Classification ascendante hiérarchique, contrainte d'ordre : conditions d'applicabilité, interprétabilité des dendrogrammes

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

[1]MIAT, Université de Toulouse, INRA, Castanet Tolosan, France
[2]Institut de Mathématiques de Toulouse, UMR 5219, Univ. de Toulouse, CNRS, France

9 avril 2019

Résumé

La classification ascendante hiérarchique (CAH) avec lien de Ward est une approche fréquemment utilisée pour partitionner des objets d'un espace multi-dimensionnel. Nous discutons ici de la validité théorique des extensions de cette approches à différents types d'entrées (distances, dissimilarités, noyaux, similarités) ainsi que de la pertinence de l'utilisation de sa version contrainte (parfois appelée « CONISS ») dans ces divers cas. En particulier, nous étudions les conditions garantissant la cohérence entre les résultats de la CAH et leur représentation graphique classique sous forme de dendrogramme. Nous présentons également une étude de simulations montrant que la version contrainte de la méthode permet, outre des gains computationnels, d'obtenir de meilleurs résultats de classification lorsque la contrainte est cohérente avec la structure des données. Beaucoup de ces résultats sont présents dans une littérature hétéroclite : l'objet de cette communication est de proposer donc un cadre de présentation uniforme et de compléter les résultats existants.

Mots-clefs : Classification ascendante hiérarchique, lien de Ward, classification ascendante hiérarchique sous contrainte, dendrogramme, croissance.

1 Classification Ascendante Hiérarchique (CAH)

Cadre euclidien standard

Dans le cadre standard de la Classification Ascendante Hiérarchique (CAH), on suppose que l'ensemble des objets $\Omega = \{x_1, \dots, x_n\}$ est un sous-ensemble de $\mathbb{R}^p$, et que la relation de proximité entre ces objets est encodée par la distance $D = (d_{ij})_{1 \leq i,j \leq n}$ induite par la norme euclidienne de $\mathbb{R}^p$ avec $d_{ij} = \|x_i - x_j\|_{\mathbb{R}^p}$.

L'algorithme de CAH part de la partition triviale en singletons $\mathcal{P}_1$, puis par fusions successives, se termine sur une autre partition triviale, $\mathcal{P}_n = \Omega$. L'étape t, permettant de passer de la partition $\mathcal{P}_t$ à la partition $\mathcal{P}_{t+1}$, fusionne les deux classes de $\mathcal{P}_t$ les plus proches au sens d'une dissimilarité entre classes appelée *critère de lien* et notée δ.

Le *lien de Ward* [War63] est fréquemment utilisé car il a une interprétation simple. Pour un ensemble quelconque $G \subset \Omega$, l'inertie de G est définie par $I(G) = \sum_{x \in G} \|x - \bar{x}_G\|^2$, où $\bar{x}_G = |G|^{-1} \sum_{x \in G} x$ est le centre de gravité de G, et $|G|$ le cardinal de G. On appelle inertie intra-classes d'une partition $\mathcal{P} = (G_1, G_2, ..., G_K)$ la somme des inerties des classes : $I(\mathcal{P}) = \sum_{k=1}^{K} I(G_k)$. Le lien de Ward entre deux sous-ensembles disjoints G et G' est défini par : $\delta(G, G') = I(G \cup G') - I(G) - I(G')$ et correspond alors à la variation d'inertie intra-classes induite par cette fusion.

Extensions de la CAH

La CAH est couramment utilisée sur des données plus générales que celles décrites dans le cadre euclidien de la section précédente. On suppose désormais que les objets $(x_i)_i$ appartiennent à un espace arbitraire (non nécessairement euclidien).

Dissimilarités. Considérons tout d'abord le cas où les objets sont décrits par une mesure de dissimilarité entre paires, $d_{ij} = d(x_i, x_j)$, qui n'est ici plus nécessairement la distance euclidienne. On suppose alors que la matrice $D = (d_{ij})_{1 \leq i,j \leq n}$ est une matrice symétrique à coefficients positifs et à diagonale nulle. Remarquons que ce cas est une extension du

cas euclidien décrit dans la section précédente, dans lequel l'inertie $I(G)$ s'exprime directement à partir des éléments de la matrice de distance D comme $I(G) = (2|G|)^{-1} \sum_{(x_i,x_j)\in G^2} d_{ij}^2$. On peut alors définir, par analogie au cas euclidien, une extension de l'inertie et du lien de Ward associée à D (voir par exemple [SR05] ou [CKSLS18]). Cependant, dans ce cas, on perd l'interprétabilité en termes de centre de gravité.

Noyaux. Dans un certain nombre de cas pratiques, 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 [STV04], c'est-à-dire, par une matrice symétrique définie positive $K = (k(x_i, x_j))_{1\leq i,j\leq n}$. [Aro50] montre que le noyau peut être interprété comme une matrice de produit scalaire dans un certain espace de représentation $\mathcal{H}$. Ce cadre-ci est donc équivalent au cadre euclidien et on peut montrer que l'expression du lien de Ward s'écrit à partir des valeurs du noyau uniquement en utilisant l'*astuce noyau* [Deh15] :

$$\delta(G, G') = \frac{K(G)}{|G|} + \frac{K(G')}{|G'|} - \frac{K(G \cup G')}{|G \cup C'|}, \quad (1)$$

avec $K(G) = \sum_{x_i,x_j\in G^2} k(x_i, x_j)$ pour un sous-ensemble, G, d'éléments de $(x_i)_i$.

Similarités. On peut interpréter la notion de similarité comme une généralisation de celle de noyau. Bien qu'il n'y ait pas de consensus sur la définition exacte d'une similarité, on appellera similarité une matrice $S = (s_{ij})_{1\leq i,j\leq n}$ symétrique à diagonale positive. [MAET15] ont montré que l'approche « noyau » décrite ci-dessus pouvait être généralisée à toute matrice de similarité, par l'argument suivant : considérons la matrice $S + \lambda I_n$, c'est-à-dire la matrice S dont les éléments diagonaux sont translatés de λ. Alors, d'une part, pour λ suffisamment grand, S_λ est définie positive, et peut donc s'interpréter comme un noyau. D'autre part, appliquer (formellement) l'algorithme de CAH à la matrice de similarité S_λ en utilisant l'équation (1) avec $K = S_\lambda$ aboutit exactement à la même suite de fusions, quelle que soit la valeur de λ. En effet, l'équation (1) implique : $\delta_{S_\lambda} = \delta_{S_{\lambda'}} + (\lambda - \lambda')$.

Dans la suite, nous distinguerons simplement deux cas : le cas euclidien (qui contient d'après ce qui précède les cas « noyau » et « similarité quelconque »), et le cas non-euclidien, qui correspond aux dissimilarités quelconques.

CAH sous contrainte d'ordre

Dans un certain nombre de contextes applicatifs, il existe une information *a priori* sur les relations entre les objets. C'est le cas, en particulier, en statistique spatiale où les objets spatiaux sont en relation de voisinage, ou bien dans le contexte génomique, où les loci (positions) génomiques sont ordonnés le long d'une ligne (le chromosome, voir [ADN+19] pour des applications aux données de génétique, SNP, et aux données de conformation spatiale de la chromatine, Hi-C). La classification ascendante hiérarchique sous contrainte de contiguïté [FB82] restreint les fusions possibles aux objets dits contigus.

Dans la suite, on s'intéresse au cas particulier de la Classification Ascendante Hiérarchique sous Contrainte d'Ordre (CAHCO) : les objets initiaux sont reliés par une relation d'ordre total (temps, position génomique), et deux classes sont dites contiguës si et seulement si on peut en extraire un couple d'objets contigus. Cette version est souvent appelée « CONISS » pour CONstrained Incremental Sum of Squares [Gri87] lorsque le lien de Ward est utilisé pour la fusion des classes.

L'intérêt de l'ajout d'une contrainte est double : d'une part, elle permet de répercuter au mieux les relations existantes entre les objets au cours de la procédure, et ainsi de rendre les résultats plus interprétables. La CAHCO peut ainsi être vue comme une méthode de segmentation de données ordonnées linéairement. D'autre part, la complexité en temps de la CAH sans contrainte est cubique ($O(n^3)$) et celle de la CAHCO est seulement quadratique ($O(n^2)$) [Deh15]. Lorsque les données d'entrées sont une similarité creuse dont les éléments non nuls sont proches de la diagonale, [ADN+19] ont proposé une approche permettant d'obtenir les résultats d'une CAHCO avec une complexité quasi-linéaire. Dans ce cas particulier, la CAHCO peut être utilisée comme une heuristique intéressante pour la segmentation.

2 Dendrogrammes

Les résultats d'une CAH sont fréquemment représentés à l'aide d'un *dendrogramme*, comme sur la figure 1. Un dendrogramme est un arbre binaire dont les nœuds sont les classes de la hiérarchie. Dans le cas particulier de la CAHCO, les feuilles, qui correspondent aux n objets à classer, sont représentées selon l'ordre naturel de ces objets. La hauteur du nœud correspondant à la t-ème fusion est notée h_t (les feuilles sont à la hauteur $h_0 = 0$). On utilise souvent comme hauteur la valeur du lien de Ward à l'étape t, notée m_t.

Une propriété naturelle attendue pour le dendro-

gramme d'une CAH est que la suite de partitions induite par la CAH coïncide avec celle décrite par le dendrogramme. Cette propriété est équivalente à la croissance de la suite $(h_t)_t$. En effet, lorsqu'un dendrogramme est construit avec une suite de hauteurs non croissante (comme sur la figure 1 entre les fusions 3 et 4), il est difficilement interprétable et l'obtention d'une partition par coupe du dendrogramme a une hauteur donnée n'est pas définie correctement et n'est pas cohérente avec la suite de partitions fournies par la CAH. En particulier, lorsque la hauteur est définie par le lien de Ward, la croissance n'est pas nécessairement vérifiée pour la CAHCO [Gri87].

[Gri87] décrit des choix de hauteurs alternatifs au lien de Ward pour pallier ce problème. Parmi ceux-ci, on trouve **l'inertie intra-classes**, qui est fréquemment utilisée dans le cadre de la CAHCO (c'est notamment cette hauteur qui est fournie par la version de la CAHCO implémentée dans le package R **rioja**). Pour ce critère, la hauteur à l'étape t est définie par $\mathrm{ESS}_t = \sum_{u=1}^{n-t} I(G_u^{t+1})$. Comme pour m_t, cette hauteur est croissante pour la CAH [War63, Bat81]. Sa croissance pour la CAHCO n'est pas non plus assurée dans le cas non-euclidien, mais elle l'est dans le cas euclidien [Gri87]. La figure 1 donne un exemple simple où la CAHCO est appliquée à six objets décrits par la dissimilarité $D = \begin{pmatrix} 0 & \sqrt{1.99} & \sqrt{1.99} & \sqrt{1.99} & 0.1 & 1 \\ \sqrt{1.99} & 0 & \sqrt{2} & \sqrt{1.99} & 0.1 & 1 \\ \sqrt{1.99} & \sqrt{2} & 0 & \sqrt{2} & 0.1 & 1 \\ \sqrt{1.99} & \sqrt{1.99} & \sqrt{2} & 0 & \sqrt{2} & 1 \\ 0.1 & 0.1 & 0.1 & \sqrt{2} & 0 & \sqrt{2} \\ 1 & 1 & 1 & 1 & \sqrt{2} & 0 \end{pmatrix}$, pour lesquels on obtient des hauteurs non croissantes : $h_1 < h_2 < h_4 < h_3 < h_5$.

3 Comparaison de CAH et CAHCO

La CAH est souvent perçue comme une approche gloutonne permettant de déterminer une partition des données avec une inertie intra-classes, ESS_t, minimale pour un nombre donné, k ($\in \{1, \ldots, n\}$) de classes. Aussi, il est raisonnable d'attendre que l'inertie des partitions obtenues avec la CAH standard soit inférieure à l'inertie des partitions obtenues par une CAH contrainte comme la CAHCO, dans laquelle l'ensemble des fusions possibles est restreint. Dans cette section, nous montrons que dans le cas où les données sont structurées de manière « compatible » avec la contrainte imposée à la CAHCO, non seulement l'algorithme contraint est plus efficace que l'algorithme standard (en termes de complexité en temps et espace), mais il permet aussi d'obtenir des partitions de plus faible inertie intra-classes, c'est-à-dire meilleures. Pour ce faire, nous proposons un processus aléatoire de perturbation de la structure originale de données génomiques et quantifions l'impact de cette perturbation sur la qualité des résultats des classifications.

Données et méthode de simulation

Pour cela, nous avons analysé des données Hi-C [DSY+12], qui présentent une forte structure d'ordre. Les données Hi-C permettent d'étudier les relations de proximité dans l'espace (3D) de la chromatine en mesurant la fréquence d'interactions physiques entre paires de positions génomiques par du séquençage haut-débit. De manière formelle, les données Hi-C peuvent être représentées sous la forme d'une matrice symétrique, $S = (s_{ij})_{i,j}$ dans laquelle chaque valeur s_{ij} correspond au nombre d'interactions mesurées entre les positions génomiques i et j. Les positions génomiques i et j représentent des intervalles de longueur identique le long de la séquence d'ADN, et les valeurs s_{ij} sont par définition des entiers positifs (ou nuls). La matrice présente une structure diagonale forte, qui reflète l'organisation linéaire le long des chromosomes, comme illustré dans la figure 2 qui représente la partie triangulaire supérieure d'une carte Hi-C. Comme décrit dans [ADN+19], la classification contrainte des positions génomiques selon cette matrice de similarité est liée à des questions biologiques de compréhension de la structure de la chromatine (détection de TADs [ZTOC18] ou de compartiments actifs/inactifs [LAvBW+09]).

Dans les simulations qui suivent, nous avons utilisé un chromosome (le chromosome 3) d'une expérience sur des cellules souches publiée dans [DSY+12] [1]. Les données ont été log-transformées avant les classifications, pour limiter l'asymétrie des valeurs ; la correction de la similarité décrite dans la section 1 pour transformer cette similarité en noyau a également été utilisée.

Pour quantifier l'influence de la structure des données sur la qualité des partitions obtenues, nous avons utilisé un processus de perturbation des données initiales qui gomme progressivement la structure diagonale. Ce processus consiste à échanger, de manière aléatoire, des paires d'entrées, s_{ij} et $s_{i'j'}$, où (i, j) et (i', j') sont tirés de manière uniforme parmi l'ensemble des paires (u, v), $(u', v') \in \{1, \ldots, n\}$ telles que

1. La carte Hi-C utilisée est la carte des données normalisée, disponible sur le site internet des auteurs : `http://chromosome.sdsc.edu/mouse/hi-c/download.html`.

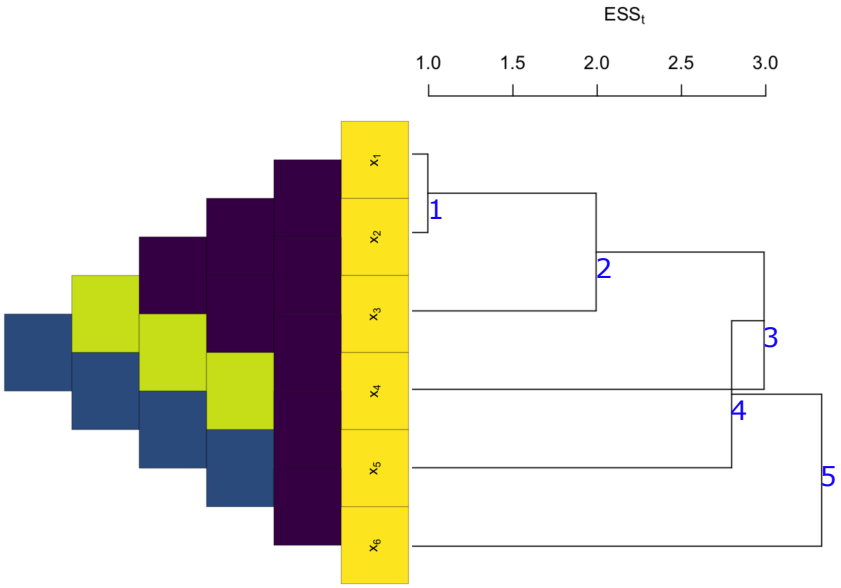

FIGURE 1 – **Un croisement dû à la non-croissance de la hauteur définie par l'inertie intra-classes, ESS$_t$, pour la CAHCO avec données non euclidiennes**. À gauche : représentation de la dissimilarité associée (les couleurs sombres correspondent à de grandes valeurs donc à des objets distants). À droite : dendrogramme correspondant aux résultats de la CAHCO (l'ordre des fusions est noté en bleu).

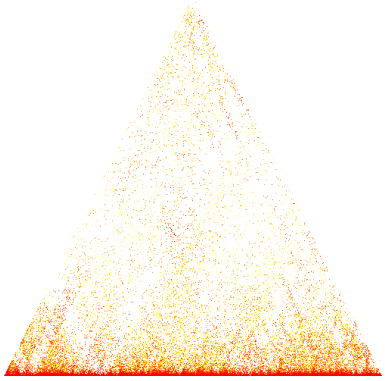

FIGURE 2 – **Partie triangulaire supérieure d'une carte Hi-C**. L'axe horizontal donne la position sur le chromosome et les niveaux de rouge indiquent l'intensité de la valeur s_{ij}.

$s_{uv} > 0$ (pour assurer de ne pas échanger deux valeurs nulles). La proportion des entrées de la matrice perturbée par rapport aux entrées de la matrice initiale varie de 1 à 30% [2]. Ce processus de simulation est répété 50 fois pour évaluer la variabilité des résultats. Tous les résultats ont été obtenus avec R [R C19]. Les résultats de la CAH standard ont été obtenus avec la fonction `hclust` (du package **stats**) et les résultats de la CAHCO ont été obtenus avec la fonction `adjClust` (du package **adjclust**).

Résultats

La figure 3 montre l'évolution de m_t (normalisée par sa valeur maximale pour la simulation) et de ESS$_t$ (normalisée par l'inertie totale, ESS$_n$) pour la CAH et la CAHCO lorsque le pourcentage de perturbation de la matrice initiale croît.

Pour les données originales (non perturbées), dans lesquelles l'organisation est cohérente avec la contrainte linéaire d'ordre, les hauteurs de la CAH standard et de la CAHCO sont très similaires. De manière

2. À cause du grand nombre de 0 dans les entrées des matrices Hi-C, 30% correspond approximativement à la proportion maximale.

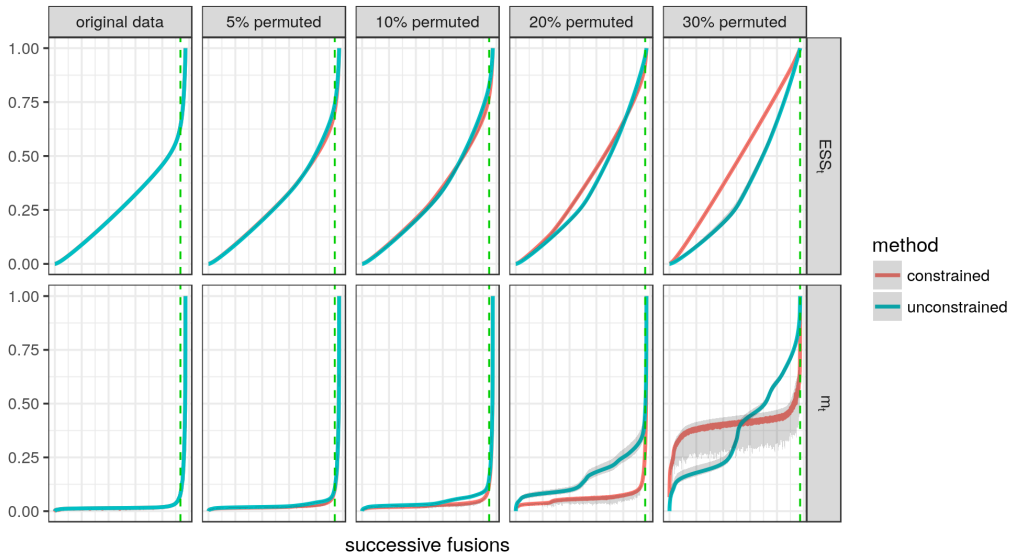

FIGURE 3 – **Comparaison de la séquence de hauteurs pour OCHAC (bleu) et HAC standard (rouge)** pour m_t (en haut) et ESS_t (en bas) pour différents niveaux de perturbation de la matrice Hi-C initiale. Les courbes correspondent à la moyenne du critère pour 50 simulations et les ombres grises au minimum et au maximum du critère sur les 50 simulations. La ligne verticale correspond au nombre de classes choisi par l'heuristique « *broken stick* ».

intéressante, la CAHCO améliore le critère ESS_t pour de faibles pourcentages de perturbations (de 5 à 10%), ce qui montre une meilleure robustesse de l'approche aux petites perturbations lorsque les données originales sont bruitées. Notons, en outre, que l'utilisation d'un critère classique de choix du nombre de classes, comme le critère « *broken stick* » [Ben96] induit la coupe du dendrogramme dans la zone où précisément ESS_t pour la CAHCO est inférieur à ESS_t pour la CAH standard. Cette propriété est intéressante pour les données rencontrées en biologie par exemple, qui présentent de nombreux biais et bruits (effets expérimentaux, biais de GC par exemple). La CAHCO est donc à privilégier dans les contextes où la contrainte d'ordre est pertinente. Enfin, pour des pourcentages élevés de perturbations (supérieurs à 20%), la CAH standard a, de nouveau, une valeur de ESS_t partout inférieure à celle obtenue avec la CAHCO. Ceci s'explique par le fait que la contrainte d'ordre n'est plus en accord, à ce niveau de perturbations, avec la structure des données. Dans ce cas, les résultats de la CAHCO ne sont plus pertinents.

Enfin, la comparaison des structures des dendrogrammes ou des classifications induites par la coupe du dendrogramme à une hauteur donnée (non montré faute de place) conduit aux mêmes observations. Les résultats sont très similaires pour CAH et CAHCO pour des données dont la structure est cohérente avec la contrainte mais deviennent rapidement différents lorsque la matrice initiale est perturbée. En particulier, l'indice γ de Baker [Bak74], permettant de comparer deux dendrogrammes, décroît très fortement à partir de 10% de perturbation et devient quasiment nul au delà de 20% de perturbation.

4 Conclusion

La CAH ainsi que sa version sous contrainte d'ordre peuvent s'appliquer de façon justifiée dans un cadre plus vaste que le cadre euclidien. En revanche, certaines propriétés théoriques qui garantissent la cohérence entre les résultats de la méthode et sa représentation graphique ne sont pas toujours garanties, en particulier, lorsque les données ne sont pas euclidiennes. Par ailleurs, nos simulations montrent que la CAH sous contrainte d'ordre peut donner des résultats de meilleure qualité que la CAH standard lorsque les données présentent une structure cohérente avec la contrainte. La complexité de la version contrainte est également inférieure à la version standard, ce qui en fait une approche pertinente pour la classification non supervisée.

Remerciements

Les auteurs remercient Marie Chavent pour des discussions stimulantes autour de ce travail. Celui-ci a été effectué dans le cadre du projet SCALES, financé par la Mission pour les Initiatives Transverses et Interdisciplinaires du CNRS. La thèse de N.R. est financée par le programme INRA/Inria.

Références

[ADN+19] Christophe Ambroise, Alia Dehman, Pierre Neuvial, Guillem Rigaill, and Nathalie Vialaneix. Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics. arXiv preprint arXiv :1902.01596, 2019.

[Aro50] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3) :337–337, 1950.

[Bak74] Frank B. Baker. Stability of two hierarchical grouping techniques case I : sensitivity to data errors. *Journal of the American Statistical Association*, 69(346) :440–445, 1974.

[Bat81] Vladimir Batagelj. Note on ultrametric hierarchical clustering algorithms. *Psychometrika*, 46(3) :351–352, 1981.

[Ben96] Keith D. Bennett. Determination of the number of zones in a biostratigraphical sequence. *New Phytologist*, 132(1) :155–170, 1996.

[CKSLS18] Marie Chavent, Vanessa Kuentz-Simonet, Amaury Labenne, and Jérôme Saracco. ClustGeo2 : an R package for hierarchical clustering with spatial constraints. *Computational Statistics*, 33(4) :1799–1822, 2018.

[Deh15] Alia Dehman. *Spatial Clustering of Linkage Disequilibrium Blocks for Genome-Wide Association Studies*. PhD thesis, Université Paris Saclay, 2015.

[DSY+12] J.R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J.S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485 :376–380, 2012.

[FB82] Anuška Ferligoj and Vladimir Batagelj. Clustering with relational constraint. *Psychometrika*, 47(4) :413–426, 1982.

[Gri87] Eric C. Grimm. CONISS : a FORTRAN 77 program for stratigraphically constrained analysis by the method of incremental sum of squares. *Computers & Geosciences*, 13(1) :13–35, 1987.

[LAvBW+09] E. Lieberman-Aiden, N.L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B.R. Lajoie, P.J. Sabo, M.O. Dorschner, R. Sandstrom, B. Bernstein, M.A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L.A. Mirny, E.S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950) :289–293, 2009.

[MAET15] Sadaaki Miyamoto, Ryosuke Abe, Yasunori Endo, and Jun-Ichi Takeshita. Ward method of hierarchical clustering for non-Euclidean similarity measures. In *Proceedings of the VIIth International Conference of Soft Computing and Pattern Recognition (SoCPaR 2015)*, Fukuoka, Japan, 2015. IEEE.

[R C19] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[SR05] Gábor J. Székely and Maria L. Rizzo. Hierarchical clustering via joint between-within distances : extending Ward's minimum variance method. *Journal of Classification*, 22(2) :151–183, 2005.

[STV04] Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert. *Kernel Methods in Computational Biology*. MIT PR, 2004.

[War63] Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301) :236–244, 1963.

[ZTOC18] Marie Zufferey, Daniele Tavernari, Elisa Oricchio, and Giovanni Ciriello. Comparison of computational methods for the identification of topologically associating domains. *Genome biology*, 19(1) :217, 2018.