Analyse et modélisation du repliement spatial de l'épigénome

Noëlle Haddad

Numéro National de Thèse : 2016LYSEN042

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON
opérée par
l'Ecole Normale Supérieure de Lyon

Ecole Doctorale N° 52
Physique et Astrophysique de Lyon (PHAST)

Spécialité de doctorat :
Physique

Soutenue publiquement le 17/11/2016, par :
Noelle Haddad

Analyse et modélisation du repliement spatial de l'épigénome

Devant le jury composé de :

Victor, Jean Marc	DR, CNRS	Rapporteur
Rosa, Angelo	CR, SISSA	Rapporteur
Bystricky, Kerstin	Professeure, Université Toulouse III	Examinatrice
Everaers, Ralf	Professeur, ENS Lyon	Examinateur
Cavalli, Giacomo	DR, CNRS	Examinateur
Mozziconacci, Julien	Maître de conférences, UPMC	Examinateur
Vaillant, Cédric	CR, CNRS	Directeur de thèse
Jost, Daniel	CR, CNRS	Co-encadrant de thèse

RÉSUMÉ

L'ADN chromosomique des cellules eucaryotes est fortement condensé au sein d'un complexe nucléoprotéïque, la chromatine. Aussi bien l'organisation spatiale que la composition biochimique (état "épigénomique") de la chromatine jouent un rôle fondamental dans la régulation des gènes. Grâce aux récents développements des techniques de séquençage à haut-débit, il est possible de déterminer l'état épigénomique local de la chromatine ainsi que la probabilité de contact entre deux sites génomiques (technique dite de "Hi-C"). Ces deux techniques ont permis de mettre en évidence l'existence de domaines d'interaction dont les positions corrèlent fortement avec la segmentation épigénomique de la chromatine. Cependant, les mécanismes responsables de ce couplage sont encore mal compris. L'objectif de cette thèse est de bâtir des modèles physiques permettant de valider l'hypothèse que l'épigénome est un acteur majeur dans le repliement 3D de la chromatine. Pour cela, nous avons tout d'abord développé "IC-Finder", un algorithme permettant de segmenter les cartes Hi-C en domaines d'interaction. Nous avons alors pu quantifier précisément l'association entre épigénome et organisation de la chromatine. Les corrélations trouvées justifient l'idée de modéliser la chromatine par un copolymère par bloc dont les monomères ont chacun un état épigénomique. Dans ce cadre, nous avons développé une méthode d'inférence des potentiels d'interaction entre sites génomiques à partir des cartes Hi-C expérimentales. Ce travail permettra à plus long terme de prévoir l'organisation de la chromatine sous différentes conditions, ce qui permettra d'étudier en particulier les changements de structure résultant de l'altération de l'épigénome.

Mots-clés : Chromatine, Organisation nucléaire, Compartimentation 3D, Hi-C, Physique des polymères, Copolymère par bloc, Inférence.

ABSTRACT

DNA of eukaryotes is highly condensed in a nucleoprotein complex called chromatin. Both the spatial organization and the biochemical composition ("epigenomic" state) of the chromatin are fundamental for gene regulation. Remarkably, recent studies indicate that1D epigenomic domains tend to fold into 3D topologically associated domains (TADs) forming specialized nuclear chromatin compartments. In this thesis, we address the question of the coupling between chromatin folding and epigenome. We first built a software called IC-finder to segment HiC maps into interacting domains. We next used it to quantify correlations between the TADs and epigenomic partitions of the genome. This led us to develop a physical model of the chromatin with the working hypothesis that chromatin organization is driven by physical interactions between epigenomic loci. We modeled chromatin as a block copolymer where each block corresponds to an epigenomic domain. With this framework, we developed a method to infer interaction parameters between chromatin loci from experimental Hi-C map. An outcome of such inference process would be a powerful tool to predict chromatin organization in various conditions, allowing investigating in silico changes in TAD formations and long-range contacts when altering the epigenome.

Keywords : Chromatin, Nuclear organization, 3D compartimentalization, Hi-C, Polymer physics, Block copolymer, Inference.

REMERCIEMENTS

Parmi les personnes qui m'ont accompagnée au cours de cette thèse, les principales sont bien sûr mes directeur et co-encadrant de thèse, Cédric Vaillant et Daniel Jost. Apprendre aux côtés de deux personnes aussi compétentes et expérimentées fut une grande chance et un réel plaisir. Je vous remercie sincèrement pour votre accueil, votre patience et toute votre implication dans mon travail de thèse. Tout au long de ces trois années, vous m'avez guidée, encouragée et conseillée avec bienveillance. Vous m'avez beaucoup apporté et je vous en suis extrêmement reconnaissante. On m'avait dit que l'écriture de la thèse est souvent un moment difficile à passer mais grâce à votre science, votre disponibilité et votre bonne humeur cette période redoutée de l'écriture restera pour moi un des meilleurs souvenirs. Je vous souhaite le meilleur dans la suite de vos projets et bien sûr aussi le meilleur dans vos vies personnelles !

Je remercie aussi les membres du jury : merci à Jean Marc Victor et Angelo Rosa d'avoir accepté le rôle de rapporteur, merci à Ralf Everaers, Giacomo Cavalli et Julien Mozziconacci pour tous les échanges que nous avons pu avoir au cours de ces trois années, merci aussi à Kerstin Bystricky que j'ai pu écouter lors de différents séminaires d'avoir accepté d'être membre du jury.

Je remercie ensuite Thierry Dauxois et Christophe Dujardin, respectivement les directeurs du laboratoire de physique de l'Ecole Normale Supérieure de Lyon et de l'école doctorale PHAST pour m'avoir acceptée en tant que doctorante au sein de leur laboratoire et école. Je remercie aussi Jean-Christophe Géminard pour sa réactivité et sa gestion de l'équipe 1.

Je souhaite vivement remercier pour leur gentillesse, leur disponibilité et pour toutes les discussions fructueuses que nous avons eu, Max Kolb qui m'a initiée à la dynamique sur réseau, Ralf Everaers (tout les chapitres de cette thèse comprennent des idées de Ralf !) et Pascal Carrivain qui m'a présenté ses travaux de dynamique moléculaire et m'a permis de les décrire dans le chapitre 4, merci beaucoup Pascal d'avoir pris le temps de me générer toutes ces figures et d'avoir relu à plusieurs reprises cette partie de la thèse.

Au cours de ces trois années, l'enseignement a occupé une bonne partie de mon temps.

J'ai appris beaucoup de choses et sans surprise j'ai énormément apprécié cette activité. Je remercie en particulier Cendrine Moskalenko, Stéphane Roux et Abdelaziz Ait Amer avec qui j'ai vraiment aimé travailler.

Plus généralement merci à toutes les personnes avec qui j'ai partagé de bons moments à l'ENS. Merci à vous pour tout les gestes du quotidien qui en font un lieu de travail très agréable. Merci à Cendrine Moskalenko(pour qui j'ai eu un gros coup de coeur dès mon arrivée à l'ENS :)), Pascal Carrivain, Laurence Lemelle, Rolf Walder, Doris Folini, Tina, Samira Riahi, Christophe Place, Antoine Naert, Zakia Mokhtari, et tout les autres ! Merci aussi aux exceptionnelles secrétaires, Laurence Mauduit, Sylvie Flores et Myriam Friat qui se sont toutes occupées avec une grande efficacité de toutes mes missions et dossiers. Je n'oublie pas non plus les responsables informatiques du CBP et du PSMN, Cerasela, Emmanuel et Loïs pour leur bonne humeur, leur disponibilité et leur efficacité.

En dehors du travail, de nombreuses personnes m'ont accompagnée ces trois dernières années. Merci à tous ceux avec qui j'ai partagé de bons moments de sport : Yacine, Guillaume Rachid, Abdel, aussi Yasser et Chaima pour la course à pied ; Eric, Ambra, Guillaume, Anthony, Géraldine et tout les autres membres de la Compagnie d'Armes de Lyon pour l'escrime.

Je tiens aussi à exprimer ma gratitude envers mes frères et soeurs pour leur soutien indéfectible et leur amour inconditionnel : Merci Yacine de toujours répondre « oui » à la question « T'es où ? Tu fais quoi ? Tu peux me rendre un service ? », Merci Rachid pour m'avoir appris à faire les additions, Merci Ratiba pour toute ta bienveillance, Merci Abdelghani avec toi rien ne semble impossible et Merci Fatma pour ta joie de vivre et ton envie de toujours chercher à tout faciliter. Je remercie bien sûr aussi leurs époux, épouse, enfants, la « grande famille » comme dirait certains. Yasser, Sawsen, Thais, Tasnime, Anas, Ilef, Wael & Hajer, vous êtes tous la fraîcheur de mes yeux, des étincelles de joie. Que la vie vous comble de bonheur. Merci à Chaima, Aurore, Mohamed, Julie et Stéphanie pour m'avoir soutenue et pour vous être toujours réjouis de mes réussites. Un grand merci aussi à mon mari Guillaume pour ses mises à jour casi quotidiennes pendant ces trois années ! Merci aussi à Fanny, Mireille, Robert, sa maman, et aussi Jean, Annie, Monique et Nicole qui m'ont accueillie très chaleureusement et qui m'ont encouragée au cours de cette thèse.

Mes plus profonds remerciements vont à mes parents. Tout au long de ma vie, ils m'ont toujours soutenu, encouragé, aidé. Ils m'ont toujours conseillé dans mes choix pour que j'en arrive jusque là. Ils ont su me donner toutes les chances pour réussir. Ma réussite est la leur. Aujourd'hui, maman et papa, vous pouvez trouver, dans la réalisation de ce travail, un aboutissement de vos efforts ainsi que l'expression de ma plus affectueuse gratitude.

Lyon, le 31 octobre 2016

À Rati, à mon père et à ma mère,

Table des matières

LISTE DES SYMBOLES

ADN	Acide désoxyribonucléique
bp	Paire de base (acide nucléique)
CDF	Fonction de répartition (en anglais, Cumulative distribution function)
ChIP	Immunoprécipitation de chromatine
DI	Indice de directionnalité
ENCODE	Encyclopédie d'élements ADN (ENCyclopedia Of DNA Elements)
FDR	Taux de faux positif (False Discovery Rate)
FISH	Hybridation in-situ fluorescente
HMM	Modèle de Markov caché
IBI	Inversion de Boltzmann itérative
ICE	Iterative correction and eigenvector decomposition (technique de normalisation des cartes Hi-C)
kb	Kilobase (= 1000 bp)
m.s.d.	mean squared displacement
Mb	Mégabase (= 1000 kb)
MPS	Microphase separation
RI	Région intermédiaire
TAD	Topologically Associating Domain
TPR	Taux de vrai positif (True Positive Rate)

CHAPITRE 1

INTRODUCTION

1.1 Question générale : chromatine et régulation des gènes

L'ADN chromosomique des cellules eucaryotes est fortement condensé au sein d'un complexe nucléoprotéïque, la chromatine. Le premier niveau de compaction, le nucléosome, correspond à un enroulement de 146 paires de bases autour d'un octamère d'histone. L'arrangement linéaire de ces nucléosomes le long de la chaine ADN forme le chapelet nucléosomal ou « fibre de 10 nm ». Cette fibre, notamment grâce à la fixation d'histones de liaison ou autre proteines architecturales pourrait se compacter en une fibre plus compacte , dite « fibre de 30 nm » et/ou s'organiser, à plus grande échelle, soit dans des phases polymériques de type globules ou brosses. Par ailleurs que ce soit au niveau de l'ADN, avec la méthylation, ou au niveau des histones, avec les modifications covalentes des queues ou l'insertion de variants, la chromatine se caractérise localement par une signature biochimique. Or, ces marqueurs biochimiques sont impliqués soit directement dans la structuration de la fibre (par exemple en modulant la stabilité des nucléosomes, ou l'interaction entre nucléosomes..) soit dans le recrutement de facteurs auxiliaires régulateurs de la chromatine comme les facteurs de remodellage. Il apparaît ainsi, qu'aussi bien l'organisation spatiale que la composition biochimique de la chromatine, en modulant l'affinité des différents complexes enzymatiques à leurs sites nucléiques peuvent jouer un rôle fondamental dans la régulation du programme transcriptionnel (quels gènes actifs et quand ?) des cellules : à temps court, dans le cas de la réponse au stress et à temps long, dans le cas de la spécification et maintenance (heritabilité) d'un type cellulaire au cours du développement ou lors de maladie comme le cancer.

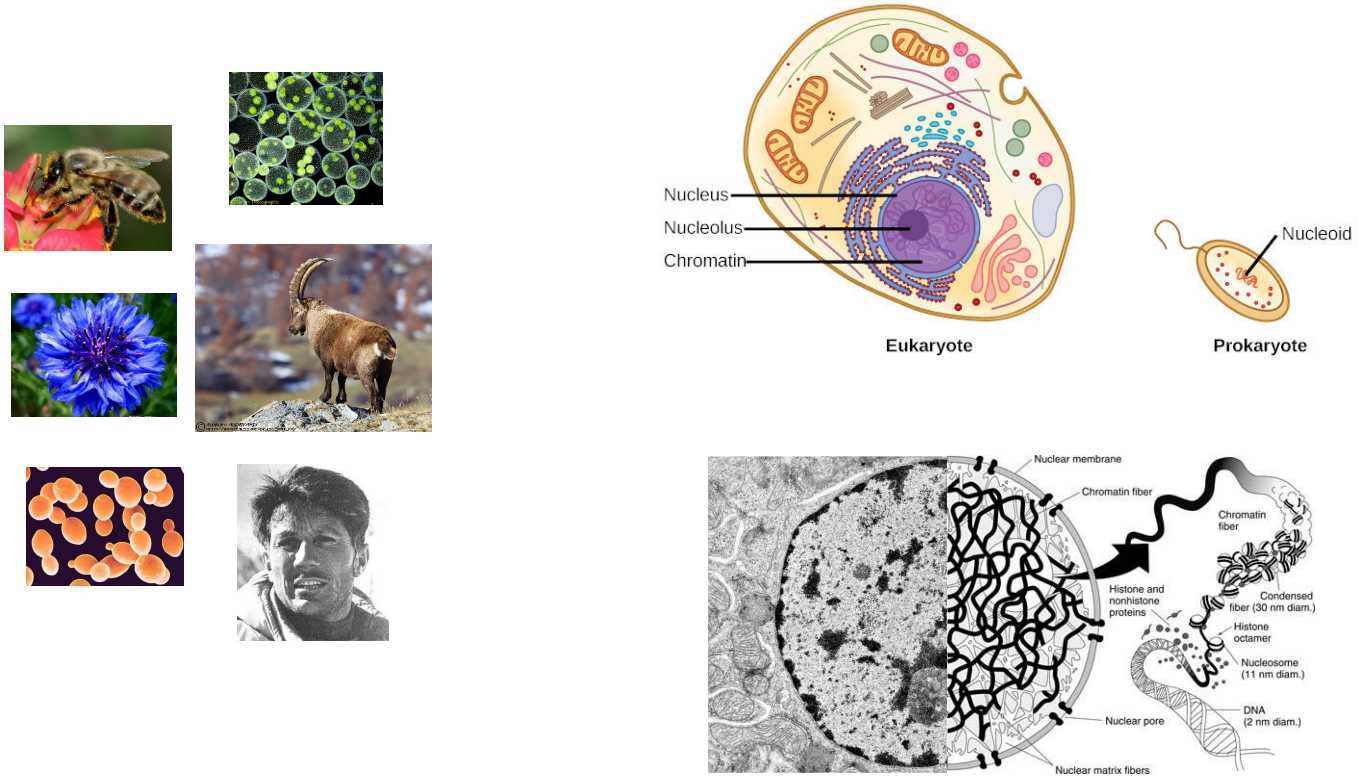

Figure 1.1.1 – **Eucaryotes : régulation des gène et chromatine.**

L'objectif du groupe dans lequel j'ai effectué ma thèse est d'identifier et de modéliser par des approches de physique statistique, de l'équilibre et du hors-équilibre, les mécanismes de régulation « chromatinienne » de la transcription : comment la structure et la dynamique de la chromatine à toute échelle est-elle impliquée dans la dynamique d'activation/répression des gènes ? Quels sont les mécanismes épigénétiques chromatiniens à l'origine de la plasticité cellulaire, c'est-à-dire les mécanismes qui puissent à partir d'un même génotype conduire à différents phénotypes stables et héritables ? (Fig. 1.1.1).

1.2 États chromatiniens : épigénome

L'expression d'un gène est caractérisée par quatre étapes principales : (i) l'initiation avec l'assemblage du complexe transcriptionnel (la polymerase + cofacteurs) au niveau du promoteur (ii) l'élongation avec la production de l'ARN (iii) la terminaison, la maturation et le transport de l'ARN messager du nucléoplasme dans le cytoplasme et (iv) la traduction de cet ARN en protéines par les ribosomes. L'étape (i) est assurée par la fixation de facteurs de transcription à leurs sites génomiques de reconnaissance (sites régulateurs) au niveau du pro-

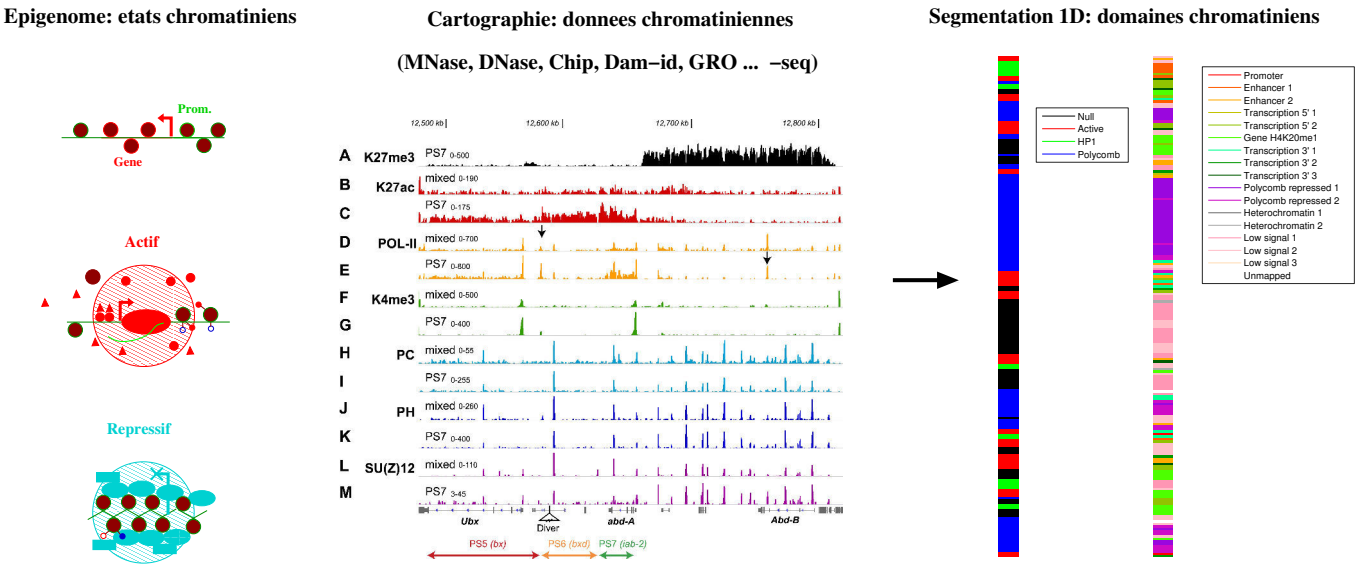

FIGURE 1.3.1 – **compartimentation 1D de l'épigénome.** Localement, la chromatine se caractérise par différents types d'assemblages macro-moléculaires définissant différents types d'états chromatiniens. L'épigénome, pour un type cellulaire donné, est la distribution de ces différents types d'états chromatiniens le long du génome. Le développement de techniques de cartographie à haut débit de la distribution des marqueurs biochimiques (modification ADN, queues d'histones, protéines architecturales, enzymes, polymerases etc...) et leur analyse statistique permettent d'établir des segmentations en domaines épigénomiques. Selon la finesse avec laquelle on définit un état chromatinien, il est possible d'avoir une segmentation avec soit 5 (4) ou 17 types chromatiniens pour un même épigénome.

moteur, qui ensuite recrutent d'autres cofacteurs pour *in fine* favoriser la formation stable du complexe pré-transcriptionnel. Cette étape est donc dépendante du niveau d'accessibilité des sites régulateurs. Or, la chromatine joue un rôle crucial dans le contrôle de cette accessibilité. Au delà des nucléosomes, dont les positions le long du génome constituent un premier niveau de régulation de cette accessibilié, il existe tout un ensemble de marqueurs biochimiques et de protéines architecturales qui vont être déposés et assemblés au niveau de l'ADN et des nucléosomes (par des enzymes dédiées) pour produire des assemblages macro-moléculaires définissant localement des états chromatiniens qui vont être plus ou moins permissif à la transcription : pour simplifier, il y a des états actifs et des états répressifs. Ces états forment le long du génome ce qu'on appelle l'épigénome, à savoir une information qui se place au dessus de l'information génomique et qui en quelque sorte prescrit la part accessible et donc traductible du génome.

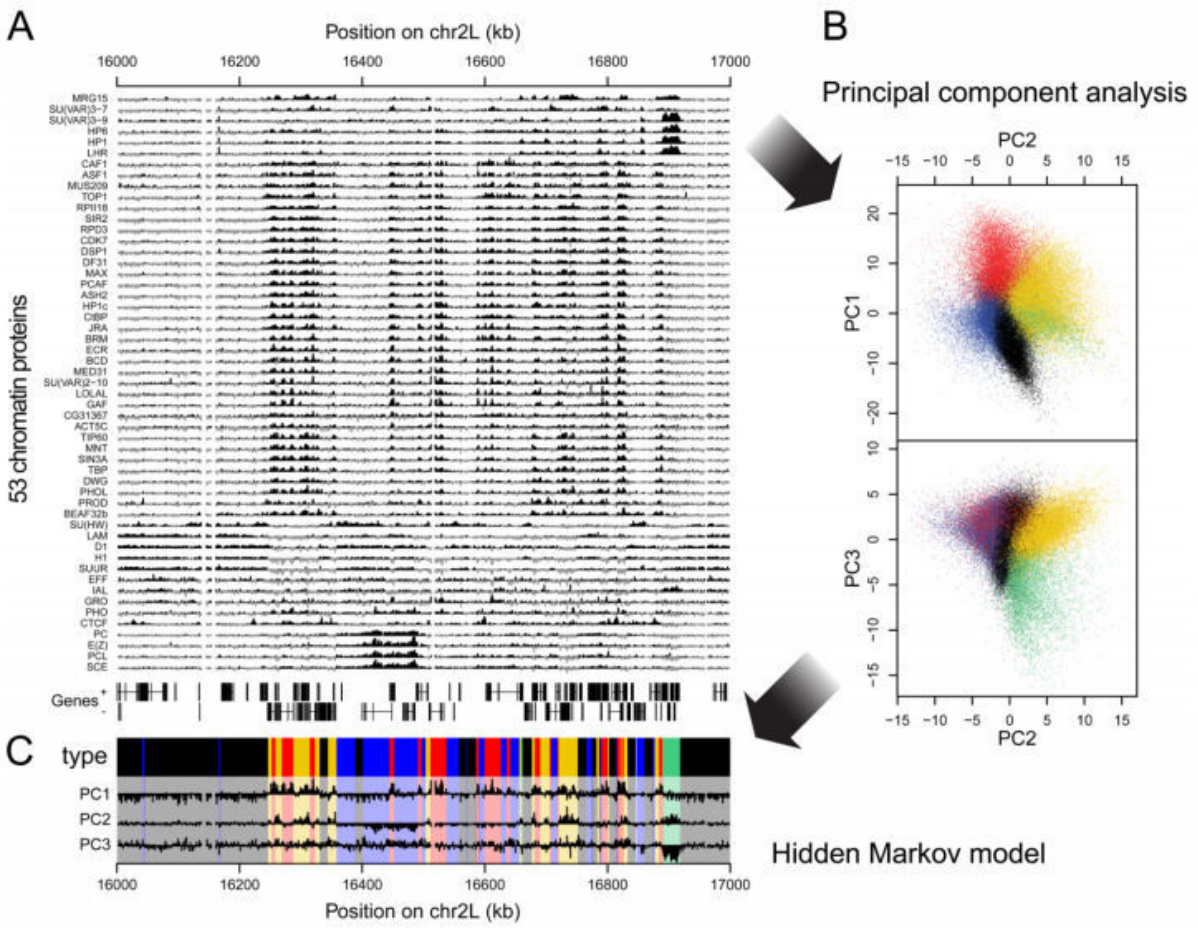

Figure 1.3.2 – **Exemple de segmentation de la chromatine en 5 états résultant des profils de protéines de liaison chez la Drosophile.** (A) Graphe représentant les 53 profils DamID. Les valeurs positives sont représentées en noir et les négatives en gris. En dessous des profils, les gènes des deux brins sont représentés sous forme de lignes avec des blocs indiquant les exons. a (B) Projections 2D des données sur les trois premières composantes principales. La couleur des points indique le type chromatinien des loci sondés commé inféré par un HMM à 5 états. (C) Valeurs des trois premières composantes principales le long de la région définie en (A), accompagnées des domaines des différents types chromatinien obtenus après segmentation par HMM. Le code couleur est le même qu'en (B). [Source : Filion et al., 2010].

1.3 Compartimentation 1D de l'épigénome

Les chromosomes eucaryotes sont ainsi composés de deux types de domaines épigénomiques structurels et fonctionnels chromatiniens : l'euchromatine, ouverte et généralement accessible, où l'on retrouve la plupart des gènes actifs et l'hétérochromatine, plus fortement condensée, et répressive. L'étude des profils de distributions de protéines régulatrices et des marqueurs épigénétiques obtenus par les récentes méthodes haut-débit ont permis de préciser un peu plus cette compartimentation 1D des génomes [Van Steensel, 2011] : de la drosophile [Filion et al., 2010] (Fig. 1.3.2) à l'homme [Ernst et al., 2011], en passant par les plantes [Roudier et al., 2011] et C. elegans [Gerstein et al., 2010] on peut distinguer quatre types de domaines chromatiniens (Fig. 1.3.2) : l'euchromatine qui contient les gènes actifs (gènes toujours actifs, «jaune» / gènes spécifiques à certains tissus, «rouge»), l'hétérochromatine constitutive de type « HP1/H3K9me » plutôt enrichie en éléments transposables et en séquences répétées (« verte »), une hétérochromatine facultative dite « Polycomb » enrichie en gènes impliqués dans la régulation de la différenciation et du développement (« bleue ») et une hétérochromatine dite nulle enrichie en gènes qui ne sont exprimés que dans très peu de tissus («noire»). L'hétérochromatine constitutive est présente de façon permanente, tandis que l'hétérochromatine facultative permute entre un état hétérochromatinien et un état euchromatinien selon le contexte biologique. D'un point de vue fonctionnel, l'hétérochromatine contrôle plusieurs aspects fondamentaux du fonctionnement nucléaire : (i) assemblage du kinetochore (ii) cohésion des chromatides soeurs assurant ainsi la bonne ségrégation durant la division cellulaire (iii) recombinaison : inhibition de toute recombinaison inopinée au niveau des séquences répétées garantissant ainsi une stabilité génomique (iv) expression des gènes : répression de la transcription des séquences sous-jacentes et voisines (v) activation/répression de certaines interactions à longue distance impliquées notamment dans la régulation du développement. L'hétérochromatine est en effet associée à la différenciation cellulaire, même dans les organismes unicellulaires où elle contrôle le type cellulaire et la reproduction sexuée. Dans les organismes multicellulaires l'hétérochromatine est impliquée dans la maintenance de l'identité cellulaire au cours du développement. D'un point de vue moléculaire, l'hétérochromatine se caractérise par des signatures biochimiques et structurelles particulières [Beisel and Paro, 2011] : l'hypoacetylation des histones, la méthylation spécifique H3K9me (resp. H3K27me), l'association avec des protéines structurales de la famille HP1 (resp. PcG) (ou H1/lamine) et par une distribution des nucléosomes très périodique. Cette organisation épigénétique du génome est dynamique : en effet, le développement des cellules germinale et la différenciation des cellules souches en cellules somatiques sont associés à une reprogrammation de ces domaines chromatiniens [Hawkins et al., 2011] ; la reprogrammation pathologique de ces do-

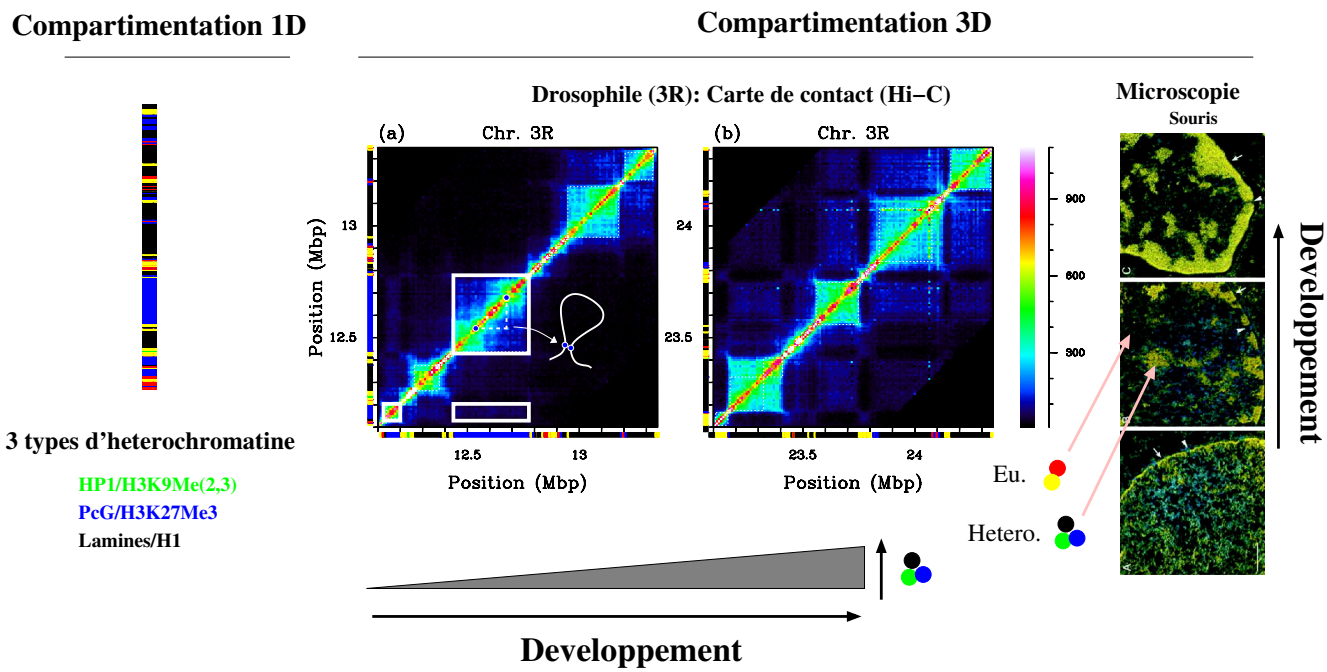

Figure 1.4.1 – **compartimentation 3D de de l'épigénome.** Les cartes Hi-C [Sexton et al., 2012] des cellules d'embryons tardifs de la drosophile révèlent une compartimentation du génome aux échelles sub-chromosomiques avec l'existence de domaines d'interaction préférentielle, les « TADs ». Ces domaines sont fortement corrélés aux domaines épigénomiques suggérant un mode de repliement par auto-association des états chromatiniens de même type (ou couleur). De façon consistante, ces cartes révèlent des interactions à longue distance entre TADs de même couleur épigénomiques. Cette compartimentation s'observe en microscopie électronique avec la ségrégation hétérochromatine / euchromatine qui se renforce au cours du développement.

maines contribue souvent à la cancérogénèse ou à d'autres pathologies [Feinberg, 2007]. La régulation « épigénétique » de ces domaines est confrontée à un double challenge : à la fois permettre une plasticité chromatinienne pour garantir une plasticité phénotypique au cours du développement et assurer une robustesse de ces états chromatiniens pour maintenir l'identité phénotypique des cellules dans un environnement fluctuant [Pujadas and Feinberg, 2012].

1.4 Compartimentation 3D de l'épigénome

De façon assez remarquable, les récentes expériences de «Capture de conformation de la chromatine» [Sexton et al., 2012] et de microscopie [Chandra et al., 2012; Cheutin and Ca-

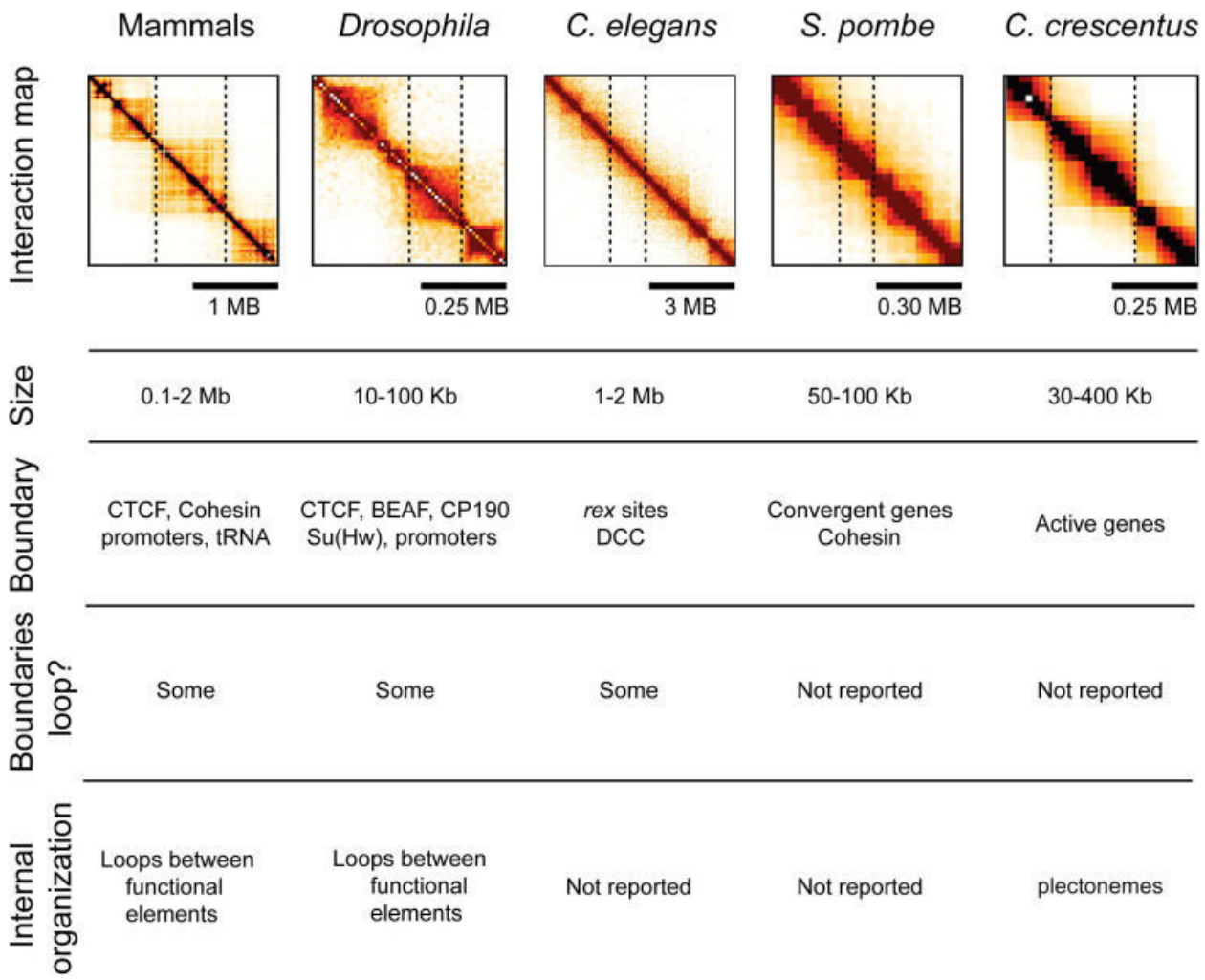

FIGURE 1.4.2 – **Universalité des TADs : des unicellulaires aux mammifères [Source : Dekker and Heard 2015].**

valli, 2012; Boettiger et al., 2016] ont permis de montrer que cette partition 1D du génome se retrouve également dans l'organisation spatiale de la chromatine [Naumova and Dekker, 2010 ; Van Steensel, 2011] : comme le montre la figure 1.4.1 les domaines épigénomiques (1D) s'organisent en effet en domaines topologiques (micro-phases 3D, appelés « TAD » pour « Topological Associating Domains ») caractérisés par des interactions spatiales essentiellement « intra-domaine » : les domaines adjacents apparaissent ainsi « isolés » les uns par rapports aux autres. Les domaines topologiques de même « couleur » épigénétique ont de plus tendance à interagir entre eux [Sexton et al., 2012] suggérant ainsi un mécanisme moléculaire d'interaction spécifique entre mêmes états épigénétiques locaux. L'organisation de ces domaines n'est par ailleurs pas aléatoire au sein du noyau [Ahmed et al., 2010; Naumova and Dekker, 2010; Rapkin et al., 2012; Chandra et al., 2012] : l'hétérochromatine a tendance à se

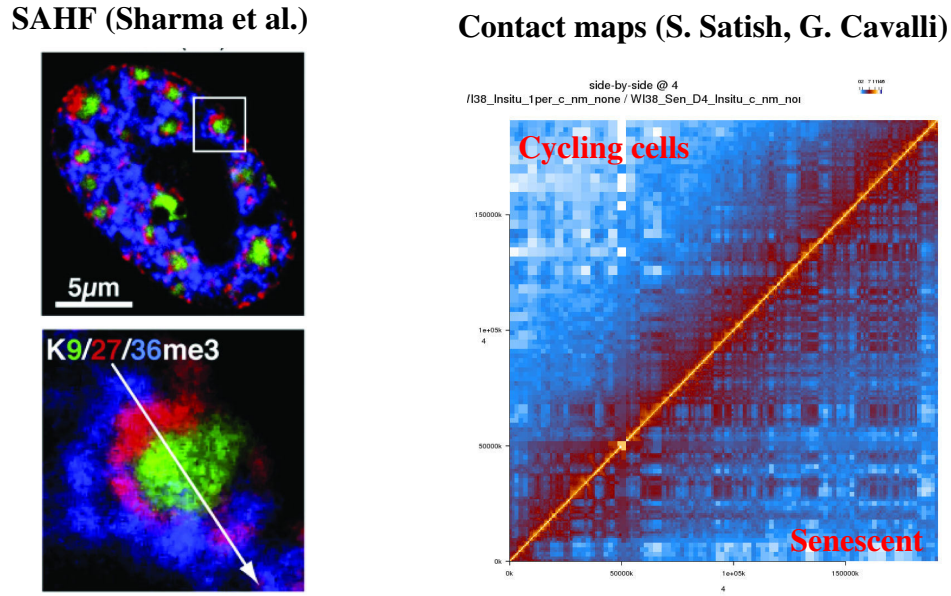

Figure 1.4.3 – **Réorganisation de l'hétérochromatine dans les cellules humaines en sénescence.** (Gauche) Images par microscopie d'immuno-fluorescence de l'organisation spatiale de la chromatine active H3K36me3 (bleue), inactive H3K27me3 (hétérochromatine de type PolyComb, rouge) et inactive H3K9me3 (Hétérochromatine constitutive, vert) : cette dernière forme des clusters caractéristiques appelés « SAHF » à l'intérieur du noyau contrastant avec sa colocalisation au niveau de la membrane dans les cellules somatiques non sénescentes [Source : Sharma et al., 2014]. (Droite) Cette réorganisation se traduit au niveau des cartes Hi-C par une augmentation des contacts à longue distance entre TADs hétérochromatiniens comparé à ce qui est observé dans les cellules non sénescentes (données Hi-C non publiées, lab. G. Cavalli).

former plutôt à la périphérie du noyau ainsi qu'autour du nucléole tandis que l'euchromatine est localisée plutôt à l'intérieur [Zullo et al., 2012] (Fig. 1.4.1 à droite). Récemment, il a été montré chez les mammifères qu'une partie (50 %) seulement des TADs pouvait être associée à des domaines épigénomiques [Sanborn et al., 2015]. En fait, et ce indépendamment des états chromatiniens, une grande partie des TADs sont caractérisés à leur frontières par un enrichissement en complexes CTCF/cohésine (Fig. 1.4) ; la formation de ces TADs seraient aussi associées à l'activité de complexes « extrudeurs » et non exclusivement à l'association entre mêmes états chromatiniens (cf. 1.6). En somme, il y a globalement une forte corrélation entre la compartimentation 3D en TADs et la segmentation 1D épigénomique ou/et la distribution de sites de fixations de protéines type CTCF (ou autres protéines insulatrices) formant les frontières.

Cette organisation spatiale évolue au cours du développement, en passant d'une faible organisation globale dans les cellules pluripotentes à une ségrégation et compartimentalisation

fortes dans les cellules différenciées [Meister et al., 2011; Ahmed et al., 2010] (Fig. 1.4.1). Cette réorganisation spatiale est dans le cas de l'embryogénèse fortement corrélée à la dynamique des domaines épigénomiques [Hawkins et al., 2011] mais au cours d'autres processus développementaux une réorganisation spatiale peut intervenir indépendamment de la dynamique des marques épigénétiques [Chandra et al., 2012] ; au cours de la sénescence cellulaire (celle induite par voie oncogénique), une partie de l'hétérochromatine de type H3K9me associée normalement aux lamines de la membrane nucléaire, se délocalise de la surface pour former des clusters (les SAHF = « Senescence Associated Heterochromatin Foci ») dans le nucleoplasme [Chandra et al., 2012] 1.4. Un point remarquable est que, une fois formés et consolidés au terme de l'embryogenèse, les TADs sont globalement stables au cours du développement et même, d'une espèce à une autre entre blocs synténiques [Vietri Rudan et al., 2015] : les variations s'observent au niveau des contacts inter-TADs (par exemple, amplifiés au cours de la senescence comme illustré à la figure 1.4 ou bien plus localement avec des modifications de contacts entre sites régulateurs (« enhancers »/promoteurs) au sein des TADs [Le Dily et al., 2014]. Il a été récemment montré que les TADs constituent des environnements isolés pour les gènes contrôlant l'identité cellulaire : que ce soit pour maintenir l'activation ou la répression de ces gènes de façon appropriée (c'est-à-dire dans le bon type cellulaire), ces TADs permettent de limiter l'effet d'interactions « indésirables » entre les séquences promotrices et des sites de régulations extérieurs. Il a d'ailleurs été montré que la dérégulation des frontières peut mener à des pathologies développementales ou cancers par dérégulation du programme transcriptionnel [Dowen et al., 2014; Lupiáñez et al., 2015; Hnisz et al., 2016] (Fig. 1.4).

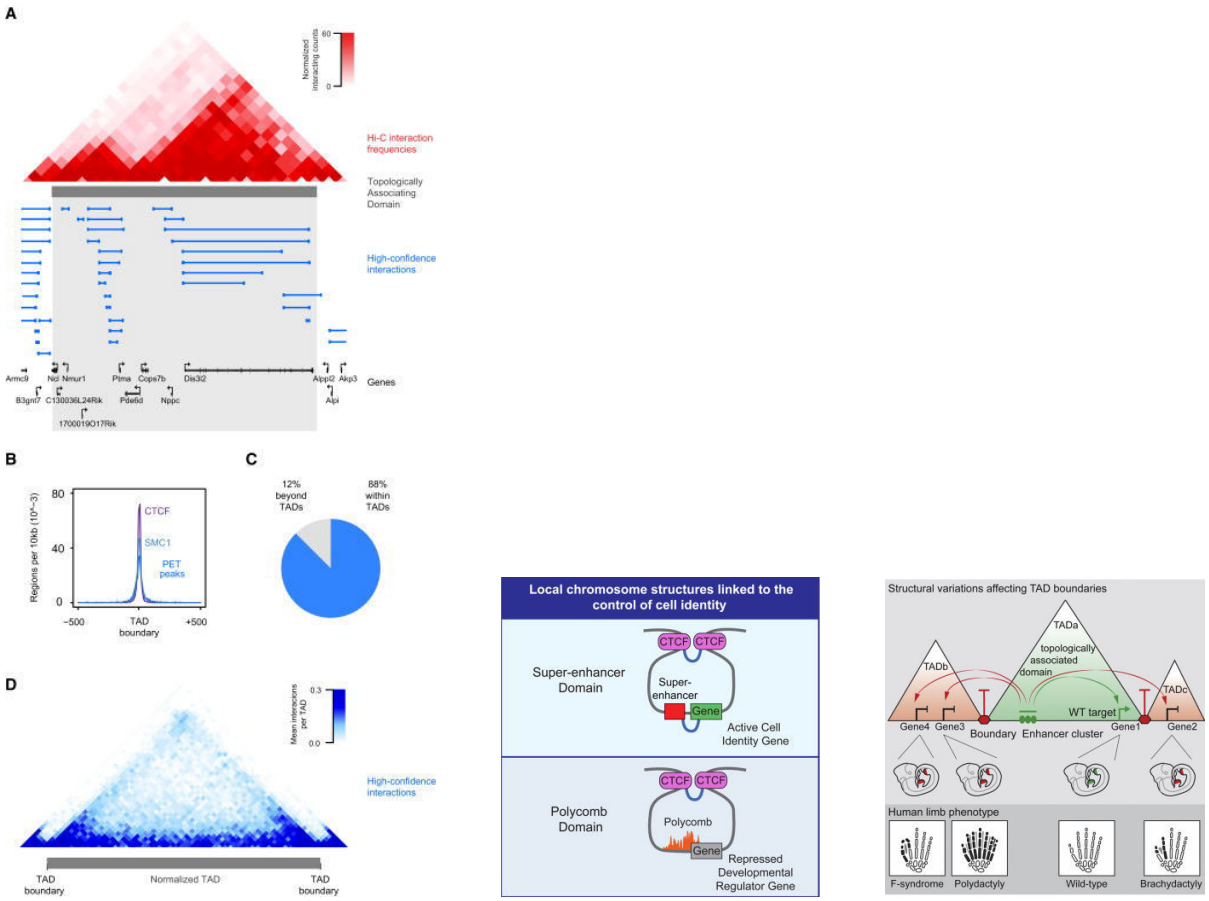

Figure 1.4.4 – **TADs et contrôle transcriptionnel.** (À gauche) Les TADS chez les mammifères sont en grand partie caractérisés par des frontières enrichies en complexes CTCF/Cohesin (Milieu). Ces TADs forment des compartiments isolés assurant la bonne régulation des gènes contrôlant l'identité cellulaire en limitant toute interaction inopinée avec d'autres sites de régulation extérieurs [Dowen et al., 2014]. (À droite) La dérégulation des ces TADs notamment par altération des frontières peut conduire à des pathologies [Lupiáñez et al., 2015]

Protocole du Hi-C

Afin d'étudier comment l'ADN se replie à l'échelle de tout le génome, il existe une technique de "Capture de Conformation de la Chromatine" dite technique de "Hi-C". Il s'agit de fixer au formaldéhyde la chromatine puis la digérer à l'aide d'une enzyme de restriction. Une ligation est alors effectuée de manière à rabouter ensemble les portions de chromatine reliées par le formaldéhyde. L'ADN est ensuite purifié. Au final, on se retrouve avec des molécules d'ADN chimériques qui reflètent les relations de proximité dans le noyau. Ces morceaux d'ADN sont analysés par séquençage haut-débit ce qui permet d'analyser l'ensemble des interactions entre locus génomiques. Le résultat de cette expérience se présente sous la forme d'une carte de

contacts indiquant le nombre d'interactions détectées entre sites génomiques deux à deux (Fig. 1.4.5). En effet, la représentation des contacts intra et inter chromosomiques peut se faire via la matrice de contact $\mathbf{C}$ où les coefficients $\mathbf{C}_{ij}$ donnent le nombre de contacts entre deux brins $\mathbf{i}$ et $\mathbf{j}$. La taille des bins vaut au minimum celle d'un fragment de restriction (typiquement long de 800pb [Sexton et al., 2012]). Actuellement, la profondeur de séquençage atteinte en laboratoire nécessite de présenter les cartes de contact avec une résolution supérieure à la taille du fragment de restriction (entre 10kb [Sexton et al., 2012] et 40kb [Dixon et al., 2012]). La matrice brute ainsi obtenue souffre de divers biais expérimentaux (GC content, taille des fragments, mappability des fragments ...). Il existe différents « pipelines » visant à corriger ces biais.

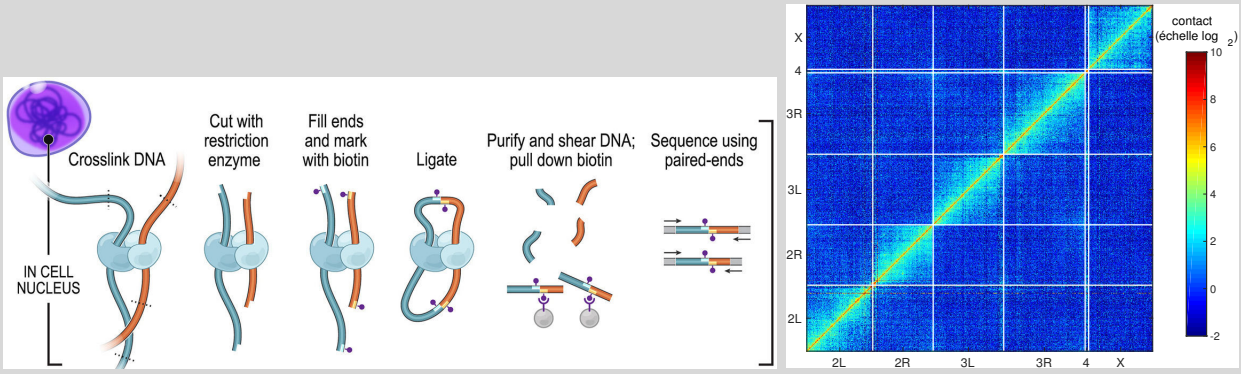

FIGURE 1.4.5 – **Technique du Hi-C.** (À gauche) Grandes lignes du principe expérimental. (À droite) Carte de contact pour tout le génome de la drosophile. On présentera les cartes de contact toujours en échelle logarithmique ($\mathtt{log}_2$) afin de mieux voir les contrastes.

Biais expérimentaux liés à la technique du Hi-C et méthodes de normalisation

La technique du Hi-C étant particulièrement complexe, elle peut induire plusieurs biais et artefacts de nature différente [Cournac et al., 2012] :

1. La profondeur de séquençage dépend de la région génomique. Par exemple, les régions riches en séquences répétées sont sous représentées au niveau de l'interactome car on sait qu'elles sont généralement non « mappable ».
2. Les régions avec beaucoup de contenu en GC sont sous représentées à cause du biais de séquençage.
3. La probabilité de se fixer sur l'ADN pour une enzyme de restriction est d'autant plus grande que cette enzyme est grande.

Plusieurs algorithmes de normalisation des cartes Hi-C ont été développés dans le but de limiter les artéfacts et biais non désirés liés à la technique du Hi-C. Ces méthodes reposent principalement sur l'une des deux approches suivantes :

1. Normalisation par utilisation d'un modèle probabiliste. Le modèle de Yaffe and Tanay, 2011 par exemple détermine la probabilité de contact Hi-C entre deux fragments en multipliant des termes de « mappability », de longueurs d'enzyme de restriction et de contenu en GC. Le coût en terme de temps de calcul est élevé. Cette approche a été présentée comme permettant de supprimer la majorité des biais systématiques (Fig. 1.4) . Une approche comparable à la précédente mais qui introduit moins de paramètres et qui est moins coûteuse en terme de temps de calcul est : une normalisation par régression de Poisson [Hu et al. (2012)]. Cette méthode consiste à estimer les biais dûs à la longueur et au contenu en GC en fixant la « mappabilité » selon une loi de Poisson.
2. Algorithmes itératifs [Cournac et al., 2012 et Imakaev et al., 2012] : ces méthodes consistent à normaliser les lignes et colonnes itérativement jusqu'à ce que la matrice Hi-C soit de nouveau symétrique. Cette méthode est généralement utilisée car a priori elle supprime également les biais inconnus contrairement aux deux autres méthodes précédentes.

Dans cette thèse, on utilisera majoritairement les données de Sexton et al., 2012 normalisées selon Yaffe and Tanay, 2011.

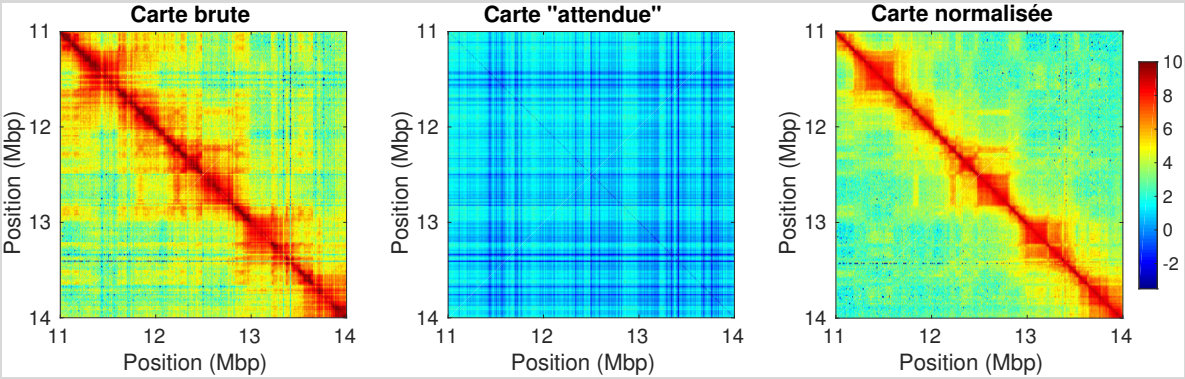

Figure 1.4.6 – **Normalisation d'une carte de contact par utilisation d'un modèle probabilité des contacts Hi-C [Yaffe and Tanay, 2011].** (À gauche) Carte de contact brute pour une région génomique du chromosome 3R de la drosophile [Sexton et al., 2012]. (Au milieu) Carte de contact prévue par le modèle probabiliste [Yaffe and Tanay, 2011]. (À droite) Carte Hi-C normalisée en divisant les contacts bruts par ceux attendus. L'échelle de couleur est commune aux trois images.

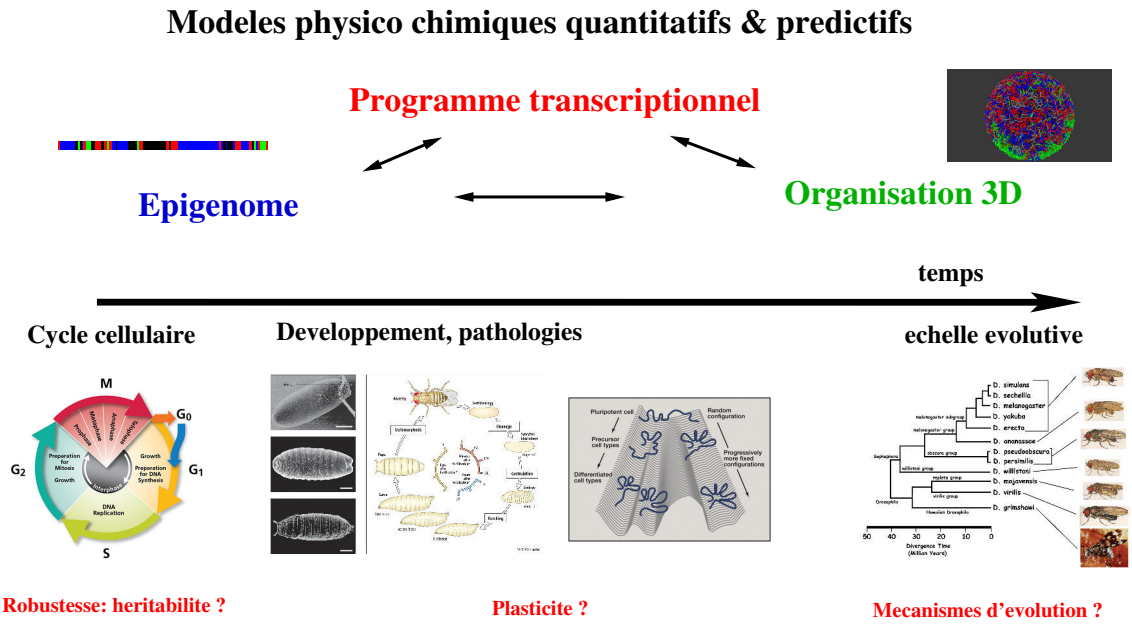

FIGURE 1.5.1 – **Objectifs :** Construire des modèles prédictifs du couplage 1D (Epigénome) /3D (Contactome) et de son implication dans la régulation des gènes. Modélisation de la dynamique de ce couplage à différentes échelles de temps : (1) à l'échelle du cycle cellulaire avec la question de la transmission mitotique, donc de la mémoire épigénétique ? Réorganisation en phase de quiescence/senescence lorsque la cellule ne cycle plus ? (2) à l'échelle du développement : mise en place du 1D/3D au cours de l'embryogenèse ? Transmission meiotique ? Vieillissement ? Dérégulations menant à des pathologies (malformations, cancers...) ? (3) à l'échelle évolutive : émergence ? conservation ? renforcement ?

1.5 Objectifs : modèles prédictifs

En collaboration avec l'équipe de Giacomo Cavalli notamment, l'objectif du groupe où j'ai effectué ma thèse est de mettre en place des modèles « fonctionnels » d'organisation des chromosomes : des modèles qui puissent décrire l'organisation spatiale des chromosomes aux échelles pertinentes (de quelques dizaines de kbp aux chromsosmes entiers) au cours du cycle cellulaire et au cours du développement. (i) A partir des expériences (de Hi-C en particulier) construire un modèle et en inférer les paramètres (ii) Analyser le lien entre ces paramètres et l'activité transcriptionnelle (l'épigénome pour commencer) et/ou la séquence génomique (donc le 1D) (iii) Prédire les réorganisations spatiales en fonction des réorganisations de l'épigénome (au cours du développement et pathologies) et de la séquence (évolution). Pour mener à bien ce projet, chez la drosophile, des expériences Hi-C menées à différents stades du cycle cellulaire et du développement, dans des mutants et dans d'autres espèces proches sont en cours.

1.6 État de l'art concernant la modélisation de la chromatine

1.6.1 Reconstruction de la structure 3D de la chromatine à partir de données Hi-C

À partir des données Hi-C, il est possible de construire des modèles 3D de la chromatine qui respectent les contraintes imposées par la distribution des contacts. Ces approches de reconstruction 3D sont nommées probabilistes, statistiques, contraintes, ou encore inverses par Rosa et al. [Rosa and Zimmer, 2014] (ou « top down ») en opposition aux modèles polymériques directs (ou « bottom up »). Elles peuvent être divisées en deux sous groupes : (1) Les méthodes du premier groupe visent à trouver une conformation consensus qui décrit au mieux les données Hi-C [Duan et al., 2010; Tanizawa et al., 2010; Zhang et al., 2013; Ay et al., 2014; Lesne et al., 2014; Varoquaux et al., 2014]. Ces méthodes tentent de déterminer les coordonnées 3D d'une telle conformation en supposant connus les distances entre loci, par exemple via la technique statistique de positionnement multidimensionnel (MDS). Ces distances sont souvent déduites des contacts moyens Hi-C en utilisant une loi d'échelle pour convertir contact en distance. Appliquées à des données Hi-C, pour des cellules uniques, ces méthodes devraient permettre d'accéder à la conformation 3D unique correspondante [Lesne et al., 2014].
(2) Les méthodes du second groupe génèrent un ensemble de structures possibles, elles sont donc adaptées à l'analyse de données Hi-C qui proviennent d'une multitude de cellules. En effet, si l'expérience de Hi-C est réalisée avec plusieurs cellules, la chromatine de chaque cellule ne présente pas forcément le même repliement [Nagano et al., 2013]. Ces méthodes peuvent être sous divisées en deux catégories selon si les conformations trouvées ont chacune vocation à reproduire les données Hi-C (finalement il s'agit d'inférer plusieurs séquences consensus ce qui est assez similaire aux méthodes du groupe (1)) [Baù et al., 2011; Rousseau et al., 2011; Giorgetti et al., 2014], ou bien si l'ensemble des conformations a pour but de reproduire les données Hi-C [Kalhor et al., 2012; Hu et al., 2013; Peng et al., 2013; Trieu and Cheng, 2014; Wang et al., 2015].
Plusieurs revues publiées aux cours de ces cinq dernières années détaillent toutes ces méthodes indirectes de reconstruction 3D de la chromatine [Mirny, 2011; Dekker et al., 2013; Rosa and Zimmer, 2014; Lajoie et al., 2015; Ay and Noble, 2015; Wang et al., 2015].
Dans les deux sous sections qui suivent, nous allons présenter des méthodes de modélisation de la chromatine basées sur la physique des polymère. Ces méthodes directes (ou bottom up) reposent sur un certain nombre d'hypothèses et de paramètres et tendent à valider si ces

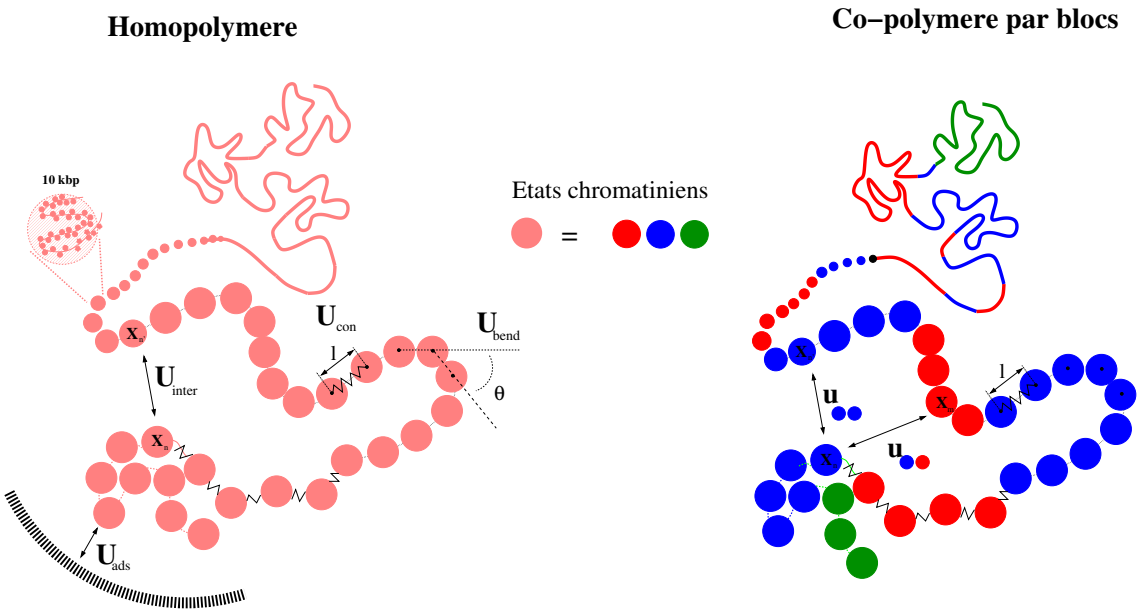

FIGURE 1.6.1 – **Modélisation de la chromatine par un copolymère par bloc.** Modéliser la chromatine par un homopolymère (à gauche) tout en tenant compte de contraintes topologiques (interactions avec la membrane, ...) permet de reproduire la structure à grande échelle de la chromatine. Toutefois, ce modèle ne tient pas compte de l'observation expérimentale selon laquelle les locus de même état épigénomique interagissent préférentiellement. De ce fait, on modélise la chromatine par un hétéropolymère ou copolymère par bloc chaque monomère une couleur représentative de son état épigénomique (à droite). Le repliement d'une telle chaîne est dirigé par des interactions non spécifiques caractéristiques du confinement ou du volume exclu et par des interactions attractives épigénome-dépendantes : les monomères de même couleur interagissent préférentiellement.

hypothèses sont compatibles avec les données expérimentales.

1.6.2 Modélisation de la chromatine par un homopolymère

Un modèle minimal permettant d'étudier l'organisation 3D et la dynamique des chromosomes est de considérer la chromatine comme un homopolymère semi-flexible et auto-évitant (Fig. 1.6.1). La conformation d'une telle chaîne est contrôlée par l'agitation thermique, les répulsions stériques et les interactions effectives entre monomères, ces dernières reflétant le confinement, les interactions spécifiques entre monomères ou encore avec le solvant. Dans le cas où les interactions attractives et répulsives se compensent parfaitement (température θ), la chaîne est dite « idéale » ou « gaussienne ». Sa conformation statistique est analogue à celle laissée par un marcheur se déplaçant aléatoirement. Ainsi, la probabilité que deux monomères soient en contact, P_c en fonction de la distance génomique s évolue selon la loi $P_c(s) \sim s^{-3/2}$.

Au dessus de la température θ (cas d'un « bon solvant »), la répulsion entre monomères domine sur l'attraction, le polymère est une chaîne « gonflée » telle que $P_c \sim s^{\alpha}$ avec $\alpha < -3/2$, typiquement $P_c \sim s^{-2}$.

En dessous de la température θ (« mauvais solvant »), les attractions entre monomères sont plus fortes que les répulsions ce qui provoque une compression du polymère qui s'équilibre en s'effondrant sur lui même. On parle de globule équilibré. Si on modélise la chromatine par un tel globule en équilibre la probabilité de contact, P_c, évoluera en fonction de la distance génomique, s, selon la loi d'échelle $P_c(s) \sim s^{-3/2}$ pour des tailles caractéristiques plus petites que celle du globule ($s < N^{2/3}$ avec N le nombre total de monomères dans la chaîne), pour des distances plus grandes ($s > N^{2/3}$), $P_c(s) \sim s^0$ [Gennes, 1979 ; Grosberg et al. (1995) ; Mirny, 2011].

Il a été montré que ces prédictions théoriques sont compatibles avec un polymère gaussien. Pour des petits chromosomes comme ceux de la levure *S. cerivisiae* (taille inférieure à 1Mb), Rosa and Everaers ont montré, par des simulations numériques, que l'état d'équilibre est rapidement atteint (temps inférieur à la durée d'un cycle cellulaire qui est d'environ 1 heure).

Par contre, pour des chromosomes beaucoup plus longs que ceux de la levure, comme ceux de l'homme ou de la drosophile, les données expérimentales (FISH et Hi-C) mènent à un exposant différent : $P_c(s) \sim s^{-1}$ [Lieberman-Aiden et al., 2009]. Ainsi, la décroissance de la probabilité de contact en fonction de la distance génomique est plus lente pour les grands chromosomes que pour les petits. Afin d'expliquer ce régime de « crumpling », il est possible de voir dans ce cas la chromatine, non pas comme un globule équilibré, mais comme un globule intermédiaire de longue durée de vie précédant l'état d'équilibre. Il s'agit là du modèle de « globule fractal » non équilibré au sein duquel la chromatine ne forme pas de nœuds et est agencée en sous-territoires consécutifs non entremêlés [Grosberg et al., 1988]. L'idée se trouvant derrière ce modèle est qu'à la mitose la chromatine s'organise sous forme de chromosomes mitotiques, bien séparés et sans noeuds, ainsi, durant l'interphase la décondensation de la chromatine débutera sans noeuds et ces derniers n'auraient donc pas le temps de se former étant donné que le temps nécessaire pour atteindre l'équilibre serait très long par rapport au cycle cellulaire.

A l'aide de simulations numériques, il a été montré que pour un globule fractal la probabilité de contact évolue selon la loi $P_c(s) \sim s^{-1}$ [Lieberman-Aiden et al., 2009 ; Mirny, 2011 ; Tamm et al., 2015]. Le globule fractal permet donc de reproduire les bonnes lois d'échelle chez des organismes ayant des chromosomes des grands chromosomes (entre 700kb et 7Mb). Aussi, ces simulations ont permis de mettre en évidence que le globule fractal présente une organisation en territoire très nette ce qui est en accord avec les images réalisées par microscopie durant l'interphase [Cremer and Cremer, 2010]. Cette ségrégation est beaucoup plus prononcée dans

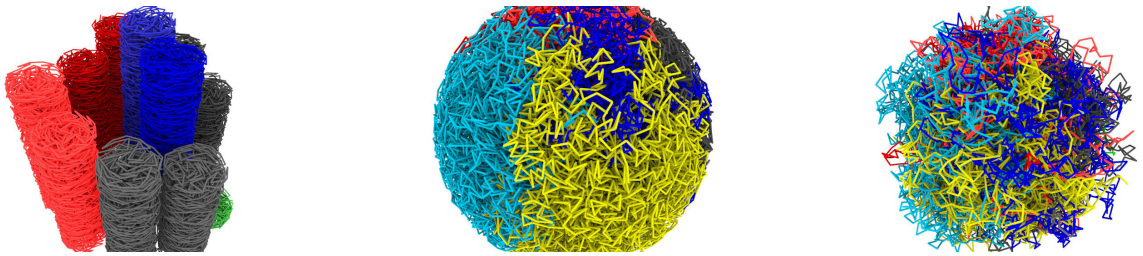

FIGURE 1.6.2 – **Modèles du globule fractal et du globule équilibré.** (À gauche) Configuration « initiale » en Rabl des chromsosomes de la drosophile à l'entrée en interphase. (Au milieu et à droite) Après relaxation à partir de cette configuration initiale, deux exemples de configurations instantanées obtenues par simulations de dynamique moléculaire [Source : Pascal Carrivain] en prenant en compte le volume exclu (au milieu) ou non (à droite). Le globule fractal (au milieu) est dans une phase intermédiaire de longue durée de vie qui évolue (lentement) en un globule équilibré (à droite). On observe sur l'image du milieu que les territoires chromosomiques sont bien définis par rapport au cas non équilibré.

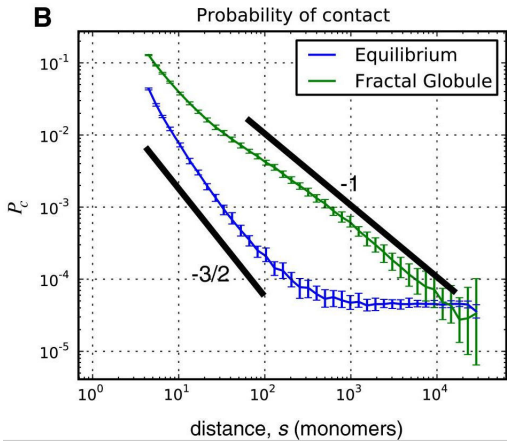

FIGURE 1.6.3 – **La modélisation de la chromatine sous forme d'un homopolymère dans la phase de globule fractal permet de reproduire les observations FISH et Hi-C réalisées chez l'homme ou la drosophile en reproduisant par exemple la loi d'échelle** $P_c \sim s^{-1}$ [Source : Mirny, 2011].

un globule fractal que dans un globule équilibré où les différentes régions chromosomiques sont assez enchevétrées (Fig. 1.6.2). Notons que les chaînes évoluent selon le modèle de reptation introduit par de Gennes [de Gennes, 1979]. Ainsi, le globule fractal évolue en globule équilibré de façon lente, selon une durée proportionnelle à N^3, avec N le nombre de monomères composant la chaîne [Rosa and Everaers, 2008 ; Mirny, 2011]. Dans le cas des grands de chromosomes, la durée d'un cycle cellulaire ne permet pas d'atteindre l'équilibre, le globule fractal est donc un état hors équilibre mais de longue durée de vie. Ce modèle de globule fractal est d'autant plus intéressant qu'il permet non seulement de retrouver la ségrégation en territoire et le résultat expérimental selon lequel $P_c(s) \sim s^{-1}$ mais il permet aussi d'expliquer, (i) comment la chromatine est stockée de façon compacte tout en évitant les noeuds et enchevêtrements et (ii) comment la chromatine peut facilement s'enrouler, se dérouler se replier en boucle ou au contraire déplier des boucles selon les besoins de la cellule (réplication, transcription, répression...).

Toutefois, il est important de préciser que la loi de décroissance de la probabilité de contact $P_c(s) \sim s^{-1}$ observée expérimentalement peut aussi être reproduite sans faire appel au globule fractal. En effet, des simulations numériques reposant sur un modèle d'anneaux sans noeuds en équilibre dans une solution semi-diluée [Halverson et al., 2011 ; Rosa and Everaers, 2014 (Fig. 1.6.3)] ou encore avec un modèle d'homopolymère linéaire tenant compte explicitement du phénomène de boucle à longue portée via l'introduction d'interactions attractives dynamiques [Bohn and Heermann, 2010 ; Barbieri et al., 2012] permettent d'aboutir à une décroissance de la probabilité de contact selon $P_c(s) \sim s^{-1}$.

En résumé, les travaux présentés ci-dessus (et dont une description plus complète pourra être trouvée dans la revue réalisée par Rosa and Zimmer (2014)) montrent qu'un modèle homopolymèrique associé à des contraintes topologiques bien choisies (pas de noeuds afin de maintenir les territoires chromosomiques, présence de boucles de chromatine pour expliquer l'indépendance entre la distance physique et la séparation génomique pour des distances supérieures à quelques Mb, ...) permet de reproduire les **propriétés génériques** de la chromatine à grande échelle (loi d'échelle pour la décroissance de la probabilité de contact, existence des territoires chromosomiques, ...). Toutefois, ces modèles construits à partir d'homopolymères ne permettent pas d'expliquer les **propriétés spécifiques** de la chromatine à petite échelle (formation des TADs, ...). Ceci peut s'expliquer par le fait que ces modèles ne tiennent pas compte (i) de la compartimentation épigénomique de la chromatine [Filion et al., 2010 ; Kharchenko et al., 2011 ; Ho et al., 2014], (ii) de l'existence d'interactions spécifiques entre états chromatiniens de même type [Sexton et al., 2012] et (iii) n'intègrent pas non plus l'effet de protéines (CTCF, cohésine, lamina, ...) jouant pourtant un rôle fondamental dans le repliement 3D et dans la dynamique de la chromatine [Cavalli and Misteli, 2013 ; Lieberman-Aiden

et al., 2009 ; Dekker et al., 2013 ; Bickmore and van Steensel, 2013 ; Ciabrelli and Cavalli, 2015 ; Holwerda and de Laat, 2012]. En effet, plusieurs preuves suggèrent que ces protéines architecturales promeuvent la formation de ponts physiques entre différentes régions chromatinienne [Canzio et al., 2013 ; Isono et al., 2013 ; Phillips-Cremins et al., 2013]. Dans ce contexte, plusieurs modèles physiques tenant compte du couplage entre structure 3D et fonction ont récemment émergé [Jost et al., 2014 ; Barbieri et al., 2012 ; Brackley et al., 2013 ; Tark-Dame et al., 2014 ; Doyle et al., 2014 ; Giorgetti et al., 2014 ; Nazarov et al., 2015 ; Sanborn et al., 2015 ; Fudenberg et al., 2016 ; Ulianov et al., 2016 ; et la revue Imakaev et al., 2015].

1.6.3 La chromatine est un hétéropolymère : repliement d'un copolymère par bloc

Comme précisé auparavant, un modèle d'homopolymère ne peut rendre compte de la compartimentation de régions génomiques spécifiques. Il est assez évident que la chromatine, notamment du fait de la présence d'états chromatiniens différents, peut être décrite par un modèle gros grain d'hétéropolymère, à savoir un polymère dont les monomères ne sont pas identiques et qui ont des propriétés géométriques, d'interaction, de diffusivité différentes. De manière générale, les monomères de type différent se mélangent peu, ce qui induit une séparation en microphases contenant les monomères de même type [Grosberg and Khokhlov, 1994] : on peut donc souvent décrire le repliement d'hétéropolymère en introduisant des interactions attractives effectives entre monomères de même type. Cependant, la question est de savoir dans quelle mesure ces variations locales de composition, de structures, diffusivité etc... locales peuvent rendre compte des compartiments 3D observés : en somme quelle est l'intensité des interactions effectives attractives ?

Un premier travail préliminaire mené par D. Jost et collaborateurs [Jost et al., 2014] a donc été de comprendre à travers un modèle de polymère comment la compartimentalisation en domaines épigénomiques pouvait rendre compte de l'organisation spatiale de la chromatine et notamment de la compartimentation 3D en TADs observée chez la drosophile. Ils ont introduit un modèle de copolymère par blocs où chaque bloc correspond à un domaine épigénomique (Fig. 1.6.1). Les interactions de volume entre monomères sont de deux types : interactions non-spécifiques (pour rendre compte du confinement global) et interactions spécifiques entre monomères de même « couleur » épigénétique. Cette spécificité est motivée par de nombreuses données expérimentales révélant une interaction effective entre fragment de même « état » chromatinien (couleur). En effet, il a été montré que les complexes protéiques du groupe Polycomb [Francis et al., 2004; Lo et al., 2012; Isono et al., 2013] et HP1 [Canzio et al.,

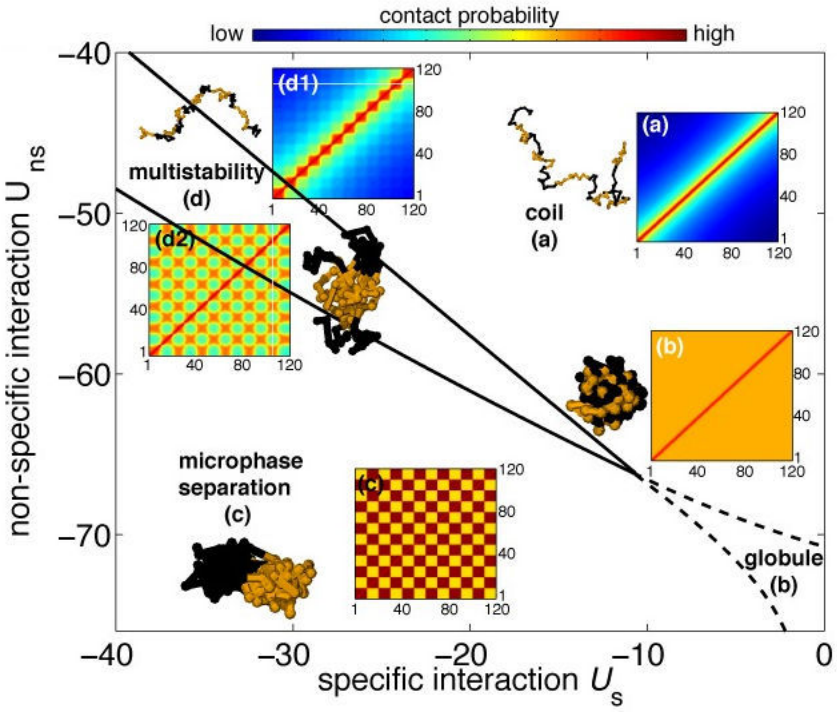

Figure 1.6.4 – **Diagramme de phase d'un copolymère par bloc $(A_{10}B_{10})_6$ en fonction des interactions spécifiques et non spécifiques en unité k_BT.** Les inserts représentent des structures et des cartes de probabilité de contact typiques (en échelle log) pour chacune des phases : (a) pelote, (b) globule, (c) séparation en microphase, (d) multistabilité. Source : [Jost et al., 2014].

2013] peuvent créer (par oligomerisation) des pontages physiques entre sites distants. De plus, la mutualisation des polII (usines de transcription), la formation de boucles entre sites de régulation par des protéines insulatrices du type CTCF, ou l'ancrage à des pores nucléaires peuvent conduire aussi à une interaction effective entre sites actifs [Phillips-Cremins et al., 2013]. Enfin, la chromatine dite « noire » (Fig. 1.3.2) est souvent associée aux « lamines » [Filion et al., 2010] suggérant une interaction effective via l'ancrage à la membrane.

Comme première approximation, Jost et al. ont en effet considéré que les interactions spécifiques avaient la même valeur quelle que soit le type épigénomique en question (i.e, par exemple deux monomères « Polycomb » interagissent avec la même intensité que deux monomères « HP1 »...). Ils ont modélisé la dynamique de la chaîne polymérique par une équation de Langevin non-linéaire, résolue par une méthode d'approximation de champ gaussien autoconsistant sur laquelle je reviendrai aux chapitres 4 et 5. Ils ont aussi mis en place des simulations de dynamique moléculaire. La méthode approchée permet de façon très efficace d'obtenir des diagrammes de phases, comme celui reporté à la figure 1.6.4 dans l'exemple pédagogique du copolymère $(A_{10}B_{10})_6$. Dans ce cas simple on voit déjà apparaître la com-

plexité du diagramme de phase avec notamment une zone de paramètres correspondant à de la multistabilité.
Appliqué à la drosophile, à savoir en considérant en entrée la séquence 1D des domaines épigénomiques obtenue par [Filion et al., 2010], ils ont montré que ce modèle de copolymère « minimal » pouvait déjà très bien rendre compte de l'organisation 3D telle qu'elle a été mesurée par les cartes Hi-C de [Sexton et al., 2012] (Fig. 1.4.1). Les résultats indiquent par ailleurs que les données expérimentales sont compatibles avec des configurations multistables à savoir des configurations caractérisées par les domaines épigénomiques repliés en « TADs » mais avec des interactions à longue portée entre TADs de même type epigénomique, et ce de façon dynamique (cf exemple pédagogique 1.6.4).
N.B. : Je reviendrai en détail sur ce modèle de copolymère par bloc et son application à la modélisation du repliement de l'épigénome chez la drosophile au chapitre 4.

1.7 Thèse : Objectifs et stratégie

Questions et objectifs

L'objectif est de modéliser les mécanismes d'organisation spatiale de la fibre de chromatine et notamment les mécanismes conduisant au repliement des domaines épigénomiques en domaines topologiques. A l'échelle d'un domaine génomique (par ex. les domaines regroupant les clusters de genes Hox [Noordermeer et al., 2011; Bantignies and Cavalli, 2011] impliqués au cours de l'embryogenèse dans la spécification de l'axe antérieur-postérieur du corps), à l'échelle d'un ensemble de domaines voire à l'échelle d'un chromosome, on s'intéressera au couplage dynamique entre l'état épigénetique et la structure 3D : en fonction des marques épigénétiques, de la présence de protéines architecturales ou insulateurs, on caractérisera les phases de condensation et leur dynamique.

Stratégie

Afin d'apporter des éléments de réponses à ces questions, la stratégie est de mettre en place des modèles d'architecture à grande échelle de la chromatine qui, puissent, via des simulations numériques et/ou analytico-numériques, rendre compte des données expérimentales obtenues entre autre par le groupe de biologistes de G. Cavalli à Montpellier, via des expériences de Hi-C et 3C [Bantignies and Cavalli, 2011; Sexton et al., 2012] et de microscopie optique haute-résolution [Cheutin and Cavalli, 2012]. Ces modèles permettront de tester certaines hypothèses en engageant d'autres expériences : organisation et dynamique de la chromatine dans des souches mutantes vs. souches sauvages, à différents stades du cycle cellulaire, à différents stade du développement. Par cet aller-retour constant entre expériences et théorie l'objectif sera donc de révéler certains principes génériques concernant la régulation de la structure

et de la dynamique 3D de l'épigénome au cours de la différenciation et du développement (embryogenèse, sénescence, reprogrammation...), et sa dérégulation lors de pathologies du type cancers.

Plan

L'objectif de cette thèse est de valider l'hypothèse selon laquelle l'épigénome est un acteur majeur dans le repliement 3D à partir d'une analyse statistique et à partir de prédiction réalisées avec un modèle physique de copolymère par bloc.

Dans le chapitre 2 on présentera « IC-Finder » un algorithme que nous avons développé afin de segmenter les cartes de contact en domaines d'interaction. Ensuite, dans le chapitre 3 nous verrons une analyse statistique quantifiant les corrélations entre contactome et épigénome. Dans le chapitre 4, on expliquera pourquoi et comment modéliser la chromatine par un copolymère par blocs. Nous introduirons dans le chapitre 5 différentes méthodes visant à inférer les paramètres du modèle développé dans le chapitre 4 à partir de cartes de contact expérimentales. Suite à cela, nous conclurons dans le chapitre 6.

CHAPITRE 2

SEGMENTATION DU GÉNOME EN TADS, IC-FINDER

Comme nous l'avons vu en introduction la technique du Hi-C a permis de révéler que la chromatine peut s'organiser sous forme de TADs caractérisés par des interactions spatiales essentiellement intra domaine, les domaines adjacents apparaissent ainsi isolés les uns des autres. Dans ce chapitre, on propose un outil, nommé IC-Finder permettant d'identifier la positions des TADs à partir de cartes de contact. La méthode de segmentation utilisée repose sur le principe de regroupement hiérarchique que nous avons adapté afin de prendre en compte la nature polymérique de la chromatine. À partir d'un ensemble de cartes de contact de référence (obtenues in silico et expérimentalement) on comparera les résultats obtenus avec IC-Finder et avec d'autres méthodes de segmentation. On verra que IC-Finder est l'une des meilleures techniques en terme de fiabilité et en terme de temps de calcul. De plus, on présentera deux options originales proposées par IC-Finder : une description probabiliste des TADs inférés et la possibilité d'explorer l'organisation de la chromatine de façon hiérarchique, c'est-à-dire selon plusieurs échelles.

2.1 Motivations

2.1.1 Importance des TADs

L'organisation de l'ADN des cellules eucaryotes sous forme de fibre chromatinienne hétérogène contribue à la régulation des gènes en contrôlant l'accessibilité des promoteurs et des séquences régulatrices à la machinerie de transcription [Allis et al., 2007]. L'organisation de la chromatine a longtemps été essentiellement étudiée localement en considérant le génome comme un objet unidimensionnel dont la structure locale est modulée par des informations épigénomiques telles que la méthylation de l'ADN, les marques sur les histones ou autres protéines se liant à la chromatine [Allis et al., 2007]. Toutefois, comme nous l'avons vu en introduction, la technique de Hi-C suggère que les chromosomes sont linéairement repliés en domaines 3D sous nucléaires, appelés TADs [Dixon et al., 2012]. Les TADs sont caractérisés par des fréquences de contact élevées à l'intérieur même des domaines et par des insulations partielles entre domaines adjacents consécutifs. Leur taille est variable, de quelque kb à des Mb et même plus dans le cas du chromosome X inactivé chez les mammifères [Deng et al., 2015]. Il a été montré que les TADs sont principalement conservés entre tissus et entre espèces voisines [Dixon et al., 2012 ; Rao et al., 2014 et Dixon et al., 2015]. Les légères différences que l'on observe sont généralement associées au développement et à la différenciation cellulaire [Dixon et al., 2015]. Le fait que les frontières des TADs soient enrichies en protéines architecturales ou insulatrices commes les cohésines ou CTCF et le fait que le contenu épigénomique des TADs soit relativement uniforme [Rao et al., 2014 ; Sexton et al., 2012 et Ho et al., 2014] suggèrent que les TADs ont un rôle important dans la régulation de l'expression des gènes. La régulation via les TADs peut par exemple se faire en favorisant des interactions promoteur/enhancer (ou amplificateur en français) [Lupiáñez et al., 2015].
De plus, les TADs eux mêmes s'organisent selon une hiérarchie de compartiments d'interaction de plus en plus grands, allant jusqu'aux territoires chromosomiques [Junier et al., 2015 ; Fraser et al., 2015 et Weinreb and Raphael, 2015]. Comprendre le rôle fonctionnel d'une telle compartimentation hiérarchique est un sujet de recherche actuel.

2.1.2 Méthodes de segmentations existantes

Différentes approches ont été développées afin de segmenter les cartes de contact. Une importante famille de méthodes se base sur une conversion des cartes de contact (données 2D) en un signal 1D présentant des extrema ou des variations brutales qui peuvent être associés à la présence d'une frontière de TAD [Dixon et al., 2012 ; Rao et al., 2014 ; Sexton et al., 2012 ; Crane et al., 2015 et Shin et al., 2016]. La méthode la plus utilisée est probablement

celle de Dixon et al. qui consiste à calculer un indice de directionnalité (ou « DI » de l'anglais Directionality Index). Cet indice donne pour chaque locus la différence de contacts entre amont et aval. Ainsi, l'indice de directionnalité subit une variation brutale à chaque frontière de TAD. Une autre méthode 1D (TopDom) est celle de Shin et al., qui consiste à trouver les minima locaux de la fréquence de contact moyenne dans le voisinage d'un locus [Shin et al., 2016].

D'autres approches utilisent un programme dynamique permettant de segmenter les chromosomes en TADs de manière optimale [Rao et al., 2014 ; Weinreb and Raphael, 2015 ; Filippova et al., 2014 et Lévy-Leduc et al., 2014]. Par exemple, HiCseg développé par Levy-Leduc et al. applique aux cartes Hi-C des techniques de segmentation 2D initialement utilisées en traitement de l'image. La carte Hi-C est alors segmentée en blocs diagonaux représentant les TADs [Lévy-Leduc et al., 2014].

En plus de déterminer la position des TADs, certaines méthodes renseignent sur les différents niveaux d'organisation de la chromatine [Junier et al., 2015 ; Fraser et al., 2015 ; Weinreb and Raphael, 2015 et Filippova et al., 2014]. Par exemple, TADtree, en approchant l'enrichissement des contacts par un modèle linéaire, infère la meilleure hiérarchie des TADs et permet la détection des sous TADs imbriqués dans les plus grands [Filippova et al., 2014].

Les différentes approches de segmentation mentionnées ci-dessus ont permis de mettre en évidence certaines caractéristiques de la chromatine, comme par exemple l'enrichissement des sites CTCF à la frontière des TADs chez les mammifères [Dixon et al., 2012] ou encore la caractérisation des régions inter-TAD chez la drosophile [Ulianov et al., 2016].

Toutefois, ces approches souffrent toutes d'un ou plusieurs inconvénients : (1) le programme de segmentation n'est pas téléchargeable ; (2) l'obtention d'une bonne segmentation nécessite d'ajuster finement des paramètres parfois nombreux et/ou parfois non intuitifs pour les non initiés ; (3) l'algorithme est coûteux en terme de temps de calcul ; (4) la robustesse des prédictions n'est pas estimée ; (5) la méthode infère la position des TADs mais pas la hiérarchie à différentes échelles.

Dans ce chapitre, on introduit IC-Finder qui est un algorithme de segmentation des cartes de contact en compartiments d'interaction (d'où l'acronyme IC, de l'anglais « Interaction Compartments »). Cet algorithme robuste et peu coûteux en terme de temps est basé sur une approche de type regroupement hiérarchique. Il dépend de deux paramètres intuitifs qui ne nécessitent pas un réglage de la part de l'utilisateur car les valeurs par défaut ont été apprises de sorte à ce que la segmentation soit optimale pour une large variété de cartes de contact expérimentales.

2.2 L'algorithme IC-Finder

IC-Finder est un programme permettant de segmenter des cartes Hi-C en compartiments d'interaction (Fig. 2.2.1A). Étant donnée une matrice de contact obtenue à partir d'une expérience de Hi-C, IC-Finder infère les frontières entre compartiments d'interaction consécutifs. L'algorithme est basé sur deux points (1) des locis appartenant à un même compartiment doivent avoir des interactions similaires avec le reste du génome ; (2) en raison de la nature polymérique intrinsèque à la chromatine, les interactions au sein d'un compartiment d'interaction doivent être « homogènes » au sens polymérique, c'est-à-dire que la fréquence de contact entre une paire de monomères ne doit seulement dépendre que de la distance génomique entre ces deux monomères. Afin de vérifier le premier point, on base IC-Finder sur un algorithme de type regroupement hiérarchique avec la contrainte que seul les plus proches voisins peuvent être groupés et avec une mesure de distance de corrélation afin de déterminer à chaque étape quelle paire de groupes doit être fusionnée. Concernant le deuxième point, on définit une condition d'arrêt à chaque étape : si l'hétérogénéité dans le groupe que l'on souhaite créer est trop grande, on ne fusionne pas les deux compartiments sélectionnés, au contraire, on fixe la frontière entre eux. À l'inverse, si l'hétérogénéité dans le groupe que l'on souhaite créer est faible, on accepte la fusion (Fig. 2.2.1A). Pour les cas intermédiaires des tests additionnels basés sur l'indice de directionnalité [Dixon et al., 2012] sont réalisés (Fig. 2.2.2). Les seuils définissant si l'hétérogénité est grande, petite ou intermédiaire sont appris à partir plusieurs cartes expérimentales dont la segmentation cible est déterminée manuellement (Fig. 2.2.4). À partir d'un ré-échantillonnage statistique de la carte de contact à segmenter, IC-Finder permet de quantifier la précision des compartiments trouvés (Fig. 2.2.1B). De plus, le programme offre la possibilité d'inférer l'organisation de la chromatine à différentes échelles (Fig. 2.2.1C). IC-Finder est libre d'accès, simple à utiliser et est disponible sur internet, `http://membres-timc.imag.fr/Daniel.Jost/DJ-TIMC/Software.html`.

2.2.1 Préliminaire : comparaison statistique entre deux partitions

Nous présentons dans cette section un outil permettant de mesurer la similarité entre deux segmentations car nous allons en avoir besoin dans ce chapitre 2 ainsi que dans le chapitre 3. Nous aurons par exemple besoin de comparer deux segmentations obtenues avec deux algorithmes différents (chapitre 2), ou encore besoin de comparer partition topologique et partition épigénomique (chapitre 3).

Soit deux partitions P_p et P_t du même ensemble, le taux de domaine vrai positif TPR_d (de l'anglais, True Positive Rate), aussi appelé sensibilité de la partition P_p par rapport à la partition P_t est défini comme étant la probabilité que deux loci quelconques appartenant au

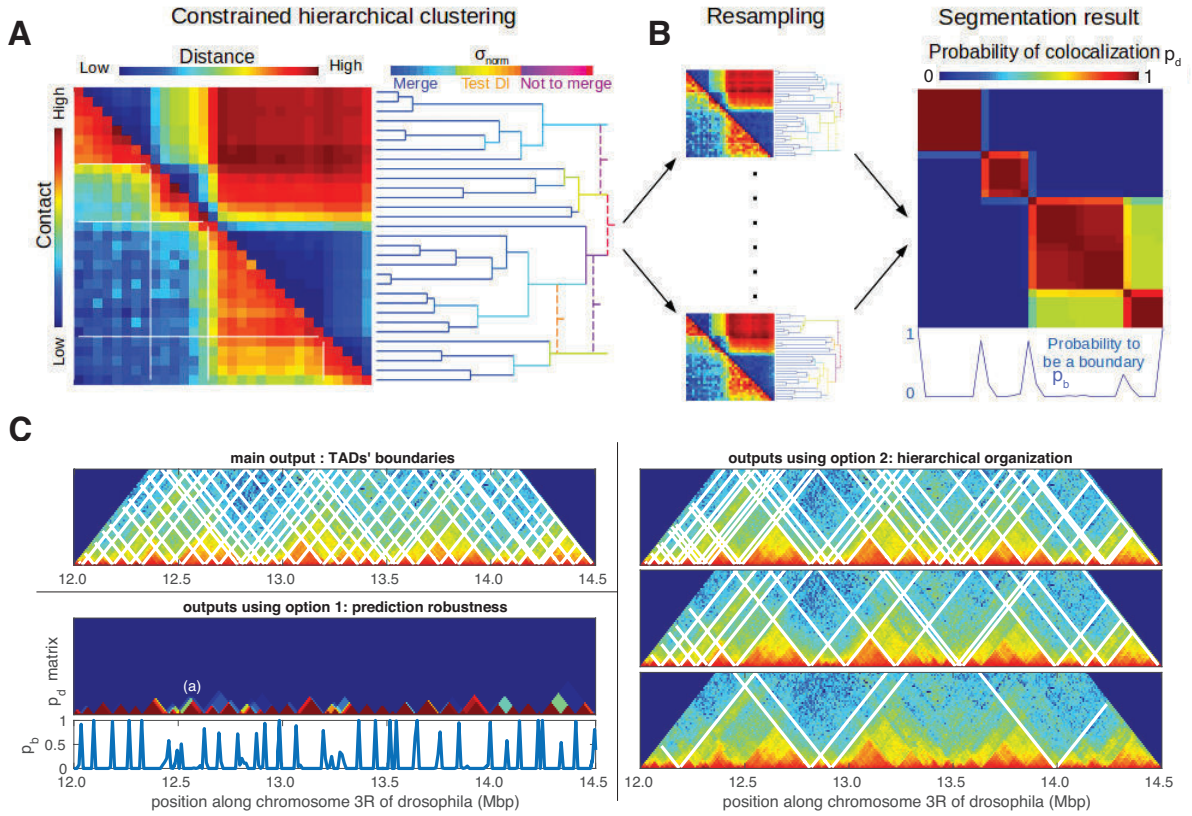

FIGURE 2.2.1 – **Vue d'ensemble d'IC-Finder.** (A) « Pipeline » de l'algorithme IC-Finder : une carte de contact (partie triangulaire inférieure) est transformée en carte de distance entre paires de colonnes (partie triangulaire supérieure). Ces distances permettent la construction de regroupements hiérarchiques ce qui est représenté par un dendrogramme (à droite de la carte). À chaque étape du regroupement, on calcule σ_{norm} qui mesure l'hétérogénéité dans le futur groupe à éventuellement former par la fusion de deux sous groupes. Si $\sigma_{\mathrm{norm}} < \sigma_-$, les deux sous groupes fusionnent pour former un nouveau groupe, si $\sigma_{\mathrm{norm}} > \sigma_-$ la frontière entre les deux sous groupes est fixée jusqu'à la fin. Si $\sigma_{\mathrm{norm}} \in [\sigma_-, \sigma_+]$, des tests additionnelles basés sur l'indice de directionnalité (Éq. 2.11) sont réalisés (Fig. 2.2.2). (B) La robustesse des prédictions est estimée en ré-échantillonnant la carte de contact de départ selon une distribution de Poisson. On applique l'algorithme de segmentation IC-Finder à ces cartes ré-échantillonnées, ce qui permet de calculer $\mathtt{p}_{\mathtt{d}}$ la probabilité que deux loci appartiennent au même compartiment d'interaction et $\mathtt{p}_{\mathtt{b}}$, la probabilité qu'un loci soit prédit comme étant une frontière (indice $\mathtt{b}$ pour boundary en anglais). (C) Différents types de résultats donnés par IC-Finder : (À gauche) IC-Finder est lancé en mode par défaut, on obtient de haut en bas, les frontières des TADs, la matrice des probabilités $\mathtt{p}_{\mathtt{d}}$ et probabilité $\mathtt{p}_{\mathtt{b}}$ en fonction du locus. (À droite) Résultats donnés par IC-Finder pour différentes hiérarchies (de bas en haut, $\sigma_- = \sigma_+ = 2$; 5 et 10). Les exemples présentés ici correspondent à la région entre 12 et 14.5Mb du chromosome 3R de la drosophile.

même domaine dans P_t, appartiennent aussi au même domaine dans P_p :

$$TPR_d\left(P_p||P_t\right) = \frac{\sum_{i<j} \delta_{ij}\eta_{ij}}{\sum_{i<j} \eta_{ij}} \tag{2.1}$$

avec $\delta_{ij} = 1$ (resp. 0) si i et j appartiennent (resp. ou pas) au même groupe dans la partition P_p et idem pour η_{ij} relativement à la partition P_t.
Le taux de fausses découvertes de domaine, FDR_d (de l'anglais, False Discovery Rate) de la partition P_p relativement à la partition P_t est défini comme étant la probabilité que deux loci appartiennent au même domaine dans la segmentation P_p, sachant qu'ils n'appartiennent pas au même domaine dans P_t :

$$FDR_d\left(P_p||P_t\right) = \frac{\sum_{i<j} \delta_{ij}\left(1-\eta_{ij}\right)}{\sum_{i<j} \delta_{ij}} \tag{2.2}$$

Dans le cas de deux segmentations parfaitement identiques, $TPR = 1$ et $FDR = 0$.
De la même manière, on définit le taux TPR_b (b comme boundary en anglais) comme étant la probabilité qu'un locus soit une frontière dans la partition P_p ($\pm$ 1 bin) sachant que ce locus correspond à une frontière dans la partition P_t. On définit aussi le taux de fausses découvertes de frontière FDR_b comme étant la probabilité qu'un loci soit une frontière dans P_p, sachant qu'il ne l'est pas dans P_t.

2.2.2 Regroupement hiérarchique contraint

À partir d'une carte de contact C, de taille $N \times N$, on cherche à regrouper les loci consécutifs le long du génome qui partagent le même motif d'interaction. En analyse de données, parmi les algorithmes de classification non supervisée, une méthode standard efficace est le regroupement hiérarchique. Le regroupement hiérarchique est une méthode itérative consistant à regrouper des objets deux par deux afin d'obtenir une hiérarchie de groupes. Au début de l'algorithme, chaque colonne de la matrice C représente un groupe. Les deux groupes les plus « similaires » sont fusionnés. Ce processus est répété jusqu'à ce qu'il n'y ait plus qu'un seul groupe. Dans la méthode standard de regroupement hiérarchique, n'importe quelle paire de groupe peut être fusionnée. Dans notre cas, afin de respecter la connectivité linéaire des chromosomes, on contraint l'algorithme à ne regrouper que des classes plus proches voisines le long du génome. La « similarité » S entre deux groupes composés chacun d'une colonne uniquement se définit simplement par la distance D entre ces deux colonnes. Cette distance entre colonnes peut être définie selon différentes métriques. On testera la distance de Manhattan (notée L1, Éq. 2.3), la distance euclidienne (notée L2, Éq. 2.4) et la distance de corrélation

obtenue en divisant la covariance entre les deux colonnes par le produit de leur déviation standard (Éq. 2.5).

$$D_{L1}(C_i,\ C_k) = \sum_{j=1}^{N} |C_{ij} - C_{kj}| \tag{2.3}$$

$$D_{L2}(C_i,\ C_k) = \sqrt{\sum_{j=1}^{N} (C_{ij} - C_{kj})^2} \tag{2.4}$$

$$D_C(C_i,\ C_k) = 1 - \frac{cov(C_i,\ C_k)}{\sqrt{Var(C_i)\,Var(C_k)}} \tag{2.5}$$

avec $cov(C_i,\ C_k)$ la covariance entre les colonnes C_i et C_k de la matrice C et avec $Var(C_i)$ la variance de la colonne C_i.

Lorsque deux groupes plus proches voisins G_1 et G_2 sont composés de plusieurs colonnes, il existe de multiples stratégies qui permettent de calculer la similarité entre ces deux groupes, notée $S(G_1,\ G_2)$. On testera les mesures de similarité S_{min}, S_{max}, $S_{moyenne}$ et $S_{moyenneP}$ qui retiennent respectivement le minimum, le maximum, la moyenne et la moyenne pondérée des distances entre les colonnes de G_1 et les colonnes de G_2 (resp. Éqs. 2.6 ; 2.7 ; 2.8 et 2.9).

$$S_{min}(G_1,\ G_2) = \min_{C_i \in G_1, C_k \in G_2} (D(C_i,\ C_k)) \tag{2.6}$$

$$S_{max}(G_1,\ G_2) = \max_{C_i \in G_1, C_k \in G_2} (D(C_i,\ C_k)) \tag{2.7}$$

$$S_{moyenne}(G_1,\ G_2) = \underset{C_i \in G_1, C_k \in G_2}{\text{moyenne}} (D(C_i,\ C_k)) \tag{2.8}$$

$$\begin{aligned} S_{moyenneP}(G_1,\ G_2) &= \underset{C_i \in G_1, C_k \in G_2}{\text{moyenneP}} (D(C_i,\ C_k)) \\ &= \frac{\sum_{C_i \in G_1, C_k \in G_2} (N - |C_i - C_k|)^2 D_c(C_i, C_k)}{\sum_{C_i \in G_1, C_k \in G_2} (N - |C_i - C_k|)^2} \end{aligned} \tag{2.9}$$

avec N la taille de la matrice C et avec $D(C_i,\ C_k)$ la distance entre les colonnes C_i et C_k pouvant être calculée avec les métriques D_{L1}, D_{L2} ou D_c (Éqs. 2.3 ; 2.4 et 2.5). La moyenne pondérée est construite de telle sorte à donner plus d'importance aux distances calculées entre colonnes proches.

N.B. : Il arrive souvent que sur les cartes Hi-C le nombre de contacts soit nul. Ce zéro peut représenter un manque d'information ou bien une réelle absence de contacts. Ne pouvant pas distinguer ces deux cas, on décide d'ignorer les lignes et colonnes avec plus de 75% de zéros

parmi les 20 coefficients encadrant la diagonale. Concrètement, ces lignes et colonnes sont supprimées de la carte avant de commencer le processus de segmentation. De plus, on décide que pour que le calcul d'une distance entre deux colonnes soit valide, il faut qu'il y ait au minimum un chevauchement de 10 valeurs non nulles entre ces deux colonnes.

2.2.3 Détermination de la segmentation optimale

Une difficulté importante du regroupement hiérarchique est de trouver un critère systématique permettant de sélectionner la segmentation la plus judicieuse parmi toutes les possibilités hiérarchiques inférées par la méthode. Ici, on base notre critère sur l'observation que les contacts dans un compartiment doivent être « homogènes » au sens polymérique, c'est-à-dire que la fréquence de contact entre paires de monomères doit seulement dépendre de la distance génomique entre les monomères. En pratique, pour chaque étape du regroupement hiérarchique, on définit une condition d'arrêt afin d'évaluer si deux groupes sélectionnés pour être fusionnés doivent effectivement l'être ou pas. Si c est une sous matrice de C composée des deux groupes voisins sélectionnés (Fig. 2.2.2A), on définit la matrice normalisée c_n :

$$c_n(i,j) = \frac{c(i,j)}{\bar{c}\left(|i-j|\right)}$$

avec

$$\bar{c}(k) = \frac{\sum_{|i-j|=k} c(i,j)}{\sum_{|i-j|=k} 1}$$

On estime alors la variance σ des éléments de c_n :

$$\sigma = \mathrm{var}\left(c_n\right) \tag{2.10}$$

Afin de corriger les effets engendrés par le bruit expérimental, on normalise σ par la variation locale médiane de c_n dans la région proche de la diagonale (pour des distances génomiques comprises entre 3 et 40 bins). On obtient alors le paramètre σ_{norm}. Si $\sigma_{norm} < \sigma_-$ (faible hétérogénéité), on fusionne les deux groupes sélectionnés. Si, au contraire, $\sigma_{norm} > \sigma_+$ (forte hétérogénéité), on fixe la frontière entre les deux groupes sélectionnés. Afin d'obtenir de bonnes prédictions et d'éviter les nombreux faux positifs (Fig. 2.2.2B), au lieu d'avoir seulement un seuil séparant les faibles des fortes variances, on introduit une zone tampon $[\sigma_-,\ \sigma_+]$ au sein de laquelle des tests supplémentaires sont réalisés (Fig. 2.2.2A). Dans cette zone tampon, un indice de directionnalité [Dixon et al., 2012] est calculé autour de la frontière dont on cherche à savoir si elle doit être supprimée ou au contraire fixée.

$$DI = sign\,(B - A)\,\frac{(B-A)^2}{(B+A)} \tag{2.11}$$

avec A (resp. B) le nombre de contacts que fait un locus avec ses voisins de la région testée en amont (resp. en aval) (Fig. 2.2.2A).

La frontière restera fixe (i) si au moins 2/3 des indices de directions avant (resp. après) la frontière sont négatifs (resp. positifs), (ii) si la variation de l'indice de directionnalité est positive au moment du passage à la frontière et (iii) si pour au moins 2/3 des bins la variation relative $2 \times {}^{(A-B)}/_{(A+B)}$ est supérieure à 0.1. Cette condition (iii) a été introduite pour augmenter la robustesse de l'algorithme vis-à-vis du bruit.

2.2.4 Détermination des paramètres pour IC-Finder

2.2.4.1 Métrique et mesure de la similarité entre groupes

Ci-dessus, nous avons présenté trois métriques différentes et quatre moyens de mesurer la similarité entre groupes, soit un total de douze stratégies possibles pour réaliser le regroupement hiérarchique. On décide de comparer les résultats obtenus avec ces douze stratégies sur six cartes expérimentales [Dixon et al., 2012 ; Rao et al., 2014 et Sexton et al., 2012] chacune de taille 1500x1500 (en bin) que nous segmentons manuellement. On compare sur la figure 2.2.3 les douze segmentations obtenues à la segmentation cible (faite à la main) en étudiant le taux de vrais positifs (Éq. (2.1)) en fonction du taux de fausses découvertes (Éq. 2.2). On observe que les mesures de similarité réalisées avec $S_{moyenneP}$ (Éq. 2.9) couplé à D_c (Éq. 2.5) donne le meilleur équilibre entre spécificité (TPR) et sensibilité (FDR) pour les partitions prédites.

2.2.4.2 Optimisation des paramètres σ_- et σ_+

Nous avons optimisé les valeurs seuil σ_- et σ_+ à partir de portions de cartes Hi-C segmentées manuellement (portions de 1500×1500 bins du chromosome 3R d'embryons tardifs de drosophile avec une résolution de 10kb [Sexton et al., 2012, GSE34453] ; du chromosome 12 de la lignée cellulaire humaine IMR90 avec une résolution de 40kb [Dixon et al., 2012, GSE35156] et du chromosome 3 de la lignée cellulaire humaine GM12878 avec une résolution de 10kb [Rao et al., 2014, GSE63525]) (Fig. 2.2.4) de telle sorte à avoir la meilleure correspondance entre les prédictions de IC-Finder et la segmentation manuelle.

N.B. : Les code GSE indiqués sont les numéros d'accès aux données à partir du site internet « Gene Expression Omnibus » [Barrett et al., 2013].

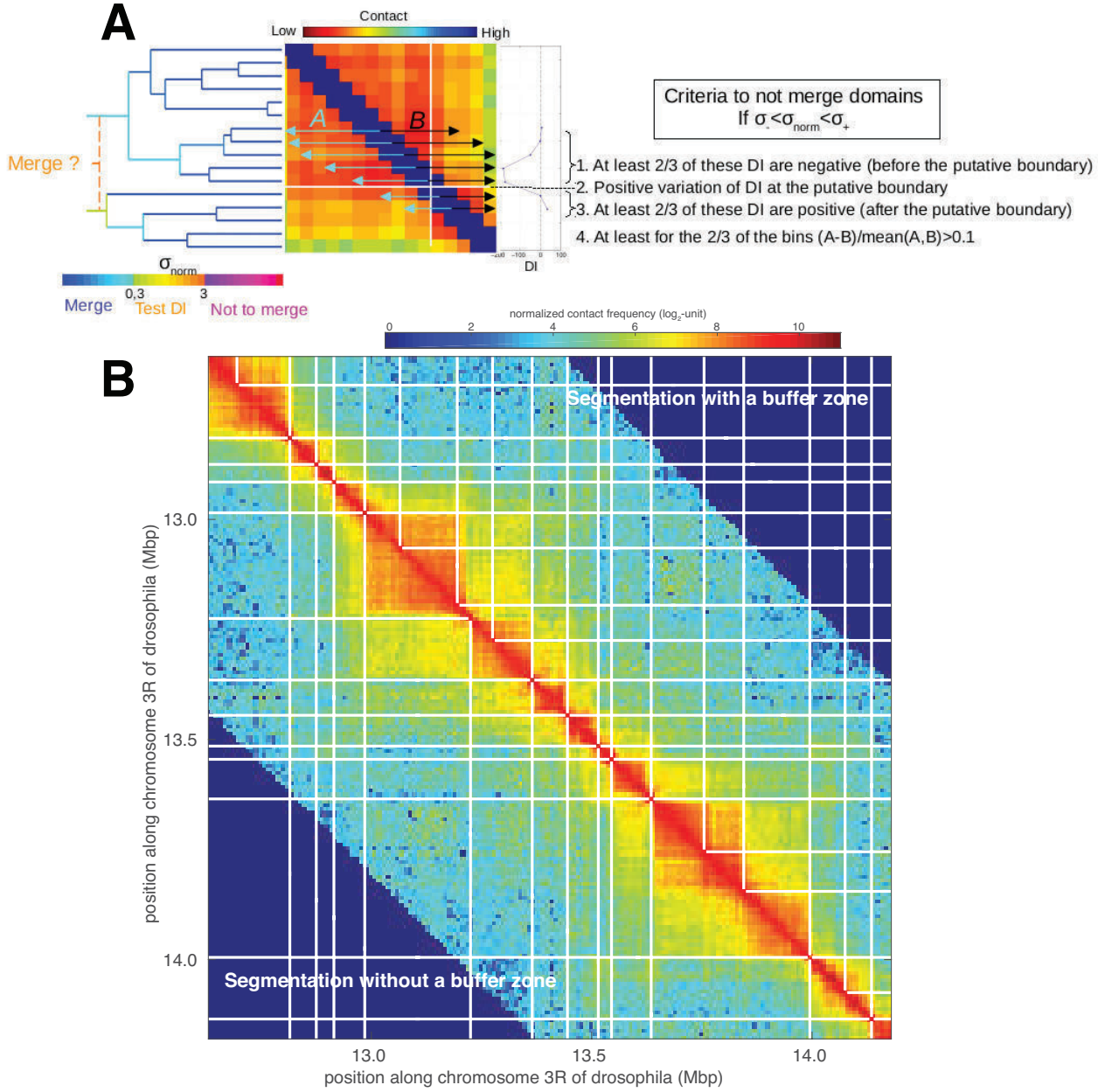

Figure 2.2.2 – **Test complémentaire avant de supprimer une frontière dans la zone tampon où** $\sigma_{norm} \in [\sigma_-, \sigma_+]$. (A) Matrice $\mathbf{c}$ (sous matrice de $\mathbf{C}$) à partir de laquelle des tests additionnels basés sur l'indice de directionnalité (Éq. 2.11) sont réalisés si $\sigma_{norm} \in [\sigma_-, \sigma_+]$. (B) Illustration de l'amélioration d'une segmentation en introduisant la zone tampon $[\sigma_-, \sigma_+] = [0.3, 3]$ par rapport à un unique seuil $\sigma_- = \sigma_+ = 1.5$. (Les cartes sont tronquées car les données publiées ne donnent pas les contacts au delà de 800kb [Sexton et al., 2012]).

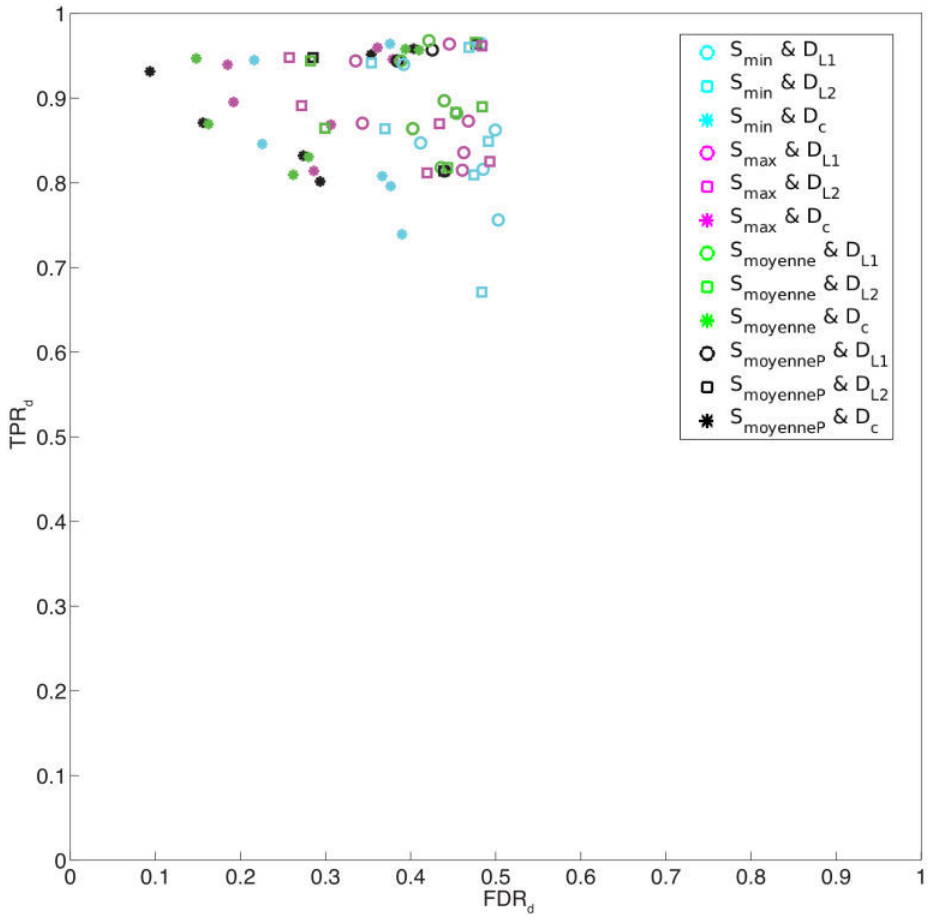

FIGURE 2.2.3 – Taux de domaines vrai positif ($\mathsf{TPR_d}$) en fonction du taux de fausses découvertes ($\mathsf{FDR_d}$) prédit par l'algorithme de regroupement hiérarchique avec différentes stratégies de mesure de similarité entre deux groupes. Les tests ont été réalisés à partir de six cartes Hi-C expérimentales dont la segmentation cible a été faite manuellement.

L'optimisation a mené aux paramètres $\sigma_- = 0.3$ et $\sigma_+ = 3$ qui sont donc les valeurs par défaut dans IC-Finder. Notons toutefois que ces deux paramètres peuvent être ajustés par l'utilisateur d'IC-Finder afin d'éventuellement améliorer la segmentation pour certaines régions génomiques spécifiques.

2.2.5 Les options d'IC-Finder : Ré-échantillonnage et Hiérarchie

2.2.5.1 Robustesse des prédictions

Sur les cartes Hi-C, de nombreuses valeurs de contact sont faibles (entre zéro et quelques milliers pour une résolution de 10kb) ce qui fait que les cartes Hi-C sont fortement bruitées (voir plus loin section 3.1.1). Afin d'estimer comment les incertitudes sur les nombres de contact se propagent sur les segmentations prédites, IC-Finder réalise plusieurs ré-échantillonnages

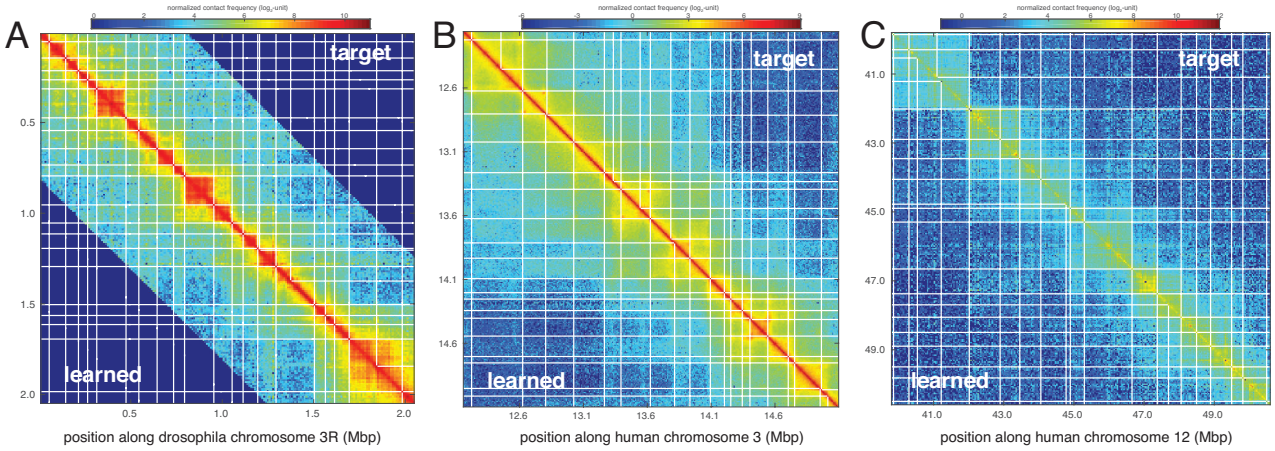

Figure 2.2.4 – **Apprentissage des paramètres σ_- et σ_+ à partir de cartes de contact expérimentales.** Ces figures comparent les prédictions d'IC-Finder avec les paramètres par défaut $\sigma_- = 0.3$ et $\sigma_+ = 3$ (partie triangulaire inférieure) aux segmentations obtenues manuellement (partie triangulaire supérieure). Ces segmentations manuelles ont été utilisées comme cibles pour l'apprentissage des paramètres σ_- et σ_+. Les cartes de contact présentées ici correspondent à trois régions génomiques différentes, à gauche, données de drosophile [Sexton et al., 2012], au milieu et à droite, données humaines [Rao et al., 2014 et Dixon et al., 2012].

(100 par défaut) de la carte de contact initiale et applique l'algorithme de segmentations à ces nouvelles cartes (Fig. 2.2.1B). En combinant les résultats obtenus, on peut calculer la probabilité $p_d(i,j)$ que deux loci appartiennent au même compartiment d'interaction et la probabilité $p_b(i)$ pour un locus i d'être une frontière entre deux domaines. Cette option permet donc de quantifier la précision des prédictions de IC-finder. Au final, bien que cette option demande du temps de calcul (100 cartes à segmenter plutôt qu'une seule), elle donne une image plus claire et plus fiable de la compartimentation. Par exemple, le compartiment d'interaction (a) sur la figure 2.2.1C est bien défini mais la position de sa frontière gauche est floue. La connaissance de ces caractéristiques est bien sûr cruciale lorsque l'on cherche à comparer des observables telles que la position d'une frontière ou bien la colocalisation de deux loci dans un même compartiment à des informations génomiques ou épigénomiques dépendant de la position.

Pour réaliser le ré-échantillonnage de la carte de contact étudiée, on commence par la multiplier par un facteur constant $f = \frac{N_c}{\sum_i C(i,i+1)}$ avec N_c le nombre total de contact sur la première diagonale des données brutes pour la même région génomique (c'est-à-dire avant tout processus de normalisation). Cette étape de multiplication par le facteur f est importante puisqu'elle permet de normaliser les données en terme de contact. Après cette multiplication, on modélise les fréquences de contact de cette matrice modifiée comme des processus de Pois-

son indépendants. Ainsi, pour chaque paire de deux loci (i, j), on tire aléatoirement selon une distribution de Poisson de moyenne $f \times C(i, j)$ une nouvelle fréquence de contact.
Si le nombre N_c est inconnu de l'utilisateur, IC-Finder utilise une valeur par défaut, à savoir, $N_c = 904$ ce qui représente la valeur typique calculée à partir de récentes expériences de Hi-C résolues à 10kb [Rao et al., 2014 et Sexton et al., 2012].

2.2.5.2 Organisation hiérarchique

En plus de la partition élémentaire du génome en TADs [Dixon et al., 2012 et Rao et al., 2014], les expériences de Hi-C révèlent clairement l'existence de niveau d'organisation supérieurs ce qui souligne le caractère hiérarchique du repliement de la chromatine : les TADs consécutifs s'organisent en compartiments d'interaction plus grands qui eux mêmes forment de plus grands groupes [Junier et al., 2015 ; Fraser et al., 2015 ; Weinreb and Raphael, 2015]. En permettant de varier le degré minimal d'hétérogénéité nécessaire pour fusionner ou pas deux groupes adjacents, via le paramètre $\sigma_0 = \sigma_- = \sigma_+$, IC-Finder offre la possibilité d'étudier l'organisation hiérarchique de la chromatine. La figure 2.2.1C illustre la segmentation hiérarchique inférée par IC-Finder avec $\sigma_0 = 2$; 5 et 10 pour la région se situant entre 12 et 14.5 Mb du chromosome 3R de cellules d'embryons tardifs de drosophile [Sexton et al., 2012]. Il est possible de coupler cette information sur la hiérarchie et le ré-échantillonnage présentés ci-dessus, de manière à avoir une image complète et fiable du repliement hiérarchique de la chromatine (Fig. 2.3.5).

2.3 Résultats

2.3.1 Fiabilité des segmentations obtenues avec IC-Finder et comparaison avec d'autres méthodes

2.3.1.1 Préliminaire : Génération des benchmarks

Afin de tester l'algorithme, nous avons construit un ensemble de matrices de type cartes Hi-C dont la segmentation optimale est connue. Ces cartes ont été simulées à partir d'un modèle polymérique capable de décrire semi-quantitativement la formation et la dynamique des TADs en se basant seulement sur l'information épigénomique. Ce modèle qui est décrit en détail dans le chapitre 4 nous a permis de simuler 100 cartes de contact de référence à partir de 100 séquences épigénomiques différentes. Nous avons pris soin de choisir les séquences épigénomiques et les paramètres d'interaction de telle sorte à reproduire les motifs typiques observés sur les cartes expérimentales de drosophile (TADs le long de la diagonale,

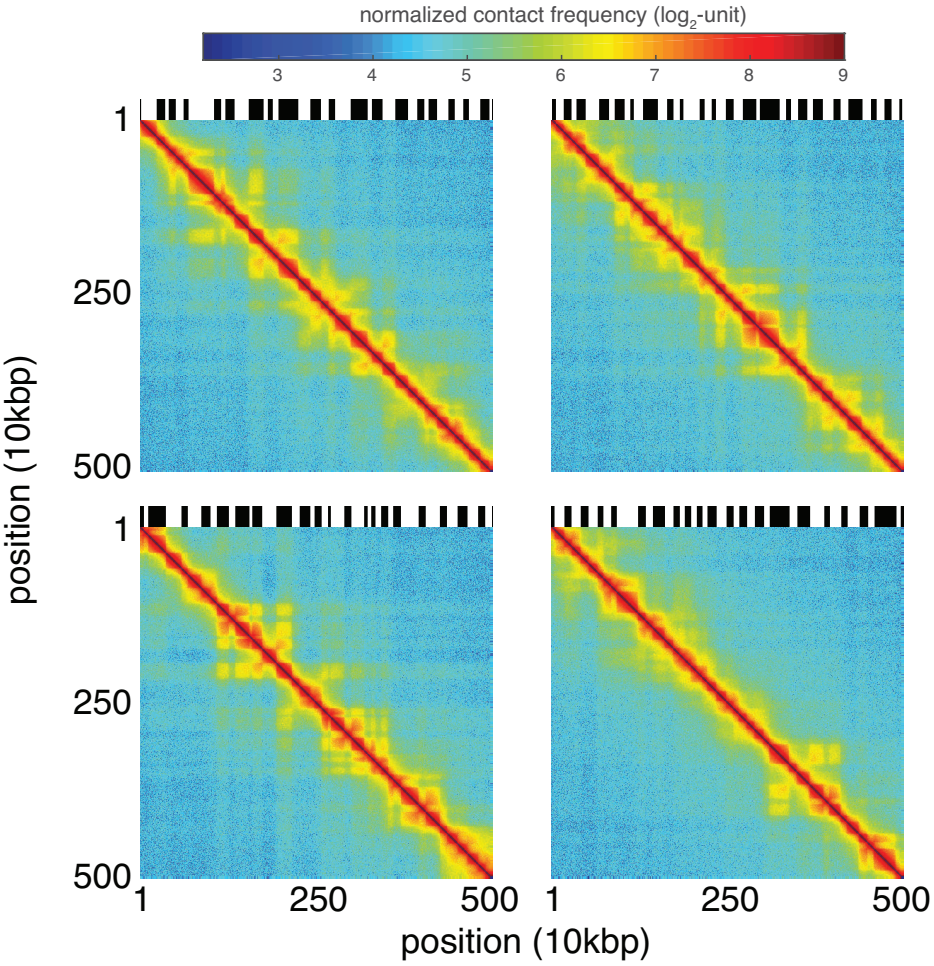

Figure 2.3.1 – **Quatre exemples de matrices de type cartes Hi-C , de taille** 500×500**, simulées avec le modèle polymérique développé dans le chapitre 4.** Au dessus de chaque carte se trouve la séquence épigénomique (en noir et blanc) à partir de laquelle la carte a été simulée. Cette séquence correspond à la segmentation irréductible de la carte. Les quatre matrices présentées ici appartiennent à un ensemble de 100 cartes de référence que nous avons construit.

interactions à longue portée et décroissance de la probabilité de contact avec une loi d'échelle du type s^{-1} avec s la distance génomique entre deux loci.

Ces cartes de contact obtenues in silico ont ensuite été bruitéee selon une distribution de Poisson de manière à simuler les variations locales observées sur les cartes expérimentales (Fig. 2.3.1).

2.3.1.2 Pouvoir de prédiction des TADs pour différentes méthodes de segmentation sur un ensemble de 100 cartes de référence

Afin de valider notre approche et afin de comparer IC-Finder à d'autres méthodes de segmentation, nous allons utiliser nos 100 cartes Hi-C de référence simulées de telle sorte à reproduire les motifs typiques des cartes Hi-C expérimentales et pour lesquelles nous connaissant la segmentation cible à viser (cf. section 2.3.1.1). Ces cartes de contact obtenues in silico n'ont pas été utilisées pour calibrer les paramètres par défaut utilisés par la méthode.

Pour chaque carte de contact on calcule les taux de vrais positifs (TPR) et taux de fausses découvertes (FDR) de la segmentation prédite par rapport à la segmentation cible. Un parfaite concordance entre prediction et cible mène à TPR=1 et FDR=0. Sur la figure 2.3.2A,B (région rouge), on représente les lignes de contour qui englobe 98% des 100 points (TPR, FDR). Ce graphe montre que notre algorithme avec les paramètres par défaut donne d'excellents résultats avec une très bonne sensibilité et un taux de fausses découvertes très faible. De plus, IC-Finder prédit presque parfaitement la distribution des tailles des domaines (Fig. 2.3.2C). Ensuite, on compare la performance d'IC-Finder à d'autres algorithmes permettant la segmentation de cartes Hi-C en TADs (Armatus [Filippova et al., 2014], Directionality Index [Dixon et al., 2012], HICSeg [Lévy-Leduc et al., 2014], Insulation method [Crane et al., 2015], TADtree [Weinreb and Raphael, 2015] and TopDom [Shin et al., 2016]). Pour chaque méthode, exceptés IC-Finder et Armatus, nous avons ajusté manuellement les paramètres (parfois nombreux) pour avoir une segmentation optimale (Tab. 2.3.3). On observe que TopDom et IC-Finder donnent des résultats similaires et surpassent les autres méthodes en terme de fiabilité (Fig. 2.3.2). En terme d'efficacité numérique, IC-Finder est plus rapide que les autres méthodes pour des cartes de grande taille (Fig. 2.3.2D). Par exemple, pour un ordinateur cadencé à 3GHz, il faut 4 minutes pour segmenter le génome humain entier (avec une résolution de 40kb), avec TopDom il faut 6 minutes, avec l'indice de directionnalité, il faut plus de 3 heures. En conclusion, la famille des cent cartes de référence a permis de montrer que IC-Finder est l'une des meilleures méthodes de segmentation en terme de fiabilité avec une détection des TADs à la fois sensible et spécifique et en terme de temps de calcul.

2.3.1.3 Pouvoir de prédiction des TADs pour différentes méthodes de segmentation sur un ensemble de cartes expérimentales

En guise de deuxième test de comparaison, on applique IC-Finder sur des cartes de contact expérimentales utilisées par les autres méthodes pour illustrer leur pouvoir de prédiction (cellules souche embryonnaire de souris et lignée cellulaire humaine IMR90 [Dixon et al., 2012]). La position des TADs obtenues avec les différentes méthodes sont téléchargés sur

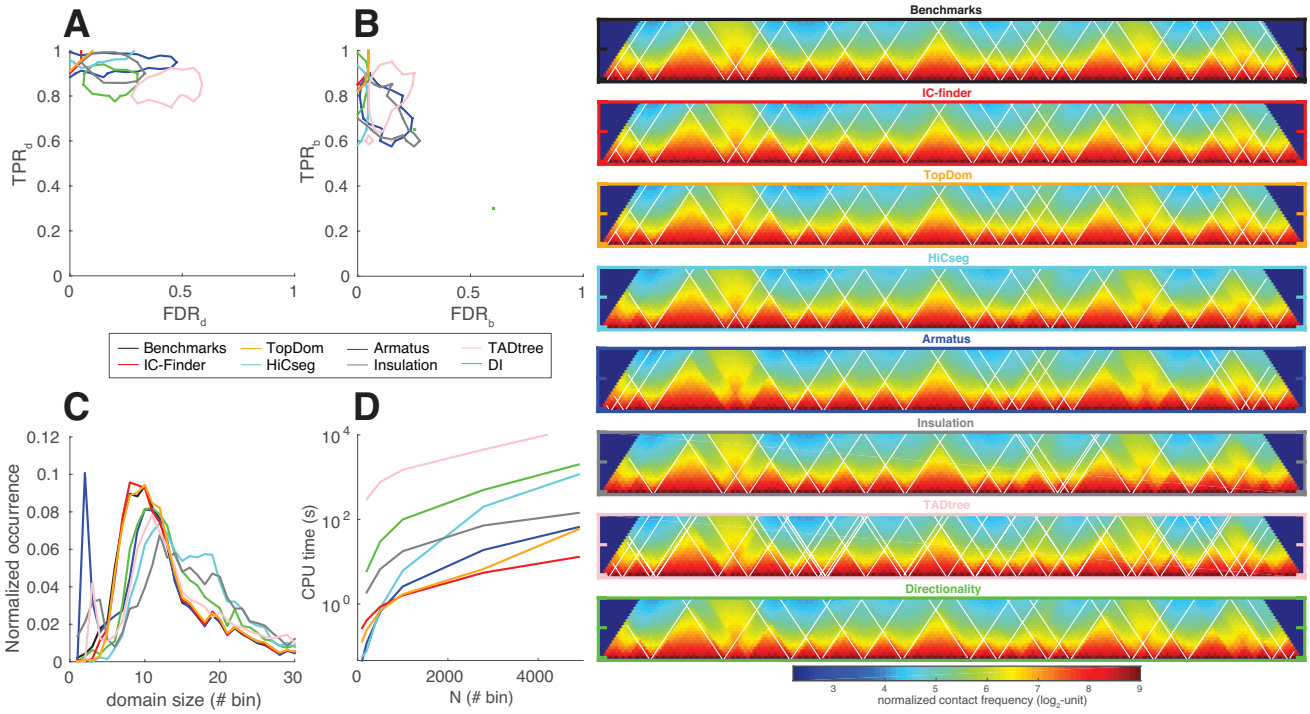

FIGURE 2.3.2 – **Comparaison statistique entre segmentations cibles et segmentations prédites avec IC-Finder et d'autres méthodes sur un ensemble de 100 cartes de référence.**(A) Probabilité que deux loci soient correctement classés dans le même domaine, $\mathsf{TPR_d}$, en fonction de la probabilité que deux loci soient classés à tort dans le même domaine, $\mathsf{FDR_d}$ calculés à partir de sept programme de segmentation explicités en légende. Pour chaque programme, on représente la ligne de contour qui englobe 98% des 100 points (TPR, FDR) obtenus à partir des 100 cartes. (B) Similaire à (A) sauf qu'on s'intéresse aux prédictions sur la position des frontières, $\mathsf{TPR_b}$ en fonction de $\mathsf{FDR_b}$. (C) Distribution des tailles des domaines inférés avec les différentes méthodes. La ligne noire concerne la segmentation cible. (D) Temps CPU nécessaire pour segmenter une carte de contact de taille $\mathsf{N} \times \mathsf{N}$ en fonction de N (temps mesurés avec un ordinateur portable classique). À droite : En haut segmentation cible et en bas exemples de segmentations prédites avec les différents programmes testés.

Softwares	Parameters
Armatus	GammaMax=0,5 (highest resolution at which domains are to be generated)
Directionality	**Window_size=1e5** (how far we need to look at the interaction patterns of a given bin in bp) **M=1:7** (number of mixtures in the HMM)
HiCseg	**nb_change_max=250** (maximal number of change-points) distrib= "G" (distribution of the data) model="D" (model Type)
IC-finder	th='Default' (threshold below which domains are merged in hierarchical clustering)
Insulation	**is=100000** (insulation square in bp) **ids=80000** (insulation delta window) im=mean (how to aggregrate signal within insulation square) **bmoe=0** (boundary margin of error added to each side of the boundary in bins) nt=0.1 (noise threshold, minimum depth of valley)
TADtree	S = 50 (max. size of TAD in bins) **M = 5** (max. number of TADs in each tad-tree) p = 3 (boundary index parameter) q = 12 (boundary index parameter) gamma = 500 (balance between boundary index and squared error in score function) N = 80 (list of numbers of TADs to use)
TopDom	**Window.size=5** (number of bins to extend)

FIGURE 2.3.3 – **Logiciels testés et paramètres correspondants utilisés pour prédire la segmentation sur nos 100 cartes de contact de références.** Les paramètres en gras sont différents des paramètres par défaut, nous les avons ajustés de manière à optimiser les segmentations.

les sites internet `http://chromosome.sdsc.edu/mouse/hi-c/download.html` [Dixon et al., 2012] ; `https://www.cs.cmu.edu/~ckingsf/software/armatus/` [Filippova et al., 2014] et `http://zhoulab.usc.edu/TopDom/` [Shin et al., 2016].

D'un point de vue globale, IC-Finder et TopDom donnent des résultats similaires avec des distributions de taille de TADs comparables (Fig. 2.3.4B,D). Concernant Armatus, ce programme a tendance à trouver de nombreux petits domaines. L'indice de directionnalité quant à lui segmente les cartes en domaines plus grands (Fig. 2.3.4A,C). Localement, IC-Finder est aussi proche de TopDom avec une meilleure correspondance à l'échelle des domaines plutôt qu'à l'échelle des frontières (Fig. 2.3.4E). Ceci est une conséquence du bruit important présent sur les cartes de contact étudiées et qui mène donc à une définition floue des frontières.

2.3.2 Amélioration de la fiabilité des prédictions

Les expériences de Hi-C sont soumises à différentes sources d'erreurs, ce qui affecte les cartes Hi-C produites. En particulier, les erreurs d'échantillonnage dues au nombre fini de cellules utilisées dans les expériences combinées à la faible efficacité du protocole Hi-C dans son ensemble sont susceptible de provoquer d'importantes fluctuations sur la carte, et spécialement

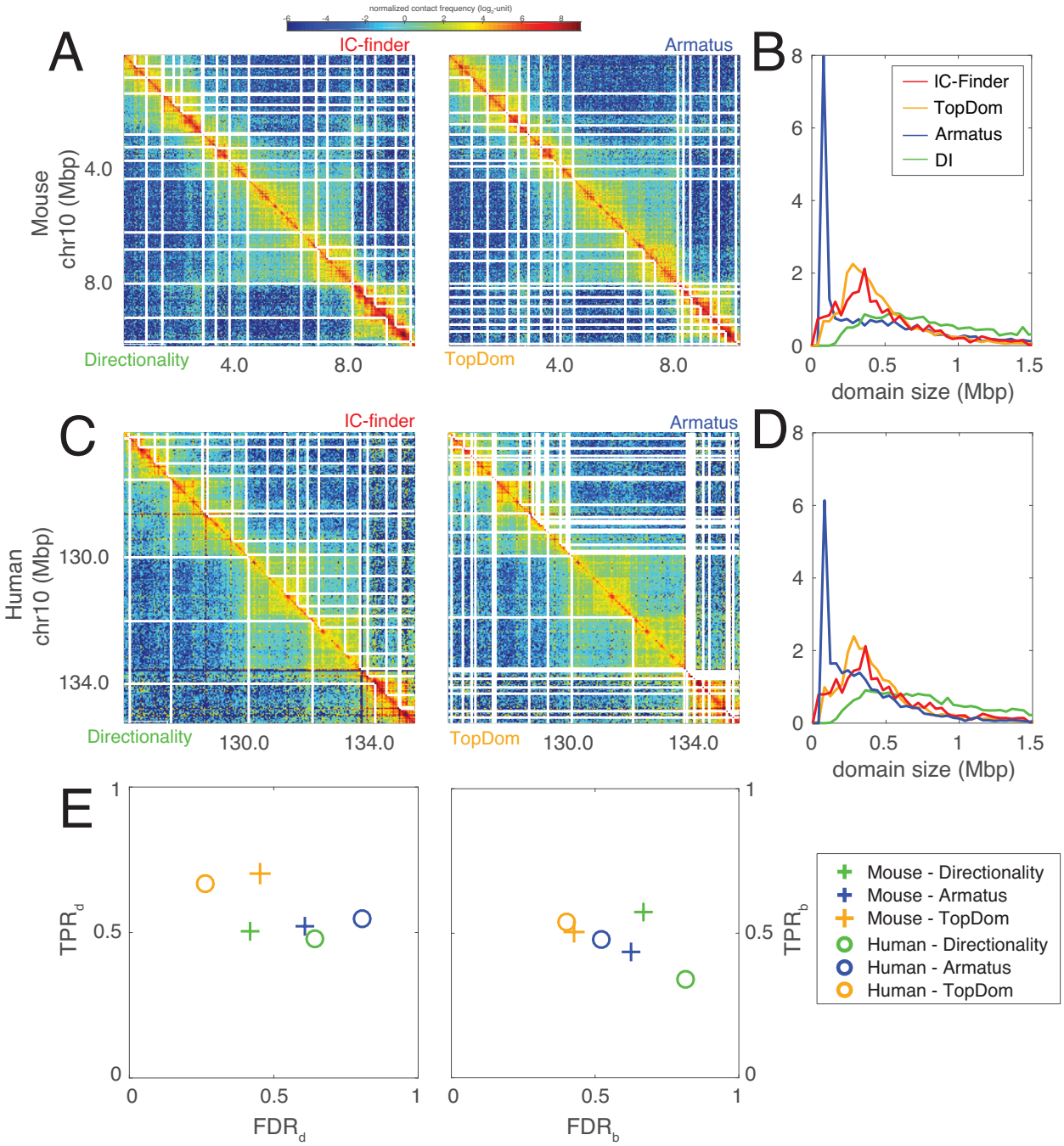

Figure 2.3.4 – **Comparaison statistique entre segmentation cible et segmentations prédites avec IC-Finder et d'autres méthodes sur des cartes expérimentales.**(A,C) Exemples de segmentations obtenues avec différentes méthodes dans le cas de cellules souches embryonnaires de souris (A) et pour la lignée cellulaire humaine IMR90 (D) [Dixon et al., 2012]. (B,D) Distribution des tailles des domaines pour l'ensemble du génome prédites avec différentes méthodes (E) TPR en fonction de FDR pour IC-Finder en fonction des partitions obtenues avec les autres méthodes.

pour les bins avec peu de contacts. Afin d'estimer à quel point les erreurs d'échantillonnage se propage sur la segmentation prédite, IC-Finder propose en option de réaliser un ré-échantillonnage statistique de la carte de contact à segmenter. Ceci va permettre d'estimer la probabilité que deux loci soient classés dans le même domaine topologique et la probabilité pour un loci qu'il corresponde à une frontière (Fig. 2.2.1B). Cette vision probabiliste des domaines est plus claire et plus juste puisqu'elle permet de quantifier à quel point les prédictions données par IC-Finder sont précises. Par exemple, le compartiment d'interaction (a) que l'on peut voir sur la figure 2.2.1C est bien défini mais la position de sa frontière gauche est floue. Connaître ces incertitudes est bien sûr crucial si on cherche à comparer les informations sur les TADs à d'autres observables qui sont fonction de la position, comme la séquence épigénomique par exemple.

2.3.3 Inférence de l'organisation hiérarchique de la chromatine

En plus de la partition élémentaire du génome en TADs [Dixon et al., 2012 et Rao et al., 2014], les expériences de Hi-C révèlent clairement l'existence de niveau d'organisation supérieurs ce qui souligne le caractère hiérarchique du repliement de la chromatine : les TADs consécutifs s'organisent en compartiments d'interaction plus grands qui eux mêmes forment de plus grands groupes [Junier et al., 2015 ; Fraser et al., 2015 ; Weinreb and Raphael, 2015]. Avec IC-Finder, le fait de pouvoir faire varier le degré minimal d'hétérogénéité nécessaire pour fusionner ou pas deux groupes adjacents permet d'étudier l'organisation hiérarchique de la chromatine (Fig. 2.2.1C). Il est possible de coupler cette information sur la hiérarchie et le ré-échantillonnage présentés ci-dessus, de manière à avoir une image complète et fiable du repliement hiérarchique de la chromatine (Fig. 2.3.5).

2.4 Conclusion

Dans ce chapitre, nous avons présenté une nouvelle méthode, IC-Finder qui permet d'extraire à partir des cartes Hi-C une segmentation 3D du génome en compartiments d'interaction. Récemment, de nombreuses preuves expérimentales ont mis en évidence le rôle fonctionnel des compartiments d'interaction qui permettent entre autre de maintenir une certaine expression des gènes soit en promouvant soit en réprimant les contacts à longue portée entre enhancer ou silencer et promoteur [Sexton and Cavalli (2015); Dekker and Mirny (2016)]. Un pré-requis pour mieux comprendre les mécanismes de formation et de contrôle des compartiments d'interaction est donc dans un premier temps de les identifier correctement. En conséquence, plusieurs méthodes ont déjà été développée au cours des années précédentes

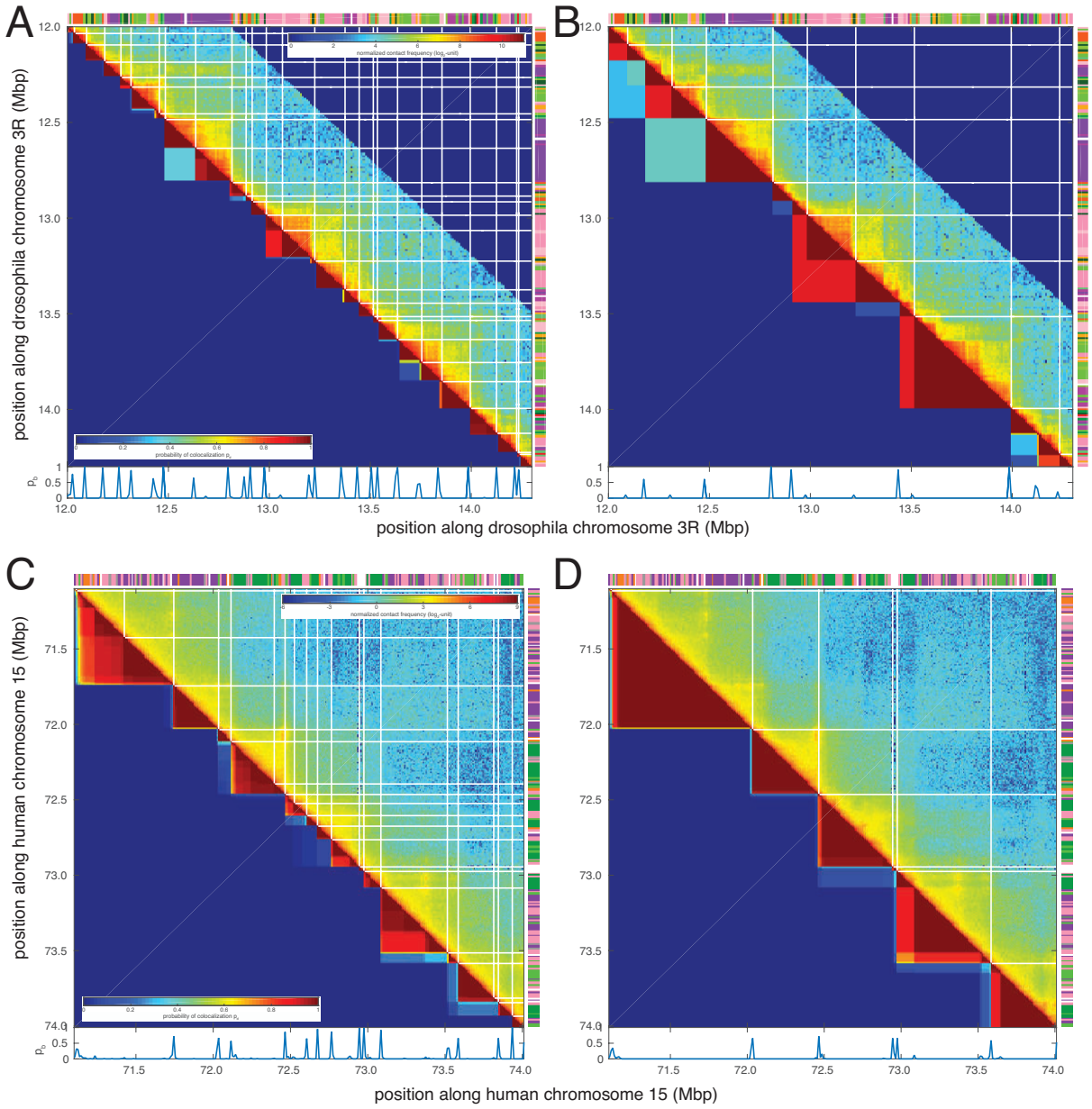

Figure 2.3.5 – **Options d'IC-Finder : ré-échantillonnage et hiérarchie.** (A,C) Differents types de résultats proposés par IC-Finder : (1) frontières des TADs déterminées par le programme en mode défaut (lignes blanches dans la partie triangulaire supérieure) ; (2) probabilité que deux locus soient dans le même compartiment $\mathtt{p_d}$ (matrice de la partie triangulaire inférieure) ; (3) probabilité pour un locus d'être une frontière $\mathtt{p_b}$ (courbe bleue en bas de chaque carte). Ces cartes sont des données de drosophile (A) [Sexton et al., 2012] et d'homme (C) [Rao et al., 2014]. (B,D) Similaires aux cartes (A,C)mais pour une hiérarchie de repliement supérieure ($\sigma_0 = 5$). Pour chaque carte, on représente également la séquence épigénomique (en haut et à droite de chaque carte) [Ho et al., 2014]. Le code couleur des états épigénomiques est donné dans le chapitre suivant, figure 3.3.2.

[Rao et al., 2014; Weinreb and Raphael, 2015; Shin et al., 2016; Filippova et al., 2014; Lévy-Leduc et al., 2014]. L'originalité et la force de notre méthode, IC-Finder est qu'elle se base sur l'hypothèse que les cartes Hi-C reflètent la nature polymérique de la chromatine [Imakaev et al., 2015]. Cette hypothèse a été motivée par de récents travaux sur le repliement de la chromatine démontrant que les données Hi-C relevées chez la drosophile sont compatibles avec un modèle de copolymère par bloc dont les blocs sont définis à partir de l'épigénome de drosophile [Jost et al., 2014; Olarte-Plata et al., 2016]. Ceci nous a mené à proposer une approche de regroupement hiérarchique dont la condition d'arrêt tient compte des propriétés polymériques de la drosophile. La méthode repose uniquement sur deux seuils dont les valeurs ont été fixées via un apprentissage sur un ensemble de cartes expérimentales. Nous avons pu vérifier la fiabilité de notre méthode sur un ensemble de 100 cartes de référence que nous avons construites in silico (avec le modèle décrit dans le chapitre 4) et pour lesquels nous connaissions donc la position des TADs. Nous avons comparé IC-Finder à d'autres méthodes de segmentation sur ces 100 cartes de référence ainsi que sur des cartes de référence. Il en ressort que IC-Finder est bien plus précis dans la détermination des TADs que les autres méthodes (excepté TopDom qui donne des résultats comparables) et que IC-Finder est dans tous les cas numériquement plus efficace pour les grandes cartes de contact (plus de 1000×1000).

Comme les cartes de contact expérimentales sont bruitées, il est préférable de localiser les compartiments d'interaction dans un sens probabiliste plutôt que déterministe. IC-finder est la première méthode de segmentation permettant d'estimer la robustesse des prédictions étant donné les erreurs expérimentales. De plus, l'observation de cartes de contact expérimentale laisse clairement apparaître qu'il n'y a pas de segmentation unique mais une organisation hiérarchique de compartiments d'interaction : plusieurs petits compartiments d'interaction adjacents forment eux-mêmes des compartiments plus grands. Il est intéressant de remarquer qu'une telle organisation hiérarchique est la marque des propriétés de repliement d'un copolymère par bloc (Fig. 2.3.1 et chapitre 4). En effet, en raison des interactions spécifiques entre monomères du même type, les domaines épigénomiques se replient sur eux mêmes : c'est le premier niveau d'organisation. Ensuite, l'existence d'attractions spécifiques entre compartiments d'interaction de même type à longue portée induit la formation de motifs d'interaction d'ordre supérieur. Par défaut, IC-Finder identifie le premier niveau de hiérarchie. Toutefois, le programme offre en option la possibilité d'accéder à différents niveaux de hiérarchie en égalant les deux paramètres de seuil et en ajustant la valeur commune. Accéder aux différentes hiérarchies dans le repliement de la chromatine est aussi possible avec TADtree [Weinreb and Raphael, 2015].

Nous utiliserons dans les chapitres suivants IC-Finder, dans un premier temps, à des fins d'analyses statistiques (chapitre 3) et dans un second temps, dans le cadre d'une inférence

de paramètres d'interaction à l'échelle des TADs (chapitre 5). Nous verrons que IC-Finder de révéler les principes généraux du repliement 3D de la chromatine et leur influence sur la régulation des gènes.
N.B. : Un article sur IC-Finder est actuellement en cours de révision dans le journal « Nucleic Acids Research ».

CHAPITRE 3

ANALYSE STATISTIQUE DE DONNÉES HI-C ET ÉPIGÉNOMIQUES

Afin de modéliser la chromatine dans le chapitre 4, nous nous baserons sur des données de Hi-C et des données épigénomiques. De ce fait, on propose dans ce chapitre une analyse statistique de ces données. On commencera dans un premier temps par estimer l'ordre de grandeur des erreurs sur les cartes Hi-C et par étudier le choix du schéma de normalisation de ces cartes. Dans un second temps, on examinera la distribution des tailles des domaines épigénomiques et on comparera plusieurs séquences épigénomiques obtenues chez la drosophile. Enfin on cherchera à quantifier les corrélations entre épigénome (données 1D) et contacts Hi-C (information 3D) ce qui nous renseignera sur le couplage 1D/3D opérant dans la chromatine. Pour étudier dans quelle mesure la séquence épigénomique 1D implique un repliement 3D du génome, on s'intéressera à la compaction locale en fonction de l'état épigénomique.

3.1 Analyse statistique de données HiC

3.1.1 Estimation des erreurs expérimentales sur les cartes Hi-C

Afin d'estimer les erreurs relatives dont sont entachées les coefficients $C(i,j)$ d'une carte de contact, C, nous allons effectuer une mesure de dispersion relative via le coefficient de variation local $c_v(i,j)$. Le coefficient de variation aussi appelé écart type relatif est le rapport entre écart type et moyenne d'un échantillon :

$$c_v(i,j) = \left| \frac{\mathrm{std}(c)}{\mathrm{moy}(c)} \right| \times 100 \tag{3.1}$$

$$\text{avec } \forall (i,j) \in [\![2, N-1]\!],\ c = \begin{pmatrix} C(i-1,j-1) & C(i-1,j) & C(i-1,j+1) \\ C(i,j-1) & C(i,j) & C(i,j+1) \\ C(i+1,j-1) & C(i+1,j) & C(i+1,j+1) \end{pmatrix}$$

avec $\mathrm{std}(c)$ l'écart type de la matrice c et $\mathrm{moy}(c)$ la moyenne de la matrice c. La multiplication par 100 permet d'avoir un pourcentage. Si i et/ou j valent 1 ou N, la matrice c est tronquée de sorte que tout ces coefficients soient définis. Si un coefficient de la matrice c est nul on l'ignore car nous n'avons pas un moyen de savoir s'il s'agit d'un vrai zéro ou d'une absence d'information.

À partir de données Hi-C obtenues chez la drosophile [Sexton et al., 2012], on représente sur la figure 3.1.1 la carte de contact d'une portion du chromosome 3R ainsi que les erreurs relatives correspondantes. On peut voir que dans les TADs caractérisés par des nombres de contact élevés le coefficient de variation local est au plus bas. L'erreur relative médiane à l'intérieur des TADs est d'environ 8% chez la drosophile. En revanche, plus on s'éloigne de la diagonale, plus les coefficients de variation $c_v(i,j)$ sont élevés, ceci s'explique par la diminution de l'échantillonnage en fonction de la distance génomique. Sur l'ensemble du génome de la drosophile, l'erreur relative médiane est de 23% ce qui est élevé mais en accord avec le protocole HiC qui est soumis à différents biais (cf. chapitre 1). On pourrait penser que la faible erreur relative dans les TADs s'explique en partie par le processus de normalisation des cartes Hi-C (données de Sexton et al., 2012 normalisées selon Yaffe and Tanay, 2011). Toutefois, si on réalise les mêmes calculs sur les données de Hi-C brutes (sans normalisation), on trouve que l'erreur relative dans les TADs est de 10% et qu'à l'échelle du génome elle s'élève à 26%. Ces résultats ne sont que légèrement supérieurs au cas normalisé ce qui suggère finalement que la normalisation n'influe pas tellement sur le bruit en lui même.

Si on applique la même méthode d'estimation des erreurs aux données Hi-C humaines de [Rao et al., 2014], on trouve que la médiane du coefficient de variation pour les données brutes est 8% au niveau des TADs et de 30% sur l'ensemble des chromosomes. Pour les

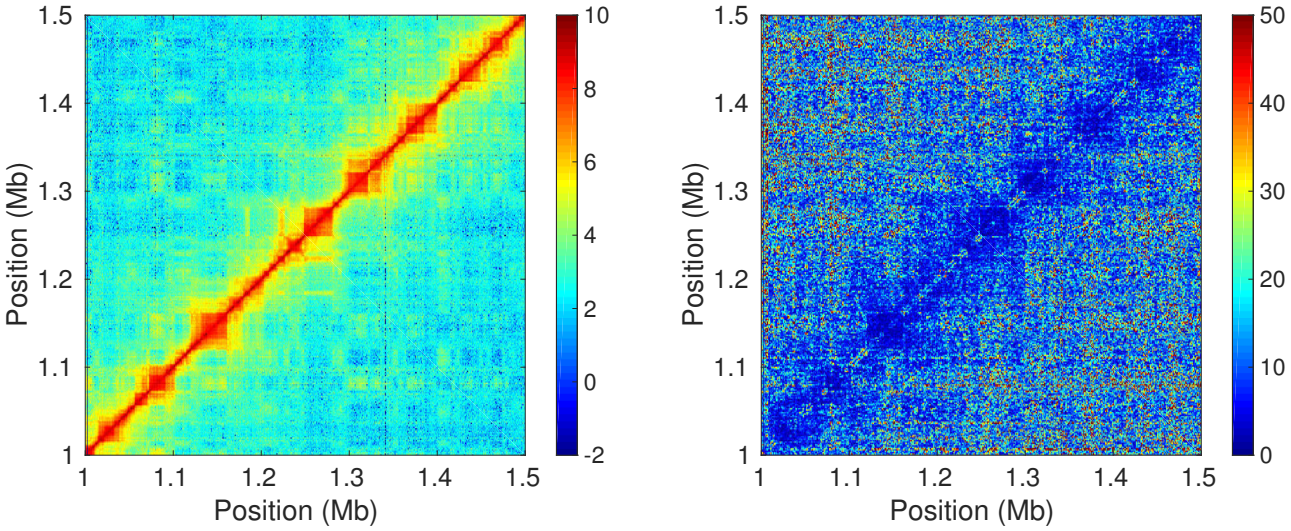

FIGURE 3.1.1 – **Estimation du bruit sur une carte de contact.** (À gauche) Carte de contact expérimentale, $\mathbf{C}$, en échelle $\log_2$, pour une région du chromosome 3R de la drosophile [Sexton et al., 2012]. (À droite) Carte des variations locales correspondantes en %, $\mathbf{c}_v$, calculées avec la formule équation (3.1).

données normalisées selon la méthode de [Knight and Ruiz, 2012], la médiane du coefficient de variation est de 6% au niveau des TADs et de 28% sur l'ensemble des chromosomes.

Finalement, pour deux expériences de Hi-C réalisées dans des laboratoires différents, sur des organismes différents et avec deux procédures de normalisation des données différentes, on trouve des coefficients de variation du même ordre de grandeur. On retiendra, pour la suite, que les données Hi-C normalisées concernant l'homme et la drosophile [Rao et al., 2014; Sexton et al., 2012] présentent des erreurs relatives de l'ordre de 25%.

3.1.2 Effet de la normalisation sur la détermination des TADs

En plus des erreurs d'échantillonnage, le protocole de Hi-C induit plusieurs biais systématiques vus dans le chapitre 1 (distance entre site de coupure des enzymes de restriction, contenue en GC des lectures, ... [Yaffe and Tanay, 2011]). Nous avons vu qu'au cours des dernières années plusieurs stratégies de normalisation ont émergé afin de normaliser les données Hi-C [Yaffe and Tanay, 2011; Hu et al., 2012; Kalhor et al., 2012; Cournac et al., 2012; Imakaev et al., 2012; Sauria et al., 2015; Knight and Ruiz, 2012]. Parmi ces méthodes, on peut distinguer deux familles : (i) les approches paramétriques qui modélisent les biais explicitement [Yaffe and Tanay, 2011; Hu et al., 2012] ; et (ii) les approches de renormalisation qui suppose que chaque fragment devrait être observé le même nombre de fois [Kalhor et al., 2012; Cournac et al., 2012; Imakaev et al., 2012; Sauria et al., 2015; Knight and Ruiz, 2012].

L'application de ces différentes procédures de normalisation aux cartes de contact brutes mène à des différences quantitatives au niveau des cartes Hi-C finales (Fig. 3.1.2C). Dans cette section, on se demande comment le processus de normalisation affecte la détection des compartiments d'interaction et si les différentes méthodes mènent à des résultats consistants les unes par rapport aux autres ou pas. Nous testons trois schémas de normalisation (Yaffe, Khalor et ICE, [Yaffe and Tanay, 2011; Kalhor et al., 2012; Imakaev et al., 2012]) qu'on applique aux données Hi-C chez la drosophile [Sexton et al., 2012] et nous comparons les segmentations obtenues avec celle obtenue dans le cas des données Hi-C brutes. L'analyse statistique des résultats montre qu'environ 70% des frontières détectées sur les données sont conservées après normalisation et inversement environ 30% de frontières prédites à partir des données normalisées ne sont pas présentes sur les données brutes (Fig. 3.1.2A). Les biais systématiques observés sur les données brutes mènent à plus d'irrégularités dans les motifs d'interaction le long de la diagonale (Fig. 3.1.2C) ce qui se manifeste au niveau de la détections des compartiments par la prédiction de nombreux petits domaines (Fig. 3.1.2B). Les schémas de normalisation mènent tous vers la détection de domaines plus grands. Pour des méthodes de normalisation appartenant à une même famille, les résultats sont hautement cohérents (voir ICE et Khalor Fig. 3.1.2A) avec des nombres et tailles de compartiments obtenus très proches (Fig. 3.1.2B). Entre familles de méthodes, les résultats sont aussi globalement cohérents, même si on observe de légères différences (Fig. 3.1.2A), dues principalement à la détection d'un plus grand nombre de petit domaine avec la méthode de Yaffe (Fig. 3.1.2B). Cette étude confirme que la normalisation est un facteur clef pour l'analyse des cartes Hi-C et elle montre que si l'on souhaite comparer des segmentations, il est nécessaire que ces dernières aient été obtenues avec des méthodes de normalisation similaires.

3.1.3 Nombre de contacts total cumulé en fonction de la distance génomique

Dans cette sous section on cherche à savoir à quelle échelle de distance génomique se font la majorité des contacts mesurés avec la technique de Hi-C. On présente sur la figure 3.1.3 le nombre de contacts entre loci total cumulé en fonction de la distance génomique. On peut voir que plus de 50% des contacts intra-chromosomiques (courbe en bleu) sont détectés pour des distances génomiques inférieures à 1Mb. La figure nous renseigne aussi sur les contacts inter chromosomiques (en rouge) : on constate qu'ils représentent environ 1/3 des contacts totaux. Ceci implique que, pendant l'interphase, malgré l'existence de territoires chromosomiques [Cremer and Cremer, 2010] il existe une interpénétration entre chromosomes non négligeable. Pour approfondir l'analyse de ces résultats, on peut se demander à quelle répartition des

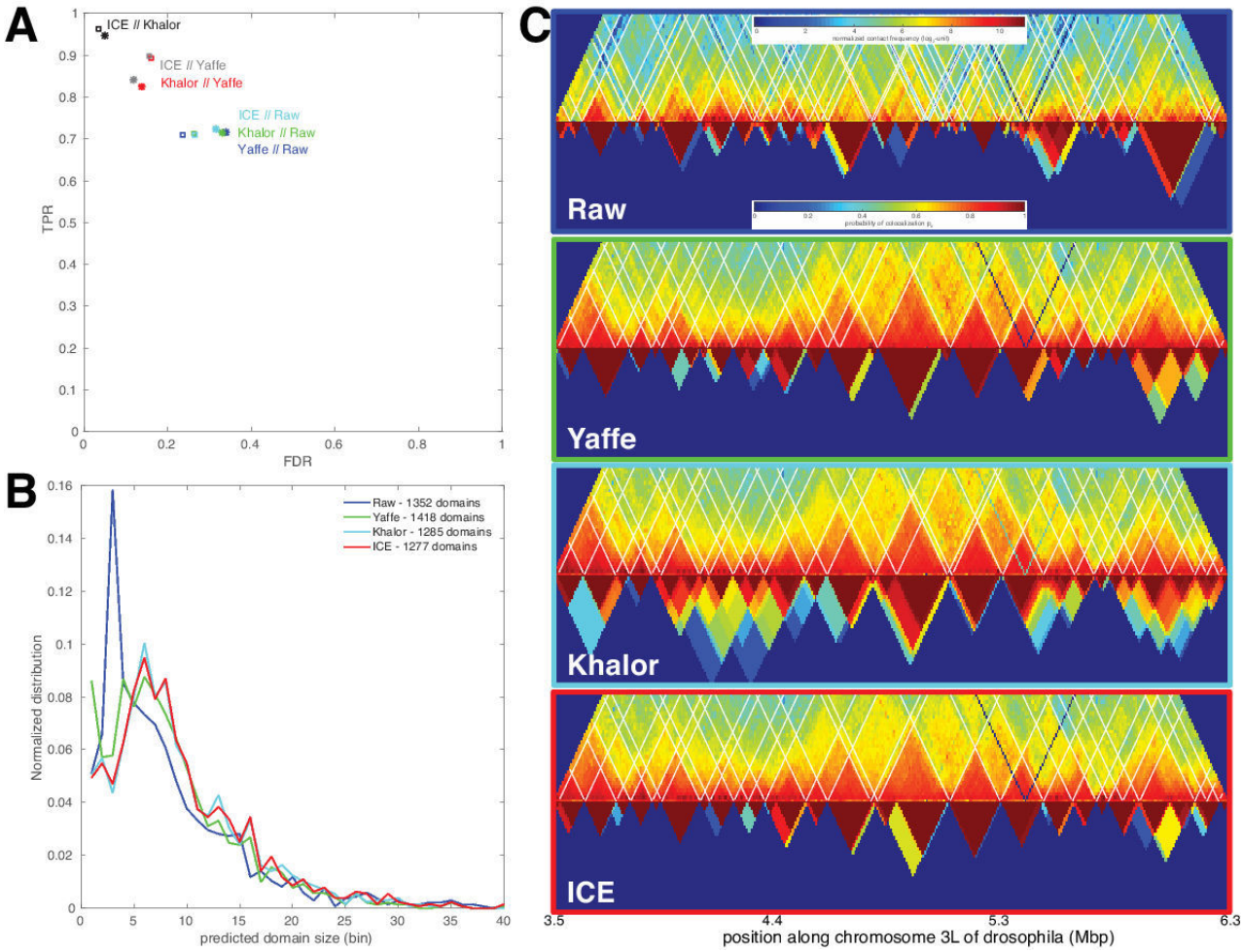

FIGURE 3.1.2 – **Segmentations prédites en fonction du schéma de normalisation.** (A) Taux de vrai positif, TPR, en fonction du taux de fausse découverte, FDR, pour les domaines (symboles carrés) et pour les frontières (astérisques) des segmentations obtenues à partir de schémas de normalisation différents indiqués sur la figure. (B) Distribution des tailles des domaines prédits par IC-Finder à partir de cartes de contact normalisées différemment. Le nombre total de domaine obtenue avec chaque méthode de normalisation est indiqué en légende. (C) Exemples de segmentations prédites avec IC-Finder pour une région génomique du chromosome 3L de drosophile [Sexton et al., 2012]. En haut : carte Hi-C avec en lignes blanches les frontières prédites avec IC-Finder. En bas : Probabilité $\mathrm{p_d}$ que deux loci soient prédits comme appartenant au même TAD. Pour obtenir ces résultats IC-Finder est lancé avec les paramètres σ_- et σ_+ par défaut.

contacts doit on s'attendre dans un cas idéal. Pour répondre à cette question, on considère un modèle théorique simple consistant à modéliser les 10 chaînes chromosomiques présentes dans le noyau cellulaire de la drosophile par des polymères purement gaussien (c'est-à-dire avec des interactions de connectivité seulement). Les 10 chaînes dont on parle sont les 2 paires des bras suivants : 2L (23Mb), 3L (25Mb)), 2R (21Mb)), 3R (28Mb)), ainsi que le chromosome X (22Mb) éventuellement en double dans le noyau ou bien accompagné du chromosome Y (40Mb). Le chromosome 4 de taille 1.4Mb est négligé. On cherche à calculer, dans ce cadre, le nombre total de contact entre bins de 10kb, $N_{total} = N_{intra} + N_{inter}$ avec N_{intra} (resp. N_{inter}) le nombre de contacts intra (resp. inter) chromosomique.

Dans la partie 4.2.1, on présentera les distributions gaussiennes multivariées et on montrera en particulier que la probabilité de contact entre deux monomères i et j, P_{ij}, peut être approchée par la relation $P_{ij} \approx a^3\sqrt{\frac{2}{9\pi}}D_{ij}^{-3/2}$ (Éq. 4.10), avec $a \approx 100nm$ le rayon d'interaction pour des bins de 10kb et avec D_{ij} la distance quadratique moyenne entre bin. On suppose ici que : $D_{ij} \approx (s \times b)^2$, avec s la distance génomique entre les bins et avec b la taille typique d'un bin de $10kb$. Ceci permet d'estimer le nombre total de contact intra chromosomique :

$$N_{intra} \approx \#c \times \sum_{s=1}^{N-1} \sqrt{\frac{2}{9\pi}} \times \left(\frac{a}{b}\right)^3 \times s^{-3/2} \times (N-s) \tag{3.2}$$

avec $\#c = 10$ le nombre total de chaînes dans le noyau de drosophile et $N = 2500$ le nombre moyen de bin de $10kb$ dans une chaîne. On estime, en première approximation, que $b = a = 100nm$

Concernant le terme N_{inter}, on peut supposer que le mélange est parfait. Ainsi, deux bins i et j sur des chaînes différentes sont en contact s'ils se trouvent dans une même sphère d'interaction de volume $\frac{4}{3}\pi a_0^3$. D'où l'estimation de N_{inter} :

$$N_{inter} \approx \left(\frac{\frac{4}{3}\pi a_0^3}{\frac{4}{3}\pi R^3}\right) N^2 \frac{\#c\,(\#c-1)}{2}$$

avec $R = 2\mu m$ le rayon du noyau cellulaire de drosophile et avec le facteur $\frac{\#c(\#c-1)}{2}$ représentant le nombre de paires de chaînes.

L'application numérique donne $N_{intra} = 1.3 \times 10^4$ et $N_{inter} = 2.2 \times 10^4$ ce qui signifie que les contacts inter chromosomiques représentent 62% des contacts totaux alors que nous avons trouvé 33% à partir des données expérimentales. Ce calcul suggère que, dans le noyau cellulaire, les chromosomes ne sont pas très mélangés. Notons qu'il a été montré qu'un modèle polymèrique tenant compte de contraintes topologiques telles que les interactions avec la membrane permet de reproduire l'existence des territoires chromosomiques [Rosa and Everaers, 2008].

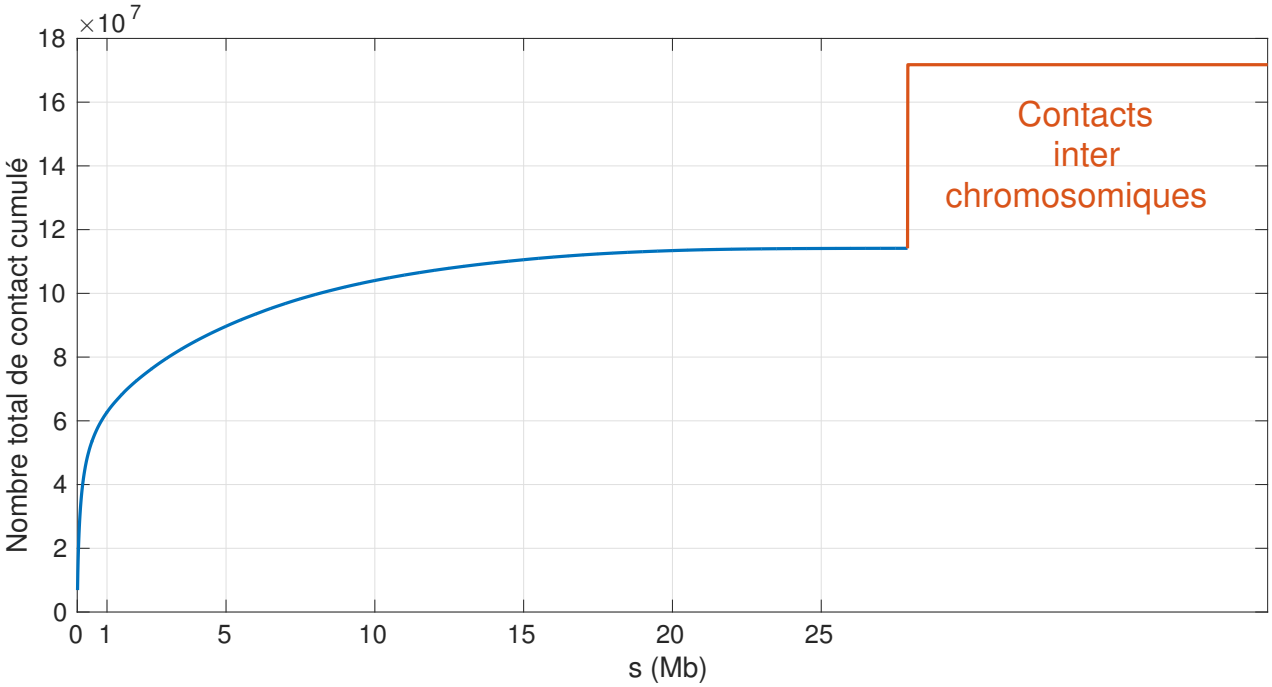

FIGURE 3.1.3 – **nombre de contacts entre loci total cumulé en fonction de la distance génomique, s, en Mb.** La courbe en bleu concerne les contacts intra chromosomiques alors que le palier en rouge indique le nombre toal de contact inter chromosomique sur l'ensemble du génome cartographiable de la drosophile [Sexton et al., 2012].

De la même manière qu'on a calculé $\mathsf{N}_{\mathtt{intra}}$ (Éq. 3.2), on peut calculer le nombre de contacts cumulé à n'importe quelle échelle (il suffit de sommer jusqu'à l'échelle choisie plutôt que jusqu'à $\mathsf{N}-1$). Si on somme les contacts jusqu'à 20kb, soit deux bins, on trouve 7.2×10^3 contacts. La moitié des contacts intra chromosomiques se produit donc entre très proches voisins. Dans le cas expérimental, on a vu qu'environ 50% des contacts concernaient des bins espacés de distances génomiques allant jusqu'à 1Mb. Ceci suggère que le confinement dans le noyau cellulaire de drosophile est si fort qu'il conduit à avoir beaucoup plus d'interactions entre loci génomiquement éloignés.

3.1.4 Nombre de contacts moyen en fonction de la distance génomique

Dans cette partie, on présente le nombre de contacts moyen, N_c, en fonction de la distance génomique, s, calculé dans les cas de la drosophile (chromosome 3R) et de l'homme (chromosome 2) [Sexton et al., 2012; Rao et al., 2014]. On observe sur la figure 3.1.4 en échelle log que pour les deux organismes différents régimes se succèdent (4 pour la drosophile et 3 pour l'homme).

Dans le cas de la drosophile, avant 50kb, N_c décroît très peu ce qui est probablement dû à l'existence d'interactions à courte portée mais aussi dû à la ligation opérée par le formaldéhyde ce qui rapproche artificiellement les monomères les uns des autres. Ensuite, jusqu'à 1Mb, la loi d'échelle est $N_c \sim s^{-1}$ ce qui peut entre autre être dû à un effet de crumpling ou encore à la coexistence de plusieurs états multistables (pelote ou globule) au voisinage du point θ. Entre 1Mb et 5Mb, la décroissance du nombre de contacts moyen est plus lente, $N_c \sim s^{-0.6}$ ce qui est probablement la conséquence d'un confinement hors équilibre des chromosomes dans leur territoire chromosomique. Enfin, au delà de 5Mb, on observe de nouveau une décroissance selon $N_c \sim s^{-1}$. Cette loi est sûrement le fruit d'un confinement longitudinal dû à un effet de mémoire de l'état initial caractérisé par une configuration en Rabl.
Pour l'homme, au cours du premier régime, jusqu'à 0.8Mb environ, la loi d'échelle est $N_c \sim s^{-0.8}$ ce qui est cohérent avec un modèle d'extrusion [Sanborn et al., 2015; Fudenberg et al., 2016]. Ensuite, jusqu'à 8Mb environ, la décroissance est plus forte, $N_c \sim s^{-1.3}$, ce qui peut être interprété par l'effet de crumpling. Enfin, au delà de 8Mb, la loi est $N_c \sim s^{-0.6}$ ce qui est imputable à un fort confinement des chromosomes dans leur territoire chromosomique. On peut noter que les territoires chromosomiques sont plutôt cylindriques chez la drosophile alors qu'ils sont sphériques chez l'homme [Cremer and Cremer, 2010]. Ceci peut expliquer la différence de pente entre les derniers régimes de chaque organisme. Toutefois, pour s'assurer de cette explication, il aurait fallu tracer sur la figure 3.1.4 l'observable N_c jusqu'à la fin du chromosome, soit 243Mb, ce qui n'a pas été possible en raison d'un manque de données.
Enfin, on peut ajouter qu'il existe une différence de comportement entre les deux organismes à très petite échelle. Cette différence n'est pas évidente à interpréter, elle peut avoir pour origine la différence de protocole expérimental qui induit des artefacts plus ou moins important via le formaldéhyde .

3.2 Analyse statistique de données épigénomiques

Comme vu dans le chapitre 1, il a été montré qu'à partir de données Chip-Seq de plusieurs marques d'histones, l'information épigénomique locale peut être caractérisée par un certain nombre d'états épigénomiques différents et que leur répartition locale le long du génome peut être inférée par des méthodes de HMM . Plusieurs laboratoires ont publié les partitions épigénomiques obtenues à partir de cellules de drosophile [Ho et al., 2014; Filion et al., 2010; Kharchenko et al., 2011]. Nous avons téléchargé ces informations épigénomiques sur les pages internet `https://www.encodeproject.org/comparative/chromatin/` [Ho et al., 2014], `http://www.ncbi.nlm.nih.gov/geo/` avec le numéro d'accés GSE22069 [Filion et al., 2010] et `http://www.modencode.org` avec le numéro GEO GSE25321 [Kharchenko et al.,

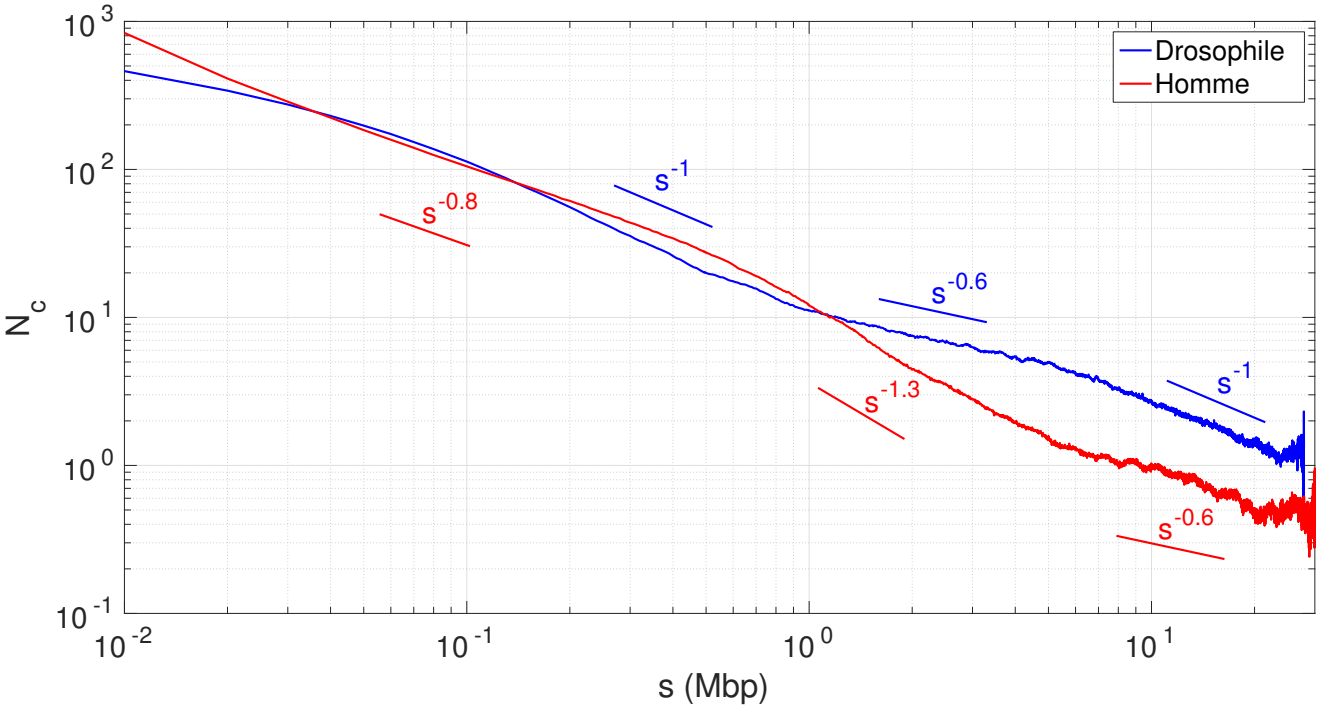

FIGURE 3.1.4 – **nombre de contacts moyen N_c en fonction de la distance génomique s (en Mb) pour la drosophile (chromosome 3R) et pour l'homme (chromosome 2) [Sexton et al., 2012; Rao et al., 2014].** N.B. : Le chromosome 2 humain est long de 243Mb, l'observable N_c est sur cette figure représentée jusqu'à 30Mb car nous n'avons pas eu accès aux données au delà.

2011].

Données [Ho et al., 2014]

Elles ont été obtenues à partir d'embryons tardifs de drosophile et révèlent que l'information épigénomique locale peut être caractérisée par 16 états épigénomiques : (1) Promoteur ; (2) Enhancer 1 ; (3) Enhancer 2 ; (4) Transcription 5' 1 ; (5) Transcription 5' 2 ; (6) Gène H4K20me1 ; (7) Transcription 3' 1 ; (8) Transcription 3' 2 ; (9) Transcription 3' 3 ; (10) Répression polycomb 1 ; (11) Répression polycomb 2 ; (12) Hétérochromatine 1 (13) Hétérochromatine 2 ; (14) Faible signal 1 ; (15) Faible signal 2 ; (16) Faible signal 3.

Les marques (12) et (13) correspondent à de l'hétérochromatine constitutive alors que les marques (14), (15) et (16) correspondent à l'hétérochromatine dite nulle (car peu de signal mesuré dans ces régions).

Données [Filion et al., 2010]

Elles ont été obtenues à partir de la lignée cellulaire embryonnaire Kc167 et une analyse en composante principale sur ces données révèle l'existence de 5 états : (1) Gène actif 1 (toujours actif) ; (2) Gène actif 2 (spécifique à certains tissus) ; (3) Répression polycomb ; (4)

Hétérochromatine de type HP1 ; (5) Hétérochromatine nulle.

Données [Kharchenko et al., 2011]

Elles ont été obtenues à partir des lignées cellulaires S2-DRSC et ML-DmBG3-c2, dérivées de tissus d'embryons tardifs mâles. Ces données mettent en avant 9 états épigénomiques : (1) Promoteur et TSS ; (2) Exon de gènes transcrits ; (3) Intron ; (4) H3K36me1 et absence de H3K27ac ; (5) H4K16ac ; (6) Répression polycomb ; (7) Hétérochromatine péricentromérique ; (8) Hétérochromatine H3K9me2 / me3 ; (9) Hétérochromatine noire.

Concernant les trois partitions épigénomiques, on s'attend bien sûr à ce qu'elles présentent quelques différences les unes par rapport aux autres puisque les cellules et les protocoles dont elles sont issues sont différents. Aussi, on peut remarquer que les trois partitions sont composées de nombres d'états différents (respectivement 16, 5 et 9). La détermination du nombre d'états épigénomiques optimal revêt en effet un caractère subjectif selon la granularité recherchée.

3.2.1 Estimation des erreurs sur l'information épigénomique

Afin de comparer et afin d'estimer les différences entre les trois partitions épigénomiques présentées ci-dessus, on groupe chacune d'elle en 4 grandes familles d'état : **Actif** ; **Répression polycomb** ; **Hétérochromatine constitutive** et **Hétérochromatine nulle**. Pour les données [Ho et al., 2014], on regroupe dans la famille des actifs, les marques **(1) à (8)** ; la famille des polycomb est composée des marques **(10) et (11)** ; l'hétérochromatine constitutive des marques **(12) et (13)** et enfin les marques **(14), (15) et (16)** appartiennent à la famille d'hétérochromatine nulle. Pour les données [Filion et al., 2010], on a simplement besoin de regrouper les états **(1) et (2)** dans une unique famille de gènes actifs. Enfin, pour les données [Kharchenko et al., 2011], on regroupe dans la famille des actifs, les marques **(1) à (5)** ; l'hétérochromatine constitutive est composée des états **(7) et (8)** ; enfin les familles polycomb et hétérochromatine nulle ne nécessitent pas de regrouper des états, elles correspondent respectivement aux états **(6)** et **(9)**.

Ainsi les partitions [Ho et al., 2014; Filion et al., 2010; Kharchenko et al., 2011] composées initialement respectivement de 36809, 8428 et 21743 domaines épigénomiques se retrouvent après le processus de regroupement en quatre familles avec 11232, 6250 et 16225 domaines. Sur la figure 3.2.1 on peut voir une carte de contact pour une région génomique du chromosome 3R de la drosophile accompagnée des segmentations épigénomiques prédites par [Ho et al., 2014; Filion et al., 2010; Kharchenko et al., 2011] et regroupés en quatre familles. Cet exemple

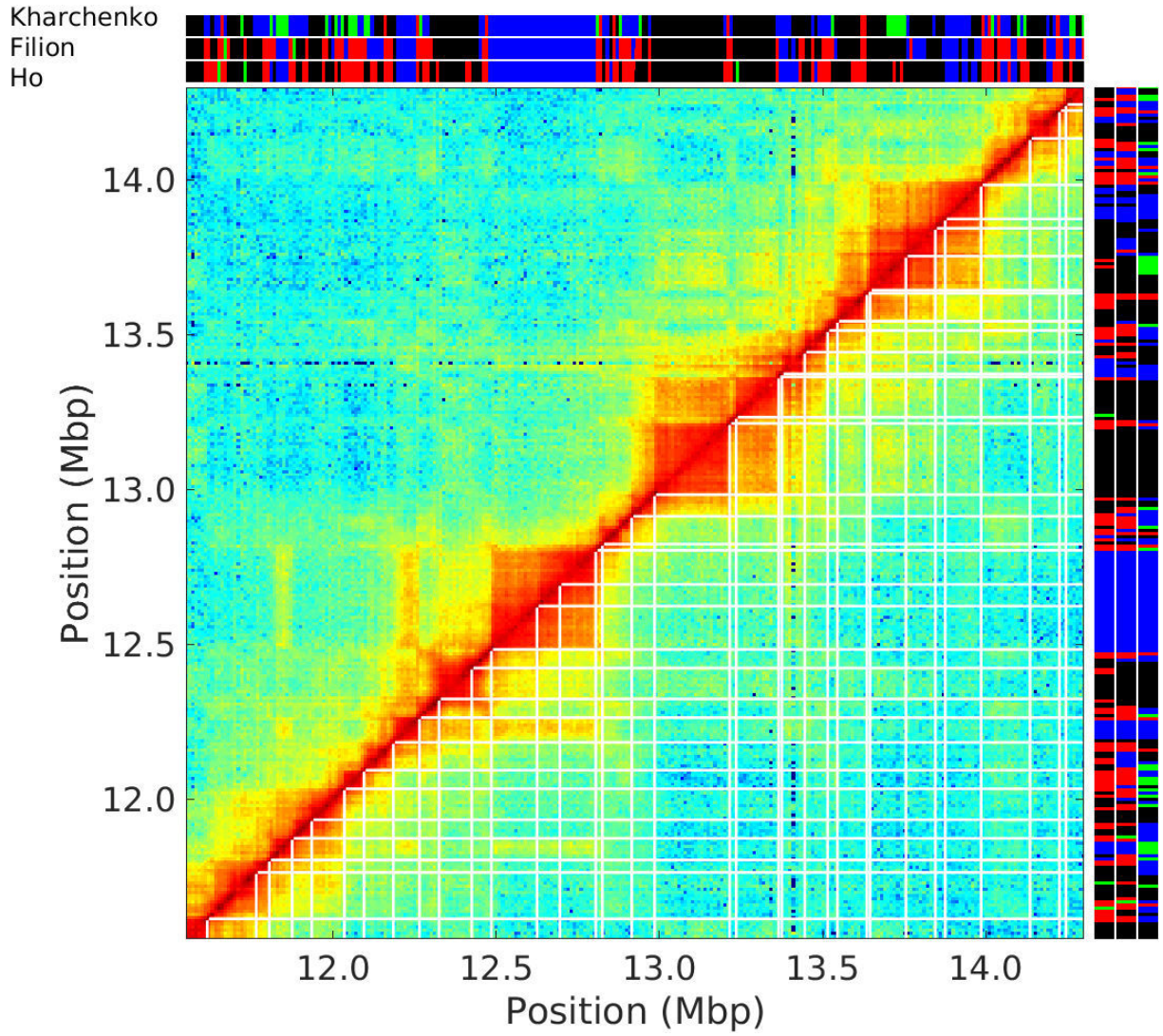

FIGURE 3.2.1 – **Carte de contact pour une région génomique du chromosome 3R de la drosophile accompagnée d'informations épigénomiques.** Les partitions épigénomiques représentées en haut et à droite de la carte de contact sont issues de [Ho et al., 2014; Filion et al., 2010; Kharchenko et al., 2011], respectivement de la partition la plus proche de la carte à la plus éloignée. Les contacts sont représentés en échelle $\log_2$, l'échelle de couleur est identique à celle donnée figure 3.1.1 (à gauche). Dans la partie triangulaire inférieure on superpose à la carte des lignes blanches représentant les frontières des TADs trouvées avec IC-Finder.

montre que les trois segmentations sont globalement similaires malgré quelques différences lorsque les domaines sont petits.

Afin de comparer les trois partitions épigénomiques à l'échelle du génome, on présente sur la figure 3.2.2A la répartition des différents états épigénomiques à l'échelle du génome et les distributions des tailles des domaines épigénomiques (sans tenir compte des régions non

cartographiées qui sont sensiblement similaires pour les trois segmentations). On observe que dans les trois cas l'hétérochromatine nulle domine en concernant environ la moitié du génome cartographié. Les proportions concernant les autres états sont relativement du même ordre d'une partition à l'autre. Le fait que la proportion de l'état polycomb est plus faible dans la segmentation 1 (celle de Ho et al., 2014) ou le fait que l'état d'hétérochromatine constitutive soit plus représenté dans la segmentation 3 (celle de Kharchenko et al., 2011) n'est pas évident à expliquer étant donné les variations dans le protocole utilisé par chaque laboratoire. Du côté de la distribution des tailles des domaines, on voit que dans les trois cas, le pic se situe autour de 4 ou 5kb. Aussi, les domaines ont une taille moyenne du même ordre de grandeur : environ 12kb. Ces observations suggèrent que les trois partitions sont significativement corrélées. Ce résultat est de plus confirmé par la figure 3.2.2B qui présente les couples (FDR,TPR) d'une partition I par rapport à une partition J avec I et J valant respectivement 1, 2 et 3 pour les partitions de Ho et al., Filion et al., et Kharchenko et al. En effet, on peut voir qu'il existe des corrélations entre les différentes partitions puisque les différents carrés sont plus proches du point (0,1) (point qui représente deux segmentations parfaitement identiques) que les diamants calculés de la même manière que les carrés mais en mélangeant la séquence épigénomique de référence. Cette comparaison des couples (FDR,TPR) réel et (FDR,TPR) moyen obtenu en mélangeant la séquence épigénomique de référence est un moyen simple de vérifier que les corrélations entre les trois partitions ne sont pas des artefacts qui s'expliqueraient par des distributions de taille des domaines similaires. En conclusion, les trois partitions donnent des résultats cohérents les uns par rapport aux autres. On peut considérer que les partitions obtenus par les trois laboratoires sont complémentaires, dans le sens où elles peuvent fournir une estimation des incertitudes sur les données. Pour cela, on peut remarquer que le pourcentage du nombre de loci (de 1kb) étant classé dans la même famille épigénomique par les trois segmentations est de 65%. Autrement dit, on peut considérer que 65% de l'épigénome cartographié est fiable. Pour les 35% loci restant, il est préférable de considérer que leur état épigénomique est incertain. Cette incertitude peut s'expliquer par le fait que les trois partitions sont issues de protocoles différents et surtout réalisées à partir de types cellulaires différents dont les états de différenciation peuvent donc présenter des variations.

3.2.2 Composition épigénomique d'un bin de 10 kb

Dans le 4, afin de modéliser la chromatine, nous allons introduire un modèle de copolymère par bloc dont chaque bloc est caractérisé par une couleur épigénomique. Nous travaillerons avec des monomères de 10kb. Or, on a vu sur la figure 3.2.2 que la distribution des tailles des domaines épigénomiques présente un pic autour de 5kb et une moyenne d'environ 12kb. On se

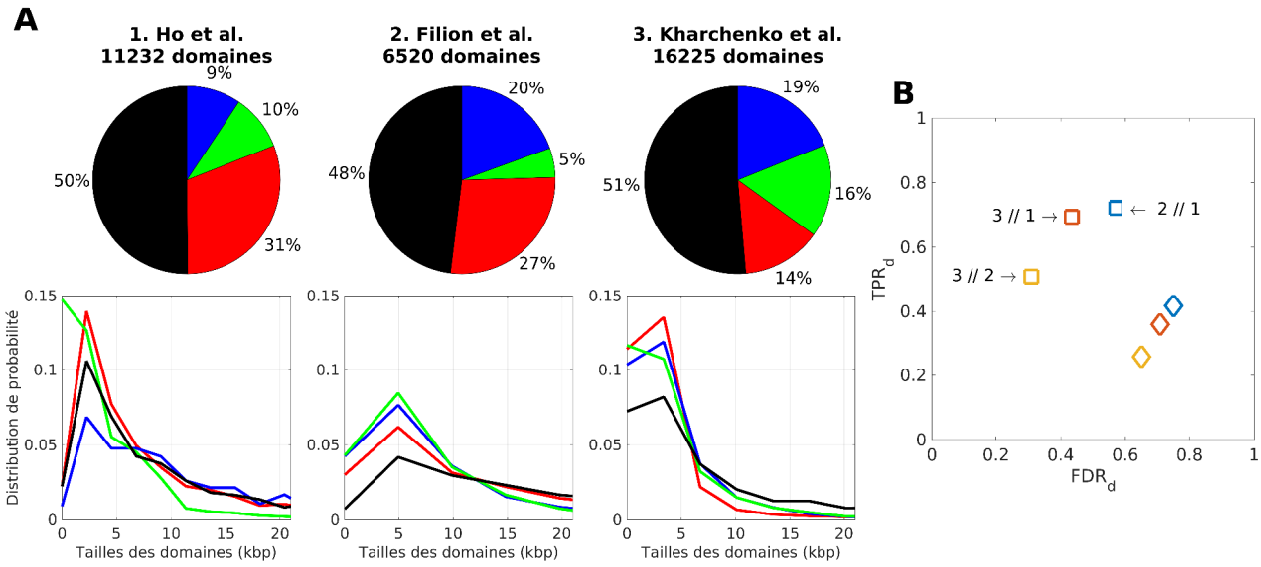

FIGURE 3.2.2 – **Comparaison de trois partitions épigénomiques obtenues à partir de cellules embryonaires de drosophile** [Ho et al., 2014; Filion et al., 2010; Kharchenko et al., 2011] **et après regroupement en 4 familles.** (A) En haut, répartition des différents états épigénomiques à l'échelle du génome, en bas, distributions des tailles des domaines épigénomiques. Les couleurs noire, rouge, verte et bleue représentent respectivement les états nul, actif, hétérochromatinien et polycomb. (B) Taux de vrai positif, TPR, en fonction du taux de fausse découverte, FDR, pour les domaines épigénomiques de la séquence I par rapport à la séquence J indiquées sur la figure (I = 1, 2, et 3 [Ho et al., 2014; Filion et al., 2010; Kharchenko et al., 2011]). Les diamants sur la figure indiquent les valeurs (FDR, TPR) dans le cas où on inverse l'ordre de la séquence de référence (couleurs respectivement bleue, rouge et jaune pour 2//1, 3//1 et 3//2).

demande donc ici s'il est légitime de considérer qu'un bin de 10kb est constitué d'une couleur majoritaire ou bien si cette hypothèse est trop grossière et qu'il faille donc mieux considérer un bin de 10kb comme constitué d'une certaine proportion de chaque état. Pour répondre à cette question, on représente sur la figure 3.2.2 la distribution cumulée des proportions épigénomiques majoritaires dans un bin de 10kb. Dans le cas idéal où le bin est constitué d'un unique état, la proportion est de 1 et dans le cas opposé où le bin est composé uniformément des quatre états épigénomques, la proportion est de 0.25. La ligne noire en tirets montre que presque tout les bins sont caractérisés par un état épigénomique majoritaire couvrant plus de 50% du bin, soit 5kb. De plus, 56% des bins de 10kb sont composés d'un unique état épigénomique. Cette figure nous apprend aussi que les bins les plus purs sont les bins dont l'état épigénomique majoritaire est polycomb. En effet, plus de 60% de ces bins sont exclusivement composés de polycomb.

En résumé, pour 56% des bins, considérer qu'ils sont composés d'une unique couleur est parfaitement légitime, pour les 44% bins restants, travailler de la sorte constitue une approxima-

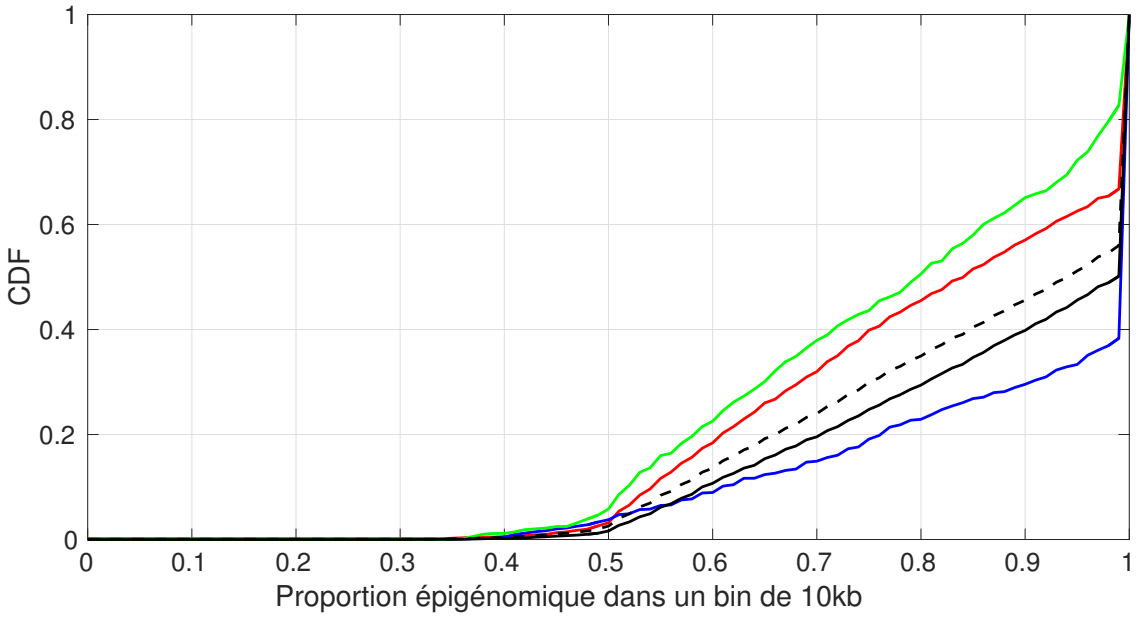

Figure 3.2.3 – **Distribution cumulée de la proportion épigénomique majoritaire d'un bin de 10 kb en fonction de l'état épigénomique.** Les courbes bleue, noire, rouge et verte représentent respectivement les états polycomb, nul, actif et hétérochromatine constitutive. La ligne noire en pointillée représente le cas moyen indépendemment de la couleur épigénomique.

tion. Cette approximation n'est toutefois pas grossière puisque ces bins sont constitués d'une marque majoritaire les couvrant à plus de 50%.

3.3 Corrélations entre compartimentation 3D et compartimentation 1D

On se demande dans cette section si les propriétés du repliement de la chromatine corrèlent avec l'épigénome.

3.3.1 Corrélations entre partition topologique et partition épigénomique

Sur la figure 3.2.1, on a pu voir « à l'oeil » que la position des domaines topologiques semble être corrélée à la position des domaines épigénomiques. Ici, on présente pour la drosophile le taux de vrai positif, TPR, en fonction du taux de fausse découverte, FDR, entre partition épigénomique et partition topologique donnée par IC-Finder. On réalise cette étude avec

les trois partitions épigénomiques présentées ci-dessus [Ho et al., 2014; Filion et al., 2010; Kharchenko et al., 2011] et avec les données Hi-C de Sexton et al.

Les points (FDR_b, TPR_b) de la figure 3.3.1 (astérisques) nous apprennent qu'à plus ou moins 10kb d'une frontière de TAD, il y a presque systématiquement une frontière épigénomique (car TPR_b proche de 1). Par contre, le fait que FDR_b soit d'environ 0.55, nous apprend qu'en l'absence de frontière de TAD, il arrive assez fréquemment que des frontières épigénomiques soient présentes. Pour vérifier que ces deux observations ne soient pas uniquement dues au fait que les domaines épigénomiques sont en moyenne bien plus petits que les domaines topologiques (respectivement 12kb contre 75kb), on calcule les mêmes couples (FDR_b, TPR_b) mais en mélangeant aléatoirement l'ordre de la séquence topologique (croix sur la figure 3.3.1). On peut voir que les résultats (croix et astérisques) sont assez proches mais avec toujours une meilleure corrélation dans le cas où la séquence topologique n'est pas mélangée. Pour s'assurer que cette observation ne soit pas un cas particulier, on a calculé des couples (FDR_b, TPR_b) pour 1000 partitions topologiques identiques à celle obtenue avec IC-Finder mais dont l'ordre a été mélangé de manière aléatoire. Quand on compare ces couples, au « vrai » couple (FDR_b, TPR_b), on trouve pour les deux paramètres des p-valeurs inférieures à 10^{-8}. Ceci nous apprend donc qu'il existe une corrélation entre partition épigénomique et partition topologique qui n'est pas imputable aux tailles de domaines formant les partitions. Sur la figure 3.3.1, l'étude des couples (FDR_d, TPR_d) mènent exactement à la même conclusion mais vue sous un angle différent : deux loci appartenant au même domaine épigénomique appartiennent aussi, la plupart du temps, au même domaine topologique (FDR autour de 0.2), par contre, souvent, deux loci appartenant au même domaine topologique n'appartiennent pas au même domaine épigénomique (TPR autour de 0.3) ce qui est dû à cette différence de taille moyenne entre domaines topologiques et domaines épigénomiques.

3.3.2 Contacts préférentiels entre loci de même état épigénomique

L'observation des cartes Hi-C révèlent la présence de TADs et de compartiments d'interaction à longue portée, suggérant que les monomères de même état épigénomique interagissent entre eux préférentiellement (cf. par exemple sur la figure 3.2.1 interactions entre deux TADs polycomb respectivement à 12.3Mb et 12.7Mb). Afin de quantifier ce phénomène, on s'intéresse au nombre de contacts réalisés par les bins i composé d'une proportion S_i^μ de l'état épigénomique μ et qui interagissent avec des bins j composé d'une proportion S_j^ν de l'état ν et situé à une distance épigénomique s. En sommant sur tout les bins i et j on obtient le nombre $N^{\mu\nu}(s)$:

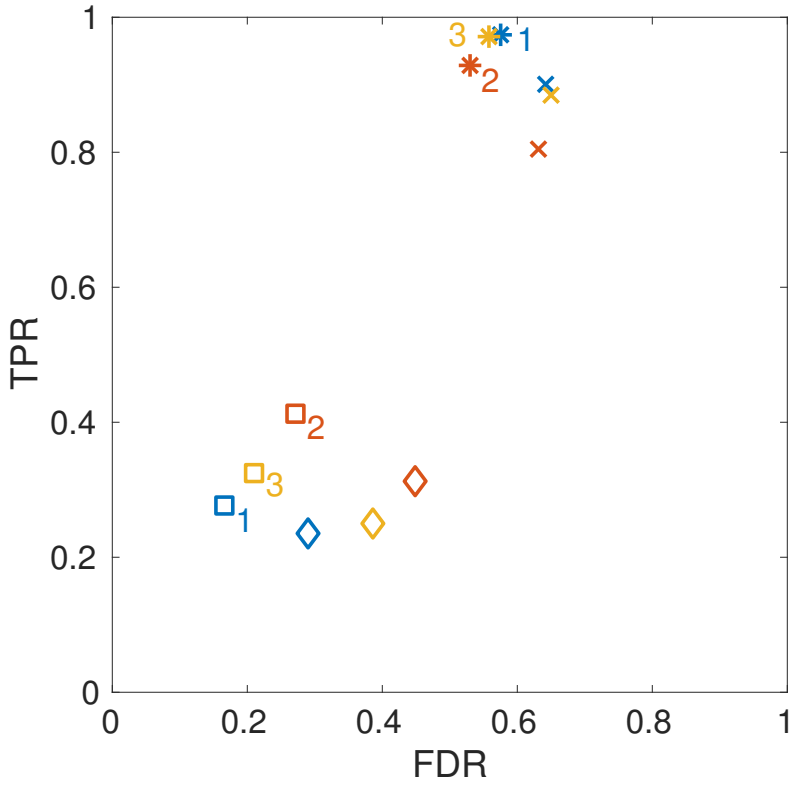

Figure 3.3.1 – **Taux de vrai positif, TPR, en fonction du taux de fausse découverte, FDR, entre la partition épigénomique** I **et la partition topologique donnée par IC-Finder** (I = 1, 2, *et* 3 **[resp. Ho et al., 2014; Filion et al., 2010; Kharchenko et al., 2011]).** Les carrés et les diamants sur la figure indiquent les valeurs (FDR_d, TPR_d) respectivement dans le cas exact et dans le cas où l'ordre des domaines dans la partition topologique est inversé. De la même manière, les astérisques et les croix correspondent aux couples (FDR_b, TPR_b) respectivement dans le cas exact et dans le cas où l'ordre des domaines dans la partition topologique est inversé. Les résultats obtenus respectivement avec [Ho et al., 2014; Filion et al., 2010; Kharchenko et al., 2011] sont en bleu, rouge et jaune.

$$N^{\mu\nu}(s) = \sum_{|i-j|=s} C(i,j)\, S_i^{\mu} S_j^{\nu} \tag{3.3}$$

avec C les contacts Hi-C expérimentaux donnés par [Sexton et al., 2012], on considère tout les chromosomes sauf le 4 dont l'état épigénomique est majoritairement de type HP1. Pour les données épigénomiques, on considérera les 16 états définis par [Ho et al., 2014] qu'on regroupera ensuite en 4 familles. La proportion de chacun des 16 états pour le génome de la drosophile se trouve figure 3.3.2.

Sur la figure 3.3.3A, on présente quatre lignes correspondant du haut vers le bas aux valeurs de $N^{nul\,\nu}(s)$, $N^{actif\,\nu}(s)$, $N^{hétérochromatine\,\nu}(s)$, $N^{polycomb\,\nu}(s)$. Selon une idée de Ralf Everaers, ces valeurs $N^{\mu\nu}(s)$ sont représentées sous forme de diagrammes circulaires : les portions

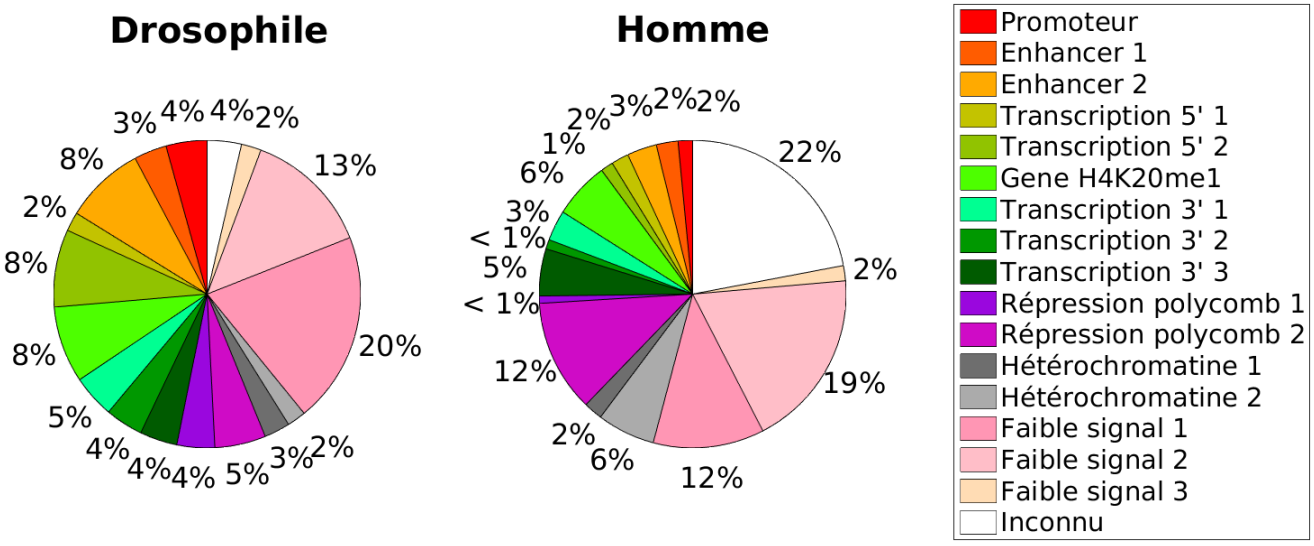

FIGURE 3.3.2 – **Répartition des 16 états épigénomiques identifiés par [Ho et al., 2014] pour l'ensemble du génome de la drosophile (à gauche) et de l'homme (à droite).**

respectivement noire, rouge, verte et bleue représentent les valeurs de ν. En transparence, on voit la répartition moyenne des contacts (idem à celle de la figure 3.2.2 mais en excluant le chromosome 4 de notre étude). L'information sur la distance génomique s est représentée par les anneaux : le premier anneau (le plus proche du centre) correspond à la somme des contacts ayant lieu pour $s \leqslant 20\text{kb}$, les 2ème, 3ème, 4ème et 5ème anneaux correspondent respectivement à des valeurs de s telles que, $20 < s \leqslant 100\text{kb}$, $100 < s \leqslant 500\text{kb}$, $0.5 < s \leqslant 5\text{Mb}$ et $5\text{Mb} < s$. Le 6ème anneau représente les contacts inter chromosomiques. De plus, la surface de chaque anneau est proportionnelle au nombre de contact. On retrouve nos observations précédentes selon lesquelles environ 1/3 des contacts se font pour des distances génomiques inférieures à 1Mb et 1/3 des contacts sont inter-chromosomiques. Ce résultat est indépendant de l'état épigénomique local.

Ces diagrammes (à gauche) obtenus avec les données expérimentales [Sexton et al., 2012] sont à comparer aux diagrammes (à droite) réalisés en moyennant le nombre de contacts expérimental pour chaque distance génomique, c'est ce qu'on appelle le cas générique (cf. Fig. 3.1.4). Pour les modèles expérimental et générique, à très petite échelle on observe pour les bins dans l'état μ un net enrichissement en contacts avec d'autre bins d'état μ aussi, ceci est une conséquence de l'existence des domaines épigénomiques. Aux échelles plus grandes cet enrichissement est toujours présent dans le cas du modèle expérimental alors qu'il l'est moins dans le cas du modèle générique, ceci suggère qu'il existe des interactions préférentielles entre bins de même état epigénomique. On peut tout de même noter des particularités selon l'état

épigénomique, en particulier, les bins majoritairement polycomb présentent, à petite échelle (jusqu'à 0.5Mb), un enrichissement de contacts avec d'autres bins polycomb alors que pour les bins majoritairement actifs, l'enrichissement est principalement présent, à grande échelle (au delà de 0.5Mb).

Concernant, les contacts inter chromosomiques il ne semble pas y avoir de nette différence entre les modèles expérimental et générique.

En conclusion, ces diagrammes circulaires nous ont renseignés sur la composition locale autour d'un site génomique. On a vu qu'il y avait des interactions préférentielles entre loci de même état épigénomique ce qui fait qu'on ne peut pas considérer la chromatine comme objet homogène. Cependant, il faut noter que cet effet n'est pas très fort, il n'y a clairement pas de séparation en microphase totale. Ceci suggère que les cartes expérimentales présentent une perturbation faible par rapport à un modèle neutre (c'est-à-dire sans interactions spécifiques entre loci de même état).

On propose maintenant de calculer les facteurs d'amplification de contact entre modèle expérimental et modèle générique en sommant les contacts quel que soit la distance génomique s (toujours selon une idée de Ralf Everaers). Le coefficient d'amplification $A_{\mu\nu}$ entre les états épigénomiques μ et ν est donné par la formule ci-dessous :

$$A_{\mu\nu} = \frac{1}{\alpha} \frac{\sum_{s=0}^{N-1} N^{\mu\nu}_{\text{expérimental}}(s)}{\sum_{s=0}^{N-1} N^{\mu\nu}_{\text{générique}}(s)} \tag{3.4}$$

$$\text{avec } \alpha = \frac{\sum_{\mu\nu} \sum_{s=0}^{N-1} N^{\mu\nu}_{\text{expérimental}}(s)}{\sum_{\mu\nu} \sum_{s=0}^{N-1} N^{\mu\nu}_{\text{générique}}(s)}$$

La matrice des $A_{\mu\nu}$ est présentée 3.3.3B. On retrouve proche de la diagonale des coefficients positifs en échelle $\log_{10}$ ce qui confirme l'enrichissement des contacts. Aussi, on remarque que certains états épigénomiques présentent des motifs d'interaction similaires et qu'il semble y avoir 6 grandes familles épigénomiques : un groupe actif comprenant les promoteurs et les états de transcription, un groupe d'enhancers, un groupe actif riche en introns (états « transcription 5' 2 » et « gene, H4K20me1 »), un groupe polycomb, un groupe nul et un groupe d'hétérochromatine. On verra dans la sous section suivante qu'à l'échelle des TADs aussi on retrouve ces mêmes familles (Fig. 3.3.5).

Au lieu de calculer la matrice $A_{\mu\nu}$ pour l'ensemble du génome, nous pouvons le faire en ne considérant que les contacts entre loci à moins de 1Mb l'un de l'autre (Fig. 3.3.4). On observe que dans les deux cas, les résultats sont sensiblement identiques avec toutefois moins d'amplification des contacts entre marques actives si on ne considère pas tout le génome.

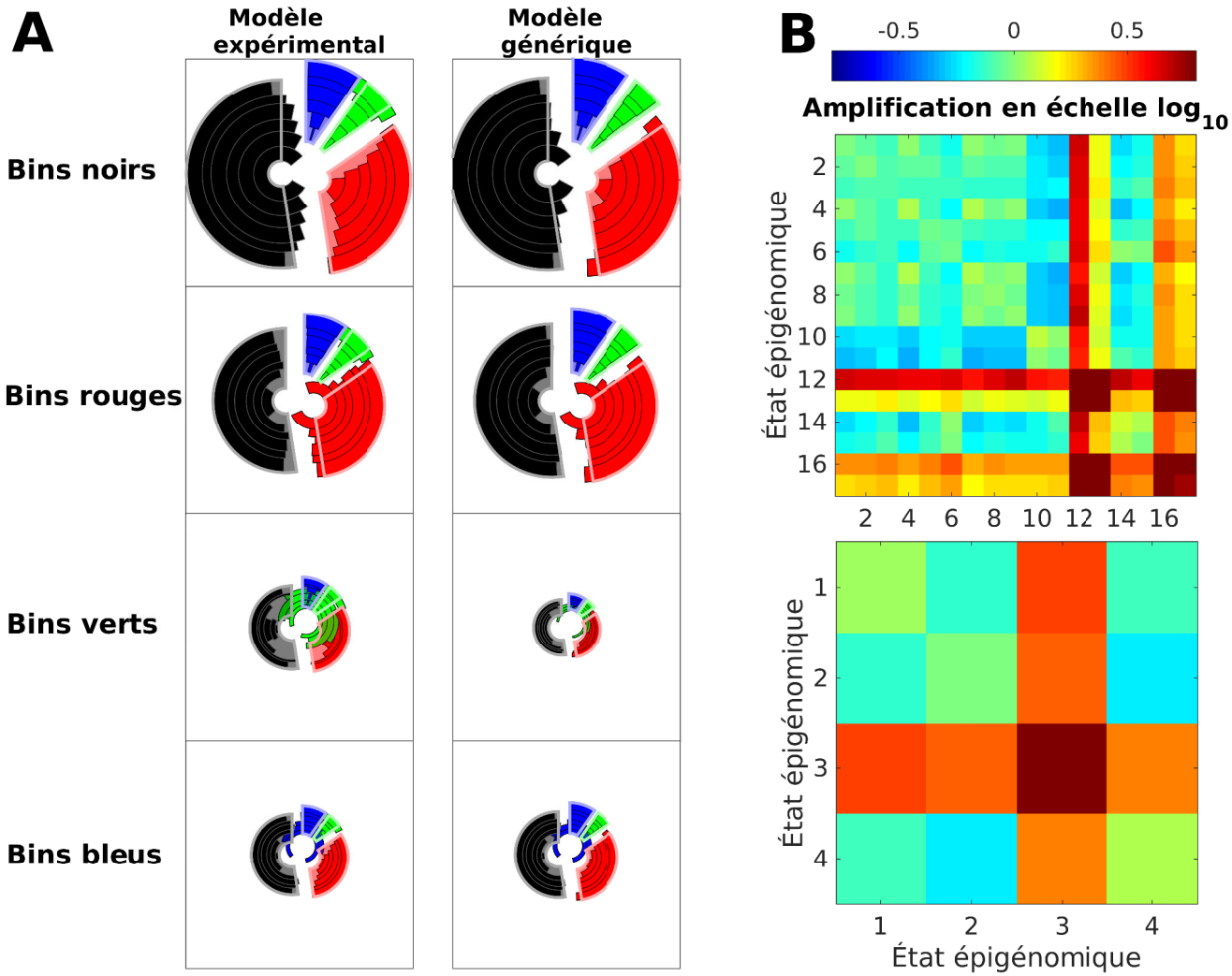

FIGURE 3.3.3 – **Enrichissement des contacts en fonction de l'état épigénomique dans le cas de la drosophile (chromosome 4 exclu de l'étude) [(Sexton et al., 2012; Ho et al., 2014)].** (A) Nombre $N^{\mu\nu}(s)$ (Éq. 3.3) représenté selon l'idée de Ralf Everaers sous forme d'anneaux dans des diagrammes circulaires avec s la distance génomique entre bins en contact. Ces nombres permettent d'évaluer l'environnement 3D des bins en fonction de l'épigénome et en fonction de s. Les cinq premiers anneaux du centre vers la périphérie correspondent respectivement à des valeurs de s telles que $s \leqslant 20\text{kb}$, $20 < s \leqslant 100\text{kb}$, $100 < s \leqslant 500\text{kb}$, $0.5 < s \leqslant 5\text{Mb}$ et $5\text{Mb} < s$. Le 6ème anneau renseigne sur les contacts inter chromosomiques. L'aire des portions est proportionnelle au nombre de contacts. La répartition moyenne des nombres de contact en fonction de l'état épigénomique, pour l'ensemble des chromosomes sauf le 4, est représentée en transparence sur chaque diagramme circulaire. (B) Matrices d'amplification (Éq. 3.4) en haut pour les 16 états épigénomiques dont les numéros sont définis sur la figure 3.3.2 et en bas pour les quatre familles obtenues après regroupement de ces 16 états.

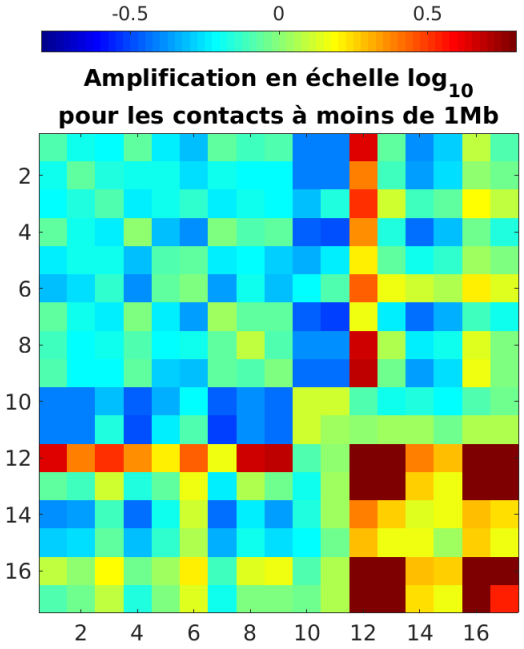

Figure 3.3.4 – **Carte d'amplification des contacts dans le cas de la drosophile (chromosome 4 exclu de l'étude) [(Sexton et al., 2012; Ho et al., 2014)] et en ne considérant que les contacts entre loci espacés de moins de 1Mb en distance génomique.**

3.3.3 Corrélation à l'échelle des compartiments d'interaction (TADs et hiérarchie supérieure)

Précédemment on a étudié les corrélations entre épigénome et contacts génomiques à l'échelle des monomères, dans cette sous section, on étudie comment l'information épigénomique est associée à la compartimentation 3D (TADs et compartiments de hiérarchie supérieure). Pour des embryons tardifs de drosophile et pour la lignée cellulaire humaine GM12878, on récupère les cartes Hi-C de chaque chromosome respectivement dans [Sexton et al., 2012] et dans [Rao et al., 2014]. On applique IC-Finder à ces cartes de contact avec l'option de ré-échantillonnage dans un premier temps avec les paramètres par défaut et dans un second temps avec $\sigma_0 = 5$ ce qui permet d'accéder à une hiérarchie de compartimentation supérieure (Fig. 2.3.5). On se demande, ici, si, pour les deux organismes étudiés, les caractéristiques du repliement de la chromatine obtenues avec IC-Finder corrèlent avec l'information épigénomique donnée par [Ho et al., 2014] (Fig. 3.3.2). Dans notre analyse, on ne considère pas les régions centromériques et pericentromeriques principalement composées d'hétérochromatine constitutive et nulle.

Pour chaque paire d'états épigénomiques (μ, ν), on estime l'enrichissement en colocalisation intra-TAD $C_{\mu\nu}$ caractérisant si un locus d'état μ est susceptible de partager le même com-

partiment d'interaction avec un autre locus d'état ν. On détaille ci-dessous la définition de la matrice $C_{\mu\nu}$:
Pour un locus donné i, on définit S_i^μ la proportion de l'état épigénomique μ dans le locus ($0 \leqslant S_i^\mu \leqslant 1$). Pour une paire d'états épigénomiques (μ, ν), l'enrichissement en colocalisation intra-compartiment $C_{\mu\nu}$ est défini par l'équation 3.5 :

$$C_{\mu\nu} = \frac{1}{N_{\mu\nu}^{intra}} \frac{\sum_{i\neq j} S_i^\mu S_j^\nu p_d(i,j)}{\sum_{i\neq j} p_d(i,j)} \qquad (3.5)$$

$$\text{avec } N_{\mu\nu} = \frac{\sum_{i\neq j} S_i^\mu S_j^\nu}{\sum_{i\neq j} 1}$$

avec $p_d(i,j)$ la probabilité que i et j soient prédits comme appartenant au même compartiment d'interaction. La matrice $C_{\mu\nu}$ représente la valeur moyenne de $S_i^\mu S_j^\nu$ pour les loci colocalisant dans le même compartiment normalisée par la valeur correspondante le long du génome $N_{\mu\nu}$. Une valeur $C_{\mu\nu} > 1$ (resp. $C_{\mu\nu} < 1$) indique que les associations, dans un même compartiment d'interaction, de loci d'états épigénomiques μ et ν sont enrichies (resp. appauvries).
On calcule donc les matrices $C_{\mu\nu}$ avec les 16 états épigénomiques données dans [Ho et al., 2014] et avec les probabilités de colocalisation dans un même TAD donnés par IC-Finder (Fig. 3.3.5A,D). Un regroupement hiérarchique standard (avec distance de corrélation et mesure de similarité entre classes moyenne) montre que des groupes d'états se forment et que les interactions dans les TADs entre ces différents groupes sont appauvris (Fig. 3.3.5B,E gauche). Cela suggère fortement que la composition épigénomique des TADs est relativement homogène dans de tels groupes. Pour la drosophile, nous avons trouvé 6 familles épigénomiques : un groupe actif comprenant les promoteurs et les états de transcription, un groupe d'enhancers, un groupe actif riche en introns (états « transcription 5' 2 » et « gene, H4K20me1 »), un groupe polycomb, un groupe nul (peu de signal) et un groupe d'hétérochromatine. Pour l'homme, quatre familles seulement sont ressorties de notre analyse : un groupe actif, un groupe d'enhancers, un groupe polycomb et un groupe d'hétérochromatine et de signal faible. Ces familles sont donc très similaires entre les deux espèces avec toutefois des exceptions soulignant les spécificités de régulation épigénomique propre à chaque espèce. Par exemple, H4K20me1 est présent dans les introns de longs gènes actifs de drosophile alors qu'il est associé à des gènes réprimés par polycomb chez l'homme [Ho et al., 2014].
Pour chaque locus d'une famille épigénomique donnée (six familles dans le cas de la drosophile et quatre pour l'homme), on calcule la distance relative à la frontière du compartiment la plus proche : une distance relative nulle signifie donc que le locus se situe au niveau d'une frontière alors qu'une distance relative de 0.5 signifie que le locus se trouve au centre d'un

compartiment d'interaction. Les figures 3.3.5C,F à gauche montrent la distribution cumulée de ces distances pour chacune des familles. En général, on observe que les familles actives sont enrichies au niveau des frontières alors que les familles inactives sont plus enrichies au coeur des TADs. Avec tout de même une notable exception chez la drosophile pour qui on voit que la famille hétérochromatine est enrichie aux frontières, ce qui illustre la partition de l'hétérochromatine constitutive en petits domaines (Fig. 3.3.6) pour cet organisme (à l'exception des régions centromériques et péricentromériques que nous ne considérons pas dans notre étude) [Sexton et al., 2012]. On observe que la corrélation entre épigénome et position des TADs (enrichissement ou appauvrissement aux frontières) est plus prononcée chez la drosophile.

Afin d'examiner si l'organisation du génome à plus haute échelle que les TADs est aussi corrélée à l'épigénome, on calcule de nouveau l'enrichissement en colocalisation intra-compartiment $C_{\mu\nu}$ et la distribution des distances relatives à la frontière la plus proche pour les six et quatre familles épigénomiques mais cette fois avec des compartiments d'une hiérarchie supérieure à celle des TADs (on ne lance plus IC-Finder avec les paramètres par défaut mais avec $\sigma_0 = 5$). On peut voir sur les figures 3.3.5B,E à droite (et aussi figure 2.3.5B,D pour une illustration) une perte nette de colocalisation pour les familles actives chez la drosophile alors que les autre familles sont toujours auto-enrichies dans les compartiments. De manière assez surprenante, excepté la famille hétérochromatine qui reste isolée dans de petits compartiments d'interaction, les autres familles sont maintenant pauvrement isolées les unes des autres par des compartiments d'interaction ($C_{\mu\nu} \approx 1$). Cette observation suggère qu'à ce degré d'organisation, les petits domaines actifs ont été fusionnés avec des plus grands inactifs, résultant en des compartiments moins bien épigénomiquement définis (Fig. 2.3.5B dans le chapitre précédent). Toutefois, les frontières de tels compartiments restent enrichis en marques actives. Chez l'homme, à ce niveau d'organisation, les familles restent partiellement auto-colocalisées et isolées, alors que les frontières sont seulement faiblement enrichies en familles actives. Il est intéressant de noter que l'enrichissement en colocalisation intra-TAD diminue légèrement pour les familles inactives, alors qu'il augmente (légèrement aussi) pour les familles actives, ceci suggère qu'à ce niveau de hiérarchie, les compartiments d'interaction sont mieux associés avec la chromatine active (Fig. 2.3.5D).

3.4 Conclusion

Nous avons vu dans ce chapitre que les données Hi-C combinées aux données épigénomiques (malgré leurs incertitudes) suggèrent chez la drosophile et chez l'homme que les loci de même état épigénomique interagissent spécifiquement entre eux. Dans le chapitre suivant

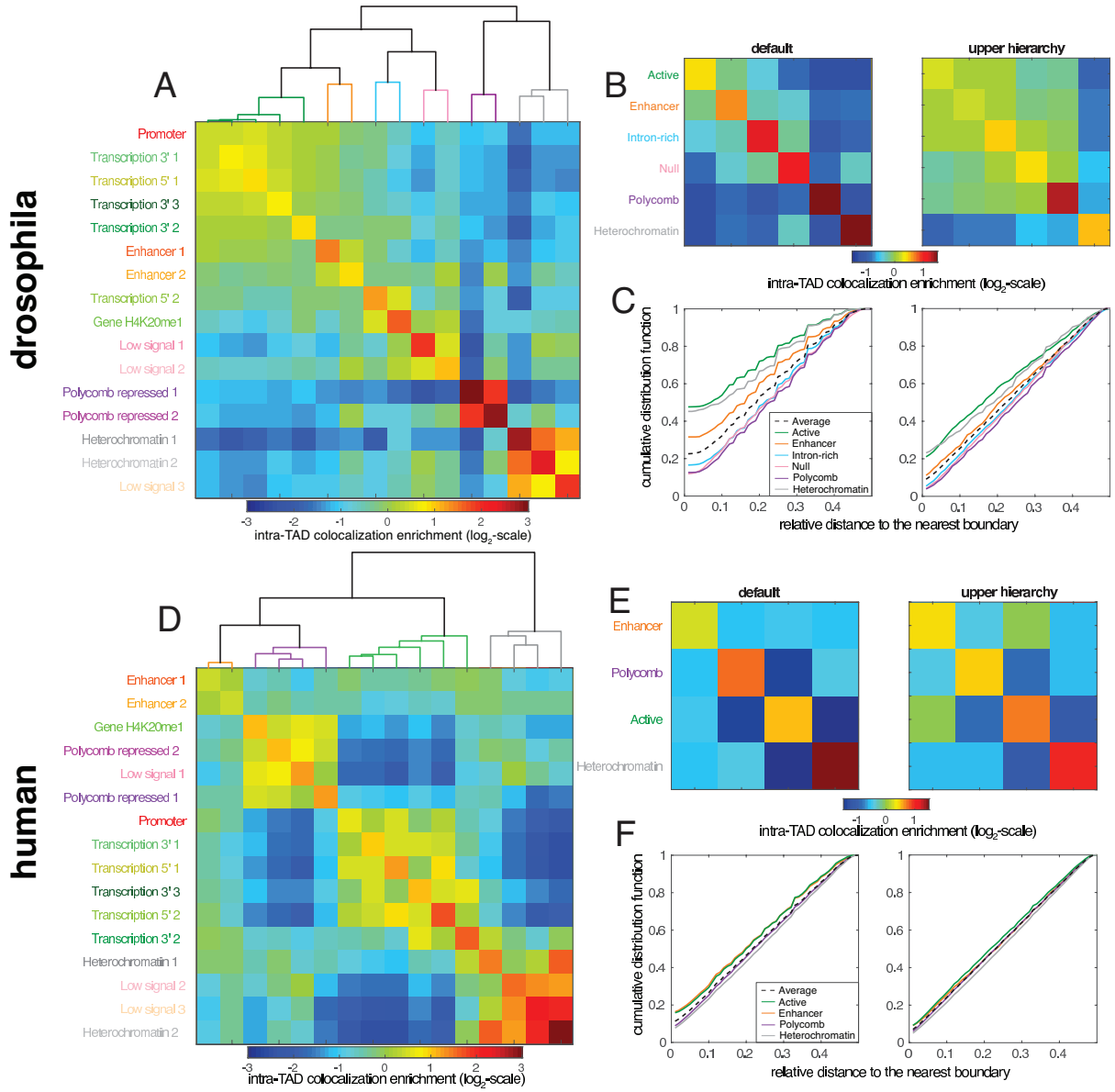

FIGURE 3.3.5 – **Couplage entre compartimentation 3D et information épigénomique.** (A,D) Enrichissement en colocalisation intra-TAD des 16 états épigénomiques, $C_{\mu\nu}$, pour la drosophile (A) et pour l'homme (D). Au dessus, de chaque matrice se trouve un dendrogramme groupant hiérarchiquement les différentes colonnes (c'est-à-dire les différents états), on retient 6 familles pour la drosophile et 4 pour l'homme. (B,E) Matrices $C_{\mu\nu}$ calculées par famille épigénomique plutôt que par état, à l'échelle des TADs à gauche et à une échelle plus grande à droite (compartiments obtenues avec $\sigma_0 = 5$ dans IC-Finder). (C,F) Distribution cumulée des distances relatives à la frontière la plus proche pour chaque famille. Les courbes noires représentent le comportement moyen.

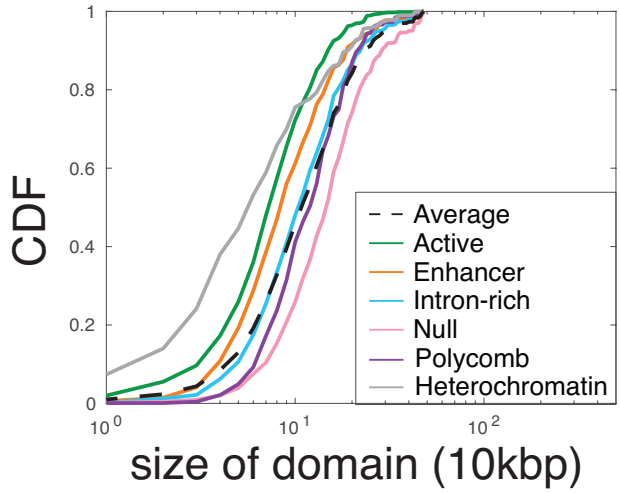

Figure 3.3.6 – Fonction de distribution cumulée (CDF) de la taille des domaines pour des loci appartenant à une famille épigénomique donnée, dans le cas de la drosophile. La courbe noire représenter la CDF moyenne.

on se demandera si l'existence d'interactions pilotées par l'épigénome permet d'expliquer l'organisation 3D de la chromatine.

CHAPITRE 4

MODÉLISATION DE LA CHROMATINE PAR UN COPOLYMÈRE PAR BLOC

Nous avons décrit dans les chapitres précédents l'existence d'une corrélation entre domaines topologiques (ou TADs) et domaines épigénomiques. Nous avons vu qu'un locus dans un certain état épigénomique colocalisera majoritairement avec des loci de même état au sein d'un TAD. Ainsi, dans ce chapitre, on s'intéresse à la modélisation de la structure tridimensionnelle de la chromatine sous l'hypothèse que la séquence épigénomique joue un rôle majeur dans l'organisation 3D de la chromatine. On commence par présenter succintement quelques propriétés des hétéropolymères justifiant l'idée de modéliser la chromatine par un copolymère par bloc dont les propriétés dépendent de l'état épigénomique de chacun des monomères. On détaille alors le formalisme et les équations correspondantes, décrivant la dynamique de la chromatine. Enfin, on présente trois approches permettant la résolution ou la simulation de ces équations dans le but d'être prédictif. La première consiste à réaliser une approximation « gaussienne auto-cohérente » permettant l'étude rapide et fiable de petits segments chromosomiques isolées. Les deux approches suivantes relèvent de la dynamique sur réseau avec simulations de Monte-Carlo cinétique et de la dynamique moléculaire. Ces deux dernières approches vont permettre de valider la première en précisant son cadre d'application.

4.1 Introduction

4.1.1 Motivations

Comme nous l'avons vu lors de l'état de l'art dans le chapitre 1, modéliser la chromatine par un homopolymère, tout en tenant compte de certaines contraintes topologiques, permet de reproduire la structure à grande échelle de la chromatine. Toutefois, ce modèle ne permet pas de reproduire certaines observations expérimentales telle que la présence de TADs pourtant visibles sur les cartes Hi-C expérimentales (cf. chapitre 1). L'existence des TADs peut s'expliquer par le fait que les interactions entre différents loci de la chromatine ne soient pas homogènes, en particulier, les loci de même état épigénomique interagissent préférentiellement (cf. chapitre 3). De récentes expériences biochimiques suggèrent que ces interactions préférentielles ont pour intermédiaires des protéines architecturales telles que des insulateurs ou des cohésines et d'autres médiateurs [Sofueva et al. (2013)] qui se regroupent dans l'espace et forment alors des ponts physique entre des sites de régulation génomiquement distants [Isono et al., 2013 ; Francis et al., 2004 ; Lo et al., 2012 pour les protéines du groupe polycomb ; Canzio et al., 2013 pour HP1]. Ces constats nous motivent à complexifier le modèle d'homopolymère en modélisant la chromatine par un hétéropolymère.

Un hétéropolymère a la particularité d'être constitué de plusieurs types de monomères. Le cas le plus simple que l'on puisse envisager est un copolymère par bloc constitué de deux types de monomères A et B dont les propriétés d'interactions, A-A, B-B et/ou A-solvant, B-solvant diffèrent, on considérera pour l'exemple que les monomères de type A (resp. B) sont hydrophiles (resp. hydrophobes). Selon l'intensité des différentes interactions sus-citées, ces copolymères peuvent se replier selon différentes phases. En particulier, comme les monomères de type A et B sont en solution très peu miscibles, une démixtion peut se produire, le copolymère se replie alors selon une organisation en micro-phases « pures » où les interactions entre même type de monomères (ou entre un type de monomère et le solvant) sont maximisées. Des images de cette phase à l'équilibre obtenues par microscopie électronique en transmission présentent des ressemblances avec les images de la chromatine obtenues expérimentalement (Fig. 4.1.1). Ceci suggère que l'organisation de la chromatine en micro-domaines 3D (par exemple les TADs) est caractéristique du repliement d'un copolymère par blocs sous certaines conditions.

Finalement, ces observations et ces quelques connaissances théoriques sur les hétéropolymères associées au résultat expérimental selon lequel les loci de la chromatine de même état épigénomique interagissent préférentiellement nous ont fortement motivés à supposer que le repliement de la chromatine est en partie dirigé par des interactions effectives dépendant de l'épigénome. Afin de tester cette hypothèse, on se propose de modéliser la chromatine

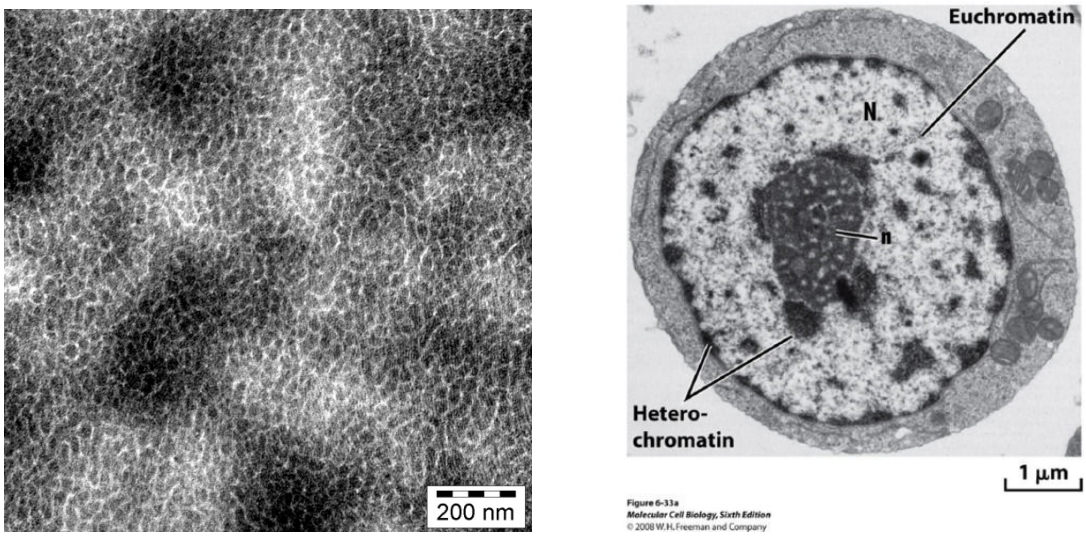

FIGURE 4.1.1 – **Ressemblance entre un copolymère par bloc et la chromatine.** A gauche : image d'un long hétéropolymère amphiphile ou autrement dit d'un copolymère par bloc dont les monomères sont hydrophiles ou hydrophobes réalisée par microscopie électronique en transmission [Scherble et al. 2001]. A droite : matériel nucléaire observé aussi par microscopie électronique en transmisssion (technique de cryo-microscopie électronique pour préparer l'échantillon). Les zones qui paraissent plus denses, en gris foncé, correspondent à l'hétérochromatine, les régions plus claires constituent l'euchromatine. Les notations « n » et « N » sur la figure correspondent respectivement au nucléole et au noyau.

par un copolymère par bloc dont les propriétés d'interaction chaîne-chaîne et chaîne-solvant dépendent des propriétés épigénétiques locales. Chaque bloc correspond à un domaine épigénomique et chaque monomère interagit préferentiellement avec des monomères de même état épigénomique (Fig. 4.1.2). Cette modélisation de la chromatine par un hétéropolymère permet de tenir compte d'une part des effets polymériques qui comme nous l'avons vu dans le chapitre 1 permettent d'expliquer la structure à grande échelle de la chromatine et permet d'autre part de tenir compte des interactions spécifiques entre états chromatiniens de même type ce qui permettra entre autre de reproduire la formation des TADs.

4.1.2 Modélisation du copolymère

On décrit la chromatine par un modèle de copolymère linéaire, non homogène auto-évitant et dont les monomères ont chacun une couleur caractéristique de leur état épigénomique (Fig. 4.1.2). La conformation d'une telle chaîne est contrôlée d'une part par des effets polymériques (connectivité, flexibilité, volume exclu) et d'autre part, par des interactions effectives à courte portée entre monomères. La particularité majeure du modèle est que les monomères de même état épigénomique interagissent préférentiellement. Nous allons décrire dans les sous sections

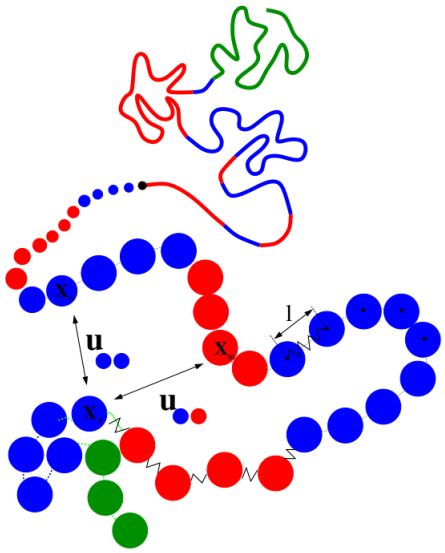

Figure 4.1.2 – **Modélisation de la chromatine par un copolymère par bloc.** On modélise la chromatine par un hétéropolymère ou copolymère par bloc dont chaque monomère a une couleur représentative de son état épigénomique, par exemple rouge, bleu, noir ou vert sur cette figure. On choisit de travailler avec des monomères de 10kb. Ils sont successivement liés les uns aux autres par des interactions de type ressort et ils sont typiquement espacés d'une longueur $\mathfrak{l}$. On modélise la chaîne comme étant auto-évitante et en interaction avec elle même. Le repliement d'une telle chaîne est dirigé par des interactions non spécifiques (par exemple $\mathfrak{u}_{\bullet\bullet}$ sur la figure) caractéristiques du confinement ou du volume exclu et par des interactions attractives spécifiques dépendant de l'épigénome (par exemple $\mathfrak{u}_{\bullet\bullet}$ sur la figure) : les monomères de même état épigénomique interagissent préférentiellement.

qui suivent trois approches permettant la description de la dynamique de la conformation 3D d'une telle chaîne.

(1) La première consiste à réaliser une approximation « gaussienne auto-cohérente ». J'introduirai tout d'abord des propriétés sur les distributions gaussiennes multivariées qui nous seront utiles. Ensuite, je présenterai le formalisme associé à cette approche « gaussienne auto-cohérente ». Ce formalisme a été développé par Jost et al. avant que ne débute ma thèse et a mené à une publication dans la revue Nucleic Acids Research [Jost et al. 2014]. Au cours de cette thèse, j'ai amélioré le modèle de 2014 avec notamment une meilleure prise en compte du volume exclu et avec l'introduction d'un champ générique modélisant en particulier les effets de crumpling jusque là mis de côté. Aussi, nous avons développé une méthode de résolution de l'équation de la dynamique bien plus rapide que la méthode précédemment utilisée. Nous verrons que ces améliorations mènent globalement aux mêmes résultats qualitatifs que ceux publiés dans Jost et al. 2014 mais permettent d'obtenir des résultats quantitatifs bien plus réalistes. La description de ce modèle « amélioré » ainsi que les résultats que j'ai obtenu avec

(diagramme de phase, prévision de structures, ...) ont été soumis en vue d'être publiés dans Chromosome Research.
(2) La seconde approche repose sur l'étude de la dynamique sur réseau par des simulations de Monte-Carlo cinétique. Je présenterai tout d'abord le formalisme associé à cette approche puis quelques travaux préliminaires que j'ai codé avec le soutien de Max Kolb et qui m'ont permis de vérifier des lois caractéristiques des homopolymères. Ensuite, je décrirai les diagrammes de phase d'hétéropolymères réalisés par Juan Olarte Plata qui était un étudiant M2 ayant repris la suite de ce travail dans le cadre de son stage sous la direction de Cédric Vaillant et Daniel Jost. Enfin, dans le but de valider le modèle de copolymère sur réseau, on comparera prédictions et expériences via une analyse de la compaction en fonction de la taille des domaines que j'ai réalisée à partir de cartes Hi-C expérimentales de drosophile. L'ensemble de ce travail a été publié dans Physical Biology [Olarte-Plata et al., 2016].
(3) La troisième approche consiste en la réalisation de simulations de dynamique moléculaire. Je présenterai la méthode développée par Pascal Carrivain au cours de ses post-doctorats (actuellement en post-doc à l'ENS de Lyon). J'exposerai ensuite une analyse statistique que j'ai réalisée avec les données qu'il m'a fournies. Comme nous le verrons, cette analyse des simulations sera accomplie avec les mêmes outils que ceux introduits dans le chapitre 3 pour l'analyse des données expérimentales. Plusieurs de ces outils ont été proposés par Ralf Everaers. Enfin, je présenterai les résultats de Pascal concernant le temps d'équilibration car cela nous permet de justifier qu'avec l'approche gaussienne auto-cohérente on décide de ne s'intéresser qu'aux solutions stationnaires.
Dans les trois cas, chaque monomère représentera 10kb. Cette taille de monomères est un bon compromis entre résolution du système et efficacité numérique. Cela nous permettra de comparer les cartes de contact obtenues in silico avec les cartes expérimentales dont les résolution à 10kb présente un bon rapport signal/bruit [Sexton et al., 2012].

4.2 Approche gaussienne auto-cohérente

Dans cette partie, on propose d'étudier le repliement de la chromatine avec le modèle de copolymère par bloc accompagné d'une approximation qu'on appelle « approximation gaussienne auto-cohérente ». Cette dernière consiste à approximer, à chaque pas de temps, la distribution de probabilité de l'ensemble des positions des monomères, $\mathsf{P}\left(\{\mathsf{X}_i\}_{1\leqslant i\leqslant N}\right)$, par une distribution gaussienne multivariée, ce qui permet de réécrire l'équation de Fokker-Planck non linéaire vérifiée par $\mathsf{P}\left(\{\mathsf{X}_i\}_{1\leqslant i\leqslant N}\right)$ en une équation auto-cohérente dont la résolution est facilitée. Cette méthode a été initialement développée par Jost et al., 2014. On commencera dans cette partie par présenter quelques propriétés des distributions gaussiennes multivariées

qui nous seront utiles. Ensuite, on mettra en place l'approximation gaussienne auto-cohérente qui nous mènera à l'équation de la dynamique auto-cohérente vérifiée par la matrice des distances quadratiques moyennes, D. Cette équation dépend directement de l'hamiltonien H du système, on présentera plusieurs formes d'hamiltonien envisageables pour décrire le copolymère. Enfin, pour clôturer le descriptif de la méthode, on présentera deux algorithmes permettant la résolution de l'équation de la dynamique dans le cas stationnaire. On exposera alors après des résultats obtenus avec ce formalisme. On développera en particulier les propriétés génériques d'un homopolymère, d'un hétéropolymère simple composé de deux types de blocs et d'un hétéropolymère dont l'état des blocs est défini selon l'épigénome de la drosophile. Les résultats développés seront des propriétés d'ensemble (pas d'informations sur les trajectoires individuelles des monomères), à l'équilibre, et concerneront des chaînes de l'ordre de $1\mathrm{Mb}$. L'avantage de l'approche gaussienne auto-cohérente est sa capacité à donner des résultats rapidement en ne nécessitant aucunes simulations.

4.2.1 Distributions gaussiennes multivariées

Soit une chaîne polymérique de N monomères. L'ensemble des positions des monomères est noté $Y = \{X_i\} = \{X_1^x, X_1^y, X_1^z, X_2^x, ..., X_N^z\}$, avec $i \in [\![1; N]\!]$. Comme nous le verrons par la suite, l'approche gaussienne auto-cohérente consiste à intégrer les équations décrivant la dynamique de la chaîne polymérique en approchant à chaque pas de temps la distribution de probabilité de l'ensemble des positions des monomères, $P(Y = \{X_i\}, t)$, par une distribution gaussienne multivariée de moyenne nulle, P_G (idée initialement développée dans Jost et al. (2014)).

$$P(Y = \{X_i\}, t) \approx P_G = \frac{1}{\sqrt{(2\pi)^{3N} \times |\det(\mathcal{C}(t))|}} \exp\left(-X^\dagger \mathcal{C}^{-1}(t) X/2\right) \tag{4.1}$$

avec $\mathcal{C}(t)$ la matrice de covariance de $X = \{X_i\}$ à l'instant t. Par isotropie du système dans les trois directions de l'espace, $\mathcal{C}(t) = \{\langle X_i X_j \rangle / 3\}$ et $\langle X \rangle = 0$. Dans le cadre de cette approximation gaussienne, la matrice de covariance $\mathcal{C}(t)$ caractérise entièrement notre chaîne.
A une distribution gaussienne multivariée, on peut associer un hamiltonien, H_G explicité ci-dessous :

$$H_G = {}^t X \frac{\mathcal{K}}{2} X = \frac{1}{2} \sum_j \left(\sum_i X_i \mathcal{K} \right) X_j \tag{4.2}$$

avec $\mathcal{K}$ les constantes d'interaction entre monomères et avec $\mathcal{K} = \mathcal{C}^{-1}$ par définition.
Cet hamiltonien décrit une chaîne purement gaussienne, c'est-à-dire une chaîne dont les monomères ne sont soumis qu'à des interactions de type ressort avec leur(s) voisin(s). La

conformation d'une telle chaîne est assimilable à une marche aléatoire. La matrice $\mathcal{K}$ correspondante à cette chaîne est de la forme symétrique suivante :

$$\mathcal{K} = 3 \begin{pmatrix} 1 & -1 & 0 & \cdots & \cdots & 0 \\ & 2 & \ddots & \ddots & 0 & \vdots \\ & & \ddots & \ddots & \ddots & \vdots \\ & & & \ddots & \ddots & 0 \\ & & & & 2 & -1 \\ & & & & & 1 \end{pmatrix}$$

Dans cette section, nous allons voir quelques propriétés fondamentales des distributions gaussiennes multivariées qui nous seront utiles par la suite.

4.2.1.1 Relation entre D et $\mathcal{C}$

On définit D comme étant la matrice des tiers des distances au carré moyennes entre monomères, soit $\mathsf{D}_{ij} = \left\langle (\mathsf{X}_i - \mathsf{X}_j)^2 \right\rangle /3$. On sait que les coefficients de la matrice de covariance sont définis par $\mathcal{C}_{ij} = \langle \mathsf{X}_i \mathsf{X}_j \rangle /3$. De ce fait, les matrices C et D par la relation 4.3.

$$\mathsf{D}_{ij} = \mathcal{C}_{ii} + \mathcal{C}_{jj} - 2\mathcal{C}ij \tag{4.3}$$

Par souci de simplification, par la suite, on pourra appeler D la matrice des distances sans préciser qu'il s'agit en fait du tiers des distances au carré moyennes. La convention d'inclure le facteur $1/3$ dans la matrice D permettra d'alléger les expressions par la suite. Cette matrice D est le principal objet de l'approche gaussienne auto-cohérente.

4.2.1.2 Choix de la jauge

La matrice des potentiels d'interaction gaussiens, $\mathcal{K}$, de taille $\mathsf{N} \times \mathsf{N}$ avec N le nombre total de monomères est non inversible puisqu'elle comporte une valeur propre nulle caractéristique du degré de liberté de translation de la chaîne polymérique. On peut donc fixer une contrainte spatiale sur la matrice $\mathcal{K}$ afin d'obtenir une matrice inversible, K, de taille $\mathsf{N}-1 \times \mathsf{N}-1$. Dans le cas général, la jauge peut être ainsi formulée :

$$\textstyle\sum_{k=1}^{N} a_k X_k = 0 \text{ avec } a_1 = 1 \text{ et les autres coefficients } a_k \text{ pour } 2 \leqslant k \leqslant N \text{ libres.}$$

L'expression de la matrice K dépendra de la jauge choisie. On propose ci-dessous de démontrer la relation entre $\mathcal{K}$ et K (Éq. 4.4).

$$\forall (i,j) \in [\![1; N-1]\!]^2 \quad K_{ij} = \frac{1}{2}\left(a_{i+1}a_{j+1}\mathcal{K}_{1,1} + \mathcal{K}_{i+1,j+1} - a_{i+1}\mathcal{K}_{1,i+1} - a_{j+1}\mathcal{K}_{1,j+1}\right) \tag{4.4}$$

Démonstration de la relation 4.4 :

Le principe de la démonstration est d'expliciter l'hamiltonien gaussien, H_G, en fonction de $\mathcal{K}$ et des coefficients de jauge a_i. On introduit la jauge dans l'expression de H_G en remplaçant le coefficient X_1 par $X_1 = -\sum_{k=2}^{N} a_k X_k$.

$$\begin{aligned}
H =_G^t X\frac{\mathcal{K}}{2}X =& \frac{1}{2}\sum_{j=1}^{N}\left(\sum_{i=1}^{N} X_i\mathcal{K}_{i,j}\right)X_j \\
=& \frac{1}{2}X_1^2\mathcal{K}_{1,1} + \sum_{i=2}^{N} X_i^2\frac{\mathcal{K}_{i,i}}{2} + \sum_{j=2}^{N} X_1X_j\mathcal{K}_{1,j} + \sum_{2<i<j}^{N} X_iX_j\mathcal{K}_{i,j} \\
=& \sum_i a_i^2X_i^2\frac{\mathcal{K}_{1,1}}{2} + \sum_{i<j} a_iX_i.a_jX_j\mathcal{K}_{1,1} + \sum_i X_i^2\frac{\mathcal{K}_{i,i}}{2} - \sum_i a_iX_i^2\mathcal{K}_{1,i} - \sum_{i<j} X_iX_j(a_i\mathcal{K}_{1,i} + a_j\mathcal{K}_{1,j}) \\
&+ \sum_{2<i<j} X_iX_j\mathcal{K}_{i,j} \\
=& \sum_{i=2}^{N} X_i^2\left(a_i^2\frac{\mathcal{K}_{1,1}}{2} + \frac{\mathcal{K}_{i,i}}{2} - a_i\mathcal{K}_{1,i}\right) + \sum_{i<j} X_iX_j\left(a_ia_j\mathcal{K}_{1,1} + \mathcal{K}_{i,j} - a_i\mathcal{K}_{1,i} - a_j\mathcal{K}_{1,j}\right) \\
=& \frac{1}{2}\sum_{j=2}^{N}\left(\sum_{i=2}^{N} X_i\left(a_ia_j\mathcal{K}_{1,1} + \mathcal{K}_{i,j} - a_i\mathcal{K}_{1,i} - a_j\mathcal{K}_{1,j}\right)\right)X_j \\
=& \frac{1}{2}\sum_{j=2}^{N}\left(\sum_{i=2}^{N} X_iK_{i,-1,j-1}\right)X_j
\end{aligned}$$

Par identification entre les deux dernières lignes obtenues, on trouve effectivement l'équation 4.4.

Le choix de jauge le plus simple, c'est-à-dire celui qui allège au plus les calculs, consiste à fixer $a_1 = 1$ et $\forall k \in [\![2;\, N]\!]$, $a_k = 0$. Ceci revient à contraindre la position du premier monomère , $X_1 = 0$. On notera donc K_1 la matrice des potentiels d'interaction gaussiens tenant compte de cette contrainte. En utilisant, l'équation 4.4, on a l'expression de K_1 :

$$\forall (i,j) \in [\![1;\, N-1]\!]^2\,,\; K_{1_{i,j}} = \mathcal{K}_{i+1,j+1} \tag{4.5}$$

On peut remarquer que pour une chaîne purement gaussienne, K_1, de taille $N-1 \times N-1$ est de la forme suivante :

$$K_1 = 3\begin{pmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 \\ & 2 & \ddots & \ddots & 0 & \vdots \\ & & \ddots & \ddots & \ddots & \vdots \\ & & & \ddots & \ddots & 0 \\ & & & & 2 & -1 \\ & & & & & 1 \end{pmatrix} \tag{4.6}$$

4.2.1.3 Relation entre D et K^{-1}

On présente ici une relation entre la matrice des distances quadratiques moyennes, D, et la matrice des potentiels d'interaction gaussien, K :

$$\begin{aligned} D_{ij} &= K^{-1}_{i-1,i-1} + K^{-1}_{j-1,j-1} - 2K^{-1}_{i-1,j-1} \text{ si } i \neq 1 \text{et } j \neq 1 \\ &= \sum_{k=2}^{N} \left(a_k^2 K^{-1}_{k-1,k-1} + 2a_k K_{k-1,j-1}\right) + 2\sum_{2\leqslant k<l} a_k a_l K^{-1}_{k-1,l-1} + K^{-1}_{j-1,j-1} \text{ sinon } (4.7) \end{aligned}$$

Démonstration de la relation 4.7 :
Soient $(i, j) \in [\![1, N]\!]^2$,
Par définition, $D_{ij} = \frac{1}{3}\left\langle (X_i - X_j)^2 \right\rangle = \frac{1}{3}\left(\langle X_i^2\rangle + \langle X_j^2\rangle - 2\langle X_i X_j\rangle\right)$.
Or pour tout $i \neq 1$ et $j \neq 1$, $< X_i X_j > /3 = K^{-1}_{i-1,j-1}$ par définition de la matrice des potentiels d'interaction gaussiens.
Donc si $i \neq 1$ et $j \neq 1$, $D_{ij} = K^{-1}_{i-1,i-1} + K^{-1}_{j-1,j-1} - 2K^{-1}_{i-1,j-1}$
Dans le cas où $i = 1$ (le cas $j = 1$ étant le même cas par symétrie), on a $\forall j \in [\![2, N]\!]$:

$$\begin{aligned} D_{1j} &= \langle X_1^2\rangle + \langle X_j^2\rangle - 2\langle X_1 X_j\rangle \\ &= \left\langle \left(-\sum_{k=2}^{N} a_k X_k\right)^2 \right\rangle + K^{-1}_{j,j} - 2\left\langle \left(-\sum_{k=2}^{N} a_k X_k\right) X_j \right\rangle \\ &= \left\langle \sum_{k=2}^{N} a_k^2 X_k^2 \right\rangle + \left\langle \sum_{2\leqslant k<l}^{N} a_k X_k a_l X_l \right\rangle + K^{-1}_{j,j} + 2\left\langle \sum_{k=2}^{N} a_k X_k X_j \right\rangle \\ &= \sum_{k=2}^{N} \left(a_k^2 K^{-1}_{k-1,k-1} + 2a_k K^{-1}_{k-1,j-1}\right) + 2\sum_{2\leqslant k<l}^{N} a_k a_l K^{-1}_{k-1,l-1} + K^{-1}_{j-1,j-1} \end{aligned}$$

Dans le cas particulier de la matrice K_1 (jauge $X_1 = 0$), la relation 4.7 s'écrit simplement :

$$\begin{cases} \forall\, i \neq j \in [\![2; N]\!]\,,\ D_{ij} = K^{-1}_{1_{i-1,i-1}} + K^{-1}_{1_{j-1,j-1}} - 2K^{-1}_{1_{i-1,j-1}} \\ \forall\, j \in [\![2; N]\!]\,,\ D_{1j} = K^{-1}_{1_{j-1,j-1}} \end{cases} \tag{4.8}$$

4.2.1.4 Relations entre D, C et P

Par la suite, il nous sera souvent utile de transformer des matrices de distance moyenne, D, en matrice de probabilité de contact, P et/ou en matrice de contact C, cette dernière étant une observable accessible expérimentalement avec la technique du Hi-C.
N.B. : On fera attention de ne pas confondre cette matrice de contact, C, avec la matrice de covariance, $\mathcal{C}$, en raison de leur notation ressemblante.

4.2.1.4.1 Relation entre P et D

Dans le cadre d'un modèle gaussien, la relation entre probabilité de contact P et distances moyennes entre monomères, D est donnée par la relation 4.9 :

$$P_{ij}(D_{ij},\ a) = \operatorname{erf}\left(\frac{a}{\sqrt{2D_{ij}}}\right) - \frac{2a}{\sqrt{2\pi D_{ij}}} e^{-a^2/2D_{ij}} \tag{4.9}$$

Démonstration de la relation 4.9 :
On considère que deux monomères i et j sont en contact s'ils sont au maximum à une distance a l'un de l'autre (Fig. 4.2.7). La probabilité de contact entre i et j est alors donnée par :

$$P_{ij}(a) = 4\pi \int_0^a dr.r^2.P(r)$$

Or pour toute chaîne gaussienne à l'équilibre décrit par un hamiltonien quadratique, la probabilité que deux monomères i et j soient « espacés » d'une distance r est donnée par :

$$P(r_{ij}) = \left(\frac{1}{2\pi D_{ij}}\right)^{3/2} e^{\frac{-r_{ij}^2}{2D_{i,j}}}$$

avec r_{ij} la distance entre les monomères i et j et avec D_{ij} le tiers de la distance carré moyenne entre i et j. Par intégration par partie, on trouve l'expression littérale recherchée entre probabilités de contact P_{ij} et distances quadratiques moyennes D_{ij} et du paramètre a (Éq. 4.9).
L'allure de cette fonction est représentée sur la figure 4.2.1. On remarque que si $\sqrt{D_{ij}}/a < 1$ la probabilité de contact est plate ce qui est dû au fait que deux monomères dont la distance moyenne est inférieure à a seront presque systématiquement vus comme étant en contact. Si, au contraire, $\sqrt{D_{ij}}/a > 1$, les monomères sont en moyenne distants de plus d'une longueur a ce qui implique que la probabilité d'observer un contact entre ces deux monomère décroît avec D_{ij}. La décroissance se fait en échelle $\log$ avec une pente de $-3/2$.

4.2.1.4.2 Approximation de la relation entre P et D

Dans la limite où $\frac{\sqrt{D_{ij}}}{a} \gg 1$, l'équation 4.9 peut être approximée par son développement au premier ordre :

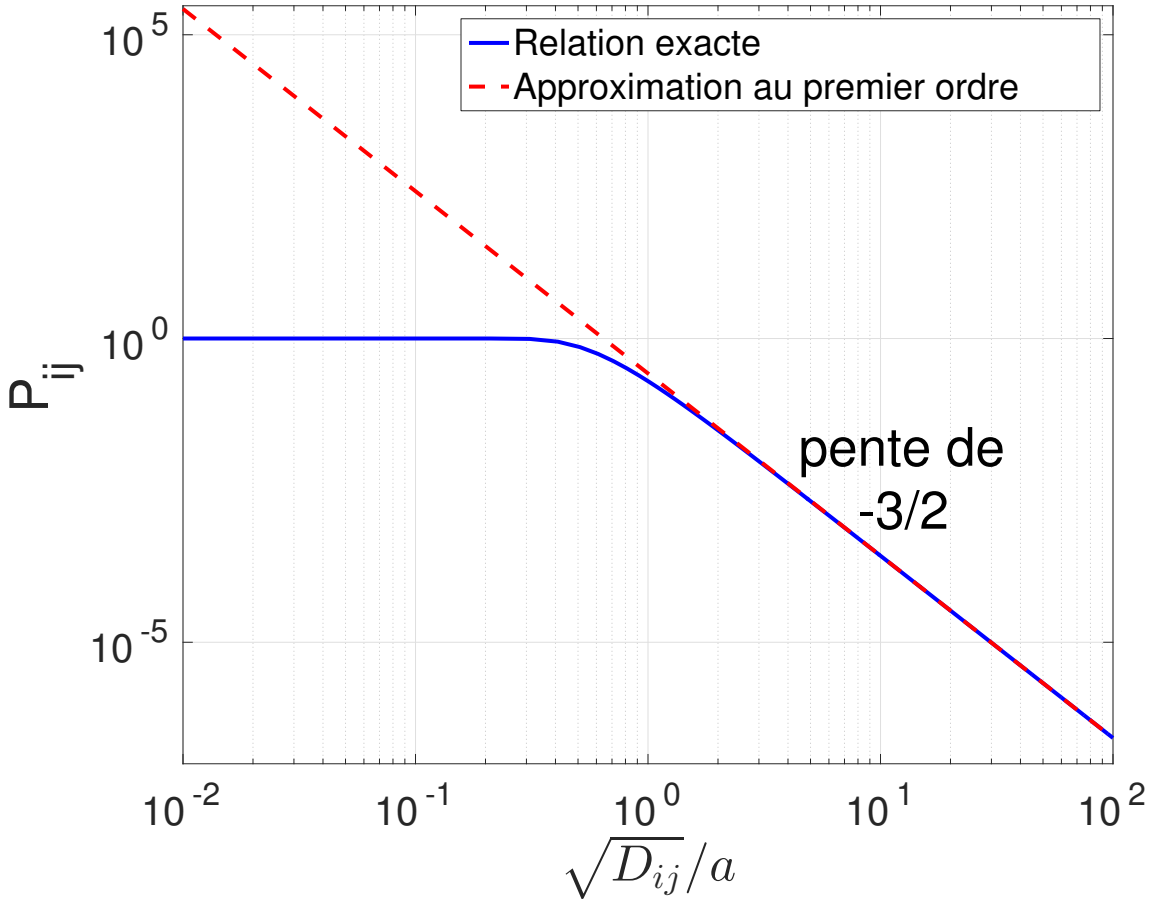

FIGURE 4.2.1 – **Probabilité de contact P_{ij} entre deux monomères i et j pour une chaîne gaussienne à l'équilibre (en bleu) et son approximation au premier ordre (en rouge) en fonction de $\sqrt{D_{ij}}/a$, avec a la portée du contact.** On considère que deux monomères i et j espacés d'une distance moyenne D_{ij} sont en contact si la distance les séparant, r_{ij} est inférieure à a. L'approximation de la relation 4.9 par son développement au premier ordre 4.10 est valable lorsque $\sqrt{D_{ij}}/a > 1$.

$$P_{ij} \approx \frac{a^3}{3}\sqrt{\frac{2}{\pi}}\,(D_{ij})^{-3/2} \tag{4.10}$$

Cette fonction est représentée en rouge sur la figure 4.2.1. On peut voir que relation exacte et approximée se croisent puis se confondent à partir de $\sqrt{D_{ij}}/a = 1$. Afin de vérifier si dans le cas de la chromatine cette approximation est valide, on se propose d'estimer le paramètre a.

Ordre de grandeur de a

La valeur de la distance en deçà de laquelle on considère que deux monomères sont en contact, a, dépend de la portée, a_0, du pontage chimique entre fibres chromatiniennes. En notant $P_{il,jk}$ et $D_{il,jk}$ respectivement la probabilité de contact et la distance carré moyenne entre le $l^{ème}$

fragment de restriction du monomère i et le $k^{ème}$ fragment de restriction du monomère j et en réutilisant les équations 4.12 et 4.10, on a :

$$P_{ij}(a) = \sum_{l=1}^{N}\sum_{k=1}^{N} P_{il,jk}(a_0)$$
$$\iff \frac{a^3}{(D_{i,j})^{3/2}} = \sum_{l=1}^{N}\sum_{k=1}^{N} \frac{a_0^3}{(D_{il,jk})^{3/2}}$$

Si on suppose dans cette dernière relation que les $D_{il,jk}$ sont tous égaux et que les nucléosomes sont des entités indépendantes, alors on peut estimer que :

$$a^3 \approx a_0^3\, N^2 \tag{4.11}$$

On estime que, a_0, la portée du pontage chimique entre fibres chromatiniennes est de l'ordre de 20 nm. En effet, le pontage peut se faire entre deux nucléosomes et on sait qu'un nucléosome peut être assimilé à un cylindre de diamètre 11nm et de hauteur 5.5nm [Alberts et al., 2002] ou il peut se faire sur une échelle plus grande allant probablement jusqu'à 30 nm (taille caractéristique de la fibre d'ADN). De plus, on sait qu'il se trouve environ 12.5 fragments de restriction dans un monomère de 10kb puisque un fragment de restriction est typiquement long de 800pb [Sexton et al., 2012]. De ce fait, $a \approx (20^3 \times 12.5^2)^{1/3} \approx 100$ nm. En conclusion, deux monomères de 10kb sont considérés comme étant en contact lorsque leurs centres respectifs se trouvent à des distances inférieures ou égales à $a \approx 100$ nm.

Ordre de grandeur de D_{ij}

D'après des expériences de FISH, chez la drosophile ou chez la levure, on sait que la distance moyenne entre les centres de masse de deux monomères de 10kb est de $\sqrt{D} \gtrsim 100$nm [Lowenstein et al. 2004 ; Bystricky et al., 2004].

Conclusion

Pour la chromatine, on vient d'estimer que $a \approx 100$ nm et que $\sqrt{D} \gtrsim 100$ nm pour des monomères de 10kb, on en déduit alors $\sqrt{D_{ij}}/a \gtrsim 1$. Ces estimations nous permettent de justifier que pour toute la suite, on se servira de la relation approximée 4.10.

4.2.1.4.3 Relation entre C et P et entre C et D

En supposant qu'un grand nombre d'expériences ait été réalisé, le nombre de contacts C_{ij}et la probabilité de contact P_{ij} entre monomères i et j sont simplement liés par la relation 4.12

$$C_{ij} = N_{exp} P_{ij} \tag{4.12}$$

avec N_{exp} le nombre d'expériences effectives réalisées pour obtenir la carte de contact.
Ainsi, si dans cette dernière expéression, on remplace P par on expression approximée donnée par la relation 4.10, on obtient la relation suivante :

$$C_{ij} \approx \frac{N_{exp}\, a^3}{3}\sqrt{\frac{2}{\pi}}\,(D_{i,j})^{-3/2}$$

Dans le cadre de l'approximation gaussienne, on retiendra donc que la relation entre contacts et distances moyennes entre monomères peut être approximée par la formule 4.13.

$$C \approx AD^{-3/2} \tag{4.13}$$

avec $A = \frac{N_{exp}\, a^3}{3}\sqrt{\frac{2}{\pi}}$ une constante multiplicative dépendant des conditions expérimentales (N_{exp} et a, la portée d'un contact). En raison, des biais expérimentaux liés à la technique de Hi-C (cf. chapitre 1), l'ordre de grandeur de la grandeur expérimentale N_{exp} est difficile à estimer. On présente ci-dessous une estimation de la constante A.

Ordre de grandeur de A

On sait que $A = \frac{N_{exp}\, a^3}{3}\sqrt{\frac{2}{\pi}}$. On a pu ci-dessus estimer $a \approx 100$ nm. Quant à la valeur de N_{exp} il n'y a pas de moyen simple permettant d'en donner un ordre de grandeur. On propose donc ici une estimation de A ne nécessitant pas la connaissance de N_{exp}. Pour cela, on remarque que l'équation **4.13** peut se réécrire $\log(C) = \log(A) - \frac{3}{2}\log(D)$. Le terme $\log(A)$ apparaît ainsi comme étant l'ordonné à l'origine de la droite de pente $-3/2$ déterminant les coefficients $C_{i,j}$ en fonction des $D_{i,j}$. Par conséquent, afin de trouver la valeur de A, on ajuste par une droite de pente $-3/2$ des valeurs expérimentales $\log(C_{i,j})$ obtenues à partir d'expériences de Hi-C aux distances moyennes correspondantes, $\log(D_{ij})$, obtenues par hybridation in situ en fluorescence (FISH). On utilise des données FISH et HiC réalisées chez la drosophile [Sexton et al., 2012 et données FISH non publiées fournies par le laboratoire de Giacomo Cavalli]. De façon assez surprenante, il a été difficile de trouver des distances spatiales entre sites génomiques mesurées par FISH dans la littérature. En effet, même si dans les papiers on peut voir que de telles mesures ont souvent été effectuées, les données n'y figurent pas forcément et ne sont pas non plus téléchargeable [Lowenstein et al., 2004]. Finalement, nous avons récupéré six couples (D_{ij}, C_{ij}), ce faible nombre de données compromet fortement la précision de notre ajustement qui a un coefficient de corrélation médiocre même après suppression des points aberrants : 0.8 (Fig. 4.2.2). Cet ajustement donne $\log(A) \approx 0.6$ soit $A \approx 4$.
De plus, sachant que $A = \frac{N_{exp}\, a^3}{3}\sqrt{\frac{2}{\pi}} \approx 4$, on peut estimer le nombre d'expériences effectives :

$$N_{exp} = \frac{3 \times A}{a^3\sqrt{2/\pi}} \approx 15000 \tag{4.14}$$

avec $\mathfrak{a}$ à exprimer en $\mu\mathfrak{m}$.

Ce résultat justifie notre hypothèse selon laquelle la statistique est bonne (hypothèse effectuée pour écrire l'équation 4.12).

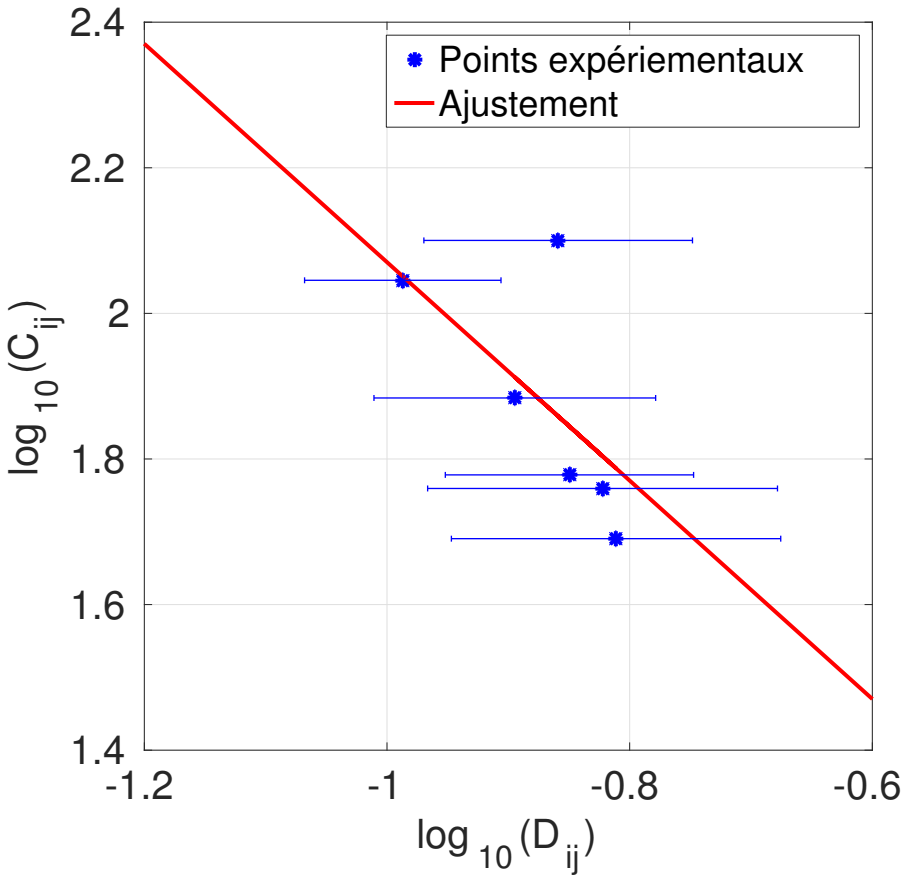

Figure 4.2.2 – **Ajustement linéaire entre les données en échelle logarithmique Hi-C et FISH (en $\mu\mathfrak{m}^2$) de six paires de sites génomiques.** Le coefficient directeur est contraint et vaut -1.5. L'ordonnée à l'origine est laissé libre. L'ajustement donne une ordonnée à l'origine de 0.6 ce qui permet d'en déduire que le facteur A défini dans l'équation **4.13** vaut $A \approx 4$. [Données FISH non publiées fournies par le laboratoire de G. Cavalli].

4.2.1.5 Générer des structures à partir de la matrice des distances D

Sous l'approximation gaussienne, il est possible de générer et de représenter des structures d'un polymère sous réserve de connaître sa matrice des distances quadratique moyennes (Fig. 4.2.3). Ces représentations sont donc des structures possibles compte tenu des contraintes sur les distances moyennes. Les représentations obtenues ne correspondent pas à la conformation

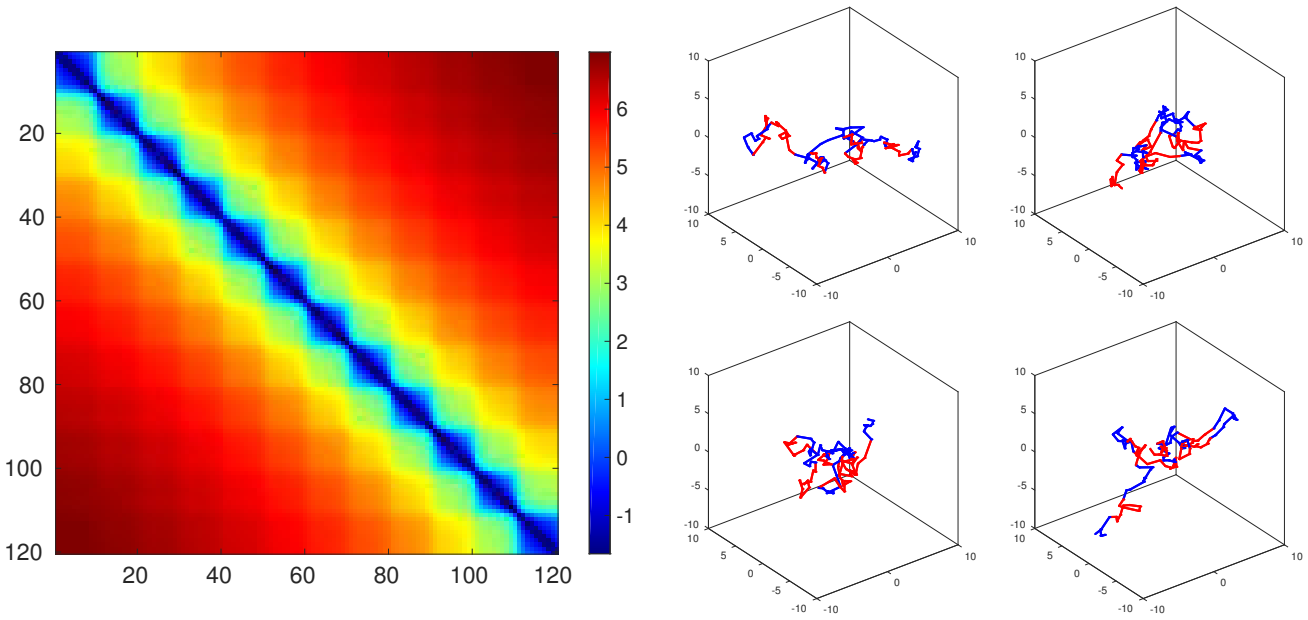

FIGURE 4.2.3 – **Carte des distances moyennes et exemples de structures tridimensionnelles pour un hétéropolymère de 120 monomères.** L'hétéropolymère présenté ici peut se noter, $(A_{10}B_{10})_6$, il est composé de six successions de deux types de bloc, A (en rouge sur la figure) et B (en bleu), chaque bloc étant formé de dix monomères chacun. Ces structures sont tirées aléatoirement dans le cadre de l'approximation gaussienne et avec comme contrainte la matrice des distances moyennes présentée à gauche.

moyenne du polymère mais elles permettent en revanche de se faire une idée concernant la dynamique d'équilibre (sous l'approximation gaussienne).

Concrètement, pour tirer une structure d'un polymère donné, on commence par calculer la matrice de covariance $\mathcal{C}$ à partir de la matrice des distances moyennes, D et de la relation 4.3 qui est une propriété des distrbutions gaussiennes multivariées. A l'aide de la matrice $\mathcal{C}$, il est alors possible de tirer aléatoirement suivant une distribution gaussienne multivariée de moyenne nulle et de covariance, $\mathcal{C}$, un ensemble de vecteurs positions $X = \{X_i\}_{1 \leqslant i \leqslant N}$ avec X_i la position du monomère i et N le nombre total de monomères. Ce tirage se fait à l'aide d'un générateur de nombre aléatoire standard. Dans notre cas, on utilise un générateur développé par *Matlab* (fonction « mvnrnd »).

4.2.2 Approximation gaussienne auto-cohérente

On souhaite dans cette partie caractériser le repliement d'une chaîne isolée composée de N monomères. La dynamique d'une telle chaîne est modélisée par une équation de Langevin (Équation de Newton sans terme d'inertie mais avec une force induite par la viscosité du milieu proportionnelle à la vitesse de déplacement du monomère et une force aléatoire

modélisant l'effet de l'agitation thermique) pour chaque monomère :

$$\zeta \frac{dX_i}{dt} = -\frac{\partial H}{\partial X_i} + \eta_i(t)$$

avec ζ_b une constante positive caractéristique de la force de viscosité s'appliquant sur les monomères et $\eta_i(t)$ un bruit blanc caractéristique des fluctuations stochastiques du système, cette force a une distribution de probabilité gaussienne avec une fonction de corrélation, $\left\langle \eta_i^\alpha(t)\eta_j^\beta(t') \right\rangle = 2\zeta k_B T \delta(t-t')\delta_{ij}\delta_{\alpha\beta}$ avec k_B la constante de Boltzmann, T la température, $\alpha,\ \beta \in \{x,\ y\ z\}$ et δ la fonction de Dirac.
La distribution de probabilité de l'ensemble des positions X_i vérifie alors l'équation de Fokker-Planck suivante :

$$\frac{\partial P\left(\{X_i\}, t\right)}{\partial t} = \frac{1}{\zeta} \sum_i \left[\frac{\partial}{\partial X_i} \left(P \frac{\partial H}{\partial X_i} \right) + k_B T \frac{\partial^2 P}{\partial X_i^2} \right] \tag{4.15}$$

Cependant la structure d'un hamiltonien H décrivant la chaîne est en général trop complexe pour pouvoir intégrer directement ces équations non linéaires. Ainsi, pour étudier ce système, une idée développée dans Jost et al. (2014) et inspirée de Timoshenko et al., 1998 est d'appliquer à chaque pas de temps une approximation « gaussienne auto-cohérente ». Il s'agit d'approximer la distribution de probabilité de l'ensemble des positions P par une distribution gaussienne multivariée (Éq. 4.1). Afin de trouver l'équation d'évolution de la matrice de covariance $\mathcal{C}(t)$ en fonction du temps, on va appliquer une méthode initialement développée dans le cadre de réseau de réactions biochimiques [Ramalho et al., 2013]. L'idée générale est de supposer qu'au départ $P\left(Y = \{X_n\}, t\right)$ est une distribution gaussienne, on détermine alors après un temps infinitésimal, δt, $P\left(Y = \{X_n\}, t + \delta t\right)$ grâce à l'équation de Fokker-Planck 4.15. Le résultat obtenu ne sera toutefois plus forcément gaussien étant donné la possible non-linéarité de l'équation. On cherche alors $P'\left(Y = \{X_n\}, t + \delta t\right)$ une distribution gaussienne la « plus proche possible » de $P\left(Y = \{X_n\}, t + \delta t\right)$, c'est-à-dire telle que la divergence de Kullback-Leibler entre $P\left(Y = \{X_n\}, t + \delta t\right)$ et $P'\left(Y = \{X_n\}, t + \delta t\right)$ soit minimale. La minimisation de cette divergence mène à l'équation 4.16 ci-dessous [Jost et al., 2014].

$$\zeta \frac{d\mathcal{C}}{dt} = \langle J \rangle\, \mathcal{C} + \mathcal{C}\, \langle J \rangle^\dagger + N_r \tag{4.16}$$

avec $N_r = 2k_B T \times I_N$ la matrice de covariance des processus aléatoires η_i, I_N étant la matrice identité d'ordre N et avec J, la matrice opposée de la matrice hessienne de l'hamiltonien H. Les coefficients de cette matrice sont $J_{ij} = -\frac{\partial^2 H}{\partial X_i \partial X_j}$. La valeur moyenne de J sur la distribution gaussienne $P(Y)$ est :

$$\langle J \rangle = -\int dY P(Y) \frac{\partial^2 H}{\partial X_i \partial X_j} \tag{4.17}$$

On a vu que $P(Y)$ ne dépend que de la matrice de covariance $\mathcal{C}$ (Éq. 4.1) et que cette dernière est liée a la matrice D par la relation 4.3. Ainsi $\langle J \rangle$ ne dépend que de la matrice D et des potentiels d'interaction définissant l'hamitonien H. Dans la section suivante où l'on proposera d'expliciter H, nous attacherons une attention particulière à choisir H de sorte à ce qu'il soit intégrable. En effet, ceci est indispensable pour obtenir la matrice $\langle J \rangle$.
Notons qu'il est possible d'aboutir à cette même équation 4.16 sans minimiser la divergence de Kullback-Leibler mais en utilisant l'inégalité de Gibbs–Bogoliubov [Timoshenko et al., 1998].
Étant donné la relation 4.3 entre $\mathcal{C}$ et D, on peut réécrire l'équation de la dynamique 4.16 en une équation dépendant de D uniquement (Éq. 4.18).

$$\zeta \frac{dD_{ij}}{dt} = 4k_B T - \sum_k \left(\langle J_{ik} \rangle - \langle J_{jk} \rangle \right) \left(D_{ik} - D_{jk} \right) \tag{4.18}$$

Nous avons vu ci-dessus que $\langle J \rangle$ ne dépend que de la matrice D et des potentiels d'interaction définissant l'hamiltonien du système. Cette nouvelle équation 4.18 est donc auto-cohérente, elle peut être intégrée numériquement. La résolution de cette équation permet de calculer la dynamique des distances moyennes entre monomères, D, sans passer par des simulations de trajectoires individuelles de monomères. La détermination de la matrice D nous permettra d'accéder à d'autres grandeurs moyennes caractérisant la chaîne (probabilité de contact (Éq. 4.9), rayon de gyration $R_g = \sqrt{\frac{1}{2N^2} \sum_{i,j} D_{ij}}$, ...).

Cas particulier du régime stationnaire

Dans le cas stationnaire, $\frac{d\mathcal{C}}{dt} = 0$ (Éq. 4.16) et $\frac{dD}{dt} = 0$ (Éq. 4.18). L'équation 4.16 se réécrit alors $\langle J \rangle \mathcal{C} + \mathcal{C} \langle J \rangle^\dagger + N_r = 0$ ce qui implique une relation simple entre $\langle J \rangle$ et $\mathcal{C}$ dont on se servira par la suite :

$$\langle J \rangle = -\mathcal{C}^{-1} k_B T = -\mathcal{K} k_B T \tag{4.19}$$

De plus, il peut exister plusieurs solutions pour le système 4.18. L'état d'équilibre du système est une combinaison linéaire des différentes solutions stationnaires. Pour chaque état stationnaire, on peut calculer le poids de Boltzmann correspondant. En effet, on peut écrire l'hamiltonien, H de notre système sous la forme :

$$H = H_G + (H - H_G)$$

avec H_G l'hamiltonien gaussien (Éq. 4.2). L'énergie libre associée au système d'hamiltonien H_G, notée, $\mathcal{F}_G$, est $\mathcal{F}_G = -k_B T \log(Z)$ avec Z la fonction de partition du système. La fonction de partition Z est le terme de normalisation de chaque facteur de boltzmann défini dans l'équation 4.1. Ainsi :

$$\mathcal{F}_G = -k_B T \log\left(\sqrt{(2\pi)^{3N} \times |\det(\mathcal{C}(t))|}\right) \tag{4.20}$$

D'après l'inégalité de Gibbs-Bogoliubov, l'énergie libre du système $\mathcal{F}$ vérifie l'inégalité 4.21 :

$$\mathcal{F} \leqslant \mathcal{F}_G + \langle H - H_G \rangle_0 \tag{4.21}$$

avec la notation $\langle \ldots \rangle_0$ signifiant :

$$\langle H - H_G \rangle_0 = \frac{Tr\left[(H - H_G)\ exp\left(-\beta H_G\right)\right]}{Tr\left[exp\left(-\beta H_G\right)\right]}$$

avec $\beta = (k_B T)^{-1}$. Ainsi, on peut réécrire l'inégalité 4.21 sous la forme :

$$\mathcal{F} \leqslant \mathcal{F}_G + \frac{Tr\left[H\ exp\left(-\beta H_G\right)\right] - Tr\left[H_G\ exp\left(-\beta H_G\right)\right]}{Tr\left[exp\left(-\beta H_G\right)\right]}$$

Connaissant la matrice de covariance $\mathcal{C}$, le premier terme, $\mathcal{F}_G$ est donné par l'équation équation (4.20) . Quant au second terme $\frac{Tr[H_G\ exp(-\beta H_G)]}{Tr[exp(-\beta H_G)]}$, tout comme $\langle J \rangle$, il peut s'exprimer facilement en fonction de la matrice des distances moyennes D et des paramètres de l'hamiltonien H. Le dernier terme est une constante $\frac{Tr[H\ exp(-\beta H_G)]}{Tr[exp(-\beta H_G)]} = \frac{3}{2} N k_B T$. Avec l'inégalité de Gibbs-Bogoliubov, on peut donc estimer une majoration de $\mathcal{F}^*$, ce qui permet d'associer à chaque point fixe de l'équation 4.18, un poids de Boltzmann $Z = exp\left(-\mathcal{F}^*/k_B T\right)$. La moyenne pondérée de toutes les solutions stationnaires permet de reconstruire l'état d'équilibre.
La limite qui s'impose à nous est que nous ne connaissons pas a priori le nombre de point(s) fixe(s) de cette équation 4.18. La résolution de l'équation avec différentes conditions initiales permet de sonder l'éventuelle existence de plusieurs point(s) fixe(s) mais en aucun cas cette méthode permettra d'affirmer que le nombre de point(s) fixe(s) trouvé est le nombre de point(s) fixe(s) total de l'équation. En conclusion, la résolution de l'équation 4.18 va nous permettre d'étudier des états stationnaires mais pas l'état d'équilibre.

4.2.3 Définition de l'hamiltonien du copolymère

On définit dans cette section l'hamiltonien caractérisant notre copolymère par bloc. On prendra en compte la connectivité de la chaîne, le volume exclu, les effets génériques pouvant avoir pour origine l'environnement ou des effets hors équilibre et les interactions entre monomères qu'elles soient spécifiques à la séquence ou non. L'hamiltonien peut donc s'écrire sous la forme :

$$H = H_{connectivité} + H_{ve} + H_{générique} + H_{interactions}$$

Forme de l'hamiltonien	Type de polymère
$H = H_{connectivité}$	Homopolymère gaussien
$H = H_{connectivité} + H_{ve}$	Homopolymère auto-évitant
$H = H_{connectivité} + H_{ve} + H_{générique}$	Homopolymère auto-évitant "crumpled"
$H = H_{connectivité} + H_{ve} + H_{générique} + H_{interactions}$	Hétéropolymère auto-évitant "crumpled"

TABLE 4.1 – **Quelques types de polymères et leur hamiltonien associé.**

On détaillera dans les sous sections ci-dessous les 4 termes définissant H et on proposera différentes formes de potentiels pour H_{ve} et $H_{générique}$ afin de pouvoir étudier par la suite si les propriétés qualitatives du système en dépendent (cf. section 4.2.5). On peut d'ores et déjà remarquer que selon les contributions formant l'hamiltonien, le polymère peut être de plusieurs types comme le résume le tableau 4.1.

4.2.3.1 Connectivité

On suppose que les monomères sont liés les uns aux autres par des interactions de type ressort.

$$H_{connectivité} = \frac{k}{2} \sum_{i=2}^{N} (X_i - X_{i-1})^2 \tag{4.22}$$

avec X_i la position du monomère i, N le nombre de monomères formant le polymère, $k = 3k_B T/l^2$, avec l la taille typique d'un monomère. On choisira de travailler avec $l = 1$, ainsi la constante l fixe l'échelle de longueur.
Connaissant l'expression de l'hamiltonien $H_{connectivité}$ et en utilisant la relation 4.17, il est possible de calculer la valeur moyenne de l'opposé du hessien de $H_{connectivité}$, notée

$\langle J_{connectivité}\rangle$ ou $\langle J_{gaussien}\rangle$. Cette matrice $\langle J\rangle$ est importante car c'est elle qui intervient dans l'équation de la dynamique de la chaîne 4.18. L'expression des coefficients de $\langle J_{gaussien}\rangle$ est donnée ci-dessous :

$$\langle J_{connectivité}\rangle = \langle J_{gaussien}\rangle_{ij} = -k\left(2\delta_{i,j} - \delta_{i-1,j} - \delta_{i+1,j}\right) \tag{4.23}$$

Dans notre cas, $J_{gaussien} = \begin{pmatrix} -3 & 3 & 0 & \cdots & 0 \\ 3 & -6 & 3 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 3 & -6 & 3 \\ 0 & \dots & 0 & 3 & -3 \end{pmatrix}$.

4.2.3.2 Volume exclu

Potentiel de type Lennard-Jones

On peut modéliser l'auto-évitement entre monomères par un potentiel répulsif de type « coeur dur » de type Lennard-Jones tronqué :

$$H_{ve1} = \sum_{i<j} U_{cd\,ij}(r_{ij}) = \sum_{i<j} U^0_{cd}\left[\left(\frac{\sigma}{r_{ij}}\right)^{5/2} - \left(\frac{\sigma}{r_{ij}}\right)^{5/4} + \frac{1}{4}\right] \text{ si } r \leqslant 2^{4/5}\sigma \text{ et 0 sinon} \tag{4.24}$$

avec r_{ij} la distance relative entre les monomères i et j, $U_{cd\,i,j}$ le potentiel répulsif de type « coeur dur » caractérisant l'auto-évitement des monomères i et j. On choisi $U^0_{cd} = 20k_BT$ et $2^{4/5}\sigma = l$. La troncature est imposée au minimum du potentiel afin d'assurer la continuité. Ce potentiel est fortement répulsif pour $x_{ij} = r_{ij}/\sigma < 1$ (Fig. 4.2.4). Le choix des puissances dans la définition de U_{cd} est en partie fait de telle sorte à ce que l'opposé de l'hessien de H que l'on a noté J soit intégrable. Ce dernier est donné par la relation ci-dessous :

$$\begin{aligned} \langle J_{ve_1}\rangle_{ij} &= \frac{5\sigma^{5/4}U^0_{cd}}{2^{1/4}12\sqrt{\pi}}\left(\frac{1}{D_{ij}^{13/8}}\right)\left[\frac{2\sigma^{5/4}}{D_{ij}^{5/8}}\left(\Gamma_{inc}\left[1/4,\ 2^{3/5}\sigma^2/D_{ij}\right] - \Gamma\left[1/4\right]\right)\right. \\ &\quad \left. 2^{5/8}\left(\Gamma\left[7/8\right] - \Gamma_{inc}\left[7/8,\ 2^{3/5}\sigma^2/D_{ij}\right]\right)\right] \text{ si } i \neq j \\ \langle J_{ve_1}\rangle_{ii} &= -\sum_{k\neq i}\langle J_{ki}\rangle \end{aligned} \tag{4.25}$$

avec $\Gamma_{inc}\left[a,z\right] = \int_z^\infty t^{a-1}e^{-t}dt$ la fonction de gamma incomplète.

Potentiel gaussien

On propose ici de modéliser le volume exclu par un deuxième potentiel, plus simple, un potentiel de type gaussien répulsif :

$$H_{ve2} = \sum_{i<j} U_{ve} \times exp\left(\frac{-r_{ij}^2}{2r_e^2}\right) \tag{4.26}$$

avec $U_{ve} = 10 k_B T$ et $r_e = 0.3 \times l$, valeur choisies de telle sorte que le point d'inflexion du potentiel soit autour de $x_{ij} = \frac{r_{ij}}{\sqrt{2} \times r_e} \approx 1$ (Fig. 4.2.4). Dans la partie « Résultats » on cherchera à savoir si les propriétés qualitatives dépendent de la forme des potentiel choisis (Lennard-Jones ou gaussien).
L'expression de $\langle J_{ve_2} \rangle$ est donnée ci-dessous :

$$\langle J_{ve_2} \rangle = -\frac{r_e^3 U_{ve}}{(D_{ij} + r_e^2)^{5/2}} \tag{4.27}$$

Comparaison des deux potentiels

Au cours de cette description du terme H_{ve}, on a proposé deux formes de potentiels (type Lennard-Jones et gaussien) afin de pouvoir étudier par la suite si les propriétés qualitatives d'un copolymère dépendent de la forme de potentiel choisi pour décrire le volume exclu (cf. section 4.2.5). Sur la figure 4.2.4 on peut voir que les potentiels n'ont pas la même allure et surtout, pas du tout les mêmes ordres de grandeur. On peut donc d'ores et déjà s'attendre à ce que les propriétés quantitatives pour un polymère donné ne seront pas les mêmes selon le potentiel utilisé pour modéliser l'auto-évitement.
Avec les termes $H_{connectivité}$ et H_{ve} définis ci-dessus, on peut construire un homopolymère auto-évitant. Pour cela, on résout l'équation équation (4.18) avec les méthodes décrites dans la section à venir section 4.2.4. La probabilité de contact en fonction de s d'un tel polymère est représentée sur la figure 4.2.4. On voit que quel que soit le potentiel utilisé pour modéliser le volume exclu, l'homopolymère auto-évitant obéit à loi du type $P_c \sim s^{-2}$ ce qui est caractéristique d'une marche aléatoire auto-évitante [Gennes, 1979].

4.2.3.3 Effets génériques

Expérimentalement, on observe plutôt que la probabilité de contact évolue selon une loi de type $P_c \sim s^{-1}$ [Lieberman-Aiden et al., 2009]. Par conséquent, nous introduisons dans cette partie un terme noté $H_{générique}$ ayant pour but d'ajouter des interactions au sein du polymère de sorte que la probabilité de contact évolue selon $P_c \sim s^{-1}$. Ce terme $H_{générique}$ modélise de façon *ad hoc* des effets génériques qui ne peuvent pas être modélisés simplement dans l'approximation gaussienne, comme l'environnement (confinement par éventuellement la présence d'autres chaînes) et comme le crumpling dû au caractère hors équilibre de la chaîne. Ces effets ne sont pas du tout triviaux à modéliser dans le cadre de l'approximation

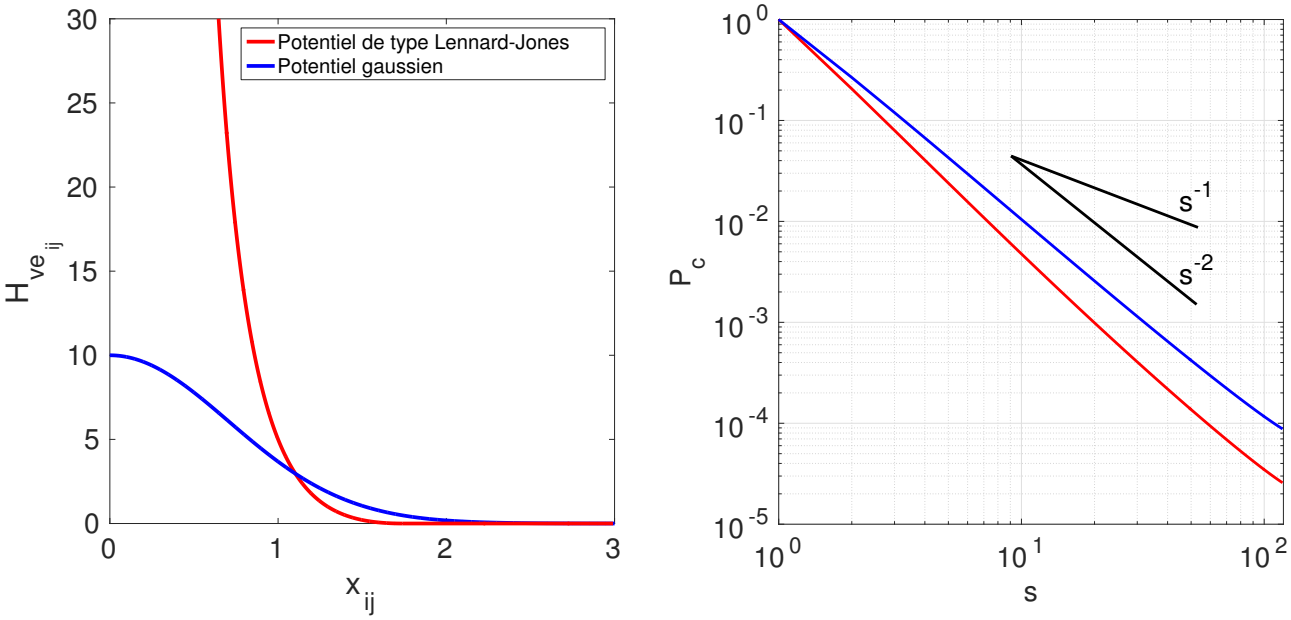

Figure 4.2.4 – **Modélisation du volume exclu à l'aide de deux hamiltoniens différents.** À gauche : $H_{ve1_{ij}}$ en fonction de $x_{1_{ij}} = r_{ij}/\sigma$ en rouge et $H_{ve2_{ij}}$ en fonction de $x_{2_{ij}} = \frac{r_{ij}}{\sqrt{2}r_e}$ en bleu calculés respectivement avec les équations équation (4.24) et équation (4.26). À droite : Probabilité de contact en fonction de la distance génomique. Pour les deux types de potentiel, la loi est proche de $P_c \sim s^{-2}$ et pas de $P_c \sim s^{-1}$ qui est la loi observée expérimentalement pour des longs chromosomes non équilibrés.

gaussienne auto-cohérente mais sont pourtant d'une grande importance chez l'homme ou la drosophile.

Dans le cas de la chromatine, le but de $H_{\text{générique}}$ est de modéliser le comportement générique du copolymère en intégrant les effets de séquence. Ainsi, on impose que la probabilité de contact, P_c, en fonction de la distance génomique évolue selon $P_c \sim s^{-1}$. Cette loi est caractéristique des longs chromosomes non équilibrés, topologiquement contraint [Rosa and Everaers, 2008].

On modélise $H_{\text{générique}}$ par un potentiel effectif entre monomères de type gaussien :

$$H_{\text{générique}} = \sum_{i<j} U_{g_{ij}} \exp\left(\frac{-r_{ij}^2}{2r_0^2}\right) \tag{4.28}$$

avec r_0 la longueur typique des interactions courte portée et avec U_g une matrice donnant l'intensité de l'interaction pour chaque couple de monomère (i, j). Nous pouvons en déduire les coefficients de la matrice $\langle J_{\text{générique}} \rangle$:

$$\langle J_{\text{générique}} \rangle_{ij} = -\frac{r_0^3 U_{g_{ij}}}{(D_{ij} + r_0^2)^{5/2}} \tag{4.29}$$

On détaille ci-dessous deux méthodes permettant de déterminer le champ générique U_g.

Méthode A : Ajustement du champ générique avec un paramètre

Le modèle le plus simple pour déterminer U_g est de considérer qu'il peut être modélisé avec un seul paramètre : $\forall (i, j), U_{g_{ij}} = u_g$. On va chercher à ajuster le paramètre u_g de telle sorte que la matrice U_{ga} mène à une probabilité de contact $P_c \sim s^{-1}$. On sait a priori que pour un polymère, selon l'intensité des interactions entre monomères, on observe une transition de pelote à globule. Pour ces deux phases, la probabilité de contact est respectivement $P_c \sim s^{-3/2}$ et $P_c \sim s^0$, donc on s'attend à ce qu'il soit effectivement possible de trouver un paramètre u_{ga} au niveau de la transition tel que $P_c \sim s^{-1}$ [Barbieri et al., 2012].

Avec les termes $H_{connectivité}$, H_{ve} et $H_{générique}$ définis ci-dessus, on peut construire un homopolymère auto-évitant « crumpled ». Pour cela, on résout l'équation équation (4.18) avec les méthodes décrites dans la section à venir (section 4.2.4). Les probabilités de contact, P_c, en fonction de s pour des homopolymères auto-évitants soumis à différentes intensités d'interactions génériques, u_g, sont représentées sur la figure 4.2.5. On peut voir que - quel que soit le type de potentiel choisi pour modéliser le volume exclu - pour des valeurs faibles de u_g en valeur absolue, la chaîne est relativement ouverte avec une probabilité de contact qui en échelle log décroît avec une pente plus raide que -1. Au contraire, lorsque les interactions u_g sont plus fortes, la chaîne se replie avec une probabilité de contact qui tend à devenir constante. Aux échelles d'interaction u_g intermédiaires, il existe une valeur u_{ga} qui approche au mieux la loi générique $P_c \sim s^{-1}$. Selon la forme du potentiel choisie pour modéliser l'auto-évitement, cette valeur vaut $u_{ga_1} \approx -71.6 k_B T$ (Lennard-Jones) ou $u_{ga_2} \approx -2.36 k_B T$(Gauss) (Fig. 4.2.5) . On remarque que ces deux valeurs ne sont pas du tout du même ordre de grandeur. Dans le premier cas, la valeur obtenue est bien plus grande ce qui est cohérent avec nos observations précédentes selon lesquelles compenser les répulsions engendrées par le potentiel de Lennard-Jones est très coûteux en énergie comparé au cas du potentiel gaussien (Fig. 4.2.4). Les paramètres u_{ga_1} et u_{ga_2} obtenus ici vont nous permettre d'intégrer un champ générique lorsque nous étudierons des copolymères par bloc dans la partie « Résultats » (section 4.2.5). Toutefois, ces paramètres ne reproduisent pas la loi $P_c \sim s^{-1}$ à toutes les échelles. En effet, on voit que pour des valeur de s très petites, c'est-à-dire pour des monomères dans un même voisinage génomique, la pente reste très proche de -1.5 (Fig. 4.2.5) ce qui est dû aux interactions répulsives de type ressort auxquelles sont soumis les monomères plus proche voisins. Par conséquent, afin de reproduire la loi $P_c \sim s^{-1}$ de manière uniforme, nous allons inférer dans le paragraphe suivant les $\frac{N(N+1)}{2}$ paramètres du champ U_g.

Méthode B : Inférence du champ générique avec $\frac{N(N+1)}{2}$ paramètres

La deuxième méthode pour déterminer la matrice U_g consiste à inférer les $\frac{N(N+1)}{2}$ coefficients qu'elle contient de sorte à imposer la loi $P_c \sim s^{-1}$ à toutes les échelles. Pour cela, on commence

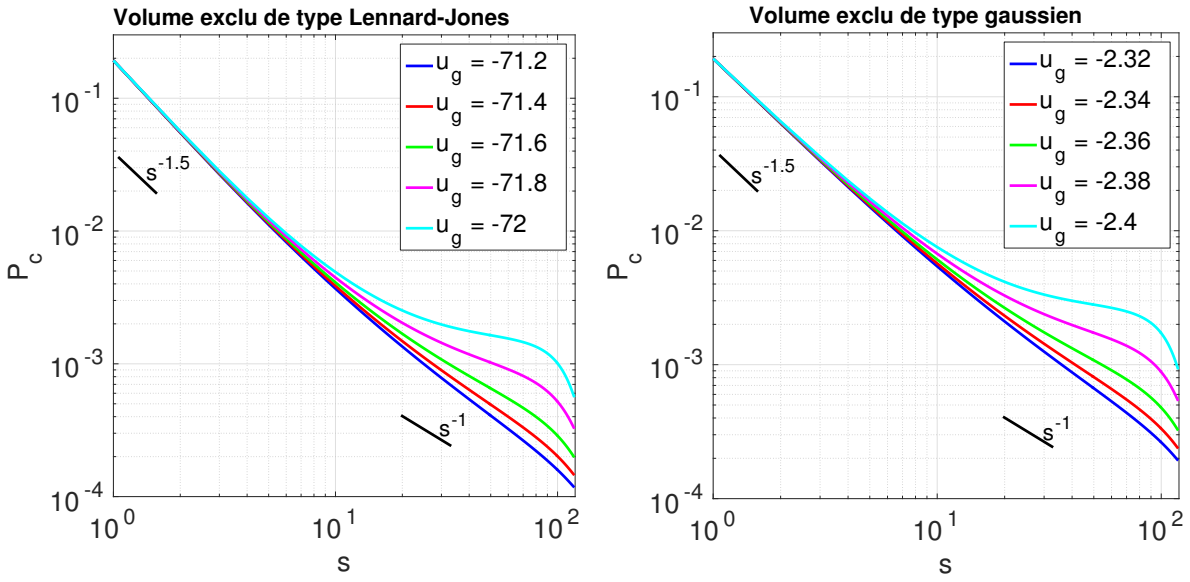

Figure 4.2.5 – **Détermination du champ générique U_g par optimisation d'un paramètre u_g.** Les deux graphes (à gauche si le volume exclu est modélisé par un potentiel de type Lennard-Jones, à droite s'il l'est par un potentiel gaussien) présentent la probabilité de contact, P_c, en fonction de la distance génomique s en bin pour un homopolymère auto-évitant de 120 monomères auquel on ajoute un champ de crumpling uniforme d'intensité en k_BT donnée en légende. Dans les deux cas, la pente de la droite en échelle log dépend de l'intensité du champ générique ajouté. La pente est globalement la plus proche de -1 dans le cas de la courbe verte, correspondant à $u_g = -71.6 k_BT$ à gauche et $u_g = -2.36 k_BT$ à droite.

tout d'abord par calculer la matrice des distances, D, à partir de la matrice des contacts C pour laquelle on aura imposer une décroissance en s^{-1} (Éq. 4.13). Nous pouvons alors en déduire la matrice K (Éq. 4.7) puis la matrice $\langle J \rangle$ (Éq. 4.19). Le champ générique peut alors être calculé en utilisant la relation 4.30 ci-dessous liant $\langle J \rangle$ et D.

$$\begin{aligned} \langle J \rangle &= \langle J_{gaussien} \rangle + \langle J_{ve} \rangle + \langle J_{générique} \rangle \\ \langle J \rangle_{ij} &= -k\left(2\delta_{i,j} - \delta_{i-1,j} - \delta_{i+1,j}\right) + \langle J_{ve} \rangle - \frac{r_0^3 U_{g_{ij}}}{\left(D_{ij} + r_0^2\right)^{5/2}} \end{aligned} \tag{4.30}$$

En effet, dans cette dernière expression, il suffit d'isoler U_g, pour trouver la matrice U_{gb} recherchée :

$$U_{gb} = \left(\langle J_{gaussien} \rangle + \langle J_{ve} \rangle - \langle J \rangle\right) \left(\frac{\left(r_0^2 + D\right)^{5/2}}{r_0^3} \right) \tag{4.31}$$

En conclusion, l'inférence du champ générique à partir d'une matrice de probabilité de contact, P ne pose pas de contraintes mathématiques ou numériques, à part éventuellement

l'inversion de la matrice K^{-1} dans le cas où celle-ci est bruitée. En effet, inverser une matrice bruitée est toujours délicat car c'est un processus qui amplifie les erreurs. Les matrices U_{gb_1} et U_{gb_2} obtenues à l'issue de l'inférence en modélisant respectivement un volume exclu du type Lennard-Jones ou de type gaussien sont présentées figure 4.2.6. On observe dans les deux cas qu'à part pour des valeurs de s très petites ou très grandes pour lesquelles des effets de bords apparaissent, le champ générique est quasiment uniforme. Les plateaux ont pour ordonnées environ $-71.6 k_B T$ et $-2.36 k_B T$. Ce sont des valeurs très proches de celles que l'on avait trouvées dans le paragraphe précédent où l'on cherchait à ajuster le champ générique à l'aide d'un unique paramètre. Les deux approches donnent donc presque le même résultat. Les différences se situent à très petites et très grandes échelles. En effet, on peut observer sur la figure 4.2.6 qu'entre plus proches voisins le champ générique est très négatif, donc très attractif. Cette forte correction est due au fait que les monomères adjacents sont repoussés les uns des autres à cause des interactions répulsives type ressort introduites avec le terme $H_{connectivité}$. Ainsi, le terme très négatif du champ générique entre plus proche voisins permet d'une certaine manière d'augmenter la raideur des ressorts et donc de rapprocher les monomères. Concernant les grandes échelles, on observe un effet de bord entre les deux bouts du polymère. Afin d'étudier cet effet de bord, on infère le champ générique pour différentes tailles de chaînes (Fig. 4.2.6 en bas). En remarquant que pour toutes les tailles de chaînes, l'effet de bord se manifeste uniquement sur les derniers monomères, on décide de s'affranchir de cet artefact en intégrant la chaîne étudiée de taille N dans un champ générique calculé à partir d'un polymère de taille $1.25 \times N$. Par exemple, si on souhaite étudier une chaîne de 120 monomères, le champ générique sera celui représenté dans l'encadré en tirets noirs sur la figure 4.2.6. On peut y voir que les premiers et derniers monomères ne subissent pas d'interactions attractives de façon disproportionnée.
On peut remarquer que le fait que le champ soit globalement composé de valeurs négatives est caractéristique d'interactions attractives. Il est normal que les interactions soient attractives car le champ moyen décroît moins vite qu'un champ gaussien auto-évitant.

4.2.3.4 Interactions entre monomère

Comme expliqué en introduction de ce chapitre, on souhaite modéliser la chromatine par un copolymère par bloc dont les blocs ont la particularité d'interagir préférentiellement s'ils sont de même état épigénomique. Dans cette partie, nous allons donc introduire le terme $H_{interaction}$ qui fixe les interactions spécifiques entre monomères. Toutefois, il est important de remarquer qu'en ajoutant des interactions spécifiques, on va perturber la loi de puissance $P_c \sim s^{-1}$ imposée précédemment avec le terme d'interactions génériques, $H_{générique}$. On peut se rendre compte de cette perturbation en observant le comportement d'un polymère

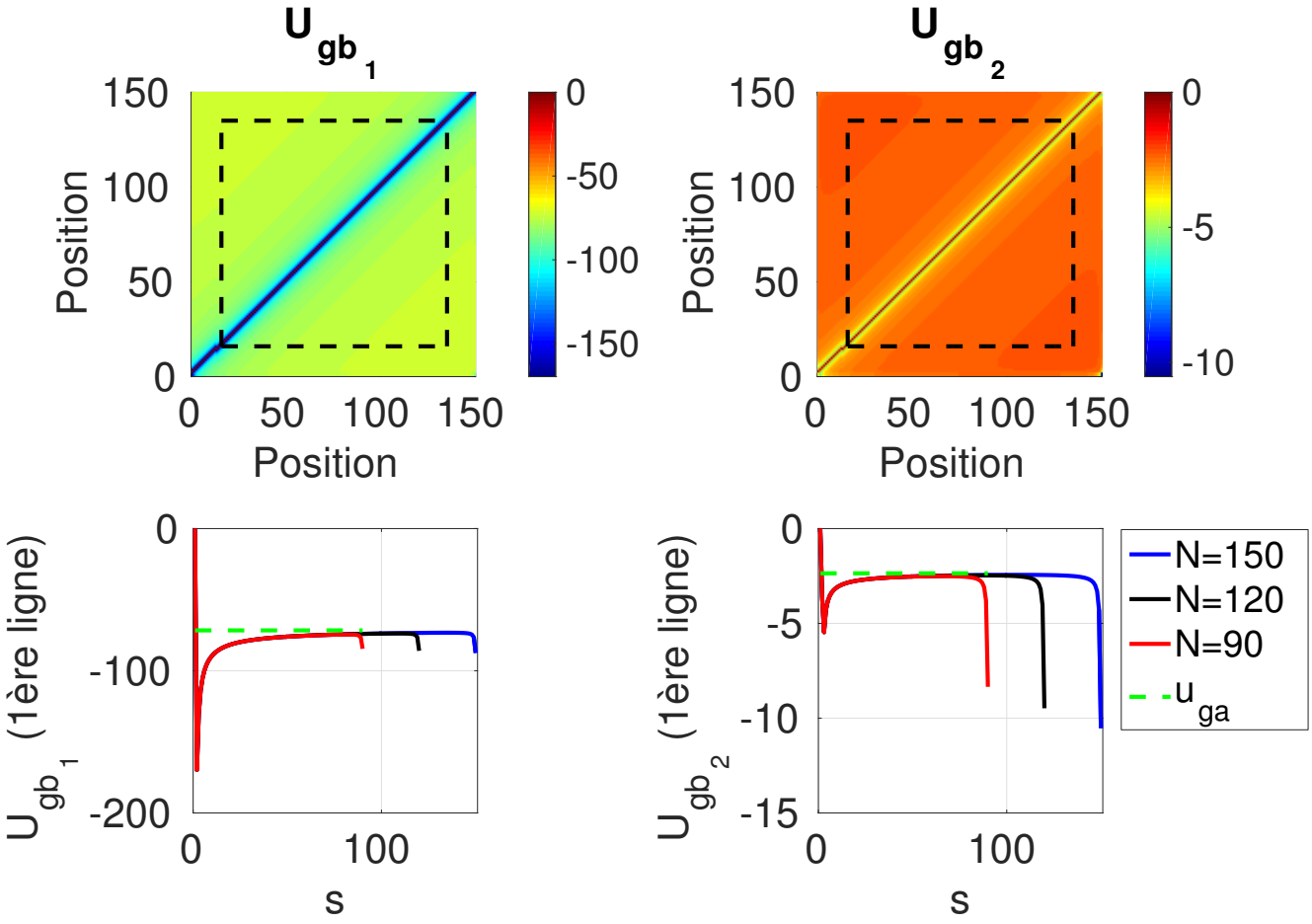

Figure 4.2.6 – **Champ générique, U_{gb}, obtenu pour un homopolymère auto-évitant à l'équilibre.** Les $\frac{N(N+1)}{2}$ paramètres de ce champ sont calculés de manière à imposer que la probabilité de contact évolue selon $P(s) = s^{-1}$. En haut : Matrices U_{gb} représentant le champ générique pour un homopolymère auto-évitant de taille $N = 150$ tel que le volume exclu est modélisé avec un potentiel de type Lennard-Jones (U_{gb1} à gauche) ou un potentiel gaussien (U_{gb2} à droite). Les encadrés en tirets noirs délimitent dans chaque cas le champ générique que l'on utilisera lors de l'étude d'un copolymère de taille $N = 120$ monomères. Ceci permet de s'affranchir des effets de bords. En bas : Évolution des coefficients de la première ligne de la matrice U_{gb} en unités k_BT pour différentes longueurs de chaînes, $N = 90, 120, 150$ respectivement représentées en rouge, noir, bleu. Sur chaque graphe, la ligne verte en tirets situe la valeur du paramètre u_{ga} trouvé avec la méthode A qui consistait à ajuster le champ générique avec un unique paramètre ($u_{ga1} = -71.6k_BT$ et $u_{ga2} = -2.36k_BT$).

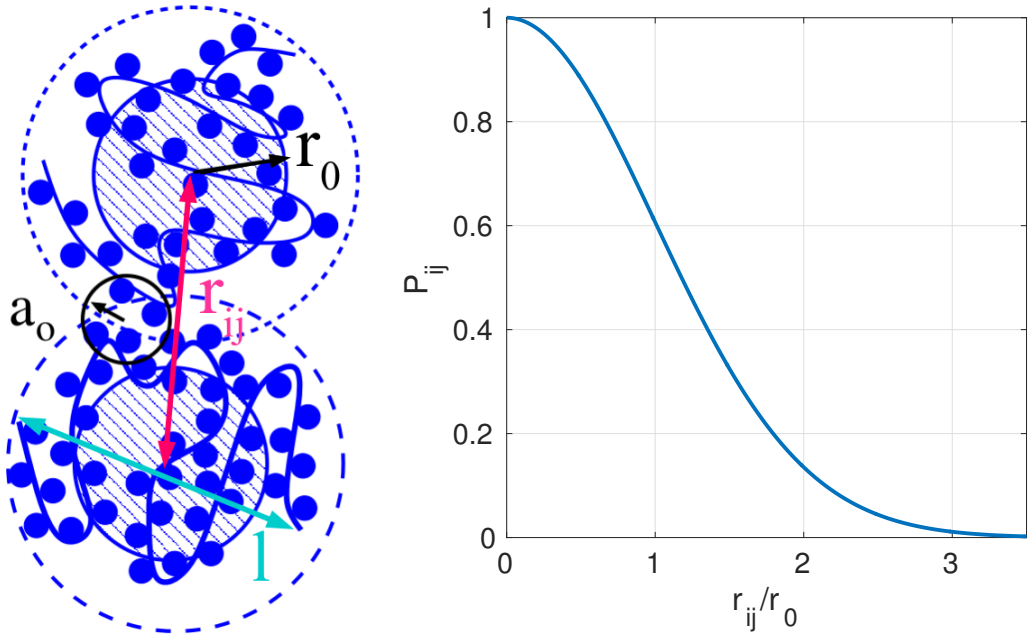

FIGURE 4.2.7 – **Modélisation du nombre de contacts entre deux monomères.** On modélise chaque monomère de 10kb par une pelote gaussienne de rayon de gyration $r_0 = \frac{l}{\sqrt{6}}$ avec l la longueur bout-à-bout de la chaine. Sur l'image de gauche, on peut voir deux tels monomères i et j hachurés en bleu et espacés d'une longueur r_{ij} (entre leurs centres de masse respectifs). La constante a_0 est la portée du pontage chimique entre deux nucléosomes représentés par des ronds bleus ici. Le nombre de contacts entre deux monomères i et j ainsi modélisés est alors proportionnel à la probabilité de contact, $P_{ij} = exp\left(\frac{-r_{ij}^2}{2r_0^2}\right)$, représentée sur l'image de droite.

auto-évitant crumpled auquel on ajoute des interactions spécifiques. Pour construire un tel polymère, on résout l'équation équation (4.18) avec l'hamiltonien $H = H_{connectivité} + H_{ve2} + H_{générique} + H_{interactions}$ avec $H_{interactions} = \sum_{i<j} u_s \delta_{e_i e_j} exp\left(\frac{-r_{ij}^2}{2r_0^2}\right)$ avec u_s l'intensité des interactions spécifiques et avec $\delta_{e_i e_j} = 1$ si les monomères i et j sont dans le même état. On choisit de modéliser les interactions avec un potentiel attractif de type gaussien. En effet, si on modélise chaque monomère de 10kb par une pelote gaussienne de rayon de gyration $r_0 = \frac{l}{\sqrt{6}}$ alors le nombre de contacts les deux chaînes gaussiennes est proportionnel à $exp\left(\frac{-r_{ij}^2}{2r_0^2}\right)$ avec r_{ij} la distance entre les deux centres de masse (Fig. 4.2.7) .
On a réalisé cette résolution pour un copolymère composé de six alternances entre blocs A et B, chacun d'entre eux étant formé de 10 monomères, copolymère noté $(A_{10}B_{10})_6$. On présente sur la figure 4.2.8 les résultats obtenus en fixant $u_s = -0.2$, $u_s = -0.4$ et $u_s = -0.6$. On peut voir que la probabilité de contact moyenne en fonction de la distance génomique dans le cas où $u_s = 0$ (homopolymère) évolue bien selon $P_c \sim s^{-1}$ alors que dans les trois autres cas ($u_s < 0$) le pente est moins raide. La raison de cette différence est visible sur les cartes

de contact moyennes et sur les exemples de structures 3D sur lesquelles on voit que l'ajout d'interaction spécifique à un homopolymère auto-évitant crumpled, implique la formation de TADs et d'interaction à longue portée, ceci implique inéluctablement une augmentation de la probabilité de contact moyenne à toutes les échelles. Afin de ré-imposer la loi $P_c \sim s^{-1}$, nous allons ajouter des interactions non spécifiques répulsives entre tout les monomères. On a vu lors de la description du terme $H_{générique}$, qu'à part dans le cas des monomères plus proches voisins, le potentiel $U_{générique}$ peut être considéré comme constant (Fig. 4.2.5). Il est donc judicieux de conserver comme base le terme $H_{générique}$ qui contient une information dépendante de la distance génomique entre monomères, s, et d'ajouter dans l'hamiltonien du système un paramètre d'interaction non spécifique constant, u_{ns}, qui va permettre de ré-imposer la loi $P_c \sim s^{-1}$. D'une certaine manière, le terme u_{ns} va corriger (par une translation) le champ générique qui a été calculé sans prendre en compte les interactions spécifiques.
En résumé, la contribution $H_{interactions}$ contient un terme qui décrit les interactions directes entre monomères (éventuellement dépendant de l'état épigénomique) et un terme non spécifique permettant la correction du champ générique. Ces interactions, modélisées par un potentiel attractif de type gaussien, constituent des déviations par rapport au modèle générique. Par conséquent, le terme $H_{interactions}$ s'écrit :

$$H_{interactions} = \sum_{i<j} \left(u_{ns} + u_s \delta_{e_i e_j}\right) exp\left(\frac{-r_{ij}^2}{2r_0^2}\right) \tag{4.32}$$

avec $\delta_{e_i e_j} = 1$ si les états épigénétiques e_i et e_j des monomères i et j sont identiques et 0 sinon, r_0 la longueur typique des interactions courte portée. Les paramètres u_{ns} et u_s sont des constantes, les matrices correspondantes sont notées avec une majuscule, respectivement, U_{ns} et U_s avec $\forall(i,j)$, $U_{ns_{ij}} = u_{ns}$ et $U_{s_{ij}} = u_s \delta_{e_i e_j}$.
L'expression de $\langle J_{interactions}\rangle$ associée à cet hamiltonien est :

$$\langle J_{interactions}\rangle_{ij} = -\frac{r_0^3\left(u_{ns} + u_s \delta_{e_i e_j}\right)}{\left(D_{ij} + r_0^2\right)^{5/2}} \tag{4.33}$$

4.2.3.5 Conclusion

On va étudier le repliement de copolymères par bloc modélisés par quatre hamiltoniens différents. Ces quatre hamiltoniens ont pour point commun la définition des termes $H_{connectivité}$ et $H_{interactions}$. Par contre, entre H_1 et H_2 le terme H_{ve} différera : le volume exclu sera respectivement modélisé par un potentiel de Lennard-Jones ou par un potentiel gaussien. Enfin, entre les hamiltoniensH_{1A} et H_{1B} ou H_{2A} et H_{2B}, la contribution $H_{générique}$ ne sera pas la même. Pour H_{1A} et H_{2A}, le potentiel U_{ga} est obtenu par l'ajustement d'un paramètre alors que pour H_{1B} et H_{2B}, le potentiel U_{gb} résulte de l'inférence de $\frac{N(N+1)}{2}$ paramètres. Les

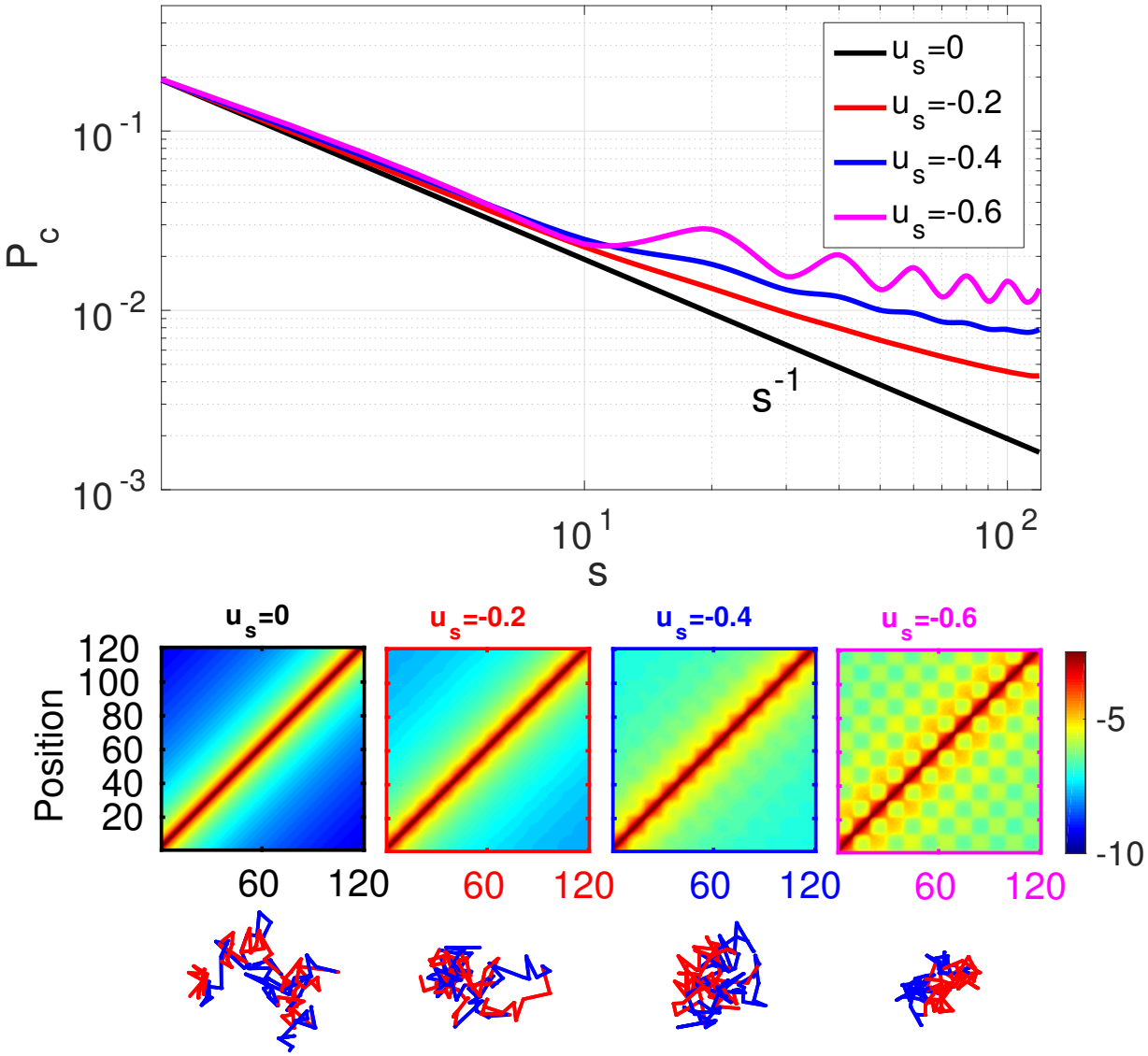

FIGURE 4.2.8 – **Effet de l'ajout d'interactions spécifiques entre monomères du même type à un copolymère auto-évitant « crumpled », $(A_{10}B_{10})_6$, formé de six successions de blocs A et de blocs B, chacun composé de dix monomères.** Son hamiltonien est $H = H_{connectivité} + H_{ve2} + H_{générique} + H_{interactions}$. En haut : Probabilité de contact moyenne, P_c, en fonction de la distance génomique, s, en bin, pour quatre paramètres d'intensité spécifique u_s différents. Lorsqu'il n'y a pas d'interaction spécifique entre monomères, la probabilité de contact évolue selon la loi $P_c \sim s^{-1}$. En bas : Cartes de contact obtenues avec les différentes valeurs de u_s accompagnées en dessous d'un exemple de structure avec en rouge les monomères de type A et en bleu les B. Ces figures montrent que l'ajout d'interactions spécifiques engendrent des TADs et des interactions à longue portée ce qui augmente la probabilité de contact moyenne à toutes les échelles.

équations 4.34, 4.35, 4.36 et 4.37 résument respectivement les expressions de H_{1A}, H_{1B}, H_{2A} et H_{2B}.

$$H_{1A} = \frac{k}{2}\sum_{i=2}^{N}(X_i - X_{i-1})^2 + \sum_{i<j} U^0_{cd}\left[\left(\frac{\sigma}{r_{ij}}\right)^{5/2} - \left(\frac{\sigma}{r_{ij}}\right)^{5/4} + \frac{1}{4}\right] + \sum_{i<j}\left(u_{ga} + u_{ns} + u_s\delta_{e_ie_j}\right)\exp\left(\frac{-r_{ij}^2}{2r_0^2}\right) \tag{4.34}$$

$$H_{1B} = \frac{k}{2}\sum_{i=2}^{N}(X_i - X_{i-1})^2 + \sum_{i<j} U^0_{cd}\left[\left(\frac{\sigma}{r_{ij}}\right)^{5/2} - \left(\frac{\sigma}{r_{ij}}\right)^{5/4} + \frac{1}{4}\right] + \sum_{i<j}\left(u_{gb_{ij}} + u_{ns} + u_s\delta_{e_ie_j}\right)\exp\left(\frac{-r_{ij}^2}{2r_0^2}\right) \tag{4.35}$$

$$H_{2A} = \frac{k}{2}\sum_{i=2}^{N}(X_i - X_{i-1})^2 + \sum_{i<j} U_{ve}\times\exp\left(\frac{-r_{ij}^2}{2r_e^2}\right) + \sum_{i<j}\left(u_{ga} + u_{ns} + u_s\delta_{e_ie_j}\right)\times\exp\left(\frac{-r_{ij}^2}{2r_0^2}\right) \tag{4.36}$$

$$H_{2B} = \frac{k}{2}\sum_{i=2}^{N}(X_i - X_{i-1})^2 + \sum_{i<j} U_{ve}\times\exp\left(\frac{-r_{ij}^2}{2r_e^2}\right) + \sum_{i<j}\left(u_{gb_{ij}} + u_{ns} + u_s\delta_{e_ie_j}\right)\times\exp\left(\frac{-r_{ij}^2}{2r_0^2}\right) \tag{4.37}$$

Par la suite, la notation H_{1X} représentera H_{1A} ou H_{1B}, la notation H_{XA} représentera H_{1A} ou H_{2A} (les notations H_{2X} et H_{XB} seront utilisées sur le même principe).
Les opposés des hessiens des hamiltoniens H_{1X} et H_{2X}, notés $\langle J_1\rangle$ et $\langle J_2\rangle$ interviennent dans l'équation de la dynamique de la chaîne 4.18. Ils différent l'un de l'autre au niveau du terme caractérisant le volume exclu. Leur expression est résumée ci-dessous :

$$\begin{aligned}
\langle J_1\rangle_{ij} &= -k\left(2\delta_{i,j} - \delta_{i-1,j} - \delta_{i+1,j}\right) \\
&\quad + \frac{5\sigma^{5/4}U^0_{cd}}{2^{1/4}12\sqrt{\pi}}\left(\frac{1}{D_{ij}^{13/8}}\right)\left[\frac{2\sigma^{5/4}}{D_{ij}^{5/8}}\left(\Gamma_{inc}\left[1/4,\ 2^{3/5}\sigma^2/D_{ij}\right] - \Gamma\left[1/4\right]\right)\right. \\
&\quad \left. + 2^{5/8}\left(\Gamma\left[7/8\right] - \Gamma_{inc}\left[7/8,\ 2^{3/5}\sigma^2/D_{ij}\right]\right)\right] + \frac{r_0^3\left(U_{ns} + U_s\delta_{e_ie_j} + U_g\right)}{\left(D_{ij} + r_0^2\right)^{5/2}} \quad \text{si } i \neq j \qquad (4.38) \\
\langle J_1\rangle_{ii} &= -\sum_{k\neq i}\langle J_{ki}\rangle \qquad (4.39)
\end{aligned}$$

avec $\Gamma_{inc}\left[a, z\right] = \int_z^\infty t^{a-1}e^{-t}dt$ la fonction de gamma incomplète.

$$\langle J_2\rangle_{ij} = -k\left(2\delta_{i,j} - \delta_{i-1,j} - \delta_{i+1,j}\right) - \frac{r_e^3 U_{ve}}{\left(D_{ij} + r_e^2\right)^{5/2}} - \frac{r_0^3\left(U_{ns} + U_s\delta_{e_ie_j} + U_g\right)}{\left(D_{ij} + r_0^2\right)^{5/2}} \tag{4.40}$$

4.2.4 Résolution du système d'équations différentielles non linéaires d'inconnu D

Pour un hamiltonien fixé, l'équation 4.18 d'inconnue D, la matrice des distances quadratiques moyennes entre monomères, nous donne un moyen d'accéder aux propriétés structurelles du copolymère. Comme on s'intéresse à l'état d'équilibre, on va chercher à résoudre cette équation dans le cas stationnaire, soit avec $\zeta \frac{dD}{dt} = 0$. Dans cette section, on développe deux méthodes permettant la détermination des solutions stationnaires : la première est celle de Runge-Kutta qui résout l'équation différentielle ordinaire, la seconde est une méthode itérative.

4.2.4.1 Algorithme de Runge-Kutta

La résolution de l'équation différentielle ordinaire 4.18 peut être réalisée avec la méthode standard de Runge-Kutta. Avec cette méthode, il s'agit de résoudre l'évolution temporelle des variables dynamiques d'un système d'équations différentielles ordinaires par discrétisation de l'équation dynamique correspondante [Dormand and Prince, 1980]. La méthode (4,5) assure un bon rapport précision/temps de calcul car le pas de temps est adaptatif. Nous l'avons mise en place en utilisant la routine préintégrée de *Matlab* [Shampine and Reichelt, 1997]. Pour une condition initiale donnée, l'algorithme résout l'évolution de $D(t)$ jusqu'à la convergence vers un point fixe. On estime l'état stationnaire atteint lorsque l'écart ϵ entre les 15 dernières matrices trouvé est plus petit que 10^{-3}. On définit l'écart ϵ_{AB} entre deux matrices A et B par la formule 4.41 :

$$\epsilon_{AB} = \sqrt{max_{i,j} \left(A_{ij} - B_{ij}\right)^2} \tag{4.41}$$

Le critère de convergence à une étape $n > 15$ peut donc se formuler ainsi :

$$\forall \left(k,\, l\right) \in [\![n-15;\, n]\!]^2 ,\; max \left(\epsilon_{D_k D_l}\right) < 10^{-3}$$

La complexité de la méthode de Runge-Kutta est pour chaque pas de temps en $\mathcal{O}\left(N^3\right)$. Étant donné que le temps de convergence vers l'état stationnaire peut être relativement long dans certaines conditions et en particulier pour les grandes échelles (Fig. 4.2.9), l'obtention de l'état stationnaire peut être assez coûteuse numériquement. Nous avons donc développé une autre méthode itérative décrite ci-après.

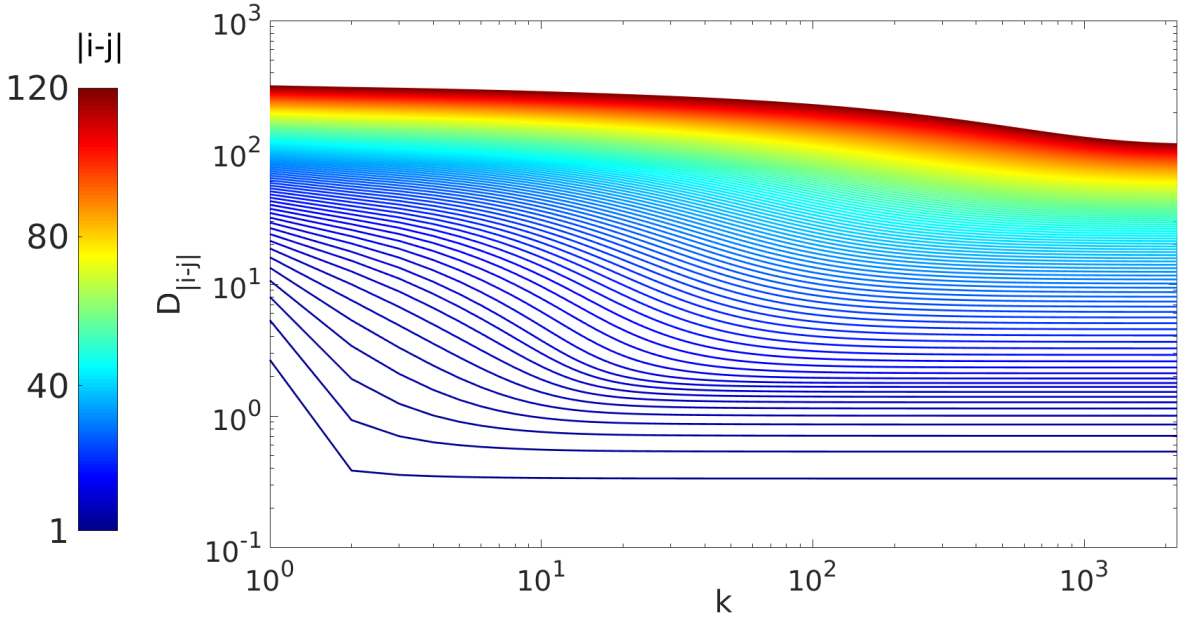

Figure 4.2.9 – **Équilibration pour différentes échelles observée grâce à l'algorithme de Runge-Kutta pour une chaîne de 120 monomères modélisée avec l'hamiltonien H_{2A}.** On voit sur cette figure l'évolution des coefficients $D_{|i-j|}$ obtenus avec le processus de Runge-Kutta en fonction de l'itération k (proportionnelle au temps réel) et en fonction de la distance génomique définie par l'échelle de couleur à gauche. On observe que les D_{ij} convergent d'autant plus vite que la distance genomique entre les monomeres est petite (couleur bleue). On retrouve là le fait que les grandes échelles s'équilibrent beaucoup plus lentement que les petites.

4.2.4.2 Résolution itérative

Dans cette méthode, on se limite à la résolution de D dans l'état stationnaire. Dans ce cas, on sait que D et $\langle J \rangle$ sont liés par la relation 4.42 :

$$\begin{aligned} D_{1j} &= \langle J \rangle_{jj} \\ D_{ij} &= \langle J \rangle_{ii} + \langle J \rangle_{jj} - 2 \times \langle J \rangle_{ij} \text{ si } i \neq j \end{aligned} \tag{4.42}$$

En effet, on connaissait une relation entre D et K_1 (Éq. 4.8), puis entre K_1 et $\mathcal{K}$ (Éq. 4.5), ce qui permet de conclure puisque à l'équilibre $\langle J \rangle = -\mathcal{K}$ (Éq. 4.19).
On établie un schéma itératif basé sur cette relation 4.42 :
Connaissant la matrice D à l'étape n, on peut calculer le $\langle J \rangle$ correspondant avec la relation 4.38 ou 4.40 (selon le choix de modélisation du volume exclu) qui explicite $\langle J \rangle$ en fonction de l'hamiltonien. Puis avec la relation d'équilibre 4.42, on peut calculer D_{n+1} à l'étape $n+1$. On définit ainsi une relation de récurrence qui a pour point fixe $D_{stationnaire}$. Le schéma itératif est résumé ci dessous (Éqs 4.43) :

$$
\begin{aligned}
&\langle J\rangle_0 = f\left(D^{initial}\right)\\
&D_1 = g\left(\langle J\rangle_0\right)\\
&\langle J\rangle_1 = f\left(D_1\right)\\
&D_2 = g\left(\langle J\rangle_1\right)\\
&\vdots\\
&\langle J\rangle_{n-1} = f\left(D_{n-1}\right)\\
&D_n = g\left(\langle J\rangle_{n-1}\right)\\
&\vdots\\
&D_{stationnaire}
\end{aligned}
\tag{4.43}
$$

avec $\langle J\rangle_n = f\left(D_n\right)$ faisant référence aux relations de définition de $\langle J\rangle$ (Éqs. 4.38 ou 4.40 selon le choix de modélisation du volume exclu) et avec $D_n = g\left(\langle J\rangle_{n-1}\right)$ faisant référence à la formule 4.42 valable uniquement à l'équilibre.

Numériquement, on réitère les opérations décrites ci-dessus (Éqs. 4.43) jusqu'à ce que l'on converge vers un point fixe. On estime que c'est le cas si les 50 dernières matrices D successivement trouvées sont presque identiques. Arbitrairement, on dit qu'elles sont presque identiques lorsque l'écart entre chacunes d'entre elles, ϵ, est inférieur à 10^{-6} (Éq. 4.41). Le critère de convergence à une étape $n > 50$ de l'algorithme itératif peut donc s'écrire :

$$
\forall\left(k,\, l\right) \in \left[\!\left[n-50;\, n\right]\!\right]^2,\ max\left(\epsilon_{D_k D_l}\right) < 10^{-6} \tag{4.44}
$$

Ce choix de seuil de 10^{-6} fait suite à deux observations : (i) nous avons remarqué sur une multitude d'exemples que lorsque ϵ devient inférieur à 10^{-4} la convergence était systématique et (ii) choisir un seuil de 10^{-6} plutôt que 10^{-4} n'allonge pas significativement le temps de calcul.

Notre problème étant non linéaire, nous avons remarqué que dans un bon nombre d'exemples le processus de résolution laisse apparaître, à partir d'un certain nombre d'itérations des cycles limites stables impliquant des oscillations maintenues d'un ensemble de matrices D. Ce cas est représenté par la courbe bleue de la figure 4.2.10 qui présente en fonction de l'itération k, l'écart entre la solution stationnaire trouvée avec l'algorithme de Runge-Kutta et la matrice D_k de l'itération k. Un autre problème rencontré est l'apparition éventuellement dans les premières itérations de coefficients négatifs dans la matrice D (courbes cyan, noire et rouge de la figure 4.2.10). Ceci se produit si la matrice D ou la matrice $\langle J\rangle$ de l'itération en cours est trop éloignée de l'équilibre, au point que la solution de l'équation $D = g\left(\langle J\rangle\right)$ (Éq. 4.42) ne correspond pas à une matrice de distance car la matrice de covariance associée $\mathcal{C}$ n'est pas définie positive (ce qui se traduit par des coefficients négatifs au sein de D). On comprend

donc que si on initialise le processus itératif avec une matrice $D^{initiale}$ trop loin de l'équilibre, il y a de forts risques que l'algorithme ne converge pas.

Afin d'éviter les problèmes sus-cités et afin donc d'assurer au mieux la convergence de notre processus itératif, on introduit un paramètre λ dans le processus itératif décrit par le schéma 4.43 : la matrice D_k calculée lors de la $k^{ème}$ itération ne sera plus utilisée telle quelle pour l'itération suivante, $k+1$, mais sera pondérée par λ de cette façon : $D_k = \lambda D_k + (1-\lambda) D_{k-1}$. Ce paramètre λ s'il est choisi comme étant inférieur à 1 va ralentir la vitesse d'évolution de la matrice D d'une itération à la suivante. Autrement dit, ce paramètre permet d'imposer que d'une itération à la suivante l'écart entre la matrice D_k et la matrice D_{k+1} soit faible. Étant donné nos observations sur plusieurs exemples, on choisit de travailler avec $\lambda = 0.1$ et dans le cas où l'algorithme ne convergerait pas, on recommence le processus depuis le début avec un paramètre λ deux fois plus petit. Si nécessaire, on laisse la possibilité de diviser λ d'un facteur 2 cinq fois, c'est-à-dire jusqu'à ce que $\lambda = 3 \times 10^{-3}$ au minimum. Si malgré cette très faible valeur de λ, il n'y a toujours pas convergence, alors on ne retient aucune solution pour le cas étudié. En conclusion, l'introduction de ce paramètre λ facilite la convergence. On voit par exemple sur la figure 4.2.10 que sans l'introduction de λ (c'est-à-dire dans le cas $\lambda = 1$), la convergence n'est pas possible car au bout de trois itérations seulement on s'est déjà trop éloigné de l'équilibre ce qui implique que la relation 4.42 n'étant pas valide, elle donne un résultat aberrant (une distance négative). Les cas $\lambda = 0.5$ et $\lambda = 0.25$ présentent les mêmes défauts de convergence. Avec $\lambda = 0.125$, on n'obtient plus de distance négative, mais on rentre dans un cycle limite. Enfin, avec $\lambda = 0.0625$, il n'y a plus de défauts de convergence : le critère 4.44 est vérifié et la matrice trouvée, D, présente un écart de $\epsilon_{D_s D} \approx 5 \times 10^{-5}$ par rapport à D_s la solution (trouvée avec l'algorithme de Runge-Kutta).

Remarquons que dans le cas où l'algorithme itératif converge, la solution obtenue ne dépend pas de λ. En effet, ceci peut être montré à partir d'un échantillon de 1000 copolymères de tailles tirées aléatoirement dans l'intervalle [50; 150] et de potentiel d'interaction u_s et u_{ns} tirés respectivement dans les intervalles $[-5; 0]$ et $[-1; 1]$. On calcule pour chacun de ces copolymères la solution de l'équation 4.18 avec l'algorithme itératif associé à deux paramètres λ_1 et λ_2. Les deux valeurs λ_1 et $\lambda_2$sont choisies aléatoirement dans l'intervalle $]0, 0.1]$, dans le cas où un paramètre tiré ne permet pas la convergence, on en tire un autre. On obtient alors deux solutions, que l'on note $D_{\lambda 1}$ et $D_{\lambda 2}$ et on calcule l'écart $\epsilon_{D_{\lambda 1} D_{\lambda 2}}$ entre ces deux matrices. Nous avons trouvé que la distribution des $\epsilon_{D_{\lambda 1} D_{\lambda 2}}$ obtenus présente un pic principale en 10^{-7}, la variance est de 5×10^{-8} (Fig. 4.2.11). Ce résultat indique qu'en cas de convergence avec l'algorithme itératif, le résultat ne dépend pas du choix de λ.

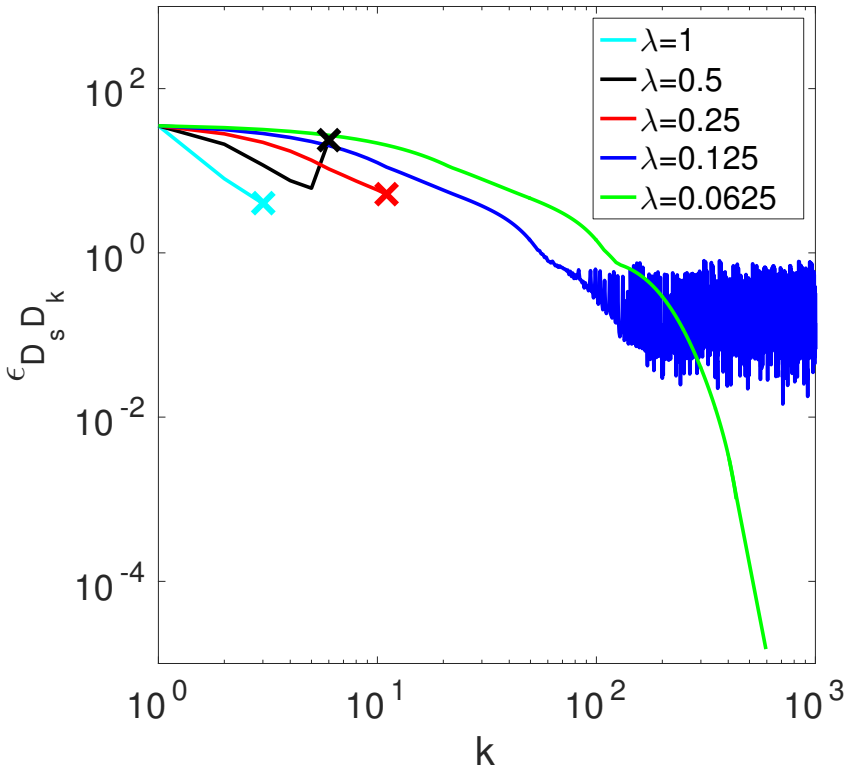

FIGURE 4.2.10 – **Intérêt du paramètre λ pour lutter contre les défauts de convergence rencontrés avec l'algorithme itératif.** Sur cette figure, on présente, en fonction de l'itération k, l'écart $\epsilon_{D_s D_k}$ (Éq. 4.41) entre la solution stationnaire, D_s, (calculée avec l'algorithme de Runge-Kutta qui ne présente pas d'instabilité numérique) et la matrice D_k trouvée lors de l'itération k avec l'algorithme itératif. Dans les cas où $\lambda = 1$, $\lambda = 0.5$ et $\lambda = 0.25$, respectivement en cyan, rouge et noir, une croix symbolise que la résolution est interrompue car une distance négative apparaît. Dans le cas où $\lambda = 0.125$, en bleu, un cycle limite se présente. Dans le cas où $\lambda = 0.0625$, en vert, la convergence se produit. En regardant la pente à l'origine des 4 courbes, on voit que plus λ est petit, plus la convergence démarre lentement. Dans le cas présenté ici, le ralentissement de la convergence permet d'éviter de rencontrer des problèmes numériques et donc de converger vers la même solution stationnaire que celle trouvée avec l'algorithme de Runge-Kutta.

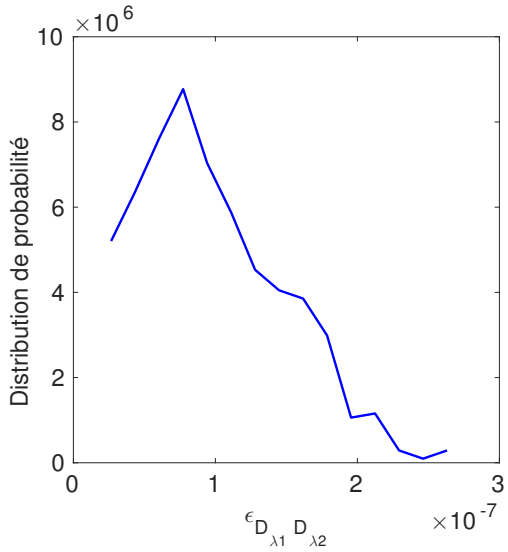

Figure 4.2.11 – **Distribution des écarts $\epsilon_{D_{\lambda 1} D_{\lambda 2}}$entre deux matrices $D_{\lambda 1}$ et $D_{\lambda 2}$ obtenues avec l'algorithme itératif et des paramètre $\lambda 1$ et $\lambda 2$ différents.** Cette distribution a été obtenue avec un échantillon de 1000 copolymères de tailles tirées aléatoirement dans l'intervalle [50; 150] et de potentiel d'interaction u_s et u_{ns} tirés respectivement dans les intervalles $[-5; 0]$ et $[-1; 1]$. Les paramètres λ_1 et λ_2 on été tirés aléatoirement dans l'intervalle $]0, 0.1]$. La moyenne étant proche de 10^{-7} et la variance environ égale à 5×10^{-8}, on peut affirmer qu'en cas de convergence avec l'algorithme itératif, le résultat ne dépend pas du choix de λ.

4.2.4.3 Comparaison des deux méthodes

La première chose que l'on puisse noter est que les deux algorithmes de résolution présentés ci-dessus, lorsqu'ils convergent, mènent à des solutions identiques. En effet, à partir d'un échantillon de 100 copolymères de tailles tirées aléatoirement dans l'intervalle [50; 150] et de potentiel d'interaction u_s et u_{ns} tirés respectivement dans les intervalles $[-5; 0]$ et $[-1; 1]$, on a calculé l'écart $\epsilon_{D_{RK} D_{AI}}$ entre la matrice solution obtenue avec l'algorithme de Runge-Kutta, D_{RK} et celle obtenue avec l'algorithme itératif, D_{AI}. Nous avons trouvé que la distribution des écarts $\epsilon_{D_{RK} D_{AI}}$ obtenus présente un pic principale en 10^{-4}, et une variance de 6×10^{-5} (Fig. 4.2.12).

Sur la figure 4.2.14 qui présente, pour les deux méthodes, les écarts entre les matrices D_k et D_{k+1} calculées d'une itération k à la suivante $k+1$, on peut voir que les matrices D_{RK} et D_{AI} obtenues à l'issue de la dernière itération (respectivement avec l'algorithme de Runge-Kutta et l'algorithme itératif) sont identiques, l'écart $\epsilon_{D_{RK} D_{AI}}$ est dans ce cas de 1.5×10^{-4}.

La différence majeure entre les deux méthodes de résolution réside dans le temps de calcul nécessaire pour obtenir la solution stationnaire. En effet, pour chaque pas de temps, la complexité du processus itératif et de l'algorithme de Runge-Kutta sont respectivement en

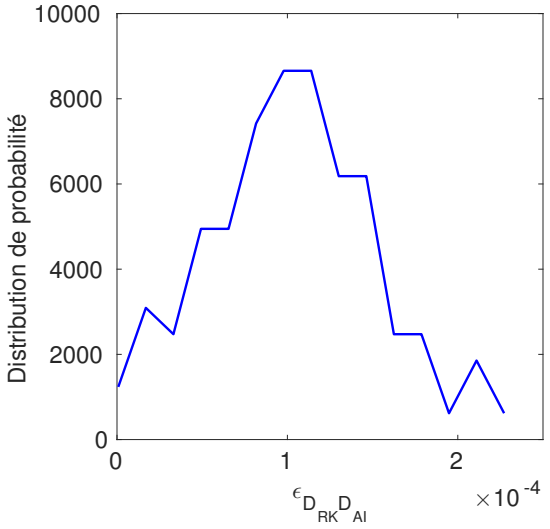

FIGURE 4.2.12 – **Distribution des écarts $\epsilon_{D_{RK}D_{AI}}$entre deux matrices D_{RK} et D_{AI} solutions de l'équation 4.18 obtenues respectivement avec l'algorithme de Runge-Kutta et avec l'algorithme itératif.** Cette distribution a été obtenue à partir d'un échantillon de 100 copolymères de tailles tirées aléatoirement dans l'intervalle [50; 150] et de potentiel d'interaction u_s et u_{ns} tirés respectivement dans les intervalles $[-5; 0]$ et $[-1; 1]$. La moyenne étant proche de 10^{-4} et la variance environ égale à 6×10^{-5}, on peut affirmer que les solutions obtenues avec les deux algorithmes sont identiques.

$\mathcal{O}(N^2)$ (coût de l'inversion d'une matrice) et $\mathcal{O}(N^3)$ (calcul de chaque terme) (Fig. 4.2.13). Le processus itératif est donc beaucoup plus rapide et permetrait de typiquement gagner un facteur 100 pour les longueurs de polymères auxquelles on s'intéresse avec notre approche d'approximation gaussienne auto-cohérente.

Sur la figure 4.2.14, on présente pour un copolymère de 120 monomères le résultat de l'intégration avec les deux méthodes. Pour cet exemple, l'écart entre les deux matrices obtenues est environ de 10^{-4}.

La convergence est plus rapide dans le cas de l'algorithme itératif mais comme on peut le voir sur la figure 4.2.14, une itération avec la méthode de Runge-Kutta apporte bien plus de changements à la matrice des distances. Cet effet est principalement dû au paramètre λ introduit plus haut afin d'éviter des problèmes numériques (distance négative ou cycle limite). Comme nous l'avons vu précédemment, il existe des cas tels que même avec un λ très faible, l'algorithme itératif ne permet pas d'obtenir la solution stationnaire, l'algorithme de Runge-Kutta sera dans ces moments là une bonne alternative. Il ne présente en effet pas d'instabilité numérique et converge donc toujours vers la solution recherchée. De plus, un avantage de l'algorithme de Runge-Kutta est qu'en résolvant pas à pas le système d'équation différentielle, il permet de suivre la dynamique de repliement (exemples de matrices D sur la figure 4.2.14 et de coefficients sur la figure 4.2.9).

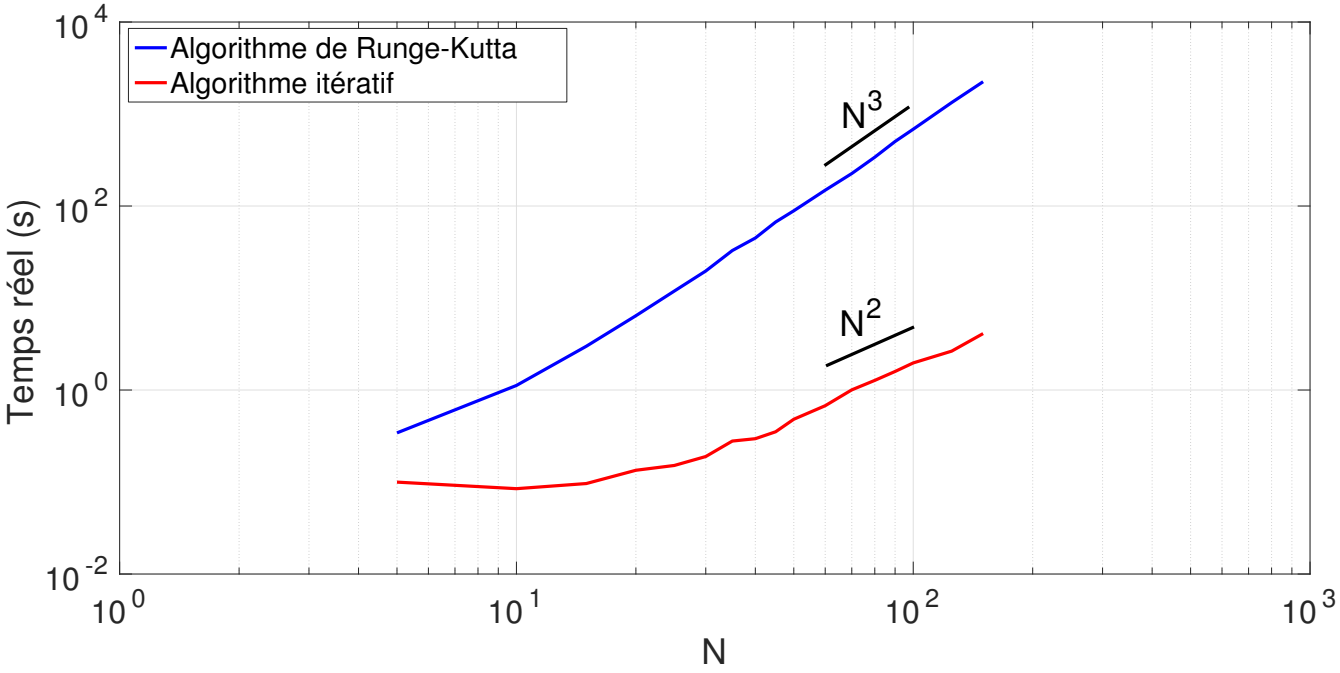

Figure 4.2.13 – **Temps de calcul nécessaire pour la résolution de l'équation 4.18 dans le cas stationnaire avec l'algorithme d'intégration itératif (en rouge) et avec celui de Runge-Kutta (en bleu) en fonction du nombre de monomères N de la chaîne.** Les deux méthodes ont respectivement un temps fonction de N^2 et N^3 avec N le nombre total de monomères du polymère étudié.

4.2.5 Résultats : Applications du modèle

Dans cette sous section nous allons illustrer à partir de notre modèle le comportement générique à l'équilibre d'un homopolymère, d'un hétéropolymère simple composé de deux types de blocs et d'un hétéropolymère dont l'état des blocs est défini selon l'épigénome de la drosophile. Les polymères seront constitués de 120 monomères de 10kb, soit 1.2Mb. Le fait d'étudier des bouts de 1.2Mb n'est pas si déraisonnable qu'il peut y paraître puisque l'on a introduit dans la partie 3.1.3 que plus de 50% des contacts intra chromosomiques sont détectés pour des distances génomiques inférieures à 1Mb et puisque l'on introduit dans notre modèle un champ générique rendant compte de l'environnement de manière effective.

Pour les différents hamiltoniens définis en section 4.2.3, on résoudra le système d'équations différentielles défini par l'équation 4.18 dans le cas stationnaire, donc avec $\zeta \frac{dD_{ij}}{dt} = 0$. Cette résolution est réalisée avec l'algorithme itératif ou dans le cas où celui-ci ne converge pas avec l'algorithme de Runge-Kutta.

Comme expliqué dans la section précédente, la matrice des distances carrés moyennes D à l'état stationnaire obtenue après intégration de l'équation 4.18, nous donne la possibilité de générer des configurations typiques (cf. 4.2.1.5) et nous permet de calculer la matrice des

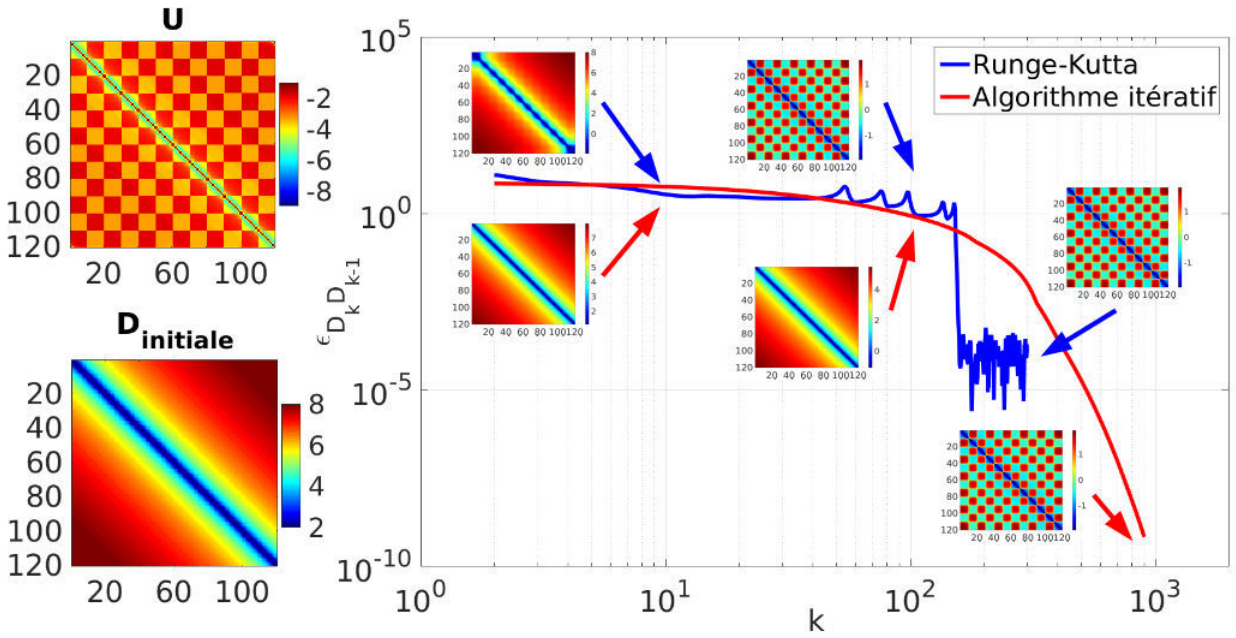

FIGURE 4.2.14 – **Exemple de résolution du système 4.18 pour le copolymère $(A_{10}B_{10})_6$ avec deux algorithmes différents.** En haut à gauche : matrice des potentiels d'interaction U caractérisant le polymère étudié ici. En bas à gauche : matrice des distances initiales choisie pour débuter la résolution de l'équation **4.18**. À droite : évolution de l'écart entre la matrice D_k calculée lors de l'itération k et la matrice D_{k-1} calculée lors de l'itération précédente (Éq. 4.41) en fonction de l'itération k et dans le cas de la résolution avec la méthode de Runge-Kutta et avec l'algorithme itératif ($\lambda = 0.025$) (resp. en bleu et en rouge). On peut voir en insert des exemples de matrices D_k correspondant aux itérations $k = 10$, $k = 100$ et k correspondant à la dernière itération. L'algorithme itératif et celui de Runge-Kutta ont respectivement convergé en 3000s et 3s, donc une différence d'un facteur 1000 pour ce cas précis.

contacts, C. Cette conversion de C à D se fait avec la relation 4.13. Dans tout cette section « Résultats », la constante A impliquée dans cette équation 4.13 est calculée en imposant arbitrairement que la distance moyenne entre deux monomères plus proches voisins soit de 1. Cette condition implique qu'en moyenne pour tout monomère i, $D_{i\,i+1} \approx 1/3$ puisque pour rappel, $D_{ij} = \left\langle (X_i - X_j)^2 \right\rangle /3)$. La constante est donc $A = \frac{1}{N} \frac{\sum_i C_{i\,i+1}}{(1/3)^{-3/2}}$. Comme C et D sont donc liés par une relation mathématique claire, on ne présentera dans cette partie que les matrices C, et pas D afin d'éviter une redondance.

4.2.5.1 Conditions initiales et multistabilité mathématique

Nous avons vu dans la partie 4.2.4 que la résolution du système 4.18 nécessite au départ une matrice des distances initiales, $D_{initiale}$. Afin de sonder l'éventuelle existence de plusieurs points fixes pour le système dynamique 4.18, on débutera alors systématiquement la

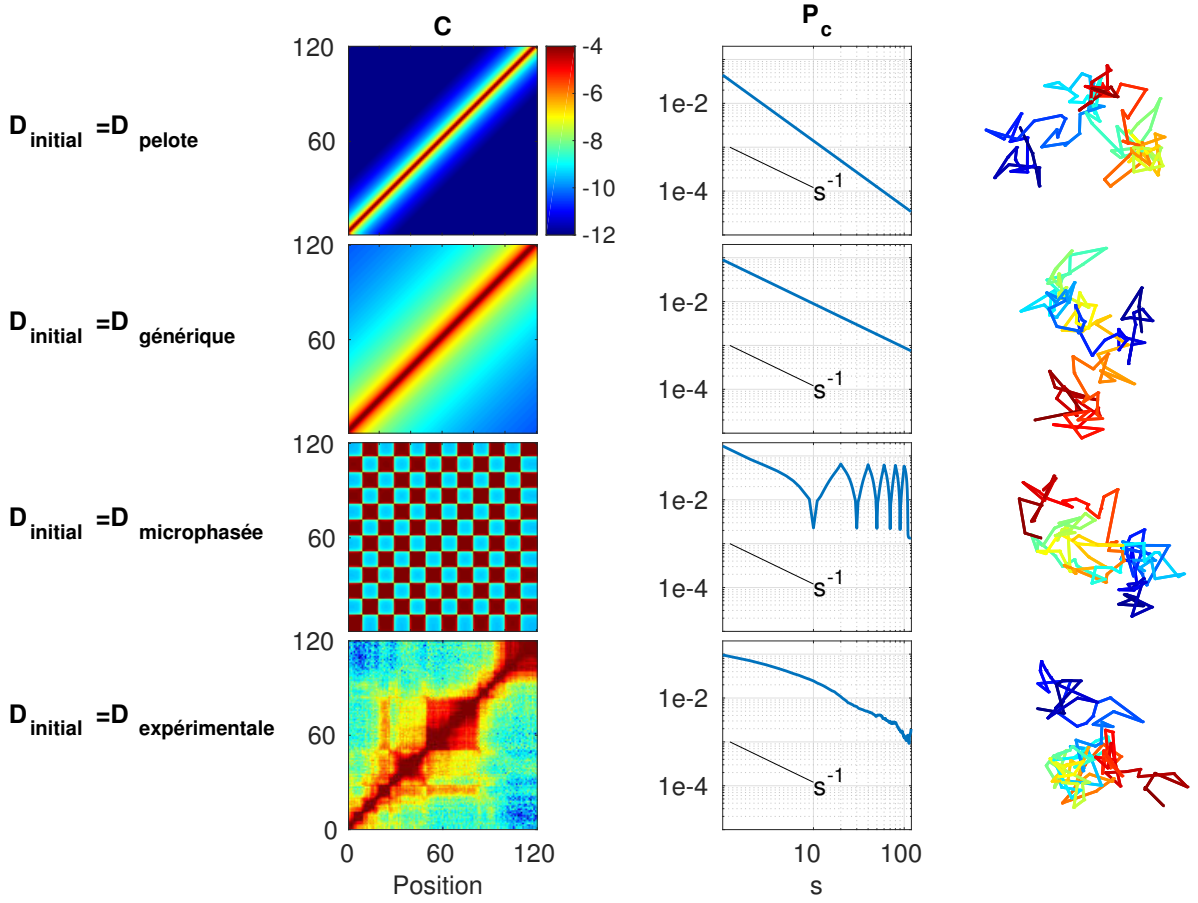

Figure 4.2.15 – **Configurations initiales pour la résolution de l'équation 4.18.** De gauche à droite, carte des contacts, C, (avec même échelle de couleur pour toutes), probabilité de contact P_c en fonction de la distance génomique s et exemple d'une structure obtenue avec l'approximation gaussienne. De haut en bas, les quatre configurations choisies comme état initial : polymère gaussien sous forme de pelote, polymère dans sa phase générique (construit à partir d'une probabilité de contact évoluant en s^{-1}), polymère dans un état microphasé et polymère issu d'un cas expérimental.

résolution avec quatre conditions initiales très différentes : polymère dans sa phase gaussienne, générique, microphasée et un cas expérimental (Fig. 4.2.15). Ces matrices $D_{initiale}$ sont construites avec le formalisme auto-cohérent. Concrètement, il s'agit de résoudre l'équation 4.18 pour obtenir les matrices $D_{initiale}$ recherchées en commençant, pour ces cas là, la résolution avec comme matrice des distances initiales, une matrice gaussienne calculée avec l'équation **4.7** qui relie D et K_1^{-1} avec la matrice K_1 donnée par la formule **4.6**. Pour obtenir la matrice $D_{initiale}$ correspondant au polymère sous forme de pelote, on résout le système avec l'hamiltonien H_{2A} et une matrice d'interaction entre monomères totale $U = U_{ns} + U_s = 0$. Pour obtenir la forme générique, on résout 4.18 avec H_{2I} et $U = 0$. Pour la microphase, on utilise H_{2A} et $U = U_{ns}$ avec $U_{ns_{ij}} = -6\delta_{e_i e_j}$ avec $\delta_{e_i e_j} = 1$ si i et j sont dans le même état

épigénomique. Enfin, la dernière matrice initiale a été obtenue à partir de la carte de contact expérimentale de la drosophile entre 12.5 et 13.7Mb [Sexton et al., 2012]. La transformation de la carte de contact en carte de distance s'est faite avec la relation 4.13.
Pour un copolymère donné, on comparera les quatre solutions obtenues que l'on note D_1, D_2, D_3, D_4 en calculant l'écart $\epsilon_{D_iD_j}$ entre chaque couple de solution (Éq. 4.45).

$$\epsilon_{D_iD_j} = \sqrt{max_{k,l}\left((D_{i_{kl}} - D_{j_{kl}})^2\right)} \tag{4.45}$$

Après observation de plusieurs exemples, et de manière cohérente avec la précision donnée par les algorithmes de résolution (cf. section 4.2.4 et en particulier Fig. 4.2.12), on choisit de dire que deux matrices solutions D_i et D_j sont identiques si l'écart entre elles est inférieur à 10^{-3}.
Dans le cas où, pour le même système 4.18, des solutions différentes sont trouvées selon la condition initiale utilisée, on dira que le système présente une multistabilité mathématique (c'est-à-dire l'existence de plusieurs points fixes pour le système dynamique 4.18). On considère que si, ϵ_4, le maximum des écarts calculés, ($\epsilon_4 = max_{i,j\in[\![1,4]\!]}\left(\epsilon_{D_iD_j}\right)$), est supérieur à 10^{-3}, il existe au moins deux points fixes. Le critère pour repérer la multistabilité sera donc le suivant :

$$\epsilon_4 = max_{(i,j)\in[\![1,4]\!]}\left(\epsilon_{D_iD_j}\right) > 10^{-3} \tag{4.46}$$

On présente sur la figure 4.2.16, le résultat de la résolution du système 4.18 avec l'hamiltonien H_{1A} et avec comme paramètres d'interaction spécifique et non spécifique : $u_s = -40k_BT$ et $u_{ns} = -40k_BT$. On observe que pour une condition initiale, la solution stationnaire obtenue est différente de celle obtenue avec les trois autres conditions initiales. Pour ce système, on observe donc le phénomène de multistabilité mathématique avec $\epsilon_4 \approx 50$. On peut préciser ici que le fait de trouver des solutions différentes n'a pas pour origine une instabilité numérique. En effet, que l'on résolve l'équation 4.18 avec l'algorithme de Runge-Kutta ou avec l'algorithme itératif (cf. section section 4.2.4), les résultats sont identiques, c'est-à-dire que la figure 4.2.16 est la même pour les deux algorithmes. Ainsi, le fait de trouver des solutions différentes s'explique par l'existence de plusieurs points fixes pour le système 4.18.

4.2.5.2 Homopolymère

Afin de valider l'algorithme développé, on commence par étudier un homopolymère auto-évitant isolé non soumis à un champ générique et sans interaction entre monomères. Son hamiltonien H est donc composé de deux contributions $H = H_{connectivité} + H_{ve}$. Son hamiltonien peut être H_1 ou H_2 selon si on modélise le volume exclu avec un potentiel de

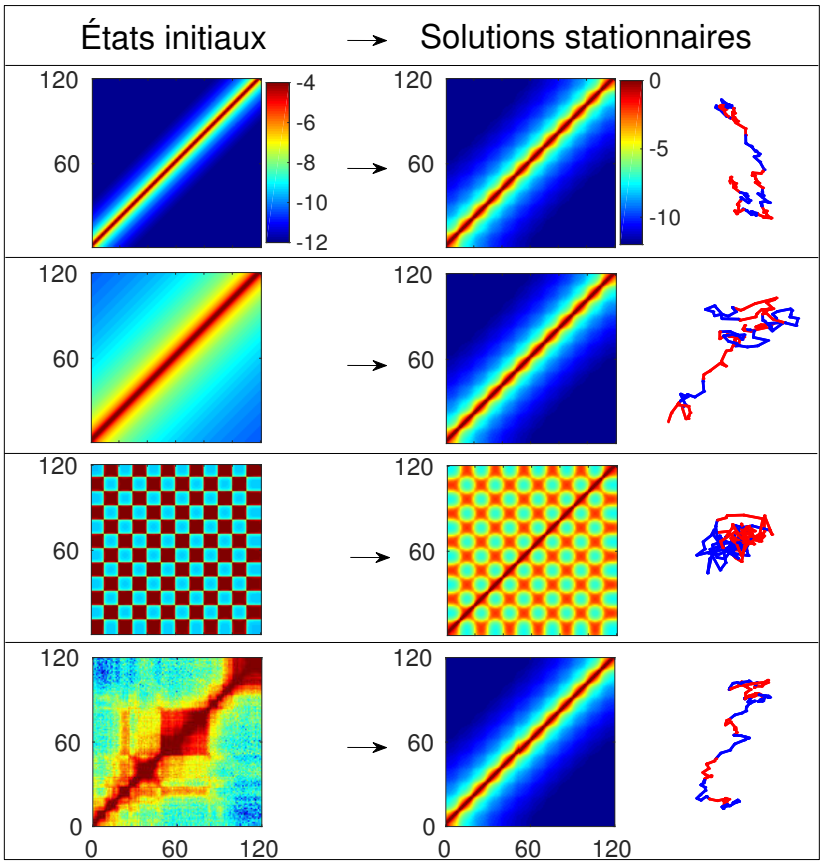

Figure 4.2.16 – **Exemple de multistabilité mathématique obtenue pour un hétéropolymère $(A_{10}B_{10})_6$ modélisé avec l'hamiltonien H_{1A} et avec comme paramètres d'interaction spécifique et non spécifique : $u_s = -40k_BT$ et $u_{ns} = -40k_BT$.** Chacune des quatre lignes présente la résolution du système 4.18 débutée la condition initiale présentée dans la colonne de gauche. Colonne de gauche : cartes de contact initiales utilisées dans l'algorithme itératif pour débuter la résolution de l'équation 4.18 (l'échelle de couleur est donné par la première carte). Colonne de droite : cartes de contact des solutions obtenues à l'issue de la résolution (l'échelle de couleur est donné par la première carte) accompagnées d'un exemple de structure où les blocs A sont représentés en rouge et les blocs B en bleu. On observe que la solution obtenue à la troisième ligne est différente des trois autres : ceci indique l'existence d'une multistabilité mathématique pour le cas étudié ici.

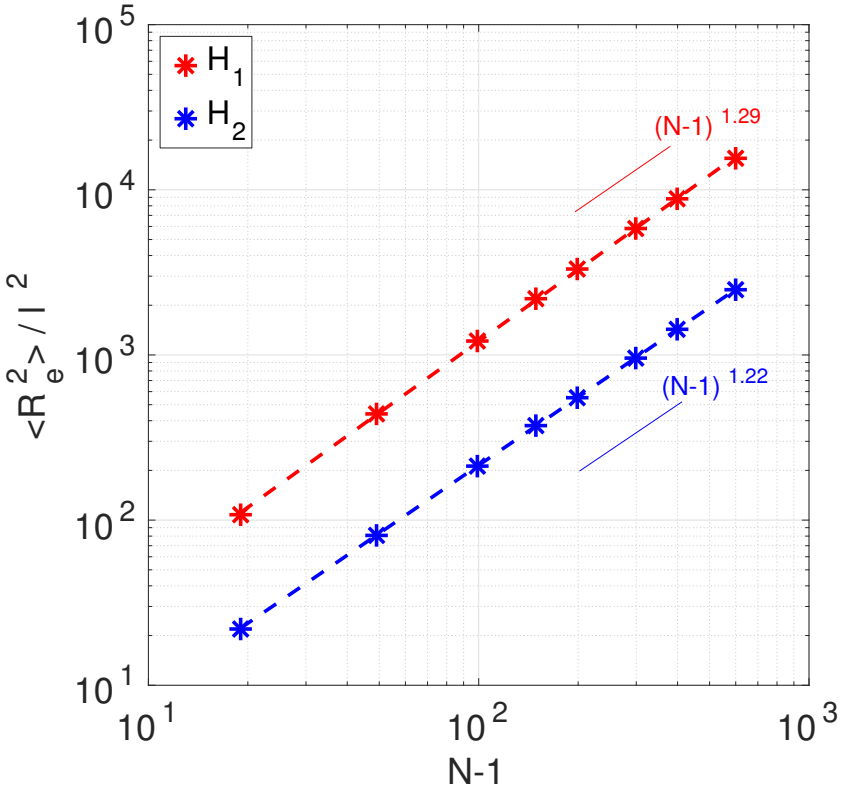

FIGURE 4.2.17 – **Rayon bout-à-bout quadratique moyen, $\langle R_e^2 \rangle$, en fonction du nombre de segments dans la chaîne, $N-1$ dans le cas d'un homopolymère auto-évitant isolé non soumis à un champ générique et sans interaction entre monomères modélisé dans le cadre de l'approximation gaussienne auto-cohérente.** Le modèle reproduit quelque que soit la modélisation du volume exclu (H_1, potentiel de type Lennard-Jones et H_2, potentiel gaussien) la loi de puissance $R_e \sim (N-1)^{2\nu}$ avec $\nu \approx 0.6$ représentée en tirets [Gennes, 1979].

type Lennard-Jones (H_{ve1} donné par l'équation 4.24) ou de type gaussien (H_{ve2} donné par l'équation 4.26). Dans les deux cas, on peut voir sur la figure 4.2.17 que le rayon bout-à-bout quadratique moyen, $\langle R_e^2 \rangle$, en fonction du nombre de segments dans la chaîne, $N-1$, évolue selon $R_e \sim (N-1)^{2\nu}$ avec $\nu \approx 0.6$ ce qui est une propriété générique des chaînes auto évitantes [Gennes, 1979]. Ce résultat constitue une première validation de l'approche gaussienne auto-cohérente. Par ailleurs, on peut remarquer que c'est en modélisant le volume exclu avec un potentiel gaussien que le coefficient est plus proche de 0.6.

On va maintenant ajouter le terme $H_{générique}$ de manière à étudier un homopolymère ($U_S = 0$) auto-évitant isolé « crumpled » composé de $N = 120$ monomères, soit 1.2Mb. L'hamiltonien de ce système peut être H_{1A}, H_{2A}, H_{1B} ou H_{2B} selon la modélisation du volume exclu et du champ générique (Tab. 4.2). L'idée derrière l'étude de ce cas simple est de premièrement vérifier de nouveau la validité de notre modèle qui repose sur une approxima-

		Volume exclu modélisé par un potentiel de type :	
Champ générique		Lennard-Jones tronqué	gaussien
	ajusté avec 1 paramètre	H_{1A} (Éq. 4.34)	H_{2A} (Éq. 4.36)
	ajusté avec $\frac{N(N+1)}{2}$ paramètres	H_{1B} (Éq. 4.35)	H_{2B} (Éq. 4.37)

Table 4.2 – Résumé des différences entre les quatre hamiltoniens H_{1A}, H_{1B} H_{2A} et H_{2B} utilisés pour modéliser la chaîne.

tion gaussienne auto-cohérente et de deuxièmement analyser quelles différences engendrent le choix de l'hamiltonien.

Dans le cadre de l'approximation gaussienne auto-cohérente mise en place, on peut observer le repliement d'un polymère d'un point de vue dynamique en utilisant l'algorithme de Runge-Kutta (cf. section 4.2.4) (Fig. 4.2.18). Sur la figure 4.2.18 par exemple, on observe l'évolution d'un homopolymère de 120 monomères modélisé avec l'hamiltonien H_{2B} , étant sous phase de pelote à l'état initial ($u_{ns} = 0.5 k_B T$) et auquel on impose des interactions non spécifiques entre monomères de $-0.5 k_B T$. Ces attractions entre monomères ont pour effet de confiner la chaîne qui va subir une transition θ. On voit sur la figure 4.2.18 que la nucléation se fait par les bouts qui sont les régions de la chaîne les plus libres donc les plus mobiles pour pouvoir interagir avec les autres. Suite à cela, deux compartiments se forment de manière transitoire. Ils fusionnent et la chaîne se retrouve alors sous phase globulaire, ce qui constitue l'état stationnaire. Cette dynamique de repliement, non triviale, ne dépend pas de l'hamiltonien (parmi les quatre que l'on propose, Tab. 4.2) et elle a été précédemment étudiée en détail par Byrne et al., 1995. L'enjeu qui découle de ces simulations est de pouvoir accéder à la relation entre temps simulé et temps réel. Nous en parlerons succintement dans la partie 4.4 où on présentera des simulations de dynamique moléculaire concernant la drosophile et réalisées par Pascal Carrivain.

En faisant varier l'intensité des interactions non spécifiques, u_{ns}, entre monomères on va pouvoir observer la transition θ de la chaîne distinguant la phase pelote de la phase globule. Cette transition peut être repérée en étudiant le rayon de gyration de la chaîne, R_g^2 (Éq. 4.47).

$$R_g^2 = \frac{1}{2N^2} \sum_{i,j} D_{ij} \tag{4.47}$$

Comme on peut le voir sur la figure 4.2.19, le rayon de gyration quadratique de la chaîne en fonction du paramètre u_{ns} présente un point d'inflexion caractéristique de la transition θ.

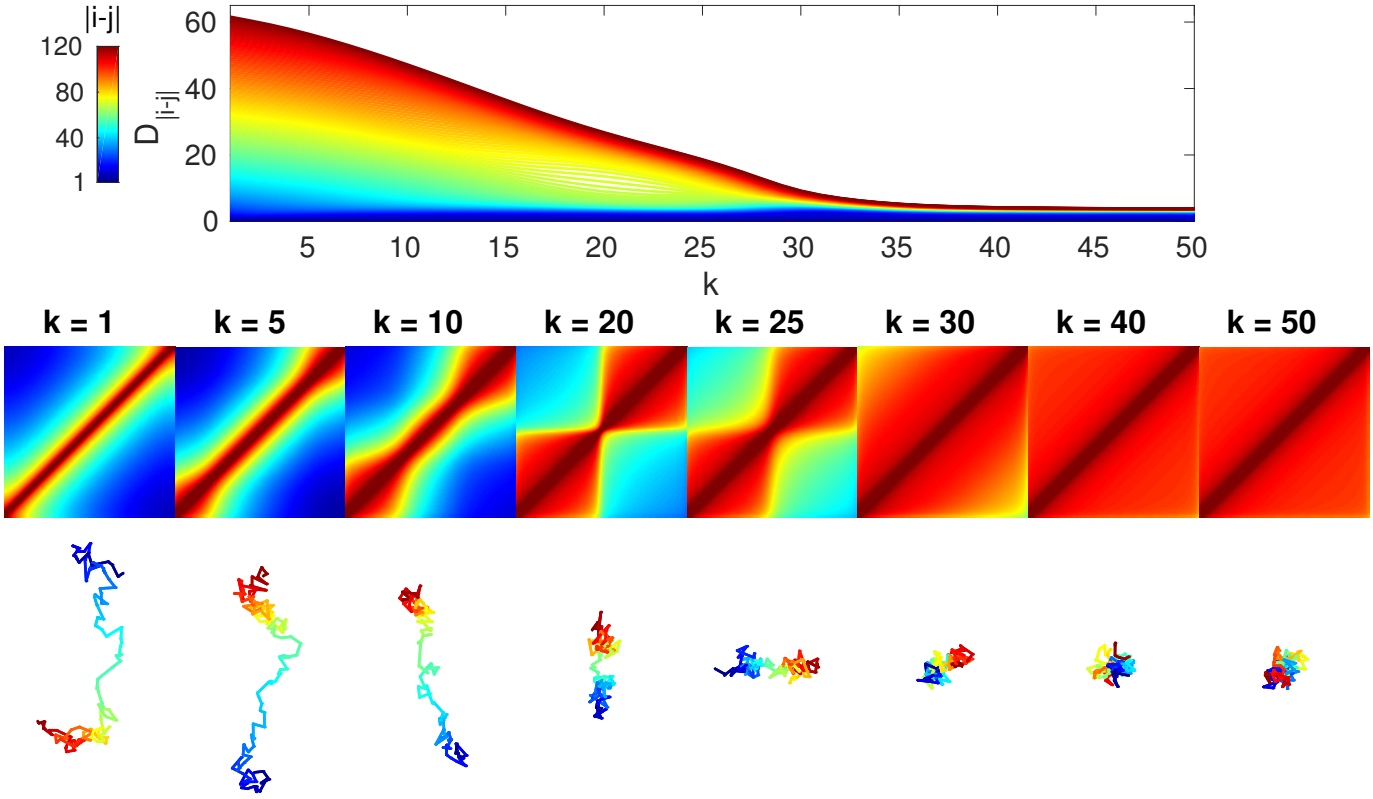

FIGURE 4.2.18 – **Dynamique de repliement d'une pelote en globule.** Le système est un homopolymère de 120 monomères associé à l'hamiltonien H_{2B} de paramètre $u_{ns} = -0.5 k_B T$. L'équation de la dynamique est intégrée avec l'algorithme de Runge-Kutta. Le graphe du haut représente l'évolution des coefficients $D_{|i-j|}$ en fonction de l'itération k (proportionnelle au temps réel) et en fonction de la distance génomique $|i-j|$ définie par l'échelle de couleur à gauche. En dessous, on représente les cartes de contact pour les itérations 1 (état initial), 5, 10, 20, 25, 30, 40 et 50 (état stationnaire). L'échelle de couleur de ces cartes est donnée figure 4.2.21. On accompagne chaque carte d'une structure représentative générée dans le cadre de l'approximation gaussienne.

Ce point est atteint pour les quatre hamiltoniens pour des valeurs de u_{ns} proche de 0.

Nous allons maintenant illustrer la transition θ de cet homopolymère en observant son diagramme de phase élaboré en faisant varier le paramètre u_{ns} qui est l'intensité des interactions non spécifiques en unité $k_B T$ ($\forall (i, j)$, $U_{ns_{ij}} = u_{ns}$). Les diagrammes de phase obtenus avec les hamiltoniens H_{1X} et H_{2X} sont respectivement représentés sur la figure 4.2.20 et la figure 4.2.21. Chaque diagramme de phase est composé des cartes de contact simulées, de la probabilité de contact moyenne P_c en fonction de la distance génomique s, et d'un exemple de structure représenté dans le cadre de l'approximation gaussienne. On peut constater qu'avec les quatre hamiltoniens modélisant la chaîne, l'évolution des configurations en fonction de l'intensité de l'interaction non spécifique, u_{ns}, est qualitativement identique. En effet, dans

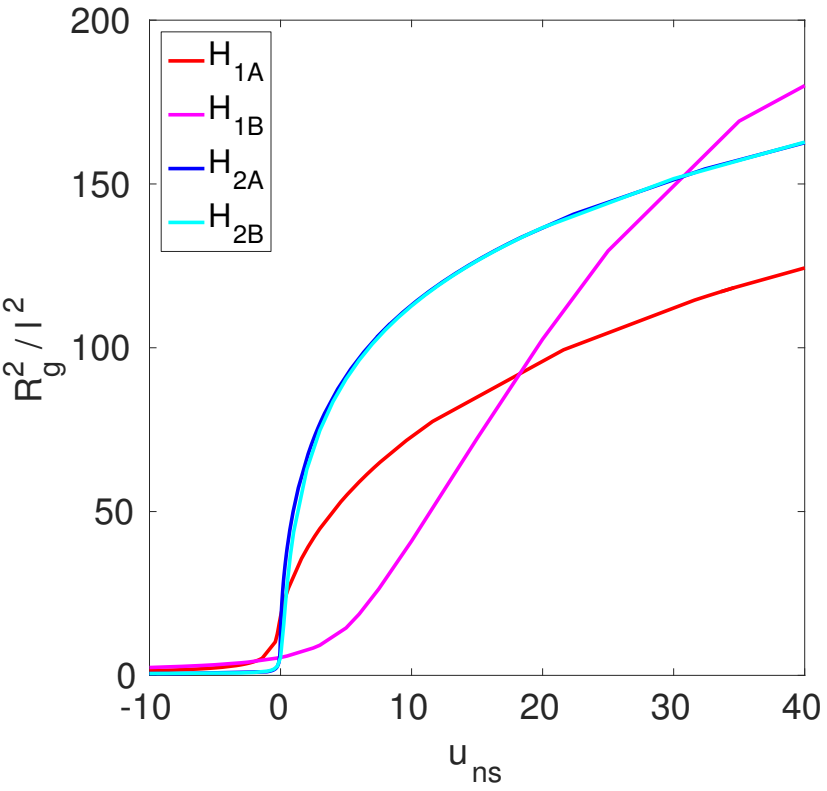

Figure 4.2.19 – **Rayon de gyration d'un homopolymère auto-évitant isolé crumpled composé de $N = 120$ monomères, en fonction de l'intensité des interactions non spécifiques u_{ns} en unité k_BT.** Les courbes respectivement rouge, magenta, bleu et cyan, ont été obtenues à partir des hamiltoniens H_{1A}, H_{1B} , H_{2A} et H_{2B}. Le point d'inflexion de chaque courbe correspond au point θ marquant la transition entre globule et pelote.

tous les cas, lorsque $u_{ns} > 0$, (interactions répulsives) on observe toujours une pelote d'autant plus allongée que u_{ns} est grand, avec une probabilité de contact qui décroît selon une loi type $P_c \sim s^{-\alpha}$ avec $\alpha > 1$. La diminution des interactions non spécifiques répulsives au contraire impose un confinement de la chaîne plus fort. La chaîne se replie alors sur elle même. Une transition θ se produit pour $u_{ns} \approx 0$. A ce moment là, la probabilité de contact moyenne évolue selon $P_c \sim s^{-1}$. Si on continue à augmenter le confinement après le point θ, la chaîne passe en phase globulaire. Dans ce dernier cas, aux grandes échelles, la probabilité contact ne dépend plus de s, $P_c \sim s^0$. Ces diagrammes de phase montrent que le paramètre u_{ns} permet de retrouver le résultat selon lequel le passage de la pelote au globule s'accompagne d'une variation continue de la valeur de α intervenant dans la loi $P_c \sim s^{-\alpha}$ [Barbieri et al., 2012]. Il semblerait en effet que α évolue continuement de -2 à 0.

Concernant les différences entre les quatre diagrammes de phase, premièrement, on remarque que l'ordre de grandeur de u_{ns} permettant d'observer des différences de structures notables doit être plus élevé (environ d'un facteur 10) dans le cas de la modélisation avec H_{1X} que

dans le cas de H_{2X}. Ceci est une conséquence du fait que le volume exclu est plus intense avec H_{1X} plutôt qu'avec H_{2X} (cf. Fig. 4.2.4). Deuxièmement, qu'on observe les cartes de contact ou bien qu'on observe la décroissance de la probabilité de contact, on voit que les variations de structures en fonction de u_{ns} sont plus abrupte avec H_{XA} plutôt qu'avec H_{XB}. En effet, lorsque l'on modélise l'homopolymère avec un hamiltonien du type H_{XB} (c'est-à-dire lorsque le champ générique est inféré avec $\frac{N(N+1)}{2}$ paramètres) le comportement de P_c reste toujours assez proche de la loi $P_c \sim s^{-1}$ en comparaison avec son comportement lorsque la modélisation est réalisée avec H_{XA}. On en conclut que l'ajout d'interactions non spécifiques éloigne moins la fonction $P_c(s)$ de la loi générique $P_c \sim s^{-1}$ si le champ générique est ajusté avec $\frac{N(N+1)}{2}$ paramètres plutôt qu'un seul. La modélisation avec H_{XB} permet donc de mieux contrôler le comportement de P_c.
En conclusion, dans le cadre de l'approximation gaussienne auto-cohérente, on a pu d'une part simuler la dynamique de relaxation au cours d'une transition θ et d'autre part retrouver les propriétés génériques d'un homopolymère isolé auto-évitant. Les quatre hamiltoniens ont menés à des résultats qualitatifs identiques. Par contre, les échelles d'énergie en jeux ne sont pas comparable d'un hamiltonien à l'autre, en particulier, (i) les ordres de grandeur d'interaction non spécifiques nécessaires pour observer des changements de structures significatifs doivent environ être dix foix plus grands si la modélisation est réalisée avec un hamiltonien type H_{1X} par rapport au cas de H_{2X} et, (ii) les variations de la probabilité de contact, P_c, sont plus douces avec les hamiltoniens type H_{XB} par rapport à H_{XA} . Au final, il semble que l'hamiltonien H_{2B} soit le plus adapté pour modéliser la chromatine puisque d'une part il donne une loi d'échelle pour $\langle R_e^2 \rangle$ proche de $(N-1)^{2\times 0.6}$, ce qui est théoriquement attendu (Fig. 4.2.17) et il permet d'autre part de mieux contrôler la décroissance de P_c (proche de s^{-1}) par rapport au cas d'une modélisation avec H_{2A}.

4.2.5.3 Hétéropolymère formé de deux types de bloc

Afin d'étudier les effets d'interactions spécifiques entre monomères, nous allons commencer par étudier un cas simple, à savoir les propriétés de repliement d'un copolymère par bloc noté, $(A_{10}B_{10})_6$, de taille 120, soit 1.2Mb, composé de 12 blocs : où chaque bloc représente un domaine épigénomique actif (A) ou inactif (B) de taille, 100kb ce qui représente 10 monomères. Les interactions $A-A$ et $B-B$ seront attractives avec la même intensité définie par le paramètre u_s. L'intensité du confinement sera modélisée avec le paramètre u_{ns}. Cette étude sera menée avec les quatre hamiltoniens (Tab. 4.2).

4.2.5.3.1 Caractérisation des frontières entre les différentes phases

On étudie ici le copolymère par bloc $(A_{10}B_{10})_6$ en faisant varier l'intensité des interactions

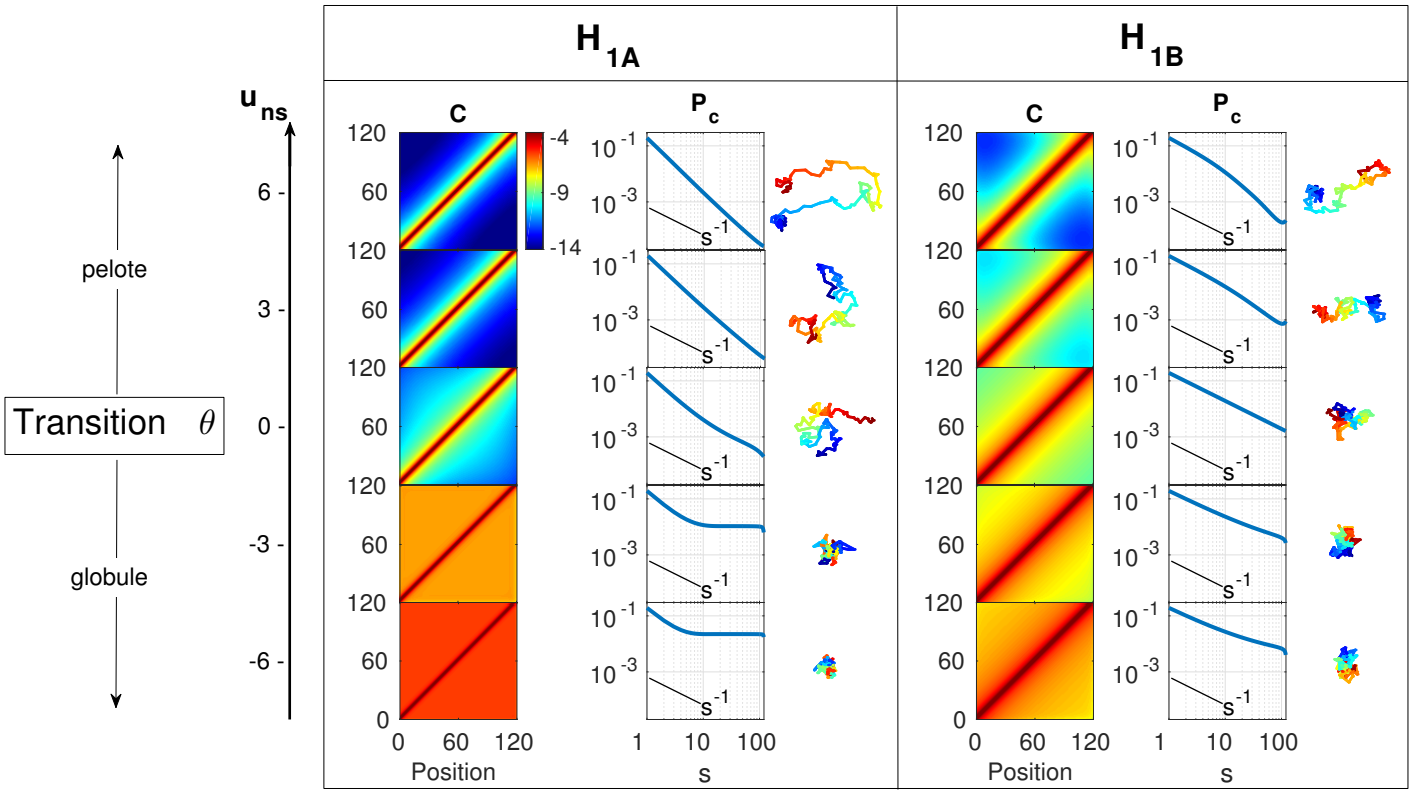

Figure 4.2.20 – **Diagrammes de phase d'une chaîne homopolymèrique isolée auto-évitante « crumpled » en fonction du paramètre d'interaction non spécifique u_{ns} et pour les deux hamiltoniens H_{1A} et H_{1B}.** Dans chacun des deux panels, on présente de gauche à droite, la carte des contacts, C, (l'échelle de couleur indiquée à droite de la première carte est aussi valable pour toutes les autres), la probabilité de contact moyenne P_c en fonction de la distance génomique s, en bins, et un exemple de structure générée dans le cadre de l'approximation gaussienne. La structure est représentée du premier monomère au dernier avec un dégradé respectivement du bleu vers le rouge. Pour des interactions non spécifiques de plus en plus attractives (du haut vers le bas), la chaîne se replie et passe par la transition θ de pelote vers globule autour de $u_{ns} \approx 0$.

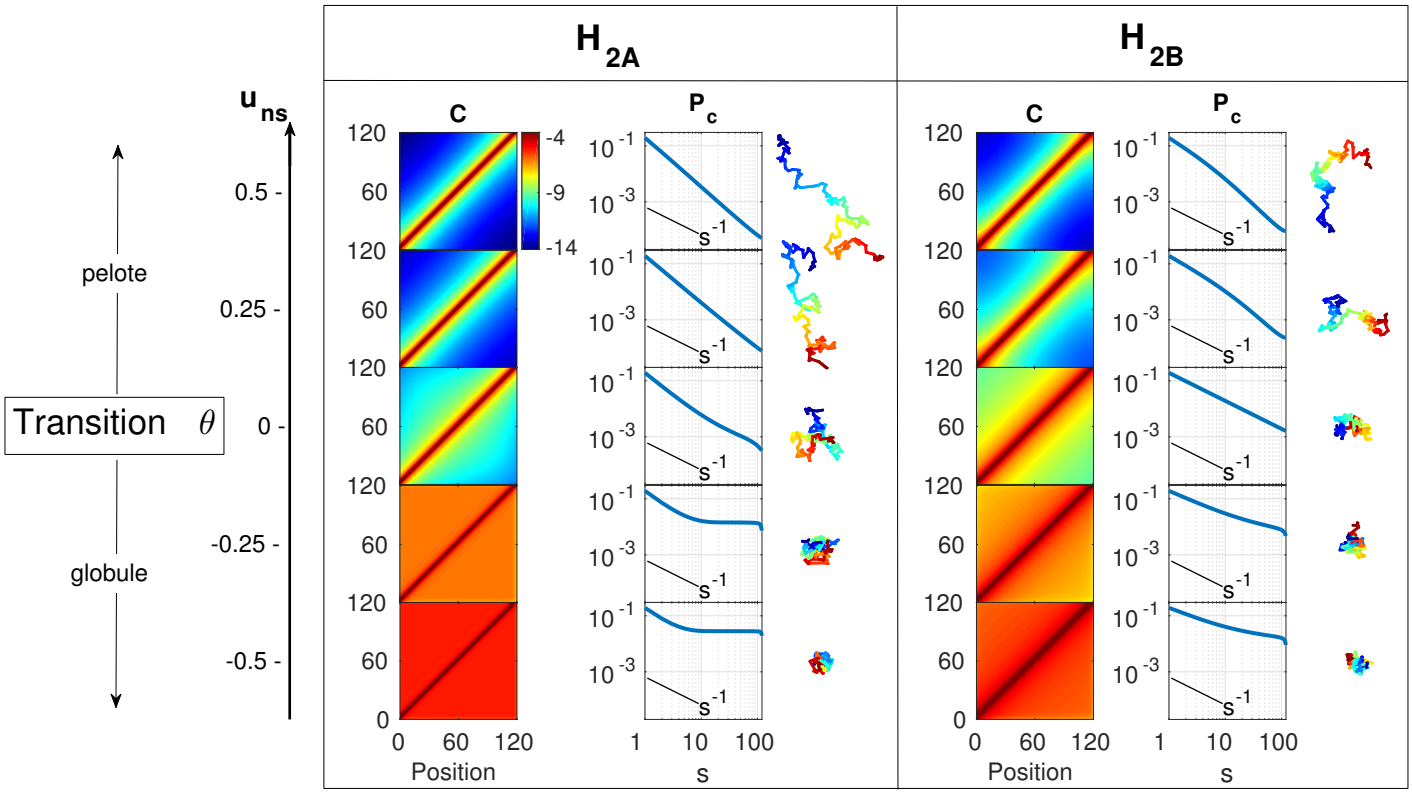

FIGURE 4.2.21 – **Diagrammes de phase d'une chaîne homopolymèrique isolée auto-évitante « crumpled » en fonction du paramètre d'interaction non spécifique u_{ns} et pour les deux hamiltoniens H_{2A} et H_{2B}.** La légende de cette figure est identique à celle de la figure 4.2.20.

non spécifiques, u_{ns} et l'intensité des interactions spécifiques entre monomères de même état épigénomique, u_s. Avec les hamiltoniens, H_{1X}, on simule le repliement du copolymère pour une grille de paramètres (u_s, u_{ns}) telle que $u_s \in [-40, 0]$ et $u_{ns} \in [-10, 30]$ avec un pas de $2k_BT$. Avec les hamiltoniens, H_{2X}, les ordres de grandeur sont différents : on réalise des simulations sur la grille de paramètres (u_s, u_{ns}) telle que $u_s \in [-5, 0]$ et $u_{ns} \in [-1, 4]$ avec un pas de $0.1k_BT$. Quelques cartes de contact obtenues à partir de ces deux grilles sont respectivement représentées sur les figures 4.2.22, 4.2.23, 4.2.24 et 4.2.25. Pour chaque couple de paramètres, on représente sur ces figures quatre cartes correspondant aux quatre conditions initiales testées. On peut observer sur ces deux figures que malgré la simplicité du copolymère $(A_{10}B_{10})_6$ qui n'est construit qu'à partir de deux états (A et B), les diagrammes de phase obtenus sont riches avec l'existence de quatre régions distinctes selon les intensités d'interaction.

Dans la première région, à faible spécificité / faible confinement (coin en haut à droite), le copolymère se replie sous forme de pelote. L'augmentation du confinement avec le paramètre

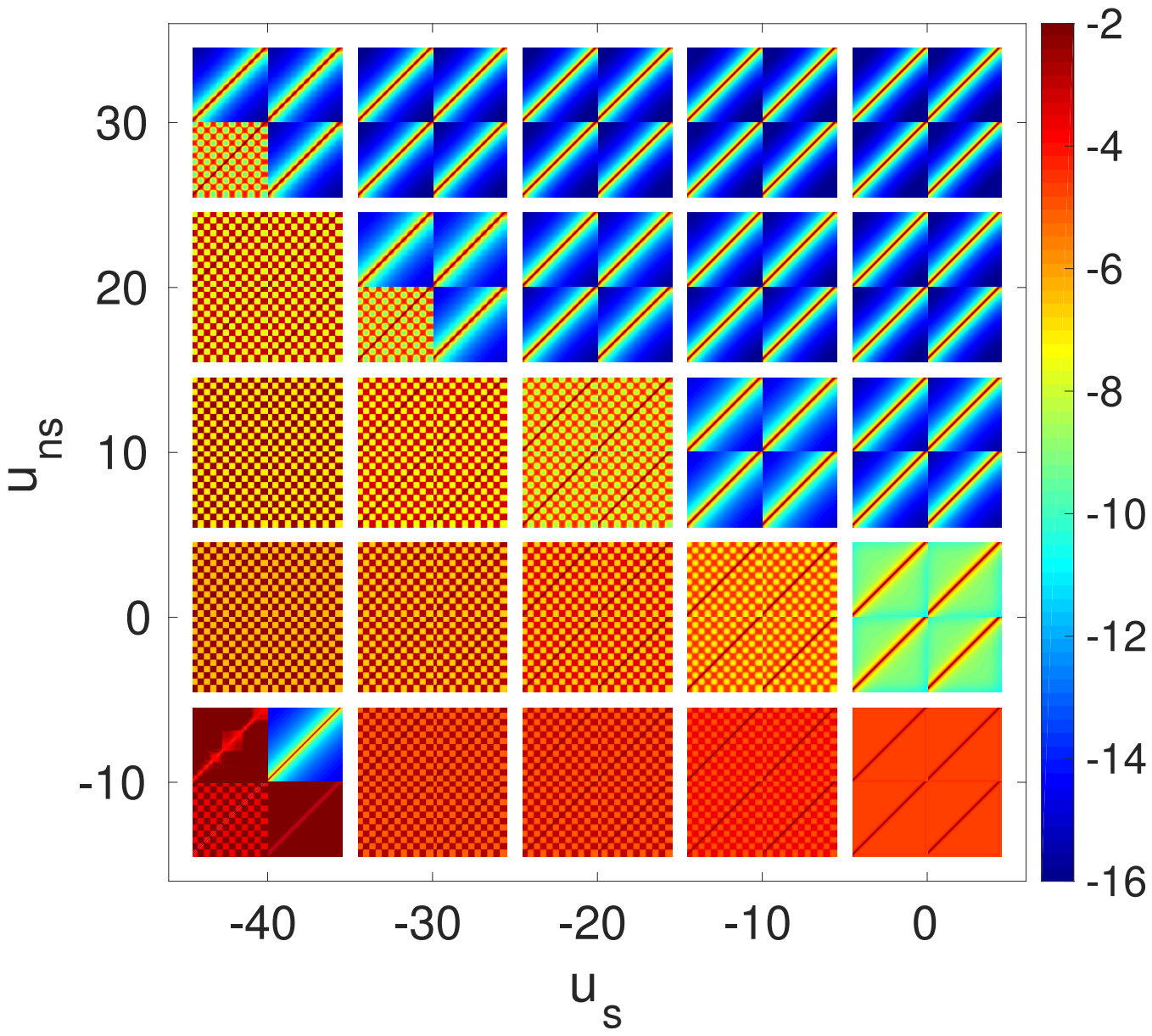

Figure 4.2.22 – **Cartes de contact à l'état stationnaire du copolymère $(A_{10}B_{10})_6$ associé à l'hamiltonien H_{1A} en fonction de l'intensité des interactions spécifiques u_s et non spécifiques u_{ns}.** Ces prédictions sont obtenues dans le cadre de l'approximation gaussienne auto-cohérente. La résolution de l'équation 4.18 est réalisée à partir de quatre états initiaux différents dont les cartes de contact sont représentées dans le coin en bas à gauche. Il existe une région où selon la condition initiale, plusieurs états stationnaires sont possibles. L'équilibre thermodynamique est une moyenne pondérée de tout les états stationnaires.

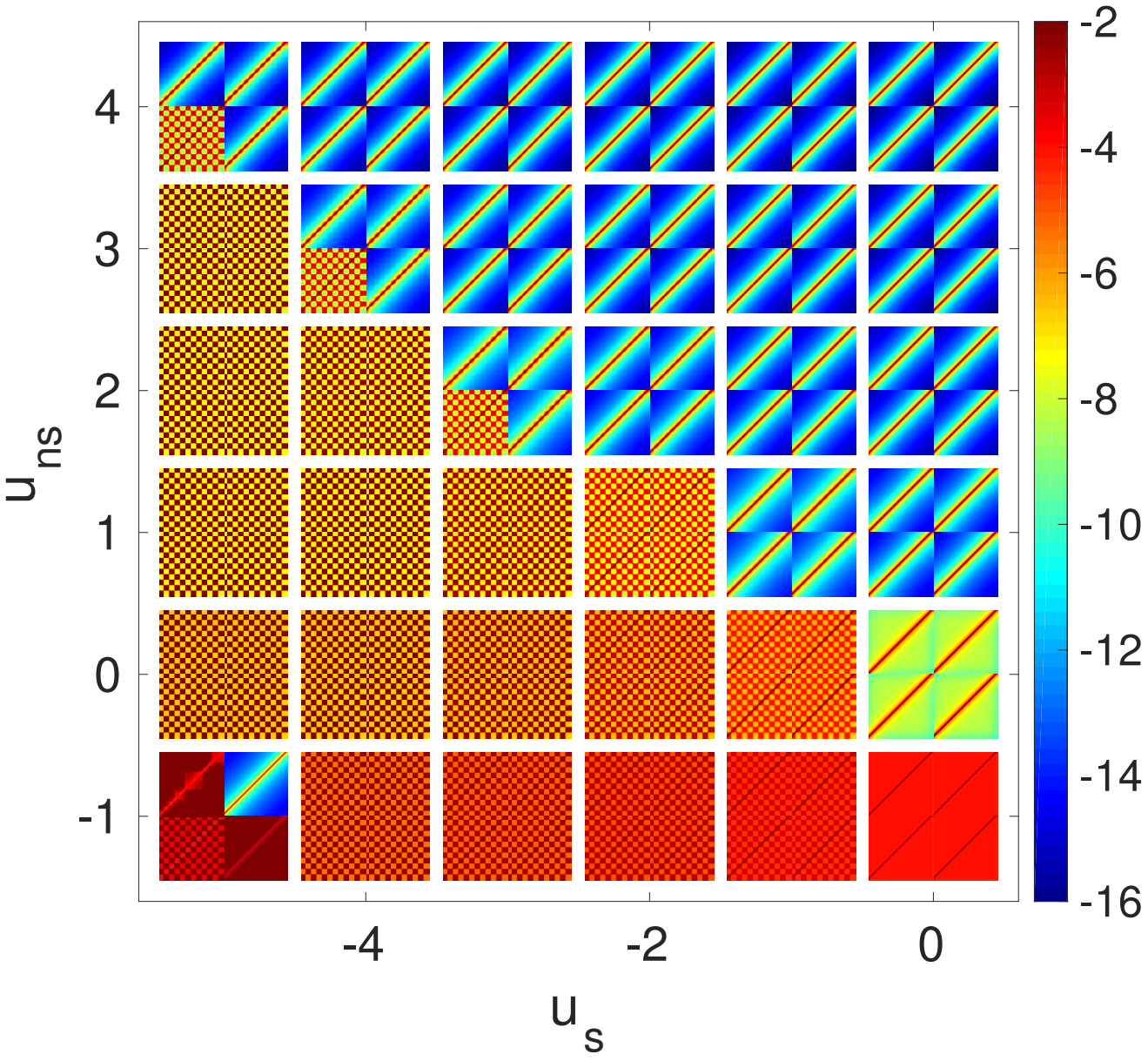

FIGURE 4.2.23 – **Cartes de contact à l'état stationnaire du copolymère $(A_{10}B_{10})_6$ associé à l'hamiltonien H_{2A} en fonction de l'intensité des interactions spécifiques u_s et non spécifiques u_{ns}.** Légende identique à celle de la figure 4.2.22.

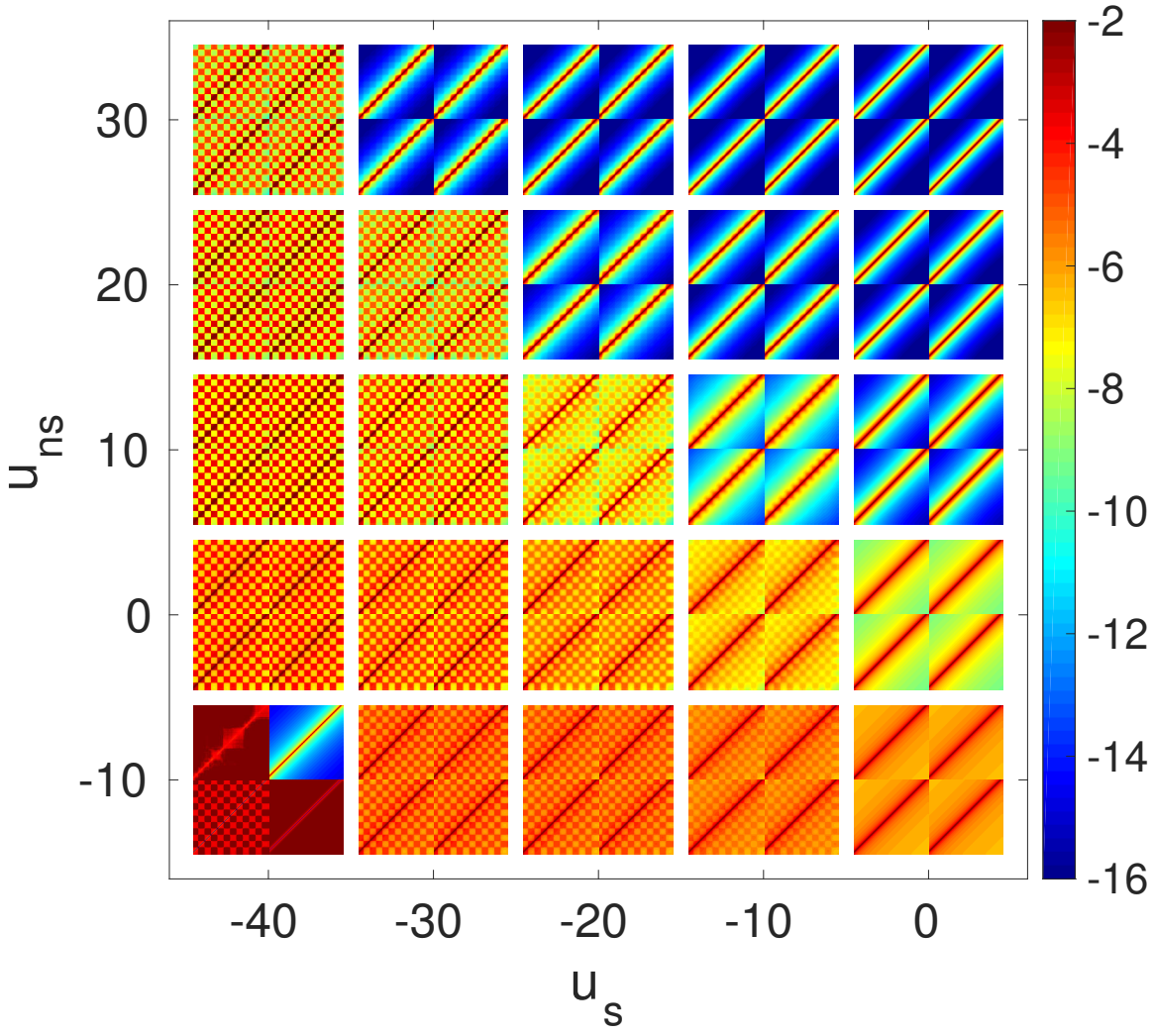

Figure 4.2.24 – **Cartes de contact à l'état stationnaire du copolymère $(A_{10}B_{10})_6$ associé à l'hamiltonien H_{1B} en fonction de l'intensité des interactions spécifiques u_s et non spécifiques u_{ns}.** Légende identique à celle de la figure 4.2.22.

u_{ns} va provoquer une transition θ de pelote vers globule. Si le confinement et la spécificité sont élevés, les monomères de même type se regroupent, une séparation en microphase apparaît. Enfin, dans la région de faible compaction et forte spécificité (en haut à gauche des figures 4.2.22 et 4.2.23), c'est-à-dire entre la séparation en microphase et la région pelote, il existe une région intermédiaire à l'intérieur de laquelle se trouve des configurations intermédiaires metastables entre polymère replié en pelote et séparé en microphases (Fig. 4.2.26). On peut y trouver par exemple des conformations en collier de perle où les domaines épigénomiques forment des TADs isolés les uns des autres.

À l'intérieur de cette région intermédiaire, il existe une zone où l'équation 4.18 peut présenter plusieurs points fixes, zone que l'on appelle « zone de multistabilité ». Des simulations

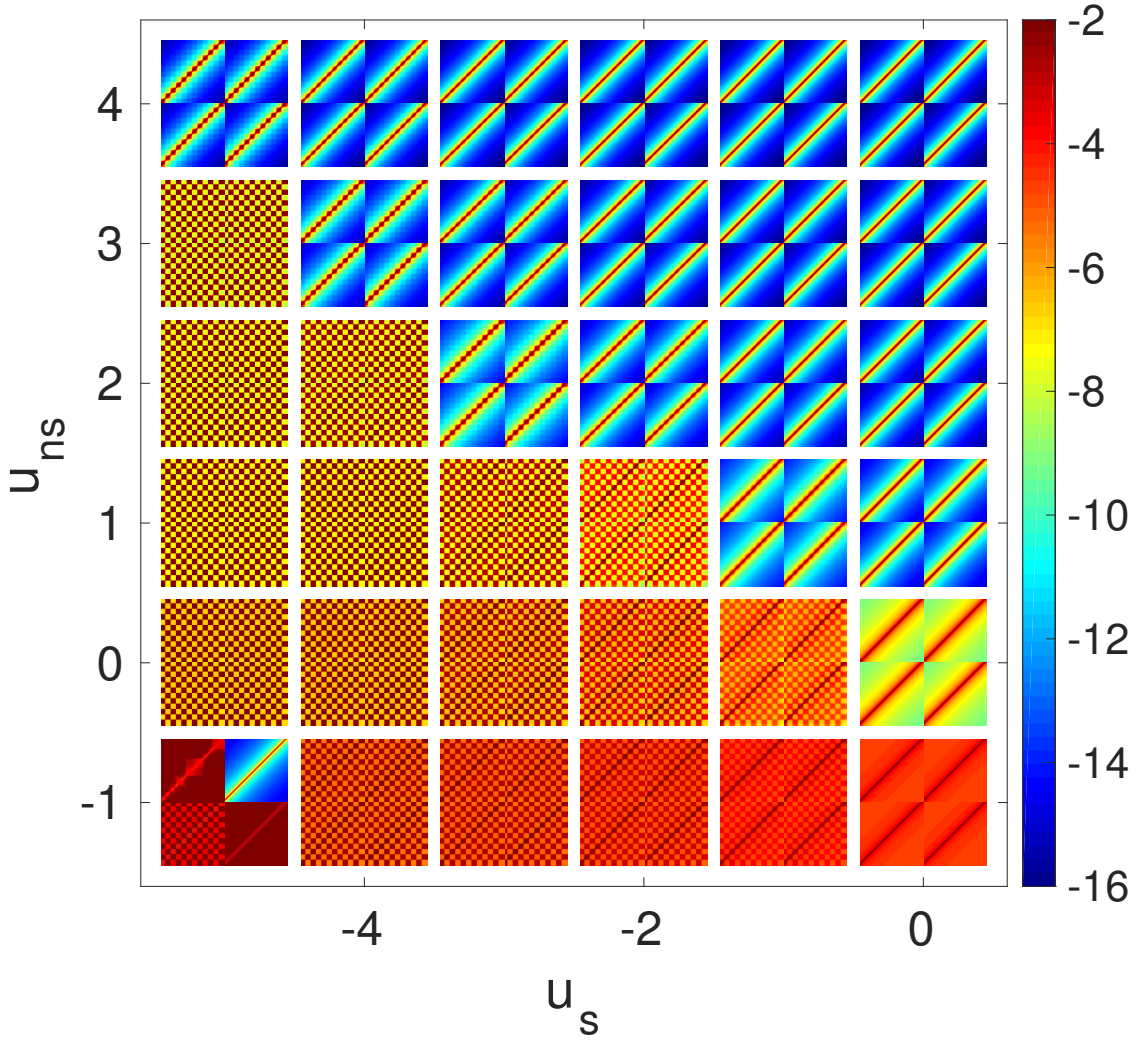

FIGURE 4.2.25 – **Cartes de contact à l'état stationnaire du copolymère $(A_{10}B_{10})_6$ associé à l'hamiltonien H_{2B} en fonction de l'intensité des interactions spécifiques u_s et non spécifiques u_{ns}.** Légende identique à celle de la figure 4.2.22.

numériques basées sur le même modèle montrent que la dynamique d'une telle chaîne est composée de sauts stochastiques entre plusieurs familles de conformations metastables [Jost et al., 2014]. Cette existence d'états multistables est cohérente avec de récentes expériences de HiC à cellule unique et avec des images par microscopie à haute résolution montrant que les TADs sont conservés entre cellules mais que ce n'est pas le cas des interactions à longue portée [Nagano et al., 2013].

Nous présenterons dans les paragraphes qui suivent les critères permettant d'identifier les frontières entre les différentes phases.

A. Transition pelote / globule

La transition θ a été étudiée dans la sous section précédente. On peut la repérer en identifiant

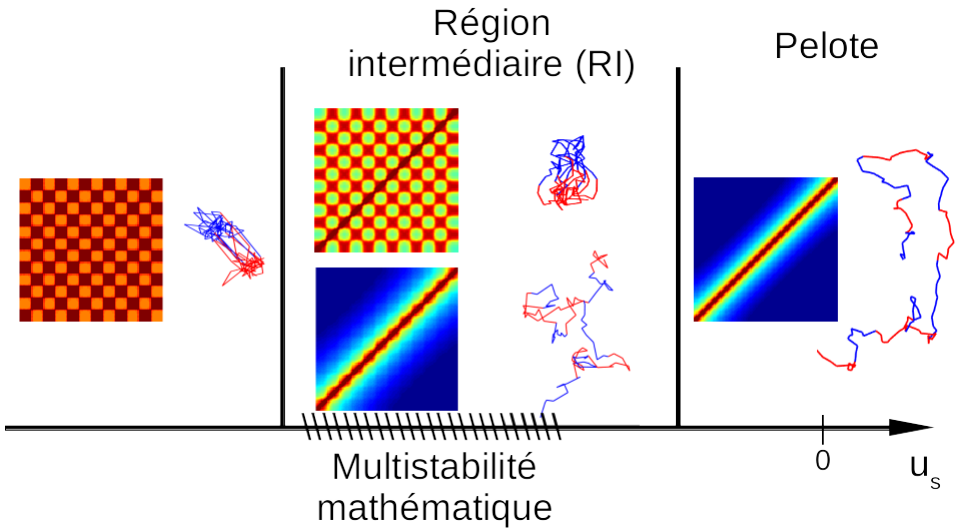

Figure 4.2.26 – **Phases du copolymère en fonction de l'intensité de la spécificité.** À faible spécificité (à droite), la chaîne se replie sous forme de pelote. A forte spécificité (à gauche), les monomères de même type se regroupent, une séparation en microphase apparaît. Entre ces deux phases, il existe une région intermédiaire comprenant des configurations metastables, intermédiaires entre pelote et séparation en microphase. On peut voir par exemple des conformations en collier de perle où les domaines épigénomiques forment des TADs isolés les uns des autres. À l'intérieur de cette région intermédiaire, il existe une zone de multistabilité mathématique, c'est-à-dire une zone où l'équation 4.18 présente plusieurs points fixes.

le point d'inflexion sur la courbe donnant l'évolution du rayon de gyration en fonction de l'intensité des interactions, u_{ns} (Fig. 4.2.19).

B. Transition vers la microphase

Afin de caractériser la transition vers la séparation en microphase qui peut se produire à partir d'une pelote, d'un globule ou bien à partir de configurations appartenant à la région intermédiaire, on définit le rayon de gyration intra TAD moyen R_g^{TAD} :

$$R_g^{TAD} = \sqrt{\frac{1}{2\left(N_A \times n_A^2 + N_B \times n_B^2\right)} \sum_{i,j} D_{ij}\delta_{b_i b_j}}$$

avec N_A (resp. N_B) le nombre de blocs A (resp. B), n_A (resp. n_B) le nombre de monomère dans un bloc A (resp. B), avec $\delta_{b_i b_j} = 1$ si les monomères i et j sont dans le même bloc et 0 sinon et avec $\delta_{e_i e_j} = 1$ si les monomères i et j sont dans le même état (A ou B) et 0 sinon. . On définit alors le ratio, r^{TAD} :

$$r^{TAD} = \frac{R_g^{TAD}}{R_g^{TAD\ aléatoire}}$$

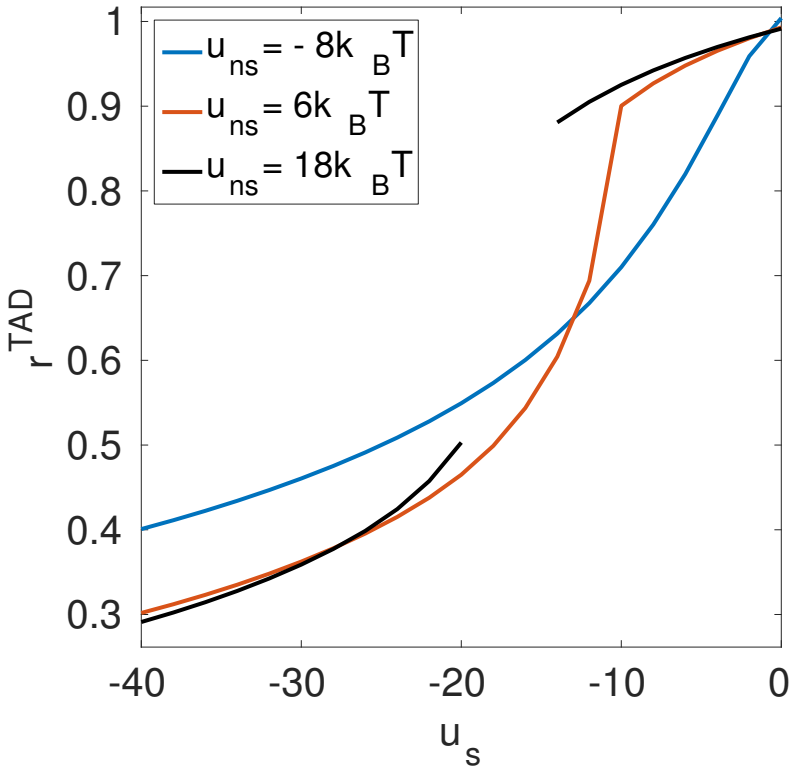

FIGURE 4.2.27 – **Évolution du ratio r^{TAD} en fonction de la spécificité, u_s pour trois valeurs d'intensité non spécifique, u_{ns}.** Pour des valeurs de u_s nulles, les TADs ne sont pas présents donc le ratio r^{TAD} est proche de 1. Si on augmente la spécificité (lecture des abscisses de la droite vers la gauche), les monomères de même type se regroupent, le ratio r^{TAD} diminue. Il présente un point d'inflexion au moment de la séparation en microphase. On voit que si les interactions non spécifiques sont attractives (en bleu), le point d'inflexion caractérisant l'apparition de la séparation en microphase arrive très tôt, (pour $u_s = -0.4 k_B T$). Si au contraire, les interactions non spécifiques sont très répulsives (en noir), le calcul de r^{TAD} n'est pas possible dans la région où l'inflexion se situerait en raison de la détection de plusieurs états stationnaires. Enfin pour des interactions non spécifiques intermédiaires, par exemple $u_{ns} = 6 k_B T$ (en rouge) on observe clairement le point d'inflexion pour $u_s = -11 k_B T$.

avec $R_g^{TAD\ aléatoire}$ le rayon de gyration intra TAD moyen si les TADs sont placés aléatoirement le long de la diagonale. Ce ratio r^{TAD} est environ égale à 1 pour les phases pelote et globule. Si on augmente la spécificité, les monomères de même type se regroupent, le ratio r^{TAD} diminue. Il présente un point d'inflexion au moment de la séparation en microphase (Fig. 4.2.27). Le repérage systématique de ce point d'inflexion va permettre de délimiter la région de séparation en microphase.

C. Transition pelote / « Pelote avec TADs et interactions à longue portée »

Pour identifier la frontière entre la région intermédiaire et la zone de pelote, on introduit le paramètre σ^{TAD} qui est la variance du nombre de contacts normalisé à l'intérieur d'un

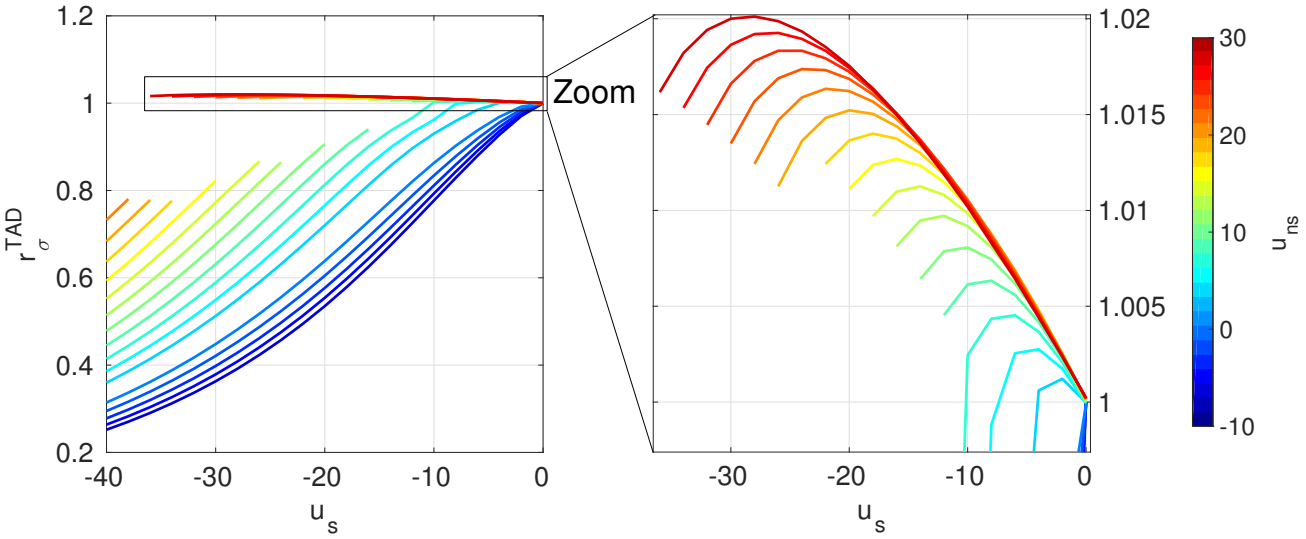

Figure 4.2.28 – **Évolution de r_σ^{TAD} en fonction de la spécificité, u_s pour plusieurs valeurs d'intensité non spécifique, u_{ns} définies par l'échelle de couleur à droite du graphe.** Lors du passage d'une pelote vers la région intermédiaire, le ratio r_σ^{TAD} passe par un maximum.

supposé TAD (idem à la définition de σ donné lors de la description de IC-Finder, Éq. équation (2.10)). On rappelle que le nombre de contacts normalisé, $C_n(i,j)$ entre deux monomères i et j est donné par la formule ci-dessous :

$$C_n(i,j) = \frac{C(i,j)}{\bar{C}(|j-i|)}$$

avec $\bar{C}(|j-i|) = \frac{\sum_{|j-i|=k} C(i,j)}{\sum_{|j-i|=k} 1}$.

À partir de σ_{TAD} , on définit le ratio $r_\sigma^{TAD} = \frac{\sigma^{TAD}}{\sigma^{TAD\ aléatoire}}$. Son évolution en fonction de u_s est représentée pour plusieurs valeurs d'intensité non spécifique, u_{ns} sur la figure 4.2.28. Pour des valeurs de u_s nulles (à droite du graphe), les TADs ne sont pas présents donc le ratio est proche de 1. Si on augmente la spécificité (lecture des abscisses de la droite vers la gauche), les monomères de même type cherchent à se regrouper, en conséquence, la variance dans un TAD augmente et est même supérieure à celle d'un compartiment choisi aléatroirement le long de la diagonale. Ceci implique que r_σ^{TAD} augmente dans un premier temps. Une fois les TADs mis en place, le nombre de contacts dans un TAD devient de plus en plus uniforme, r_σ^{TADs} diminue. Ce paramètre passe donc par un maximum caractéristique de l'apparition des TADs. Ce comportement permet de localiser la transition entre pelote et pelote intermédiaire (avec TADs et éventuellement interactions à longue portée).
(Fig. 4.2.28).

D. Transition vers une région de multistabilité mathématique

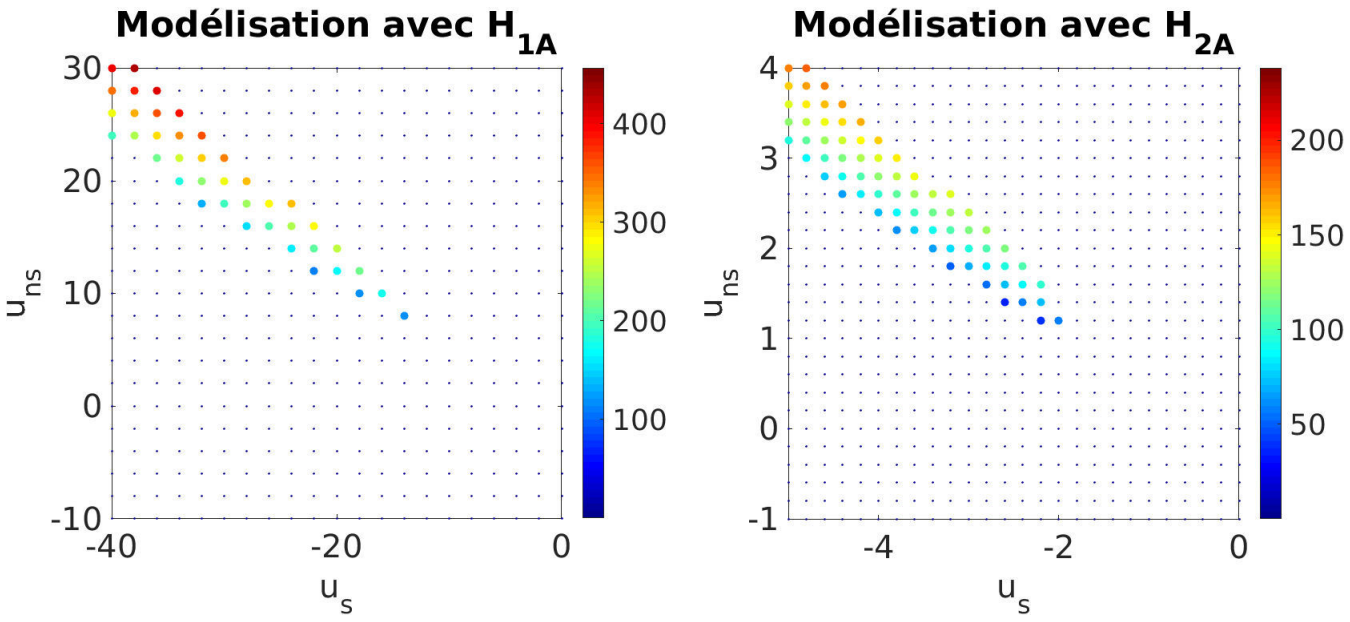

FIGURE 4.2.29 – **Valeurs de ϵ_4 , (Éq. équation (4.46)) obtenues à partir d'une grille de paramètres (u_s, u_{ns}) pour un hétéropolymère $(A_{10}B_{10})_6$ modélisé avec les hamiltoniens H_{1A} (à gauche) et H_{2A} (à droite).** La couleur de chaque point placé sur cette grille représente la valeur de ϵ_4 selon l'échelle de couleur indiquée à droite de chaque carte. Le point est gros si $\epsilon_4 > 10^{-3}$ et petit sinon. Dans le cas où $\epsilon_4 > 10^{-3}$, le phénomène de multistabilité mathématique est présent (critère (4.46)). Ainsi l'espace occupé par les gros points, c'est-à-dire ceux dont la couleur est autre que bleu foncé, définit la région de multistabilité.

Nous avons défini dans la partie 4.2.5.1 un critère permettant de détecter la multistabilité. Il s'agit de calculer l'écart maximal entre les cartes de contact obtenues à partir de conditions initiales différentes (Éq. 4.46). La figure 4.2.29 présente les valeurs de cet écart maximal, ϵ_4, pour une grille de paramètres (u_s, u_{ns}). On peut voir que les couples (u_s, u_{ns}) tels que ϵ_4 est supérieur à 10^{-3} ne sont pas distribués de manière aléatoire mais sont au contraire concentrés dans une région particulière. Quel que soit l'hamiltonien du système, cette région est située dans la même zone de forte spécificité / faible compaction. Il est important de préciser ici que la région de multistabilité trouvée est inclus dans la zone de multistabilité totale (et au mieux égale). En effet, nous n'avons résolue l'équation 4.18 qu'avec quatre conditions initiales, nous ne sommes donc pas en mesure de dire si d'autres conditions initiales auraient permis de détecter de la multistabilité ailleurs ou pas.

4.2.5.3.2 Région telle que $P_c \sim s^{-1}$

Étant donné que notre but est la modélisation de la chromatine, on souhaite reproduire la loi d'échelle $P_c \sim s^{-1}$ observée expérimentalement dans le cas des longs chromosomes. Grâce à l'introduction du champ générique, cette loi est vérifiée pour le polymère $(A_{10}B_{10})_6$ sans interaction ($u_{ns} = 0$ et $u_s = 0$). Si on introduit des interactions spécifiques, la probabilité de contact P_c dévie du champ générique. Dans ce paragraphe, on se demande si pour un paramètre u_s donné, il est possible d'agir sur le paramètre u_{ns} de telle sorte à s'approcher de la loi générique $P_c \sim s^{-1}$. Pour répondre à cette question, on ajuste la probabilité de contact moyenne en fonction de la distance génomique par une droite en échelle log pour les grille de couples (u_s, u_{ns}) précédemment définies et avec la méthode des moindres carrés. Notons que dans le cas de la modélisation avec les hamiltoniens H_{XA} (contrairement à H_{XB}), on exclut de la procédure d'ajustement les données telles que $s < 5$ car on sait que le champ générique inféré avec un unique paramètre ne permet pas d'imposer la loi $P_c \sim s^{-1}$ (Fig. 4.2.5). On représente sur la figure 4.2.30 les valeurs, α des pentes trouvées.

A. Propriétés communes à tout les hamiltoniens (Tab. 4.2)

La figure 4.2.30 nous confirme qu'effectivement, pour une interaction spécifique, u_s donnée, le paramètre u_{ns} permet d'agir sur le coefficient α de la loi $P_c \sim s^{\alpha}$. Cette puissance évolue environ de -2 (cas d'une chaîne « gonflée ») à 0 (globule). Il semble que l'évolution de α soit continue. On peut voir que si la spécificité est faible (proche de 0), il existe une région dans les quatre cas pour laquelle α est proche de -1. Si la spécificité est forte, le phénomène de multistabilité mathématique apparaît rendant impossible le calcul d'une pente puisque l'on ne connaît pas le nombre exact d'états stationnaires. On peut imaginer que dans cette région de multistabilité, l'état d'équilibre soit caractérisé par une loi de type $P_c \sim s^{-1}$.

B. Comparaison des résultats en fonction de l'inférence du champ générique

Lorsque le champ générique est inféré avec $\frac{N(N+1)}{2}$ paramètres plutôt qu'avec un seul, les variations du coefficient α de la loi $P_c \sim s^{\alpha}$ sont plus douces. Avec les hamiltoniens du type H_{XB} il est donc plus aisé de s'approcher de la loi d'échelle $P_c \sim s^{-1}$ observée expérimentalement. Ceci est cohérent avec les conclusions que nous avions tirées à partir de l'étude de l'homopolymère : les variations du coefficient α en fonction de u_s et de u_{ns} sont plus douces avec les hamiltoniens type H_{XB} par rapport à H_{XA} .

4.2.5.3.3 Diagrammes de phase

Maintenant que nous avons défini les critères permettant de caractériser toutes les transitions de phase observées sur les figures introductives 4.2.22, 4.2.23 4.2.24 et 4.2.25, on peut représenter précisément les diagrammes de phases du copolymère $(A_{10}B_{10})_6$ obtenues avec les différents hamiltoniens (Fig. 4.2.31). Les régions où de la multistabilité mathématique a

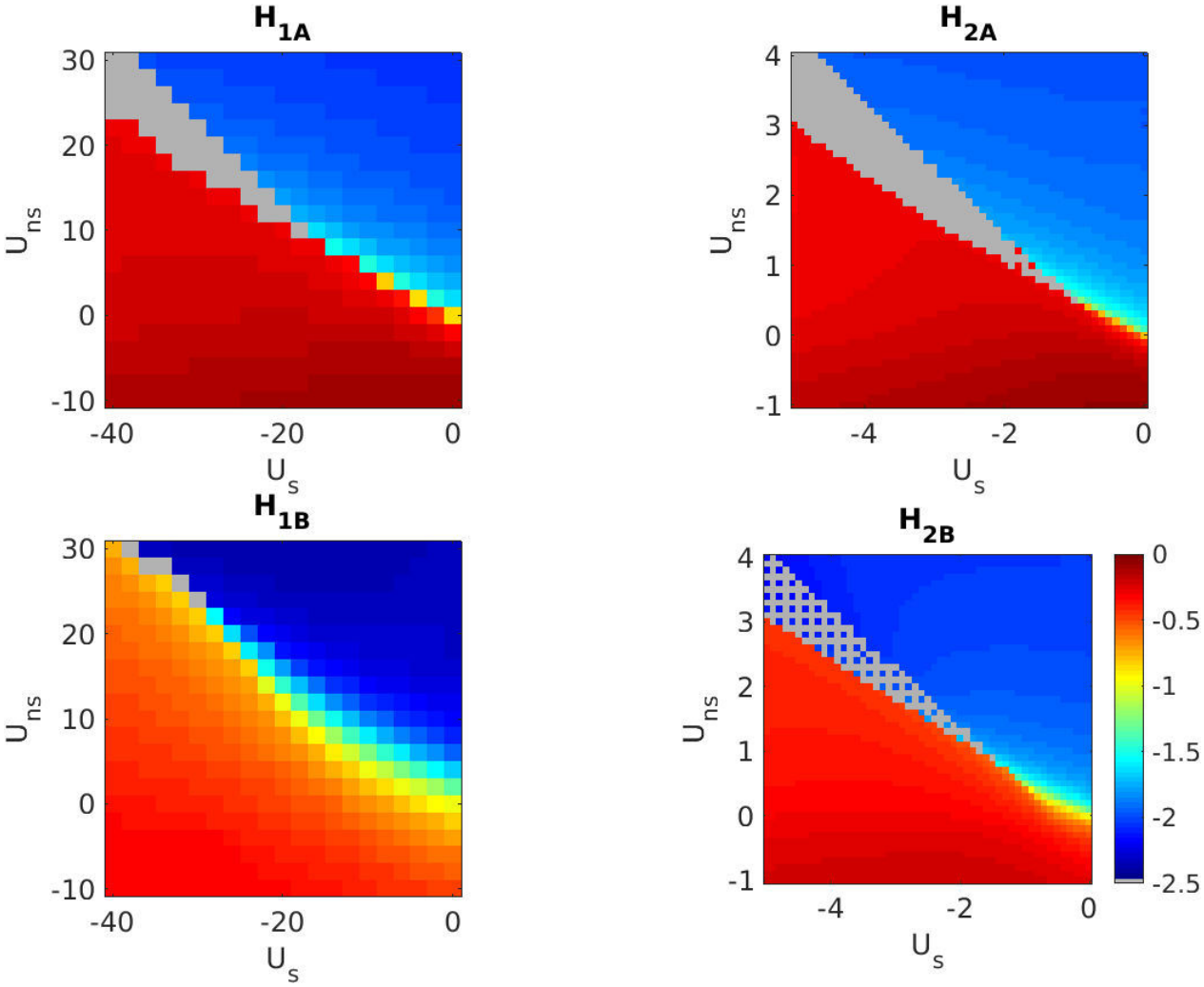

FIGURE 4.2.30 – **Pentes α obtenues après ajustement par une droite en échelle log de la probabilité de contact moyenne en fonction de la distance génomique ($P_c \sim s^{\alpha}$).** Les calculs ont été réalisés pour l'hétéropolymère $(A_{10}B_{10})_6$ modélisé avec les quatre hamiltoniens résumés dans le tableau4.2. L'échelle de couleur donnant la valeur de la pente (à droite de la dernière carte) est commune à toutes les images. Une case grisée signifie que l'ajustement n'a pas pu être réalisé en raison du phénomène de multistabilité.

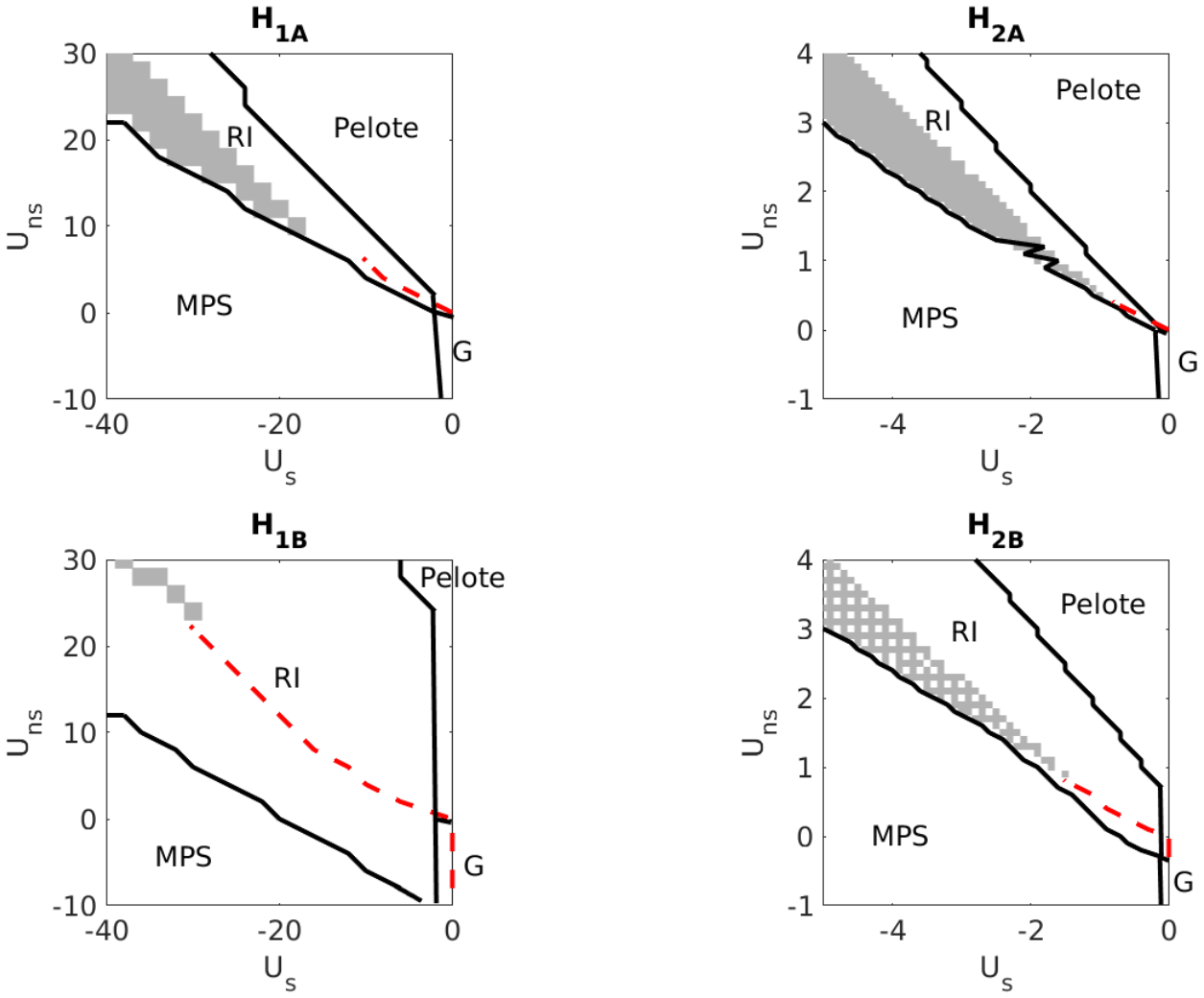

Figure 4.2.31 – **Diagramme de phase du copolymère par bloc $(A_{10}B_{10})_6$ en fonction des interactions spécifiques, u_s et non-spécifiques u_{ns} entre monomères (en unité k_BT).** La modélisation est effectuée avec les quatre hamiltoniens résumés dans le tableau 4.2. Les abréviations « G », « MPS » et « RI » signifient respectivement globule, séparation en microphase et globule. Les régions grisées signalent la présence de multistabilité mathématique. La ligne rouge met en evidence la zone où le coefficient α de la loi $P_c \sim s^{\alpha}$ est le plus proche de -1.

été détectée sont grises. Bien sûr, comme nous avons résolu l'équation 4.18 avec un nombre de conditions initiales fini, l'ensemble formé par ces régions grises est inclus (ou au mieux égal) à la zone totale de multistabilité mathématique. De plus, sur les diagrammes de phase, on signale, pour chaque paramètre u_s, le point tel que le coefficient α de la loi $P_c \sim s^{\alpha}$ est le plus proche de -1, avec la contrainte que $\alpha \in [-1.2, -0.8]$. Ces points forment une ligne représentée en tirets rouge. Cette ligne ajouté au diagramme de phase permet de situer les configurations reproduisant au mieux la loi $P_c \sim s^{-1}$ observée expérimentalement.

A. Propriétés communes à toutes les hamiltoniens (Tab. 4.2)

On peut observer que malgré la simplicité du copolymère $(A_{10}B_{10})_6$ qui n'est construit qu'à partir de deux états (A et B), les quatre diagrammes de phase obtenus sont riches avec l'existence de quatre régions distinctes selon les intensités d'interaction : une région pelote, une région globule (« G »), une région de séparation en microphases (« MPS ») et une région intermédiaire entre pelote et MPS (« RI »). Cette région est particulièrement intéressante car son existence démontre qu'un hétéropolymère composé de deux types de bloc permet de reproduire des cartes de contacts avec des TADs et des domaines d'interaction à longue portée tels que ceux observés expérimentalement avec la technique de Hi-C.

On peut aussi voir sur les diagrammes que les couples (u_s, u_{ns}) tels que le phénomène de multistabilité est détecté ne sont pas distribués de manière aléatoire mais sont au contraire concentrés dans une région particulière. Quel que soit l'hamiltonien du système, cette région est caractérisée par une forte spécificité / faible compaction et est inclus dans la région intermédiaire. Précédemment, il a été montré qu'une multistabilité mathématique implique l'existence de conformations multistables pour lesquelles les TADs de même état épigénomique interagissent de façon dynamique les uns avec les autres [Jost et al., 2014 ; Olarte-Plata et al., 2016].

B. Comparaison des résultats en fonction de la modélisation du volume exclu

La différence dans le choix du potentiel de volume exclu se traduit principalement par des ordres de grandeurs d'énergies mises en jeu différents. La modélisation du volume exclu par un potentiel de type Lennard-Jones tronqué aboutit à des résultats qualitatifs en accord avec les expériences mais à des résultats quantitatifs non réalistes : les potentiels d'interaction spécifiques dans la région qui reproduit au mieux les données expérimentales sont de l'ordre de $30k_BT$. Avec le potentiel gaussien, ces ordres de grandeur sont environ dix fois plus petits et donc plus réalistes.

C. Comparaison des résultats en fonction de l'inférence du champ générique

De manière assez surprenante, en modélisant le copolymère $(A_{10}B_{10})_6$ avec les hamiltoniens H_{1B} et H_{2B} - qui ont pour particularité de contenir un terme $H_{générique}$ dont les $\frac{N(N+1)}{2}$ paramètres ont été inférés - la résolution de l'équation 4.18 avec les quatre conditions initiales présentées dans la partie 4.2.5.1 n'a pas mené à la détermination d'une région de multistabilité aussi large et aussi bien définie que celle obtenue avec les hamiltoniens H_{1A} et H_{2A}. En effet, pour une raison mathématique obscure à nos yeux, nous avons observé avec les hamiltoniens H_{XB} de la multistabilité de façon clairsemée (Fig. 4.2.31).

Par ailleurs, on remarque que la modélisation avec H_{XB} par rapport à celle avec H_{XA}, laisse apparaître une région intermédiaire bien plus large ce qui montre que la transition de pelote à microphase est dans ce cas plus lente et continue.

4.2.5.3.4 Conclusion L'intégration de l'équation 4.18 dans le cas du copolymère $(A_{10}B_{10})_6$ avec les quatre hamiltoniens H_{1A}, H_{2A}, H_{1B} et H_{2B} a mené à des résultats qualitativement identiques et forts prometteurs puisque l'on n'observe que pour certains couples d'interaction (u_s, u_{ns}) les matrices de contact obtenues in silico ressemblent fortement aux cartes expérimentales. En effet, si le confinement est faible et la spécificité forte, on retrouve des cartes de contact avec des TADs le long de la diagonale ainsi que des domaines domaines d'interactions à longue portée. Toutefois, pour obtenir de telles cartes, avec les hamiltoniens H_{1A} et H_{1B}, il a fallu utiliser des paramètres d'interaction de l'ordre de 30 à $40k_BT$. Ces ordres de grandeurs sont trop élevés pour pouvoir représenter les interactions en jeu dans le cas de la chromatine. Ceci nous amène donc à la conclusion que la définition du terme H_{ve1} (Éq. 4.24) à conduit à une surestimation du volume exclu. De plus, nous avons vu qu'avec les hamiltoniens de type $H_{\chi B}$, il est plus facile de reproduire la loi $P_c \sim s^{-1}$. De ce fait, on se propose de poursuivre dans la partie suivante avec uniquement l'hamiltonien H_{2B}.

4.2.5.4 Application à la drosophile : Heteropolymère dont les blocs sont définis à partir de l'épigénome de la drosophile

Ici, on étudie le comportement d'un segment chromosomique de la drosophile constitué successivement de domaines épigénomiques , petits et actifs, et, grands et inactifs. La région à laquelle on s'intéresse se trouve entre 23.06 à 24.36Mb du chromosome 3R, elle est composée d'une alternance entre domaines actifs et inactifs [Filion et al., 2010]. Le diagramme de phase obtenue à partir de cette séquence épigénomique révèle l'existence de conditions telle que les cartes de contact simulées ressemblent fortement à celles obtenues expérimentalement. Ces conditions se situent dans la région intermédiaire (Fig. 4.2.32 et 4.2.33). En particulier, le repliement de cette région chromosomique est bien modélisé en choisissant comme paramètre $u_s = -2.6 k_B T$ et $u_{ns} = 2 k_B T$ (Fig. 4.2.34 partie triangulaire supérieure). Ces valeurs ont été choisies en minimisant la différence quadratique entre carte de contact expérimentale et simulée.
On peut noter que l'aspect dynamique semble bien reproduit avec notre approche gaussienne auto-cohérente puisque la résolution de l'équation de la dynamique 4.18 avec l'algorithme de Runge-Kutta révèle que la naissance des TADs est rapide et qu'elle précède la formation des interactions longues portées ce qui a été vérifié expérimentalement [Naumova et al., 2013]. En effet, on peut voir sur la figure 4.2.35 qu'au bout d'une cinquantaine de pas de temps les TADs sont formés ; par contre, pour que toutes les interactions à longues portées soient mises en place, 800 unités de temps sont nécessaires. Autrement dit, les interactions à longues portées mettent 16 fois plus de temps à se former par rapport aux TADs.

4.2.6 Conclusion

La modélisation de la chromatine par un copolymère par bloc dans le cadre de l'approche gaussienne auto-cohérente a permis de reproduire des cartes de contact fortement ressemblantes à celles obtenues expérimentalement avec la technique de Hi-C. Nous avons travaillé avec des petits segments équilibrés tout en intégrant les effets de crumpling mimant l'environnement et/ou le caractère hors équilibre. Les deux approches suivantes basées sur des simulations de dynamique vont permettre de confirmer les résultats obtenus ici et vont donc valider l'approximation gaussienne auto-cohérente.

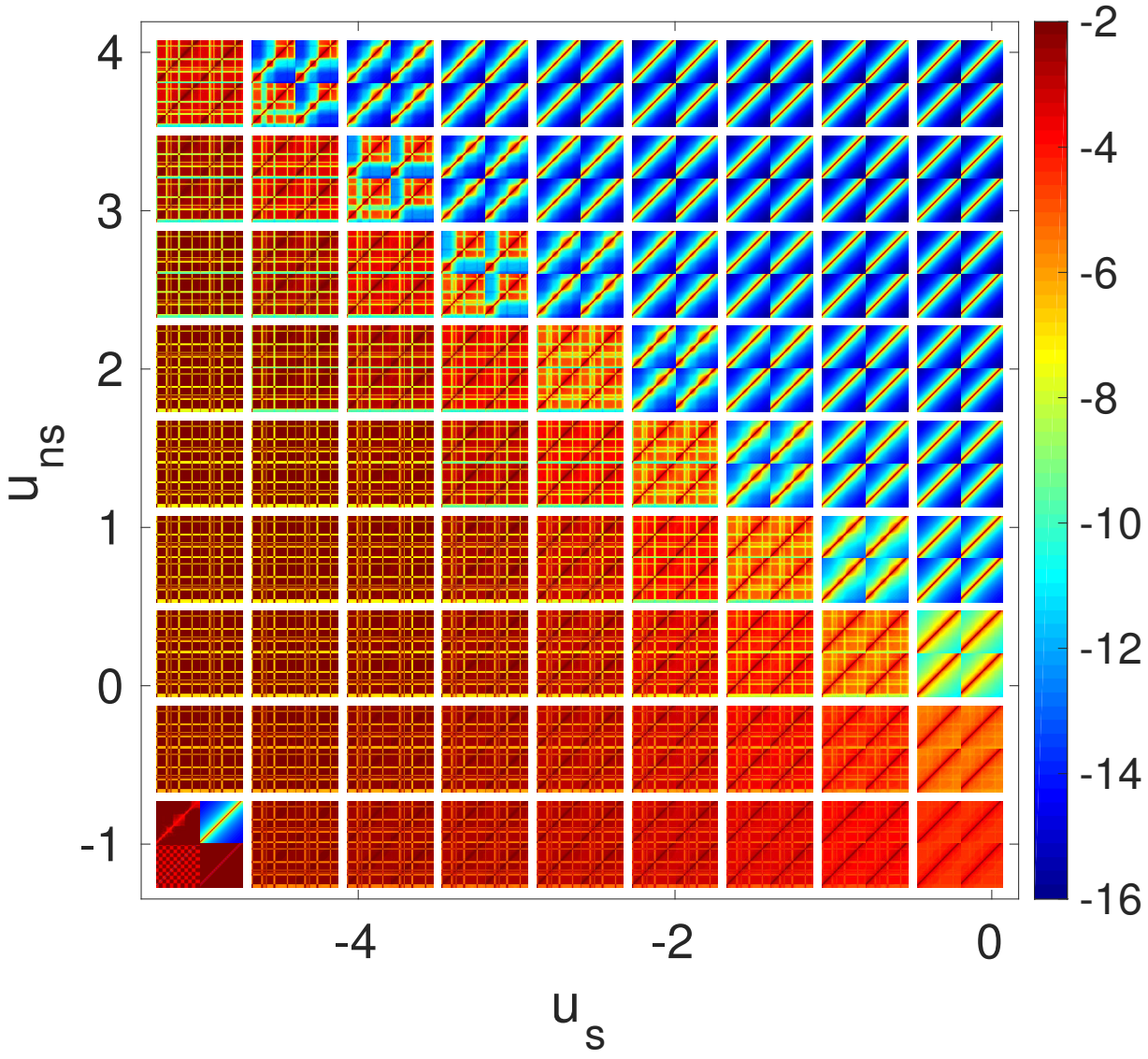

Figure 4.2.32 – **Cartes de contact à l'état stationnaire de l'hétéropolymère dont les blocs sont définis à partir de l'épigénome de la drosophile (région entre 23.06Mb et 24.36Mb du chromosome 3R) associé à l'hamiltonien H_{2B} en fonction de l'intensité des interactions spécifiques u_s et non spécifiques u_{ns}.** Ces prédictions sont obtenues dans le cadre de l'approximation gaussienne auto-cohérente. La résolution de l'équation 4.18 est réalisée à partir de quatre états initiaux différents dont les cartes de contact sont représentées dans le coin en bas à gauche. Il existe une région où selon la condition initiale, plusieurs états stationnaires sont possibles.

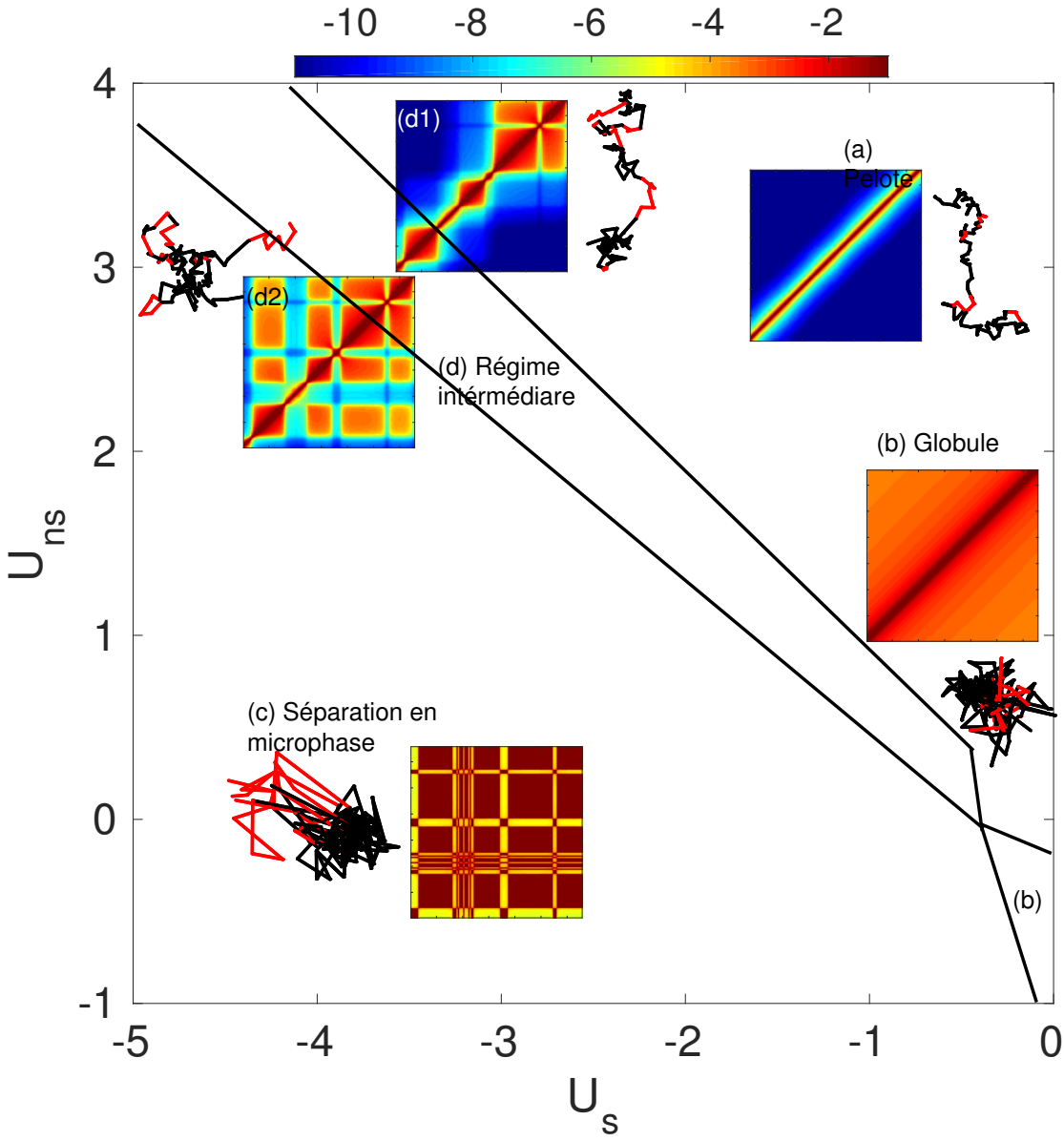

FIGURE 4.2.33 – **Diagramme de phase de l'hétéropolymère dont les blocs sont définis à partir de l'épigénome de la drosophile (région entre 23.06Mb et 24.36Mb du chromosome 3R) associé à l'hamiltonien H_{2B} en fonction de l'intensité des interactions spécifiques u_s et non spécifiques u_{ns} (en unité k_BT).** Les inserts représentent des structures et des cartes de contact typiques (en échelle log donnée au dessus du diagramme de phase) pour chacune des phases.

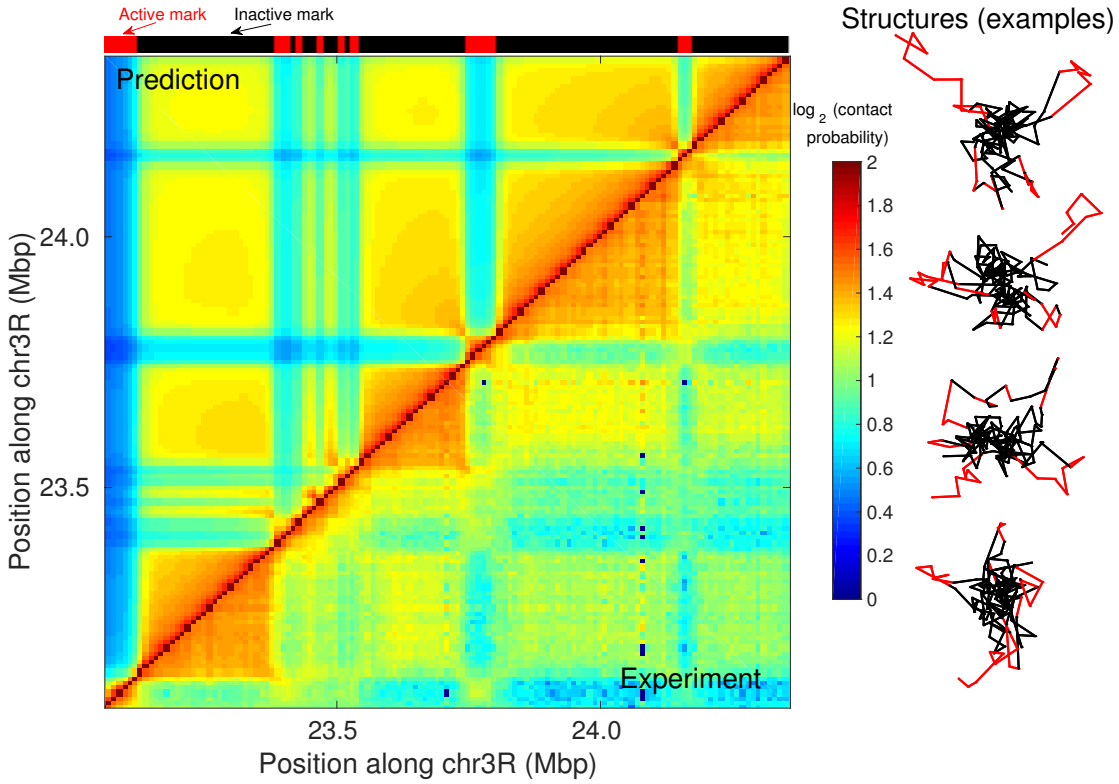

Figure 4.2.34 – **Cartes de contact expérimentale en bas [Sexton et al., 2012] et obtenue avec l'approximation gaussienne auto-cohérente en haut.** Ce segment chromosomique est simulé avec l'hamiltonien H_{2B}. La séquence épigénomique utilisée est indiquée au dessus de la carte (noir=inactif, rouge=actif) [Filion et al., 2010]. Les paramètres utilisés sont $u_s = -2.6 k_B T$ et $u_{ns} = 2 k_B T$.

4.3 Dynamique sur réseau

4.3.1 Motivation

L'étude de la chromatine avec l'approximation gaussienne auto-cohérente nous a permis de montrer que la modélisation de la chromatine par un copolymère par bloc permet de reproduire des propriétés à petites échelles observées expérimentalement. On se propose donc dans cette partie de poursuivre l'étude du repliement de la chromatine avec le modèle de copolymère par bloc mais en abandonnant l'approximation gaussienne auto-cohérente au profit de simulations de dynamique sur réseau. La dynamique sur réseau permet d'obtenir la dynamique des trajectoires des monomères de façon individuelle, données auxquelles nous n'avions pas accès avec l'approximation gaussienne auto-cohérente. On va donc pouvoir étudier les propriétés dynamiques et structurelles de la chromatine ce qui va notamment nous permettre de caractériser la multistabilité qu'on a discutée avec l'approche gaussienne auto-cohérente. Réaliser des simulations de dynamique sur réseau se révèle donc être pertinent

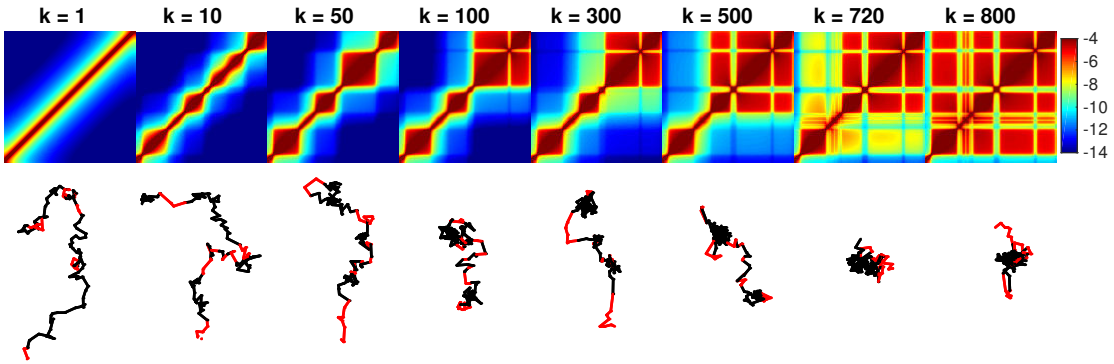

FIGURE 4.2.35 – **Dynamique de repliement d'une pelote (homopolymère de 131 monomères associé à l'hamiltonien H_{2B} de paramètre $u_{ns} = -0.5k_BT$) en hétéropolymère dont les blocs sont définis à partir de l'épigénome de la drosophile (région entre 23.06Mb et 24.36Mb du chromosome 3R) associé à l'hamiltonien H_{2B} ($u_s = -2.6k_BT$ et $u_{ns} = 2k_BT$).** L'équation de la dynamique est intégrée avec l'algorithme de Runge-Kutta. On représente les cartes de contact pour les itérations 1 (état initial), 5, 10, 50, 100, 300, 500, 720 et 800 (état stationnaire). L'échelle de couleur de ces cartes est donnée à droite de la figure. On accompagne chaque carte d'une structure représentative générée dans le cadre de l'approximation gaussienne.

d'une part pour valider ou non l'approche gaussienne auto-cohérente et éventuellement en fixer les limites et d'autre part pour accéder à de nouvelles propriétés. Dans un premier temps, on considérera une dynamique sur réseau. Il a été montré que la dynamique d'un polymère sur réseau est la même que celle d'un polymère hors réseau pour des échelles d'observation supérieures à 3 pas de réseau [Hugouvieux et al., 2008]. C'est pourquoi nous présenterons dans cette section l'étude du copolymère par bloc via des simulations de Monte-Carlo cinétique par chaînes de Markov sur réseau. On présentera ici des résultats obtenus sur des petits segments chromosomiques, de l'ordre de 1Mb, (comme dans l'approche gaussienne auto-cohérente) mais l'algorithme a vocation à être utilisé sur des plus grandes échelles. Ces résultats ont été publiés dans Olarte-Plata et al. 2016.

4.3.2 Formalisme

La chromatine se trouve ici dans un espace discrétisé. Sa conformation sera caractérisée par son énergie totale $H = H_{chaine} + H_{interactions}$, avec :

$$H_{interactions} = \sum_{i<j} U_{e_i e_j)} \delta_{ij}$$

avec $\delta_{ij} = 1$ si les monomères i et j sont plus proches voisins sur le réseau et 0 sinon, avec e_i et e_j respectivement les états chromatiniens des monomères i et j et avec $U_{e_i e_j}$ l'intensité

des interactions entre une paire de monomères i et j plus proche voisin sur le réseau. Comme précédemment on suppose que $U_{e(i),e(j)} \neq 0$ seulement dans le cas où les monomères i et j sont dans le même état épigénomique. La chaîne est formée de N monomères et est confinée dans une boite cubique de côté $L \times b$, avec b la taille typique d'une unité du réseau. Les conditions aux bords de la boite sont périodiques. Cette hypothèse ne modifie pas les propriétés du polymère pour des échelles supérieures à une maille élémentaire. Cette façon de modéliser l'environnement dans lequel évolue la chromatine (boite avec condition aux bords périodiques) constitue la principale différence de modélisation par rapport à l'approche gaussienne auto-cohérente. En effet, avec cette dernière approche le confinement était modélisé par des interactions non spécifiques entre monomères. La fraction volumique typique occupée par la chromatine dans un noyau cellulaire est de 0.1 [Milo et al., 2010]. On fixe donc les paramètres N et L de telle sorte à ce que la densité $\rho = \frac{N}{N_{sites}} = 0.1$, avec N_{sites} le nombre total de sites dans le réseau. Une densité de chromatine typique de 4×10^{-3} pb.nm^{-3} [Milo et al., 2010] impose un diamètre de 75nm pour un monomère de 10kb. L'unité de temps de la simulation est déterminée en faisant correspondre le déplacement quadratique moyen, g_1, pour une chaine neutre (sans interaction spécifique entre monomère) à des valeurs expérimentales mesurées chez la levure [Hajjoul et al., 2013] pour lesquelles on trouve $g_1 \approx 0.01t^{0.5}$ avec g_1 en μm^2 et t en s et ce qui donne qu'un pas de simulation Monte Carlo correspond envrion à 0.3 s en temps réel. On a vu en introduction (cf. chapitre 1) que les contraintes topologiques (dont on tient compte dans la modélisation) ralentissent drastiquement la dynamique de la chaîne au point que l'équilibration pour un chromosome eucaryote de taille comprise entre 20 et 100Mb peut être bien supérieur à la durée d'un cycle cellulaire [Rosa and Everaers, 2008]. Le chromosome est donc hors équilibre. Toutefois, localement, pour des régions de chromatine de l'ordre de 1Mb, les contraintes topologiques sont à priori faibles et les segments peuvent donc être considérés comme équilibrés [Rosa and Everaers, 2014]. Étudier un segment de 1Mb environ présente donc l'avantage de travailler à l'équilibre ce qui permet d'éviter d'avoir à tenir compte d'effets de mémoire dus à une condition initiale inconnue. La validité de cette hypothèse sera testée à l'aide de simulations de dynamique moléculaire à plus grande échelle (cf. section §4.4). Enfin cette petite taille de l'ordre de 1Mb va permettre l'obtention de résultats rapides et permettre la comparaison avec les résultats obtenus avec l'approximation gaussienne auto-cohérente au cours de laquelle nous avions aussi étudié des segments d'environ 1Mb. Plus précisément, toutes les simulations de cette partie concerneront des segments de 1.2Mb, soit 120 monomères.

La dynamique de la chaîne se fait sur un réseau cubique face centrée et a pour particularité de se dérouler selon les contraintes décrites par Hugouvieux et al. (2008) (Fig. 4.3.1) et pouvant se résumer ainsi : *deux monomères peuvent occuper le même site sur le réseau si et seulement*

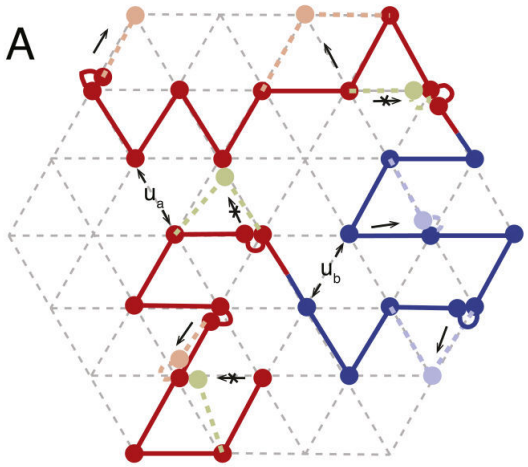

FIGURE 4.3.1 – **Projection 2D d'une section du réseau Cubique Face Centrée.** Les points représentent des monomères (actifs en rouge, inactifs en bleu). Dans ce cadre de modélisation de la chromatine par un copolymère par bloc sur réseau, chaque monomère peut se déplacer sur les sites plus proches voisins occupés par au plus un monomère. Ce déplacement ne doit pas rompre la connectivité de la chaîne. Sur la figure les mouvement en bleu clair et rouge clair sont possibles alors que les mouvement en vert clair sont interdits. Les interactions spécifiques notées U_a et U_b sur la figure se font entre monomères de même états épigénétiques plus proche voisins. [Source : figure tirée de Olarte-Plata et al. (2016) ayant été redessinée à partir de celle faite par Hugouvieux et al. (2008)].

si ils sont consécutifs le long de la chaîne. Les contraintes découlant de ce principe assurent la connectivité du polymère, définissent le volume exclu et empêchent l'entrecroisement de brins de chromatine. Ceci permet de simuler efficacement la dynamique de reptation des systèmes denses. À chaque essai de mouvement, un monomère et un déplacement sont choisis aléatoirement. Si ce mouvement ne rompt pas la connectivité de la chaîne alors le choix d'accepter ou pas le mouvement reposera sur un schéma de Metropolis standard pour les énergies d'interaction.

4.3.3 Résultats

4.3.3.1 Propriétés génériques d'une chaîne isolée auto-évitante sans interactions

Afin de valider l'algorithme développé, on commence par vérifier certaines lois de puissance concernant les homopolymères auto-évitant isolé et sans interaction entre monomères. On peut voir sur la figure 4.3.2 que le rayon bout-à-bout quadratique moyen, $\langle R_e^2 \rangle$, en fonction du nombre de monomères total formant la chaîne, N évolue selon $\langle R_e^2 \rangle \sim (N-1)^{2\nu}$ avec $\nu \approx 0.59 \pm 0.01$ ce qui est une propriété générique des chaînes idéales à l'équilibre [Gennes,

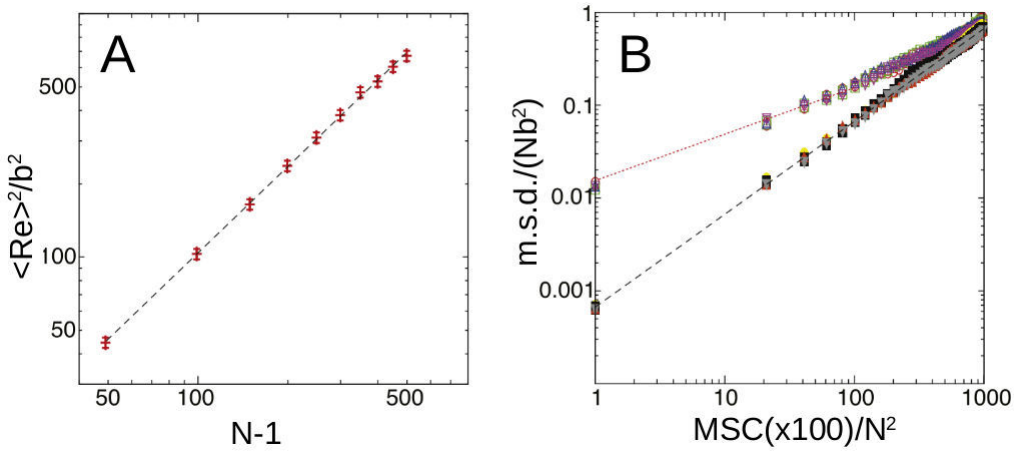

Figure 4.3.2 – **Validation de l'approche de dynamique sur réseau avec l'étude d'un homopolymère auto-évitant isolée sans interactions.** (A) Distance bout-à-bout carré moyenne, $\langle R_e^2 \rangle$ prédite par le modéle sur réseau en fonction du nombre de segments $N-1$ avec N. Le modèle reproduit la loi $\langle R_e^2 \rangle \sim (N-1)^{2\nu}$ [Gennes, 1979], ligne avec tirets sur la figure avec $\nu \approx 0.59 \pm 0.01$. (B) Déplacement quadratique moyen normalisé (abrégé en m.s.d.) du centre de masse g_3 (symboles fermés) et de la bille du milieu g_1 (symbole ouvert) en fonction de temps de simulation normalisé pour $N = 50$, 100, 150, 200 (resp. symbolisés par des cercles, carrés, triangles tournés vers le haut et vers le bas). Le modèle reproduit l'évolution de g_3 en $g_3 \sim t/N$ (droite avec tirets) et de g_1 en $g_1 \sim t^{1/2}$ (droite avec pointillés) aux temps courts, et reproduit aux temps longs $g_1 \sim g_3$ [Jost and Everaers, 2010].

1979]. Aussi, sur la même figure, à droite, on constate que le déplacement carré moyen des monomères par rapport au centre de masse de la chaîne, g_1, évolue en fonction du temps t, en $\sqrt{t}$ puis en t^1 pour de plus longues échelles. La dynamique simulée est donc conforme au modèle de Rouse [Doi and Edwards, 1988]. Le modèle sur réseau permet donc de reproduire les propriétés génériques statique et dynamique d'une chaîne polymérique. Ces deux résultats constituent une première validation de l'approche sur réseau et permettent de poursuivre en introduisant deux types de blocs : actifs et inactifs.

4.3.3.2 Propriétés génériques d'un hétéropolymère

On étudie ici le copolymère par bloc $(A_{10}B_{10})_6$ constitué de six successions de dix monomères actifs et dix inactifs. En faisant varier l'intensité de l'interaction spécifique u_e entre monomères de même état épigénomique, on va pouvoir dégager des propriétés génériques concernant ce copolymère. Sur la figure 4.3.3, on présente le diagramme de phase de ce copolymère $(A_{10}B_{10})_6$ composé des cartes de contact simulées, de la probabilité de contact P_c en fonction de la distance génomique s, des cartes des distances carré moyennes, des distances carré moyenne en fonction de s, du déplacement carré moyen des monomères par rapport au

centre de masse de la chaîne, g_1, en fonction du temps t, et de la probabilité $P_d(n)$ pour un domaine dans un état donné (A ou B) d'appartenir à un cluster composé de n domaines de même état. Pour estimer $P_d(n)$, pour chaque paire de domaines de même état épigénomique, on calcule la proportion de paires de monomères appartenant à différents domaines qui sont plus proche voisins sur le réseau. Si cette proportion est supérieure à 10%, on considère que les deux domaines sont couplés. Sur ce diagramme de phase, on retrouve quatre phases différentes tout comme c'était le cas avec l'approche gaussienne auto-cohérente (Fig. 4.2.31).

Pelote

Pour de faibles valeurs de spécifité, $u_e \sim 0$, le copolymère est dans une conformation allongée de pelote caractérisée par une probabilité de contact $P_c(s) \sim s^{-2.1}$ qui décroît rapidement avec s et avec de rapides mouvement des monomères. On retrouve un comportement similaire à celui d'une chaîne isolée auto-évitante avec une dynamique de Rouse : g évolue en effet en $t^{1/2}$ aux petites échelles de temps et en t^1 ensuite. Comme la chaine est allongée, la densité est faible, ce qui se retrouve avec $P_d(n)$ qui est presque toujours égal à 1 ce qui signifie que chaque monomère n'est entouré que de ses voisins génomiques, appartenant donc la plupart du temps à son propre domaine.

« Pelote avec TADs »

Pour des intensités d'interaction u_e comprises entre -0.1 et $-0.2 k_B T$, on observe une phase caractérisée par une augmentation de la probabilité de contact P_c pour les monomères de même type, et qui sont particulièrement dans le même domaine. Ceci se manifeste d'une part sur les cartes de contact avec la formation de TADs le long de la diagonale et d'autre part sur la probabilité de contact qui évolue selon $P_c(s) \sim s^{-1.5/-1}$. On note donc qu'une loi de type $P_c \sim s^{-1}$ n'est pas forcément caractéristique d'un état non équilibré. On voit en effet, que cette loi de puissance peut être reproduite à l'équilibre avec des interactions spécifiques d'une certaine valeur ($u_e = -0.2 k_B T$ dans notre cas). Par contre, sur les cartes de distances et le déplacement quadratique moyen on n'observe pas de changements. La dynamique est donc toujours du type Rouse. Les domaines épigénomiques se replient sur eux-mêmes en TADs avec tout de même quelques interactions avec des autres domaines du même type. On voit en effet avec $P_d(n)$ que des clusters de taille 2 et 3 apparaissent.

Multistabilité

Pour des intensités d'interaction u_e comprises entre -0.3 et $-0.4 k_B T$, les cartes de contact présente des structures en damiers : les TADs sont maintenus et il se forme maintenant des interactions à longue portée entre domaines de même type épigénomique. Ces contact spécifiques commencent à être visible au niveau des cartes de distance La distribution $P_d(n)$ devient uniforme ce qui signifie que toutes les tailles de clusters sont équiprobables : le système est dans un état multistable. L'ensemble des configurations du copolymère dans ces

conditions est composé de globules de différentes tailles qui sont sans cesse remodelés. On peut noter que cette échelle d'énergie est compatible avec de récentes simulations numériques sur la transition pelote/globule de la chromatine modélisée un copolymère par bloc isolé de taille fini et évoluant hors réseau [Caré et al., 2015].

Séparation en microphase

Pour des intensités d'interaction u_e comprises entre -0.5 et $-0.6 k_B T$, la chaine se sépare en micro phases. La plupart des conformations consiste en la présence de deux compartiments, un pour A et un pour B. De façon plus rare, quelque petits clusters peuvent exister (cf. $P_d(n)$). La probabilitié de contact entre monomère reste élevée avec des valeurs encore plus grande pour les paires de monomères de même type ce qui explique les oscillations observée sur $P_c(s)$. Quant à la dynamique, le modèle de Rouse n'est plus vérifié au profit d'un modèle de reptation avec $g \sim t^{1/4}$ caractéristique des phases denses [Doi and Edwards, 1988].

4.3.3.3 Effets de la taille des domaines

Dans cette section, on cherche à caractériser comment le diagramme de phase précédent (Fig. 4.3.3) se trouve modifié si on change la taille des domaines épigénomiques. Pour cela, on étudie deux copolymère de même nombre de monomères au total ($N = 120$) mais répartis différemment : $(A_5B_5)_{12}$ et $(A_{20}B_{20})_3$. Pour ces deux copolymères, les cartes de contact, de distance et la probabilité $P_d(n)$ introduite précedemment sont représentés figure 4.3.4. La variable permettant de suivre la dynamique, g_1, n'est pas représenté car son comportement ne change presque pas en fonction de la taille des domaines. Pour chaque taille de domaine, on observe au niveau de la carte de contact la même évolution des phases : pelote, formation des TADs, apparition des interactions à longue portée, multistabilité et séparation en microphases. D'un point de vue quantitatif, les TADs apparaissent pour des énergies d'interaction plus faible dans le cas de gros domaines ce qui est cohérent avec de récents résultats analytiques sur des chaînes de taille finie montrant que des chaînes plus longues commencent leur transition θ à de plus faibles énergies d'interaction [Caré et al., 2015]. Pour le copolymère $(A_5B_5)_{12}$, lorsque l'on augmente l'interaction avec le paramètre u_e, les domaines se replient sur eux-mêmes en globule et une transition θ se produit à l'échelle de tout le polymère ($u_e \approx 0.4 k_B T$). Ceci s'explique par le fait qu'à cette échelle, comme les domaines sont petits, le polymère dans son ensemble peut être vu comme un homopolymère en interaction avec lui même. Par contre, l'organisation interne du globule n'est pas aléatoire, il se forme en effet des petits groupes de différentes tailles. En regadant $P_d(n)$ (Fig. 4.3.4), on peut remarquer que si l'interaction u_e est forte, les groupes de 6 ou 12 domaines sont majoritaires, ce qui signifie que tout les monomères de même état épigénomique forment soit un unique groupe, soit deux sous groupes. Concernant le copolymère $(A_{20}B_{20})$, nous observons toujours le repliement in-

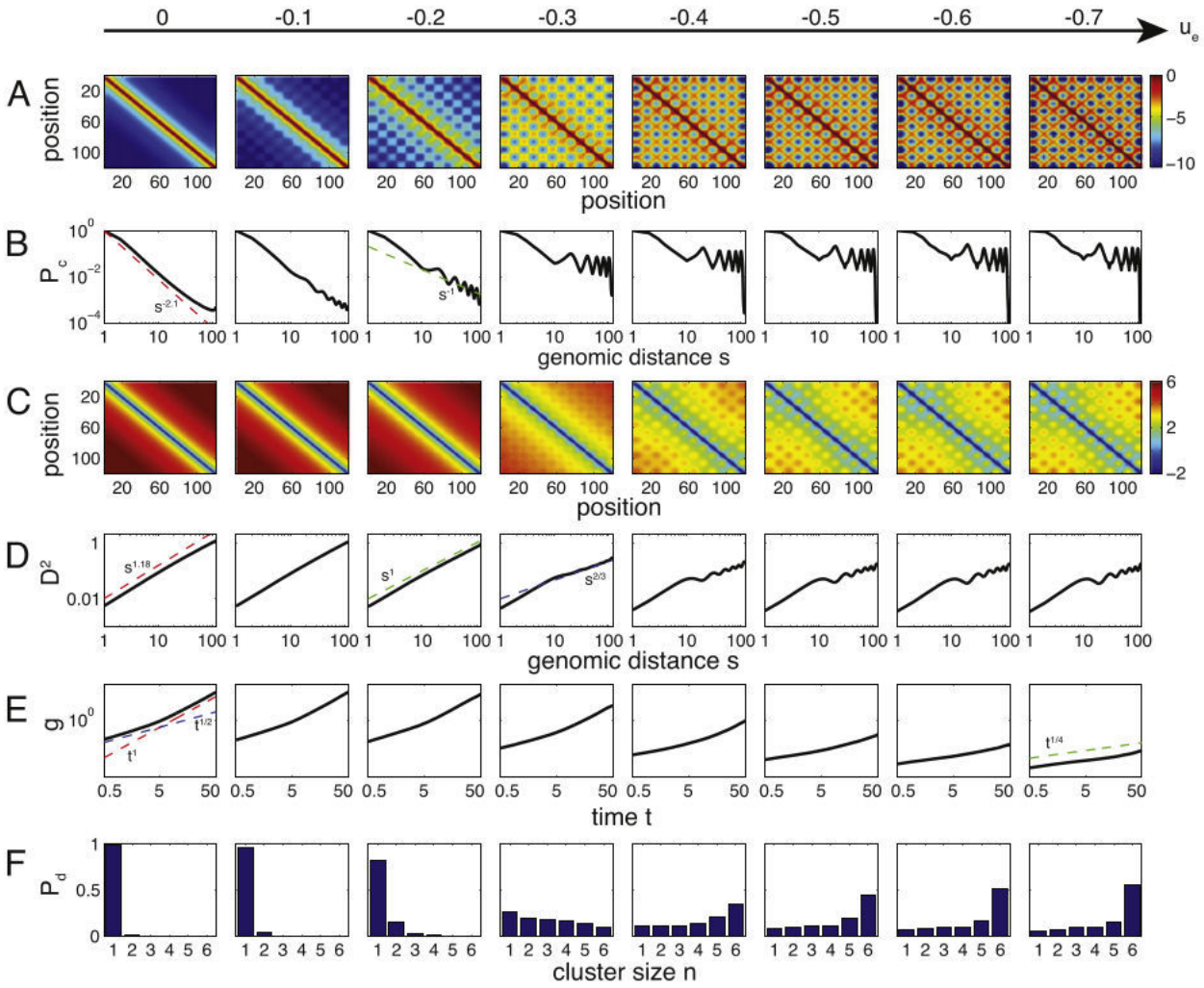

FIGURE 4.3.3 – **Diagramme de phase du copolymère par bloc $(A_{10}B_{10})_6$ obtenu en faisant varier l'intensité de l'interaction spécifique entre monomères de même état épigénomique.** (A) Carte de contact simulée (en échelle $\log_2$). (B) Fréquence de contact moyenne P_c en fonction de la distance génomique s. (C) Carte des distances carré moyennes. (D) Distances carré moyennes en fonction de s. (E) Déplacement quadratique moyen g_1 en fonction du temps t. (F) Probabilité $P_d(n)$ pour un domaine dans un état donné (A ou B) d'appartenir à un cluster composé de n domaines de même état.

terne des domaines mais il n'y a plus de transition θ à grande échelle, le polymère ne peut donc pas être considéré comme homogène. Pour des valeurs de u_e élevées, on observe une répartition homogène des différentes tailles de groupe : les monomères de même état peuvent se réunir pour former un seul groupe, ou bien ils peuvent être séparés en 2 ou 3 groupes.

Si on s'intéresse maintenant au polymère asymétrique en terme de taille de domaines, $(A_5B_{15})_6$, on observe qu'en augmentant l'interaction u_e les TADs et compartiments d'interaction à longue portée se forment d'abord avec les monomères de type B et ensuite avec les A. Cet effet est dû à la différence de taille de domaine entre A et B. Sur les cartes de distance de ce copolymère on peut voir qu'une transition θ se produit à l'échelle de la chaîne, le polymère apparaît comme homogène car les monomères A sont négligeables. Pour des énergies très fortes, une séparation en microphase est observée avec la formation d'un compartiment pour les monomères de type A et un ou plusieurs groupes de B.

4.3.3.4 Comparaison avec l'approche gaussienne auto-cohérente

Le diagramme de phase obtenu avec la dynamique sur réseau (Fig. 4.3.3) et ceux réalisés avec l'approximation gaussienne auto-cohérente se valident mutuellement (Fig. 4.2.31). En effet, on a pu observer l'existence des quatre même phases avec les mêmes caractéristiques. L'approche gaussienne auto-cohérente est plus simple à mettre en oeuvre mais la dynamique apporte plus de données. En particulier, avec cette dernière approche on obtient les distances et nombres de contact entre monomères indépendemment alors qu'avec la première méhode on utilisait une relation de puissance approximée (Éq. 4.13). En observant sur la figure 4.3.3 ces deux types de données obtenus avec la dynamique sur réseau, on voit clairement que distance et nombre de contacts ne peuvent pas être simplement liés par une loi de puissance et ce particulièrement à grande échelle. En effet, on peut voir pour des valeurs de u_e de $-0.1k_BT$ ou $-0.2k_BT$ que les TADs sont présents sur les cartes de contact mais pas sur les cartes de distance (A et C sur la figure), une relation en puissance entre les deux grandeurs ne peut pas reproduire cette observation. On peut aussi voir ce résultat en regardant toujours sur la même figure la probabilité de contact, P_c et la distance quadratique moyenne notée ici, D^2, en fonction de la distance génomiqe, s : on voit que si par exemple $u_e = -0.3k_BT$, pour P_c il existe des oscillations à grande échelle alors que pour D^2 elles ne sont pas présentes, ceci confirme que la relation entre distance moyenne et nombre de contacts ne peut pas être parfaitement décrite par une loi de puissance. Toutefois, dans le cadre de l'approche gaussienne auto-cohérente, on utilisera cette relation approximée 4.13 à défaut d'avoir une meilleur relation.

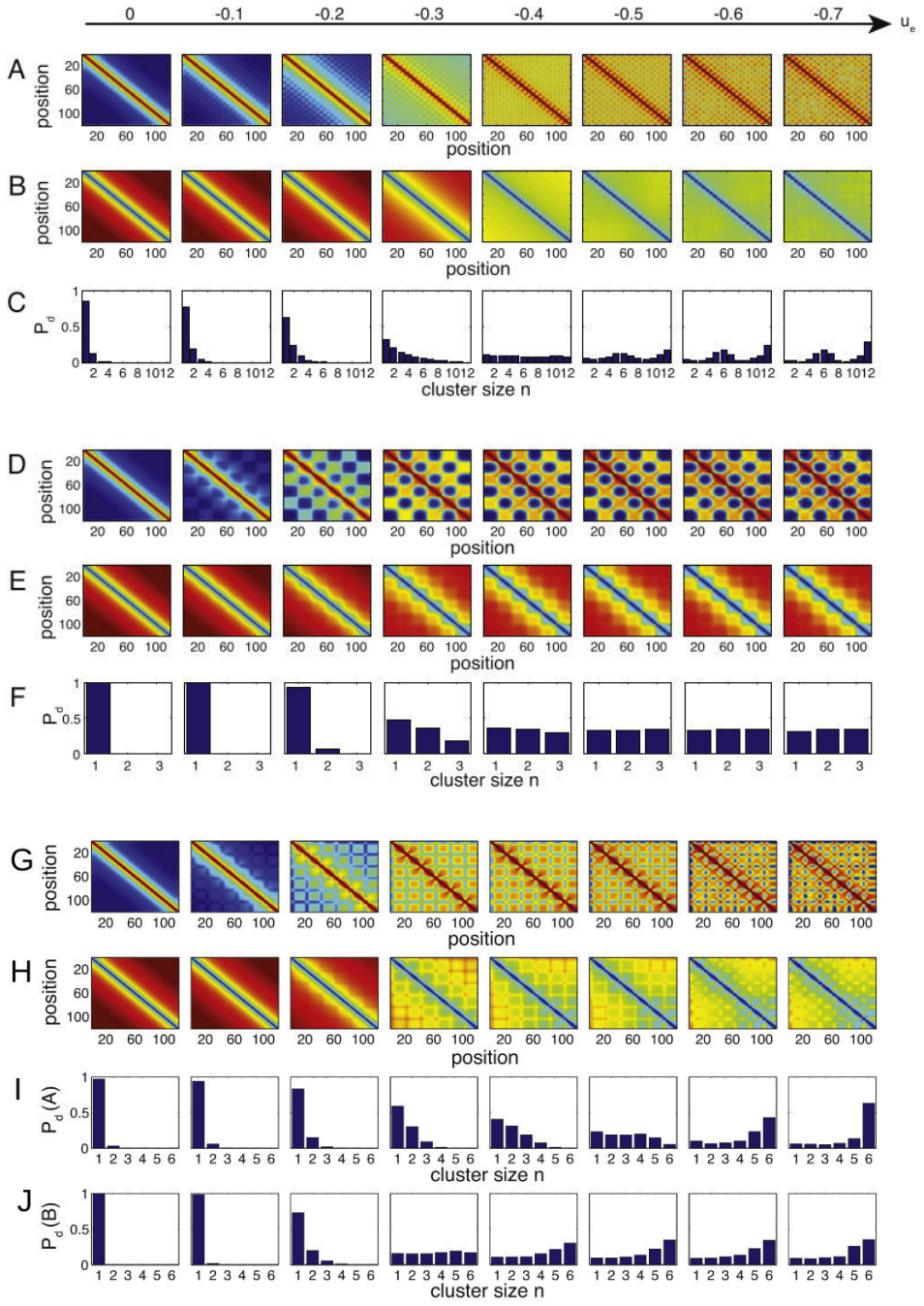

FIGURE 4.3.4 – Propriétés structurales d'un copoloymère en fonction du paramètre $\mathfrak{u}_e$ et en fonction de la taille de ses domaines. Les lignes A, B et C présentent le copolymère $(A_5B_5)_{12}$, les lignes D, E et F s'intéressent à $(A_{20}B_{20})_3$ et enfin les lignes G, H, I et J correspondent au copoymère asymétrique $(A_5B_{15})_6$. (A,D,G) Matrice de contact avec la même échelle de couleur que figure 4.3.3(A). (B,E,H) Matrice des distances quadratiques moyennes avec la même échelle de couleur que figure 4.3.3(C). (C,F,I,J) Probabilité $P_d(n)$ pour un domaine dans un certain état épigénomique (A ou B) d'appartenir à un groupe composé de n domaines du même état.

4.3.3.5 Application à la drosophile

Le diagramme de phase obtenue révèle l'existence de conditions telles que les cartes de contact simulées ressemblent fortement à celles obtenues expérimentalement. On s'intéresse donc ici à la modélisation d'un segment chromosomique de la drosophile constitué successivement de domaines épigénomiques , petits et actifs, et, grands et inactifs (Fig. 4.3.3.5 partie triangulaire supérieure). Le repliement de cette région est bien modélisé en choisissant comme paramètre $u_{actif} = -0.1 k_B T$ et $u_{inactif} = -0.3 k_B T$ (Fig. 4.3.3.5 partie triangulaire inférieure). Ces valeurs ont été choisies en minimisant la différence quadratique entre carte de contact expérimentale et simulée. On peut noter que le jeu de paramètre $u_{actif} = -0 k_B T$ et $u_{inactif} = -0.4 k_B T$ fonctionne tout aussi bien. On constate que l'énergie d'interaction entre domaines actifs est plus faible qu'entre domaines inactifs ce qui est cohérent avec le fait que l'euchromatine est souvent moins dense que l'hétérochromatine [Gilbert et al., 2004] et cohérent avec une récente application du formalisme du copolymère au cas du repliement de la chromatine de la drosophile [Ulianov et al. (2016)]. Pour ces échelles d'énergie spécifique, le modèle prédit que les interactions à longue portée observées entre domaines épigénomiques reflètent un état multistable avec la formation multistable de clusters de différentes tailles en permanence remodelés. Cette existence d'états multistables est cohérente avec de récentes expériences de HiC à cellule unique et avec des images super résolues montrant que les TADs sont conservés entre cellules mais que ce n'est pas le cas des interactions à longue portée [Nagano et al., 2013].

Une prédiction forte du modèle est que pour une intensité d'interaction, u_e, fixée, la taille des domaines joue sur la compaction. Afin de tester la validité de cette prédiction, nous allons tout d'abord définir la compaction et ensuite la calculer à partir de cartes de contact expérimentales et simulées.

Pour une paire de monomères (i, j) donnée le ratio $r_c(i,j) = \frac{C(i,j)}{P_c(|i-j|)}$ entre le nombre de contacts Hi-C entre i et j et entre le nombre moyen de contact à la distance $|i-j|$ quantifie si un contact entre i et j est enrichie par rapport au cas neutre (sans structure 3D particulière) ou s'il est au contraire appauvri. On estime que le niveau de compaction d'un domaine est la moyenne des ratios, $r_c(i,j)$, avec $i \neq j$ contenus dans tout le domaine. Une meilleure manière de faire serait de calculer la concentration des monomères par domaine épigénomique (nombre de monomères sur le rayon de gyration du domaine au cube) normalisé par un modèle de chaîne neutre ($u_e = 0$) mais les expériences de Hi-C ne permettent pas d'accéder au rayon de gyration des domaines. Toutefois, sur la figure 4.3.6 on voit dans l'insert que théoriquement la compaction et la concentration sont fortement corrélées. Il est donc pertinent de poursuivre avec la compaction. En choisissant des intensités d'interaction spécifiques réalistes ($u_e \sim -0.2/-0.3 k_B T$), on observe que les grands domaines sont plus compacts. Ce résultat souligne

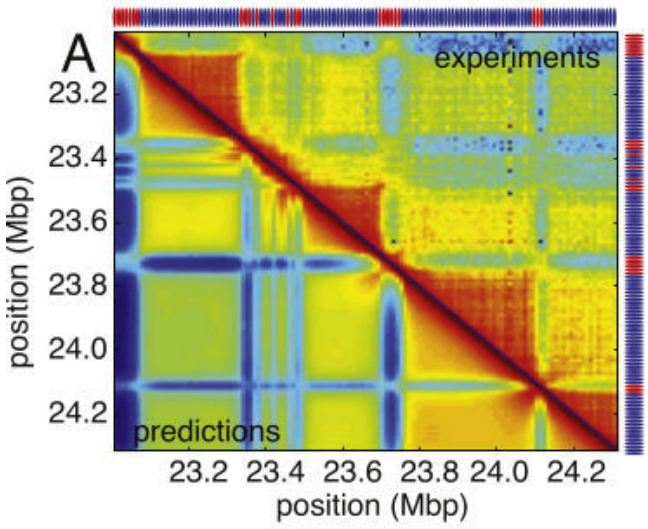

FIGURE 4.3.5 – **Cartes de contact expérimentale en haut et obtenue avec des simulation de dynamique sur réseau et Monte Carlo cinétique en bas.** Les paramètres $u_{actif} = -0.1 k_B T$ et $u_{inactif} = -0.3 k_B T$ ont été choisis de sorte à reproduire le mieux possible la carte de contact expérimentale. L'échelle de couleur est identique à celle de la figure 4.3.3(A).

et confirme l'importance des effets de taille finie dans le repliement de la chromatine [Caré et al., 2015, Cortini et al., 2016]. Afin de tester cette prédiction, on a calculé le degré de compaction des TADs observés chez la drosophile [Sexton et al., 2012] en fonction de leur taille et en fonction de leur état épigénomique (Fig. 4.3.6). On observe qu'en moyenne la compaction est plus élevée pour les grands domaines (inactifs) tout comme cela était prédit par le modèle. Par contre, pour les domaines actifs la compaction ne dépend pas de la taille du domaine. Ceci suggère que l'euchromatine n'interagit que très peu avec elle-même. Il y aurait donc deux modes d'interaction locale distincts : l'euchromatine s'organise localement de manière discrète avec des pontages à courte portée entre loci spécifiques alors que l'hétérochromatine interagit plus continuellement avec des regroupements de plusieurs sites génomiques. La définition de ces deux modes d'interaction est cohérente avec l'observation expérimentale selon laquelle les domaines inactifs présentent des contacts homogènes alors que pour les domaines actifs, l'interactome est plus complexe [Sofueva et al., 2013].

4.3.4 Conclusion

La dynamique sur réseau avec Monte Carlo cinétique permet de confirmer les propriétés précédemment démontrées avec l'approximation gaussienne auto-cohérente. Cette approche est plus coûteuse en terme de temps de calcul mais a permis d'obtenir plus de résultats en

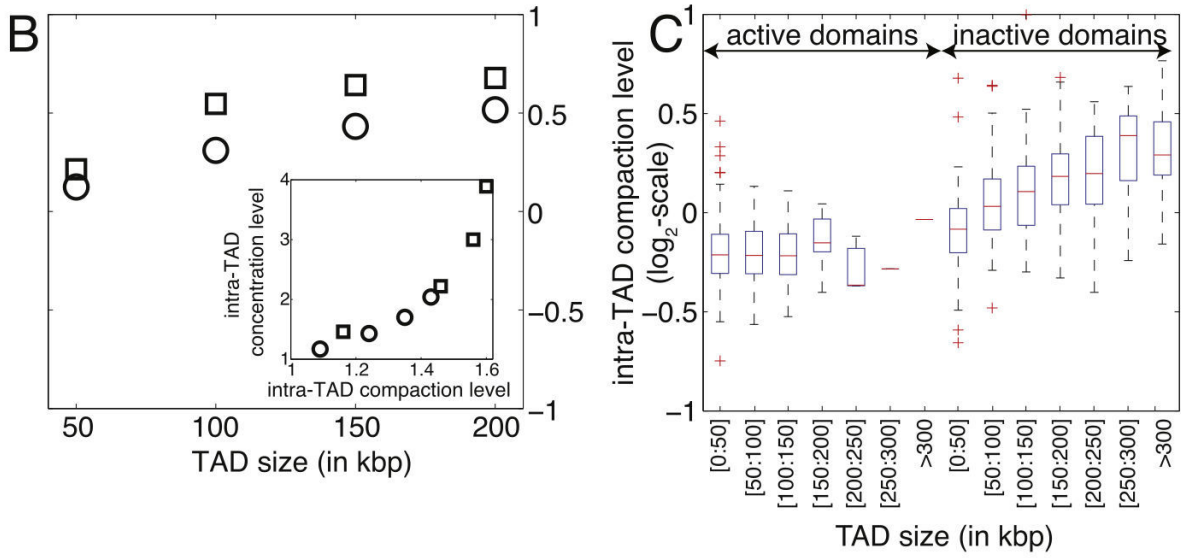

Figure 4.3.6 – **Les prédictions faites par des simulations de dynamique sur réseau sont compatibles avec les données expérimentales.** (B) Prediction de la compaction intra-TAD pour $u_e = -0.2 k_B T$ (cercle) et $u_e = -0.3 k_B T$ (carré) calculé pour les copolymères $(A_5 I_5)_{12}$, $(A_{10} I_{10})_6$, $(A_{15} I_{15})_4$ et $(A_{20} I_{20})_3$. Insert : Concentration intra-TAD en fonction du niveau de compaction intra-TAD. On observe une corrélation entre ces deux grandeurs justifiant le calcul de compaction. (C) Niveau de compaction intra-TADs observé chez la drosophile en fonction de la taille et du type de TAD (actif / inactif). Cette observable, calculée à partir des données de Sexton et al. (2012), confirme les prédictions du modèle présenté en (B). On observe en effet dans le cas des domaines inactifs que plus ils sont grands plus le niveau de compaction est élevé. Concernant les domaines actifs la compaction ne dépend pas de la taille des domaines suggérant que l'euchromatine intéragit faiblement avec elle même (suggestion développée dans le 3).

particulier concernant la dynamique et la caractérisation de la multistabilité. Toutefois, nous avons travaillé avec des petits segments équilibrés. Or, in vivo, la chromatine peut se trouver équilibré ou pas selon les organismes et les cellules. Il est donc intéressant de poursuivre avec la dynamique moléculaire à l'échelle du génome afin de caractériser plus en détail le processus d'équilibration. Dans la section suivante, on présentera des travaux de dynamique moléculaire réalisés par Pascal Carrivain lors de son post-doc à l'ENS de Lyon.

4.4 Dynamique moléculaire à l'échelle du génome

4.4.1 Motivation

L'approximation gaussienne auto-cohérente, développée plus haut, nous a permis d'obtenir des propriétés d'ensemble concernant de petites chaînes isolées. Ces propriétés d'ensemble ont ensuite été retrouvées avec la dynamique sur réseau, Monte Carlo cinétique, qui donne accès aux propriétés dynamiques de chaque chaîne individuelle. Ici, on va présenter des simulations de dynamique moléculaire à l'échelle du génome complet de la drosophile réalisées par Pascal Carrivain. Cette approche a le mérite de donner accès aux trajectoires individuelles de chaque monomère au cours du repliement et a le mérite de reposer sur moins d'approximations que les deux précédentes. En revanche, l'inconvénient de taille de la dynamique moléculaire est son coût en terme de temps. Notre but sera dans un premier temps de vérifier si cette dernière approche permet aussi de reproduire les cartes de contact expérimentales. Nous chercherons aussi à savoir jusqu'à quelle échelle en Mb et au bout de combien de temps, on peut considérer qu'une région chomosomique atteint un état stationnaire. La réponse à cette question va nous permettre de préciser dans l'approche gaussienne auto-cohérente les échelles de temps et d'espace à partir desquelles il faut tenir compte du champ de crumpling modélisant les effets hors équilibre.

4.4.2 Formalisme

Dans ce formalisme, les monomères sont modélisés par des capsules de 10kbp caractérisées par une proportion de chaque état épigénomique parmi quatre états possibles : actif, polycomb, hp1 et nul (Fig. 4.4.1). Généralement un bin de 10kb est constitué d'un état épigénomique majoritaire mais ce n'est pas systématique (Fig. 3.2.2). Ces capsules sont représentées sous forme de cylindre de rayon, $r = 11$ nm et de longueur, $l = 173$ nm, avec des bords en demi-sphère. Ces deux dernières grandeurs ont été fixées à l'aide de données FISH et d'un modèle de polymère [Rosa and Everaers, 2008] de manière à conserver la densité volumique et la longueur d'enchevêtrement. Les capsules sont des corps rigides liés successivement par leurs extrémités avec un joint mécanique. Ce joint impose le point de fixation tout en laissant libre l'orientation relative des capsules. Les capsules successives n'exercent donc pas de force élastique les unes sur les autres. Elles évoluent dans une sphère modélisant le confinement dans le noyau. La dynamique de la capsule i de masse m, de centre de gravité X_i et d'hamiltonien H_i est régie par l'équation de Langevin suivante :

$$m\frac{d^2X_i}{dt^2} = -\frac{\partial H_i}{\partial X_i} - \zeta_b\frac{dX_i}{dt} + \eta_i(t)$$

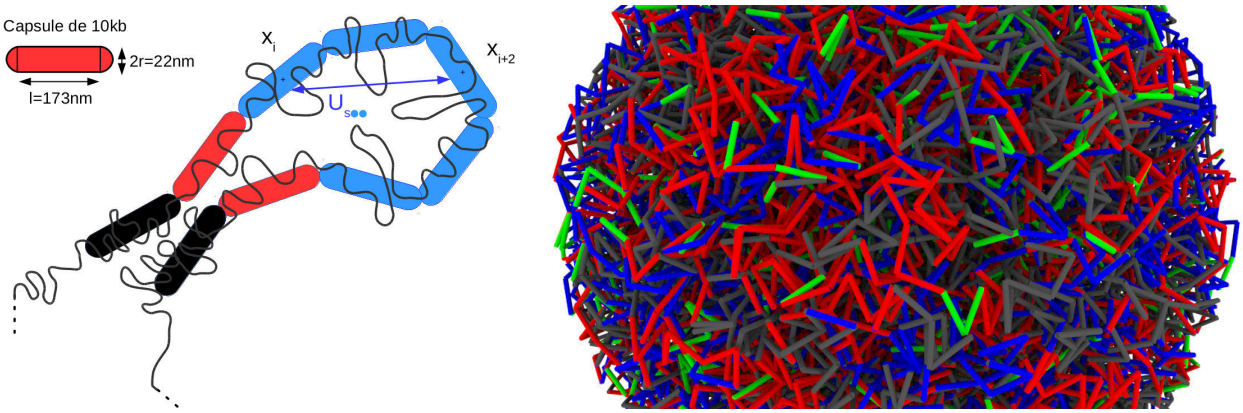

Figure 4.4.1 – **Modélisation de la chromatine par un copolymère par bloc à l'échelle du génome.** Les blocs de 10kb chacun sont modélisés par des capsules cylindriques de longueur $l = 173$ nm et dont les bords sont des demi-sphères de rayon $r = 11$ nm. Chaque capsule est caractérisée par un état ou une proportion d'états épigénétique parmi quatre états possibles : actif en rouge, polycomb en bleu, hp1 en vert et nul en noir. Les interactions entre loci dépendent de leur état épigénomique : les régions de même état interagissent de façon préférentielle. Les capsules sont successivement « collées » les unes aux autres via un point de fixation laissant libre l'orientation relative des capsules. L'ensemble des copolymères représentant différents chromosomes ou bras chromosomiques sont confinés dans une sphère modélisant le noyau cellulaire (figure de droite).

— Le terme $-\frac{\partial H_i}{\partial X_i}$ représente les forces ayant pour origine l'interaction répulsive de volume exclu et l'interaction épigénétique attractive subies par la capsule i. Ces deux forces sont modélisées comme dérivant respectivement d'un potentiel de Morse tronqué à la valeur d'équilibre (afin de gagner du temps de calcul) et d'une superposition de potentiels gaussiens. Finalement, H_i s'écrit :

$$H_i = \sum_j \left(I_s e^{-r_{ij}^2/2\sigma^2} + \alpha \left(e^{-2a(\min(r_{ij},b)-b)} - 2e^{-a(\min(r_{ij},b)-b)} \right) \right)$$

avec r_{ij} la distance entre centres de masse des capsules i et j, $\sigma = \frac{l}{\sqrt{6}}$, I_s l'intensité de l'interaction spécifique, $1/a$ la portée du potentiel de Morse, b la distance d'équilibre entre les deux capsules i et j et α l'intensité du potentiel de Morse qui a été optimisé de manière à respecter le non recouvrement des capsules

Le paramètre I_s dépend de l'état épigénomique des capsules i et j, $I_s = \sum_{c=1}^{4} p_c^i \times p_c^j I_{s_c}$ avec p_c^i (resp. p_c^j) la proportion de la marque c dans la capsule i (resp. j) et I_{s_c} une intensité choisie de telle sorte à reproduire les cartes Hi-C expérimentales : $I_{s_{actif}} \leqslant I_{s_{polycomb}} = I_{s_{HP1}} = I_{s_{nul}}$.

— Le terme $-\zeta_b \frac{dX_i}{dt}$ représente la force de friction présente dans le noyau cellulaire. Le coefficient ζ_b est choisi de façon à optimiser le temps de calcul. Un mapping du temps simulé avec des expériences permet de donner une correspondance entre itérations de la dynamique moléculaire et temps réel (Fig. 4.4.2). Plus précisément, on représente la

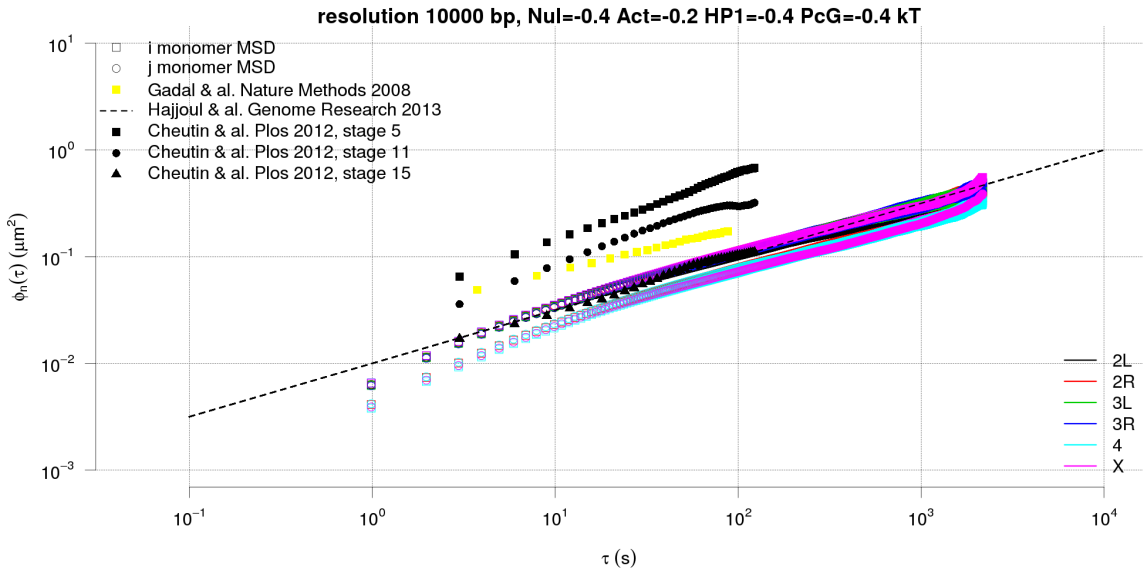

FIGURE 4.4.2 – **Distances quadratiques moyennes parcourues par un monomère, ϕ_n, en fonction du temps réel, τ.** Afin d'obtenir un mapping entre temps de simulation et temps expérimental (aussi appelé temps réel), on trace les déplacements au carré moyens, ϕ_n, en fonction du temps de simulation. Ce déplacement $\phi_n$est estimé en moyennant les déplacements des monomères sur différentes simulations et pour différents moment de départ. Grâce à un ajustement de ϕ_n en fonction du temps simulé, on détermine la meilleure correspondance entre ϕ_n expérimental et ϕ_n réel ce qui nous donne un mapping entre temps simulé et temps réel.

distance quadratique moyenne parcourue par un monomère en fonction du temps de simulation et par un ajustement on trouve la meilleure correspondance entre distances moyennes carrés simulées et expérimentales déterminées par FISH [Hajjoul et al., 2013].

— Le terme $\eta_i(t)$ correspond à un bruit gaussien caractéristique de l'agitation thermique régnant dans le noyau.

Maintenant que chaque terme de l'équation de Langevin a été détaillé, il est nécessaire de préciser le choix de conformation initiale de la chromatine pour réaliser les simulations. Ce choix est particulièrement crucial car il a été montré que pour des polymères longs confinés dans un noyau, le temps d'équilibration peut être très long, plus long que la durée du cycle cellulaire [Rosa and Everaers, 2008]. Cela implique donc que les simulations réalisées ne mèneront jamais la chromatine vers un état d'équilibre aux grandes échelles. Il faut donc partir d'un « bon » état initial car seul une perte partielle de la mémoire de cet état initial ne sera possible. Ainsi, en accord avec des observations expérimentales prouvant qu'à l'état mitotique les chromosomes sont en configuration de Rabl [Hiraoka et al., 1989], chaque chromosome est

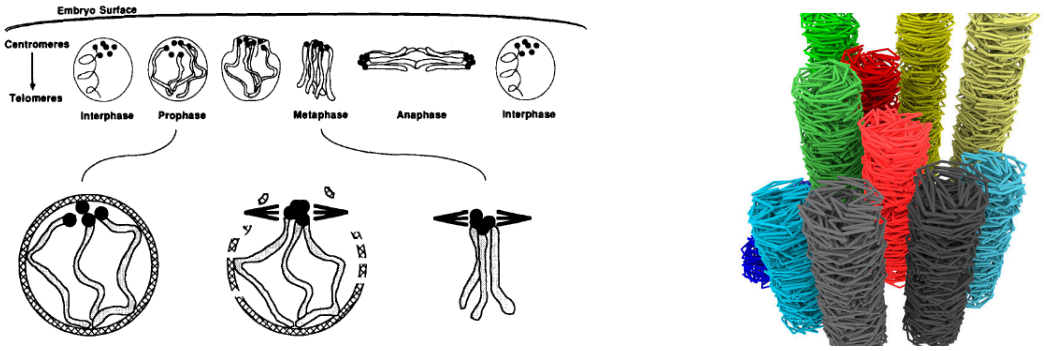

Figure 4.4.3 – **La configuration initiale des chromosomes pour les simulations de dynamique moléculaire est la configuration de Rabl à l'état mitotique.** On simule la dynamique de l'ensemble du génome à partir de cette condition initiale qui est la plus réaliste. [Source de la figure de gauche : Hiraoka et al., 1989].

au début de chaque simulation confiné dans un cylindre virtuel de longueur $\mathtt{L}_\mathtt{m} = 4000\mathtt{nm}$ et de rayon $\mathtt{R}_\mathtt{m} = 400\mathtt{nm}$ et ne présente aucun noeud (Fig. 4.4.3).
Une fois la simulation réalisée, c'est-à-dire après l'équivalent de 2 heures en temps réel, on compare l'accord avec l'expérience. Ceci peut se faire en comparant les cartes de contact expérimentales avec les cartes de contact simulées. Étant donné une simulation de dynamique moléculaire, les contacts sont calculés à partir de plusieurs simulations indépendantes. Tout les loci étant espacés de moins de $\mathtt{b}$ ($\mathtt{b}$ étant défini ci-dessus comme étant la distance d'équilibre entre deux capsules) sont considérés comme étant en contact. Afin d'être cohérent avec les expériences de Hi-C, pour chaque simulation on ne conserve qu'un seul contact par loci, ce contact est choisi aléatoirement. Le nombre de simulations indépendantes est choisi de telle sorte à ce que le nombre total de contact Hi-C simulé soit égal au nombre de contacts Hi-C expérimental brut (c'est-à-dire sans normalisation).

4.4.3 Résultats

Pascal Carrivain a réalisé au cours de ses post-doctorats des simulations de dynamique moléculaire à l'échelle du génome de la drosophile. Les résultats qu'il a obtenu à l'issue de ses simulations semblent présenter un bon accord avec les expériences réalisée quelques heures après la dernière chez mitose chez des embryons tardifs de drosophile (Fig. 4.4.4). En effet, le premier constat est que les deux types de carte se ressemblent fortement dans leur structure : on y retrouve en effet les TADs le long de la diagonale ainsi que les compartiments d'interaction inter-TADs. Ces premières observations suggèrent que les principaux mécanismes impliqués dans le repliement 3D de la chromatine ont été identifiés et modélisés convenablement. Dans le but de comparer simulation et expérience quantitativement, nous allons

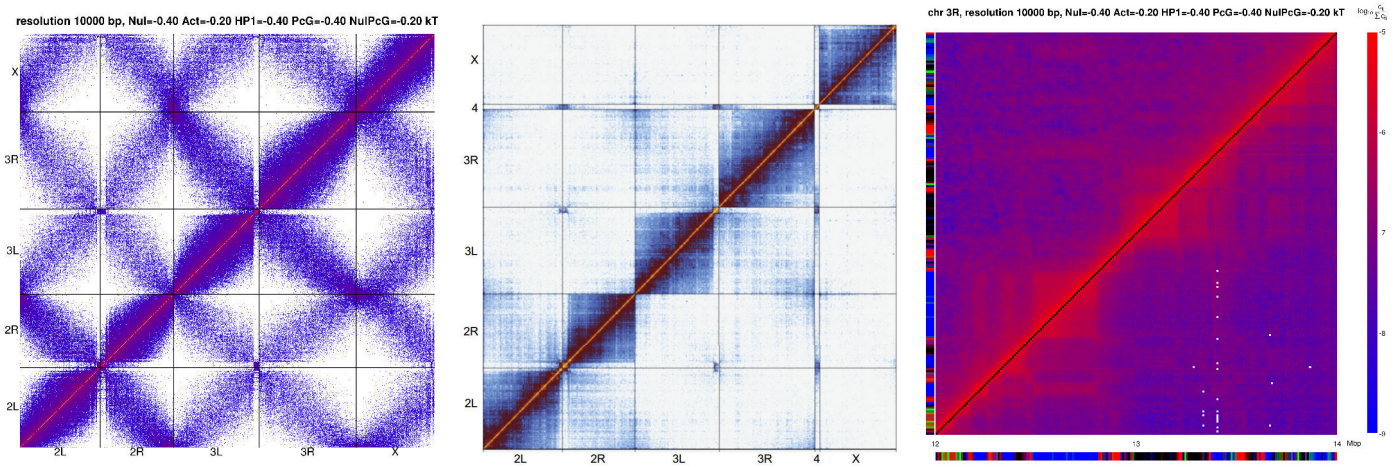

FIGURE 4.4.4 – **Accord global entre expériences et simulations de dynamique moléculaire.** (À gauche) Carte de contact à l'échelle du génome simulé par Pascal Carrivain à partir de la séquence épigénomique de la drosophile donnée par Daniel Day (non publiée). (Au milieu) Carte de contact à l'échelle du génome quelques heures après la dernière mitose chez des embryons tardifs de drosophile. (À droite) Zoom sur une région génomique du chromosome 3R, dans la partie triangulaire supérieure (resp. inférieure), données simulées (resp. expérimentales). A gauche et en dessous de la carte se trouve la séquence épigénomique. [Sources : Pascal Carrivain (pour la simulation) Sexton et al., 2012 (pour le cas expérimental)].

développer ci-dessous une analyse statistique avec les mêmes outils que ceux utilisés dans le 3. Aussi, on abordera succinctement l'aspect dynamique : on verra que les simulations de Pascal Carrivain montre que le temps nécessaire pour tendre vers un état stationnaire local est très court ce qui est un point important pour justifier l'étude de la chromatine via l'approche gaussienne auto-cohérente.

4.4.3.1 Amplification des contacts entre loci de même état

Nous avons développé dans cette partie la même analyse statistique que celle réalisée en section 3.3.2. Il s'agit d'analyser dans quelle mesure les monomères de même état épigénomique interagissent entre eux préférentiellement. Pour quantifier ce phénomène, nous avons introduit les nombres $N^{\mu\nu}(s)$ (Éq. 3.3) qui renseignent sur l'environnement spatial d'un bin dans l'état épigénomique μ interagissant avec des bins d'état ν et espacés d'une distance génomique s et que nous avons représentés, selon un idée de Ralf Everaers, sous forme de diagrammes circulaires. La comparaison entre diagrammes calculés avec le modèle simulé et diagrammes calculés avec le modèle générique (moyenne des contacts simulés par distance génomique, cf. section 3.3.2 pour plus de détails) permet de mettre en avant les effets caractéristiques du

repliement 3D de la chromatine. Nous avons aussi présenté la matrice d'amplification des contacts en fonction des états épigénomiques, $A_{\mu\nu}$ (Éq. 3.4 et idée de Ralf Everaers également). Avec les données Hi-C de [Sexton et al., 2012] et avec les informations épigénomiques de [Ho et al., 2014] nous avions produit la figure 3.3.3. On réalise ici la même figure mais avec cette fois les données Hi-C issues des simulations de dynamique moléculaire de Pascal Carrivain (Fig. 4.4.5).
En observant les figures 3.3.3 et 4.4.5, on retrouve des caractéristiques communes dont en particulier l'enrichissement des contacts entre loci de même état. Toutefois, on peut noter trois points de divergence que nous décrivons ci-dessous :

1. Le nombre de contacts à longue portée (au delà de 5Mb) et inter-chromosomiques est extrêmement faible par rapport au reste des contacts (la surface des anneaux est proportionnelle au nombre de contact) ce qui n'est pas du tout le cas expérimentalement. Ceci peut s'expliquer par un temps de simulation trop faible pour pouvoir reproduire les données de l'embryon tardif ou une définition de la condition initiale pas assez précise.

2. À toutes les échelles de distance génomique, l'amplification des contacts entre loci de même état épigénomique est bien plus prononcée dans le cas des simulations. Cet effet peut s'expliquer par le fait que les interactions entre capsules sont dirigées uniquement par les contraintes topologiques et par les attractions épigénétiques alors qu'expérimentalement d'autres effets biologiques peuvent éventuellement intervenir. On peut imaginer que des simulations réalisées avec moins d'attraction mèneront à des amplifications moins fortes, et donc plus ressemblantes à celles observées expérimentalement.

3. Dans le cas des bins majoritarement de type polycomb (4ème ligne de la figure 4.4.5A), on peut voir que l'enrichissement des contacts avec d'autres bins de type polycomb est net et est présent à toutes les échelles. Pourtant le cas expérimental (4ème ligne de la figure 3.3.3A) révèle que l'attraction spécifique entre bins polycomb ne se fait principalement que sur les trois premiers anneaux, c'est-à-dire jusqu'à 500kb. Cette observation peut être justifiée de la même manière qu'en (2).

Ces trois différences pourraient aussi avoir pour origine la définition des états épigénomiques qui, on peut l'imaginer, ont une influence majeure sur le résultat des simulations. Nous présentons donc les amplifications obtenues avec une autre séquence épigénomique : celle obtenue au sein du laboratoire de Peter Park, après un HMM réalisé par Daniel Day. D'ailleurs, Pascal Carrivain a particulièrement optimisé ses paramètres d'interaction pour cette séquence épigénomique. La figure 4.4.6 montre qu'effectivement les partitions de [Ho et al., 2014] et de Daniel Day sont différentes, avec en particulier un nombre élevé de domaines épigénomiques de type polycomb dans le cas de la deuxième segmentations. Malgré cela, les deux points de

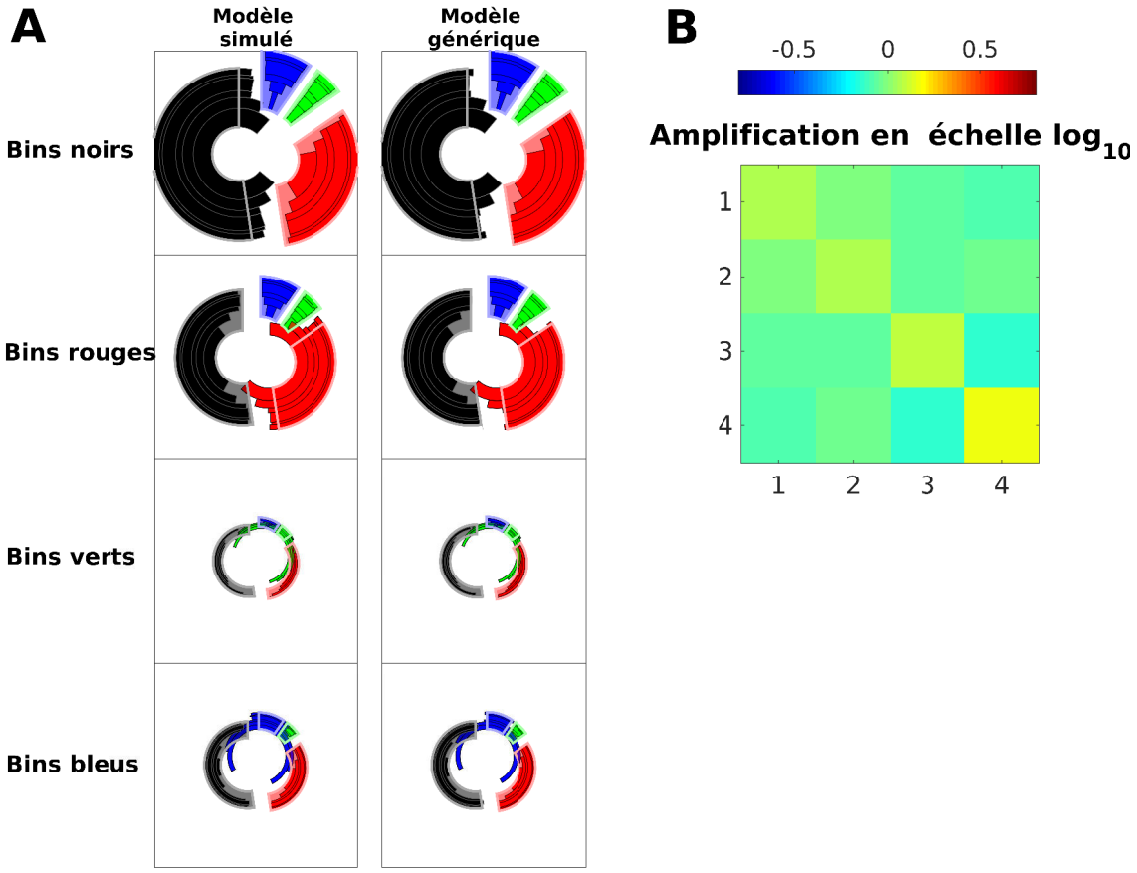

FIGURE 4.4.5 – **Enrichissement des contacts en fonction de l'état épigénomique dans le cas de la drosophile (chromosome 4 exclu de l'étude) [données Hi-C simulées fournies par Pascal Carrivain ; données épigénomiques Ho et al., 2014].** (A) Nombres $N^{\mu\nu}(s)$ (Éq. 3.3) représentés sous forme d'anneaux dans des diagrammes circulaires avec s la distance génomique entre bins en contact. Ces nombres permettent d'évaluer l'environnement 3D des bins en fonction de l'épigénome et en fonction de s. Les cinq premiers anneaux du centre vers la périphérie correspondent respectivement à des valeurs de s telles que $s \leqslant 20\text{kb}$, $20 < s \leqslant 100\text{kb}$, $100 < s \leqslant 500\text{kb}$, $0.5 < s \leqslant 5\text{Mb}$ et $5\text{Mb} < s$. Le 6ème anneau renseigne sur les contacts inter chromosomiques. L'aire des portions est proportionnelle au nombre de contacts. La répartition moyenne des nombres de contact en fonction de l'état épigénomique, pour l'ensemble des chromosomes sauf le 4, est représentée en transparence sur chaque diagramme circulaire. (B) Matrices d'amplification $A_{\mu\nu}$(Éq. 3.4), les lignes et colonnes 1, 2, 3 et 4 représentent respectivement les états nul (noir), actif (rouge), HP1 (vert) et polycomb (bleu). Les paramètres avec lesquels la simulation a été réalisée sont : Nul/Nul : -0.40 ; Actif/Actif : -0.40 ; HP1/HP1 : -0.40 ; Polycomb/Polycomb : -0.40 ; autres : 0.

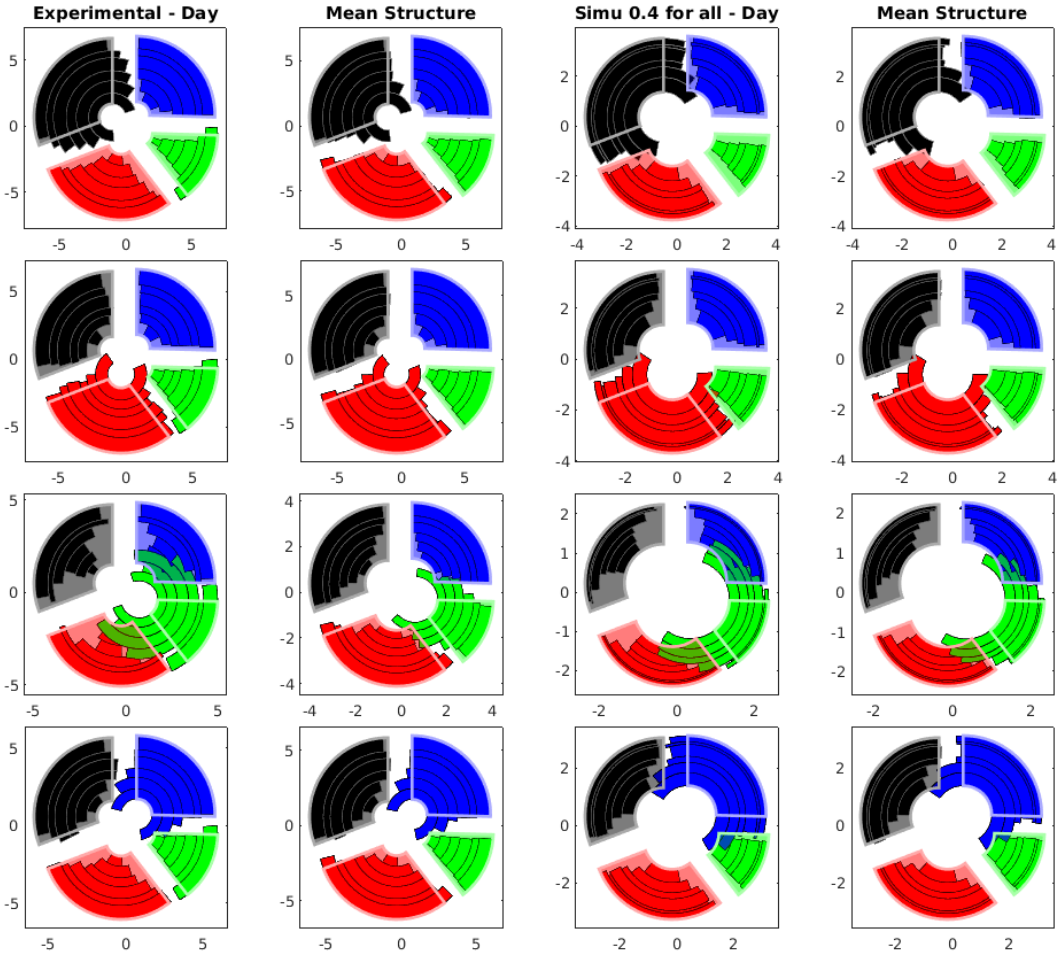

Figure 4.4.6 – **Enrichissement des contacts en fonction de l'état épigénomique dans le cas de la drosophile (chromosome 4 exclu de l'étude) [données Hi-C simulées fournies par Pascal Carrivain ; données épigénomiques de Daniel Day].** Cf. légende de la figure 4.4.5.

divergence énoncés ci-dessus restent valables ce qui renforce donc les explications évoquées pour rendre compte des différences entre simulation et expérience.

4.4.3.2 Corrélations entre domaines épigénomiques et domaines topologiques simulés

Les simulations de dynamique moléculaire pour un copolymère par bloc permettent d'aboutir à des cartes de contact présentant des TADs (Fig. 4.4.4). Ces TADs ont pour origine les attractions entre même état épigénomique introduites dans la simulation. En effet, une simulation avec des paramètres d'interaction nuls mène à des cartes de contact sans TADs. On se demande ici à quel point la séquence épigénomique dicte la formation des TADs. L'ob-

servation de la troisième carte de contact de la figure 4.4.4, par exemple, permet de réfuter l'hypothèse selon laquelle chaque domaine épigénomique donnerait naissance à un domaine topologique. En effet, il apparaît sur cette carte que certains domaines épigénomiques sont trop petits pour pouvoir se replier en TAD. À l'échelle du génome, le constat est le même puisque la simulation est réalisée avec un total de 13153 domaines épigénomique alors que le nombre de TADs déterminés avec IC-Finder est de 1085 (nombre de TADs obtenus à partir de la simulation réalisée à partir des informations épigénomiques de Daniel Day et avec les paramètres d'interaction suivants : Nul/Nul : -0.40 ; Actif/Actif : -0.40 ; HP1/HP1 : -0.40 ; Polycomb/Polycomb : -0.40 ; autres : 0). Une analyse des corrélations entre domaines épigénomiques et TADs à l'échelle du génome permet de calculer le taux de vrai positif, TPR (Éq. 2.1), et le taux de fausses découvertes FDR (Éq. 2.2). Ces taux sont calculés en choisissant que les domaines épigénomiques forment la cible et les domaines topologiques forment l'échantillon étudié.

On trouve alors $\mathrm{TPR_d}$=0.93 et $\mathrm{FDR_d}$=0.84. Le taux de vrais positifs est donc excellent alors que le taux de fausses découvertes est très mauvais car trop élevé. Ceci signifie que deux capsules dans le même domaines épigénomiques sont presque systématiquement dans des TADs identiques (vrai découverte). Par contre, deux capsules dans des domaines épigénomiques différents sont la grande majorité du temps détectés comme étant dans le même TAD (fausse découverte). Ces observations sont dues évidemment à la différence de taille moyenne entre domaines topologiques simulés et domaines épigénomiques. Toutefois, on peut voir sur la figure 4.4.7 que la corrélation est tout de même moins bonne dans le cas d'une fausse séquence épigénomique que l'on a obtenu en inversant l'ordre des domaines épigénomiques de référence (écart entre $\star$ et $\times$ et entre $\square$ et $\diamond$).

La conclusion que l'on peut tirer est que les domaines épigénomiques sont loin de former systématiquement des TADs. Ainsi, les petits domaines épigénomiques en particulier, peuvent tout simplement passer inaperçu en terme de TAD. Toutefois, remarquons que même si un petit domaine épigénomique ne forme pas de TAD, il peut avoir une influence sur l'organisation des TADs voisins. On peut citer l'exemple d'un petit domaine épigénomique entre deux plus grands domaines de même couleur : ce petit domaine peut ne pas former de TAD mais imposer par sa présence la formation de deux TADs adjacents au lieu d'un seul (ce rôle de barrière entre deux domaines de même couleur évoqué n'est bien sûr pas systématique chez les petits domaines épigénomiques).

4.4.3.3 Corrélations entre domaines topologiques expérimentaux et simulés

Afin de savoir si l'organisation des TADs entre cartes de contact expérimentale et simulée sont similaires, on applique IC-Finder aux cartes de contact simulées et on calcule les taux

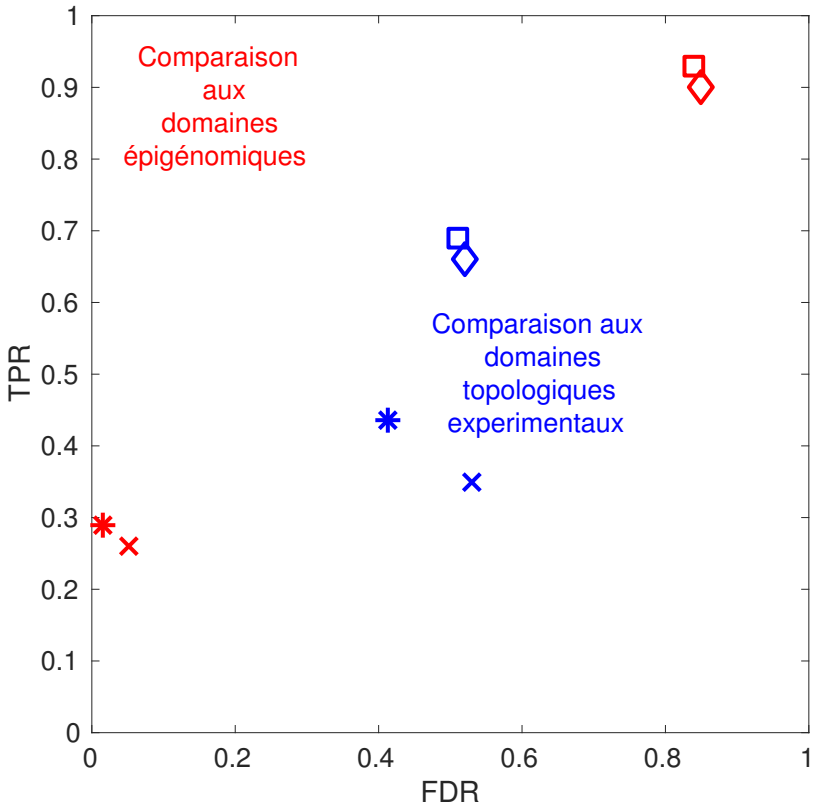

Figure 4.4.7 – **Taux de vrai positif, TPR, en fonction du taux de fausse découverte, FDR, entre les domaines topologiques trouvés par simulation de dynamique moléculaire et les domaines épigénomique utilisés pour réaliser la simulation (resp. et les domaines topologiques expérimentaux) en rouge (resp. en bleu).** Les carrés et les diamants sur la figure indiquent les valeurs ($\mathrm{FDR_d}$, $\mathrm{TPR_d}$) respectivement dans le cas exact et dans le cas où l'ordre des domaines dans la partition de référence est inversé. De la même manière, les astérisques et les croix correspondent aux couples ($\mathrm{FDR_b}$, $\mathrm{TPR_b}$) respectivement dans le cas exact et dans le cas où l'ordre des domaines dans la partition de référence est inversé. La simulation a été réalisée par Pascal Carrivain avec les domaines épigénomiques donnés par Daniel Day et avec les paramètres d'interaction suivants : Nul/Nul : -0.40 ; Actif/Actif : -0.20 ; HP1/HP1 : -0.40 ; Polycomb/Polycomb : -0.40 ; autres : 0.

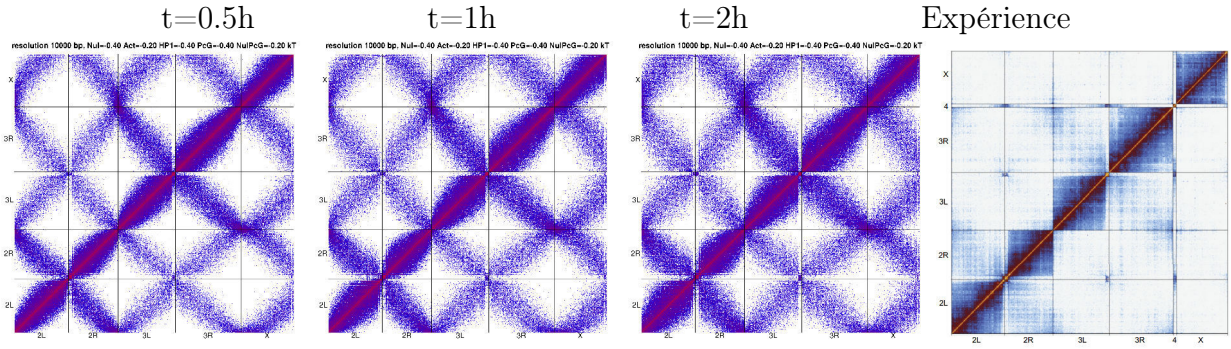

FIGURE 4.4.8 – **Cartes de contacts en fonction du temps issues de simulations de dynamique moléculaire sur l'ensemble du génome de la drosophile.** De gauche à droite, les trois premières cartes de contact représentent respectivement la situation après un temps de relaxation de 0.5h, 1h et 2h. La figure la plus à droite représente les contact à l'échelle du génome dans le cas expérimental. On observe que pour des embryons tardifs, quelques heures après la dernière mitose, la proximité spatiale entre les bras 2L/2R et 3L/3R est toujours présente. C'est aussi le cas dans les simulations.

TPR et FDR avec comme partition de référence la segmentation en TADs expérimentale (aussi obtenue avec IC-Finder et avec les données de [Sexton et al., 2012]). On peut voir sur la figure 4.4.7 qu'il existe une réelle corrélation entre les deux segmentations en TADs, dans le sens où elle n'est ni due au hasard, ni aux tailles des domaines dans les partitions (écart entre $\star$ et $\times$ en particulier). Toutefois, la corrélation pourrait être meilleure. Les écarts entre les deux partitions peuvent avoir pour origine une mauvaise normalisation de la carte de contact expérimentale rendant difficile la comparaison avec des cartes simulées puisque celles ci ne sont pas atteints de biais expérimentaux. Ces écarts peuvent aussi s'expliquer par des imprécisions dans la séquence épigénomique utilisée ou encore par un choix des paramètres d'interaction non optimal (l'optimisation des paramètres est un travail de Pascal Carrivain en cours).

4.4.3.4 Aspect dynamique / temps d'équilibration

Les simulations de dynamique moléculaire permettent de suivre l'évolution du repliement du copolymère en fonction du temps. On peut voir sur les cartes de contact des figures 4.4.8 (génome complet) et 4.4.9 (zoom sur une région du chromosome 3R) que l'aspect dynamique semble bien reproduit par les simulations de Pascal Carrivain puisque ces dernières révèlent que la naissance des TADs est rapide et qu'elle précède la formation des interactions longues portées ce qui a été vérifié expérimentalement [Naumova et al., 2013]. En effet, les cartes obtenues au bout de 30 minutes sont déjà très ressemblantes à celles obtenues au bout de deux heures. L'organisation en TADs est déjà présente. Au delà de 30 minutes, le contraste

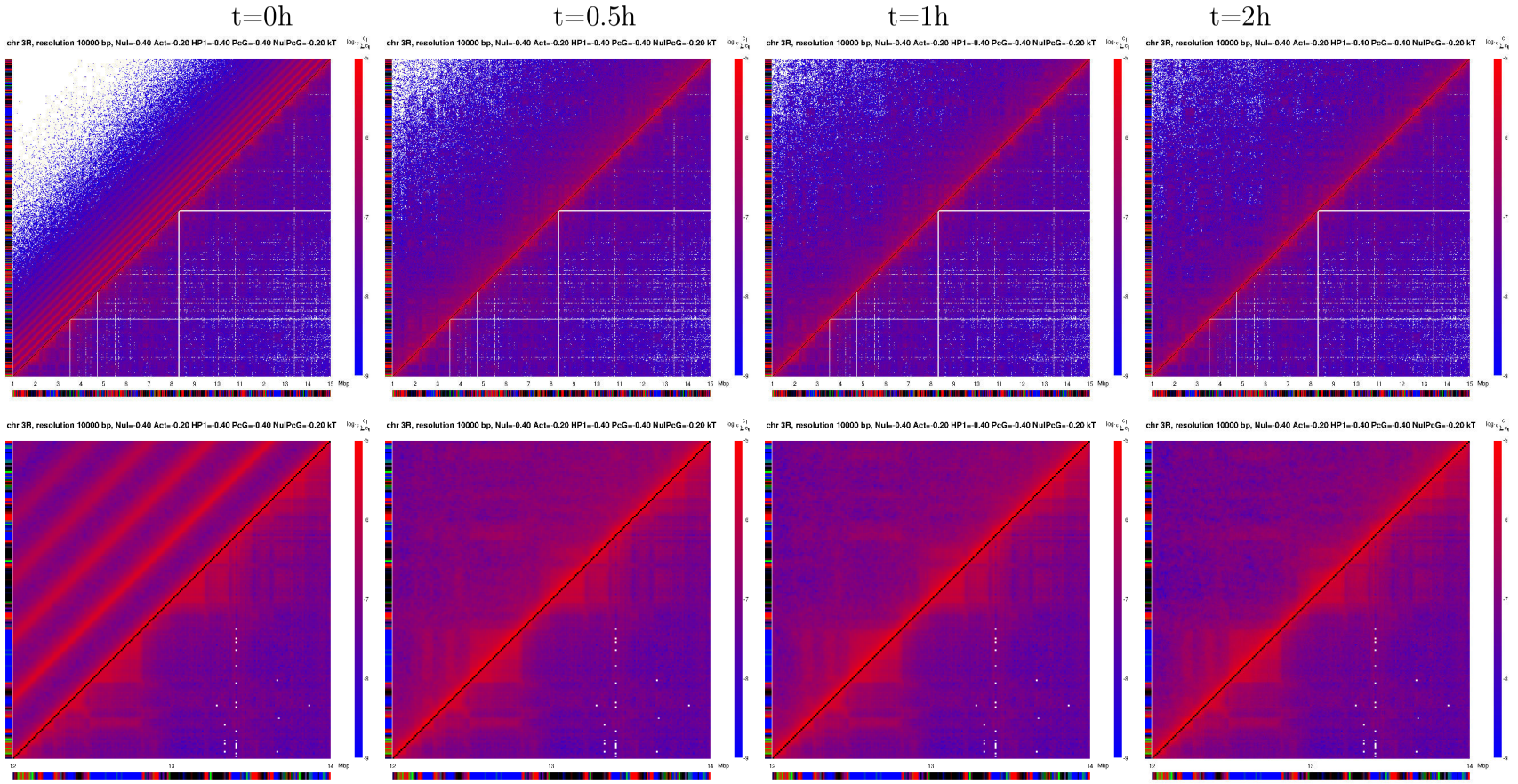

Figure 4.4.9 – **Cartes de contacts en fonction du temps issues de simulations de dynamique moléculaire sur l'ensemble du génome de la drosophile.** En haut, on présente les cartes de contact entre 1 et 15Mb du chromosome 3R de la drosophile. En bas, il s'agit d'un zoom sur la région se situant entre 12 et 14Mb. Sur chaque carte, la partie supérieure correspond à la simulation tandis que la partie triangulaire inférieure de la matrice représente les données expérimentales [Sexton et al., 2012]. Notons que ces dernières sur chacune des quatre cartes car nous n'avons pas accès aux contacts en fonction du temps (données d'embryiogenèse non publiées). Pour toutes les figures, l'échelle de couleur est la même, le blanc représente l'absence de données. Sur chacune des lignes, les quatre cartes correspondent de la gauche vers la droite à des temps de plus en plus grand (t=0h, t=0.5h, t=1h et t=2h marquant la fin de la simulation). On voit sur ces matrices qu'au début de la simulation, les chromosomes sont en configuration de Rabl, ils évoluent ensuite de sorte à former d'abord des TADs puis ensuite des interactions longues portées. Les couleurs rouges bleues vertes et noires de la séquence épigénomique en dessous et à gauche de chaque carte correspondent respectivement aux marques épigénétiques actives, polycomb, HP1 et nulles. Les proportions épigénétiques utilisées ici sont celles fournies par Daniel Day.

au niveau des contacts augmente légèrement et progressivement, il y a aussi de plus en plus d'interactions à longue portée. Ces observations sont non seulement consistantes avec les données expérimentales montrant chez l'homme une formation rapide des TADs puis un enrichissement progressif en interaction TAD/TAD [Naumova et al., 2013] mais elles sont aussi consistantes avec le résultat d'expériences de Hi-C réalisées à différents stades de l'embryogenèse de drosophile dans le laboratoire de Giacomo Cavalli (résultats en cours de publication).

L'aspect dynamique peut aussi être étudié via l'examen du nombre de contacts moyen, N_c, en fonction de la distance génomique, s. Sur la figure 4.4.10, on présente cette observable pour le chromosome 3R de la drosophile dans le cas expérimental (embryons tardifs) et pour différentes conditions de simulations. La courbe noire (cas expérimental) est à comparer à la courbe bleue obtenue après deux heures en temps réel et pour des paramètres d'interaction visant à reproduire le mieux les données (valeurs des paramètres indiquées dans la légende de la figure). On peut voir que les courbes ont globalement la même allure sauf à très petite et très grande échelle. En effet, à très petite échelle, la simulation prévoit une loi du type $N_c \sim s^{-2}$, caractéristique d'une chaîne auto-évitante, alors que les expériences ne prévoit pas du tout cet effet, mais au contraire une décroissance avec une pente très légère. Comme évoqué dans le 3, cet effet peut éventuellement être un artefact causé par la technique du Hi-C (le formaldéhyde nécessaire à la ligation rapproche artificiellement les monomères les uns des autres).

À très grande échelle, au delà de 20Mb, on peut voir sur la courbe noire qu'il existe un léger effet de mémoire de la configuration initiale. Par contre, dans le cas de la simulation, on retrouve à partir de $s \approx 5$Mb la configuration de Rabl initiale d'une manière extrêmement prononcée. Cet effet peut soit s'expliquer par un problème de mauvaise approximation de la condition intiale, soit s'expliquer par un temps de simulation trop court. On peut rappeler ici qu'avec l'approximation gaussienne auto-cohérente, on avait déjà observé que les grandes échelles mettent beaucoup plus de temps à s'équilibrer par rapport aux petites (Fig. 4.2.9). Ceci laisse à penser que les simulations devraient être plus longues en terme de temps afin de reproduire au mieux le repliement de la chromatine dans le cas des embryons tardifs de drosophile. Toutefois, on peut remarquer qu'allonger la durée des simulations n'est pas aussi simple qu'il y parait, c'est en effet un défi computationnel d'obtenir des résultats satisfaisants dans des temps raisonnables.

Aux échelles intermédiaires, entre 100kb et 5Mb, on observe une décroissance de N_c du type s^{-1} dans le cas expérimental. Dans le cas de la simulation avec paramètres optimisés (traits pleins sur la figure 4.4.10), on observe qu'au bout de vingt minutes, l'état stationnaire est atteint avant 200kb et qu'au bout de 2h environ, il est atteint en deçà de 1Mb. Cette obser-

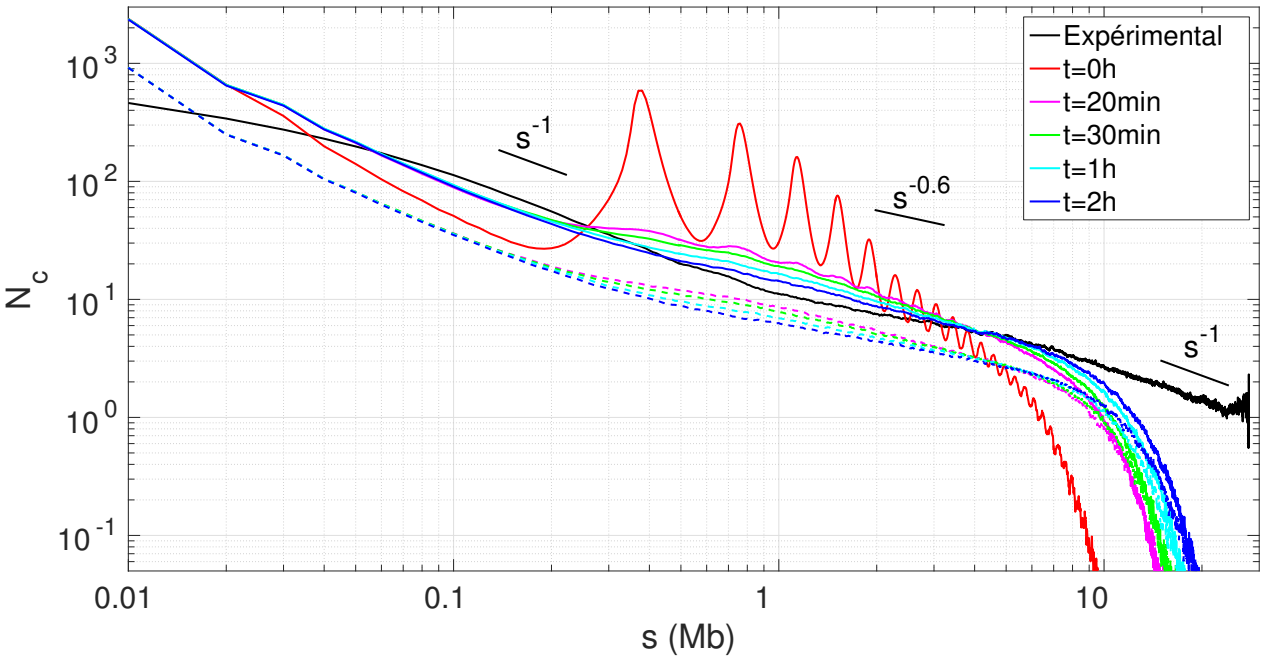

Figure 4.4.10 – **Nombre de contacts moyen simulé N_c en fonction de la distance génomique s pour quatre temps de relaxation successifs (cf. légende) et pour deux jeux de paramètres différents pour le chromosome 3R de la drosophile.** Les courbes en trait plein ont été obtenues avec les paramètres d'interaction optimisés de manière à reproduire au mieux les expériences (Nul/Nul : -0.40 ; Actif/Actif : -0.20 ; HP1/HP1 : -0.40 ; Polycomb/Polycomb : -0.40 ; Nul/Polycomb : -0.20 et autres : 0). Les courbes en tirets sont obtenues avec le modèle neutre caractérisé par des paramètres d'interaction tous nuls. La courbe noire représente le cas expérimental, elle est accompagnés de segments noirs rappelant les lois d'échelle vues précédemment (Fig. 3.1.4). Les simulations fournies par Pascal Carrivain ont été réalisées à partir de la partition épigénomique de Daniel Day.

vation valide notre modélisation de la chromatine en s'intéressant aux solutions stationnaires trouvées avec l'approximation gaussienne auto-cohérente.

Au fil du temps de simulation, la pente s'accentue légèrement. On peut imaginer que si la simulation était plus longue, on retrouverait au bout d'un certain temps la loi expérimentale $N_c \sim s^{-1}$.

On a vu en introduction qu'une loi de type $N_c \sim s^{-1}$ pouvait avoir pour origine un effet de crumpling ou bien avoir pour origine l'existence d'interactions spécifiques. Ici, en observant les résultats obtenus avec paramètres d'interaction optimisés (trait pleins) et avec paramètres d'interaction nuls (tirets), on voit que dans les deux cas, entre 100kb et 5Mb, la décroissance se fait selon $N_c \sim s^{-1}$ ce qui suggère que ce comportement n'est pas caractéristique des interactions spécifiques. Autrement dit, avec ou sans interactions, la physique du repliement générique est la même.

4.4.4 Conclusion

On a vu que les simulations de dynamique moléculaire permettent de reproduire des cartes de contact ayant les mêmes caractéristiques que les cartes expérimentales. On a vu que les TADs se formaient rapidement, puis suivaient ensuite les compartiments d'interaction à longue portée. Par contre, l'équilibration est très lente comme déjà observé dans le cadre de la modélisation avec l'approximation gaussienne auto-cohérente. Les simulations présentées avaient toutes le défaut de présenter aux grandes échelles un effet de mémoire important par rapport aux expériences, effet difficile à atténuer en un temps imparti. La dynamique moléculaire nous a permis de voir que chez la drosophile un segment chromosomique peut être considéré comme étant dans un état stationnaire déjà au bout d'une vingtaine de minutes. Ce résultat permet de justifier la pertinence de l'approche reposant sur l'approximation gaussienne auto-cohérent. On a vu que l'issue des simulations de dynamique moléculaire dépend en bonne partie du choix des paramètres d'interaction. Ce choix peut être orienté grâce à des analyses statistiques telles que celles menées au chapitre 3 et développées ici (on a par exemple vu que les domaines rouges semblent être moins attractifs entres eux) ou il peut l'être par l'approche gaussienne auto-cohérente qui est plus rapide à mettre en oeuvre. Toutefois, notons que pour trouver une correspondance entre les paramètres des deux approches, il y aura un travail technique à réaliser. Finalement, la dynamique moléculaire et l'approche gaussienne auto-cohérente s'avèrent être complémentaires, l'une pouvant fournir des informations pour l'autre. De plus, comme nous l'avons vu elles permettent de valider mutuellement leurs résultats respectifs. Enfin, on peut souligner le fait que les simulations de dynamique moléculaires sont très riches, elles permettent d'étudier bien plus d'aspects que ceux présentés dans cette section. Par exemple, on pourrait précisément caractériser le phénomène de multistabilté discuté dans les sections précédentes.

4.5 Résumé et complémentarité des trois approches

Nous avons décrit dans ce chapitre le formalisme permettant de modéliser la chromatine par un copolymère par bloc linéaire, non homogène auto-évitant et dont les monomères ont chacun une couleur caractéristique de leur état épigénomique. Ce modèle a pour avantage de s'inscrire dans le cadre de la physique des polymères - qui permet de reproduire bon nombre de propriétés de la chromatine à grande échelle - tout en intégrant des interactions biologiques pertinentes. La particularité du modèle est en effet que les monomères de même état épigénomique interagissent préférentiellement. Nous avons présenté trois approches permettant la description de la dynamique de la conformation 3D d'une telle chaîne. La première approche, la plus simple à mettre en oeuvre, consiste à réaliser une approximation « gaussienne auto-cohérente » permettant l'intégration numérique de l'équation de la dynamique portant sur les distances quadratiques moyennes entre monomères de la chaine. La seconde approche repose sur l'étude de la dynamique sur réseau par des simulations de Monte-Carlo cinétique et la dernière, celle avec le moins d'approximations, consiste en la réalisation de simulations de dynamique moléculaire. Nous résumons dans le tableau 4.3 les différences entre ces trois approches.

Les trois approches ont chacune leur limites, il est donc intéressant de les utiliser de manière complémentaire. Avec l'approximation gaussienne auto-cohérente introduite initialement dans Jost et al., 2014, on peut simuler une carte de contact à partir d'une séquence épigénomique assez rapidement et de façon raisonnable. En effet, même si on a étudié des petits bouts équilibrés dans le cadre de cette approche, on a intégré les effets d'autres chaînes et les effets de crumpling avec l'introduction d'un champ générique. Les résultats apportés avec cette approche peuvent être valider et détailler en utilisant les deux autres approches.

Sur réseau, on a présenté l'étude de petits bouts équilibrés par simulations de Monte Carlo cinétique (méthode développé dans Olarte-Plata et al., 2016). On ne s'est donc pas intéressé aux effets dus au caractère hors équilibre. Toutefois, le formalisme n'empeche pas l'étude de crumpling en travaillant avec plusieurs segments plus grands. Pour les petits bouts que nous avons étudié, nous avons tenu compte de l'environnement, de la densité, avec des conditions aux bords périodiques.

Enfin, grâce aux simulations de dynamique moléculaire réalisées par Pascal Carrivain, on a pu s'intéresser à la dynamique de repliement du copolymère. Nous avons particulièrement appris que l'état stationnaire local était très vite atteint chez la drosophile (moins de 20 minutes à 200kb et environ 2h à 1Mb) ce qui est un point important car il permet de valider la première approche développée dans ce chapitre. Aussi, l'analyse statistique des simulations réalisées à l'échelle du génome comparées aux données expérimentales obtenues chez la drosophile

	Approximation gaussienne auto-cohérente	Dynamique sur réseau	Dynamique moléculaire
Cadre d'application	Molécule isolée de petite taille, de 1 à 5Mb	D'une molécule isolée de 1Mb au génome entier	D'une molécule isolée de 1Mb au génome entier
Observables	Propriétés d'ensemble de la chaîne	Trajectoires individuelles de chaque monomère	Trajectoires individuelles de chaque monomère
Connectivité	Interactions de type ressort	Contrainte d'être plus proches voisins sur le réseau	Capsules jointes ponctuellement avec un angle relatif libre
Volume exclu	Potentiel de type Lennard-Jones tronqué ou gaussien	Site occupé par au plus deux monomères	Potentiel de Morse tronqué
Confinement	Potentiel gaussien	Boite cubique avec conditions aux bords périodiques	Sphère avec collisions sur la surface
Interactions entre monomère de même état épigénomique	Potentiel gaussien	Énergie de contact limitée aux plus proches voisins	Superposition de potentiels gaussiens
Dynamique	Équation de Langevin auto-cohérente	Monte Carlo cinétique	Équation de Langevin
Temps de calcul	Typiquement 10s pour 1Mb, 100s pour 5Mb	10^{-8} cpumin/min/pb	4×10^{-8} cpumin/min/pb
Défauts principaux	Approximations	Coût en temps de calcul.	Coût en temps de calcul. Sensible à la condition initiale
Avantages principaux	Facile à mettre en oeuvre	Compromis entre dynamique et approximation	Fidélité à la réalité

TABLE 4.3 – Résumé des différences entre les trois types d'approche permettant la modélisation de la chromatine par un copolymère par bloc.

montre de fortes corrélations.
Tous les résultats obtenus avec ces trois approches sont donc qualitativement identiques. Ces résultats mis bout-à bout prouvent que le repliement 3D de l'épigénome peut expliquer le repliement 3D adopté par la chromatine.
Dans le chapitre suivant nous allons nous concentrer sur l'approche gaussienne auto-cohérente. Le but du modèle étant d'être prédictif, nous allons chercher à en inférer les paramètres.

CHAPITRE 5

INFÉRENCE DES POTENTIELS D'INTERACTION ENTRE MONOMÈRES À PARTIR DE CARTES HI-C

À partir du modèle de copolymère par bloc et de l'approximation gaussienne auto-cohérente décrits dans le chapitre précédent, il est possible de simuler des cartes Hi-C dépendant des paramètres d'interaction entre monomères et donc possible d'élaborer un diagramme de phase du copolymère (Fig. 4.2.32). Dans ce chapitre, on se propose de développer des méthodes permettant l'inférence des paramètres d'interaction d'un modèle copolymère afin d'obtenir un accord quantitatif avec les données expérimentales. Nous allons commencer par décrire la manière la plus naturelle permettant d'obtenir la matrices des interactions, elle consiste en une inversion des équations mathématiques du modèle décrit dans le chapitre 4 et a donc pour avantage de ne nécessiter aucune simulation. Nous verrons qu'en revanche cette méthode est fortement sensible au bruit rendant impossible l'inférence, et ce, malgré la diminution du conditionnement du problème en travaillant à l'échelle des TADs plutôt qu'à l'échelle des monomères. Nous verrons alors trois autres méthodes visant à calculer itérativement les coefficients de la matrice d'interaction en égalant les nombres de contact prédits et simulés, elles seront respectivement basées sur une résolution faisant appel à des courbes d'étalonnage, une dichotomie et enfin une inversion de Boltzmann itérative.

5.1 Motivation

Afin de modéliser le repliement de la chromatine, nous avons développé dans le chapitre 4 un modèle de copolymère par bloc dont la dynamique est résolue via une approximation gaussienne auto-cohérente. On choisi dans ce chapitre de modéliser la chromatine avec l'hamiltonien H_{2B} (Éq. *4.37*) car parmi les quatre hamiltoniens présentés dans le chapitre 4, H_{2B} est celui qui reproduit le mieux la loi $P_c \sim s^{-1}$ (observée expérimentalement dans le cas des longs polymères équilibrés) tout en ayant des paramètres d'interaction d'ordre de grandeur raisonnable. Cet hamiltonien contient un terme d'interaction type ressort représentant la connectivité de la chaîne, un terme de volume exclu modélisé par un potentiel gaussien, un terme générique reproduisant de façon simple les effets de l'environnement et de crumpling et un terme d'interactions entre monomères (ces interactions peuvent être spécifiques entre les monomères de même état épigénomique ou non spécifiques). Les paramètres des deux premiers termes sont fixés par le modèle. Les paramètres des deux derniers termes dépendent, en revanche, de la région chromosomique que l'on souhaite étudiée.

Après avoir présenté le formalisme lié à ce problème direct, nous avons pu élaborer des diagrammes de phase qui décrivent comment les paramètres associés à l'hamiltonien du système se traduisent en terme de carte de contact. En particulier, nous avons montré que ce cadre associé à certains jeux de paramètres permettent de produire des cartes de contact qui ressemblent fortement aux cartes expérimentales (Fig. 4.2.34).

Ainsi, maintenant, afin d'être prédictif, nous allons chercher à résoudre le problème inverse, c'est-à-dire que nous allons chercher à inférer les paramètres du modèle qui reproduisent le mieux les cartes de contact obtenues expérimentalement, justement à partir de ces dernières. Cette résolution peut se faire de manière analytique ou avec des simulations numériques. Nous commencerons avec la méthode analytique qui a pour avantage d'être plus directe mais nous verrons que comme nous avons à faire un problème mal posé au sens mathématique du terme, la résolution mathématique est rendue difficile. Afin de simplifier le problème d'origine, on cherchera à inferer les potentiels entre TADs plutôt que les potentiels entre monomères. Ce changement d'échelle ne suffira pas à transformer notre problème en un problème bien posé donc nous verrons deux autres méthodes d'inférence locale basées respectivement sur l'utilisation de courbes d'étalonnage et sur une résolution d'équation par dichotomie. Ces deux dernières méthodes ont pour point commun de nécessiter des simulations et de reposer sur l'égalisation locale des nombres de contacts expérimentaux et prédits. Enfin on présentera une inversion de Boltzmann itérative.

N.B. : Toutes les cartes de contact présentées dans ce chapitre seront issues de Sexton et al.,

2012.

5.2 Inférence des potentiels d'interaction entre monomères par méthode directe

Dans cette section, notre but est d'inférer directement par le calcul et à partir des cartes de contact expérimentales obtenues avec la technique de Hi-C, les paramètres de l'hamiltonien H_{2B} nécessaires pour la modélisation de la chromatine dans le cadre de l'approximation gaussienne auto-cohérente. On rappelle l'expression de H_{2B} ci-dessous :

$$H_{2B} = \frac{k}{2}\sum_{i=2}^{N}(X_i - X_{i-1})^2 + \sum_{i<j} U_{ve} \times exp\left(\frac{-r_{ij}^2}{2r_e^2}\right) + \sum_{i<j}\left(U_{g_{ij}} + U_{ns_{ij}} + U_{s_{ij}}\right) \times exp\left(\frac{-r_{ij}^2}{2r_0^2}\right)$$

Les deux premiers termes de cet hamiltonien sont constitués de paramètres fixes (le terme d'interaction type ressort entre monomère et le terme de volume exclu). En revanche, le dernier terme dépend des paramètres d'interaction entre monomères $U_{ij} = U_{g_{ij}} + U_{ns_{ij}} + U_{s_{ij}}$ qui englobe respectivement les effets génériques via les coefficients $U_{g_{ij}}$, les interactions directes entre monomères (éventuellement dûes à l'épigénétique) via $U_{s_{ij}}$ et les corrections à apporter au champ générique, $U_{ns_{ij}}$ (qui on pour but de maintenir une loi pour la probabilité de contact moyenne du type $P_c \sim s^{-1}$ malgré l'ajout d'interactions spécifiques). On note $\Delta U_{ij} = U_{ns_{ij}} + U_{s_{ij}}$ puisque ces interactions constituent l'écart par rapport au champ générique de base. Ces paramètres $U_{ij} = U_{g_{ij}} + \Delta U_{ij}$ dépendent du segment chromosomique et ne peuvent donc pas être fixés, ils doivent donc être inféré au cas par cas. On commencera dans cette partie par présenter les équations mathématiques permettant de calculer la matrice U des interactions totales entre monomères. Ensuite, on appliquera ces équations mathématiques pour trouver les coefficients $U_{g_{ij}}$ dans un premier temps puis les ΔU_{ij} ensuite. On sait que les coefficients ΔU_{ij} sont composés d'une contribution spécifique entre monomères de même état et d'une contribution non spécifique,$\Delta U_{ij} = U_{ns_{ij}} + U_{s_{ij}}$, mais on cherchera à inférer la matrice ΔU sans distinguer l'origine de l'interaction (spécifique ou non spécifique).

5.2.1 Principe de la méthode : Inversion des équations mathématiques

Nous présentons dans cette partie la méthode la plus naturelle pour inférer les paramètres d'interaction entre monomères, à savoir, l'inversion directe du problème. En effet, notre modèle décrit dans le chapitre 4 permet de simuler des cartes de contact étant donné des para-

mètres d'interaction entre monomères. Ci-dessous, nous allons présenter les équations mathématiques nécessaire pour faire l'inverse. Pour cela, on réalise un « chemin » commençant avec la matrice Hi-C expérimentale, $C_{expérimentale}$, et passant ensuite par le calcul de matrices intermédiaires résumées dans la table 5.1.

La première étape de cette méthode est de convertir la matrice de contact C, en matrice de distance D. Précédemment, nous avons vu que ces deux matrices, C et D, étaient liées par la relation $C = AD^{-3/2}$ (4.13).

Dans cette section, comme le paramètre A impliqué dans cette équation n'est qu'une constante, on choisi de le calculer en imposant arbitrairement que la distance moyenne entre deux monomères plus proches voisins soit de 1. Cette condition implique que pour tout monomère i, $D_{ii+1} = 1/3$ puisque pour rappel, $D_{ij} = \left\langle (X_i - X_j)^2 \right\rangle /3)$. La constante est donc $A = \frac{1}{(N-1)} \times \frac{\sum_{i,j}(C_{|i-j|=1})}{(1/3)^{-3/2}}$.

À partir de la matrice D, la déduction de l'inverse du potentiel gaussien K^{-1} à partir de la matrice D repose sur le système 4.7. Ce système ne permet pas d'obtenir la matrice totale des potentiels gaussiens $\mathcal{K}$ directement. En effet, la matrice $\mathcal{K}$ contient une valeur propre nulle décrivant le degré de liberté de translation de la chaîne, ce qui la rend non inversible. C'est pourquoi nous calculons la matrice intermédiaire K^{-1} qui possède une dimension en moins en raison de la perte d'un degré de liberté. Cette matrice est donc de taille $N-1$ avec N le nombre de monomères formant la chaîne et elle ne possède pas de valeur propre nulle. Dans un premier temps, on choisit de supprimer le degré de liberté de la chaîne en imposant la position du premier monomère : $X_1 = 0$ (car c'est la contrainte qui allège le plus les calculs). La matrice K^{-1} ainsi contrainte sera notée K_1^{-1} (l'indice 1 rapellant que la contrainte est fixée sur le premier monomère). On utilise donc le système 4.8 (qui est un cas particulier du système 4.7) pour calculer cette matrice K_1^{-1}. Après inversion de cette dernière, on obtient la matrice K qui permet directement la détermination de la matrice des potentiels $\mathcal{K}$ (Éq. 4.4). Enfin on utilise la relation à l'équilibre $\langle J \rangle = -k_B T \mathcal{K}$ (Éq. 4.19) et on en déduit la matrice des potentiels d'interaction entre monomères U avec la relation 4.40. Cette matrice U contient trois contributions, $U = U_{générique} + U_{ns} + U_s$. Nous allons présenter ci dessous deux sous sections distinctes permettant de calculer séparément les termes $U_{générique}$ et $\Delta U = U_{ns} + U_s$.

5.2.2 Inférence du champ générique entre monomères, $U_{g_{ij}}$

Dans cette section, on cherche à inférer le champ générique $U_{générique}$ modélisant les effets de l'environnement et les effets de crumpling dû au caractère éventuellement hors équilibre de la chaîne. Pour déterminer ce champ, à partir d'une carte de contact expérimentale, on

Donnée d'entrée	$C_{experimentale}$
Étape 1	Calcul de D (Éq. 4.13)
Étape 2	Calcul de K^{-1} (Éq. 4.7)
Étape 3	Inversion de K^{-1} pour obtenir K
Étape 4	Calcul de $\mathcal{K}$ (Éq. 4.4)
Étape 5	Calcul de $\langle J \rangle$ (Éq. 4.19)
Étape 6	Calcul de U (Éq. 4.40)

TABLE 5.1 – **Méthode d'inférence par inversion directe du problème.** Étapes successives permettant d'obtenir la carte des potentiels d'interaction entre monomères, U à partir d'une carte de contact, C. Ce chemin a pour avantage principal de ne nécessiter aucune simulation. Il est valable dans le cadre de l'approximation gaussienne auto-cohérente et il est valable dans le cas de l'équilibre (à cause de la relation à l'équilibre 4.19).

commence par supprimer sur les cartes de contact expérimentales les effets engendrés par les interactions entre monomères. On réalise cela en moyennant le nombre de contacts par diagonale, c'est-à-dire le nombre de contacts à la même distance génomique (Éq. 5.1).

$$C_{générique_{ij}} = \frac{1}{N - |i-j|} \times \sum_{k \leqslant l} C_{expérimentale_{|k-l|=|i-j|}} \tag{5.1}$$

La carte de contact moyennée, que l'on note $C_{générique}$ ne présente plus de structures 3D caractéristiques des interactions entre monomères (Fig. 5.2.1). En revanche, les effets génériques sont toujours présents sur cette carte.

On peut voir sur la figure 5.2.2 que dans le cas de la drosophile, pour des échelles comprises entre 50kb et 1Mb, le nombre de contacts moyenné présente, en échelle $\log_{10}$, une décroissance avec une pente proche de -1, en fonction de la distance génomique. Cette région est justement celle qui nous intéresse puisque nous cherchons avec l'approximation gaussienne auto-cohérente à étudier les conformations de segments chromosomiques isolés de taille environ égale à 1Mb. Afin de lisser la probabilité de contact en fonction de s, on réalise un ajustement avec la fonction suivante :

$$C^{générique\,ajusté}_{|i-j|} = e^{3.9 - 2.715 \times th\left(\frac{\log(|i-j|) - 3.075}{2.57}\right)} \tag{5.2}$$

À partir de $C^{générique\,ajusté}$ nous pouvons en déduire $U_{générique}$ en utilisant le schéma présenté dans le tableau 5.1 (même méthode que celle utilisée dans la partie 4.2.3.3). Le champ générique obtenu dans le cas de la drosophile est présenté sur la figure 5.2.3. On peut clairement voir sur cette figure l'intérêt de l'ajustement (Éq. 5.2). En effet, il semble que la méthode d'inversion que nous utilisons soit sensible au bruit.

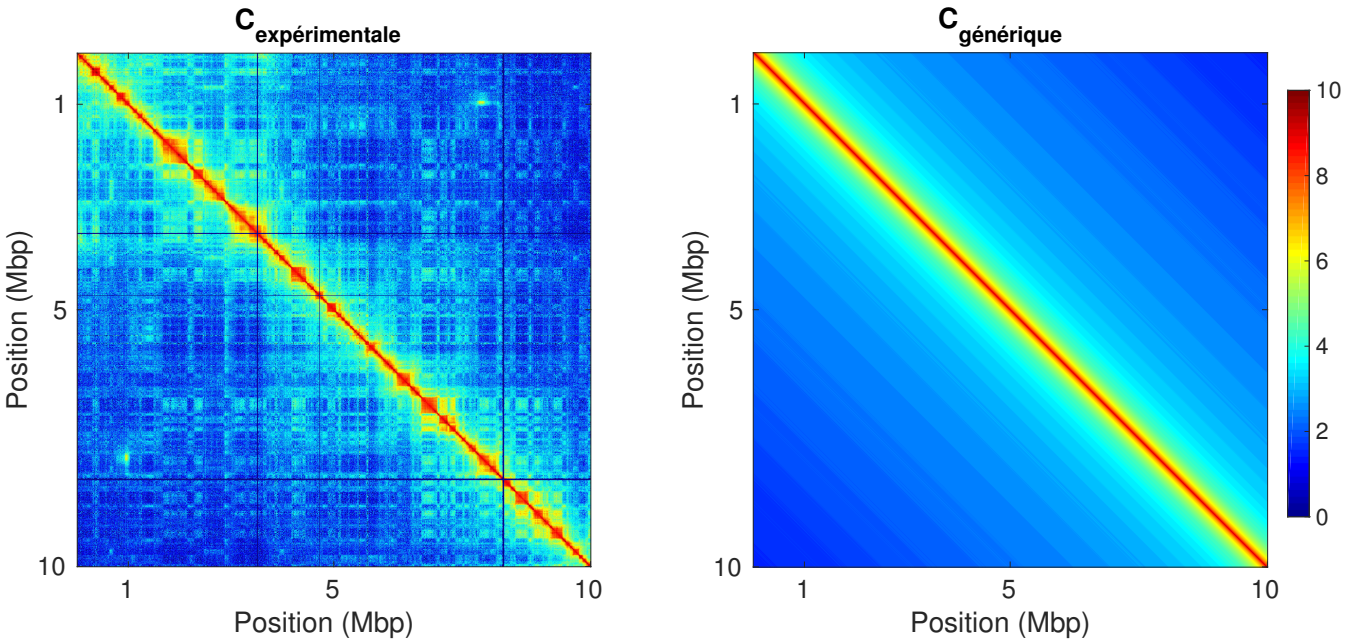

Figure 5.2.1 – **Exemple d'une carte de contact expérimentale $C_{expérimentale}$ (à gauche) et de la carte $C_{générique}$ correspondante.** Ces contacts concernent la région entre 1 et 10Mb du chromosome 3R de la drosophile [données issues de Sexton et al., 2012] .

5.2.3 Inférence des déviations par rapport au champ générique entre monomères, ΔU_{ij}

On applique maintenant le schéma du tableau 5.1 pour inférer la matrice des interactions entre monomères ΔU à partir d'une carte de contact expérimental $C_{expérimentale}$ et connaissant le champ générique $U_{générique}$. Les deux premières étapes ont été franchies sans problèmes. Par contre, la troisième étape consistant à inverser de la carte des potentiels gaussiens K_1^{-1} s'est révélée être critique du fait qu'elle amplifie dramatiquement les erreurs et compromet donc l'inférence (Fig. 5.2.4).

Pour comprendre d'où vient cette extrême sensibilité aux erreurs, nous avons construit par le calcul la carte Hi-C d'une chaîne gaussienne dont nous avons perturbé la valeur d'un coefficient, C_{ij}. Cela nous a permis de voir que cette erreur sur C_{ij} se répercute sur les coefficients (i,j) (i,i) et (j,j) de la matrice K_1^{-1} ainsi que sur leur plus proche voisins. Selon l'emplacement du coefficient C_{ij} (plus ou moins loin de la diagonale) des interactions fantômes plus ou moins importantes apparaissent (Fig. 5.2.5). Ainsi, des erreurs aléatoirement distribuées sur l'ensemble de la carte Hi-C couplent tout les coefficients de la matrice K_1 d'une manière complexe (Fig. 5.2.4).

Nous savons donc que l'étape d'inversion de la matrice K_1^{-1} amplifie énormément les erreurs.

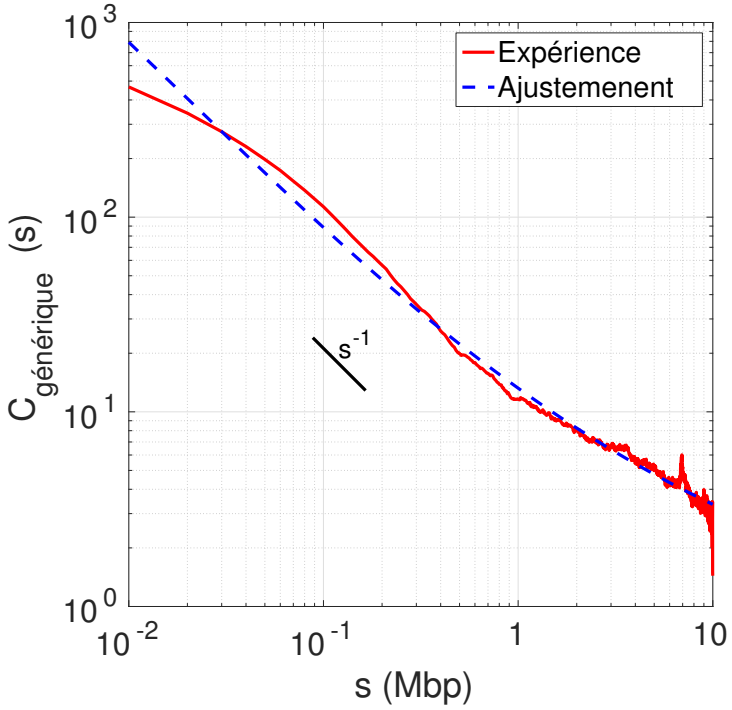

FIGURE 5.2.2 – **Évolution du nombre de contacts intra-chromosomiques moyens** (matrice $C_{générique}$ de la figure 5.2.1) en fonction de la distance génomique entre paires de loci, s, en Mb, du chromosome 3R de la drosophile (rouge). Ajustement (Éq. 5.2) en bleu. Le segment noir indique la pente de la loi d'échelle s^{-1}. On observe sur cette figure en échelle log que chez la drosophile la décroissance du nombre de contacts moyen suit une une loi type $P_c \sim s^{-1}$ entre $50kb$ et $1Mb$. Les comportements en deçà de 50 kb et au delà de 1 Mb peuvent éventuellement s'expliquer respectivement par des biais expérimentaux liés à la technique du Hi-C (par exemple, pontage avec le formaldéhyde) et par un effet de mémoire (équilibration) ou par la présence de boucle à grande échelle ce qui fait qu'à partir d'une certaine distance génomique les distances physiques décroissent moins vite.

Afin de quantifier ce problème, nous nous intéressons à la notion de conditionnement définie ci-dessous :

Soit P un problème $P : \mathbb{R}^n \to \mathbb{R}$ et soit $\hat{x_i}$ une variable perturbée telle que $\hat{x_i} = x_i(1 + \varepsilon_i)$ avec $|\varepsilon_i| < \epsilon$ où ϵ est la précision de la machine. Alors, la condition κ du problème est le plus petit nombre tel que :

$$\frac{|P(\hat{x}) - P(x)|}{|P(x)|} \leqslant \kappa\epsilon + o(\epsilon)$$

À partir de cette notion de condition, on peut définir le conditionnement d'une matrice inversible, par exemple K_1^{-1}, via la relation ci dessous [Higham, 2002] :

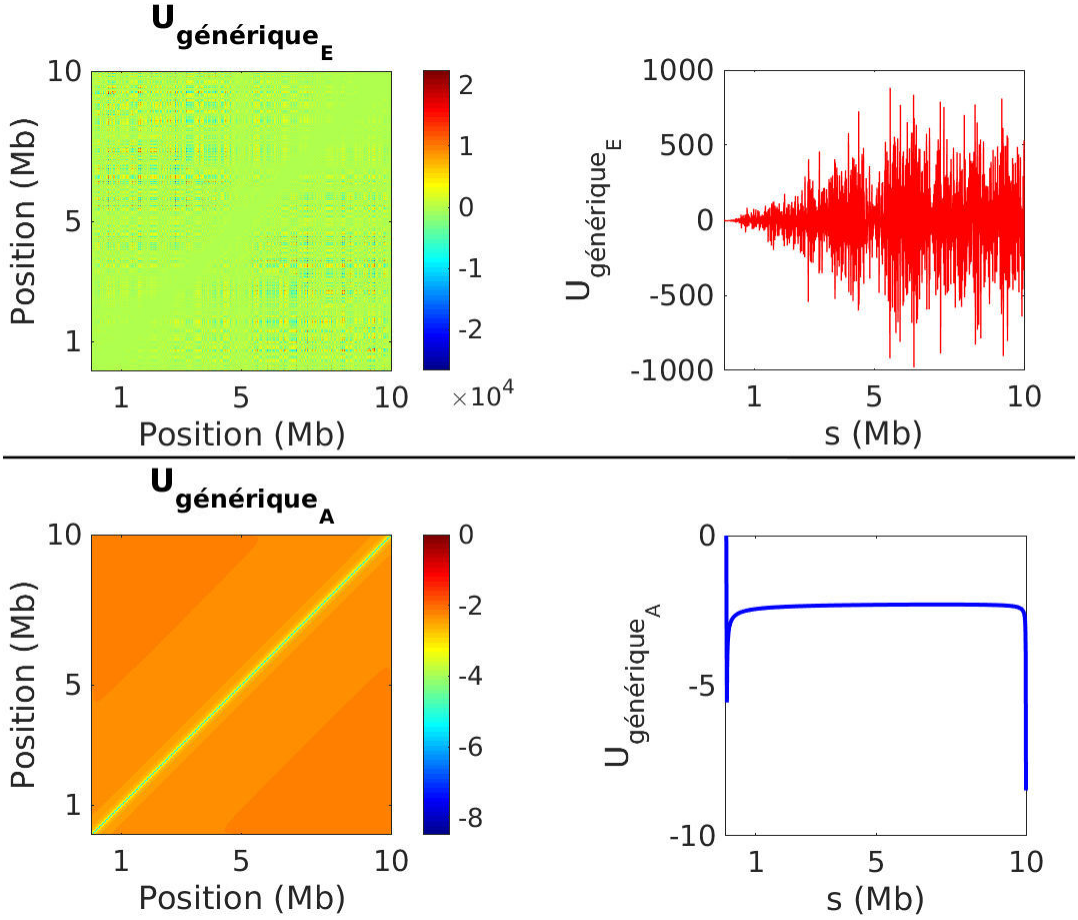

Figure 5.2.3 – **Champ générique calculé à partir de la carte de contact $C_{générique}$ de la figure 5.2.1.** En haut : Inférence sans lissage des données. En bas : inférence après ajustement de $C_{générique}$ avec la formule 5.2.

$$\kappa(K_1^{-1}) = \frac{\sigma_{max}(K_1^{-1})}{\sigma_{min}(K_1^{-1})} \tag{5.3}$$

avec σ_{min} et σ_{max}les valeurs singulières maximales et minimales de la matrice V^{-1}. Des résultats d'analyse numérique [Higham, 2002] permettent alors de majorer les erreurs sur K via la relation suivante :

$$\frac{||\Delta K_1||}{||K_1||} \leqslant \kappa(K_1^{-1}) \frac{||\Delta K_1^{-1}||}{||K_1^{-1}||}$$

Ainsi, le conditionnement, κ est une grandeur simple et rapide à calculer (par exemple avec *Matlab*) et porteuse d'une information intéressante : plus le conditionnement d'une matrice est faible, moins son inversion amplifiera les erreurs. Cette grandeur est en quelque sorte une mesure de la susceptibilité d'une matrice aux erreurs.

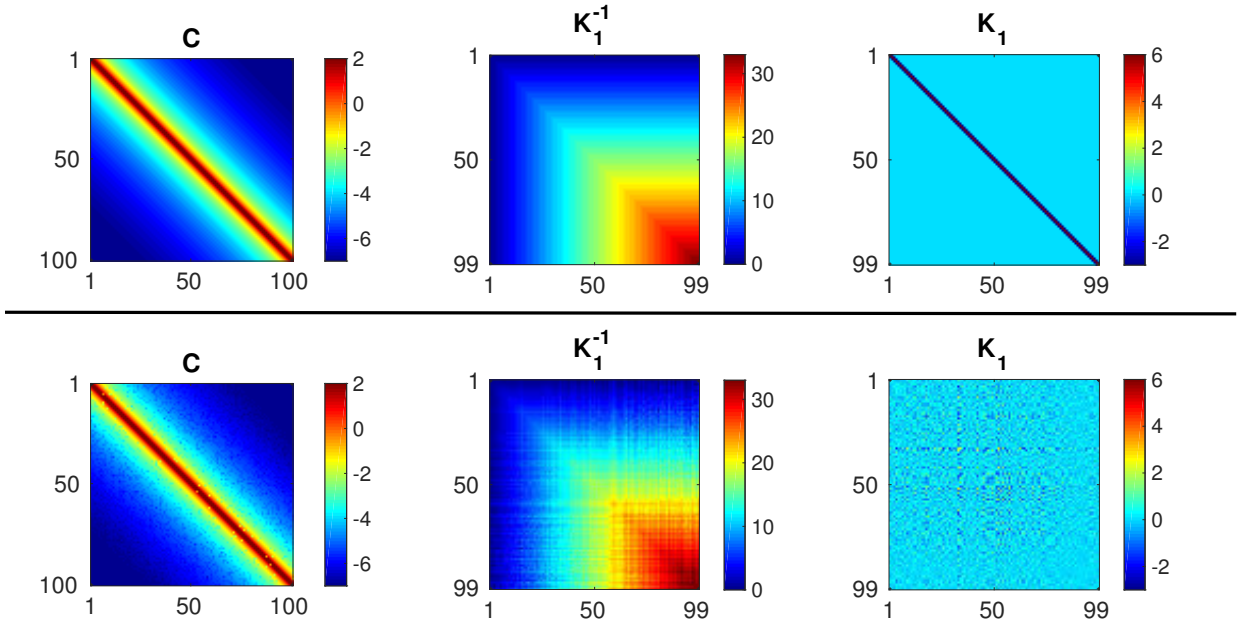

FIGURE 5.2.4 – **Inférence des potentiels d'interaction par inversion des équations mathématiques à partir de la carte de contact d'une chaîne idéale (en haut) et à partir de la même carte sur laquelle on ajoute 5% d'erreur relative sur tout les coefficients (en bas).** On voit que cette méthode est extrêmement sensible au bruit. En effet, il semble que K_1 dans le cas où la carte de contact est bruitée soit une matrice dont le signal est complexe à extraire.

Dans le cas de la figure 5.2.4, $\kappa \approx 10^5$ ce qui implique que l'erreur relative sur les coefficients de la matrice K_1 est au plus multiplié d'un facteur 100 000 par rapport aux erreurs commises sur K_1^{-1} ! L'inversion directe de K_1^{-1} n'est donc pas informative.

Sur la figure 5.2.6, on présente le seuil d'erreur relative critique sur la matrice de contact d'une chaîne gaussienne $\mathsf{C}^{gaussienne}$, noté ϵ, au delà duquel le conditionnement de la matrice K_1^{-1} est « trop élevé » pour que son inversion directe soit informative. Par « trop élevé », on entend que le conditionnement est supérieur à 10 fois celui de la matrice K_1^{-1} dans le cas où celle ci est calculée à partir de la même matrice $\mathsf{C}^{gaussienne}$ non bruitée. Ce seuil dépend de la taille de la matrice : plus elle est grande, plus le seuil est bas. Pour une carte de taille 100 monomères, il s'élève à 0.01% d'erreur relative. Cette valeur est extrêmement faible comparées aux erreurs expérimentales. En effet, par un calcul de déviation standard entre plus proche voisins, on estime que les erreurs relatives sur les cartes Hi-C pour une résolution de 10kb sont de l'ordre de 25% (cf. section 3.1.1).

En résumé, nous avons vu que le conditionnement de la matrice K_1^{-1} est trop élevé pour pouvoir réaliser une inférence satisfaisante. Nous allons donc retourner à l'étape 2 de notre méthode d'inférence (Tab. 5.1) et chercher à savoir si le conditionnement de la matrice K^{-1}

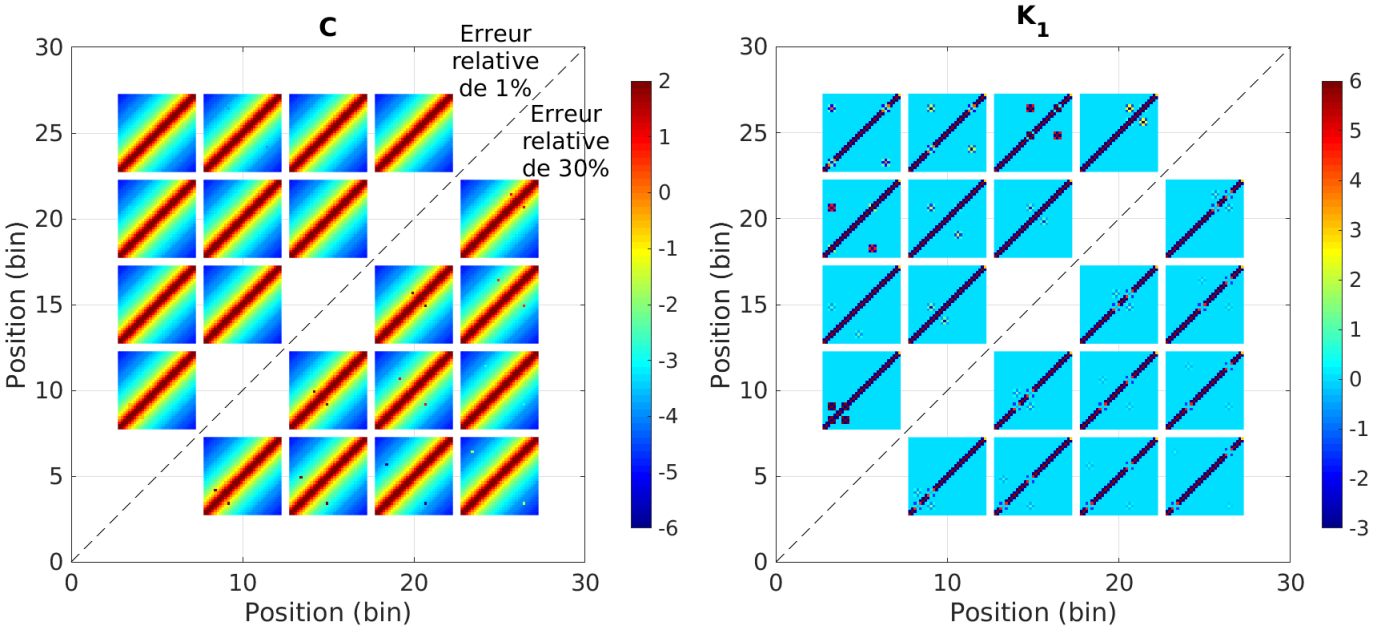

FIGURE 5.2.5 – **Propagation vers la matrice des potentiels, K, d'une erreur relative de 1% ou de 30% appliquée sur un coefficient d'une matrice de contact, C, de taille $\mathsf{N} = 30$.** À gauche : Cartes de contacts avec une erreur relative de 1% ou 30% sur l'un des coefficients. L'abscisse et l'ordonnée auxquelles chaque carte est placée représente les coordonnées du coefficients perturbé. À droite : Matrices des potentiels gaussien K_1 inférées. On voit que selon la position de la perturbation dans la matrice C, la propagation de l'erreur se répercute plus ou moins sur la matrice K_1. Dans tous les cas, l'erreur sur C_{ij} a une conséquence sur les coefficients V_{ii}, V_{jj}, V_{ij} et leur plus proche voisins. Ainsi, des erreurs aléatoirement distribuées sur l'ensemble de la carte Hi-C couplent tout les coefficients de la matrice V d'une manière complexe.

dépend de la contrainte spatiale imposée. En effet, dans toute cette partie, nous avons travaillé avec K_1^{-1}, matrice contrainte par la position du premier monomère, $\mathsf{X}_1 = 0$, mais on peut tout à fait envisager de changer de jauge. Cela ne change pas la physique du problème et on peut espérer qu'une autre jauge mènera à une matrice à inverser avec un conditionnement plus faible que celui de K_1^{-1}.

5.2.3.1 Influence de la jauge sur l'inversion de la matrice K^{-1}

Comme nous l'avons vu dans le 4 la matrice du potentiel gaussien $\mathcal{K}$ est non inversible puisqu'elle contient une valeur propre nulle caractéristique du degré de liberté de translation de la chaîne polymérique. De ce fait, nous avons utilisé la matrice K_1 à qui la contrainte spatiale $\mathsf{X}_1 = 0$ a été imposée. Cette jauge a été choisie dans un premier temps car c'est celle qui facilite le plus les calculs. Toutefois, nous avons vu que l'étape d'inversion de la matrice K_1^{-1} obtenue amplifie dramatiquement les erreurs. Ainsi, dans cette partie on va toujours

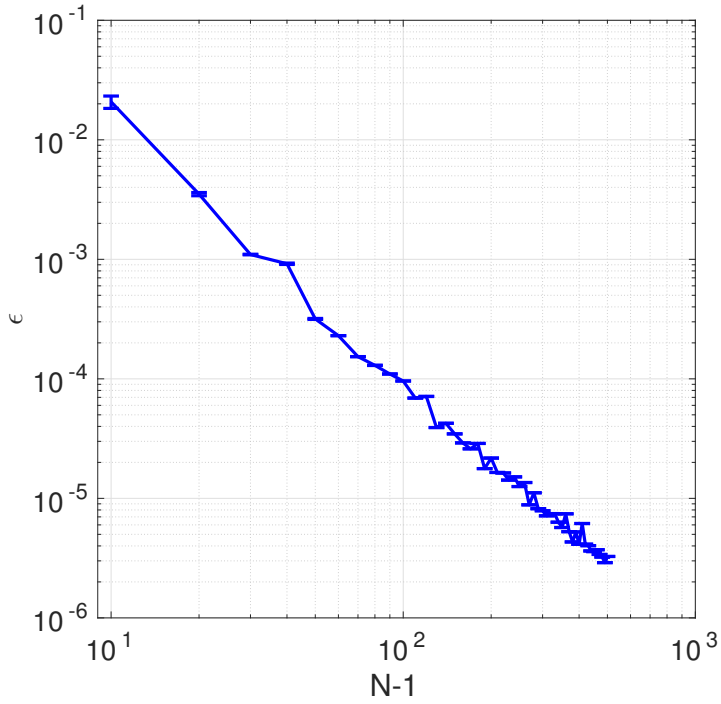

FIGURE 5.2.6 – **Seuil de l'erreur relative critique sur la matrice de contact d'une chaîne gaussiene $C^{gaussienne}$, noté ϵ, en fonction de la taille de la matrice K_1^{-1}, $N-1$.** Au delà du seuil critique ϵ , le conditionnement de la matrice K_1^{-1} est 10 fois plus grand que celui de la matrice K_1^{-1} calculé à partir de la matrice $C^{gaussienne}$ sans erreur.

chercher à développer la méthode résumée sur la figure 5.1 mais en choisissant une autre contrainte spatiale. Toutes les contraintes spatiales sont envisageables. Le cas général peut être ainsi formulé :

$$\sum_{k=1}^{N} a_k X_k = 0 \tag{5.4}$$

avec $a_1 = 1$ et les autres coefficients a_k pour $2 \leqslant k \leqslant N$ libres. On peut remarquer que pour la matrice K_1, tout les coefficients a_k avec $2 \leqslant k \leqslant N$ sont nuls. Afin d'étudier si le choix de la contrainte peut influencer la propagation des erreurs, nous définissons 10 contraintes résumées dans le tableau 5.2.7. Pour chacune de ces contraintes, nous noterons K_c, la matrice des potentiels gaussiens obtenue avec la contrainte numérotée c dans le tableau 5.2.7.

Pour chacune des contraintes c, nous pouvons calculer le conditionnement de la matrice K_c^{-1}. Rappelons que le conditionnement est une grandeur simple et rapide à calculer avec Matlab et porteuse d'une information intéressante : plus le conditionnement d'une matrice est faible, moins son inversion amplifiera les erreurs. Les résultats obtenus sont présentés figure 5.2.8. Ils ont été obtenus à partir de 100 cartes de contact de chaînes gaussiennes générés aléatoirement

Contrainte 1	$\forall k > 1,\ a_k = 0$ (ie $x_1 = 0$)	Contrainte 6	$\forall k > 1,\ 10^{-1} < a_k < 10^{-0}$ (VA)
Contrainte 2	$\forall k > 1,\ 10^{-5} < a_k < 10^{-4}$ (valeurs aléatoires ou VA)	Contrainte 7	$\forall k > 1,\ a_k = 1$ (ie $x_G = 0$)
Contrainte 3	$\forall k > 1,\ 10^{-4} < a_k < 10^{-3}$ (VA)	Contrainte 8	$\forall k > 1,\ 10^{0} < a_k < 10^{1}$ (VA)
Contrainte 4	$\forall k > 1,\ 10^{-3} < a_k < 10^{-2}$ (VA)	Contrainte 9	$\forall k > 1,\ 10^{1} < a_k < 10^{2}$ (VA)
Contrainte 5	$\forall k > 1,\ 10^{-2} < a_k < 10^{-1}$ (VA)	Contrainte 10	$\forall k > 1,\ 10^{2} < a_k < 10^{3}$ (VA)

Figure 5.2.7 – **Résumé des contraintes testées.**

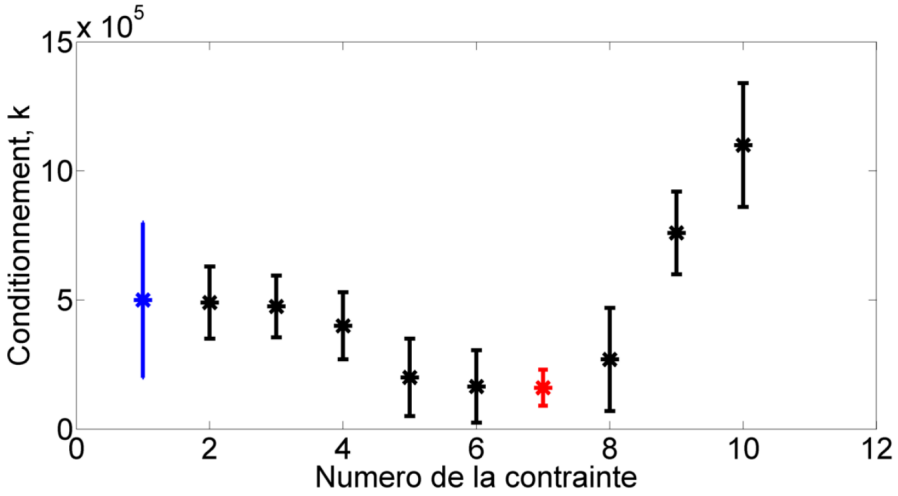

Figure 5.2.8 – **Conditionnement des matrices V_c^{-1} avec c le numéro de la contrainte définie dans le tableau 5.2.7.** La contrainte 7 fixant le centre de gravité à zéro (en rouge) est la contrainte réduisant au minimum le conditionnement de la matrice à inverser, c'est en plus la contrainte ayant le plus de sens physique. Toutefois cette contrainte (et a fortiori les autres) ne permet pas de diminuer assez le conditionnement, l'inférence par cette méthode demeure impossible.

avec des erreurs relatives de l'ordre de 5%. On peut voir que le choix de la contrainte creuse des écarts d'un facteur 10 en terme de conditionnement. La contrainte 7 fixant le centre de gravité de la chaîne à zéro est la contrainte réduisant au minimum le conditionnement de la matrice à inverser. C'est donc cette contrainte spatiale que nous retenons pour la suite. Toutefois, la méthode telle quelle ne permet toujours pas l'inférence du potentiel gaussien puisque le conditionnement de K_7^{-1} est de 10^5. De ce fait, en plus de ce choix de contrainte, nous allons imposer à ce que la matrice K_7^{-1} que l'on notera maintenant $\mathsf{K}_\mathsf{G}^{-1}$ soit définie positive comme le prévoit la théorie.

5.2.3.2 Approximation de K_1^{-1} par une matrice définie positive

D'après la théorie développée dans le 4, on sait que la matrice $\mathsf{K}_\mathsf{G}^{-1}$ est une matrice de covariance et doit donc être définie positive. Cette propriété est (tout naturellement) vérifiée

si on réalise l'inférence à partir d'une carte de contact gaussienne. Par contre, si on introduit des erreurs poissoniennes aléatoirement distribuées sur cette carte alors l'équation 4.7 mènera à une matrice K_G^{-1} dont les valeurs propres ne sont pas toutes positives. Par exemple sur la figure 5.2.4, la matrice K_1^{-1} de la deuxième ligne obtenue à partir d'une carte de contact bruitée avec 5% d'erreur relative a 42 valeurs propres négatives sur 99. Les matrices K_1^{-1} de la figure 5.2.5 calculées à partir de carte de contact avec une erreur sur un coefficient seulement ont entre 0 et 2 valeurs propres négatives selon les cas.
Ces valeurs propres négatives, engendrées par le bruit, n'ont aucun sens physique dans notre étude. Dans un premier temps, on propose donc de mettre ces valeurs propres à zéros. On construit ainsi la matrice semi-définie positive, $K_{G_{sdp}}^{-1}$. Le spectre de cette matrice est celui de la matrice K_1^{-1} à la différence près que les valeurs propres négatives sont remplacées par des zéros. Concrètement $K_{G_{sdp}}^{-1}$ est donné par la formule ci-dessous :

$$K_{G_{sdp}}^{-1} = VDV^{-1} \tag{5.5}$$

avec V la matrice des vecteurs propres de K_G^{-1} et D la matrice diagonale semblable à K_G^{-1} mais dont les coefficients négatifs ont été mis à zéro.
On présente sur la figure 5.2.9, un exemple d'application de cette méthode. On construit la carte de contact d'une chaîne gaussienne composée de 50 monomères. Sa matrice de covariance est $K_{G_{idéal}}^{-1}$. On ajoute à la carte de contact des erreurs relatives de 5% tirées selon une distribution de Poisson. La matrice de covariance associée à cette carte de contact bruitée est K_G^{-1}, son conditionnement est de 10^4. On peut voir sur cette figure que la matrice K_G^{-1} présente effectivement des valeurs propres négatives et que son inversion donne une matrice K_G peu informative. On calcule maintenant $K_{G_{sdp}}^{-1}$ à l'aide la formule 5.5. Le spectre de cette matrice est donc composé de valeurs propres positives ou nulles. Cette matrice a un conditionnement de 10^5, soit 10 fois supérieur à celui de la matrice K_G^{-1}. La matrice obtenue, $K_{G_{sdp}}$obtenue après inversion est donc encore moins informative que la matrice, K_G. Notre problème d'inversion ne peut donc pas être traité en remplaçant naïvement les valeurs propres négatives par des valeurs propres nulles.
Une autre solution pour supprimer les valeurs propres négatives de la matrice K_G^{-1} est de déterminer la matrice définie positive qui lui est la plus proche en terme de norme. On note cette matrice $K_{G_{ppdp}}^{-1}$ (ppdp pour « plus proche définie positive »). Elle peut être évaluée en calculant la matrice X minimisant le problème suivant :

$$\min \left(\sum_{ij} \left(X_{ij} - K_{G_{ij}}^{-1} \right)^2 \right) \Big/ \sigma(X) > 0$$

avec $\sigma(X)$ le spectre de la matrice X recherchée.

Afin de calculer $K^{-1}_{G_{ppdp}}$ on utilise la fonction *fmincon* de *Matlab* qui permet de trouver le minimum d'une fonction multivariée non linéaire contrainte.

On voit sur la figure 5.2.9 que le spectre de la matrice $K^{-1}_{G_{ppdp}}$ ainsi calculée ne comporte en effet que des valeurs propres strictement positives. Le conditionnement de cette matrice est de 100, l'inversion de la matrice $K^{-1}_{G_{ppdp}}$ propage moins les erreurs que l'inversion de K^{-1}_{G}.

Tout les tests que nous avons réalisé avec cette contrainte « plus proche définie positive » démontrent une nette amélioration des résultats, le conditionnement de la matrice à inverser est maintenant de l'ordre de 100 (au lieu de 10 000). En conclusion, l'inférence du potentiel gaussien par inversion du problème donne les meilleurs résultats dans le cas où on impose une contrainte spatiale, $X_G = 0$ et une contrainte mathématique, K^{-1}_G définie positive. Toutefois, cette inférence reste trop approximative comme en témoigne la figure 5.2.9. Les erreurs obtenues sur la matrice K apportent en effet trop d'interactions fantômes, perturbant complètement l'interprétation de ces matrices.

Étant donné la complexité mathématique de ce problème d'inversion, on a choisi de ne plus chercher à inférer les potentiels entre monomères de manière analytique. On va envisager des méthodes d'inférence alternatives évitant d'avoir à inverser la matrice K^{-1}.

5.3 Inférence des potentiels d'interaction intra TAD par inférence bayésienne

Nous avons vu dans les 1 et 3 que les cartes de contact expérimentales laissent apparaître des domaines topologiques appelés TADs à l'intérieur desquelles les monomères se comportent sensiblement de la même façon, c'est-à-dire qu'ils ont des motifs d'interaction similaires. De plus, nous avons vu que dans le cas de la drosophile un TAD contient généralement entre deux et trente monomères de 10 kb. Par conséquent, nous formulons ici l'hypothèse selon laquelle à l'intérieur d'un TAD tout les monomères ont les mêmes paramètres d'interaction. Automatiquement, le fait de considérer le domaine topologique comme unité de base d'étude va réduire le nombre de sites d'interaction ce qui va diminuer le nombre de degré de liberté de notre problème. On commencera par segmenter les cartes de contact en domaines topologiques à l'aide d'IC-Finder présenté dans le 2. Ainsi, au lieu de rechercher $\frac{N(N-1)}{2}$ paramètres avec N le nombre de monomère constituant la chaîne, nous en rechercherons $\frac{N_d(N_d-1)}{2}$ avec N_d le nombre de domaines topologiques présents (Fig. 5.3.1).

Chaque domaine topologique k sera caractérisé par un paramètre d'interaction, noté u_{intra_k}, que l'on va chercher à calculer en considérant que tout les monomères à l'intérieur de ce domaine interagissent de la même manière. Autrement dit, pour tout monomère i dans le

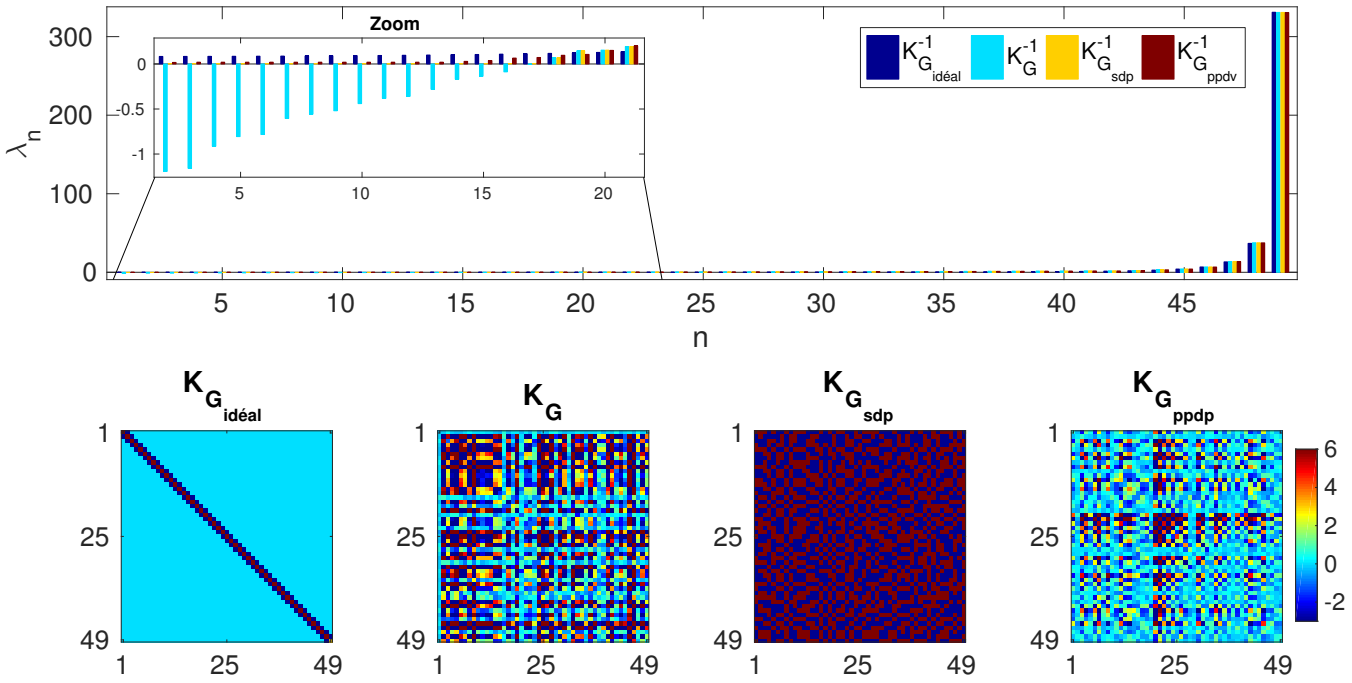

FIGURE 5.2.9 – **Étude du spectre des matrices à inverser.** En haut : Spectre des matrices $K^{-1}_{G_{idéal}}$ (cas chaîne gaussienne idéale de $N = 50$ monomères), K^{-1}_G (cas chaine gaussienne idéale de $N = 50$ monomères avec ajout de 5% d'erreur relative), $K^{-1}_{G_{sdp}}$ (matrice semi-définie positive la plus proche de K^{-1}_G) et $K^{-1}_{G_{ppd}}$ (matrice définie positive la plus proche de K^{-1}_G en terme de norme). Les valeurs propres λ_n avec n le numéro de la valeur propre sont classées par ordre croissant. L'insert en haut à gauche contient un zoom sur les 21 plus petites valeurs propres. En bas : matrices $K_{G_{idéal}}$, K_G, $K_{G_{sdp}}$ et $K_{G_{ppd}}$ dans la même échelle de couleur indiquée à droite.

TAD k, $U_{ii} = u_{intra_k}$. Nous chercherons ensuite les paramètres d'interaction inter domaines topologiques. On notera ce paramètre entre les domaines k et l, $u_{inter_{kl}}$.

N.B. : Jusqu'à la fin du chapitre, nous travaillerons à l'échelle des TADs.

Nous allons développer dans cette section une méthode d'inférence bayésienne assez naïve dans un premier temps en supposant que les paramètres d'interaction intra domaine, u_{intra} ne dépendent pas des interactions entre domaines.

N.B. : Dans toute cette section, la chromatine est modélisée avec l'hamiltonien H_{1A} (Éq. 4.34).

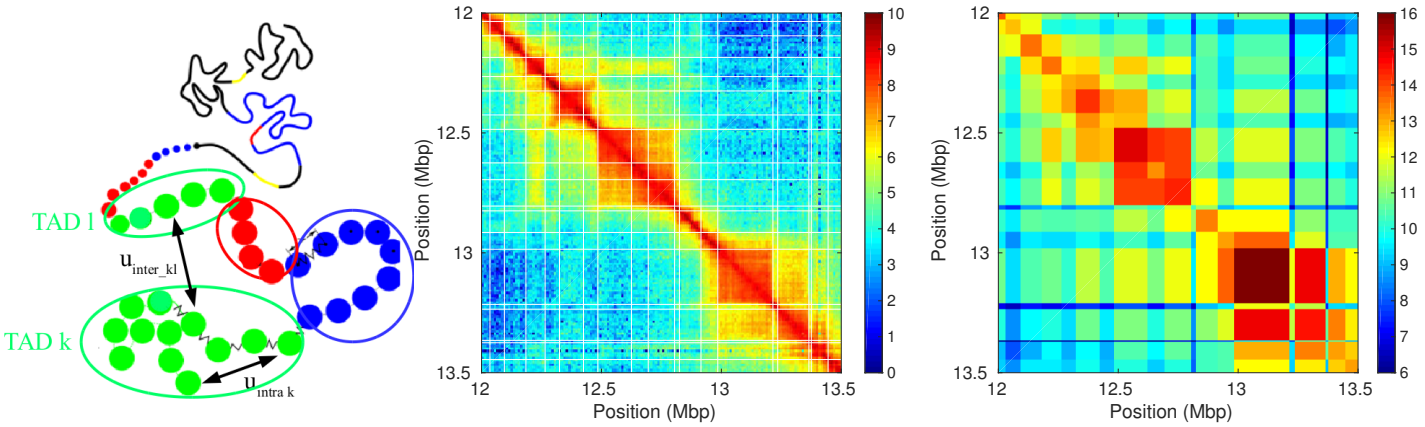

Figure 5.3.1 – **Inférence à l'échelle des TADs sous l'hypothèse qu'à l'intérieur d'un TAD les monomères se comportent sensiblement de la même façon.** (À gauche) Illustration représentant la chromatine. Chaque bille correspond à un monomère de 10kb. Chaque cercle englobe un TAD. À l'intérieur d'un TAD k , tout les monomères interagissent avec l'intensité u_{intra_k}. Entre deux TADs, tout les monomères du TAD k interagissent avec les monomères du TAD l avec l'intensité $u_{inter_{kl}}$. (Au milieu) Carte Hi-C obtenue expérimentalement avec une résolution de 10kb [chromosome 3R de la drosophile Sexton et al., 2012]. Les lignes blanches obtenues avec IC-Finder délimitent les TADs. (À droite) Carte Hi-C identique à la précédente mais telle que dans chaque TAD, le nombre de contacts est sommé. Le fait de considérer le TAD comme unité de base permet de réduire le nombre de degrés de liberté et de réduire le bruit expérimental. Ainsi, on espère faciliter l'inférence sans toutefois obtenir des résultats grossiers.

5.3.1 Principe de la méthode basée sur l'utilisation de courbes d'étalonnage

Nous allons ci-dessous présenter une inférence bayésienne permettant de déterminer les paramètres d'interaction entre monomères compris dans un même TAD, paramètres notés u_{intra} (Fig. 5.3.1). Pour un TAD k, la valeur de u_{intra_k} dépend du nombre total de contacts entre les monomères de ce TAD, c_k. La probabilité a posteriori de u_{intra_k} sachant c_k, notée $P(u_{intra_k}|c_k)$ peut être calculée à partir du théorème de Bayes :

$$P(u_{intra_k}|c_k) = \frac{P(c_k|u_{intra_k})}{P(c_k)} P(u_{intra_k}) \tag{5.6}$$

avec $P(u_{intra_k})$ la probabilité la probabilité a priori de u_{intra_k}et $P(c_k)$ la probabilité a priori de c_k. Le terme $P(c_k|u_{intra_k})$, pour un c_k connu, est appelé la fonction de vraisemblance de u_{intra_k}.

Dans notre cas, pour chaque TAD k dont on connaît le nombre total de contact, c_k, nous allons chercher le paramètre u_{intra_k} qui maximise la probabilité $P(u_{intra_k}|c_k)$. Pour calculer

$P(u_{intra_k}|c_k)$, nous utilisons donc la relation 5.6 en considérant $\frac{P(u_{intra_k})}{P(c_k)}$ comme étant une constante multiplicative. Quant au facteur $P(c_k|u_{intra_k})$, si on suppose que le nombre de contacts c_k suit une loi de Poisson, il peut se réécrire ainsi :

$$\begin{aligned} P(c_k|u_{intra_k}) &= \int dc_k^- P(c_k|c_k^-) P(c_k^-|u_{intra_k}) \\ &= \int dc_k^- P(c_k|c_k^-) \delta(c_k^- - c_k^-(u_{intra_k})) \\ &= P(c_k|c_k^-(u_{intra_k})) \\ &= \frac{c_k^-(u_{intra_k})^{c_k} e^{-c_k^-(u_{intra_k})}}{c_k!} \end{aligned}$$

avec δ la fonction de Dirac et avec la notation $\bar{.}$ pour représenter une grandeur moyenne.
Cette inférence bayésienne des paramètres d'interaction entre monomères dans un TAD k, notés u_{intra_k}, nécessite donc de connaître le nombre de contacts moyen dans un TAD de paramètre u_{intra_k}, $c_k^-(u_{intra_k})$. Nous allons calculer cette grandeur à partir de cartes de contact simulées avec l'approximation gaussienne auto-cohérente (formalisme décrit dans le chapitre 4). Concrètement, on modélise un copolymère ayant un TAD central d'intérêt, k, composé de n_k monomères qui intéragissent entre eux avec une intensité u_{intra_k}. Afin de mimer l'environnement dans lequel ce TAD évolue, on considère qu'il est lié à 5 blocs de chaque côté. Ces derniers sont chacun formés de 8 monomères interagissant entre eux avec une intensité de $-90 k_B T$. En effet, on a vu dans le chapitre 3 qu'en moyenne un TAD de drosophile est composé de 8 monomères et est relativement compact. Afin de modéliser la situation simplement, on suppose que l'inférence des u_{intra_k} ne dépend pas des interactions inter TAD. On n'introduit donc pas d'interactions entre les différents blocs.
Pour chaque copolymère ainsi modélisé, on peut calculer la distance moyenne entre monomères (cf. section 4.2.4) et donc en déduire le nombre de contacts moyen entre monomères. La figure 5.3.2 présente le nombre de contacts moyen dans le TAD central k, c_k^- en fonction des paramètres n_k et u_{intra_k} et dans le cas où le copolymère est modélisé avec l'hamiltonien H_{1A} (Éq. 4.34 page 98).
Les graphes de la figure 5.3.2 donnent le nombre de contacts moyen, c_k^-, obtenu après une expérience. Ainsi, pour pouvoir comparer c_k^- au nombre de contacts expérimental, il faut le multiplier par un facteur N_{exp} donnant le nombre d'expériences effectives. Cette grandeur expérimentale N_{exp} est difficile à estimer. Nous allons donc la fixer en maximisant le nombre de TAD dont le paramètre u_{intra} sera compris entre -100 et $0 k_B T$. La figure 5.3.3 présente pour chaque TAD de la drosophile (classé par taille croissante) en bleu la gamme des valeurs de N_{exp} permettant d'obtenir une intensité d'interaction intra TAD comprise entre -100 et

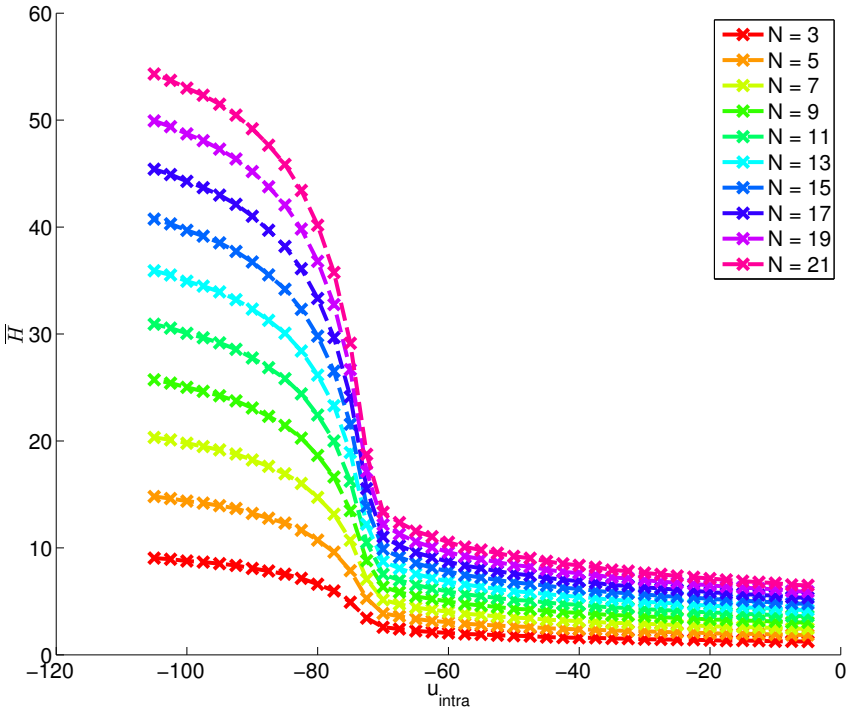

Figure 5.3.2 – **Nombre moyen de contact moyen dans un TAD k, $\bar{c_k}$ en fonction de la taille du domaine n_k et du paramètre d'interaction entre monomère de ce TAD, u_{intra_k}.** Le TAD est modélisé dans le cadre de l'approximation gaussienne auto-cohérente avec l'hamiltonien H_{1A}. Il est lié à 5 blocs de chaque côté. Ces derniers sont formés de 8 monomères interagissant entre eux avec une intensité de $-90 k_B T$.

$0 k_B T$ (car on a vu dans le chapitre 4 que c'est à ces échelles des énergie qu'on reproduit le mieux les cartes de contact expérimentales de drosophile ; Figs. 4.2.20 et 4.2.22). On voit sur cette figure que la valeur de N_{exp} permettant d'optimiser le nombre de domaines ayant une énergie d'interaction intra TAD comprise entre -100 et $0 k_B T$ est de 15000. Il est intéressant de remarquer que cette estimation du nombre d'expériences effectives donne exactement le même résultat que celle réalisée avec une toute autre approche. En effet, nous étions arrivés au même résultat (Éq. équation (4.14)) en faisant appel à l'approximation gaussienne couplée à quelques ordres de grandeur concernant la chromatine (donc sans l'intervention d'un hamiltonien contrairement à ici).

5.3.2 Résultats

Nous avons mis en oeuvre la méthode d'inférence bayésienne des u_{intra} décrite ci-dessus. Nous présentons sur la figure 5.3.4 des diagrammes en boite représentant la distribution des paramètres trouvés en fonction de l'état épigénétique et de la taille du domaine. Nous observons, en accord avec les résultats du chapitre 3, que les domaines où la chromatine

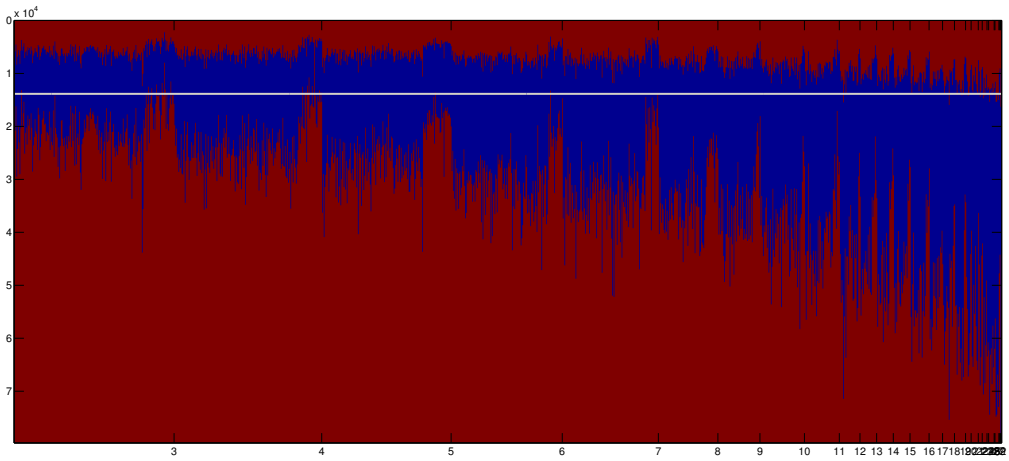

FIGURE 5.3.3 – Gamme de nombre d'expériences effectives, N_{exp}, telle que le paramètre u_{intra} inféré soit compris entre -100 et $0 k_B T$ (en bleu) pour tout les TADs de la drosophile classés par taille croissante (en abscisse). Le nombre N indiqué en abscisse est le nombre de monomères dans le domaine étudié. La ligne blanche signale la valeur de N_{exp} qui minimise le passage par des zones rouges, $N_{exp}^{optimal} \approx 15000$.

est active ont en moyenne des potentiels d'interaction moins élevés en valeur absolue, donc moins attractifs par rapport aux domaines répressifs. Les domaines actifs sont généralement « ouvert » ce qui facilite le recrutement de facteurs de transcription. Au contraire, les domaines répressifs présentent des potentiels d'interaction très négatifs ce qui caractérise des domaines « fermés », où les interactions intra domaines sont fortes. On peut aussi noter que les distributions sont plus larges dans le cas des petits domaines ce qui signifie que les erreurs d'estimations des u_{intra} sont plus fiables pour les grands domaines par rapport aux petits. Par contre, on remarque aussi sur ces diagrammes que les u_{intra} trouvés dépendent de la taille du domaine. Les paramètres u_{intra} contiennent donc une information sur le niveau de compaction du domaine. Pourtant ce ne devrait pas être le cas. Le fait d'avoir négliger les interactions entre les monomères d'un TAD et leur environnement peut expliquer ce problème. L'idée donc de chercher à inférer les u_{intra_k} sans tenir compte de l'environnement dans un premier temps, puis de poursuivre avec l'inférence des u_{inter} ne semble pas pertinente. Nous proposons dans la partie qui suit d'inférer en parallèle les paramètre d'interaction u_{intra} et u_{inter}.

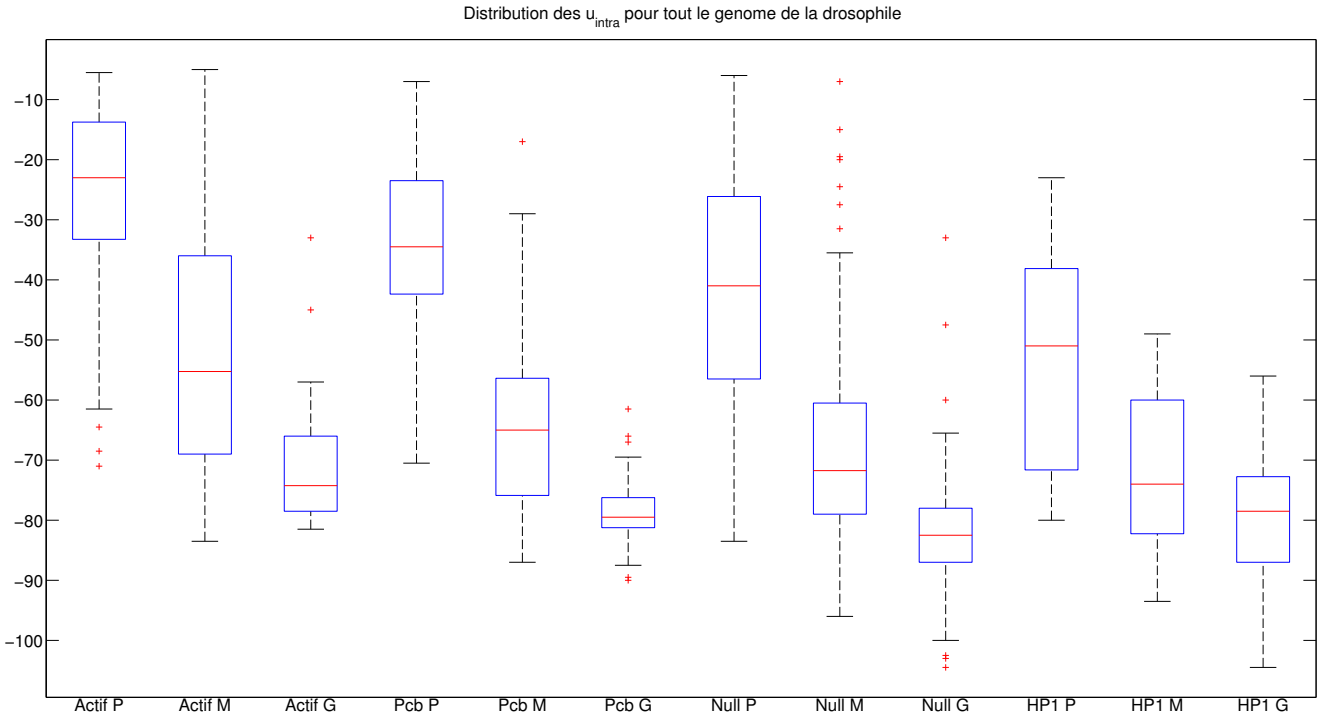

FIGURE 5.3.4 – **Paramètres u_{intra} trouvés sur l'ensemble du génome de la drosophile rangés selon la couleur épigénétique et selon la taille du TAD duquel ils proviennent.** Les initiales P, M et G signifiant petits moyens et grands correspondant à des domaines de taille inférieur à 5 monomères, entre 6 et 12 monomères et enfin de taille supérieur ou égale à 13 monomères. Quel que soit l'état épigénomique, on observe que les plus grands domaines présentent des paramètres d'attraction plus forts et moins dispersés.

5.4 Optimisation locale du nombre de contacts par dichotomie

N.B. : À partir de cette section et jusqu'à la fin du chapitre, la chromatine est modélisé avec l'hamiltonien H_{2B} (Éq. 4.37).

5.4.0.1 Principe de la méthode

Nous allons développer ici une approche fortement semblable à la précédente, qui repose toujours sur une inférence bayésienne, mais en cherchant cette fois à inferer les paramètres d'interaction intra TAD et inter TAD en parallèle. Nous travaillons toujours à l'échelle du TAD, c'est-à-dire sous l'hypothèse que les monomères d'un même TAD ont les mêmes paramètres d'interaction. On notera ces paramètres d'interaction u_{kl} entre deux TAD k et l, avec k pouvant être égal à l.

La méthode qu'on propose ici pour inférer les coefficients u_{kl} de la matrice U, à partir d'une

carte de contact C^{cible}, est une méthode itérative dont le point de départ est une matrice $U^{initial}$ (dont les coefficients sont tous nuls par exemple). À chaque itération, on choisi de façon pseudo aléatoire un compartiment (k, l) pour lequel on optimise l'intensité u_{kl} de sorte que les sommes des contacts cibles et simulés dans le compartiment (k, l) soient égaux (Fig. 5.4.1). L'optimisation du paramètre u_{kl} repose donc sur la résolution de l'équation :

$$n_{kl}(u) = n_{kl}^{cible} \tag{5.7}$$

Afin de calculer $n_{kl}(u)$ on utilisera le formalisme développé avec l'approximation auto-cohérente (cf. section 4.2.4) et nous travaillerons avec l'hamiltonien H_{2B} (Éq. 4.37). La fonction $n_{kl}(u)$ étant a priori continue et monotone (plus l'interaction est élevée en valeur absolue plus le nombre de contacts est élevé), la résolution de l'équation 5.7 peut se faire avec la méthode de Newton-Raphson ou avec une dichotomie [Press et al., 2007]. Dans notre cas, les deux algorithmes donnent les mêmes résultats avec la même précision mais nous travaillerons par dichotomie car nous avons remarqué que de cette manière la convergence est un peu plus rapide. Le principe de la dichotomoie est de rechercher une solution de façon récursive dans un intervalle de plus en plus petit (divisé en deux à charque itération). La première étape consiste donc en la recherche de deux bornes entre lesquelles la solution se trouve. Dans notre cas, une première borne peut être a priori $u_{kl}^{initial} + 0.1$ avec $u_{kl}^{initial}$ le paramètre se trouvant dans la matrice U avant optimisation . Si avec ce paramètre le nombre de contacts est sous estimé (comme c'est le cas sur le panel de gauche de la figure 5.7) (resp. sur estimé), on va rechercher une deuxième borne en soustrayant (resp. ajoutant) une unité à la première borne. Si le nombre de contacts prédit est toujours sous estimé (resp. sur estimé), on soustrait (resp. ajoute) itérativement $0.1 k_B T$ jusqu'à ce que la simulation du nombre de contact, n_{kl} soit supérieure (resp. inférieur) à la cible, n_{kl}^{cible}. Au final, les deux bornes que l'on retient pour réaliser la dichotomie sont les deux paramètres u_{kl}^{inf} et u_{kl}^{sup} permettant de simuler des nombres de contact n_{kl}^{inf} et n_{kl}^{sup} les plus proche de n_{kl}^{cible} et tels que $n_{kl}^{inf} < n_{kl} < n_{kl}^{sup}$.
Sur l'insert dans le panel de gauche de la figure 5.7, on peut voir les 5 étapes successives de la dichotomie. On estime que la solution est atteinte lorsque le nombre de contacts prédit dans le compartiment (k, l) est tel que $\left|n_{kl} - n_{kl}^{cible}\right| < 0.001 n_{kl}^{cible}$.
Le tirage des compartiments (k, l) se fait de manière pseudo-aléatoire afin d'optimiser préférentiellement les paramètres d'interaction u_{kl} les plus importants pour comprendre le repliement de la chaîne. Les compartiments les plus importants sont probablement ceux de la diagonale, c'est-à-dire les TADs. On va donc commencer par tirer uniquement et aléatoirement des compartiments (k, k) jusqu'à ce qu'à avoir visité au moins 3 fois chaque TAD.
Ensuite, on va tirer des compartiments (k, l) (avec k pouvant être égal à l ou pas) en favorisant les plus influents de manière à limiter le temps de convergence. Étant donné qu'il

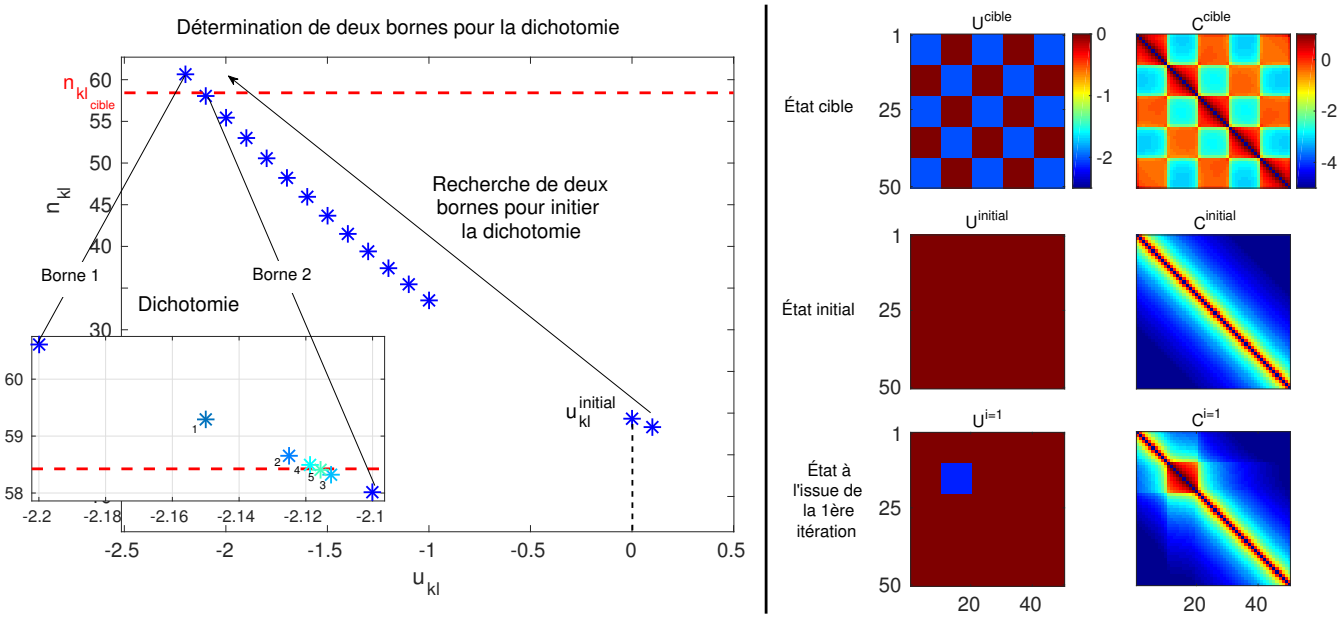

Figure 5.4.1 – **Optimisation d'un paramètre d'interaction u_{kl} entre les monomères d'un TAD k et d'un TAD l.** L'otpimisation est réalisée de telle sorte que le nombre de contact total simulé dans le compartiment (k, l) soit égal à 0.1% près au nombre de contacts cible pour ce même compartiment. (Panel de droite) Pour une carte de contact cible, C^{cible}, dont on cherche à inférer les paramètres d'interaction entre monomères, U^{cible} (première ligne), le principe est de choisir de façon pseudo aléatoire un compartiment (k, l), (sur cette figure $k = 2$ et $l = 2$) et d'optimiser l'intensité des interactions dans ce compartiment, u_{kl}, de sorte que la somme des contacts simulés et cibles dans ce compartiment, respectivement n_{kl} et soit égaux n_{kl}^{cible}. (Panel de gauche) Nombre de contact simulé entre les TAD k et l, n_{kl}, en fonction du paramètre d'interaction u_{kl}. Cette figure vise à expliquer comment se déroule une itération du processus d'inférence. La résolution de l'équation $n_{kl}(u_{kl}) = n_{kl}^{cible}$ se fait par dichotomie. Pour la recherche des deux bornes nécessaires à la dichotomie, le point de départ est $u_{kl}^{initial}$ le paramètre se trouvant dans la matrice U avant optimisation. Dans l'insert, on peut voir les étapes successives de la dichotomie. L'algorithme s'arrête lorsque $\left|n_{kl} - n_{kl}^{cible}\right| < 0.001 n_{kl}^{cible}$.

Facteurs de pondération	Compartiments privilégiés
$f_{1_{kl}} = 1$	Aucun
$f_{2_{kl}} = \#k \times \#l$	Les plus grands
$f_{3_{kl}} = n_{kl}^{cible}$ et $f_{4_{kl}} = \log\left(n_{kl}^{cible}\right)$	Les plus riches en contact
$f_{5_{kl}} = \left\lvert n_{kl}^{cible} - n_{kl} \right\rvert f_{3_{kl}}$ et $f_{6_{kl}} = \left\lvert n_{kl}^{cible} - n_{kl} \right\rvert f_{4_{kl}}$	Les plus riches en contact et moins bien simulés
$f_{7_{kl}} = \frac{n_{kl}^{cible}}{\#k \times \#l}$ et $f_{8_{kl}} = \frac{\log\left(n_{kl}^{cible}\right)}{\#k \times \#l}$	Les plus riches en contact par unité de taille
$f_{9_{kl}} = \left\lvert n_{kl}^{cible} - n_{kl} \right\rvert f_{7_{kl}}$ et $f_{10_{kl}} = \left\lvert n_{kl}^{cible} - n_{kl} \right\rvert f_{8_{kl}}$	Les plus riches en contact par unité de taille et moins bien simulés
$f_{11_{kl}} = \left\lvert \frac{n_{kl} - n_{kl}^{cible}}{n_{kl}^{cible}} \right\rvert$ et $f_{12_{kl}} = \left\lvert \frac{\log(n_{kl}) - \log\left(n_{kl}^{cible}\right)}{\log\left(n_{kl}^{cible}\right)} \right\rvert$ et $f_{13_{kl}} = \left\lvert \frac{\log(n_{kl})}{\log\left(n_{kl}^{cible}\right)} \right\rvert$	Les moins bien simulés

TABLE 5.2 – **Expressions mathématiques des matrices de pondération, f, testées, accompagnées des caractéristiques des compartiments qu'elles privilégient.** Les formules sont données avec $\#k$ le nombre de monomère dans le TAD k et avec n_{kl}^{cible} (resp. n_{kl}) la somme des contact entre les TADs k et l dans la carte de contact cible (resp. carte de contact simulée). La pondération est utilisée pour biaiser le tirage aléatoire du compartiment (k, l) dont le paramètre d'interaction u_{kl} est à optimiser.

n'est pas évident de déterminer quels caractéristiques de compartiments doivent être favorisés (grande taille, nombre de contacts élevé, etc...), on teste treize types de matrices de pondération différentes dont les expressions mathématiques sont données dans le tableau 5.2. On introduit la fonction $\log$ pour certaines matrices f car cela permet d'être sensible au contraste. D'après nos tests, ces treize manières de faire mènent aux mêmes résultats après un grand nombre d'itération mais on remarque que la convergence est la plus rapide en tirant les domaines selon la matrice de pondération f_{10} (Éq. 5.8).

$$f_{10_{kl}} = \left\lvert n_{kl}^{cible} - n_{kl} \right\rvert \times \frac{\log\left(n_{kl}^{cible}\right)}{\#k \times \#l} \tag{5.8}$$

Cette matrice f_{10} favorise les domaines ayant le plus d'interaction en échelle $\log$ par unité de taille (tout comme la matrice f_6, Fig. 5.4.2) et oriente aussi de façon dynamique le tirage vers les compartiments d'interaction dont le nombre de contacts prédit est éloigné du nombre de contacts cible visé. Le caractère dynamique est apporté par le facteur $\left\lvert n_{kl}^{cible} - n_{kl} \right\rvert$ qui évolue à chaque itération.

Ce processus de tirage de compartiment (k, l) et de résolution de l'équation 5.7 est répété jusqu'à ce que au moins tout les compartiments (k, l) aient été visités 3 fois et jusqu'à ce que la matrice de contact simulée C « ressemble » à la matrice C^{cible}. Nous mesurons le degré de

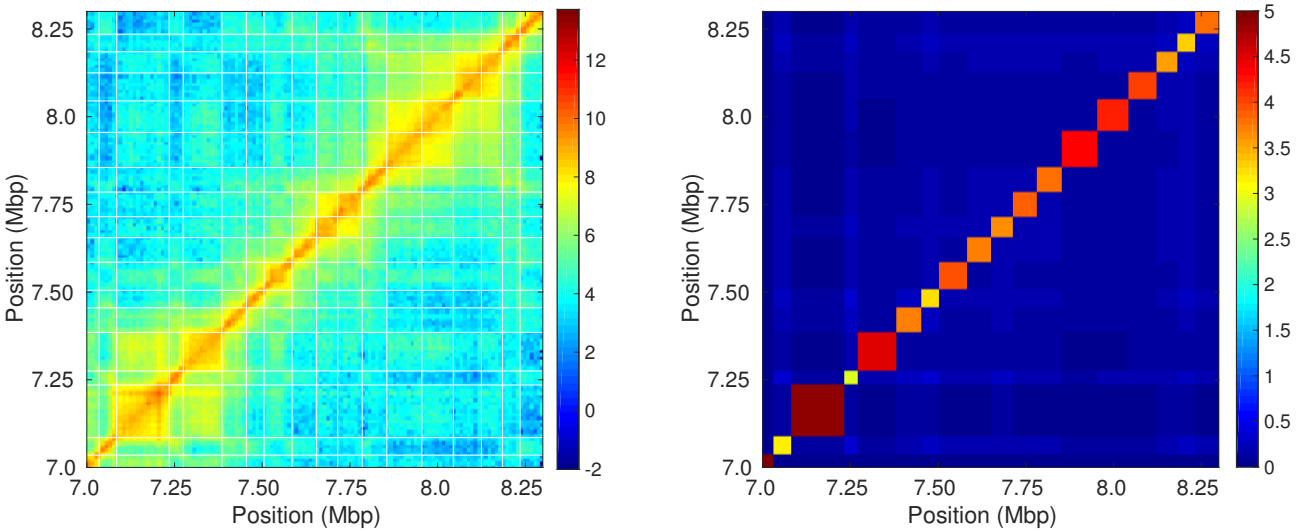

Figure 5.4.2 – **Pondération pour le tirage aléatoire d'un compartiment dont le nombre de contacts est à optimiser.** (À gauche) Carte de contact expérimentale pour une région du chromosome 3R de la drosophile. Les lignes blanches représentent la segmentation en TADs obtenue avec IC-Finder (À droite) Matrice f_6 utilisée pour pondérer le tirage aléatoire des compartiments. Cette matrice favorise les domaines ayant le plus d'interaction en échelle log par unité de taille. Plus le coefficient de pondération pour un compartiment (k, l) est grand, plus son tirage sera fréquent et donc plus le paramètre u_{kl} sera mis à jour.

ressemblance entre ces deux matrices avec un score de type χ^2 de moyenne C_{ij}^{cible} et d'écart type $\sqrt{C_{ij}^{cible}}$ car on suppose que les valeurs C_{ij} sont entachées d'une incertitude suivant une loi de Poisson. On peut remarquer que bien que l'inférence soit réalisée à l'échelle des compartiments $(k,\ l)$, le calcul du χ^2 se fait sur les cartes de contact à l'échelle des monomères (i, j). On estime que les matrices C et C^{cible} sont proche si le χ^2 défini ci-dessous passe en dessous du seuil de 1 :

$$\chi^2 = \sum_{i \leqslant j} \left(\frac{C_{ij} - C_{ij}^{cible}}{\sqrt{C_{ij}^{cible}}} \right)^2 \tag{5.9}$$

Cette méthode d'inférence a été implémentée en C afin d'optimiser le temps de calcul.
NB : Cette méthode n'a de sens que si les paramètres recherchés sont hors zone de multistabilité (Fig. 4.2.31).

5.4.0.2 Application de la méthode au copolymère $(A_{10}B_{10})_6$

Dans cette sous section, on va construire un copolymère par bloc avec le formalisme développé dans le chapitre 4 et on va chercher à savoir si on est capable de retrouver les paramètres choisis. On considère un copolymère $(A_{10}B_{10})_6$ de 6 blocs A et 6 blocs B, chacun comportant

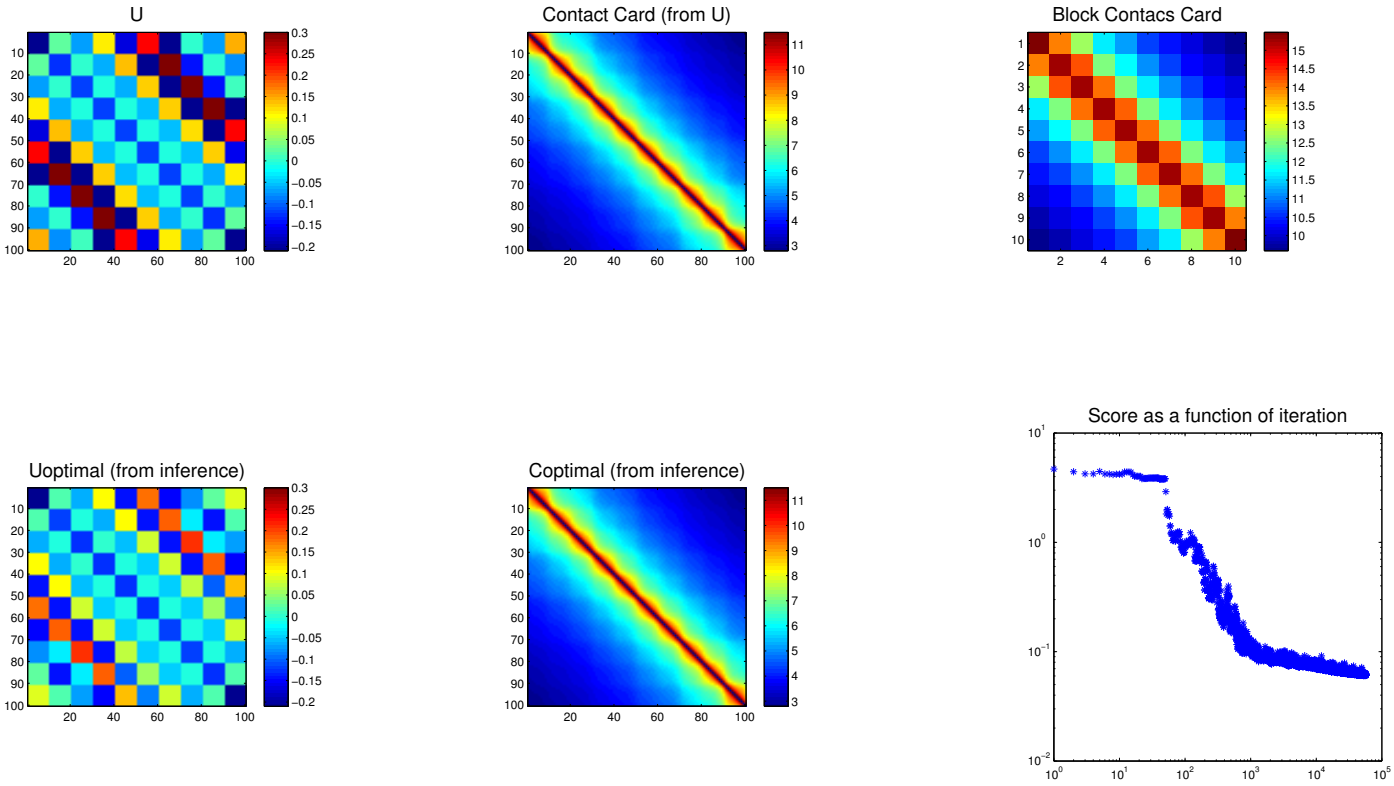

FIGURE 5.4.3 – **Inférence dans le cas d'un copolymère $(A_{10}B_{10})_6$ de paramètres $u_s = -0.1k_BT$ et $u_{ns} = 0.1k_bT$.** (En haut) On observe de la gauche vers la droite la matrice des interactions totales (comprenant le champ générique), la carte de contact qui sert de point de départ pour l'inférence et complètement à droite on voit la carte par bloc (celle utilisée pour l'optimisation par IBI). (En bas) Dans les deux premières colonnes, données identiques à au dessus mais dans le cas inféré, complètement à droite on observe l'évolution du χ^2 en fonction de l'itération.

10 monomères. On le modélise avec l'hamiltonien H_{2B} (Éq. 4.36) et on choisi comme couples de paramètres d'interaction $(u_s, u_{ns}) = (-0.1, -0.1)$ et $(u_s, u_{ns}) = (-0.5, -0.5)$ (Figs. 5.4.3 et 5.4.4). Sur le diagramme de phase élaboré dans le chapitre précédent, on peut voir que ces copolymères se situe dans la région intermédiaire mais hors zone de multistabilité (Fig. 4.2.31).

Dans les deux cas, la convergence est excellente. Il est intéressant de remarquer que l'optimisation se fait sur les cartes dont les contacts sont sommés par bloc (carte en haut à droite des deux figures) mais l'inférence permet de retrouver la structure polymérique.

Comme on voit que l'on arrive à inférer les paramètres mais que la convergence est trop lente, on va développer une inversion de Boltzmann itérative comme nous l'a suggéré Ralf Everaers. Cette méthode va permettre de s'affranchir de la dichotomie et sera donc beaucoup plus rapide.

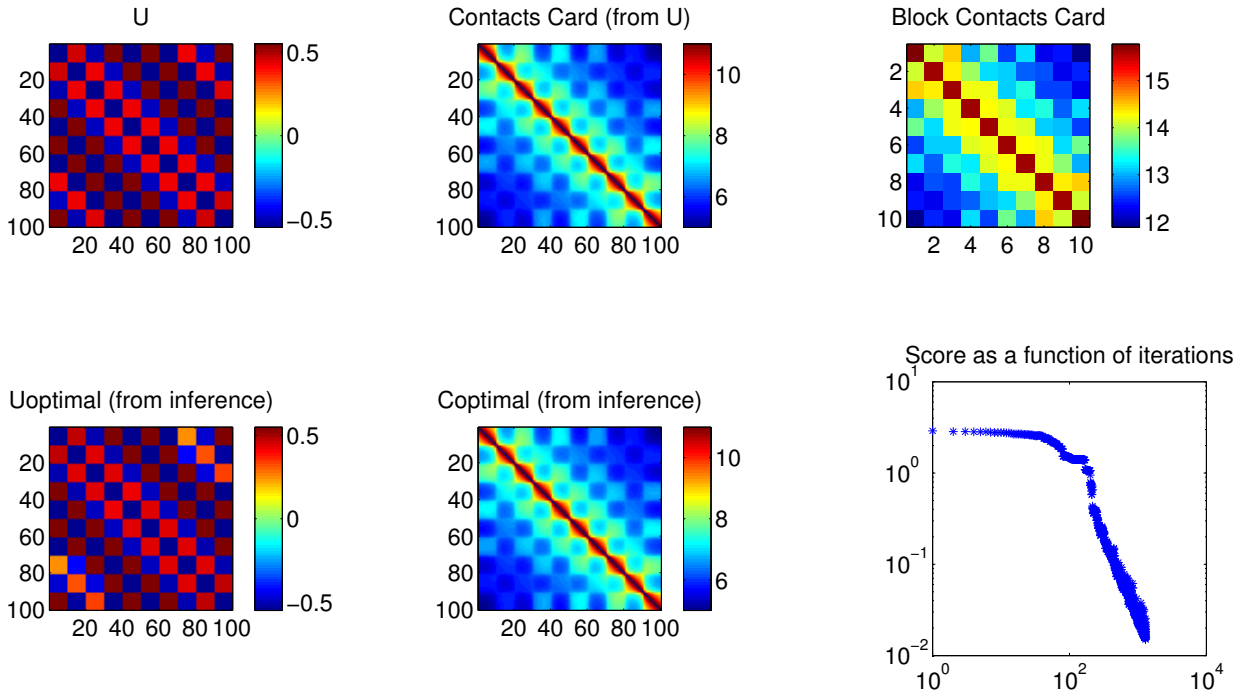

FIGURE 5.4.4 – **Inférence dans le cas d'un copolymère $(A_{10}B_{10})_6$ de paramètres $u_s = -0.5k_BT$ et $u_{ns} = 0.5k_bT$.** Légende identique à celle de la figure 5.4.3.

5.5 Inférence par inversion de Boltzmann itérative

Dans la sous section précédente, nous sommes parvenus à obtenir de bon résultats mais le temps de calcul nécessaire pour inférer la matrice U recherchée nous motive à envisager une autre méthode. Nous avons vu dans la partie 3.3.2, que les cartes de contact expérimentales présentent une perturbation faible par rapport à un modèle neutre (c'est-à-dire sans interactions spécifiques entre loci de même état). De ce fait, l'inversion de Boltzmann itérative (IBI) semble être une méthode adaptée à notre contexte de travail. C'est en effet une approche perturbative consistant à optimiser la matrice d'interaction de manière itérative de sorte que la carte de contacts simulée (carte toujours à l'échelle des TADs) se rapproche le plus de la carte expérimentale (Éq. 5.10). Cette inversion de Boltzmann, contrairement à la technique précédente, optimise à chaque itération tout les paramètres. À cela s'ajoute l'abandon de la dichotomie au profit d'une relation directe ce qui permet aussi un gain de temps important.

5.5.1 Principe de la méthode

L'inversion de Boltzmann itérative consiste à optimiser la matrice d'interaction de manière itérative de sorte que la carte de contacts simulée se rapproche le plus de la carte expérimentale. Concrètement, à partir d'une matrice d'interaction obtenue à une étape i, U_i, on

va chercher U_{i+1} en supposant que l'écart entre $C_i(k,l)$ (nombre de contacts entre les compartiments k et l simulé au cours de l'étape i) $C^{cible}(k,l)$ (nombre de contacts cible) est dû à la perturbation $U_{i+1}(k,l) - U_i(k,l)$, ce qui donne mathématiquement :

$$e^{-\alpha k_B T(U_{i+1}(k,l)-U_i(k,l))} = \frac{C_i(k,l)}{C^{cible}(k,l)}$$

avec k_B et T respectivement la constante de Boltzmann et la température effective du système (pour nous, les énergies sont en unité k_BT, donc $k_BT = 1$) et avec α un facteur d'échelle compris entre 0 et 1 permettant d'éviter les larges fluctuations, on choisit $\alpha = 0.01$. La valeur de α choisie étant faible, elle n'optimise pas le temps de calcul mais elle permet d'éviter les instabilités numériques. Par la suite, on pourra rendre le coefficient α adaptatif, c'est-à-dire qu'il pourra être plus ou moins grand selon le comportement de la convergence en fonction des itérations.
On peut donc calculer la matrice U_{i+1} avec l'expression suivante :

$$U_{i+1}(k,l) = U_i(k,l) - \alpha k_B T \ln\left(\frac{C^i(k,l)}{C^{cible}(k,l)}\right) \tag{5.10}$$

Le principe algorithmique est de partir avec une matrice U_1 dont les coefficients sont par exemple tous nuls si on n'a pas d'a priori sur la solution, puis d'appliquer la formule 5.10 pour obtenir U_2. On réitére le processus autant de fois que nécessaire. Dans notre cas, on estime qu'une solution est obtenue lorsque le χ^2 est inférieur à 1 (Éq. 5.9). Ici, on voit clairement l'avantage de l'inversion de Boltmann itérative par rapport à la méthode d'inférence par dichotomie présentée ci-dessus : il s'agit juste d'appliquer une formule par itération et pout tout les coefficients alors qu'avant on faisait une dichotomie de sorte à égaler le nombre de contacts expérimental et le nombre de contacts simulé dans un même compartiment. D'après nos test, cela permet de gagner un facteur 50 en terme de temps tout en menant aux mêmes résultats.

5.5.2 Résultats

5.5.2.1 Validité de l'IBI

Dans cette sous section, on va construire un copolymère par bloc avec le formalisme développé dans le chapitre 4 et on va chercher à savoir si avec avec l'inversion de Boltzmann itérative on retrouve les paramètres choisis. On considère un copolymère $(A_{10}B_{10})_6$ de 6 blocs A et 6 blocs B, chacun comportant 10 monomères. On le modélise avec l'hamiltonien H_{2B} (Éq. 4.36) et on choisi comme paramètre d'interaction $u_s = -0.4k_BT$ et $u_{ns} = 0k_bT$. On peut voir sur la figure le résultat de l'inférence après 20 minutes de calcul.

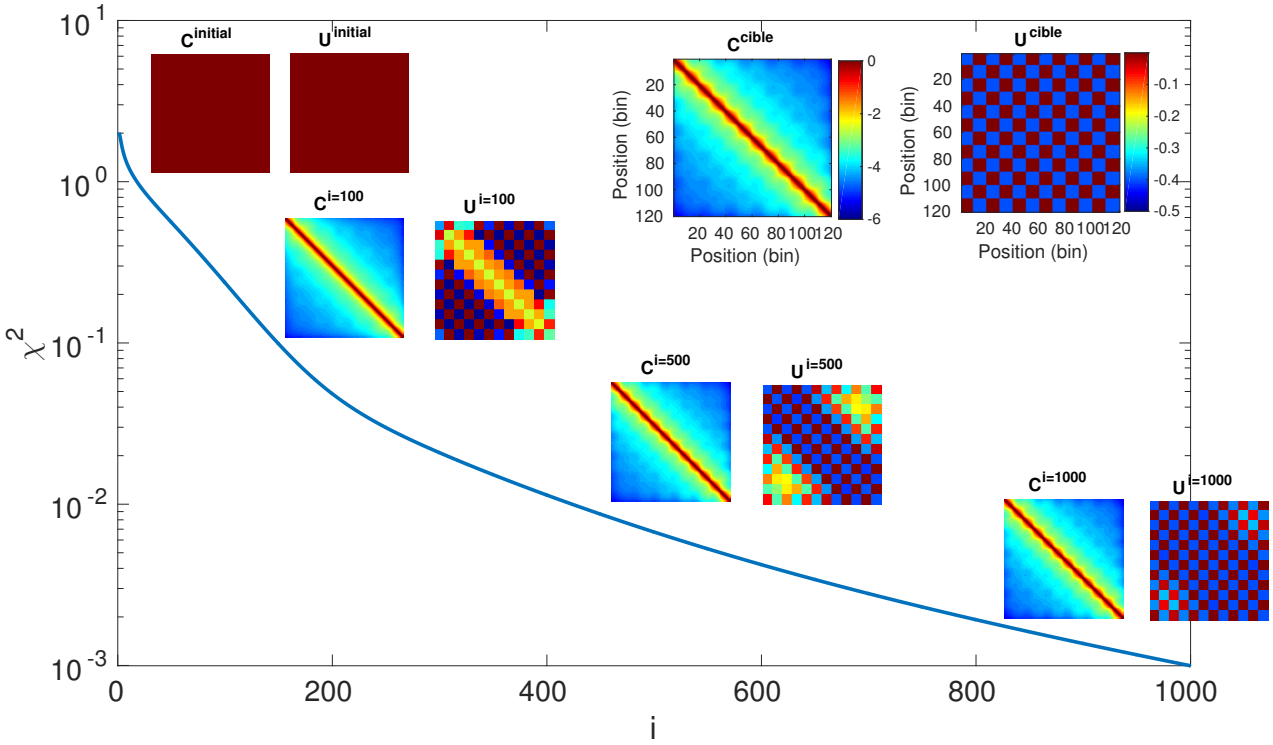

Figure 5.5.1 – **Inférence des paramètre d'interaction pour un copolymère $(A_{10}B_{10})_6$ par inversion de Boltzmann itérative.** Score χ^2 (Éq. équation (5.9)) en fonction de l'itération i. Le long de la courbe, pour les itérations $i = 0$, 100, 500 et 1000 les cartes prédites C^i avec la matrice des interactions U^i sont représentées. On vérifie ici que l'inversion itérative de Boltzmann permet de retrouver les paramètres d'interaction avec lesquels on a construit le copolymère d'étude. (En haut à droite) Carte cible, C^{cible} générée avec le formalisme de l'approximation gaussienne auto-cohérente et à partir de la matrice d'interaction U^{cible}. Les cartes C et les cartes U ont des échelles de couleur communes.

Nous avons testé de débuter l'inférence avec d'autres matrices $U^{initial}$, l'inversion de Boltzmann a toujours pour le cas de ce copolymère convergé vers la même solution (erreur relative maximale de 10^{-2}). D'une condition initiale à une autre, seul le temps de calcul a changé.

Étant donné la rapidité de la méthode IBI, on propose de vérifier sa validité sur un ensemble E_1 de 100 copolymères $(A_{10}B_{10})_6$ de paramètres $u_s \in [-4,\ 0]$ et $u_{ns} \in [-1,\ 3]$ dont les couples sont choisis de telle sorte à couvrir toutes les régions du diagramme de phase présenté figure 4.2.31. On trouve dans tous les cas que les résultats obtenus avec l'IBI sont exacts dans le sens où ils permettent de retrouver les paramètres u_s et u_{ns} choisis, à 10^{-2} près d'erreur relative maximum sur l'ensemble des coefficients.

Aussi, on valide la méthode en observant que l'inférence du champ générique avec l'IBI donne exactement les mêmes résultats que ceux obtenus avec la méthode d'inversion directe (erreur relative de 0.05 au plus).

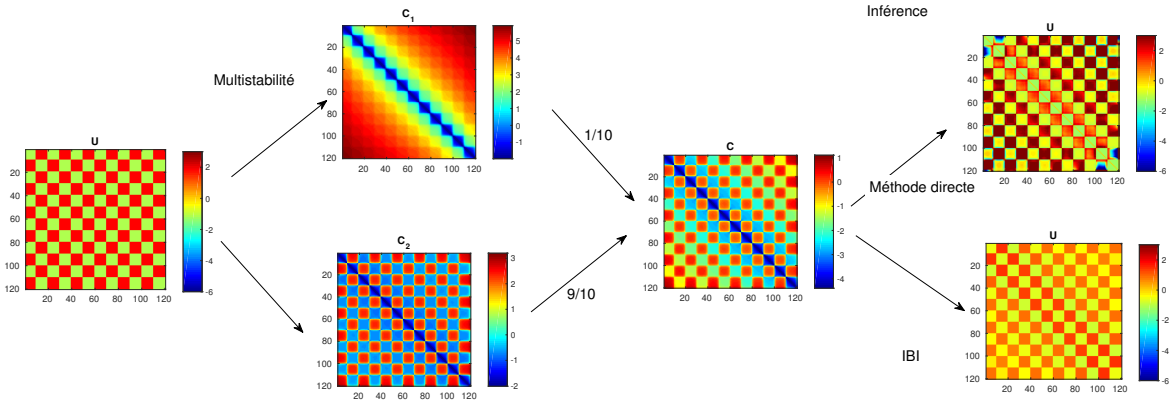

FIGURE 5.5.2 – **Inférence des paramètre d'interaction pour un copolymère $(A_{10}B_{10})_6$ dans la région de multistabilité par inversion de Boltzmann itérative.** À gauche matrice U choisie, on voit que $u_{ns} = 1.8k_BT$ et $u_s = -3k_BT$. Ensuite, C_1 et C_2 sont deux cartes simulées à partir de U (multistabilité). On calcule $C = 1/10C_1 + 9/10C_2$ et on applique l'IBI à cette carte de contact. Les matrices U trouvées sont complètement à droite : en haut calculé en résolvant le problème inverse et en bas avec l'IBI.

Pour se faire une idée du temps de calcul, on peut signaler que l'inférence par méthode directe est « instantanée » (moins d'une seconde) alors qu'avec l'IBI elle nécessite environ 8 minutes.

5.5.2.2 Cas particulier de la multistabilité

Dans cette sous section, on s'intéresse à un copolymère $(A_{10}B_{10})_6$ se situant dans la région de multistabilité mathématique sur le diagramme de phase 4.2.31. On choisit le copolymère de paramètre d'interaction non spécifique $u_{ns} = 1.8k_BT$ et de paramètre d'interaction spécifique entre monomères du même type, $u_s = -3k_BT$. Sur la figure 5.5.2 on peut voir la matrice U correspondante. Cette matrice mène à l'obtention de deux états stationnaires représentés par les cartes C_1 et C_2. Appliquer l'IBI à C_1 et à C_2 indépendemment mène à l'inférence des bons paramètres (cf. sous section précédente). Ici, on applique l'IBI à la carte de contact $C = 1/10C_1 + 9/10C_2$. On voit que proche de la diagonale l'IBI et la méthode de calcul exact donnent des résultats très ressemblants. À l'inverse, on observe que plus on s'éloigne de la diagonale, moins les paramètres inférés sont convaincants.

5.5.2.2.1 Inférence des paramètre chez la drosophile On s'intéresse ici à l'inférence des paramètres d'interaction dans le cas de la drosophile. Pour cela, on commence par segmenter l'ensemble de son génome en fenêtre de taille environ égale à 1Mb à l'aide de IC-Finder. Ensuite on segmente en TADs chacune des fenêtres puis on applique l'IBI sur chacune d'entre elles (cf. Fig. 5.5.3 qui est une fenêtre de 1.5Mb). L'inférence mène à une matrice U présentée

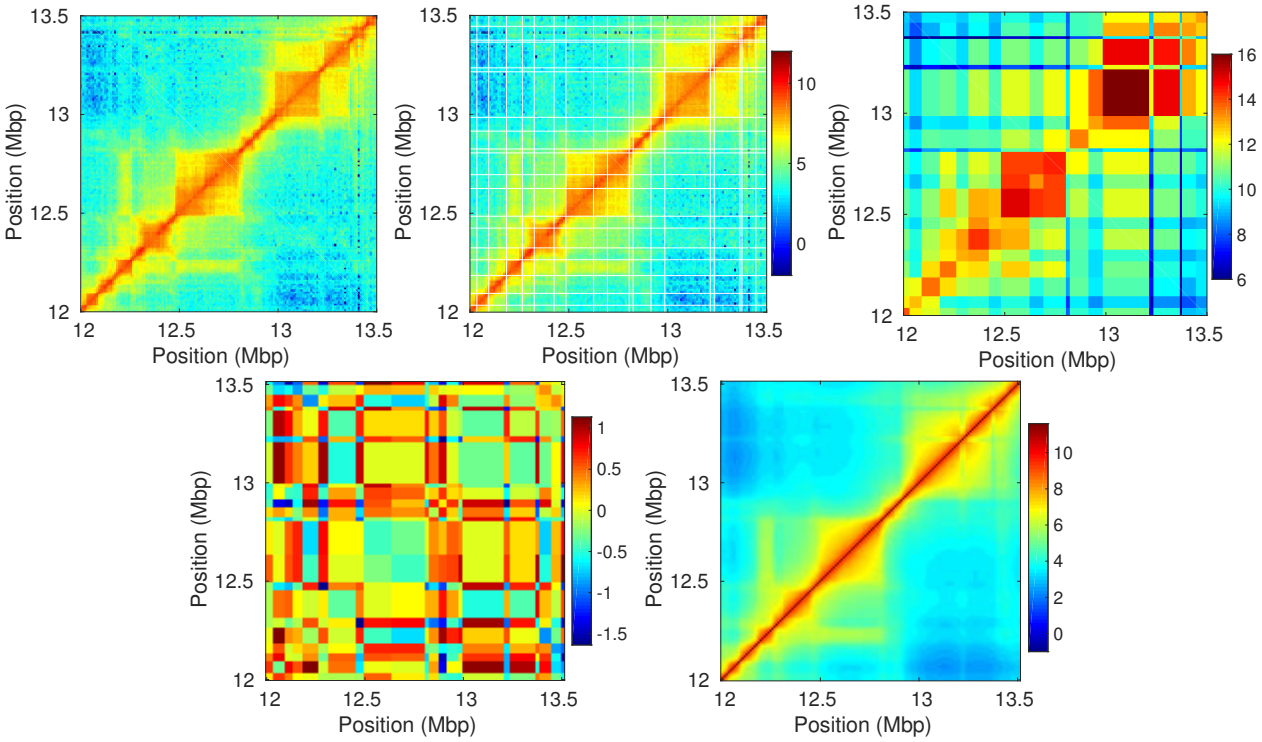

Figure 5.5.3 – **Inférence par Inversion de Boltzmann Itérative (IBI) pour la région entre 12Mb et 13.5Mb du chromosome 3R de la drosophile.** De la gauche vers la droite et du haut vers le bas, la première carte représente le bout expérimental étudié [Sexton et al. 2012], la seconde est identique à la première mais accompagnée de traits blancs définissant les TADs obtenus avec IC-Finder, la troisième carte est obtenue en sommant les contacts dans chaque compartiment, la quatrième carte donne la matrice des paramètres d'interaction inférés, U, et enfin la dernière matrice représentée est le résultat de la simulation obtenue à partir de la matrice U.

figure 5.5.3. Cette matrice U obtenue permet de reproduire fidèlement la carte de contact (5.5.3).

5.5.2.2.1.a Cas des paramètres intra domaine Comme on ne sait pas si la matrice U trouvée se situe dans la région de multistabilité mathématique et comme on a vu que dans le cas échéant l'inférence n'avait de sens que près de la diagonale, on s'intéresse tout d'abord aux paramètres $\mathsf{u}_{\mathtt{intra}}$ seulement. Sur la figure 5.5.4, on remarque que pour les domaines répressifs par rapport aux actifs les distributions des paramètres $\mathsf{u}_{\mathtt{intra}}$ trouvées sont moins étendues. Ceci peut s'expliquer de deux façons : les domaines actifs sont plus petits donc l'incertitude sur les potentiels trouvés est plus grande ou les domaines actifs sont le siège de divers processus de régulation faisant intervenir des interactions plus ou moins fortes selon les cas.

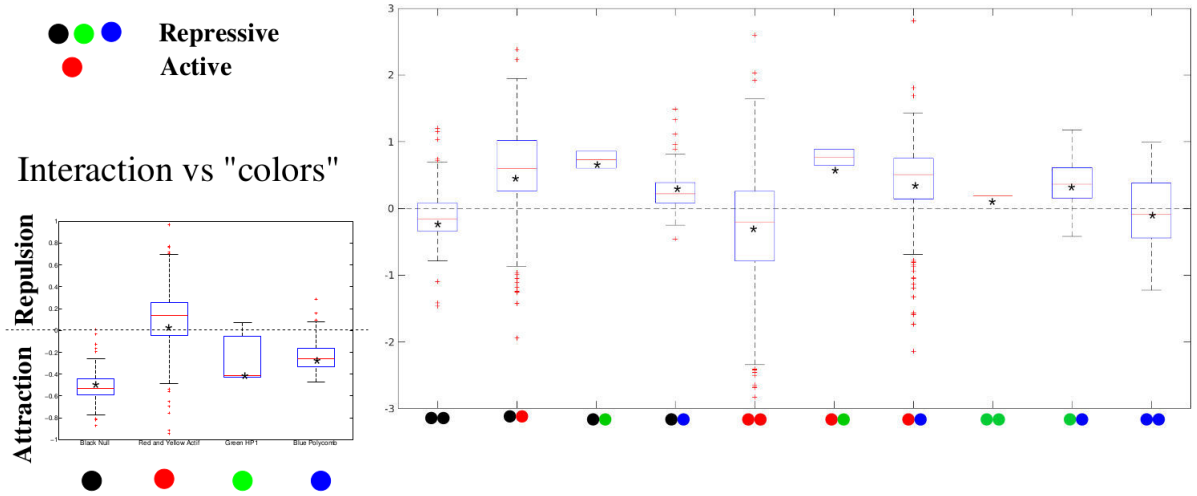

FIGURE 5.5.4 – **Diagrammes en boites des paramètres d'interaction intra domaines (à gauche) et inter domaine (à droite) inférés sur l'ensemble du génome de la drosophile sauf le chromosome 4.** L'ordonnée des deux graphes donne la valeur des paramètres en unité $k_B T$. On notera tout de même qu'entre les deux graphe, l'échelle n'est pas la même. La ligne pointillée représente une intensité d'interaction nulle. Les paramètres inter domaine présentés ici se limite aux voisins espacés d'au plus de 2 TADs. Les étoiles noires donnent la valeur des paramètres inférés en limitant le nombre de paramètres à inférer à 14.

Aussi, on observe que dans le cas de l'hétérochromatine, les paramètres intra domaines sont en moyenne négatifs (attraction) alors qu'ils sont proche de zéro, ou légèrement positifs (faible répulsion) dans le cas de l'euchromatine. Ceci suggère que l'euchromatine n'interagit que très peu avec elle-même en intra. Il y aurait donc deux modes d'interaction locale distincts : l'euchromatine s'organise localement de manière discrète avec des pontages à courte portée entre loci spécifiques alors que l'hétérochromatine interagit plus continuellement avec des regroupements de plusieurs sites génomiques. La définition de ces deux modes d'interaction est cohérente avec l'observation expérimentale selon laquelle les domaines inactifs présentent des contacts homogènes alors que pour les domaines actifs, l'interactome est plus complexe [Sofueva et al., 2013].

5.5.2.2.1.b Cas des paramètres inter domaine On s'intéresse ici aux paramètres d'interaction inter domaines. Étant donné que nous ne savons pas si les cartes de contact à partir desquelles nous inférons les paramètres sont issues de configurations multistables ou pas, nous nous concentrons uniquement sur les paramètres inter domaines proches voisins (espacés au plus de 2 domaines). La figure 5.5.4 montre que les domaines de même état épigénomique interagissent préférentiellement comparés aux domaines d'états différents (hormis

le cas vert/vert pour lequel la statistique n'est pas bonne). Cette observation va exactement dans le sens des conclusions tirées des analyses statistiques du chapitre 3.
Aussi, il est intéressant de remarquer que les domaines actifs (en rouge) présentent des paramètres d'interaction intra domaine en moyenne positifs (répulsion) alors que les paramètres inter domaines actifs sont en moyenne négatifs (attraction à longue portée entre domaines actifs). Autrement dit, en moyenne, les régions actives de la chromatine interagissent peu avec leur partenaire de TAD mais interagissent plus avec les régions actives d'autres TADs. Une interprétation possible de cette observation est que les petits domaines actifs sont "exclus" des gros domaines inactifs environnant (on le voit très bien avec les simulations sur réseau par exemple). Or, avec l'approche gaussienne auto cohérente on ne tient pas compte de ces effets, l'inférence mène donc à des interactions négatives pour reproduire ces interactions attractives à longue portée.

5.5.2.2.1.c Reproductibilité de l'IBI Si on considère que toutes les solutions telles que $\chi^2 < 1$ sont acceptables, on obtient plusieurs solutions. La distribution des solutions pour chaque coefficient de matrice U est représentée figure 5.5.5. On peut voir sur ces distributions qu'a priori il n'y a qu'un pic ce qui suggère que les solutions trouvées forment une même et unique famille. De plus, sachant que notre méthode est non déterministe en raison du tirage des compartiments à optimiser, on peut inférer 4 fois les paramètres pour le même système. C'est ce qu'on fait sur la même figure (une couleur par réplicat). On observe une très bonne reproductibilité sur la diagonale mais concernant les paramètres hors diagonaux trouvés, on peut voir que plus ceux-ci sont éloignés de la diagonale, moins ils sont fiables.
Aussi, on peut se demander si la matrice U inférée dépend de la condition initiale sur U dans notre algorithme. Pour avoir une idée de la réponse, on représente sur la figure 5.5.6 les résultats obtenus en partant de $\mathsf{U} = 0$ et en partant de $\mathsf{U} = \mathsf{U}^{\text{générique}}$. On voit que tout comme l'étude réalisée avec 4 réplicats, on voit ici que plus les coefficients sont loin de la diagonale plus leur valeur inférée varie.

5.5.2.2.1.d Vers un modèle prédictif En conclusion, nous avons vu que l'IBI dans le cadre de l'approximation gaussienne auto-cohérente permet d'inférer les paramètres intra domaine ou entre proche voisins (mais pas plus en raison de l'éventuel phénomène de multistabilité). Les paramètres que nous avons obtenus se sont avérés être cohérents avec des résultats expérimentaux. Afin de tester si les paramètres obtenus permettent de rendre le modèle prédictif, on propose de réduire la dimensionnalité du problème en inférant les paramètres entre états épigénomiques plutôt qu'entre TADs. En considérant un nombre 4 états épigénomiques (actif, polycomb, HP1 et nulle), nous avons 14 paramètres à inférer (4 intra

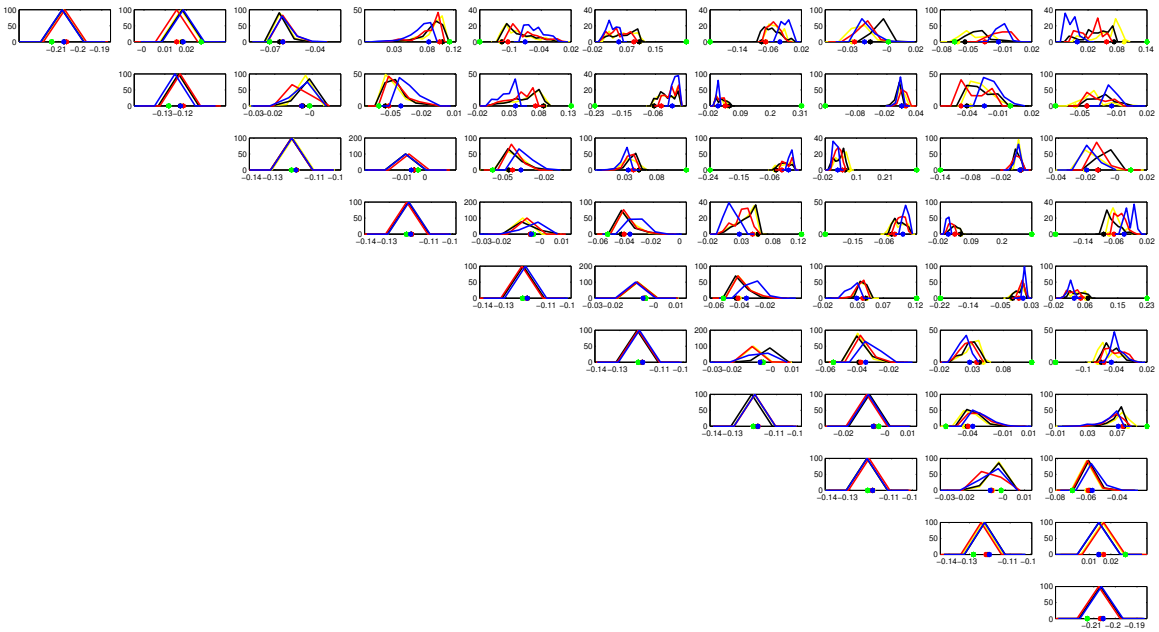

FIGURE 5.5.5 – **Reproductibilité de la méthode.** Chacune des quatre couleurs correspond à un réplicat de l'inférence. Le point de la même couleur sur l'axe abscisses correpond à la valur de U_{kl} telle que le χ^2 est minimum. Le point vert indique le paramètre utilisé pour générer le copolymère étudié.

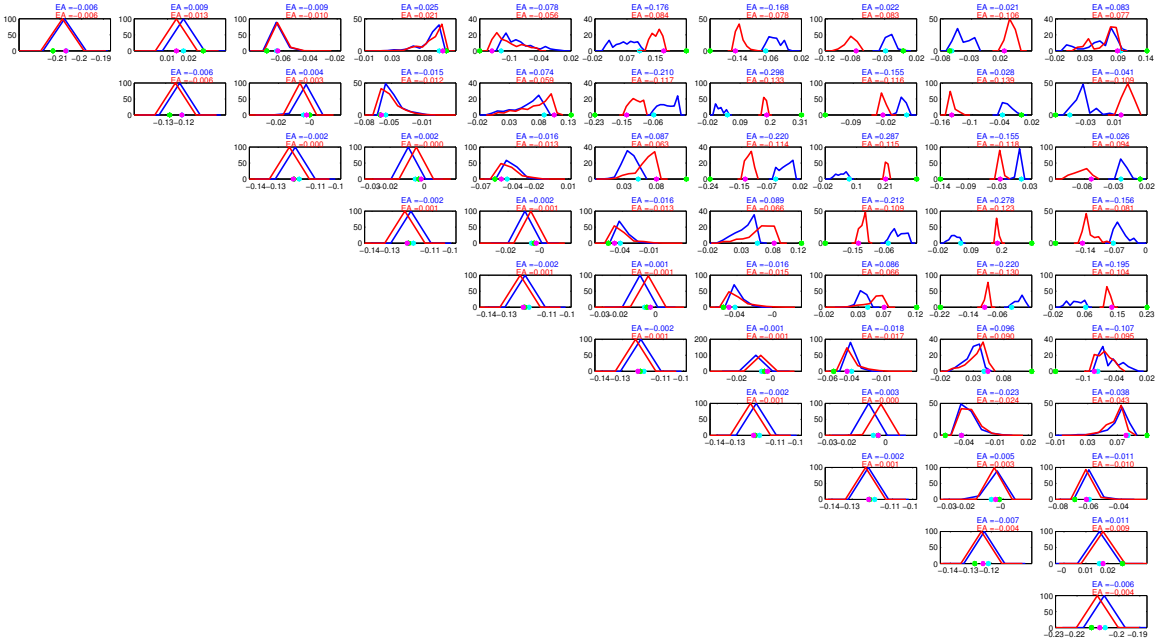

FIGURE 5.5.6 – **Distributions des U_{kl} obtenus par IBI (tels que $\chi^2 < 1$) pour deux conditions initiales différentes.** En bleu, la matrice initiale est nulle, en rouge est est égale au champ générique. En cyan et en magenta on peut respectivement voir les paramètres donnant le χ^2 le plus faible dans chaque pas. Le point vert indique le coefficient dont on s'est servi pour générer la carte de contact simulée.

et 10 inter) plutôt que $\frac{n_d(n_d-1)}{2}$ avec n_d le nombre de domaines topologiques. Les résultats obtenus sont représentées par des étoiles sur la figure 5.5.4. Les valeurs trouvés en inférant 14 paramètres seulement sont proches des moyennes trouvées en inférant $\frac{n_d(n_d-1)}{2}$ paramètres. Les deux manières de faire se valident donc mutuellement.
Sur la figure 5.5.7, on peut voir les prédictions obtenues avec les 14 paramètres inférés. Ces prédictions ressemblent aux données expérimentales. Les différences observées peuvent s'expliquer par le fait qu'on soit dans la zone de multistabilité ou encore par des défauts de segmentation épigénomique : par exemple, il arrive que la partition épigénomique indique l'existence d'un domaine polycomb de signal faible, or avec notre approche on ne tient pas compte du fait que le signal soit faible, on prévoit donc un TAD alors que celui ci n'est pas visible expérimentalement (Fig. 5.5.7). Par la suite, afin d'éviter ce problème, une amélioration du modèle sera de tenir compte pour chaque domaine épigénomique du niveau d'intensité des marques.

5.6 Conclusion

Dans ce chapitre, notre objectif a été d'extraire à partir des cartes Hi-C expérimentales les potentiels d'interaction entre monomères définis dans le chapitre précédent. Pour réaliser cela, nous avons tout d'abord envisagée la méthode la plus naturelle qui consiste à inverse le problème direct. Avec cette méthode, nous avons pu étudier des exemples pédagogiques et en particulier analyser l'effet de la multistabilité mathématique sur l'inférence des potentiels. Aussi nous avons pu inférer le champ générique. Cette méthode nous servira donc systématiquement pour l'inférence de ce champ. Par contre, il n'a pas été possible de traiter le cas de l'inférence des paramètres à partir d'une carte HiC expérimentale. En effet, bien que cette méthode d'inversion soit idéale, elle nécessite l'inversion de la matrice des potentiels gaussiens, étape qui amplifie dramatiquement les erreurs. Malgré nos efforts réduire le conditionnement de ce problème, nous sommes restés sans solutions face à cette extrême sensibilité au bruit. Nous avons alors décidé d'adopter la stratégie suivante : concernant l'inférence du champ générique, elle sera réalisée avec la méthode directe (puisque pas de problème de bruit) et pour l'inférence des interactions directes entre monomères (écarts par rapport au champ générique), nous allons développer une approche « gros grain » dont l'échelle est fixé par les TADs. Autrement dit, nous cherchons les paramètres d'interaction entre entre domaines topologiques. Pour cela nous avons proposé trois approches : (1) une inférence bayesienne avec l'hypothèse que les paramètres u_{intra} à inférer ne dépendent pas de l'environnement, (2) une méthode d'optimisation locale du nombre de contact par dichotomie, (3) une inversion de Boltzmann itérative. Les résultats obtenus avec (1) ne nous ont pas convaincus contrairement

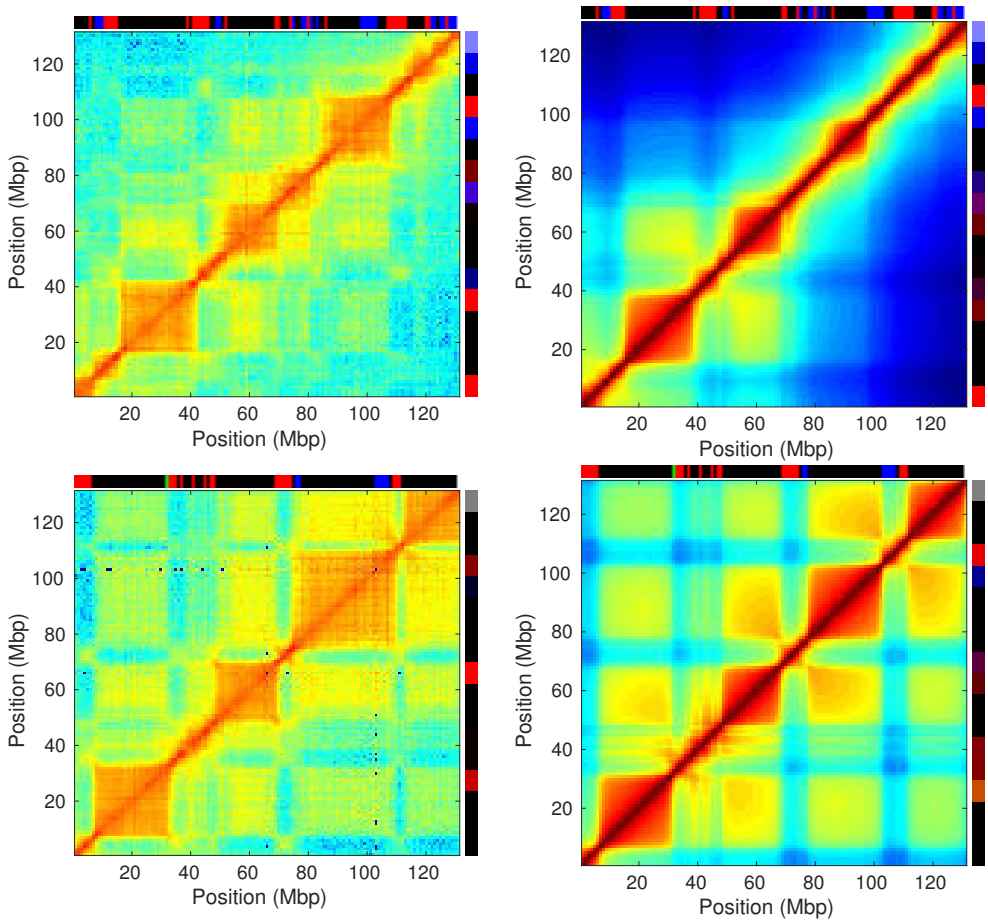

FIGURE 5.5.7 – **Prédictions obtenues à partir de l'approche gaussienne autocohérente couplée aux 14 paramètres trouvés par IBI pour deux régions génomiques du chromosome 3R. Sur chaque ligne, à gauche se trouve la carte expérimentale et à droite se trouve la carte simulée.** Les séquences épigénétiques données sont issues de Filion et al., 2010 (en groupant les domaines actifs en un unique domaine rouge).

à ceux obtenus avec (2) et (3). Ces deux dernières méthodes donnent des résultats semblables mais dans un temps divisé par 50 pour la (3).

En résumé, l'inversion de Boltzmann itérative est une approche perturbative raisonnable, non rédhibitoire par son temps de calcul et qui permet de traiter le cas de cartes bruitées. Naturellement, c'est donc cette méthode qui nous a permis d'extraire les paramètres d'interaction à l'échelle du génome chez la drosophile. Comme nous l'avons vu, le phénomène de multistabilité mathématique (observé particulièrement lorsque la carte de contact étudiée présente de nombreuses interactions à longue portée) implique que seuls les coefficients intra TAD et entre proches voisins sont fiables. Par contre, en dehors de la zone de multistabilité mathématique, l'inférence est fiable sur l'ensemble de la carte de contact.

Une perspective prometteuse de cette inférence est qu'elle peut permettre de rendre le modèle de copolymère prédictif. Il sera alors possible de simuler l'organisation de la chromatine dans différentes conditions. En particulier, on pourra étudier in silico les conséquences d'une altération de l'épigénome.

CHAPITRE 6

CONCLUSION

L'objectif de cette thèse était de valider l'hypothèse selon laquelle l'épigénome est un acteur majeur dans le repliement 3D de la chromatine. C'est ce que nous avons fait à partir d'une analyse statistique et à partir de prédiction réalisées avec un modèle physique de co-

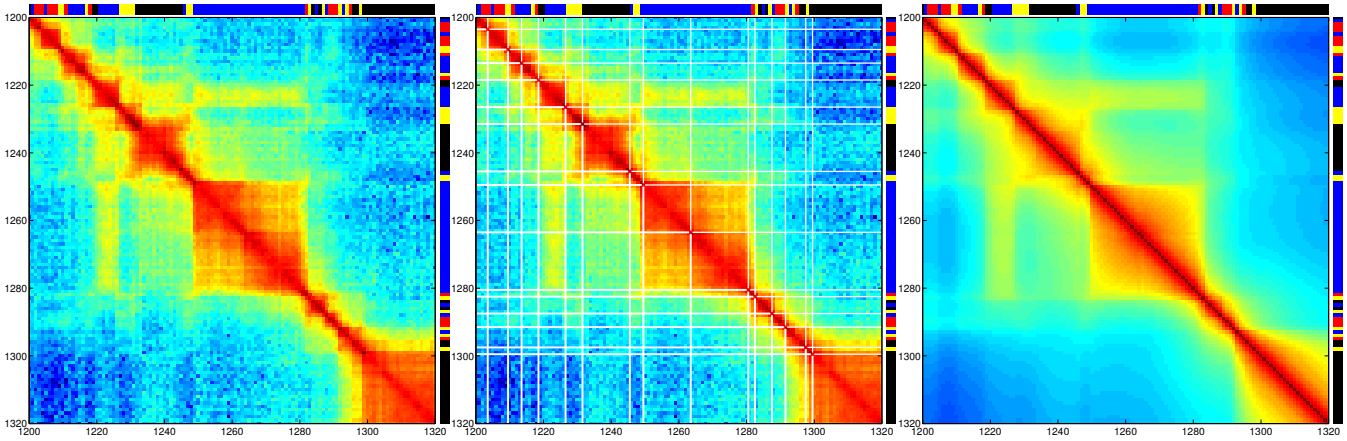

FIGURE 1 – **À partir d'une analyse statistique et à partir de prédiction réalisées avec un modèle physique de copolymère par bloc, on a montré que l'épigénome peut en grande partie expliquer le repliement de la chromatine.** (À gauche) Carte expérimentale pour une région génomique du chromosome 3R de la drosophile [Sexton et al. 2012]. (Au milieu) Carte identique à celle de gauche mais avec en plus la position des TADs déterminés avec IC-Finder. (À droite) Carte de contact simulée dans le cadre de la modélisation de la chromatine par un copolymère par bloc sous l'approximation gaussienne auto-cohérente. Les paramètres nécessaires pour simuler cette carte ont été inféré via l'inversion de Boltzmann itérative. Au dessus et à droite de chaque carte se trouve la séquence épigénomique [Filion et al., 2010].

polymère par bloc. Ces approches ont toutes deux sollicitées « IC-Finder » un algorithme que nous avons développé afin de segmenter les cartes de contact en domaines d'interaction. Dans la partie sur l'analyse statistique, nous avons vu que les données Hi-C combinées aux données épigénomiques suggèrent que chez la drosophile et chez l'homme les loci de même état épigénomique interagissent spécifiquement entre eux. Concernant la modélisation de la chromatine, nous avons approfondi le modèle de copolymère par bloc sous l'approximation gaussienne auto-cohérente initialement développée par Jost et al., 2014. En particulier, nous avons modélisé le volume exclu d'une manière plus réaliste et nous avons introduit un champ générique permettant de rendre compte de l'environnement et/ou des effets de « crumpling ». Nous avons montré que des approches de dynamique sur réseau et dynamique moléculaire justifiaient et validaient l'approche auto-cohérente. Les trois approches que nous avons présenté afin de modéliser la chromatine comme un copolymère par bloc ont montré que l'existence d'interactions pilotées par l'épigénome permet de reproduire les cartes HiC obtenues expérimentalement et donc d'expliquer l'organisation 3D de la chromatine (Fig. 1). Nous nous sommes donc ensuite concentrés sur l'inférence des paramètres de ce modèle. En raison de problèmes techniques tels que le bruit expérimental ou encore le temps calcul il n'a pas été évident de réaliser cette tache. Nous avons vu que l'inversion de Boltzmann itérative résout en partie ces problèmes rencontrés et offre de bons résultats. Aussi, nous avons vu que réaliser l'inférence à partir d'un modèle d'interaction entre états épigénomiques donne des résultats comparables à ceux obtenus avec un modèle d'interaction entre TADs et cette méthode a pour avantage de n'impliquer que peu de paramètres.

Afin de valider l'approche gaussienne auto-cohérente couplée au processus d'inférence par IBI, on pourra travailler avec les cartes de contact simulées par dynamique moléculaire par Pascal Carrivain. Cela permettra d'une part de vérifier que le processus d'inférence mène aux bons paramètres et cela permettra d'autre part de juger les prédictions obtenues. En cherchant à valider le côté prédictif de l'approche gaussienne auto-cohérente avec des résultats obtenus in silico plutôt qu'avec des données expérimentales permet de s'affranchir des incertitudes liées à la segmentation épigénétique. Ce projet nécessitera au préalable un travail théorique visant à trouver une correspondance entre les paramètres utilisé dans les deux approches. Une fois ce travail réalisé, l'interprétation des résultats obtenus par inférence sera facilitée.

La suite de ce travail de thèse, déjà entrepris par Daniel Jost et Cédric Vaillant, est de rendre le copolymère « vivant » en permettant des transitions d'états épigénomiques. Cette idée de modéliser la chromatine par un copolymère par bloc dont les blocs peuvent changer d'états a pour origine des résultats expérimentaux mettant en avant des mécanismes permettant aux marques épigénomiques de se propager après nucléation au niveau de certains sites génomiques [Beisel and Paro, 2011; Chen and Dent, 2014; Zhang et al., 2015; Soshnev et al., 2016].

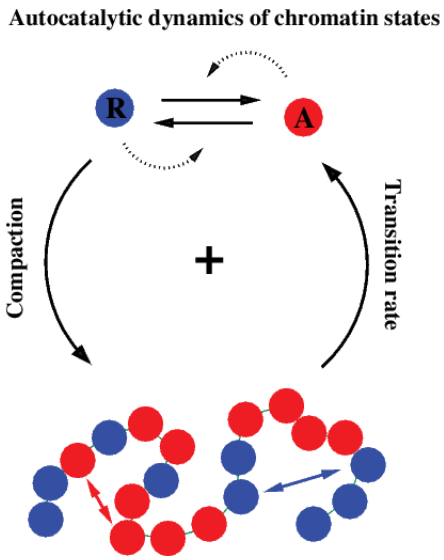

FIGURE 2 – **Modèle de « chromatine vivante ».** Ce modèle combine le cadre du copolymère par bloc (interaction attractive entre blocs de même état) avec une modélisation de la dynamique de l'épigénome (conversion auto catalytique entre états). La figure présente le cas simple d'une chaîne formée de deux types blocs : réprimé en bleu et actif en rouge.

La propagation des marques épigénomiques le long du génome ne se fait donc pas seulement en cis mais aussi en trans entre régions génomiques proche spatialement. Ainsi, le repliement 3D de la chaîne influence l'épigénome et réciproquement. L'étude de ce couplage permettra de mieux comprendre les liens entre epigénome et organisation 3D de la chromatine et en particulier l'étude de la propagation de l'hétérochromatine.

BIBLIOGRAPHIE

Ahmed, K., Dehghani, H., Rugg-Gunn, P., Fussner, E., Rossant, J. and Bazett-Jones, D. P. (2010). Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo. PloS One *5*, e10531.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2002). Molecular Biology of the Cell. 4th edition, Garland Science.

Allis, C., Jenuwein, T. and Reinberg, D. (2007). Epigenetics. Cold Spring Harbor Laboratory Press.

Ay, F., Bunnik, E. M., Varoquaux, N., Bol, S. M., Prudhomme, J., Vert, J.-P., Noble, W. S. and Le Roch, K. G. (2014). Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. Genome Research *24*, 974–988.

Ay, F. and Noble, W. S. (2015). Analysis methods for studying the 3D architecture of the genome. Genome Biology *16*, 183.

Bantignies, F. and Cavalli, G. (2011). Polycomb group proteins : repression in 3D. Trends in genetics : TIG *27*, 454–464.

Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L.-M., Dostie, J., Pombo, A. and Nicodemi, M. (2012). Complexity of chromatin folding is captured by the strings and binders switch model. Proceedings of the National Academy of Sciences of the United States of America *109*, 16173–16178.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N.,

Robertson, C. L., Serova, N., Davis, S. and Soboleva, A. (2013). NCBI GEO : archive for functional genomics data sets–update. Nucleic Acids Res *41*, D991–D995.

Baù, D., Sanyal, A., Lajoie, B. R., Capriotti, E., Byron, M., Lawrence, J. B., Dekker, J. and Marti-Renom, M. A. (2011). The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. Nature Structural & Molecular Biology *18*, 107–114.

Beisel, C. and Paro, R. (2011). Silencing chromatin : comparing modes and mechanisms. Nature Reviews. Genetics *12*, 123–135.

Bickmore, W. A. and van Steensel, B. (2013). Genome Architecture : Domain Organization of Interphase Chromosomes. Cell *152*, 1270–1284.

Boettiger, A. N., Bintu, B., Moffitt, J. R., Wang, S., Beliveau, B. J., Fudenberg, G., Imakaev, M., Mirny, L. A., Wu, C.-t. and Zhuang, X. (2016). Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. Nature *529*, 418–422.

Bohn, M. and Heermann, D. W. (2010). Diffusion-Driven Looping Provides a Consistent Framework for Chromatin Organization. PLoS ONE *5*.

Brackley, C. A., Taylor, S., Papantonis, A., Cook, P. R. and Marenduzzo, D. (2013). Non-specific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization. Proceedings of the National Academy of Sciences of the United States of America *110*, E3605–3611.

Byrne, A., Kiernan, P., Green, D. and Dawson, K. A. (1995). Kinetics of homopolymer collapse. The Journal of Chemical Physics *102*, 573–577.

Bystricky, K., Heun, P., Gehlen, L., Langowski, J. and Gasser, S. M. (2004). Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. Proceedings of the National Academy of Sciences of the United States of America *101*, 16495–16500.

Canzio, D., Liao, M., Naber, N., Pate, E., Larson, A., Wu, S., Marina, D. B., Garcia, J. F., Madhani, H. D., Cooke, R., Schuck, P., Cheng, Y. and Narlikar, G. J. (2013). A conformational switch in HP1 releases auto-inhibition to drive heterochromatin assembly. Nature *496*, 377–381.

Caré, B., Emeriau, P.-E., Cortini, R. and Victor, J.-M. (2015). Chromatin epigenomic domain folding : size matters. AIMS Biophysics *2*, 517–530.

Cavalli, G. and Misteli, T. (2013). Functional implications of genome topology. Nature Structural & Molecular Biology *20*, 290–299.

Chandra, T., Kirschner, K., Thuret, J.-Y., Pope, B. D., Ryba, T., Newman, S., Ahmed, K., Samarajiwa, S. A., Salama, R., Carroll, T., Stark, R., Janky, R., Narita, M., Xue, L., Chicas, A., Nũnez, S., Janknecht, R., Hayashi-Takanaka, Y., Wilson, M. D., Marshall, A., Odom, D. T., Babu, M. M., Bazett-Jones, D. P., Tavaré, S., Edwards, P. A. W., Lowe, S. W., Kimura, H., Gilbert, D. M. and Narita, M. (2012). Independence of repressive histone marks and chromatin compaction during senescent heterochromatic layer formation. Molecular Cell *47*, 203–214.

Chen, T. and Dent, S. Y. R. (2014). Chromatin modifiers and remodellers : regulators of cellular differentiation. Nature Reviews. Genetics *15*, 93–106.

Cheutin, T. and Cavalli, G. (2012). Progressive polycomb assembly on H3K27me3 compartments generates polycomb bodies with developmentally regulated motion. PLoS genetics *8*, e1002465.

Ciabrelli, F. and Cavalli, G. (2015). Chromatin-Driven Behavior of Topologically Associating Domains. Journal of Molecular Biology *427*, 608–625.

Cortini, R., Barbi, M., Caré, B. R., Lavelle, C., Lesne, A., Mozziconacci, J. and Victor, J.-M. (2016). The physics of epigenetics. Reviews of Modern Physics *88*.

Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R. and Mozziconacci, J. (2012). Normalization of a chromosomal contact map. BMC genomics *13*, 436.

Crane, E., Bian, Q., McCord, R. P., Lajoie, B. R., Wheeler, B. S., Ralston, E. J., Uzawa, S., Dekker, J. and Meyer, B. J. (2015). Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature *523*, 240–244.

Cremer, T. and Cremer, M. (2010). Chromosome territories. Cold Spring Harbor Perspectives in Biology *2*, a003889.

de Gennes, P.-G. (1979). Scaling Concepts in Polymer Physics. Cornell University Press.

Dekker, J. and Heard, E. (2015). Structural and functional diversity of Topologically Associating Domains. FEBS Lett *589*, 2877–84.

Dekker, J., Marti-Renom, M. A. and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes : interpreting chromatin interaction data. Nature Reviews Genetics *14*, 390–403.

Dekker, J. and Mirny, L. (2016). The 3D Genome as Moderator of Chromosomal Communication. Cell *164*, 1110–1121.

Deng, X., Ma, W., Ramani, V., Hill, A., Yang, F., Ay, F., Berletch, J. B., Blau, C. A., Shendure, J., Duan, Z., Noble, W. S. and Disteche, C. M. (2015). Bipartite structure of the inactive mouse X chromosome. Genome Biol *16*, 152.

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V. V., Ecker, J. R., Thomson, J. A. and Ren, B. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

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

Doi, M. and Edwards, S. F. (1988). The Theory of Polymer Dynamics. Clarendon Press.

Dormand, J. R. and Prince, P. J. (1980). A family of embedded Runge-Kutta formulae. Journal of Computational and Applied Mathematics *6*, 19–26.

Dowen, J. M., Fan, Z. P., Hnisz, D., Ren, G., Abraham, B. J., Zhang, L. N., Weintraub, A. S., Schuijers, J., Lee, T. I., Zhao, K. and Young, R. A. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell *159*, 374–387.

Doyle, B., Fudenberg, G., Imakaev, M. and Mirny, L. A. (2014). Chromatin Loops as Allosteric Modulators of Enhancer-Promoter Interactions. PLOS Comput Biol *10*, e1003867.

Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., Lee, C., Shendure, J., Fields, S., Blau, C. A. and Noble, W. S. (2010). A three-dimensional model of the yeast genome. Nature *465*, 363–367.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M. and Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Feinberg, A. P. (2007). Phenotypic plasticity and the epigenetics of human disease. Nature *447*, 433–440.

Filion, G. J., van Bemmel, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., Brugman, W., de Castro, I. J., Kerkhoven, R. M., Bussemaker, H. J. and van Steensel,

B. (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell *143*, 212–224.

Filippova, D., Patro, R., Duggal, G. and Kingsford, C. (2014). Identification of alternative topological domains in chromatin. Algorithms Mol Biol *9*, 14.

Francis, N. J., Kingston, R. E. and Woodcock, C. L. (2004). Chromatin compaction by a polycomb group protein complex. Science (New York, N.Y.) *306*, 1574–1577.

Fraser, J., Ferrai, C., Chiariello, A. M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B. L., Kraemer, D. C. A., Aitken, S., Xie, S. Q., Morris, K. J., Itoh, M., Kawaji, H., Jaeger, I., Hayashizaki, Y., Carninci, P., Forrest, A. R. R., , F. A. N. T. O. M. C., Semple, C. A., Dostie, J., Pombo, A. and Nicodemi, M. (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol Syst Biol *11*, 852.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N. and Mirny, L. A. (2016). Formation of Chromosomal Domains by Loop Extrusion. Cell Reports *15*, 2038–2049.

Gennes, P.-G. (1979). Scaling Concepts in Polymer Physics. Cornell University Press, Ithaca, NY.

Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., Alves, P., Chateigner, A., Perry, M., Morris, M., Auerbach, R. K., Feng, X., Leng, J., Vielle, A., Niu, W., Rhrissorrakrai, K., Agarwal, A., Alexander, R. P., Barber, G., Brdlik, C. M., Brennan, J., Brouillet, J. J., Carr, A., Cheung, M.-S., Clawson, H., Contrino, S., Dannenberg, L. O., Dernburg, A. F., Desai, A., Dick, L., Dosé, A. C., Du, J., Egelhofer, T., Ercan, S., Euskirchen, G., Ewing, B., Feingold, E. A., Gassmann, R., Good, P. J., Green, P., Gullier, F., Gutwein, M., Guyer, M. S., Habegger, L., Han, T., Henikoff, J. G., Henz, S. R., Hinrichs, A., Holster, H., Hyman, T., Iniguez, A. L., Janette, J., Jensen, M., Kato, M., Kent, W. J., Kephart, E., Khivansara, V., Khurana, E., Kim, J. K., Kolasinska-Zwierz, P., Lai, E. C., Latorre, I., Leahey, A., Lewis, S., Lloyd, P., Lochovsky, L., Lowdon, R. F., Lubling, Y., Lyne, R., MacCoss, M., Mackowiak, S. D., Mangone, M., McKay, S., Mecenas, D., Merrihew, G., Miller, D. M., Muroyama, A., Murray, J. I., Ooi, S.-L., Pham, H., Phippen, T., Preston, E. A., Rajewsky, N., Rätsch, G., Rosenbaum, H., Rozowsky, J., Rutherford, K., Ruzanov, P., Sarov, M., Sasidharan, R., Sboner, A., Scheid, P., Segal, E., Shin, H., Shou, C., Slack, F. J., Slightam, C., Smith, R., Spencer, W. C., Stinson, E. O., Taing, S., Takasaki, T., Vafeados, D., Voronina, K., Wang, G., Washington, N. L., Whittle, C. M., Wu, B., Yan,

K.-K., Zeller, G., Zha, Z., Zhong, M., Zhou, X., modENCODE Consortium, Ahringer, J., Strome, S., Gunsalus, K. C., Micklem, G., Liu, X. S., Reinke, V., Kim, S. K., Hillier, L. W., Henikoff, S., Piano, F., Snyder, M., Stein, L., Lieb, J. D. and Waterston, R. H. (2010). Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science (New York, N.Y.) *330*, 1775–1787.

Gilbert, N., Boyle, S., Fiegler, H., Woodfine, K., Carter, N. P. and Bickmore, W. A. (2004). Chromatin architecture of the human genome : gene-rich domains are enriched in open chromatin fibers. Cell *118*, 555–566.

Giorgetti, L., Galupa, R., Nora, E. P., Piolot, T., Lam, F., Dekker, J., Tiana, G. and Heard, E. (2014). Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. Cell *157*, 950–963.

Grosberg, A. Y. and Khokhlov, A. R. (1994). Statistical physics of macromolecules. AIP Press.

Grosberg, A. Y., Khokhlov, A. R., Stanley, H. E., Mallinckrodt, A. J. and McKay, S. (1995). Statistical Physics of Macromolecules. Computers in Physics *9*, 171–172.

Grosberg, A. Y., Nechaev, S. and Shakhnovich, E. (1988). The role of topological constraints in the kinetics of collapse of macromolecules. Journal de Physique *49*, 2095–2100.

Hajjoul, H., Mathon, J., Ranchon, H., Goiffon, I., Mozziconacci, J., Albert, B., Carrivain, P., Victor, J.-M., Gadal, O., Bystricky, K. and Bancaud, A. (2013). High-throughput chromatin motion tracking in living yeast reveals the flexibility of the fiber throughout the genome. Genome Research *23*, 1829–1838.

Halverson, J. D., Lee, W. B., Grest, G. S., Grosberg, A. Y. and Kremer, K. (2011). Molecular dynamics simulation study of nonconcatenated ring polymers in a melt. II. Dynamics. The Journal of Chemical Physics *134*, 204905.

Hawkins, R. D., Hon, G. C., Yang, C., Antosiewicz-Bourget, J. E., Lee, L. K., Ngo, Q.-M., Klugman, S., Ching, K. A., Edsall, L. E., Ye, Z., Kuan, S., Yu, P., Liu, H., Zhang, X., Green, R. D., Lobanenkov, V. V., Stewart, R., Thomson, J. A. and Ren, B. (2011). Dynamic chromatin states in human ES cells reveal potential regulatory sequences and genes involved in pluripotency. Cell Research *21*, 1393–1409.

Higham, N. (2002). Accuracy and Stability of Numerical Algorithms. Other Titles in Applied Mathematics, Society for Industrial and Applied Mathematics.

Hiraoka, Y., Minden, J. S., Swedlow, J. R., Sedat, J. W. and Agard, D. A. (1989). Focal points for chromosome condensation and decondensation revealed by three-dimensional in vivo time-lapse microscopy. Nature *342*, 293–296.

Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A.-L., Bak, R. O., Li, C. H., Goldmann, J., Lajoie, B. R., Fan, Z. P., Sigova, A. A., Reddy, J., Borges-Rivera, D., Lee, T. I., Jaenisch, R., Porteus, M. H., Dekker, J. and Young, R. A. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science (New York, N.Y.) *351*, 1454–1458.

Ho, J. W. K., Jung, Y. L., Liu, T., Alver, B. H., Lee, S., Ikegami, K., Sohn, K.-A., Minoda, A., Tolstorukov, M. Y., Appert, A., Parker, S. C. J., Gu, T., Kundaje, A., Riddle, N. C., Bishop, E., Egelhofer, T. A., Hu, S. S., Alekseyenko, A. A., Rechtsteiner, A., Asker, D., Belsky, J. A., Bowman, S. K., Chen, Q. B., Chen, R. A.-J., Day, D. S., Dong, Y., Dose, A. C., Duan, X., Epstein, C. B., Ercan, S., Feingold, E. A., Ferrari, F., Garrigues, J. M., Gehlenborg, N., Good, P. J., Haseley, P., He, D., Herrmann, M., Hoffman, M. M., Jeffers, T. E., Kharchenko, P. V., Kolasinska-Zwierz, P., Kotwaliwale, C. V., Kumar, N., Langley, S. A., Larschan, E. N., Latorre, I., Libbrecht, M. W., Lin, X., Park, R., Pazin, M. J., Pham, H. N., Plachetka, A., Qin, B., Schwartz, Y. B., Shoresh, N., Stempor, P., Vielle, A., Wang, C., Whittle, C. M., Xue, H., Kingston, R. E., Kim, J. H., Bernstein, B. E., Dernburg, A. F., Pirrotta, V., Kuroda, M. I., Noble, W. S., Tullius, T. D., Kellis, M., MacAlpine, D. M., Strome, S., Elgin, S. C. R., Liu, X. S., Lieb, J. D., Ahringer, J., Karpen, G. H. and Park, P. J. (2014). Comparative analysis of metazoan chromatin organization. Nature *512*, 449–452.

Holwerda, S. and de Laat, W. (2012). Chromatin loops, gene positioning, and gene expression. Frontiers in Genetics *3*, 217.

Hu, M., Deng, K., Qin, Z., Dixon, J., Selvaraj, S., Fang, J., Ren, B. and Liu, J. S. (2013). Bayesian inference of spatial organizations of chromosomes. PLoS computational biology *9*, e1002893.

Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B. and Liu, J. S. (2012). HiCNorm : removing biases in Hi-C data via Poisson regression. Bioinformatics *28*, 3131–3133.

Hugouvieux, V., Axelos, M. A. and Kolb, M. (2008). Amphiphilic multiblock copolymers : From intramolecular pearl necklace to layered structures. Macromolecules *42*, 392–400.

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., Dekker, J. and Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature Methods *9*, 999–1003.

Imakaev, M. V., Fudenberg, G. and Mirny, L. A. (2015). Modeling chromosomes : Beyond pretty pictures. FEBS letters *589*, 3031–3036.

Isono, K., Endo, T. A., Ku, M., Yamada, D., Suzuki, R., Sharif, J., Ishikura, T., Toyoda, T., Bernstein, B. E. and Koseki, H. (2013). SAM domain polymerization links subnuclear clustering of PRC1 to gene silencing. Developmental Cell *26*, 565–577.

Jost, D., Carrivain, P., Cavalli, G. and Vaillant, C. (2014). Modeling epigenome folding : formation and dynamics of topologically associated chromatin domains. Nucleic Acids Research *42*, 9553–9561.

Jost, D. and Everaers, R. (2010). Prediction of RNA multiloop and pseudoknot conformations from a lattice-based, coarse-grain tertiary structure model. The Journal of Chemical Physics *132*, 095101.

Junier, I., Spill, Y. G., Marti-Renom, M. A., Beato, M. and le Dily, F. (2015). On the demultiplexing of chromosome capture conformation data. FEBS Lett *589*, 3005–3013.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. and Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol *30*, 90–98.

Kharchenko, P. V., Alekseyenko, A. A., Schwartz, Y. B., Minoda, A., Riddle, N. C., Ernst, J., Sabo, P. J., Larschan, E., Gorchakov, A. A., Gu, T., Linder-Basso, D., Plachetka, A., Shanower, G., Tolstorukov, M. Y., Luquette, L. J., Xi, R., Jung, Y. L., Park, R. W., Bishop, E. P., Canfield, T. K., Sandstrom, R., Thurman, R. E., MacAlpine, D. M., Stamatoyannopoulos, J. A., Kellis, M., Elgin, S. C. R., Kuroda, M. I., Pirrotta, V., Karpen, G. H. and Park, P. J. (2011). Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature *471*, 480–485.

Knight, P. A. and Ruiz, D. (2012). A fast algorithm for matrix balancing. IMA Journal of Numerical Analysis *1*.

Lajoie, B. R., Dekker, J. and Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis : practical guidelines. Methods (San Diego, Calif.) *72*, 65–75.

Le Dily, F., Baù, D., Pohl, A., Vicent, G. P., Serra, F., Soronellas, D., Castellano, G., Wright, R. H., Ballare, C., Filion, G., Marti-Renom, M. A. and Beato, M. (2014). Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. Genes & Development *28*, 2151–2162.

Lesne, A., Riposo, J., Roger, P., Cournac, A. and Mozziconacci, J. (2014). 3D genome reconstruction from chromosomal contacts. Nat Methods *11*, 1141–3.

Lévy-Leduc, C., Delattre, M., Mary-Huard, T. and Robin, S. (2014). Two-dimensional segmentation for analyzing Hi-C data. Bioinformatics *30*, i386–i392.

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

Lo, S. M., Follmer, N. E., Lengsfeld, B. M., Madamba, E. V., Seong, S., Grau, D. J. and Francis, N. J. (2012). A bridging model for persistence of a polycomb group protein complex through DNA replication in vitro. Molecular Cell *46*, 784–796.

Lowenstein, M. G., Goddard, T. D. and Sedat, J. W. (2004). Long-range interphase chromosome organization in Drosophila : a study using color barcoded fluorescence in situ hybridization and structural clustering analysis. Molecular Biology of the Cell *15*, 5678–5692.

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*, 1012–1025.

Meister, P., Mango, S. E. and Gasser, S. M. (2011). Locking the genome : nuclear organization and cell fate. Curr Opin Genet Dev *21*, 167–174.

Milo, R., Jorgensen, P., Moran, U., Weber, G. and Springer, M. (2010). BioNumbers–the database of key numbers in molecular and cell biology. Nucleic Acids Research *38*, D750–753.

Mirny, L. A. (2011). The fractal globule as a model of chromatin architecture in the cell. Chromosome Research *19*, 37–51.

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A. and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*, 59–64.

Naumova, N. and Dekker, J. (2010). Integrating one-dimensional and three-dimensional maps of genomes. Journal of Cell Science *123*, 1979–1988.

Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B. R., Mirny, L. A. and Dekker, J. (2013). Organization of the mitotic chromosome. Science (New York, N.Y.) *342*, 948–953.

Nazarov, L. I., Tamm, M. V., Avetisov, V. A. and Nechaev, S. K. (2015). A statistical model of intra-chromosome contact maps. Soft Matter *11*, 1019–1025.

Noordermeer, D., Leleu, M., Splinter, E., Rougemont, J., Laat, W. D. and Duboule, D. (2011). The dynamic architecture of Hox gene clusters. Science *334*, 222–225.

Olarte-Plata, J. D., Haddad, N., Vaillant, C. and Jost, D. (2016). The folding landscape of the epigenome. Physical Biology *13*, 026001.

Peng, C., Fu, L.-Y., Dong, P.-F., Deng, Z.-L., Li, J.-X., Wang, X.-T. and Zhang, H.-Y. (2013). The sequencing bias relaxed characteristics of Hi-C derived data and implications for chromatin 3D modeling. Nucleic Acids Research *41*, e183.

Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. K., Ong, C.-T., Hookway, T. A., Guo, C., Sun, Y., Bland, M. J., Wagstaff, W., Dalton, S., McDevitt, T. C., Sen, R., Dekker, J., Taylor, J. and Corces, V. G. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281–1295.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (2007). Numerical Recipes 3rd Edition : The Art of Scientific Computing. 3 edition edition, Cambridge University Press, Cambridge, UK ; New York.

Pujadas, E. and Feinberg, A. P. (2012). Regulated noise in the epigenetic landscape of development and disease. Cell *148*, 1123–1131.

Ramalho, T., Selig, M., Gerland, U. and Enblin, T. A. (2013). Simulation of stochastic network dynamics via entropic matching. Physical Review E *87*.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. and Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–80.

Rapkin, L. M., Anchel, D. R. P., Li, R. and Bazett-Jones, D. P. (2012). A view of the chromatin landscape. Micron (Oxford, England : 1993) *43*, 150–158.

Rosa, A. and Everaers, R. (2008). Structure and Dynamics of Interphase Chromosomes. PLOS Comput Biol *4*, e1000153.

Rosa, A. and Everaers, R. (2014). Ring Polymers in the Melt State : The Physics of Crumpling. Physical Review Letters *112*, 118302.

Rosa, A. and Zimmer, C. (2014). Computational models of large-scale genome architecture. International Review of Cell and Molecular Biology *307*, 275–349.

Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., Giraut, L., Després, B., Drevensek, S., Barneche, F., Dèrozier, S., Brunaud, V., Aubourg, S., Schnittger, A., Bowler, C., Martin-Magniette, M.-L., Robin, S., Caboche, M. and Colot, V. (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. The EMBO journal *30*, 1928–1938.

Rousseau, M., Fraser, J., Ferraiuolo, M. A., Dostie, J. and Blanchette, M. (2011). Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. BMC Bioinformatics *12*, 414.

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*, E6456–E6465.

Sauria, M. E. G., Phillips-Cremins, J. E., Corces, V. G. and Taylor, J. (2015). HiFive : a tool suite for easy and efficient HiC and 5C data analysis. Genome Biol *16*, 237.

Scherble, J., Thomann, R., Béla and Mülhaupt, R. (2001). Formation of CdS nanoclusters in phase-separated poly(2-hydroxyethyl methacrylate)-l-polyisobutylene amphiphilic conetworks. Journal of Polymer Science Part B : Polymer Physics *39*, 1429–1436.

Sexton, T. and Cavalli, G. (2015). The role of chromosome domains in shaping the functional genome. Cell *160*, 1049–59.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. and Cavalli, G. (2012). Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. Cell *148*, 458–472.

Shampine, L. and Reichelt, M. (1997). The MATLAB ODE Suite. SIAM Journal on Scientific Computing *18*, 1–22.

Sharma, R., Jost, D., Kind, J., Gómez-Saldivar, G., van Steensel, B., Askjaer, P., Vaillant, C. and Meister, P. (2014). Differential spatial and structural organization of the X chromosome underlies dosage compensation in C. elegans. Genes Dev *28*, 2591–2596.

Shin, H., Shi, Y., Dai, C., Tjong, H., Gong, K., Alber, F. and Zhou, X. J. (2016). TopDom : an efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Res *44*, e70.

Sofueva, S., Yaffe, E., Chan, W.-C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., Pollard, S. M., Schroth, G. P., Tanay, A. and Hadjur, S. (2013). Cohesin-mediated interactions organize chromosomal domain architecture. The EMBO journal *32*, 3119–3129.

Soshnev, A. A., Josefowicz, S. Z. and Allis, C. D. (2016). Greater Than the Sum of Parts : Complexity of the Dynamic Epigenome. Molecular Cell *62*, 681–694.

Tamm, M. V., Nazarov, L. I., Gavrilov, A. A. and Chertovich, A. V. (2015). Anomalous diffusion in fractal globules. Physical Review Letters *114*.

Tanizawa, H., Iwasaki, O., Tanaka, A., Capizzi, J. R., Wickramasinghe, P., Lee, M., Fu, Z. and Noma, K.-i. (2010). Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. Nucleic Acids Research *38*, 8164–8177.

Tark-Dame, M., Jerabek, H., Manders, E. M. M., Heermann, D. W. and Driel, R. v. (2014). Depletion of the Chromatin Looping Proteins CTCF and Cohesin Causes Chromatin Compaction : Insight into Chromatin Folding by Polymer Modelling. PLOS Comput Biol *10*, e1003877.

Timoshenko, E. G., Kuznetsov, Y. A. and Dawson, K. A. (1998). Conformational transitions of heteropolymers in dilute solutions. Physical Review E *57*, 6801–6814.

Trieu, T. and Cheng, J. (2014). Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. Nucleic Acids Research *42*, e52–e52.

Ulianov, S. V., Khrameeva, E. E., Gavrilov, A. A., Flyamer, I. M., Kos, P., Mikhaleva, E. A., Penin, A. A., Logacheva, M. D., Imakaev, M. V., Chertovich, A., Gelfand, M. S., Shevelyov, Y. Y. and Razin, S. V. (2016). Active chromatin and transcription play a key

role in chromosome partitioning into topologically associating domains. Genome Research *26*, 70–84.

Van Steensel, B. (2011). Chromatin : constructing the big picture. The EMBO Journal *30*, 1885–1895.

Varoquaux, N., Ay, F., Noble, W. S. and Vert, J.-P. (2014). A statistical approach for inferring the 3D structure of the genome. Bioinformatics *30*, i26–33.

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A. and Hadjur, S. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Reports *10*, 1297–1309.

Wang, S., Xu, J. and Zeng, J. (2015). Inferential modeling of 3D chromatin structure. Nucleic Acids Res *43*, e54.

Weinreb, C. and Raphael, B. J. (2015). Identification of hierarchical chromatin domains. Bioinformatics *btv485*.

Yaffe, E. and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nature Genetics *43*, 1059–1065.

Zhang, T., Cooper, S. and Brockdorff, N. (2015). The interplay of histone modifications - writers that read. EMBO reports *16*, 1467–1481.

Zhang, Z., Li, G., Toh, K.-C. and Sung, W.-K. (2013). 3D chromosome modeling with semi-definite programming and Hi-C data. Journal of Computational Biology : A Journal of Computational Molecular Cell Biology *20*, 831–846.

Zullo, J. M., Demarco, I. A., Piqué-Regi, R., Gaffney, D. J., Epstein, C. B., Spooner, C. J., Luperchio, T. R., Bernstein, B. E., Pritchard, J. K., Reddy, K. L. and Singh, H. (2012). DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina. Cell *149*, 1474–1487.