Comprendre l'organisation spatiale de l'ADN à l'aide de la statistique

Pierre Neuvial, Sylvain Foissac et Nathalie Vialaneix

Les processus moléculaires qui permettent de passer de la molécule d'ADN, support de l'information génétique, au fonctionnement de la cellule et à des phénomènes visibles à l'échelle de l'individu entier sont complexes et intrinsèquement multi-échelle. Ils sont pourtant rarement abordés comme tels par les méthodes d'analyse statistique qui essaient de les comprendre au travers des données collectées par les techniques modernes de biologie. Ces méthodes tendent à se focaliser sur l'analyse d'une partie du génome, fixée de manière plus ou moins arbitraire, plutôt que sur l'organisation de l'ensemble de celui-ci. Un des objectifs du projet SCALES était de développer des méthodes pour comprendre comment l'organisation tri-dimensionnelle de l'ADN permettait d'expliquer les phénomènes de développement de la fibre musculaire qui se produisent en fin de gestation chez le porc, et permettent d'expliquer son degré de maturité, qui conditionne sa survie, à la naissance. Cette question est critique pour une meilleure prise en charge de l'immaturité des porcelets, responsable d'un taux de décès de près de 20% dans les jours qui suivent la naissance chez certaines races commerciales. Cet article explique comment, dans le cadre de ce projet, nous avons modélisé la structure multi-échelle 3D de l'ADN grâce à des objets mathématiques adaptés, appelés dendrogrammes. Le projet SCALES a été développé entre l'Institut de Mathématiques de Toulouse et les équipes MIAT et GenPhySE d'INRAE.

Un fonctionnement cellulaire multi-échelle

L'image initiale d'un ADN linéaire partitionné en chromosomes, eux-mêmes segmentés en gènes espacés par de longues séquences non fonctionnelles, a été bouleversée par les avancées du domaine de la biologie moléculaire. Celles-ci ont en effet révélé l'importance de la structure tri-dimensionnelle de l'ADN, et de la manière dont cette molécule est compactée au sein de la cellule, mettant en contact des éléments distants sur le brin génomique d'un chromosome (figure 1). D'un point de vue fonctionnel, divers ensembles topologiques, définis de manière plus ou moins stricte, contrôlent une partie de la manière dont les gènes sont activés au sein de la cellule à divers niveaux d'échelle : régions compactées (faiblement transcrites) ou ouvertes (fortement transcrites), organisation en régions génomiques contiguës le long du chromosome favorisant les interactions physiques en son sein dans l'espace tri-dimensionnel de la cellule (« TADs » pour *Topologically Associating Domains*), boucles de régulation (pouvant mettre en contact physique un amplificateur avec un gène qu'il contribue à exprimer)…. La complexité de la structure tri-dimensionnelle du génome implique une organisation hiérarchique de ces éléments fonctionnels.

Ainsi se dessine une image complexe du fonctionnement génétique, reposant sur une organisation linéaire le long du génome, elle-même structurée en régions fonctionnelles cohérentes imbriquées dans l'espace tridimensionnel de la cellule.

Un besoin de méthodes statistiques pour explorer le vivant

Le développement des techniques de microscopie dans les années 1980-90 a permis de mieux comprendre la structure spatiale du génome. En particulier, l'hybridation *in situ* en fluorescence (FISH) permet de visualiser dans la cellule la distance physique entre deux régions distinctes du génome. La nécessité de marquer spécifiquement les régions ciblées et la précision de la mesure optique limitent le nombre de régions observables par expérience FISH. La méthode Hi-C (*HIgh-throughput Chromatin Conformation Capture)*, développée au début des années 2000, constitue aujourd'hui un outil de référence pour cartographier finement la structure spatiale du génome. Elle comprend une étape de séquençage à haut débit de l'ADN et mesure la fréquence des contacts physiques entre des millions de positions génomiques. Cette technologie a permis des avancées majeures dans la compréhension de l'organisation spatiale des chromosomes dans la cellule, et l'analyse des données qu'elle produit représente actuellement un domaine de recherche en pleine ébullition. À partir d'un échantillon biologique, une expérience Hi-C identifie donc des paires de positions génomiques en interaction. Ces données peuvent être représentées sous la forme d'une grande matrice (figure 2), tableau de valeurs numériques dont le coefficient (i,j) correspond au nombre d'interactions physiques mesurées entre les deux positions génomiques i et j.

Ces matrices font apparaître des structures à différentes échelles, qui peuvent correspondre aux différents niveaux d'organisation décrits dans la section précédente. L'identification visuelle de telles structures paraît possible pour une matrice donnée, mais elle est à la fois très fastidieuse, et subjective (elle dépend de l'expertise de l'observatrice). Une solution à ces problèmes consiste à définir objectivement, c'est-à-dire par des règles mathématiques ou informatiques, ce qui caractérise ces structures, et à implémenter ces règles dans des algorithmes permettant une analyse automatique de ces données.

Plusieurs méthodes statistiques ou bio-informatiques ont été proposées dans les dernières années pour identifier ces structures à partir d'une matrice Hi-C. Cependant, ces méthodes intègrent peu ou mal le caractère multi-échelle de l'organisation spatiale du génome. Dans le cadre du projet SCALES, nous avons introduit et implémenté des méthodes qui tiennent compte et même tirent parti du caractère hiérarchique de l'organisation spatiale du génome.

La CAH contrainte comme outil d'exploration multi-échelle

Un des outils mathématiques les plus simples pour représenter une structure hiérarchique est le dendrogramme (figure 3). Celui-ci se présente sous la forme d'un arbre inversé dont les feuilles (en bas de la

représentation graphique) correspondent aux éléments que le dendrogramme met en correspondance (ici, des positions génomiques organisées le long du chromosome). Des arêtes lient ces éléments en procédant par fusion successive (en montant dans l'arbre) de paires de feuilles, puis de paires de nœuds intermédiaires qui correspondent alors à des ensembles de plusieurs feuilles, jusqu'à la racine, qui est la fusion de toutes les feuilles (en haut de la représentation graphique). La hauteur de fusion est alors une mesure de la ressemblance des objets fusionnés : plus celle-ci est grande, plus les objets qu'elles fusionnent sont différents selon un certain critère (dans le cas des données Hi-C, ce critère peut s'interpréter comme : « sont éloignés dans l'espace tri-dimensionnel de la cellule »). Ce modèle présente plusieurs avantages. D'une part, sa visualisation graphique est facilement interprétable par l'œil humain, même non expert. D'autre part, il est bien adapté à des représentations de structures hiérarchiques diverses : outre les exemples déjà cités, il est couramment utilisé en phylogénie, par exemple, pour représenter l'évolution des espèces et leurs séparations, au cours du temps, en sous-espèces. Nous l'avons aussi utilisé pour définir des structures hiérarchiques en « haplotypes », c'est-à-dire, en régions chromosomiques qui ont une plus forte probabilité qu'attendue d'être transmises en « blocs » d'un parent à sa descendance. Aussi, pour ces raisons, il est naturel de s'appuyer sur celui-ci comme étape intermédiaire pour représenter les niveaux d'organisation complexes du génome permettant le développement de méthodes d'analyse multi-échelle, ne nécessitant plus de segmentation arbitraire.

Un algorithme statistique classique pour obtenir un dendrogramme à partir de données numériques est la « Classification Ascendante Hiérarchique (CAH) ». Dans le cadre du projet SCALES, nous avons adapté cette méthode aux données génomiques, en imposant une contrainte supplémentaire par rapport à l'algorithme initial : que seuls les éléments situés de manière contigüe sur le génome puissent être fusionnés à toute étape. En particulier, nous avons étudié quels cadres et conditions mathématiques permettent d'utiliser les mesures de proximités spatiales fournies par les données Hi-C comme données d'entrée de cette méthode de CAH contrainte.

Nous avons également proposé une approche permettant de surmonter la complexité algorithmique que représente l'utilisation de cette méthode avec un très grand nombre d'éléments à traiter. De manière plus précise, la complexité (c'est-à-dire le nombre d'opérations élémentaires à réaliser) pour mettre en œuvre une CAH contrainte croît avec le carré du nombre de positions considérées, ce qui peut être prohibitif pour l'analyse d'un génome entier. La méthode de CAH contrainte que nous avons développée tire parti du très grand nombre de valeurs nulles dans les données pour proposer une implémentation dont la complexité est proportionnelle au nombre de positions. En pratique, pour obtenir un dendrogramme à partir d'une matrice Hi-C d'un chromosome standard, notre approche permet de diviser le temps de calcul par un facteur supérieur à 10. Cette méthode est implémentée dans le logiciel libre *adjclust*.

Perspective : différences de structure multi-échelle entre conditions

La modélisation de la structure multi-échelle de la chromatine par les dendrogrammes ouvre de nouvelles perspectives pour l'analyse de ces données. En particulier, il est possible d'utiliser les dendrogrammes pour réaliser des tests de différence entre deux conditions (entre deux groupes de matrices Hi-C obtenues dans deux conditions différentes) directement au niveau de la structure et non plus de l'interaction entre une paire donnée de positions génomiques. Ainsi, il est possible d'identifier les régions chromosomiques dont la structure est significativement modifiée par la condition, avec des garanties mathématiques sur la fiabilité de ces découvertes. C'est l'objet d'un travail à venir ayant pour objectif la meilleure caractérisation des processus moléculaires intervenant en fin de gestation chez le cochon et des applications potentielles à la prévention de la surmortalité périnatale qui est constatée dans les élevages. D'ores et déjà, cette approche nous a permis de mettre en valeur des modifications structurelles importantes correspondant à près de 30% des régions chromosomiques testées. L'analyse fonctionnelle de ces régions est actuellement en cours.

Références

[1] Dossier hors série « Pour la Science », n°81, octobre 2013 « L'hérédité sans gène » https://www.pourlascience.fr/sd/genetique/dossier-pour-la-science-n0-81-683.php

[2] Le génome en quatre dimensions: Introduction historique à l'organisation du génome. Edith Heard, Cours du 9 mars 2020 dans le cadre de la chaire « Épigénétique et mémoire cellulaire » au collège de France. https://www.college-de-france.fr/site/edith-heard/course-2020-03-09-10h00.htm

[3] Randriamihamison N, Vialaneix N, Neuvial P: Applicability and Interpretability of Ward Hierarchical Agglomerative Clustering With or Without Contiguity Constraints, 2020 *Journal of Classification*. https://hal.archives-ouvertes.fr/hal-02294847

[4] Ambroise C, Dehman A, Neuvial P, Rigaill G, Vialaneix N: Adjacency-constrained hierarchical clustering of a band similarity matrix with application to Genomics. *Algorithms for molecular biology* 14, 2019 https://hal.archives-ouvertes.fr/hal-02006331. Logiciel adjclust: https://pneuvial.github.io/adjclust/

Glossaire

- **chromatine** : La chromatine est l’ensemble des éléments de tous les chromosomes, ADN compris, qui permettent la compaction de la double hélice d’ADN dans la cellule. En particulier, un certain nombre de molécules sont associées à l’ADN dans la chromatine pour maintenir sa structure et lui conférer des caractères fonctionnels.

- **chromosomes** : Un chromosome est une des molécules d’ADN (en double hélice) qui compose la chromatine. Généralement, les organismes vivants ont plusieurs : l’Humain a 23 paires de chromosomes, le cochon domestique 19.

- **classification ascendante hiérarchique (CAH)** : La Classification Ascendante Hiérarchique (CAH) est une méthode statistique permettant de regrouper des objets en groupes de plus en plus grands d’objets semblables. Initialement, chaque objet est dans son propre groupe, puis les deux groupes d’objets (les plus similaires) sont fusionnés, jusqu’à n’avoir plus qu’un seul groupe.

- **dendrogramme** : Le dendrogramme est une représentation graphique permettant de visualiser l’organisation de groupes d’objets structurés de manière hiérarchique. Il permet, en particulier, d’illustrer l’organisation des groupes obtenus lors d’une CAH.

- **Hi-C** : Le Hi-C (High Chromosome Contact) est une méthode expérimentale moléculaire, basée sur le séquençage haut débit, qui permet de mesurer la proximité spatiale, dans la cellule, entre toutes les paires d’éléments (ou positions le long du génome) de la chromatine.

- **FISH** : La technique FISH (hybridation in situ en fluorescence) est une méthode expérimentale moléculaire basée sur l’utilisation de sondes fluorescentes et de microscopes pour mesurer la proximité spatiale de deux éléments (ou positions le long du génome) donnés de la chromatine dans la cellule.

- **phylogénie** : étude des liens existant entre espèces apparentées. Grâce à elle, il est possible de retracer les principales étapes de l'évolution des organismes depuis un ancêtre commun et ainsi de classifier plus précisément les relations de parenté entre les êtres vivants.

- **séquençage** : Le séquençage est une technique permettant de lire l’ordre linéaire des composants d’une molécule. Cette technique permet donc de retrouver la séquence des nucléotides (A, C, T et G) d’un morceau d’ADN. Le séquençage « nouvelle génération » ou « haut débit » permet d’obtenir ces lectures pour l’intégralité de la séquence d’ADN d’un être vivant.

Figures

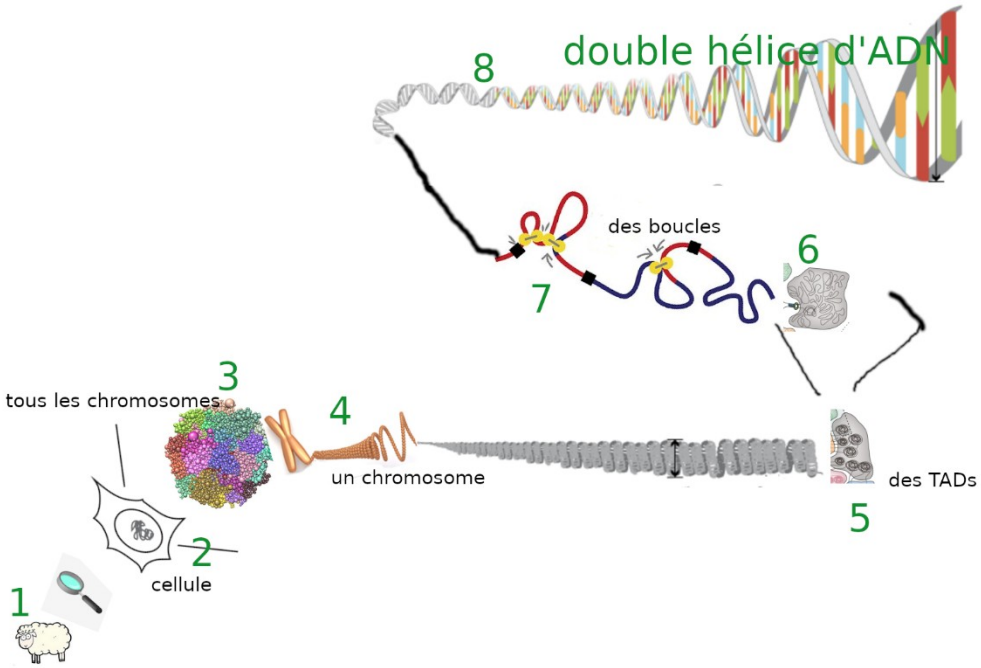

Figure 1 : Les organismes vivants (1) sont composés de cellules (2) qui contiennent, les chromosomes (4) porteurs de l'information génétique sur la double hélice d'ADN (8). Ces chromosomes, représentés par des couleurs distinctes, sont fortement compactés au sein de la cellule (3)). Lorsque l'on zoome sur un chromosome, d'autres niveaux de compaction apparaissent, appelés TADs (Topologically Associating Domains, (5)) qui correspondent à des zones où des morceaux de brins d'ADNs non contigus ont des proximités spatiales plus importantes qu'ailleurs (6). Cela permet au brin d'ADN de former des boucles (7), qui ont un rôle fonctionnel.

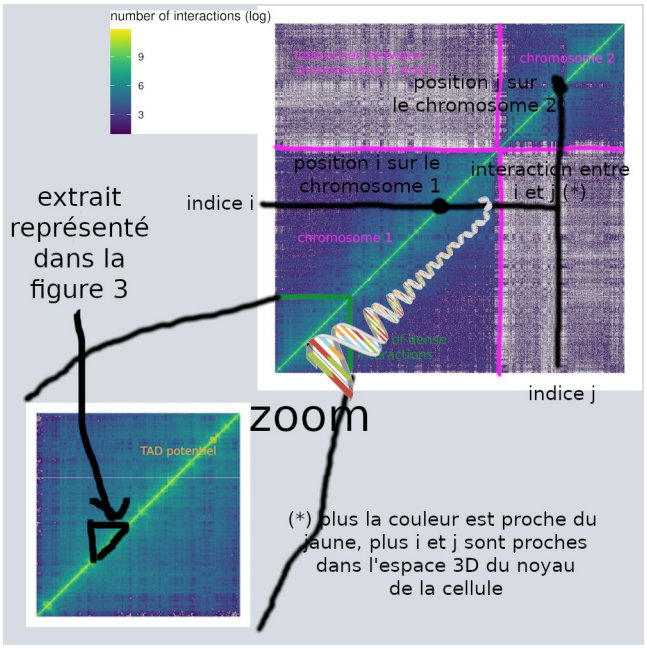

Figure 2 : Matrice de données HiC entre deux chromosomes (en haut à droite) et zoom sur une portion de celle-ci (en bas à gauche). La diagonale de la matrice correspond à la séquence d'ADN des deux chromosomes dépliés et mis côte à côte. Chaque position de la matrice (ligne *i* et colonne *j*) indique le niveau d'« interaction » (proximité spatiale) entre les parties correspondantes des chromosomes (ici la partie du chromosome 1 située à la *i*-ème position et la partie du chromosome 2 située à la *j*-ème position) : plus l'intensité de la couleur est importante et plus ces deux parties de chromosome sont proches dans la cellule. On voit sur cette figure que la proximité spatiale à l'intérieur des deux chromosomes (en haut à droite et en bas à gauche de la matrice globale) est plus forte que la proximité spatiale entre les deux chromosomes (en haut à gauche et en bas à droite de la matrice globale : couleurs moins intenses). Un zoom sur une partie du chromosome 1 met en valeur une zone où les couleurs sont particulièrement intenses comparées au reste de la matrice : cette zone pourrait correspondre à un TAD.

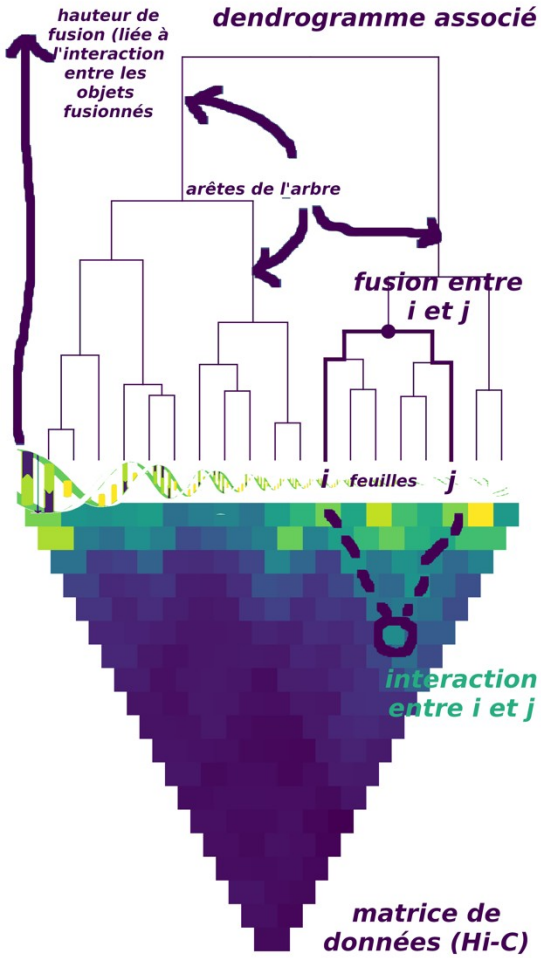

Figure 3 : En bas : Portion d'une demi matrice Hi-C (coupée le long de la diagonale, qui est représentée à l'horizontale ici). En haut : Dendrogramme (ou arbre) associé. Les paires de positions (parties) du chromosome sont les « feuilles » de l'arbre. Elles sont fusionnées successivement en montant dans l'arbre pour modéliser leur proximité dans la matrice : plus la fusion se produit à une hauteur élevée, plus les deux parties correspondantes du chromosome sont éloignées dans la cellule. Cette représentation simplifiée de la matrice permet de mieux identifier visuellement les structures très compactées comme les TADs (par exemple, l'ensemble des positions situées entre les feuilles *i* et *j* sur la figure).

Affiliations

Pierre Neuvial. Mathématiques appliquées – CNRS – Chargé de Recherche – Institut de Mathématiques de Toulouse/UMR 5219/Université de Toulouse/CNRS – UPS, F-31062 Toulouse Cedex 9 – pierre.neuvial@math.univ-toulouse.fr

Sylvain Foissac. Bioinformatique – INRAE – Chargé de Recherche – GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet-Tolosan – sylvain.foissac@inrae.fr

Nathalie Vialaneix. Statistique – INRAE – Directrice de Recherche - Université de Toulouse, INRAE, UR MIAT, F-31326, Castanet-Tolosan – nathalie.vialaneix@inrae.fr