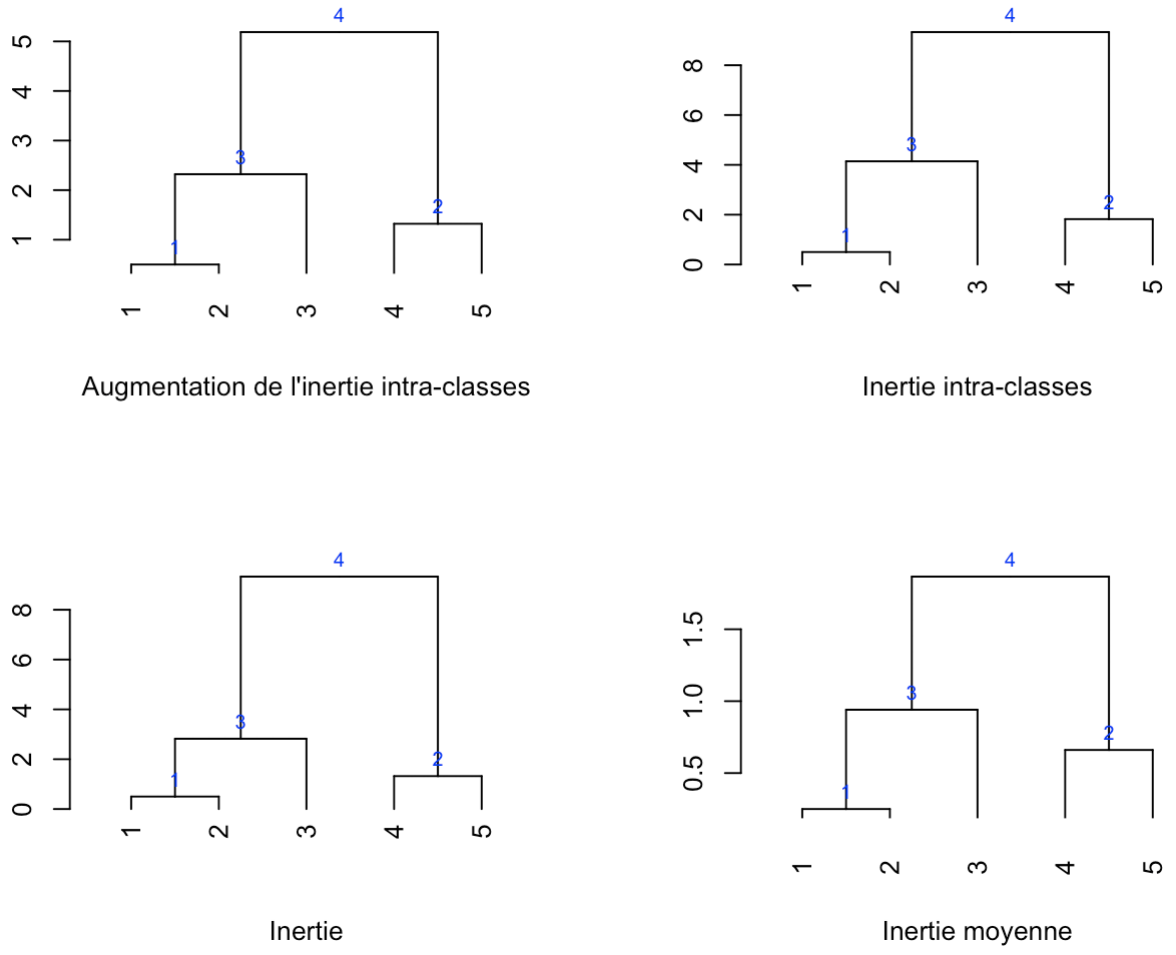

FIGURE 9 – Un illustration des différents choix de hauteurs pour une même CAH

Désormais on notera respectivement m_t, ESS_t, $\mathcal{I}_t$ et $\overline{\mathcal{I}}_t$ les hauteurs obtenues par l'approche standard, l'inertie intra-classes, l'inertie et l'inertie moyenne à l'étape t pour $t = 1, \ldots, n-1$ de la CAH.

4.3 Propriétés dans le cas non contraint

Proposition 2 (Croissance des niveaux de fusions)**.**

Soit $t \in [\![1, n-1]\!]$. On note A_t et B_t les classes fusionnées à l'étape t, permettant de passer de la partition $\mathcal{P}_t$ à $\mathcal{P}_{t+1}$, et m_t le niveau de fusion associé. Alors :

— $\forall\, G \in \mathcal{P}_t, \qquad \delta(G, A_t \cup B_t) \geq m_t$

— $0 \leq m_{t-1} \leq m_t$ *(avec la convention $m_0 = 0$)*

Démonstration. En utilisant la formule de Lance-Williams, on a que :

$$\delta(G, A_t \cup B_t) = \frac{\mu_{A_t} + \mu_G}{\mu_{A_t} + \mu_{B_t} + \mu_G}\delta(G, A_t) + \frac{\mu_{B_t} + \mu_G}{\mu_{A_t} + \mu_{B_t} + \mu_G}\delta(G, B_t) - \frac{\mu_G}{\mu_{A_t} + \mu_{B_t} + \mu_G}\delta(A_t, B_t)$$

De plus, par optimalité de $\delta(A_t, B_t)$, on a $\delta(A_t, B_t) \leq \delta(G, A_t)$ et $\delta(A_t, B_t) \leq \delta(G, B_t)$, donc

$$\delta(G, A_t \cup B_t) \geq \left(\frac{\mu_{A_t} + \mu_G}{\mu_{A_t} + \mu_{B_t} + \mu_G} + \frac{\mu_{B_t} + \mu_G}{\mu_{A_t} + \mu_{B_t} + \mu_G} - \frac{\mu_G}{\mu_{A_t} + \mu_{B_t} + \mu_G} \right) \delta(A_t, B_t) = \delta(A_t, B_t).$$

Le second point est équivalent à : $\forall\, G, G' \in \mathcal{P}_{t+1}\,, \delta(G, G') \geq \delta(A_t, B_t)$. Ceci est vrai pour $G, G' \neq A_t \cup B_t$, puisque alors $G, G' \in \mathcal{P}_t$ et, par optimalité de la fusion, $\delta(A_t, B_t) \leq \delta(G, G')$. Si $G' = A_t \cup B_t$ (ou, de façon équivalente, $G = A_t \cup B_t$), alors d'après le premier point, on a aussi que $\delta(G, G') = \delta(G, A_t \cup B_t) \geq m_t$. Ainsi, la suite des niveaux de fusions est croissante. Enfin, la positivité des niveaux de fusion à chaque étape est directement impliqué par le point précédent : la première fusion est définie par le minimum des quantités suivantes :

$$\delta(\{x_i\}, \{x_j\}) = \frac{1}{2} d^2(x_i, x_j)$$

pour $(i, j) = [\![1, n]\!]^2$ et $i \neq j$. Or toutes ces quantités sont positives et donc, $m_1 \geq 0$. □

Proposition 3. *Soient $(m_t)_{t=1,\dots,n-1}$ la suite des niveaux de fusions et $(ESS_t)_{t=1,\dots,n-1}$ la suite des inerties intra-classes, on a la relation suivante :*

$$ESS_t = \sum_{t' \leq t} m_{t'}$$

Démonstration. Cette propriété est démontrée par un argument récursif :

— La propriété est vraie pour $t = 1$ puisque, dans ce cas, toute classe est composée seulement d'un objet (et a donc une inertie égale à 0) exceptée pour une qui est composée de deux objets x_i et x_j. Sa dispersion vaut alors :

$$\begin{aligned} \mathrm{ESS}_1 = \mathcal{I}(\{x_i\} \cup \{x_j\}) &= \mathcal{I}(\{x_i\} \cup \{x_j\}) - \underbrace{\mathcal{I}(\{x_i\})}_{=0} - \underbrace{\mathcal{I}(\{x_j\})}_{=0} \\ &= \delta(\{x_i\}, \{x_j\}) = m_1; \end{aligned}$$

— Si, pour un certain $1 \leq t \leq n-2$, $\mathrm{ESS}_t = \sum_{t' \leq t} m_{t'}$ et $G_{u^*}^{t+2} = G_i^{t+1} \cup G_j^{t+1}$ sont les

deux classes fusionnées à l'étape $t+1$ alors on a que :

$$\begin{aligned}
\mathrm{ESS}_{t+1} &= \sum_{\substack{u=1\\u\neq u^*}}^{n-t-2} \mathcal{I}(G_u^{t+2}) + \mathcal{I}(G_i^{t+1} \cup G_j^{t+1}) \\
&= \sum_{\substack{u=1\\u\neq u^*}}^{n-t-2} \mathcal{I}(G_u^{t+2}) + \delta(G_i^{t+1}, G_j^{t+1}) + \mathcal{I}(G_i^{t+1}) + \mathcal{I}(G_j^{t+1}) \\
&= \sum_{v=1}^{n-t-1} \mathcal{I}(G_v^{t+1}) + \delta(G_i^{t+1}, G_j^{t+1}) \\
&= \mathrm{ESS}_t + m_{t+1} = \sum_{t'\leq t+1} m_{t'}.
\end{aligned}$$

□

Dans le cadre d'une classification ascendante hiérarchique standard, pour tout type de dissimilarité, on a donc :

1. croissance des hauteurs $(m_t)_{t=1,\dots,n-1}$ d'après la Proposition 2
2. croissance des hauteurs $(\mathrm{ESS}_t)_{t=1,\dots,n-1}$ en combinant la Proposition 3 avec le fait que $m_t \geq 0$ pour tout t.

Durant le stage, j'ai construit un contre exemple montrant que les hauteurs $(\mathcal{I}_t)_{t=1,\dots,n-1}$ et $(\overline{\mathcal{I}}_t)_{t=1,\dots,n-1}$ peuvent donner lieux à des inversions dans le dendrogramme, même dans le cas de la CAH standard pour une dissimilarité euclidienne :

Proposition 4. *Dans le cadre d'une classification ascendante hiérarchique standard,* $(\mathcal{I}_t)_{t=1,\dots,n-1}$ *et* $(\overline{\mathcal{I}}_t)_{t=1,\dots,n-1}$ *ne sont pas nécessairement croissantes.*

Démonstration.

Prouvons cela pour une dissimilarité euclidienne à l'aide d'un exemple, On considère les 5 points suivants dans $\mathbb{R}^2$: $x_1 = (1/2, \sqrt{3}/2)$, $x_2 = (-1/2, \sqrt{3}/2)$, $x_3 = (0, -1)$, $x_4 = (10, 0)$ et $x_5 = (10, d)$. En utilisant ces points, il est simple de construire un exemple où $\mathcal{I}_t$ est décroissante selon l'ordre des fusions. En effet, en choisissant dans le bon intervalle le paramètre d, on peut avoir la topologie suivante $\{\{\{x_1, x_2\}, x_3\}, \{x_4, x_5\}\}$ pour la classification, avec la fusion $\{x_4, x_5\}$ se produisant après la formation de la classe $\{x_1, x_2, x_3\}$ tandis que l'inertie de $\{x_4, x_5\}$ est plus petite que celle de $\{x_1, x_2, x_3\}$.
Plus précisément,

- $\mathcal{I}(\{x_1, x_2, x_3\}) = \frac{1}{3}(d_{12}^2 + d_{23}^2 + d_{31}^2)$
- $\mathcal{I}(\{x_4, x_5\}) = \frac{1}{2}d_{45}^2 = \frac{1}{2}d^2$
- $\delta(\{x_1, x_2\}, \{x_3\}) = \mathcal{I}(\{x_1, x_2, x_3\}) - \mathcal{I}(\{x_1, x_2\}) = \frac{1}{3}(d_{12}^2 + d_{23}^2 + d_{31}^2) - \frac{1}{2}d_{12}^2$
- $\delta(\{x_4\}, \{x_5\}) = \mathcal{I}(\{x_4, x_5\}) = \frac{1}{2}d^2$

Donc, pour avoir la configuration voulue, il ne reste plus qu'à choisir d tel que :

$$\sqrt{\frac{2}{3}(d_{12}^2 + d_{23}^2 + d_{31}^2 - d_{12}^2)} < d < \sqrt{\frac{2}{3}(d_{12}^2 + d_{23}^2 + d_{31}^2)}$$

qui est un intervalle plus large que

$$2.16 \leq d \leq 2.37$$

De plus, l'exemple précédent peut être réemployé sans grandes modifications pour l'inertie moyenne $\overline{\mathcal{I}_t}$: Dans ce cas, l'inégalité devient :

$$\sqrt{\frac{2}{3}(d_{12}^2 + d_{23}^2 + d_{31}^2 - d_{12}^2)} < d < \sqrt{\frac{4}{9}(d_{12}^2 + d_{23}^2 + d_{31}^2)}$$

Donc, on peut choisir x_1,x_2 et x_3 centrés sur l'origine tel que $d_{23} = d_{13} = R$ et $d_{12} \underset{d12<R}{\longrightarrow} R$ de sorte que la topologie soit préservée. R est arbitraire mais doit être choisi suffisamment petit de façon à ce que le classe $\{x_1, x_2, x_3\}$ apparaisse dans la CAH avant $\{x_4, x_5\}$. Dans cette situation, l'inégalité pour d devient :

$$\sqrt{R} < d < \sqrt{\frac{4}{3}R}$$

qui a bien des solutions. (Par exemple, avec $R = 0.1$, on obtient $0.1 \leq d \leq 0.115$) □

4.4 Propriétés dans le cas contraint

Les résultats négatifs de non-croissance des hauteurs $(\mathcal{I}_t)_t$ et $(\bar{\mathcal{I}}_t)_t$ pour la CAH standard sont directement généralisables à la CAHCC, puisque un contre-exemple pour la CAH peut être considéré comme un contre-exemple pour la CAHCC avec des contraintes de contiguïté définies a posteriori comme l'ordre de fusion dans le cas de la CAH standard. Par conséquent, nous nous intéresserons seulement aux hauteurs m_t et ESS_t pour la CAHCC. [Ferligoj and Batagelj, 1982] identifient des conditions nécessaires et suffisantes sur les coefficients de Lance-Williams pour que la suite des hauteurs de la CAH sous contrainte soit croissante. Ici, nous cherchons à identifier des conditions portant directement sur la matrice de dissimilarité d.

4.4.1 Niveaux de fusion

En se concentrant sur les deux premières fusions, ($t = 1$ et $t = 2$), on a la proposition suivante (dont la preuve est donnée en Section 6).

Lemme. *Supposons que* $i^* = \text{argmin}_{i=1,\dots,n-1}\, d(x_i, x_{i+1})$*, alors* $m_2 < m_1$ *(la croissance des niveaux de fusion est invalidée dès la seconde étape de la classification) si et seulement si les deux conditions suivantes sont satisfaites :*

1. $\text{argmin}_{i=1,\dots,n-2}\, \delta(G_i^2, G_{i+1}^2) \in \{i^* - 1, i^*\}$ *;*
2. *en supposant sans perte de généralité que le minimum est atteint pour* i^**,*

$$d^2(x_{i^*+1}, x_{i^*+2}) - d^2(x_{i^*}, x_{i^*+1}) < d^2(x_{i^*}, x_{i^*+1}) - d^2(x_{i^*}, x_{i^*+2})$$

Grâce au Lemme 4.4.1 on peut prouver que les deux premiers niveaux de fusion peuvent être décroissants même pour une configuration euclidienne très simple, par exemple avec les distances suivantes (qui vérifient l'inégalité triangulaire) pour le triplet $\{x_1, x_2, x_3\}$:

$$d^2(x_1, x_3) = 0.5, \qquad d^2(x_1, x_2) = 2 \text{ and} \qquad d^2(x_2, x_3) = 2.05.$$

Alors la Figure 10 illustre le résultat en termes de dendrogramme (les indices des fusions y sont représentés en bleu) :

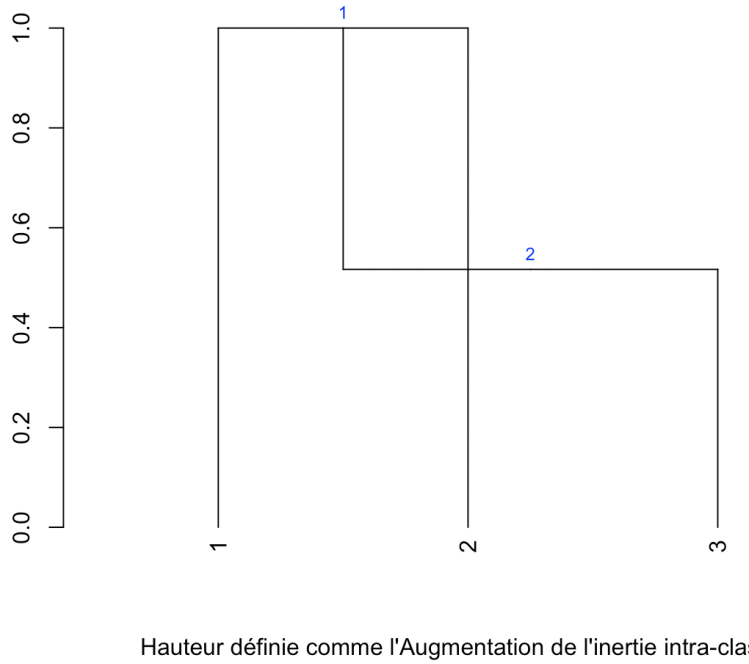

FIGURE 10 – Dendogramme obtenu à partir de la dissimilarité

On a donc le résultat suivant :

Proposition 5. *Dans le cadre d'une classification ascendante hiérarchique avec contrainte de contiguïté, les niveaux de fusion* $(m_t)_{t=1,\dots,n-1}$ *ne sont pas nécessairement croissants.*

Remarque. *La contrainte (ii) du Lemme 4.4.1 est directement généralisable à n'importe quel niveau de la hiérarchie avec la condition suivante :*

$$\left(1 - \frac{\mu^t_{i^*}}{\mu^t_{i^*} + \mu^t_{i^*+1} + \mu^t_{i^*+2}}\right)\left[\delta(G^t_{i^*+1}, G^t_{i^*+2}) - \delta(G^t_{i^*}, G^t_{i^*+1})\right] < \left(1 - \frac{\mu^t_{i^*+1}}{\mu^t_{i^*} + \mu^t_{i^*+1} + \mu^t_{i^*+2}}\right)\left[\delta(G^t_{i^*}, G^t_{i^*+1}) - \delta(G^t_{i^*+1}, G^t_{i^*+2})\right]$$

avec μ^t_i *le cardinal de* G^t_i, $G^t_{i^*}$ *et* $G^t_{i^*+1}$ *les deux classes fusionnées pour donner l'étape* t *de la hiérarchie et* $G^t_{i^*} \cup G^t_{i^*+1}$ *et* $G^t_{i^*+2}$ *les deux classes fusionnées pour donner l'étape* $t+1$ *de la hiérarchie.*

Quand d n'est pas euclidienne, en suivant l'argument donné par [Olteanu and Villa-Vialaneix, 2015], on peut prouver que :

$$(\mu_A + \mu_B)\delta(A,B) = -\frac{\mu_A \mu_B}{2}\left(\mathbf{1}_A - \mathbf{1}_B\right)^\top \mathbf{D}\left(\mathbf{1}_A - \mathbf{1}_B\right),$$

avec $\mathbf{1}_C$ un vecteur de taille n dont l'entrée z_i est égal à $\frac{1}{\mu_C}$ si $x_i \in C$ et 0 sinon. Cette quantité n'est pas nécessairement positive. J'ai même pu montrer le résultat suivant à l'aide d'un contre-exemple :

Lemme. *Dans le cas de la CAHCC avec une distance non euclidienne, le lien de Ward peut être négatif.*

Démonstration. on suppose que l'on a 4 objets à classer avec des contraintes d'adjacence avec :

$$d^2(x_1, x_2) = 1 \qquad d^2(x_1, x_3) = d^2(x_1, x_4) = d^2(x_2, x_4) = 0.1 \qquad d^2(x_2, x_3) = 2 \qquad d^2(x_3, x_4) = 1.1.$$

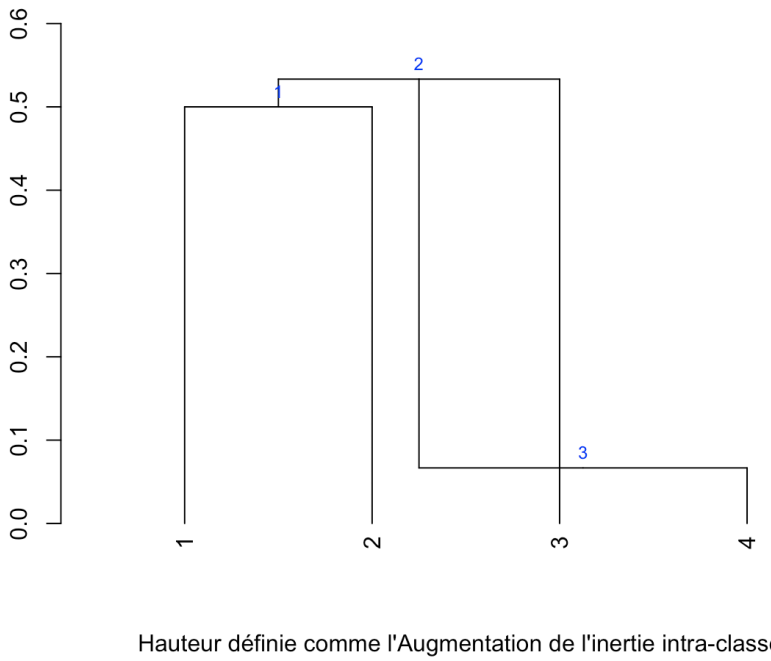

FIGURE 11 – Dendogramme obtenu à partir de la dissimilarité

Comme illustré sur la Figure 11 (où les indices des fusions sont représentés en bleu), la première étape de la méthode fusionne $\{x_1\}$ avec $\{x_2\}$:

$$A_1 = \{x_1, x_2\} \text{ avec } m_1 = 0.5 \qquad \delta(A_1, \{x_3\}) = \frac{1.6}{3} \simeq 0.53 \qquad \delta(A_1, \{x_4\}) = -0.1 < 0.$$

Cependant, la seconde étape de la méthode fusionne A_1 avec $\{x_3\}$ et on a :

$$A_2 = A_1 \cup \{x_3\} \text{ avec } m_2 = 0.53 \qquad \delta(A_2, \{x_4\}) = \frac{0.2}{3} \simeq 0.067 > 0$$

qui prouve que la dernière fusion est également positive.

□

4.4.2 Inertie intra-classes

J'ai pu prouver pendant le stage, à l'aide d'un contre exemple, que les hauteurs définies par l'inertie intra-classes ESS_t n'étaient pas non plus nécessairement croissantes dans le cas de la CAHCC.

Proposition 6. *Dans le cadre d'une classification ascendante hiérarchique avec contrainte de contiguïté, les inerties intra-classes $(ESS_t)_{t=1,\dots,n-1}$ ne sont pas nécessairement croissantes.*

Démonstration. On prouve la non croissance de ESS_t (ainsi que celles de $\mathcal{I}_t$ et $\bar{\mathcal{I}}_t$) pour une dissimilarité non euclidienne à l'aide du contre-exemple suivant. Soit la dissimilarité $D = (d_{ij})$ définie par :

$$\begin{pmatrix} 0 & \sqrt{2-\epsilon} & \sqrt{2-\epsilon} & \sqrt{2-\epsilon} & \sqrt{\epsilon} & 1 \\ \sqrt{2-\epsilon} & 0 & \sqrt{2} & \sqrt{2-\epsilon} & \sqrt{\epsilon} & 1 \\ \sqrt{2-\epsilon} & \sqrt{2} & 0 & \sqrt{2} & \sqrt{\epsilon} & 1 \\ \sqrt{2-\epsilon} & \sqrt{2-\epsilon} & \sqrt{2} & 0 & \sqrt{2} & 1 \\ \sqrt{\epsilon} & \sqrt{\epsilon} & \sqrt{\epsilon} & \sqrt{2} & 0 & \sqrt{2} \\ 1 & 1 & 1 & 1 & \sqrt{2} & 0 \end{pmatrix}$$

Dans ce cas, la classification avec lien de Ward, fusionne les éléments de la droite vers la gauche. Ci-après, on présente les dendrogrammes obtenu pour cette dissimilarité avec les ordres de fusion en bleu.

— Avec pour choix de hauteur ESS_t ou $\mathcal{I}_t$ (ces deux quantités sont égales dans ce cas particulier) dans la figure 12 :

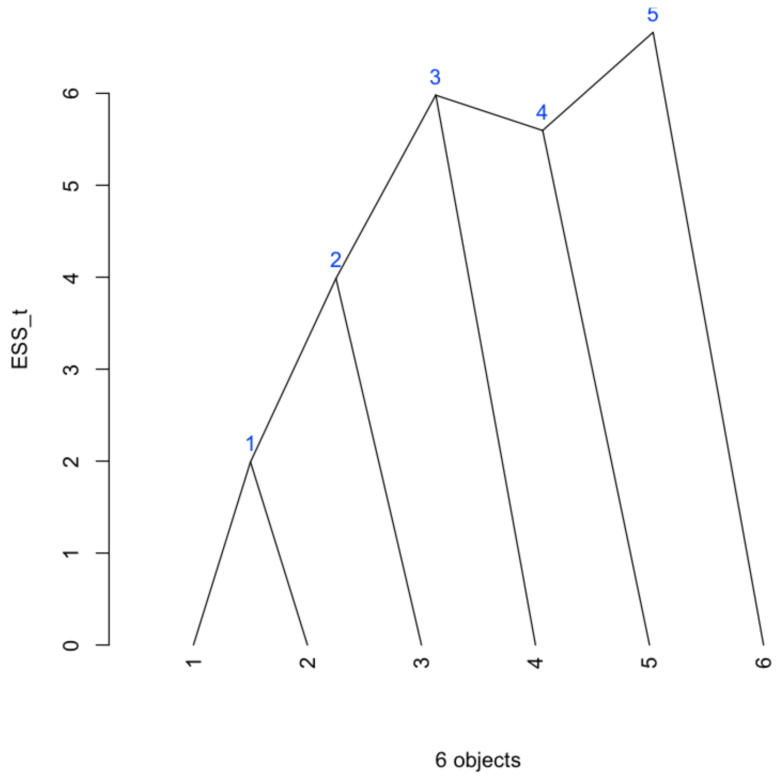

FIGURE 12 – Dendrogramme obtenu à l'aide de D avec pour choix de hauteur ESS_t

— Avec pour choix de hauteur $\bar{\mathcal{I}}_t$ dans la figure 13 :

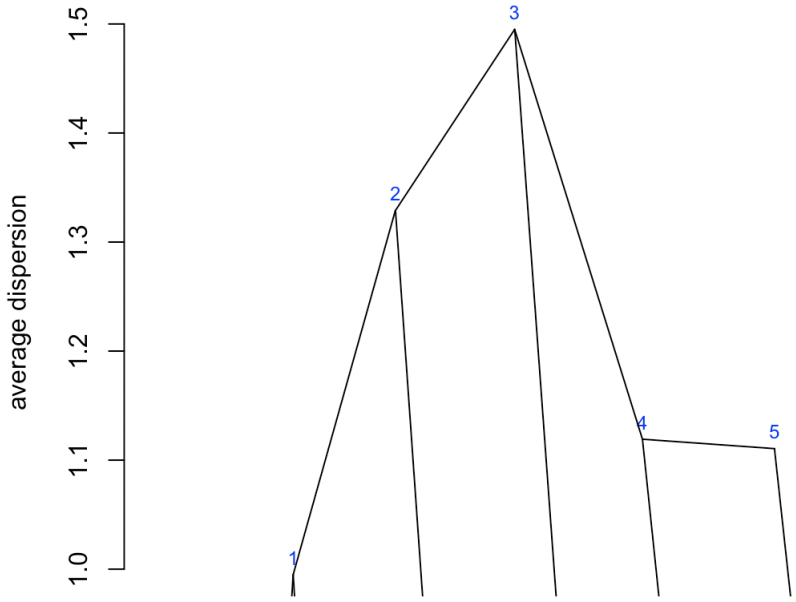

FIGURE 13 – Dendrogramme obtenu à l'aide de D avec pour choix de hauteur $\bar{\mathcal{I}}_t$

□

Le contre-exemple ci-dessus a été construit avec $n = 5$ objets à classer. On peut montrer (voir preuve en Section 6) que les niveaux ESS_t sont nécessairement croissants dans le cas de la CAHCC avec seulement $n = 4$ objets à classer.

Remarque. *Une de mes participations en termes d'implémentation a été l'ajout en bleu de l'ordre des fusions sur les noeuds du dendrogramme. Pour cela, j'ai du comprendre la structure de donnée que constitue le dendrogramme sous R et également les fonctions de bases associées à la représentation de dendrogramme. En modifiant certaines lignes du code, l'affichage des ordres de fusion a été implémenté de sorte à pouvoir être demandé en argument de la fonction* `plot`.

4.4.3 Synthèse des résultats obtenus

On va donc résumer ce que l'on sait à propos des différentes hauteurs de dendrogrammes possibles pour la classification ascendante hiérarchique sous contrainte de contiguïté dans le tableau suivant :

CAHCC	m_t		ESS_t		I_t		$\bar{I}_t$	
Dissimilarité	positivité	croissance	positivité	croissance	positivité	croissance	positivité	croissance
Euclidienne	oui	non	oui	oui	oui	non	oui	non
Non Euclidienne	non	non	oui	non	oui	non	oui	non

Les champs en rouge correspondant à des contre-exemples que j'ai pu mettre en évidence au cours du stage.

5 Comparaison et stabilité de dendrogrammes

Pouvoir étudier des conditions biologiques différentes au moyen de la classification ascendante hiérarchique nécessite d'avoir des méthodes de comparaison des résultats.

Cette partie consiste en une étude bibliographique d'un certain nombre d'articles autour de la thématique de la comparaison de classifications hiérarchiques. On y présente des articles qui ont été perçus comme utiles dans le cadre de notre problématique, presque tous provenant du champ de la phylogénétique. En effet, les dendrogrammes y sont très utilisés pour représenter les filiations entre espèces au cours du temps.

La section 5.1 a pour but de mettre en évidence des outils pour identifier des différences (éventuellement locales) entre résultats de CAH. On y définit des scores et des distances afin de comparer les classifications hiérarchiques. De plus, interpréter ces différences nécessite également de pouvoir tester leur significativité. Cette question est abordée dans la section 5.2 au travers de l'étude de la robustesse d'une CAH.

5.1 Ressemblances et différences entre dendrogrammes

Parmi les critères quantitatifs permettant de comparer deux dendrogrammes, on distingue ci-dessous ceux reposant sur la comparaison des partititions issues des dendrogrammes de ceux exploitant réellement le caractère hiérarchique des dendrogrammes. Les critères de la sectioon 5.1.1 ne prennent en compte que la topologie des classifications et revêtent un aspect local car ils sont définis au niveau des partitions. Ceux de la section 5.1.2 en revanche ont un point de vue global car ils comparent les dendrogrammes. dans leur ensemble.

5.1.1 Critères de ressemblance entre partitions

On présente ici deux scores de ressemblance entre partitions. Ils peuvent être perçus comme scores de ressemblance pour CAH en considérant la suite des scores pour chacune des partitions des hiérarchies. Cependant, par leur nature, ces approches ne peuvent prendre en compte que la topologie des classifications concernées.

Ici, on a besoin seulement de l'ordre de fusion des classes et donc, il n'y a pas de considération concernant les hauteurs. Etant donnés deux résultats de classification ascendante hiérarchique à comparer pour lesquels on dispose de deux dendrogrammes, on peut décider de couper ces arbres de sorte à définir des partitions à k classes, $\mathcal{P}^k$ et $\mathcal{Q}^k$, pour k variant de 1 à n. Pour une valeur de k donnée et chacune des deux partitions $\mathcal{P}^k$ et $\mathcal{Q}^k$ issues des coupes, on numérote les classes arbitrairement de 1 à k. On peut alors définir la matrice $M^k = (m^k_{ij})_{1\leq i,j\leq k}$ où m^k_{ij} est le nombre d'éléments en commun entre la i-ème classe de $\mathcal{P}^k$ et la j-ème classe de $\mathcal{Q}^k$ (dans la suite on omettra l'indice supérieur k pour des raisons de clarté).

Indice de Fowlkes and Mallows [Fowlkes and Mallows, 1983] définissent le critère de ressemblance suivant :

$$B_k = \frac{T_k}{\sqrt{P_k Q_k}}$$

$$\text{où } T_k = \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij}^2 - n \qquad m_{i\cdot} = \sum_{j=1}^{k} m_{ij} \qquad m_{\cdot j} = \sum_{i=1}^{k} m_{ij} \qquad m_{\cdot\cdot} = n = \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij}$$

$$P_k = \sum_{i=1}^{k} m_{i\cdot}^2 - n \qquad Q_k = \textstyle\sum_{j=1}^{k} m_{\cdot j}^2 - n$$

B_k est donc calculé pour $k \in [\![2, n-1]\!]$, et un graphe de comparaison des deux classifications est obtenu en représentant B_k en fonction de k par exemple. On a $0 \leq B_k \leq 1$ pour tout k. $B_k = 1$ lorsque M a exactement k entrée non vide, ce qui se produit lorsque les k classes de chaque classification correspondent deux à deux. De plus, $B_k = 0$ lorsque tout m_{ij} vaut 0 ou 1, de sorte que toute paire d'objets qui apparaissent dans la même classe pour la première classification sont dans des classes distinctes dans la seconde.

On peut étudier certaines propriétés de ce critère :

1. $T_k \geq 0, \quad P_k \geq 0 \quad$ et $\quad Q_k \geq 0$:
2. $B_k \leq 1$:
3. Cas $B_k = 0$:
   $B_k = 0 \Leftrightarrow T_k = 0 \Leftrightarrow m_{ij} = 1$ ou $0, \forall i, j \in [\![1, k]\!]$
4. Cas $B_k = 1$:
   $B_k = 1 \Leftrightarrow \sum_{i=1}^{k} \sum_{j_1<j_2} m_{ij_1} m_{ij_2} = 0$ et $\sum_{j=1}^{k} \sum_{i_1<i_2} m_{i_1 j} m_{i_2 j} = 0$

L'aspect hiérarchique de la classification n'est ici pas directement pris en compte, c'est seulement à travers la suite des B_k qu'on obtient une information globale.

L'indice de Rand En 1971, [Rand, 1971] introduit une mesure de ressemblance pour comparer deux partitions d'un même ensemble. On peut utiliser ce critère pour étudier les partitions successives de deux classifications ascendantes hiérarchiques. Il est basé sur trois principes :

— une classification est discrète en ce sens qu'elle assigne un objet à une classe ;
— une classe est définie autant par les points qu'elle contient que par ceux qu'elle ne contient pas ;
— tous les points sont d'importance égale dans la détermination de la partition.

Ainsi, étant donnée deux partitions $\mathcal{P}$ et $\mathcal{Q}$ d'un même ensemble, il définit la quantité $c(\mathcal{P}, \mathcal{Q})$ comme étant la fraction de paires d'éléments rangés de la même façon par les deux partitions. Plus précisément, en notant $\Omega = \{x_1, \cdots, x_n\}$ l'ensemble à classer, $\mathcal{P} = \{\mathcal{P}_1, \cdots, \mathcal{P}_{K_1}\}$ et $\mathcal{Q} = \{\mathcal{Q}_1, \cdots, \mathcal{Q}_{K_2}\}$:

$$c(\mathcal{P}, \mathcal{Q}) = \frac{1}{\binom{n}{2}} \sum_{i<j}^{n} \gamma_{ij}$$

$$\text{avec } \gamma_{ij} = \begin{cases} 1 \text{ s'il existe } k \text{ et } k' \text{ tels que } x_i \text{ et } x_j \text{ sont tous deux dans } \mathcal{P}_k \text{ et } \mathcal{Q}_{k'} \\ 1 \text{ s'il existe } k \text{ et } k' \text{ tels que } x_i \in \mathcal{P}_k \text{ mais pas } x_j \text{ et } x_i \in \mathcal{Q}_{k'} \text{ mais pas } x_j \\ 0 \text{ autrement} \end{cases}$$

On peut également exprimer c à partir de la matrice $M = (m_{ij})_{1 \leq i,j \leq k}$ définie ci-dessus, où m_{ij} est le nombre d'éléments en commun entre la i-ème classe de $\mathcal{P}$ et la j-ème classe de $\mathcal{Q}$:

$$c(\mathcal{P}, \mathcal{Q}) = \frac{1}{\binom{n}{2}} \left[\binom{n}{2} - \left(\frac{1}{2} \left\{ \sum_i \left(\sum_j m_{ij} \right)^2 + \sum_j \left(\sum_i m_{ij} \right)^2 \right\} - \sum_i \sum_j m_{ij}^2 \right) \right]$$

Cette matrice M revient de façon récurrente dans la littérature comme outil de comparaison de partition. Le critère c possède certaines propriétés. Tout d'abord, il varie de 0

(cas où une des partitions consiste en une seule classe et où l'autre est la partition en singleton par exemple) à 1 (cas où les deux partitions sont identiques). Ensuite, $1 - c$ est une distance sur l'ensemble des partitions de Ω.

Ces particularités en font un outil utile pour diverses tâches comme : l'évaluation de la qualité des méthodes de classifications, la comparaison de méthodes de classifications, l'analyse de sensibilité suite à une perturbation des données, etc.

Tout comme le score précédent, il convient de considérer la suite des scores pour chaque niveau de la hiérarchie afin d'avoir un critère global de comparaison de classification.

5.1.2 Distances entre dendrogrammes

Une autre idée peut consister à utiliser des distances sur l'espace des arbres binaires enracinés à n feuilles. En réalité, beaucoup des distances proposées pour mesurer la dissemblance entre arbres sont basées sur différentes façons de les représenter.

Ainsi, nous exposons ci-dessous deux définitions de distance entre dendrogrammes exploitant leur caractère hiérarchique, et donc ayant un aspect global. La première repose uniquement sur la topologie du dendrogramme, alors que la seconde exploite également la notion de hauteur (point de vue pondéré).

Distances sur l'espace des arbres binaires enracinés Dans cette optique, [Diaconis and Holmes, 1998] exploitent une bijection entre l'ensemble des arbres binaires enracinés à $n + 1$ feuilles et l'ensemble des couplages parfaits. Un couplage parfait à $2n$ points est un appariement en n couples des points (en notant que l'ordre à l'intérieur d'un couple ou entre couples ne compte pas). La procédure bijective est la suivante :

1. repérer toutes les paires déjà indicées ;
2. choisir la paire comportant le plus petit indice ;
3. indexer le noeud parent de cette paire avec le prochain indice disponible ;
4. répéter jusqu'à ce que tous les noeuds (excepté la racine) aient un indice.

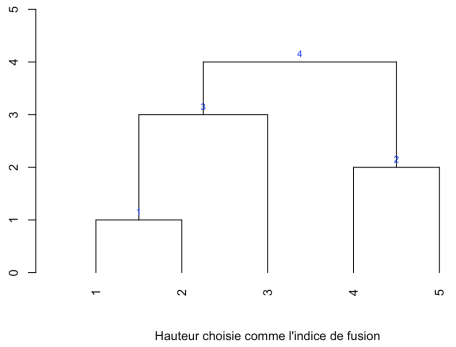

FIGURE 14 – Un exemple de dendrogramme simple

Ainsi, le dendrogramme de la figure 14 est associé au couplage parfait

$$(1, 2)(4, 5)(6, 3)(7, 8)$$

On obtient ensuite une distance en considérant la distance sur l'espace des couplages parfaits définie par le nombre minimal de transpositions nécessaires pour passer d'un

couplage à l'autre. La distance ainsi définie ne prend en compte que la topologie de l'arbre et néglige son éventuel aspect pondéré.

De plus, cette bijection permet une représentation géométrique où les arbres sont modélisés par les sommets d'un polytope convexe.

Coefficient de corrélation cophénétique [Sokal and Rohlf, 1962] définissent une méthode objective pour pourvoir comparer deux dendrogrammes, qui jusqu'alors étaient plutôt soumis au contrôle d'experts (notamment dans le domaine de la taxonomie dont est issu l'article).

Pour cela, ils introduisent le *coefficient de corrélation cophénétique*. Pour le définir, on a besoin de la notion de *distance cophénétique*, notée c_{ij}, entre deux objets x_i et x_j. À partir du dendrogramme pondéré tel que fournie par la méthode de classification ascendante hiérarchique, les auteurs divisent l'axe vertical en L intervalles $\{I_1, \cdots, I_L\}$ de même longueur, l'indice $l \in [\![1, L]\!]$ étant croissant par rapport à l'ordre des fusions. Si les deux éléments x_i et x_j apparaissent pour la première fois dans la même classe à la hauteur $h_{ij} \in I_{l^*}$, alors leur distance cophénétique est $c(i,j) = l^*$. On pose alors les notations suivantes :

- $d(i,j)$ la distance entre x_i et x_j comme donnée dans la matrice de dissimilarité ;
- $c(i,j)$ la distance cophénétique entre x_i et x_j ;
- $\bar{d}$ la moyenne des $d(i,j)$;
- $\bar{c}$ la moyenne des $c(i,j)$.

Alors, le coefficient de corrélation cophénétique $\mathcal{C}$ est défini par :

$$\mathcal{C} := \frac{\sum_{i<j}(d(i,j) - \bar{d})(c(i,j) - \bar{c})}{\sqrt{\left(\sum_{i<j}(d(i,j) - \bar{d})^2\right)\left(\sum_{i<j}(c(i,j) - \bar{c})^2\right)}}$$

qui est donc le coefficient de corrélation entre la distance cophénétique induite par le dendrogramme et la distance originale.

Avec cette définition, le coefficient de corrélation cophénétique compare en réalité les distances telles que fournies par la dissimilarité et les distances induites par la classification en termes de hauteurs. Il s'agirait donc plutôt de vérifier l'adéquation du résultat de la classification avec les données.

Pour avoir réellement un outil de comparaison entre dendrogrammes, il faut remplacer la dissimilarité originale dans la définition par la distance cophénétique induite par un second dendrogramme. On obtient ainsi une distance prenant en compte l'ensemble des informations issues de la classification ascendante hiérarchique. On peut néanmoins noter que l'aspect pondéré est discrétisé par la transposition en distance cophénétique.

Ce critère permet donc de comparer les représentations graphiques de CAH à travers les distances cophénétiques induites par la classification à l'aide d'un seul coefficient.

5.2 Fiabilité d'un dendrogramme

Dans la section précédente, on a vu un certain nombre d'outils pour quantifier des dissemblances entre résultats de classifications ascendantes hiérarchiques. Néanmoins, dans le cadre de notre problématique, il est légitime de poser la question de la significativité

d'une différence entre dendrogrammes. Cette interrogation est liée à la question de robustesse de la CAH. L'approche majoritaire que j'ai pu rencontrer pour répondre à ce problème est le ré-échantillonage bootstrap en vue de la réalisation de tests statistiques.

5.2.1 Généralités au sujet de l'approche par échantillons bootstrap

Le *bootstrap* a été introduit par [Efron, 1982, Efron and Tibshirani, 1986]. Son but est d'approcher de façon empirique la distribution d'un estimateur lorsqu'on ne connait pas la loi de l'échantillon. En pratique, cette technique permet de remplacer des hypothèses probabilistes difficilement vérifiables par des simulations.

Etant donné un échantillon d'apprentissage $\{x_i\}_{i\in[\![1,n]\!]}$ de taille n, le principe de ce ré-échantillonage est de substituer à la distribution de probabilité inconnue F, dont est issue l'échantillon d'apprentissage, la distribution empirique $\hat{F}$ qui donne un poids $\frac{1}{n}$ à chacune des n observations. Ainsi, on obtient un nouvel échantillon de taille n, dit échantillon bootstrap, selon la distribution empirique $\hat{F}$ par n tirages aléatoires avec remise parmi les n observations initiales.

L'idée est que, si la taille de l'échantillon original, n, est suffisamment grande, alors toutes les valeurs possibles x_i seront représentées dans la même proportion dans l'échantillon que dans la distribution inconnue F. Ainsi, un ré-échantillonage avec remise à partir des données initiales donnera un résultat similaire à un échantillon tiré selon la distribution inconnue F.

Etant donnés les moyens computationnels actuels, il est facile d'obtenir un grand nombre d'échantillons bootstrap à partir desquels on peut calculer l'estimateur concerné. La loi simulée de cet estimateur est une approximation asymptotiquement convergente sous des hypothèses raisonnables de la loi de l'estimateur. Ainsi, l'approche bootstrap fournit des estimations du biais, de la variance et même des intervalles de confiance de l'estimateur sans hypothèse sur la vraie loi.

5.2.2 Bootstrap Probability Test

L'article de [Felsenstein, 1985] s'inscrit dans le contexte de la phylogénie. Il a pour but d'établir des ensembles de confiance concernant l'optimalité de l'arbre obtenu en fin de procédure.

En phylogénie, les données se présentent le plus souvent sous la forme d'un tableau individus $\times$ caractères. Autrement dit, chaque individu est représenté par un vecteur de $\mathbb{R}^d$. L'approche bootstrap mise en oeuvre consiste à ré-échantilloner par rapport à d, c'est-à-dire, à ré-échantilloner les caractères. Ainsi, on génère r échantillons bootstrap à partir des données initiales donnant lieu à r arbres.

La première étape en vue d'obtenir des intervalles de confiance est alors de déterminer si une partie de l'arbre (c'est-à-dire un nœud donné et ses sous-branches) est significative du point de vue des données. On va choisir un seuil et considérer une partie comme significative si sa fréquence d'apparition dans les échantillons bootstrap est supérieure à ce seuil (typiquement 95%). Ainsi, on obtient un ensemble de parties d'arbre significatives du point de vue des données. Etant donné la fréquence d'apparition de ces parties, elles doivent nécessairement apparaître simultanément dans certains arbres et sont donc non contradictoires (soit imbriquées, soit disjointes). On peut donc construire un arbre unique à partir de ces parties.

On illustre la procédure avec les figures suivantes. La procédure choisie par l'auteur amène à plusieurs arbres potentiellement optimaux parmi lesquels les deux de la Figure 15.

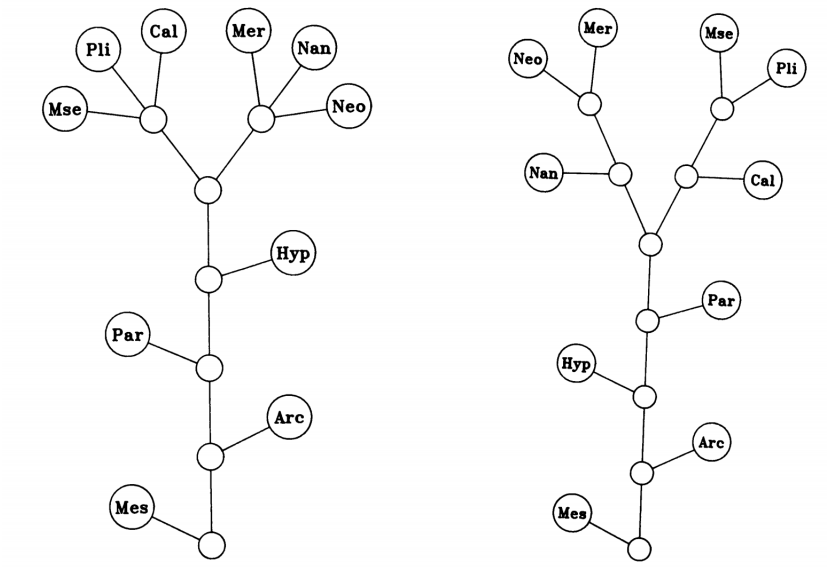

FIGURE 15 – Deux dendrogrammes possibles pour la phylogénie du cheval

Source: [Felsenstein, 1985]

Ainsi, une approche bootstrap à 50 échantillons conduit à l'arbre significatif de la figure 16 ci-dessous (sur les noeuds, on a porté le nombre d'occurence des parties associées).

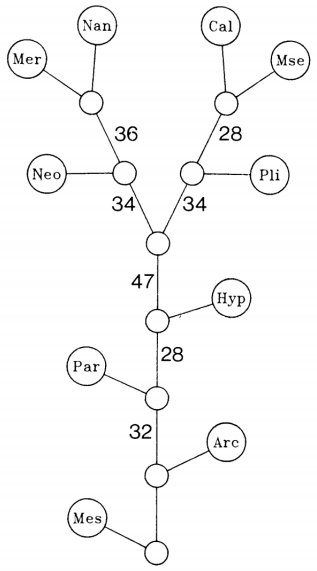

FIGURE 16 – Exemple de dendrogramme inféré à l'aide de l'approche par bootstrap

Source: [Felsenstein, 1985]

5.2.3 Approximately Unbiased Test

[Shimodaira, 2002] propose une version modifiée de l'idée précédente. Au lieu de réaliser des échantillons bootstrap de la même taille que l'échantillon initial, il suggère d'utiliser des échantillons bootstrap de taille variable. Il illustre la comparaison à l'aide d'un exemple jouet. Supposons que sur la figure 17, l'étoile rouge désigne l'arbre obtenu à l'aide de l'échantillon initial et l'étoile bleue la moyenne à estimer de la distribution inconnue.

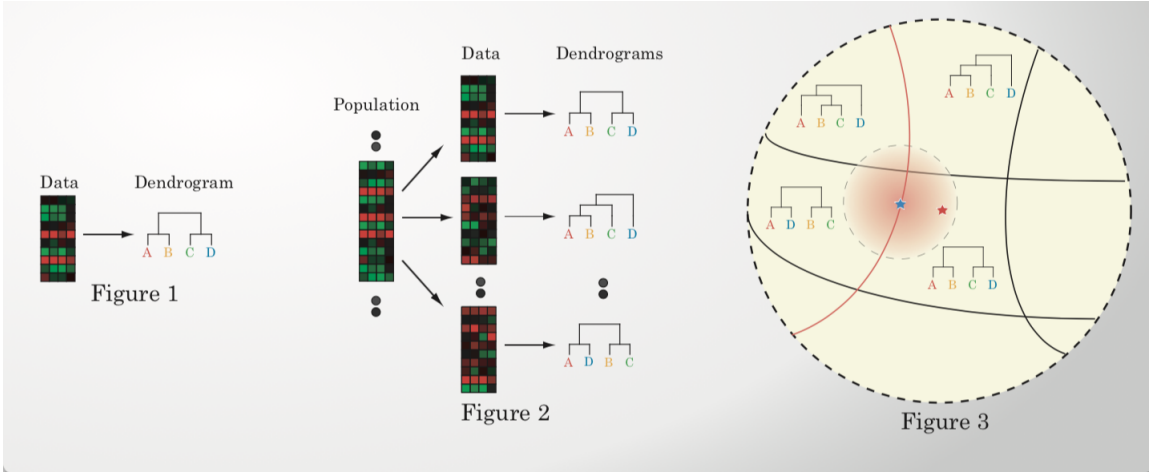

FIGURE 17 – Illustration du problème des régions

Source: [Suzuki and Shimodaira, 2004]

L'étoile bleue se situe à la frontière, dans la région où la première fusion est (B, C). Si on teste l'hypothèse (B, C), à cause du biais introduit par les données initiales, on peut être amené à rejeter cette fusion comme l'illustre la figure 18.

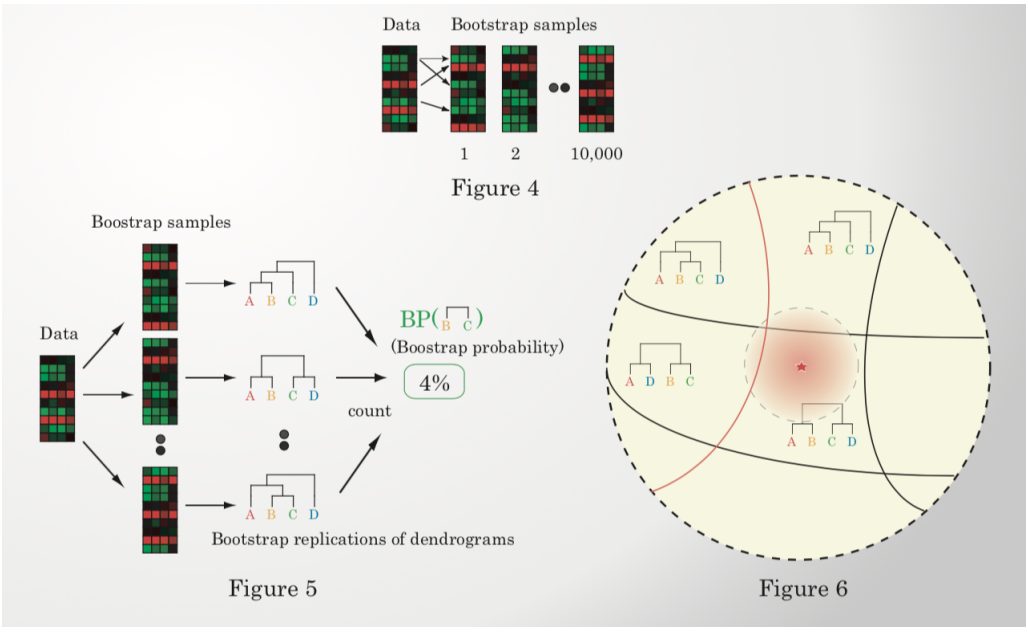

FIGURE 18 – Calcul de la Bootstrap Probability Value

Source: [Suzuki and Shimodaira, 2004]

En réalité, la bootstrap probability value de Felsenstein est une approximation de la valeur calculée par Shimodaira : la Approximately Unbiased value. Elle corrige une part du biais de la Bootstrap Probability value comme illustré dans la Figure 19.

L'algorithme de bootstrap multi-échelles est le suivant. Premièrement, on génère des échantillons bootstrap pour toutes les tailles d'échantillonnages prédéfinies. Dans un second temps, on applique la procédure de classification hiérarchique sur l'ensemble de ces échantillons bootstrap et on calcule la bootstrap probability value pour chacune des tailles d'échantillonnage. Enfin, on utilise les Bootstrap Probability values pour calculer la p-valeur. La p-valeur ainsi estimée est ce qu'on appelle l'Approximately Unbiased value.

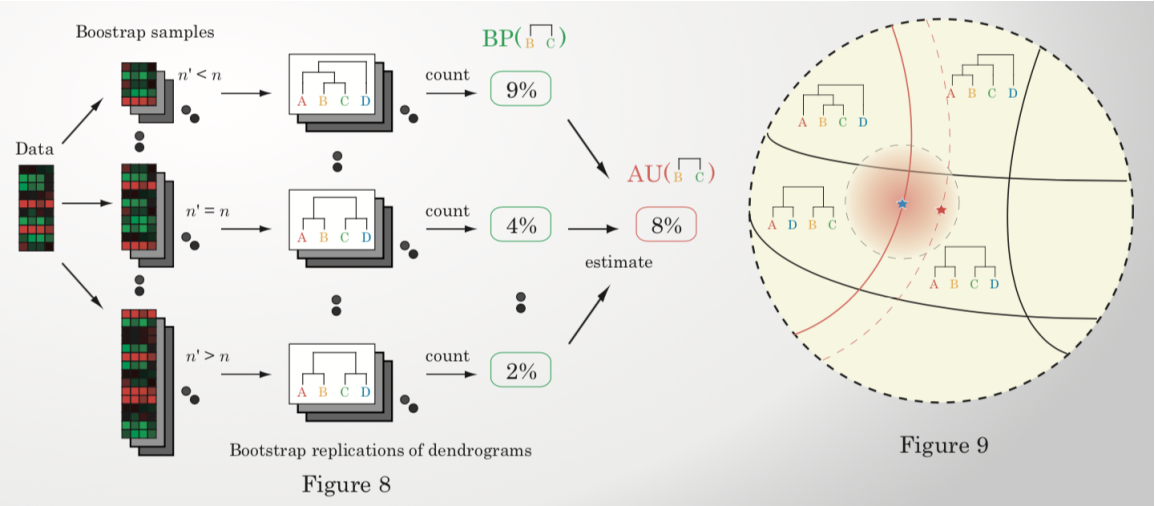

FIGURE 19 – Calcul de l'Approximately Unbiased Value

Source: [Suzuki and Shimodaira, 2004]

Ci-dessous en figure 20, une comparaison des BP et AU values pour un exemple de classification ascendante hiérarchique de tumeurs pulmonaires. Les données consistaient en les profils d'expression de 916 gènes pour 73 tumeurs. Les classes (A, B, C, D, E) ont été établies par des experts. Les valeurs supérieures à 95% désignent des classes fortement suggérées par les données.

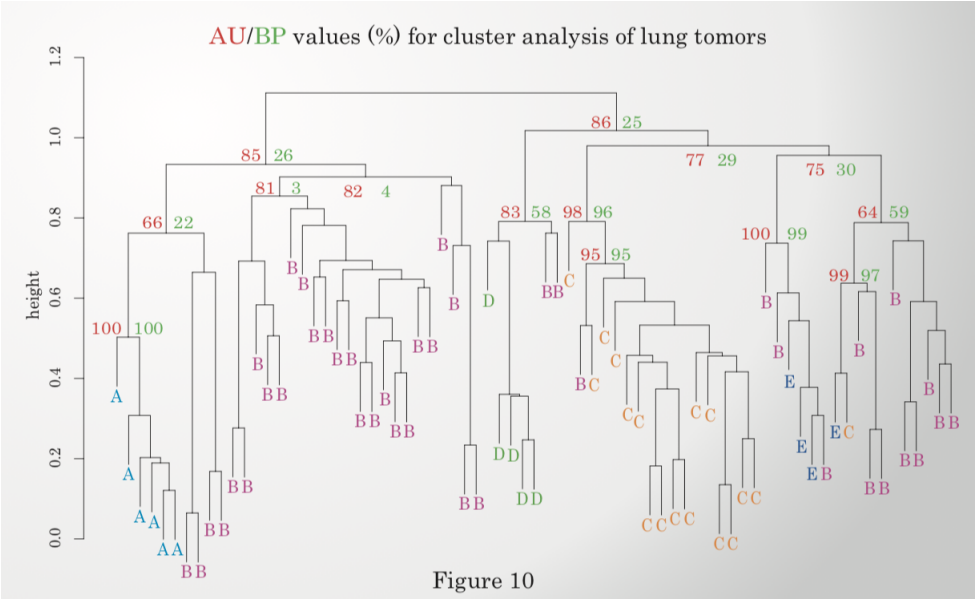

FIGURE 20 – Comparaison entre BP value et AU value

Source: [Suzuki and Shimodaira, 2004]

Conclusion

Ce stage a été pour moi l'occasion de me familiariser avec un sujet et une problématique d'origine biologique. Cette interdisciplinarité a été pour moi l'occasion de découvrir un champ d'application concret des mathématiques appliquées, me faisant ainsi percevoir la nécessité des contributions statistiques pour les autres sciences.

J'ai également pu découvrir précisément les caractéristiques et les propriétés théoriques d'un outil statistique, la classification ascendante hiérarchique, que j'avais déjà rencontré dans le cadre de ma formation, mais cette fois-ci avec un niveau de détail supplémentaire. J'ai pu manipuler une version alternative, sous contrainte de contiguïté, et j'ai eu à étudier les extensions possibles de la méthode, ce qui m'a finalement fait prendre conscience des possibilités de cet outil.

Enfin, ce stage aura été l'occasion pour moi de me former à la recherche notamment grâce à une partie exploratoire d'étude bibliographique. Certains des résultats mis en évidence à cette occasion sont applicables directement comme ceux concernant les critères de dissemblance. D'autres en revanche, ne le sont pas en l'état, et une généralisation des approches bootstrap pour les données Hi-C serait une direction de recherche future intéressante.

Ce travail est pour moi le point de départ d'une réflexion que je souhaite poursuivre pendant plusieurs années. Pouvoir poursuivre en thèse sur ce sujet est une véritable chance d'approfondir toutes les interrogations qui subsistent à l'issue de ce stage.

6 Annexe : preuves

Démonstration du Lemme 4.4.1. Si $i^* = \text{argmin}_{i=1,\dots,n-1} d(x_i, x_{i+1})$ alors,

$$G_i^2 = \begin{cases} \{x_i\} & \text{if } i \leq i^* - 1 \\ \{x_{i^*}, x_{i^*+1}\} & \text{if } i = i^* \\ \{x_{i-1}\} & \text{if } i \geq i^* + 1 \end{cases},$$

avec

$$\forall\, i \neq \{i^*, i^*+1\},\ \delta(G_{i^*}^2, \{x_i\}) = \frac{1}{3}d^2(x_{i^*}, x_i) + \frac{1}{3}d^2(x_{i^*+1}, x_i) - \frac{1}{6}d^2(x_{i^*}, x_{i^*+1}) \qquad (5)$$

et

$$\forall\, i, j \neq \{i^*, i^*+1\},\ \delta(\{x_i\}, \{x_j\}) = \frac{1}{2}d^2(x_i, x_j).$$

Maintenant supposons que $j^* := \text{argmin}_{i=1,\dots,n-2} \delta(G_i^2, G_{i+1}^2) \notin \{i^*-1, i^*\}$, cela signifie :

$$m_2 = \delta(\{x_{j^*}\}, \{x_{j^*+1}\}) \geq \delta(\{x_{i^*}\}, \{x_{i^*+1}\}) = m_1$$

par définition de i^*. Dans ce cas, les deux premiers niveaux de fusions sont croissants.

Au contraire, si $j^* = i^*$, alors les deux niveaux de fusions sont décroissants si et seulement si

$$\delta(G_{i^*}^2, \{x_{i^*+2}\}) < \delta(\{x_{i^*}\}, \{x_{i^*+1}\}) = \frac{1}{2}d^2(x_{i^*}, x_{i^*+1}).$$

En utilisant l'équation (5), la condition précédente est équivalente à

$$\frac{1}{3}d^2(x_{i^*}, x_{i^*+2}) + \frac{1}{3}d^2(x_{i^*+1}, x_{i^*+2}) - \frac{1}{6}d^2(x_{i^*}, x_{i^*+1}) < \frac{1}{2}d^2(x_{i^*}, x_{i^*+1})$$

qui est équivalent à

$$\frac{1}{3}\left(d^2(x_{i^*+1}, x_{i^*+2}) - d^2(x_{i^*}, x_{i^*+1})\right) < \frac{1}{3}\left(d^2(x_{i^*}, x_{i^*+1}) - d^2(x_{i^*}, x_{i^*+2})\right)$$

On peut noter que le membre de gauche de l'inégalité précédente est positif par définition de i^*, ce qui implique automatiquement que :

$$\frac{1}{2}d^2(x_{i^*}, x_{i^*+1}) = \delta(\{x_{i^*}\}, \{x_{i^*+1}\}) > \frac{1}{2}d^2(x_{i^*}, x_{i^*+2}) = \delta(\{x_{i^*}\}, \{x_{i^*+2}\}).$$

□

Lemme. *Pour une classification contrainte à 4 objets et avec même pondération, ESS_t est croissant.*

Preuve du Lemme. Supposons données les quantités d_{ij} pour $i, j = [\![1, 4]\!]^2$, on définit $\delta_{ij} = \frac{1}{2}d_{ij}^2$. La première étape fusionne x_i et x_{i+1} qui minimisent $\delta_{i,i+1}$. Les différents cas peuvent être :

— une fusion de bord, par exemple $A_1 = \{x_1, x_2\}$ et dans ce cas, on a que : $\delta_{12} \leq \delta_{23}$ et $\delta_{12} \leq \delta_{34}$. On également les dissimilarités suivantes :

$$\delta_{A_1,\{x_3\}} = \frac{2}{3}\delta_{13} + \frac{2}{3}\delta_{23} - \frac{1}{3}\delta_{12} \qquad \text{et} \qquad \delta_{A_1,\{x_4\}} = \frac{2}{3}\delta_{14} + \frac{2}{3}\delta_{24} - \frac{1}{3}\delta_{12}.$$

Comme on l'a montré précédemment, on peut avoir que $\delta_{A_1,\{x_4\}} < 0$ mais $\delta_{A_1,\{x_3\}} \geq \frac{2}{3}\delta_{13} + \frac{2}{3}\delta_{23} - \frac{1}{3}\delta_{23} = \frac{2}{3}\delta_{13} + \frac{1}{3}\delta_{23} \geq 0$.
La seconde fusionne soit :

— $A_2 = A_1 \cup \{x_3\}$ et on a donc que $\delta_{A_1,\{x_3\}} \leq \delta_{34}$. Dans cette situation, la dernière fusion est telle que

$$\begin{aligned}
\delta_{A_2,4} &= \frac{3}{4}\delta_{A_1,4} + \frac{2}{4}\delta_{34} - \frac{1}{4}\delta_{A_1,3} \\
&= \frac{3}{4}\left(\frac{2}{3}\delta_{14} + \frac{2}{3}\delta_{24} - \frac{1}{3}\delta_{12}\right) + \frac{2}{4}\delta_{34} - \frac{1}{4}\delta_{A_1,3} \\
&\geq \frac{1}{2}\delta_{14} + \frac{1}{2}\delta_{24} - \frac{1}{4}\delta_{34} + \frac{1}{2}\delta_{34} - \frac{1}{4}\delta_{34} \qquad \text{parce que } \delta_{A_1,3} \leq \delta_{34} \text{ et } \delta_{12} \leq \delta_{34} \\
&= \frac{1}{2}\delta_{14} + \frac{1}{2}\delta_{24} \geq 0,
\end{aligned}$$

qui montre finalement que toutes les valeurs du lien de Ward sont positives dans cette situation.

— $A_2 = \{3, 4\}$. Dans cette situation, on a donc que $\delta_{34} \leq \delta_{A_1,3}$ et la dernière valeur du lien de Ward est

$$\begin{aligned}
\delta_{A_1,A_2} &= \frac{3}{4}\delta_{A_1,3} + \frac{3}{4}\delta_{A_1,4} - \frac{2}{4}\delta_{34} \\
&\geq \frac{3}{4}\delta_{34} + \frac{3}{4}\delta_{A_1,4} - \frac{2}{4}\delta_{34} \qquad \text{parce que } \delta_{34} \geq \delta_{A_1,3} \\
&= \frac{1}{4}\delta_{34} + \frac{3}{4}\left(\frac{2}{3}\delta_{14} + \frac{2}{3}\delta_{24} - \frac{1}{3}\delta_{12}\right) \\
&\geq \frac{1}{4}\delta_{34} + \frac{1}{2}\delta_{14} + \frac{1}{2}\delta_{24} - \frac{1}{4}\delta_{34} \qquad \text{parce que } \delta_{12} \leq \delta_{34} \\
&= \frac{1}{2}\delta_{14} + \frac{1}{2}\delta_{24} \geq 0,
\end{aligned}$$

qui prouve encore que toutes les fusions ont des hauteurs positives dans cette siutation.

— une fusion centrale, $A_1 = \{2, 3\}$. Dans ce cas, on a que $\delta_{23} \leq \delta_{12}$ et $\delta_{23} \leq \delta_{34}$

$$\delta_{A_1,1} = \frac{2}{3}\delta_{12} + \frac{2}{3}\delta_{13} - \frac{1}{3}\delta_{23} \qquad \text{et } \delta_{A_1,4} = \frac{2}{3}\delta_{24} + \frac{2}{3}\delta_{34} - \frac{1}{3}\delta_{23}.$$

Les deux quantités sont positives parce que, puisque $\delta_{23} \leq \delta_{12}$, on a que $\delta_{A_1,1} \geq \frac{2}{3}\delta_{12} + \frac{2}{3}\delta_{13} - \frac{1}{3}\delta_{12} = \frac{1}{3}\delta_{12} + \frac{2}{3}\delta_{13} \geq 0$. Sans perte de généralité, on peut prouver que la prochaine fusion est $A_2 = A_1 \cup \{1\}$, ce qui signifie que $\delta_{A_1,1} \leq \delta_{A_1,4}$.On a finalement pour dernière valeur du lien de Ward :

$$\begin{aligned}
\delta_{A_2,4} &= \frac{3}{4}\delta_{A_1,4} + \frac{2}{4}\delta_{14} - \frac{1}{4}\delta A_1, 1 \\
&\geq \frac{2}{4}\delta_{A_1,4} + \frac{2}{4}\delta_{14} \qquad \text{parce que } \delta_{A_1,1} \leq \delta_{A_1,4}
\end{aligned}$$

et ceci conclut la preuve.

□

Remarque. *Pour l'exemple précédent, on a aussi les inerties intra-classes suivantes :*

$$\mathcal{I}(A_1) = 0.5 \qquad \mathcal{I}(A_2) = \frac{1}{3}(1 + 0.1 + 2) = \frac{3.1}{3} \qquad \mathcal{I}(A_3) = \frac{1}{4}(1 + 3 \times 0.1 + 2 + 1.1) = 1.1$$

qui donne

$$\overline{\mathcal{I}}(A_1) = 0.25 \qquad \overline{\mathcal{I}}(A_2) = \frac{3.1}{9} \simeq 0.34 \qquad \overline{\mathcal{I}}(A_3) = \frac{1.1}{4} = 0.275$$

qui est suffisant pour montrer que, dans ce cas, $\overline{\mathcal{I}}(t)$ *est décroissant lors de la dernière fusion.*

Références

[Aronszajn, 1950] Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3) :337–337.

[Ay and Noble, 2015] Ay, F. and Noble, W. S. (2015). Analysis methods for studying the 3d architecture of the genome. *Genome Biology*, 16(1).

[Bonev and Cavalli, 2016] Bonev, B. and Cavalli, G. (2016). Organization and function of the 3d genome. *Nature Reviews Genetics*, 17(11) :661–678.

[Cailliez, 1983] Cailliez, F. (1983). The analytical solution of the additive constant problem. *Psychometrika*, 48(2) :305–308.

[Chavent et al., 2017] Chavent, M., Kuentz-Simonet, V., Labenne, A., and Saracco, J. (2017). Clustgeo : an r package for hierarchical clustering with spatial constraints. *Computational Statistics*.

[Cormack, 1971] Cormack, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society. Series A (General)*, 134(3) :321.

[Dehman, 2015] Dehman, A. (2015). *Spatial clustering of linkage disequilibrium blocks for genome-wide association studies*. PhD thesis, Ecole Doctorale Struture et Dynamique des Systèmes Vivants - Paris Saclay.

[Diaconis and Holmes, 1998] Diaconis, P. W. and Holmes, S. P. (1998). Matchings and phylogenetic trees. *Proceedings of the National Academy of Sciences of the United States of America*, 95 :14600–14602.

[Dixon et al., 2016] Dixon, J. R., Gorkin, D. U., and Ren, B. (2016). Chromatin domains : The unit of chromosome organization. *Molecular Cell*, 62(5) :668–680.

[Dixon et al., 2015] Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V. V., Ecker, J. R., Thomson, J. A., and Ren, B. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539) :331–336.

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

[Efron, 1982] Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics.

[Efron and Tibshirani, 1986] Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1) :77–77.

[Felsenstein, 1985] Felsenstein, J. (1985). Confidence limits on phylogenies : An approach using the bootstrap. *Evolution ; international journal of organic evolution*, 39 :783–791.

[Ferligoj and Batagelj, 1982] Ferligoj, A. and Batagelj, V. (1982). Clustering with relational constraint. *Psychometrika*, 47(4) :413–426.

[Forcato et al., 2017] Forcato, M., Nicoletti, C., Pal, K., Livi, C. M., Ferrari, F., and Bicciato, S. (2017). Comparison of computational methods for hi-c data analysis. *Nature Methods*, 14(7) :679–685.

[Fowlkes and Mallows, 1983] Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383) :553–569.

[Giorgio et al., 2015] Giorgio, E., Robyr, D., Spielmann, M., Ferrero, E., Gregorio, E. D., Imperiale, D., Vaula, G., Stamoulis, G., Santoni, F., Atzori, C., Gasparini, L., Ferrera, D., Canale, C., Guipponi, M., Pennacchio, L. A., Antonarakis, S. E., Brussino, A., and Brusco, A. (2015). A large genomic deletion leads to enhancer adoption by the lamin b1 gene : a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD). *Human Molecular Genetics*, 24(11) :3143–3154.

[Gómez-Díaz and Corces, 2014] Gómez-Díaz, E. and Corces, V. G. (2014). Architectural proteins : regulators of 3d genome organization in cell fate. *Trends in cell biology*, 24 :703–711.

[Grimm, 1987] Grimm, E. C. (1987). CONISS : a FORTRAN 77 program for stratigraphically constrained cluster analysis by the method of incremental sum of squares. *Computers & Geosciences*, 13(1) :13–35.

[Kaufman and Rousseeuw, 1990] Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data : An Introduction to Cluster Analysis*. Wiley-Interscience.

[Lance and Williams, 1967] Lance, G. N. and Williams, W. T. (1967). A general theory of classificatory sorting strategies : 1. hierarchical systems. *The Computer Journal*, 9(4) :373–380.

[Levy-Leduc et al., 2014] Levy-Leduc, C., Delattre, M., Mary-Huard, T., and Robin, S. (2014). Two-dimensional segmentation for analyzing hi-c data. *Bioinformatics*, 30(17) :i386–i392.

[Lieberman-Aiden et al., 2009] Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950) :289–293.

[Lingoes, 1971] Lingoes, J. C. (1971). Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika*, 36(2) :195–203.

[Lupiáñez et al., 2015] Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., and Mundlos, S. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5) :1012–1025.

[Miyamoto et al., 2015] Miyamoto, S., Abe, R., Endo, Y., and ichi Takeshita, J. (2015). Ward method of hierarchical clustering for non-euclidean similarity measures. In *2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*. IEEE.

[Nora et al., 2012] Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., and Heard, E. (2012). Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485 :381–385.

[Olteanu and Villa-Vialaneix, 2015] Olteanu, M. and Villa-Vialaneix, N. (2015). On-line relational and multiple relational SOM. *Neurocomputing*, 147 :15–30.

[Rand, 1971] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336) :846–850.

[Schölkopf et al., 2004] Schölkopf, B., Tsuda, K., and Vert, J.-P. (2004). *Kernel Methods in Computational Biology*. MIT PR.

[Serra et al., 2016] Serra, F., Baù, D., Filion, G., and Marti-Renom, M. A. (2016). Structural features of the fly chromatin colors revealed by automatic three-dimensional modeling.

[Sexton et al., 2012] Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, 148(3) :458–472.

[Shimodaira, 2002] Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, 51(3) :492–508.

[Sokal and Rohlf, 1962] Sokal, R. R. and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, 11(2) :33.

[Suzuki and Shimodaira, 2004] Suzuki, R. and Shimodaira, H. (2004). An application of multiscale bootstrap resampling to hierarchical clustering of microarray data : How accurate are these clusters ? 2004.

[Ward, 1963] Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301) :236–244.

[Weinreb and Raphael, 2015] Weinreb, C. and Raphael, B. J. (2015). Identification of hierarchical chromatin domains. *Bioinformatics*, 32(11) :1601–1609.