Applications des ondelettes sur graphe en génomique

Rasha E. BOULOS, Nicolas TREMBLAY, Alain ARNEODO, Pierre BORGNAT, Benjamin AUDIT

Laboratoire de Physique de l'ENS-Lyon, CNRS UMR5672,
École Normale Supérieure de Lyon, 69007 Lyon, France
rasha.boulos@ens-lyon.fr, nicolas.tremblay@ens-lyon.fr,
alain.arneodo@ens-lyon.fr, pierre.borgnat@ens-lyon.fr, benjamin.audit@ens-lyon.fr,

Résumé – Les techniques expérimentales de capture de conformation des chromosomes permettent aujourd'hui de déterminer la matrice des fréquences d'interaction entre tous les sites du génome ! Ces nouvelles données constituent un challenge méthodologique pour la génomique qui doit maintenant en extraire de manière objective les *motifs structuraux* caractéristiques de l'organisation du génome. Nous appliquons la méthode rapide de détection multi-échelle de communautés basée sur les ondelettes sur graphes aux réseaux des interactions intra-chromosomiques dans deux lignées cellulaires humaines. Nous montrons que l'ensemble des communautés de structure forme une hiérarchie d'intervalles du génome, ainsi à toutes les échelles le repliement des chromosomes se fait principalement par des interactions entre sites voisins, plutôt que par la formation de liens entre sites distants. Cette méthodologie étant indépendante de l'assemblage exact du génome de référence, elle est applicable aux cellules cancéreuses dont le génome est souvent fortement réarrangé.

Abstract – Structural interaction frequency matrices between all genome loci is now experimentally achievable, thanks to high-throughput chromosome conformation capture technologies ! There is a new methodological challenge for genomics to objectively extract from these data the *structural motifs* characteristic of genome organisation. We deploy the fast multi-scale community mining algorithm based on spectral graph wavelets to characterise the networks of intra-chromosomal interactions in two human cell lines. We demonstrate that the set of structural community forms a hierarchy of genomic segments. Hence, at all scales, chromosome folding mainly involves interactions between neighbouring sites, rather than the formation of links between distant loci. The proposed methodology is independent of the precise assembly of the reference genome, it is thus directly applicable to cancer cells which often present a rearranged genome.

Introduction

Il est maintenant clairement reconnu que la dynamique et l'architecture 3D du génome chez les eucaryotes ont un rôle important pour la régulation des fonctions nucléaires telles que la réplication et la transcription [1–5]. A petite échelle (~ 200 nucléotides, nuc), la structure cristalline du premier niveau de compaction de l'ADN (formation d'un complexe nucléo-protéique appelé *nucléosome*) a été déterminée il y a près de 20 ans [6]. A l'échelle du noyau, l'imagerie de fluorescence a mis en évidence la structuration prépondérante du génome en *territoires chromosomiques*, reflétant une compartimentation sans mélange des chromosomes [2]. Par contre, aux échelles intermédiaires, notre connaissance sur l'organisation du polymère ADN reste partielle. Le développement récent de protocoles expérimentaux à haut débit pour la capture de la conformation des chromosomes (techniques Hi-C [7]) a ouvert de nouvelles perspectives dans l'étude de ces structures intermédiaires chez les eucaryotes supérieurs tels que les mammifères [7–14]. Les techniques Hi-C reposent sur le séquençage à haut-débit et permettent de mesurer quantitativement pour le génome complet les fréquences de co-localisation de toutes les paires de sites (la résolution des données les plus récentes [15, 16] est de l'ordre de quelques 10^4 nuc pour les génomes de mammifères dont la taille est de l'ordre de $3\ 10^9$ nuc). Les fréquences d'interactions inter-chromosomiques sont inférieures aux fréquences intra-chromosomiques, reflétant l'organisation du noyau en territoires chromosomiques [7]. Pour les données intra-chromosomiques, la fréquence d'interaction moyenne décroît avec la distance génomique, comme attendu pour un polymère [17]. La loi de décroissance permet de suivre les modifications de la structuration globale des chromosomes comme par exemple la condensation des chromosomes lors de l'entrée en métaphase [15]. Néanmoins, les données Hi-C font apparaître une compartimentation structurelle du génome à différentes échelles qui ne s'explique pas par des modèles de polymères homogènes simples [18]. L'analyse en composantes principales de la matrice des fréquences d'interactions révèle l'existence de deux catégories de régions interagissant préférentiellement avec elles mêmes, associées d'une part à des régions de type A riches en gènes actifs et à réplication précoce, et d'autre part à des régions de type B pauvres en gènes et à réplication tardive [7]. Projetée sur le génome, cette classification définit une partition du génome comme la succession de domaine de type A ou B de taille $\sim 10^7$ nuc. La représentation des fréquences d'interactions intra-chromosomiques sous forme matricielle fait apparaître un niveau plus fin de structuration, caractérisé par des blocs diagonaux de longueur 10^5-10^6 nuc : la fréquence d'interaction est forte entre régions d'un même bloc mais faible entre régions de blocs différents [12]

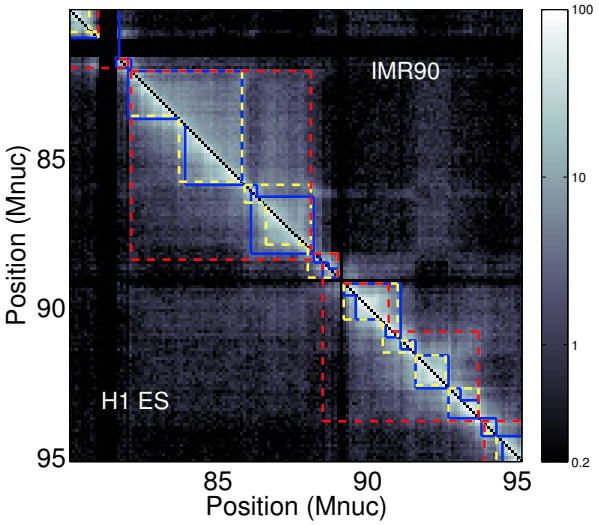

FIGURE 1 – Matrice des interactions le long d'un fragment de 15 Mnuc du chromosome 10 humain pour les lignées cellulaires H1 ES (resp. IMR90) au dessous (resp. au dessus) de la diagonale (données Hi-C de [12]). Les matrices ont été obtenues en comptant le nombre d'interactions observées entre fenêtres non-chevauchantes de 100 knuc. Les lignes bleus marquent les *domaines topologiques* identifiés dans [12]. Les pointillés jaunes (resp. rouge) marquent la partition en communautés obtenues à une petite (resp. grande) échelle. Les colonnes et lignes noires correspondent aux régions masquées.

(Fig. 1). Ces blocs sont la signature d'une compartimentation structurelles des chromosomes dont les liens avec l'organisation et la dynamique fonctionnelle du génome sont au coeur d'une intense activité de recherche [8, 12, 13, 15, 16, 19–21]. Afin de mener à bien ces recherches, diverses approches ont été développées permettant d'extraire objectivement cette compartimentation structurelle des données Hi-C [8, 12, 20–24]. Ces méthodes utilisent pour la plupart la connaissance a priori de l'agencement des sites le long du génome et supposent que les domaines structuraux recherchés sont des intervalles du génomes. Par exemple, certaines reposent sur la partition d'un signal 1D quantifiant la symétrie de la fréquence des interactions avec les régions en amont et en aval du site d'intérêt (index de directionalité) [12, 21]. D'autres utilisent des algorithmes de programmation dynamique qui utilise explicitement l'ordre génomique [23, 24]. Comme illustré sur la Fig. 1, la structuration du génome se fait sur une large gamme d'échelle [18] et est susceptible de faire intervenir des domaines emboîtés. Seule la méthode proposée par Filippova et al. [23] est construite afin de pouvoir identifier des domaines à diverses échelles d'observation.

L'objectif du travail présenté ici est de proposer une nouvelle méthode d'analyse des données Hi-C qui permette une identification multi-échelle de domaines structuraux et qui ne repose pas sur l'assemblage spécifique des génomes de référence. En effet, dû aux polymorphismes au sein d'une espèce ou au réarrangements chromosomiques caractéristiques des cancers par exemple [25], cet assemblage particulier ne représente pas nécessairement l'assemblage véritable pour la lignée cellulaire étudiée. Une matrice de fréquence d'interaction construite à partir de données Hi-C est positive et symétrique, elle peut donc être interprétée comme la matrice d'adjacence du réseau des interactions où les nœuds sont les différents fragments d'ADN et les arrêtes reflètent la fréquence d'interaction entre ces fragments, et être analysée par l'intermédiaire des outils de la théorie des graphes comme nous l'avons précédemment expérimenté [26, 27]. Cette représentation ne dépend de l'assemblage du génome qu'à l'échelle de la résolution utilisée pour construire la matrice Hi-C. Dans la théorie des graphes, un ensemble de nœuds appartient à une même *communauté* lorsqu'ils partagent plus de connections entre eux qu'avec le reste du graphe [28]. Ainsi, nous proposons de reformuler la question de recherche de domaines structuraux comme le problème de recherche de communautés dans le réseau des interactions. Cette approche a été expérimentée précédemment sur le réseau des interactions internes aux domaines chromosomiques de type A ou B [22]. Ici, afin de ne privilégier aucune échelle particulière dans l'analyse, nous effectuons le partitionnement multi-échelle en communautés des réseaux d'interaction intra-chromosomique complets en utilisant l'algorithme basé sur les ondelettes sur graphe que nous avons développé récemment [29].

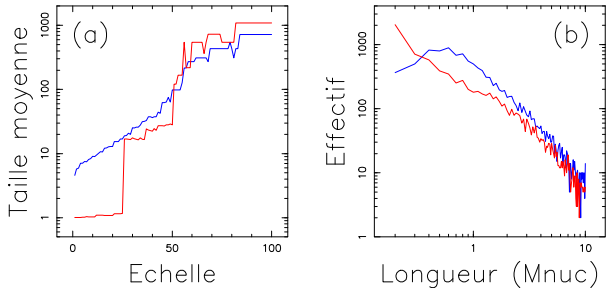

FIGURE 2 – (a) Taille moyenne des communautés pour le chromosome 1 en fonction de l'indice d'échelle dans les lignées cellulaires H1 ES (bleu) et IMR90 (rouge). (b) Histogrammes des longueurs génomiques des communautés-intervalles en représentation doublement logarithmique et calculés dans des bins de 100 knuc. Les couleurs ont la même signification qu'en (a).

Détection multi-échelle rapide de communautés

Notre méthode de recherche de communautés repose sur une construction précise d'ondelettes sur graphe afin d'introduire la notion d'échelle [30]. Les ondelettes à une échelle s caractérisent la structure locale du graphe autour de chaque nœud sur une "distance" contrôlée par s. A une échelle fixée, la similarité entre les voisinages de 2 nœuds est calculée comme la corrélation entre les ondelettes centrées sur chacun des 2 nœuds à cette échelle. On applique alors un algorithme de partitionnement hiérarchique avec la méthode de chaînage moyenné (ave-

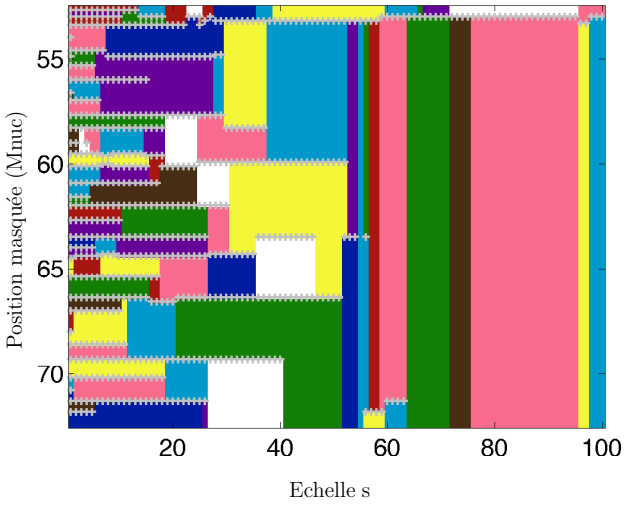

FIGURE 3 – Communautés multi-échelles le long d'un fragment de 20 Mnuc du chromosome 12 humain pour la lignée cellulaire H1 ES. À chaque échelle, les communautés-intervalles sont représentées par des segments colorés, bordés par des croix. Lorsque qu'une communauté est identique entre deux échelles successives, la même couleur de segment est utilisée.

rage linkage) [31, 32]. En coupant le dendrogramme suivant la méthode proposée dans [29], nous obtenons la partition à l'échelle d'analyse s.

Nous utilisons la mise en oeuvre rapide de cette procédure [29]. D'une part, la transformée en ondelettes sur graphe est calculée par l'algorithme rapide proposé par Hammond et al. [30]. D'autre part, pour calculer la matrice des corrélations entre ondelettes à une échelle donnée, au lieu de déterminer les ondelettes sur les N nœuds du graphe, ce qui demande N transformations en ondelettes de fonctions Dirac, celle-ci est approximée par la corrélation entre η ($\ll N$) transformées en ondelettes de vecteurs aléatoires gaussiens. Ainsi, cette procédure est applicable à des réseaux de grande taille ($\gtrsim$ 10 000 nœuds) [29], ce qui permet d'envisager son application sur des réseaux d'interactions intra-chromosomiques à très haute résolution ($\sim$ 10 knuc) [16].

Application au réseau des interactions intra-chromosomiques

Pour cette première application, nous avons utilisé les données Hi-C de deux lignées cellulaires humaines : H1 ES, une lignée embryonnaire souche et IMR90, une lignée fibroblaste de poumon de foetus [12]. Pour chacun des 22 autosomes, la matrice M des interactions entre fenêtres non-chevauchantes de 100 knuc a été déterminée (Fig. 1). Afin de s'affranchir des régions présentant un profil d'interactions aberrant, nous avons calculé la moyenne m et l'écart type σ du nombre de fenêtres n_i en interaction avec la fenêtre i ($M_{ij} > 0$) et éliminé (i) les régions à faible nombre d'interactants lorsque $n_i \leq \max(0, m - 2\sigma)$, il s'agit principalement des régions non séquencées du génome comme les centromères, et (ii) les régions à fort nombre d'interactants lorsque $n_i \geq \min(0.99N, m + 2\sigma)$ où N est le nombre de fenêtres le long du chromosome, il s'agit principalement de régions apparaissant interagir avec l'ensemble du chromosome dû à des réactions non-spéciques lors des étapes de biologie moléculaire. Ce masquage concerne $\sim 10\%$ de chaque chromosome, il reste entre 314 (chrom. 21) et 2179 (chrom. 1) (resp. 326 et 2172) fenêtres/nœuds pour les chromosomes de IMR90 (resp. H1 ES). Nous avons systématiquement appliqué la méthode de détection multi-échelle rapide de communauté aux 44 réseaux d'interactions intra-chromosomiques totalement connecté résultants, pour 100 valeurs d'échelles distribuées logarithmiquement dans la gamme d'échelle disponible (dépendante du réseau analysé, voir [29]) et $\eta = 200$. Il en résulte un jeu de 100 partitions du génome masqué pour chaque lignée cellulaire, constitué de 419 757 (resp. 89 770) communautés dans IMR90 (resp. H1 ES). Comme attendu, la taille moyenne des communautés croît avec l'échelle d'analyse (Fig. 2(a)). Pour la lignée H1 ES, cette croissance se fait de manière homogène suggérant qu'il n'y a pas de tailles caractéristiques dans la structure en communautés. Pour la lignée IMR90, on observe que sur une première gamme d'échelles la plupart des communautés se réduisent à un singleton (taille moyenne proche de 1), et que l'on passe brutalement à une situation où, pour le chromosome 1 par exemple, la taille moyenne des communautés est supérieure à 17, suggérant une structuration à petite échelle plus homogène (les plus petites communautés non triviales ont tendance à être plus grande) dans IMR90 que H1 ES. L'existence de ces singletons sur une large gamme d'échelles explique que l'on décompte au total plus de communautés dans IMR90 que dans H1 ES.

Les communautés de structure sont une hiérarchie d'intervalles du génome

Comme illustré sur la Fig. 1, les fréquences d'interactions en dehors des blocs diagonaux caractéristiques de la compartimentation structurelle précédemment décrite [12] ne sont pas a priori négligeables, ce qui suggère la possibilité qu'existe des domaines structuraux ne se réduisant pas à un intervalle du génome. Ainsi, pour chacune des communautés non triviales, nous avons calculé la proportion P_{int} couverte par le plus grand intervalle du génome masqué appartenant à la communauté : $P_{int} = 1$ signifie que la communauté est exactement un intervalle et $P_{int} = 0$ que la communauté ne contient pas de paire de fenêtres voisines sur le génome. En utilisant $P_{int} \geq 0.95$ comme critère qu'une communauté correspond principalement à un intervalle du génome, on observe pour les 2 lignées cellulaires que plus de 99% des communautés correspondent à des intervalles du génome. Cette propriété reste valable quelle que soit la taille des communautés y compris les plus grandes qui couvrent une proportion importante ($\gtrsim 30\%$) d'un chromosome ! On ne retient pour la suite que les communautés-intervalles ($P_{int} \geq 0.95$) que l'on réduit à

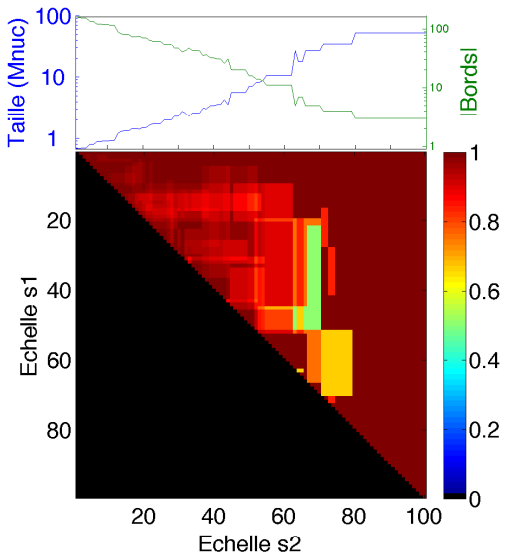

FIGURE 4 – Conservation des bords de communautés à travers les échelles pour le chromosome 12 humain dans la lignée cellulaire H1 ES. (Bas) La proportion de bords de communautés à l'échelle $s2 > s1$ qui sont aussi un bord à l'échelle $s1$ est représentée sous forme matricielle (code couleur à droite). (Haut) Taille moyenne des communautés et nombre de bords de communautés en fonction de l'échelle.

leur intervalle principal. Ceci nous permet d'adopter une représentation simple des communautés-intervalles obtenues à toutes les échelles, comme illustré dans la Fig. 3. Une propriété frappante qui émerge de cette représentation est l'organisation hiérarchique des communautés-intervalles. Au travers des échelles, les plus petites communautés fusionnent pour former les plus grandes, de telle sorte que toutes les frontières entre communautés-intervalles existent à la plus petite échelle et disparaissent à une certaine échelle résultant en une communauté plus grande. Quantitativement, la conservation des frontières des grandes vers les petites échelles est forte quelle que soit la paire d'échelles considérées (Fig. 4). Elle s'établit en moyenne à 80% et 73% pour les lignées H1 ES et IMR90, respectivement. La Fig. 3 illustre également la redondance de l'ensemble des communautés-intervalles obtenues d'une échelle à l'autre. On a donc construit une banque de données non redondante de communautés-intervalles de taille ≥ 2 de laquelle on a éliminé les domaines dont la longueur génomique est plus que doublée lorsque l'on réintègre les régions masquées du génome, il s'agit par exemple des communautés-intervalles enjambant les centromères. Au total la description multi-échelle en domaine structuraux se compose de 8827 communautés-intervalles pour la lignée IMR90 et de 12278 communautés-intervalles pour la lignée H1 ES. La distribution de la longueur génomique de ces communautés-intervalles est similaire entre IMR90 et H1 ES pour les grandes longueurs ($\gtrsim$ 2 Mnuc) (Fig. 2(b)). Pour les plus petites tailles, on observe dans IMR90 un excès de communautés-intervalles de 200 knuc mais un déficit pour celles de $\sim$ 500 knuc à $\sim$ 1.5 Mnuc, en cohérence avec la différence de structuration à petite échelle décrite précédemment (Fig. 2(a)).

Conclusion et perspectives

La méthode de détection de communautés multi-échelles basée sur les ondelettes sur graphes appliquée aux réseaux des interactions intra-chromosomiques permet de construire un ensemble de communautés structurales qui forme une hiérarchie d'intervalles du génome. Ainsi, à toutes les échelles, le repliement des chromosomes se fait principalement par des interactions entre sites voisins le long des chromosomes. Cette description de la structuration des chromosomes englobe la description en *domaines topologiques* initialement proposée par Dixon et al. [12], sans favoriser d'échelle d'observation a priori (Fig. 1). Elle donne un cadre pour la comparaison de la structuration du génome à différentes échelles d'une lignée cellulaire à l'autre, par exemple on observe une forte ($> 65\%$) correspondance entre les communautés-intervalles des lignées H1 ES et IMR90 de grande longueur ($\gtrsim$ 1 Mnuc), ce qui n'est pas le cas pour les communautés-intervalles plus petites. Ce nouveau cadre descriptif nous permet maintenant d'analyser les liens entre l'organisation structurelle et l'organisation fonctionnelle du génome en domaines de réplication et régions transcriptionnellement actives [8, 9, 16, 19, 20]. Cette nouvelle méthodologie est indépendante de l'assemblage exact du génome, évitant les artefacts dans l'analyse de données Hi-C provenant de cellules au génome réarrangé par rapport au génome de référence. Les communautés-intervalles obtenues respectant l'ordonnancement 1D des chromosomes, elles permettent en principe de faire l'assemblage de nouveaux génomes à partir de données Hi-C [33].

1. P. R. Cook, Science **284**, 1790 (1999).
2. T. Cremer et al., Nat. Rev. Genet. **2**, 292 (2001).
3. R. Berezney, Adv. Enzyme Regul. **42**, 39 (2002).
4. T. Misteli, Cell **128**, 787 (2007).
5. P. Fraser et al., Nature **447**, 413 (2007).
6. K. Luger, et al., Nature **389**, 251 (1997).
7. E. Lieberman-Aiden, et al., Science **326**, 289 (2009).
8. T. Sexton, et al., Cell **148**, 458 (2012).
9. C. Hou, et al., Mol. Cell **48**, 471 (2012).
10. Y. Zhang, et al., Cell **148**, 908 (2012).
11. R. Kalhor, et al., Nat. Biotechnol. **30**, 90 (2012).
12. J. R. Dixon, et al., Nature **485**, 376 (2012).
13. E. P. Nora, et al., Nature **485**, 381 (2012).
14. S. I. Takebayashi, et al., Proc. Natl. Acad. Sci. USA **109**, 12574 (2012).
15. N. Naumova, et al., Science **342**, 948 (2013).
16. S. S. P. Rao, et al., Cell **159**, 1665 (2014).
17. G. Fudenberg et al., Curr. Opin. Genet. Dev. **22**, 115 (2012).
18. J. H. Gibcus et al., Mol. Cell **49**, 773 (2013).
19. A. Baker, et al., PLoS Comput. Biol. **8**, e1002443 (2012).
20. F. Le Dily, et al., Genes Dev. **28**, 2151 (2014).
21. B. D. Pope, et al., Nature **515**, 402 (2014).
22. L. Liu, et al., BMC Genomics **13**, 164 (2012).
23. D. Filippova, et al., Algorithms Mol. Biol. **9**, 14 (2014).
24. C. Lévy-Leduc, et al., Bioinformatics **30**, i386 (2014).
25. S. Negrini, et al., Nat. Rev. Mol. Cell Biol. **11**, 220 (2010).
26. R. E. Boulos, et al., Phys. Rev. Lett. **111**, 118102 (2013).
27. R. E. Boulos, et al., New J. Phys. **16**, 115014 (2014).
28. S. Fortunato, Phys. Rep. **486**, 75 (2010).
29. N. Tremblay et al., EEE Trans. Signal Process. **62**, 5227 (2014).
30. D. K. Hammond, et al., Appl. Comput. Harmon. Anal. **30**, 129 (2011).
31. B. King, J. Amer. Statist. Assoc. **62**, 86 (1967).
32. A. Jain, et al., ACM Comput. Surv. (CSUR) **31**, 264 (1999).
33. J. N. Burton, et al., Genes, Genomes, Genetics (G3) **4**, 1339 (2014).