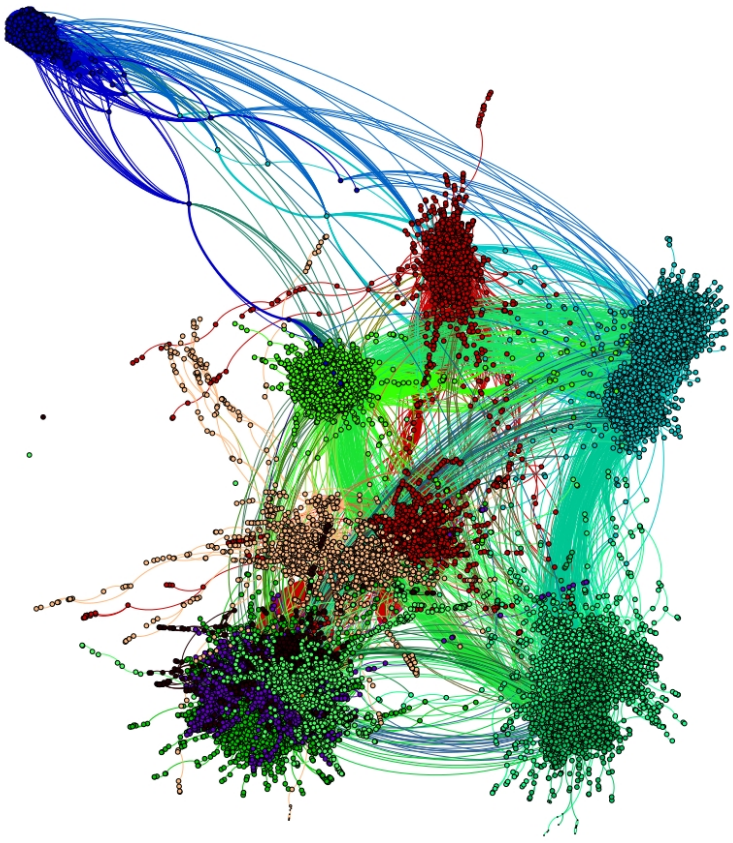
Hackathon ENGIE - Institut Pasteur

Où sont les boucles de chromosomes?

Équipe Régulation Spatiale des Génomes

Nous proposons dans ce document un sujet pour un hackathon collaboratif entre ENGIE et l'Institut Pasteur. Nous sommes une unité de recherche appelée *Régulation Spatiale des Génomes* au sein du département *Génomes et Génétique* et nous nous intéressons à la structure 3D des chromosomes chez divers organismes tels que les bactéries, les levures (du boulanger et surtout de bière) et l'humain. Notre équipe se compose à la fois de biologistes expérimentateurs et computationnels.



Contexte biologique et motivations

La structure 3D des objets biologiques a fréquemment un impact direct sur leur fonction: la structure tertiaire des protéines est intimement liée à leur activité, la structure des ARN messagers peuvent leur conférer des propriétés régulatrices. La structure 3D des génomes est également précisément organisée pour assurer certaines fonctions biologiques comme la réplication de l'information génétique et le maintien de son intégrité. La compréhension fine de la structure 3D des génomes suscite

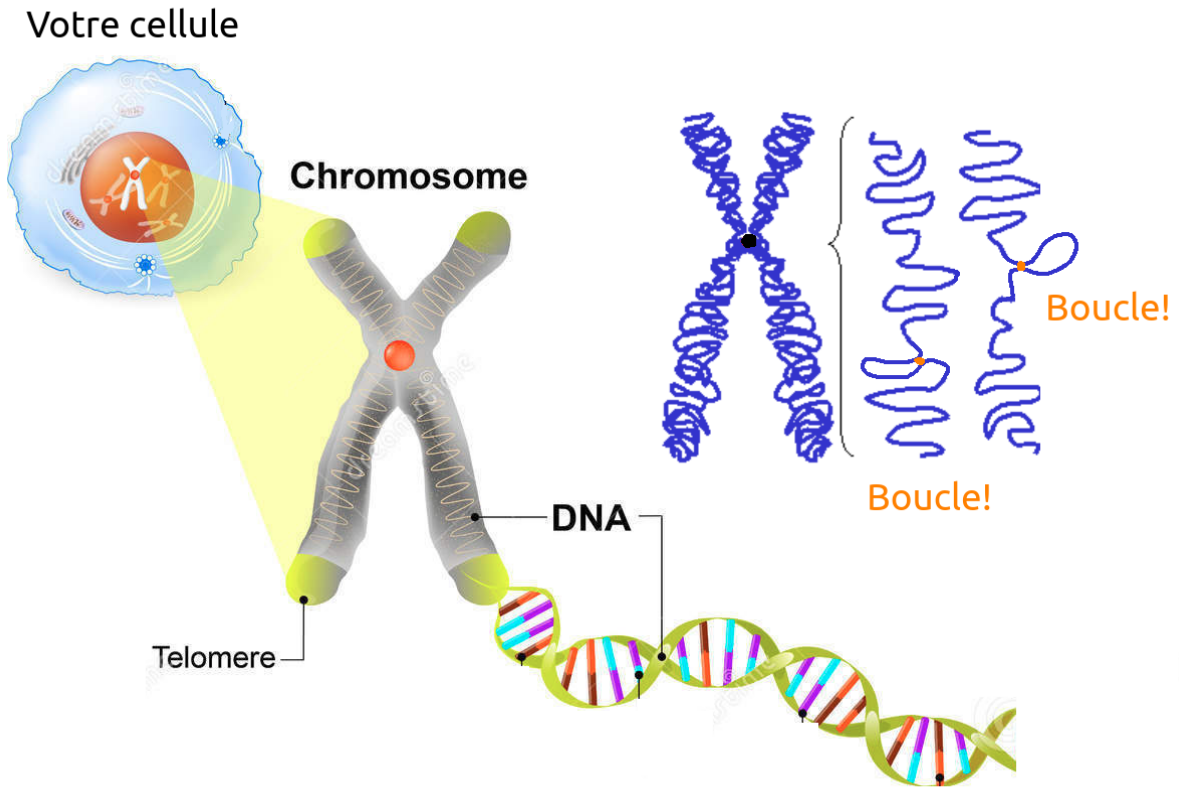


Illustration 1: Schéma montrant un exemple de cellule constituée d'un noyau qui contient les chromosomes. Ces chromosomes sont de très longues molécules d'ADN enroulées et compactées avec d'autres constituants. Certaines boucles de chromosome articulées par des protéines peuvent être très stables et détectables (points oranges).

actuellement un grand investissement dans la communauté internationale de chercheurs notamment à l'institut Pasteur. En effet, l'identification des règles dictant l'architecture des chromosomes touchent à des processus biologiques fondamentaux comme la réplication, la régulation de l'expression des gènes, le processus de réparation de l'ADN.
Ces nouvelles informations acquises au fil des années vont permettre une meilleure compréhension de certains processus et mécanismes biologiques entrant en jeu dans des pathologies humaines: développement de cellules tumorales, réplication de bactéries pathogènes, localisation de virus au sein d'un génome humain.

Ces recherches dans l'esprit pasteurien portent donc en elles de nouveaux espoirs de moyens thérapeutiques.

La technologie du 3C

Pour observer en détail l'architecture des chromosomes, des technologies dites de *contacts* (*contact genomics*) se sont développées depuis une vingtaine d'années,

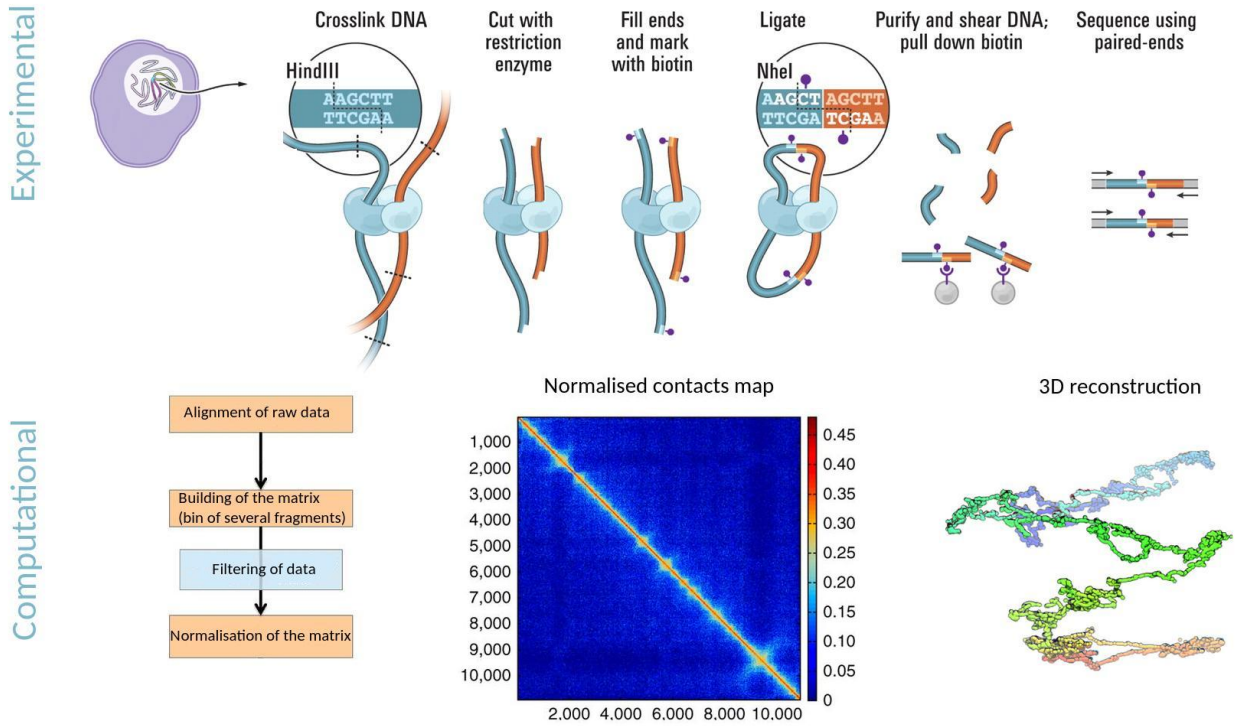


Illustration 2: Étapes expérimentales et computationnelles de la technique de 3C. La première étape consiste à piéger des segments de chromosomes proches à 3D par l'intermédiaire d'une réticulation au formaldéhyde. Des étapes de restriction, de ligation, déprotéination et de filtrage à l'aide de biotine permettent de garder les fragments d'ADN chimériques provenant de deux endroits différents du génome (Figure issue de Lieberman-Aiden et al. 2009 [2]). Un séquençage haut débit en paired-end (séquençage des 2 bouts) permet d'obtenir des millions d'événements de contacts au sein d'un génome. La partie bioinformatique comporte également plusieurs étapes. La première étant d'aligner les paires de lectures sur un génome de référence. Puis en comptant les séquences par groupes de fragments de restriction ou par kilobase, on construit une matrice de contacts d'une certaine résolution. Après plusieurs filtrages, une procédure de normalisation est appliquée (Cournac et al. 2012 [5]), celle-ci fait l'hypothèse que tous les points du génome doivent être détectés en moyenne avec la même intensité. Enfin, une structure 3D peut être reconstruite grâce à l'algorithme ShRec3D (Lesne et al. 2014 [6]). La matrice et la structure 3D correspond au chromosome 1 du génome humain (embryonic stem cells (hESCs), données issues de Dixon et al. 2012 [4]).

parallèlement aux techniques de microscopie. La principale d'entre elles, la technique dite 3C (Capture de Conformation de Chromosomes https://en.wikipedia.org/wiki/Chromosome_conformation_capture) conceptualisée dès le début des années 2000 quantifie les fréquences de contact physique entre différentes régions au sein d'un chromosome ou entre chromosomes [1]. Basée sur des procédés de biologie moléculaire standards, elle consiste à piéger ensemble les régions chromosomiques qui sont en contact au sein d'un noyau ou d'une cellule bactérienne. Plus précisément, les segments d'ADN chromosomique proches dans l'espace sont pontés entre eux grâce à un agent chimique (formaldéhyde), l'ADN présent dans ces complexes est ensuite digéré (le génome est fragmenté en milliers de segments) puis relié en condition diluée créant ainsi des molécules d'ADN chimériques contenant les loci initialement en contact proche. Couplée aux technologies de séquençage haut débit, cette technique permet de séquencer des millions d'événements de religation et de mesurer les fréquences de contact au sein d'un chromosome et d'un génome tout entier et d'en déduire sa conformation dans l'espace [2]. La technique a ainsi pu permettre l'observation de la structure des chromosomes à des échelles spatiales jusque-là inaccessibles par des techniques de microscopie directes.

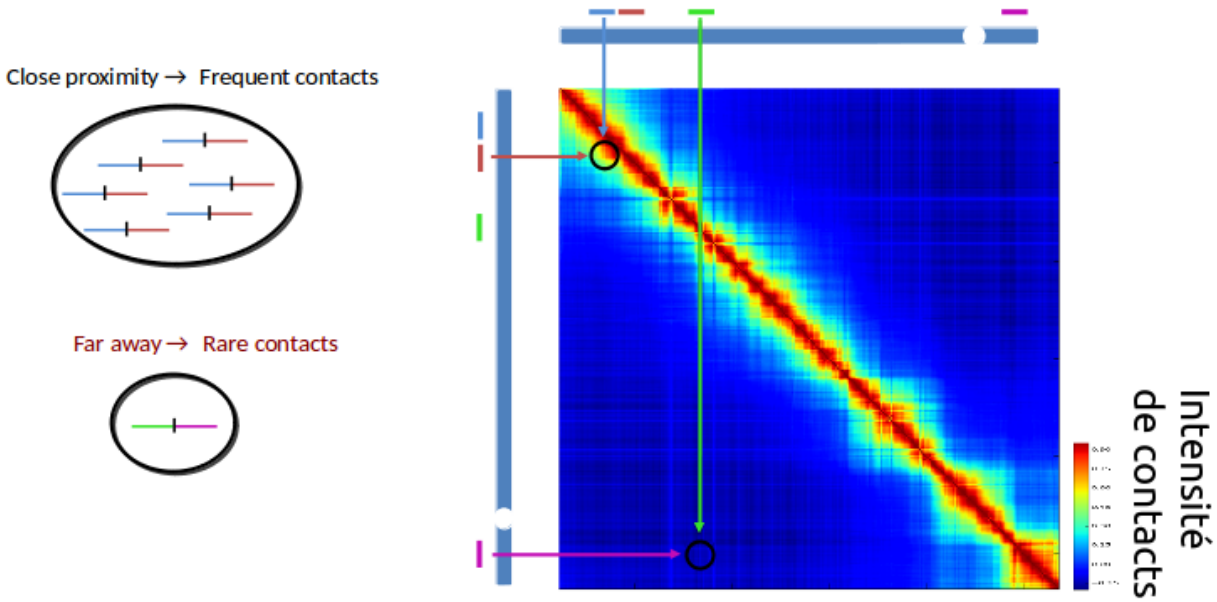


Illustration 3: Exemple de carte de contacts (pour un chromosome de bactérie). Le code couleurs de la carte de contacts indique l'intensité de la fréquence de contact entre 2 loci placés à des endroits différents sur le chromosome (faible : bleu à très fort contact, rouge).

Les boucles de chromosomes

Les boucles de chromosomes sont des contacts spécifiques et forts entre deux régions distinctes sur un chromosome. Le contact peut être articulé par certaines protéines. Notamment, tout récemment un modèle dit de *loop extrusion* a été proposé et suscite un très vif intérêt chez la communauté travaillant sur l'organisation des chromosomes.

Les condensines sont des complexes protéiniques qui ont la forme d'anneau qui pourraient accrocher ensemble deux points distincts d'un chromosome en avançant le long de l'ADN comme des moteurs moléculaires (figure 4, droite). Ce mécanisme permettrait de connecter des points a priori placés à de grandes distances sur le chromosome (> 100 000 paires de base). Une première observation in vitro par des techniques de microscopie de molécule unique de ce mécanisme a récemment été réalisée [3].
Un événement de type boucle de chromosome est visualisable sous la forme d'un spot de contact plus ou moins éloigné de la diagonale principale (figure 4, gauche).
C'est ce type d'événements que nous de souhaitons pouvoir détecter de façon automatique. Ces boucles ont un rôle majeur dans l'architecture générale des chromosomes.

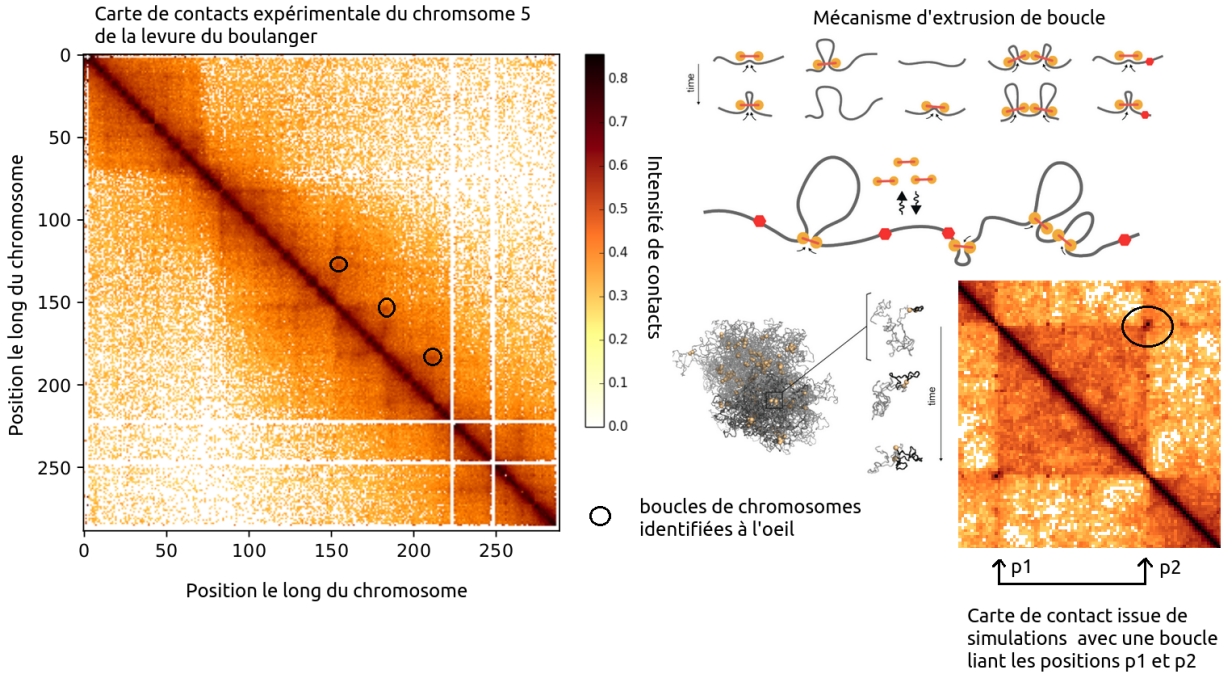


Illustration 4: Carte de contacts du chromosome 5 de la levure du boulanger. 3 exemples de boucles de chromosomes identifiées à l'œil sont montrés. Le défi proposé est de les identifier de façon algorithmique. Droite: schéma du mécanisme d'extrusion de boucle. Exemple de carte de contacts issue de simulations de polymères.

Le challenge pour ENGIE

A ce jour, il n'y a que deux algorithmes présents dans la littérature qui permettent une détection de ce type d'événements. Cependant, leur utilisation sur les données générées dans notre laboratoire s'est montrée très décevante : durée d'exécution longue et surtout beaucoup de faux positifs et de motifs non détectés.
Le défi que nous voulons lancer aux ingénieurs de ENGIE est l'implémentation d'une méthode de leur choix pour l'identification des positions de ces boucles à partir des cartes de contacts de chromosome.

Les fichiers d'entrées sont de simples matrices sous la forme de fichiers textes, par exemple : 1000 x 1000 nombres flottants compris entre 0 et 1, ou, pour une représentation sparse, une liste de trois colonnes représentant les coordonnées et la valeur du pixel correspondant. Les matrices sont toujours supposées symétriques.

Le défi est le suivant :

- Identifier les motifs de boucle de manière exhaustive sur des matrices ou sous-matrices de n'importe quelle taille (par exemple : une matrice 1000x1000 en entrée, en sortie une liste de coordonnée de type (35, 70), (125, 148), (729, 767) indiquant les positions de début et de fin de chaque motif de boucle).
- Optionnellement, un score de confiance pourrait être attribué à chaque boucle.
- Idéalement, cette méthode devrait être robuste à un niveau raisonnable de bruit. Elle pourrait être applicable à des cartes avec des données manquantes, en tenir compte et les abstraire du reste du signal.
- L'implémentation devrait être simple d'utilisation (formats d'entrée et de sortie sous forme de fichiers textes minimalistes). Une attention particulière sera attachée à la vitesse d'exécution (parallélisation sur plusieurs jeux de données envisageable) et la mémoire consommée sur des gros jeux de données.
-L'implémentation peut être soit sous la forme d'un exécutable statique et multi-plateforme, soit sous la forme d'une librairie avec des bindings dans un langage courant (idéalement Python).
Nous mettrons à disposition des participants des codes python (via une page Github https://github.com/axelcournac) pour visualiser les cartes et explorer les données.

Pistes envisagées

Nous pensons qu'une approche type *machine learning* pourrait être pertinente pour l'identification de ces événements . Notamment, les dernières applications du *deep learning* basées sur des réseaux de neurones convolutionnels se sont montrées particulièrement efficaces dans la reconnaissance d'images et de motifs; par exemple la distinction entre des images de chien et de chats. Par comparaison, les motifs que nous cherchons à détecter sont beaucoup plus moins complexes et devraient être facilement détectables de la même manière. Cependant, nous ne disposons pas pour le moment de suffisamment de matrices d'entrée générées biologiquement pour fournir une base de données conséquente pour entraîner un tel réseau. C'est pour cette raison que nous comptons générer des données simulées de boucles pour ainsi avoir un groupe de cartes pour une éventuelle phase d'apprentissage.
De plus, un groupe de cartes simulées avec des positions de boucles connues pourrait aussi permettre d'évaluer la précision (taux de faux positifs, etc) des méthodes proposées. Si plusieurs équipes sont en compétition, cela permettrait de donner un score à chacune et un classement pour insuffler un esprit positif de compétition important dans un hackathon.

Enfin, si méthode se montre efficace, elle sera utilisée en routine dans notre laboratoire. Si elle se révélait particulièrement compétitive, elle pourrait faire l'objet d'une publication dans une revue scientifique où les ingénieurs d'ENGIE seraient associés.

Références

[1] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, "Capturing chromosome conformation," *Science*, vol. 295, no. 5558, pp. 1306–1311, Feb. 2002.

[2] E. Lieberman-Aiden *et al.*, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science*, vol. 326, no. 5950, pp. 289–293, Oct. 2009.

[3] M. Ganji *et al.*, "Real-time imaging of DNA loop extrusion by condensin," *Science*, vol. 360, no. 6384, pp. 102–105, Apr. 2018.

[4] J. R. Dixon *et al.*, "Topological domains in mammalian genomes identified by analysis of chromatin interactions.," *Nature*, vol. 485, no. 7398, pp. 376–380, May 2012.

[5] A. Cournac, H. Marie-Nelly, M. Marbouty, R. Koszul, and J. Mozziconacci, "Normalization of a chromosomal contact map.," *BMC Genomics*, vol. 13, p. 436, 2012.

[6] A. Lesne, J. Riposo, P. Roger, A. Cournac, and J. Mozziconacci, "3D genome reconstruction from chromosomal contacts," *Nat. Methods*, vol. 11, no. 11, pp. 1141–1143, Sep. 2014.