Analyse différentielle de données Hi-C via la Classification Ascendante Hiérarchique sous Contrainte de Contiguïté

Nathanaël Randriamihamison, Marie Chavent, Sylvain Foissac, Nathalie Vialaneix, Pierre Neuvial

Analyse différentielle de données Hi-C via la Classification Ascendante Hiérarchique sous Contrainte de Contiguïté

Nathanaël Randriamihamison [1,2,3] & Marie Chavent [3] & Sylvain Foissac [4] & Nathalie Vialaneix [4] & Pierre Neuvial [3]

[1] *INRAE, UR875 Mathématiques et Informatique Appliquées Toulouse, F-31326 Castanet-Tolosan, France,* {*nathanael.randriamihamison,nathalie.vialaneix*}*@inrae.fr*

[2] *Institut de Mathématiques de Toulouse, Univ. Paul Sabatier, UMR 5219, pierre.neuvial@math.univ-toulouse.fr*

[3] *Inria BSO, CQFD Team, CNRS UMR5251 Institut Mathématiques de Bordeaux, marie.chavent@inria.fr*

[4] *GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet-Tolosan, France, sylvain.foissac@inrae.fr*

Résumé.

Les données Hi-C mesurent la proximité spatiale entre paires de positions génomiques et donnent des informations sur l'organisation 3D de l'ADN qui, elle-même, a un rôle important dans la régulation de l'expression des gènes. Le but de l'analyse différentielle de données Hi-C est de trouver des différences significatives entre la structure 3D du génome de deux conditions biologiques différentes à partir de plusieurs réplicats d'expériences Hi-C dans chaque condition. Ici, nous proposons une nouvelle méthode d'analyse différentielle basée sur la Classification Ascendante Hiérarchique avec Contrainte de Contiguïté (CAHCC). Celle-ci est utilisée pour représenter la structure hiérarchique des positions génomiques sous la forme d'un arbre binaire et le problème de l'analyse différentielle Hi-C est alors transformé en un problème de comparaison d'arbres, résolu en utilisant des distances entre arbres.

Mots-clés. Classification ascendante hiérarchique, classification ascendante hiérarchique sous contrainte, dendrogramme, données Hi-C, analyse différentielle, distances entre arbres.

Abstract.

The spatial proximity between pairs of genomic positions can be measured by Hi-C experiments, which give insights into the 3D organization of DNA. This organization plays an important role in the regulation of gene expression. The aim of Hi-C differential analysis is to find significant differences in the 3D structure of the genome between two biological conditions from replicates of Hi-C experiments in each condition. Here, we present a new differential analysis method based on Hierarchical Agglomerative Clustering with Contiguity Constraint (CCHAC). CCHAC is used to represent the hierarchical structure of genomic positions in the form of a binary tree. The problem of Hi-C differential analysis is then translated into a tree comparison problem and handled using tree distances.

Keywords. Hierarchical agglomerative clustering, constrained hierarchical agglomerative clustering, dendrogram, Hi-C data, differential analysis, tree distances.

1 Introduction

Données Hi-C et analyse différentielle

Les données Hi-C, issues du séquençage haut-débit de nouvelle génération, sont des données génomiques qui renseignent sur l'organisation spatiale des chromosomes dans le noyau des cellules. Elles nous permettent d'avoir accès à une mesure de la proximité spatiale entre paires de positions à travers l'ensemble du génome et elles ont permis de mettre en évidence l'existence de régions du génome très compactées [Dixon et al., 2012]. Cette organisation spatiale et ses variations ont des répercussions dans la régulation de l'expression des gènes, jouant notamment un rôle déterminant dans le développement de certaines maladies ou malformations [Lupiáñez et al., 2015].

En pratique, les données Hi-C se présentent sous la forme de matrices, aussi appelées *cartes Hi-C*, dont l'entrée à la position (i, j) correspond au comptage (obtenu par le séquençage haut débit) du nombre de fois où les intervalles génomiques i et j ont été observés en contact. L'intensité des coefficients décroît avec l'éloignement à la diagonale, ce qui est dû à l'organisation linéaire du génôme dans les molécules d'ADN. Les matrices sont donc des matrices de similarité, carrées, symétriques, à entrées entières et positives. Dans la suite, on notera $(\mathbf{H}^t)_{1\leq t\leq T}$, T matrices Hi-C, de dimension $p \times p$ (p est le nombre d'intervalles génomiques considérés), obtenues dans deux conditions biologiques différentes, $\mathcal{C}_1$ et $\mathcal{C}_2$, telles que $\mathcal{C}_1 \cup \mathcal{C}_2 = \{1, \ldots, T\}$ et $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$. On notera aussi h_{ij}^t le nombre de contacts entre les positions génomiques i et j pour la matrice $\mathbf{H}^t$ [1]. L'objet de ce travail est l'analyse différentielle de données Hi-C, c'est-à-dire, la recherche de différences, avec une garantie statistique, entre les matrices de la condition $\mathcal{C}_1$ et celles de la condition $\mathcal{C}_2$.

État de l'art et limites des méthodes existantes

La plupart des méthodes d'analyse différentielle de données Hi-C sont basées sur des comparaisons indépendantes pour chaque paire d'intervalles génomiques (i, j). Pour (i, j) donnés, elles calculent une statistique de test puis une p-valeur rendant compte de la différence entre les valeurs des $(h_{ij}^t)_t$ entre les deux conditions, ces p-valeurs étant ensuite corrigées pour contrôler le FDR (False Discovery Rate). Parmi ces méthodes, on peut citer celle de [Lun and Smyth, 2015], reposant sur une modélisation des coefficients par une distribution binomiale négative, celle de [Stansfield et al., 2018], basée sur l'utilisation d'un Z-score, ou encore celle de [Djekidel et al., 2018] utilisant un processus spatial de Poisson.

Les limites de telles approches résident dans le fait qu'elles négligent des propriétés importantes des données comme leur structure hiérarchique ou encore la dépendance entre les coefficients. Cela peut engendrer des difficultés pour interpréter les résultats en termes de différences de structure de la chromatine.

1. Pour simplifier le propos, dans ce qui suit, les matrices $(\mathbf{H}^t)_t$ sont considérées ne correspondre qu'à un seul même chromosome.

2 Méthode de comparaison structurelle de données Hi-C

L'approche que nous proposons repose sur une modélisation de la structure hiérarchique de l'organisation spatiale du génome par une méthode de classification ascendante hiérarchique (CAH) avec contrainte de contiguïté entre les classes fusionnées. Les dendrogrammes issus de cette CAH sont alors directement comparés, ce qui permet d'obenir des différences dans l'organisation structurelle et non plus simplement des différences ponctuelles. De manière plus précise, la méthode se déroule en 4 étapes principales :

1. **Étape 1 : Modélisation de la structure hiérarchique sous forme de dendrogramme.** Une adaptation de la Classification Ascendante Hiérarchique avec lien de Ward est utilisée pour obtenir un dendrogramme pour chaque matrice $\mathbf{H}^t$. Cette version permet d'utiliser les entrées log-transformées de la matrice Hi-C comme des similarités sur lesquelles le critère de Ward est calculé et respecte, de plus, l'ordre des positions génomiques grâce à l'imposition d'une contrainte de contiguïté. Les propriétés pratiques et les justifications théoriques de cette méthode, appelée Classification Ascendante Hiérarchique sous Contrainte de Contiguité (CAHCC), ont été discutées dans [Ambroise et al., 2019, Randriamihamison et al., 2019] et elle est implémentée dans le package R **adjclust** (disponible sur le CRAN). À la fin de cette étape, un dendrogramme, $\mathbf{D}^t$, est associé à chaque matrice HiC, $\mathbf{H}^t$, comme illustré dans la Figure 1.

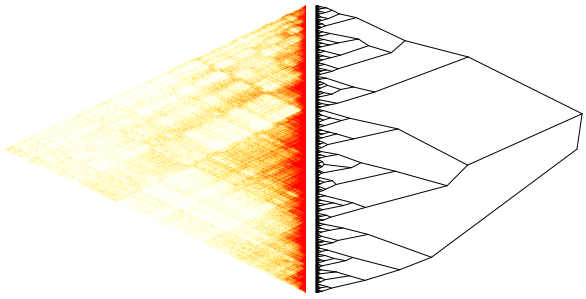

FIGURE 1 – Matrice Hi-C (gauche) et dendrogramme associé (droite).

2. **Étape 2 : Extraction de sous-arbres pour la comparaison.**
Pour déterminer des régions du génome présentant des différences significatives entre conditions, des sous-arbres, basés sur un ensemble commun de feuilles pour tous les dendrogrammes $\mathbf{D}^t$, sont extraits. De manière plus précise, on détermine L sous-ensembles d'intervalles génomiques contigus, $(I_l)_{1\leq l\leq L}$ (par exemple, l'ensemble des intervalles génomiques contigus de taille fixée) et on définit $\mathbf{D}_l^t$ le plus petit sous-arbre induit par $\mathbf{D}^t$ recouvrant I_l dont toutes les branches correspondantes à des feuilles extérieures à I_l sont élaguées (voir figure 2 pour un exemple).

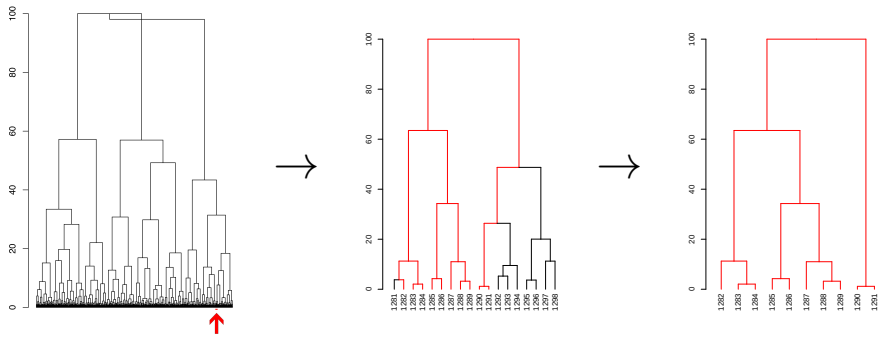

FIGURE 2 – Gauche : Dendrogramme $\mathbf{D}^t$. En rouge est indiquée la zone correspondant à l'intervalle I_l considéré. Centre : Plus petit sous-arbre induit par $\mathbf{D}^t$ recouvrant I_l. En rouge sont indiquées les branches correspondant à des feuilles dans I_l. Droite : Sous-arbre D_l^t obtenu après élagage des feuilles extérieures à I_l.

Des biais techniques variés, liés à l'acquisition des données, induisent des différences dans les hauteurs des dendrogrammes et des sous-arbres associés, qui sont préjudiciables à l'analyse. Ces biais techniques sont corrigés par une phase de normalisation qui consiste à aligner les hauteurs maximales et minimales de l'ensemble des arbres $(D_l^t)_t$ sur des valeurs communes, et uniformes pour tous l.

3. **Étape 3 : Comparaison d'arbres.**
   L'ensemble des paires d'arbres, D_l^t et $D_l^{t'}$ pour un intervalle donné sont alors comparées par le calcul d'une distance entre arbres appelée **weighted Path Difference metric (wPD)** et disponible dans le package R **phangorn**[2]. Cette distance est basée sur l'ensemble des longueurs des plus courts chemins entre deux feuilles le long du dendrogramme (distances cophénétiques pondérée, $\mathbf{d}^t \in \mathbb{R}^Q$ où $Q = n_l(n_l-1)/2$ avec n_l le cardinal de I_l) et correspond à la distance euclidienne des vecteurs de ces longueurs entre les deux dendrogrammes : $\mathbf{wPD}(\mathbf{D}_l^t, \mathbf{D}_l^{t'}) = \|\mathbf{d}_l^t - \mathbf{d}_l^{t'}\|$.
4. **Étape 4 : Obtention de garanties statistiques.**
   Par analogie avec les approches ANOVA, nous proposons de comparer les sous-arbres entre deux conditions par utilisation d'une analogie entre la distance **wPD** et la distance euclidenne et en calculant une statistique de test basée sur le ratio entre inertie inter-conditions et inertie intra-conditions :

$$F_l = \frac{(\mathcal{I}_{inter})/1}{(\mathcal{I}_1 + \mathcal{I}_2)/4}$$

où

2. Pour des questions de place, nous ne présentons pas les distances alternatives qui ont été évaluées.

— $\mathcal{I}_k$ correspond à l'inertie de la condition $\mathcal{C}_k$ calculée à partir des distances **wPD** entre sous-arbres de cette condition ;
— $\mathcal{I}_{\text{inter}}$ désigne l'inertie inter-conditions calculée à partir des distances **wPD** entre sous-arbres de deux conditions différentes.

Pour déterminer les intervalles génomiques significativement différents d'un point de vue structurel, on considère une approximation de la distribution de F_l sous l'hypothèse nulle, obtenue par mélange des conditions biologiques entre échantillons. Des travaux sont en cours pour déterminer une distribution théorique.

3 Exemples de résultats

L'approche proposée a été testée sur un ensemble de 6 matrices Hi-C de 2 conditions différentes (3 par condition) correspondant au chromosome 18 du génome humain pour deux types de cellules différentes. Ces données proviennent de [Sanborn et al., 2015] pour la condition 1 et [Darrow et al., 2016] pour la condition 2 (numéros d'accession GEO GSE63525 et GSE71831). Des exemples de sorties de la méthode sont données dans la figure 3 et illustrent la pertinence de l'approche sur données réelles.

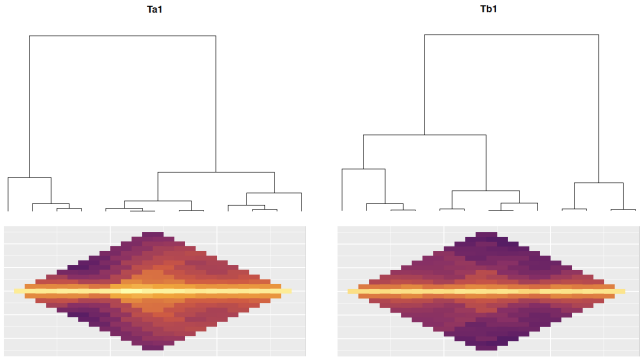

FIGURE 3 – Haut : Exemple de deux sous-arbres identifés associés, respectivement à la condition 1 (gauche) et à la condition 2 (droite). Bas : Sous-matrices Hi-C correspondantes (les niveaux de couleurs correspondent à l'intensité du contact, h_{ij}^t.

4 Conclusion

La méthode d'analyse différentielle présentée dans cette communication est basée sur l'idée de représenter la structure hiérarchique inhérente aux données Hi-C à l'aide d'un arbre (graphe binaire) construit à l'aide de la classification ascendante hiérarchique sous contrainte de contiguïté. On utilise ensuite une distance entre arbre et une statistique de

ratio des inerties inter et intra-conditions pour déterminer des intervalles génomiques où les structures sont significativement différentes.

Cette méthode permet de répondre aux limites des méthodes classiques d'analyse différentielle de données Hi-C qui ne prennent pas en compte la structure hiérarchique des données et les dépendances entre coefficients.

Remerciements

Ce travail 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 doctoral INRAE/Inria.

Références

[Ambroise et al., 2019] Ambroise, C., Dehman, A., Neuvial, P., Rigaill, G., and Vialaneix, N. (2019). Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics. *Algorithms for Molecular Biology*, 14 :22.

[Darrow et al., 2016] Darrow, E. M., Huntley, M. H., Dudchenko, O., Stamenova, E. K., Durand, N. C., Sun, Z., Huang, S.-C., Sanborn, A. L., Machol, I., Shamim, M., Seberg, A. P., Lander, E. S., Chadwick, B. P., and Aiden, E. L. (2016). Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proceedings of the National Academy of Sciences*, 113(31) :E4504–E4512.

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

[Djekidel et al., 2018] Djekidel, M. N., Chen, Y., and Zhang, M. Q. (2018). FIND : difFerential chromatin INteractions Detection using a spatial Poisson process. *Genome Research*, 28(3) :412–422.

[Lun and Smyth, 2015] Lun, A. T. and Smyth, G. K. (2015). diffHic : a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics*, 16 :258.

[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.

[Randriamihamison et al., 2019] Randriamihamison, N., Vialaneix, N., and Neuvial, P. (2019). Applicability and interpretability of hierarchical agglomerative clustering with or without contiguity constraints. Submitted for publication. Preprint arXiv 1909.10923.

[Sanborn et al., 2015] Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K. P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E. K., Lander, E. S., and Aiden, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, 112(47) :E6456–E6465.

[Stansfield et al., 2018] Stansfield, J. C., Cresswell, K. G., Vladimirov, V. I., and Dozmorov, M. G. (2018). HiCcompare : an R-package for joint normalization and comparison of Hi-C datasets. *BMC Bioinformatics*, 19 :279.